ArchSummit
全 球 架 构 师 峰 会

AIGC驱动的3D场景理解及医学图像解析

香港中文大学（深圳）
助理教授
李镇博士

1 讲者介绍

李镇
助理教授
FNII 助理院长

- 香港大学博士（师从余益州教授），芝加哥大学访问学者（师从许锦波教授）
- 香港中文大学（深圳）理工学院/未来智联网络研究院 助理院长/教授，校长青年学者
- 香港中文大学（深圳）深度比特实验室主任
  博士后：1名，博士生：8名，研究生：2名

人才荣誉

- CASP12 接触图预测全球冠军，并作为AlphaFoldV1的基线方案
- PLOS CB 2018年创新与突破奖 （一年一例）
- 中国科协 2019年青年托举人才
- 2022年05月CAMEO蛋白打分月度第一，2022 SemancticKITTI分割竞赛第一，2023 CVPR HOI4D 分割竞赛第一，2018全球气象预测大赛第一，ICCV 2022 Urban3D第二等

科研学术

- 主持国家自然科学基金青年项目1项
- 主持深港A类项目“深度学习辅助的RNA蛋白结构预测以及蛋白高亲和性RNA设计 ”（300万）
- CCF-腾讯犀牛鸟2019优秀奖，2022年犀牛鸟专项
- 参与科技部国家重点研发项目
- 合作牵头国家自然科学基金重点项目，合作牵头粤深联合基金重点项目

目录

案例简介

- 300字以内进行概括性的案例介绍（突出亮点、案例独特性等）

随着AIGC和ChatGPT等生成模型的迅速发展，我们探索出AIGC驱动的3D场景理解以及医疗场景的分析，并通过一系列自研的算法和工具，对AIGC算法辅助的下游应用进行了深入地研究，从3D场景的自动稠密描述，到室内场景的视觉定位，再到3D视觉驱动的高保真说话人脸生成，并推广到AIGC辅助的医疗场景的解析，我们均进行了深入地探讨。在本次分享中，我们将会从3D场景描述和定位，3D说话人脸生成，生成图片辅助的肠胃镜图片解析等方面，详解介绍我们应用方案的架构设计与工程实践，同时也会基于我们的经验分享在使用AIGC驱动的3D场景理解和医疗图像理解过程中的思考和对未来AIGC演进的展望。

目录

InstanceRefer: Cooperative Holistic Understanding for Visual Grounding on Point Clouds through Instance Multi-level Contextual Referring

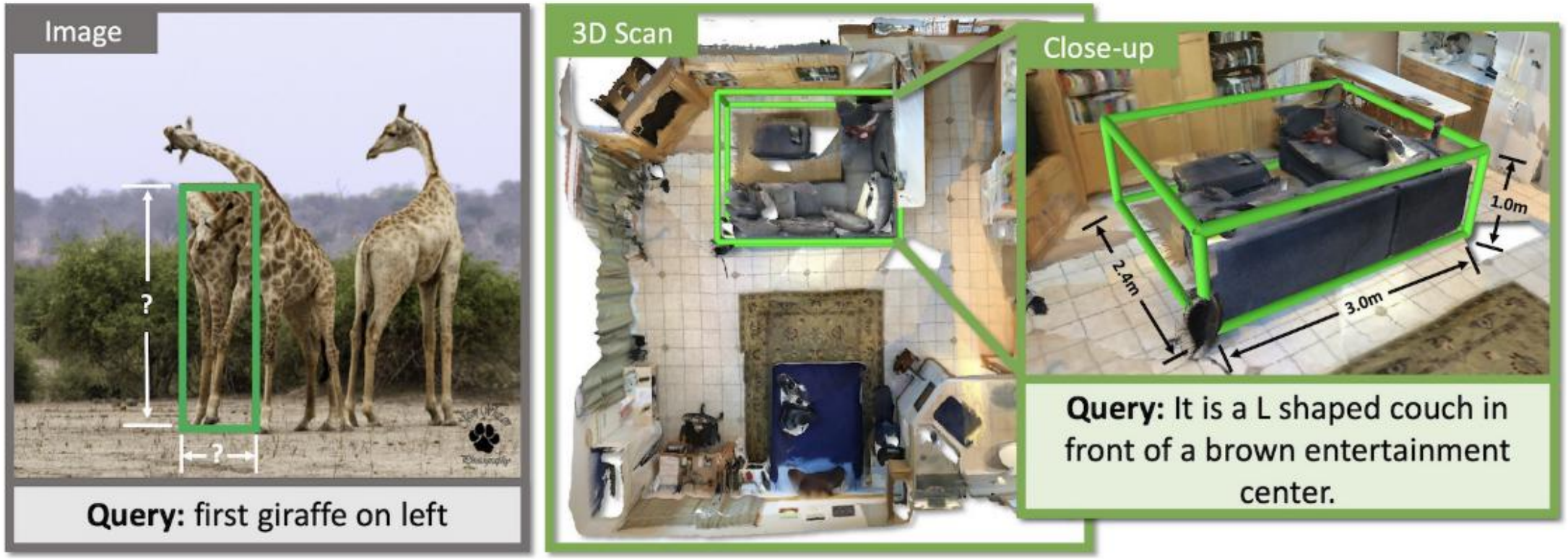
Zhihao Yuan [1,†], Xu Yan [1,†], Yinghong Liao [1], Ruimao Zhang [1]
Sheng Wang [2], Zhen Li [1,], and Shuguang Cui [1]*

[1] The Chinese University of Hong Kong (Shenzhen),
Shenzhen Research Institute of Big Data
[2] CryoEM Center, Southern University of Science and Technology

Background

Visual Grounding:

Visual grounding (VG) aims at localizing the desired objects or areas in an image or a 3D scene based on an object-related linguistic query

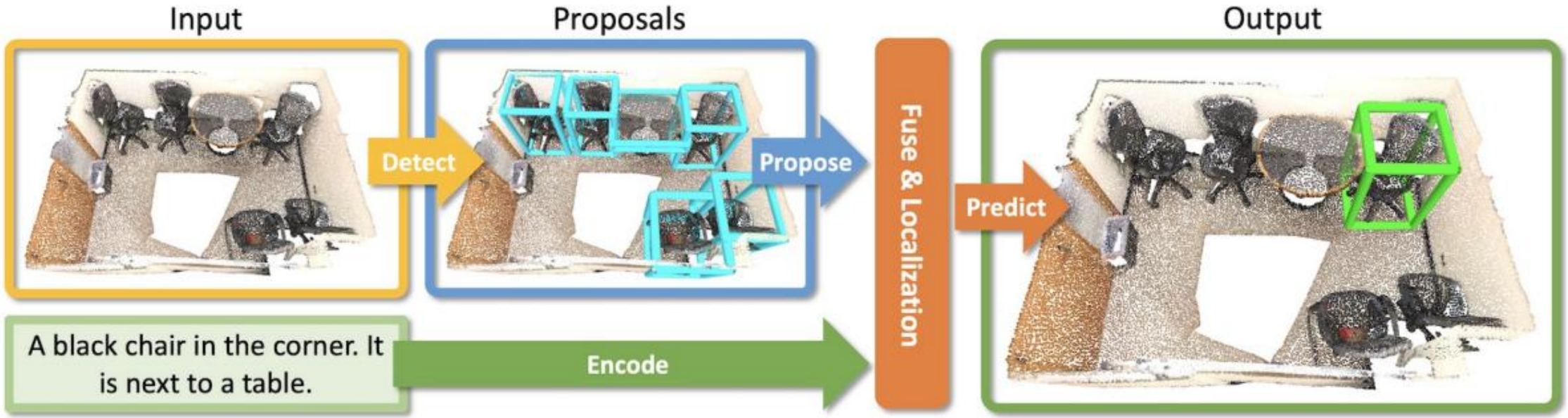


ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language

Background

ScanRefer:

1. Exploiting object detection to generate proposal candidates;
2. Localize described object by fusing language features into candidates.



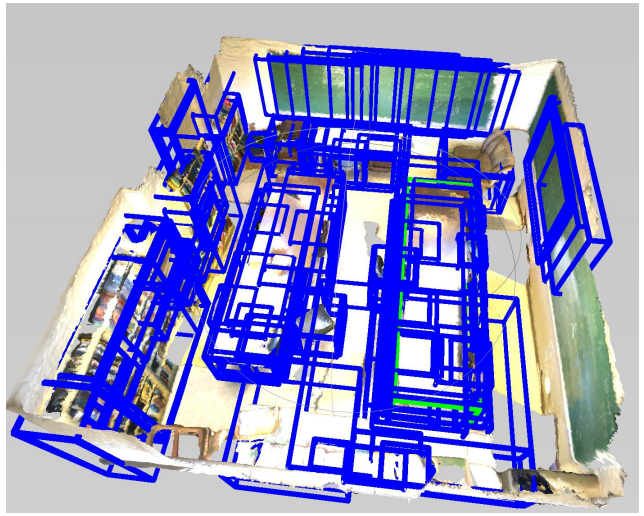
ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language

Background

ScanRefer:

Cons:

1. The object proposals in the large 3D scene are usually redundant;
2. The appearance and attribute information is not sufficiently captured;
3. The relations among proposals and the ones between proposals and background are not fully studied.

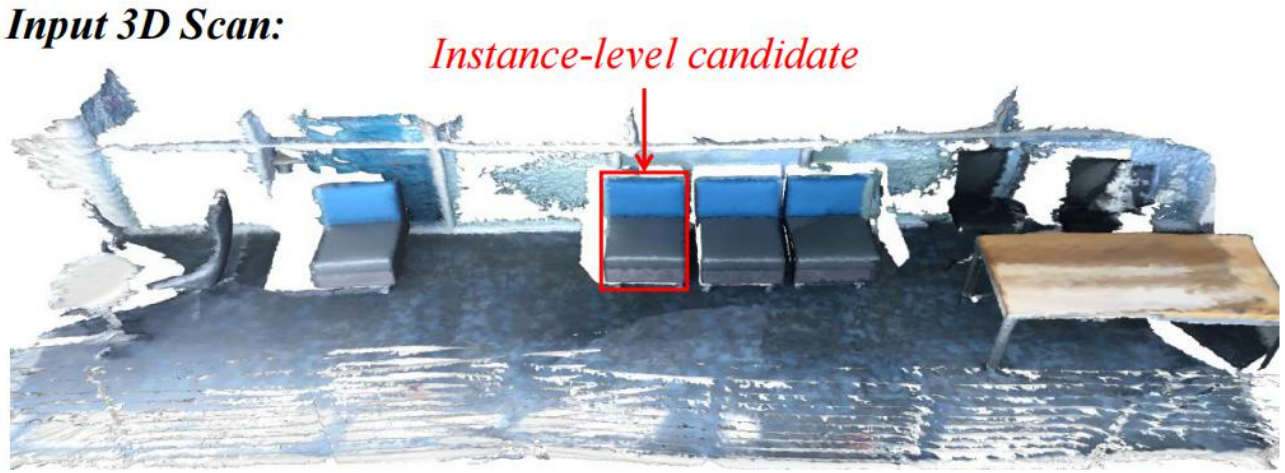


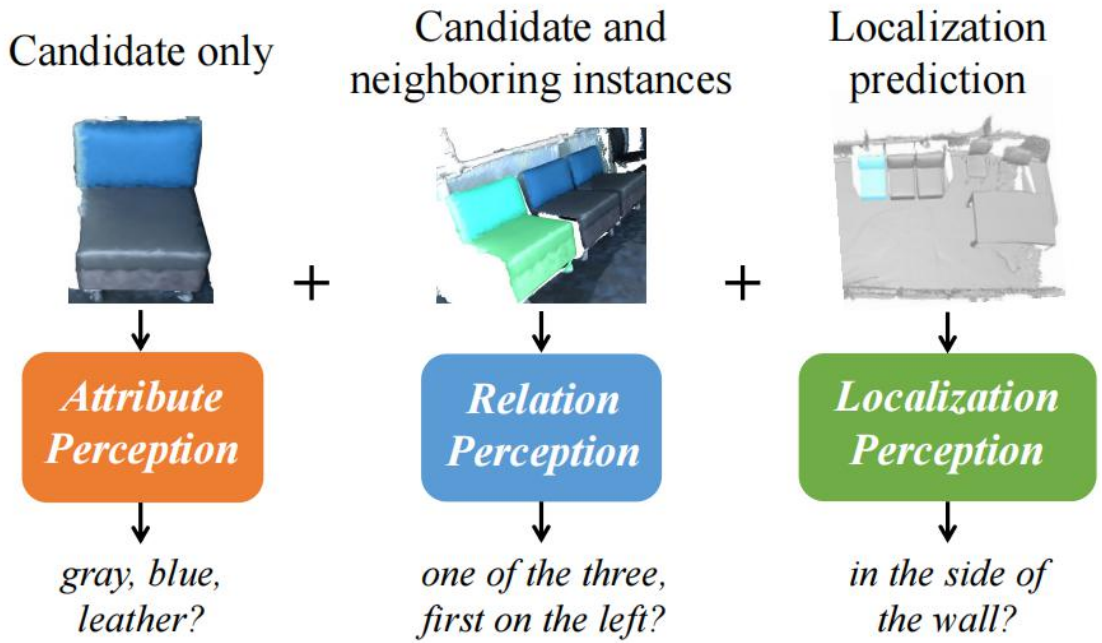
- *ScanRefer generates 114 possible candidates after filtering proposals by their objectness scores;*
- *Each proposal's feature is generated by the detection framework;*
- *There is no relation reasoning among proposals*

Method

InstanceRefer:

1. Instance-level candidate representation (small number);
2. Multi-level contextual inference (attribute, objects' relation and environment).

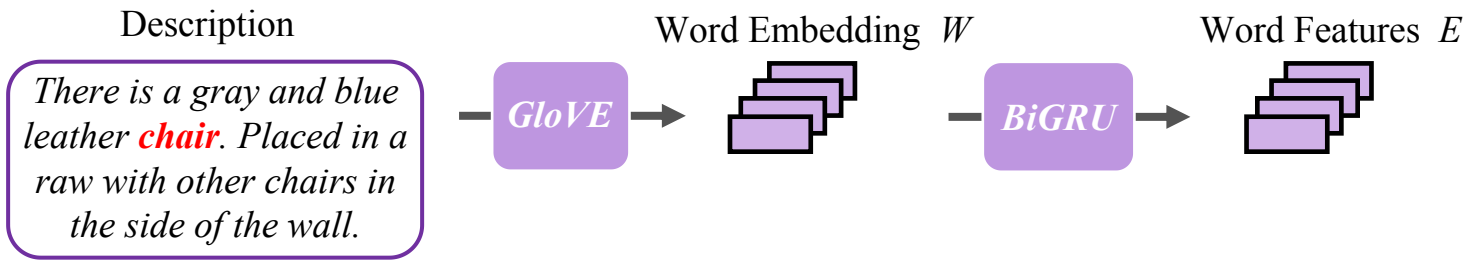




Method

InstanceRefer Architecture:

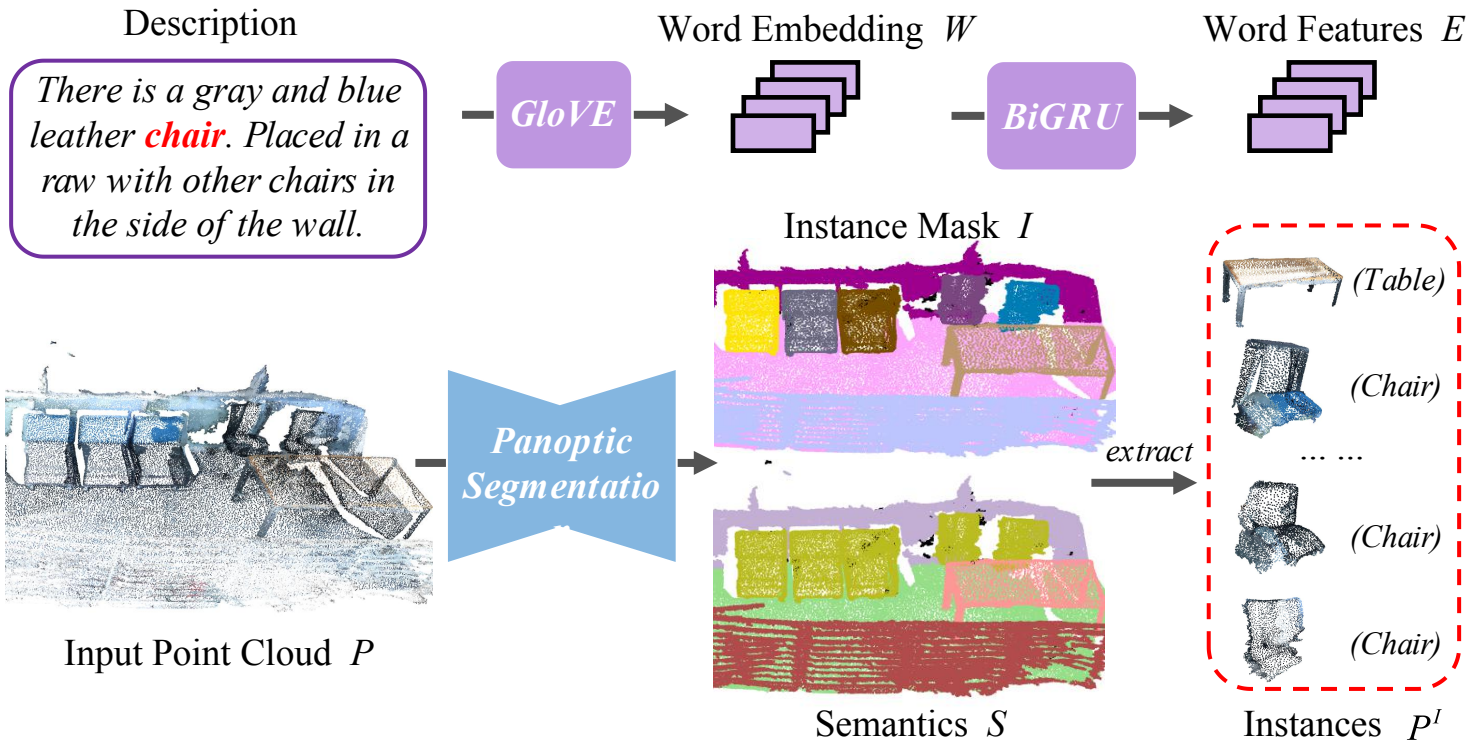
Language feature encoding (the same as ScanRefer).



Method

InstanceRefer Architecture:

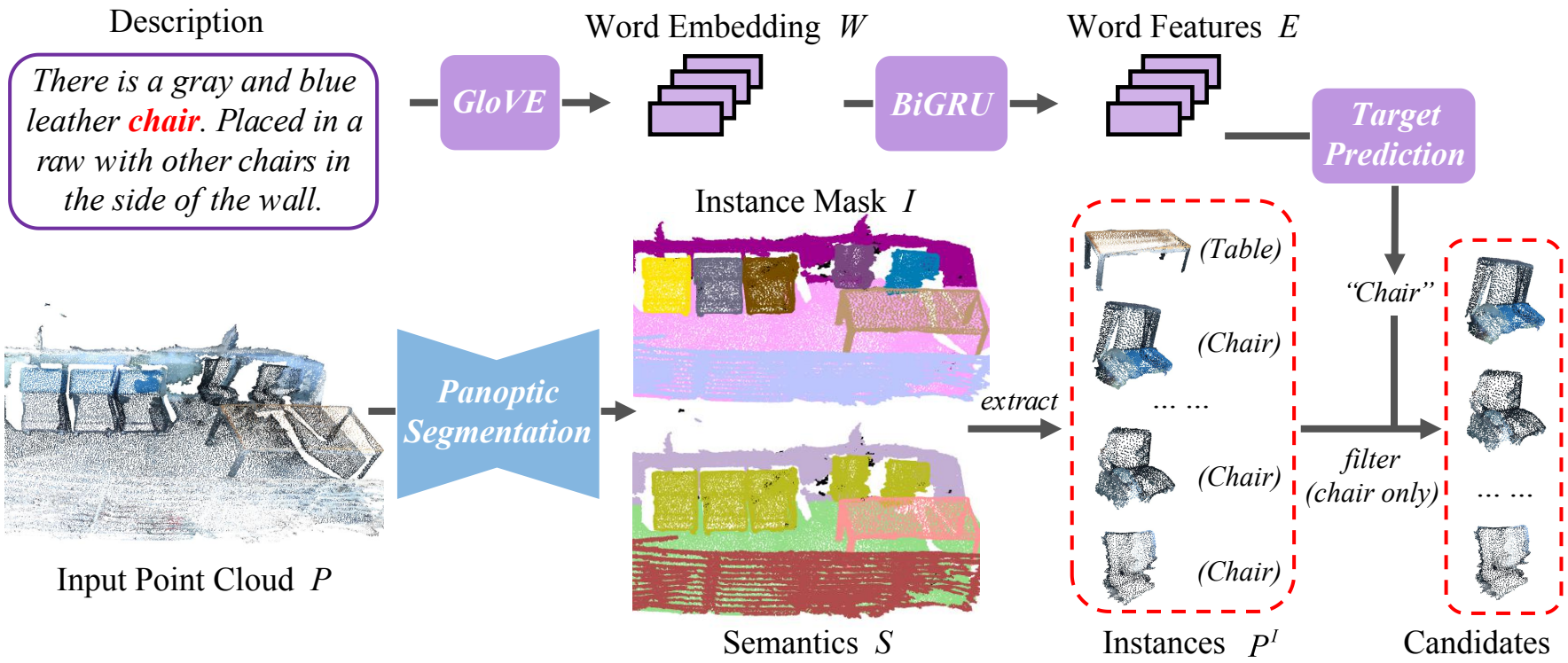
Extracting instances through panoptic segmentation (predict instance and semantics).



Method

InstanceRefer Architecture:

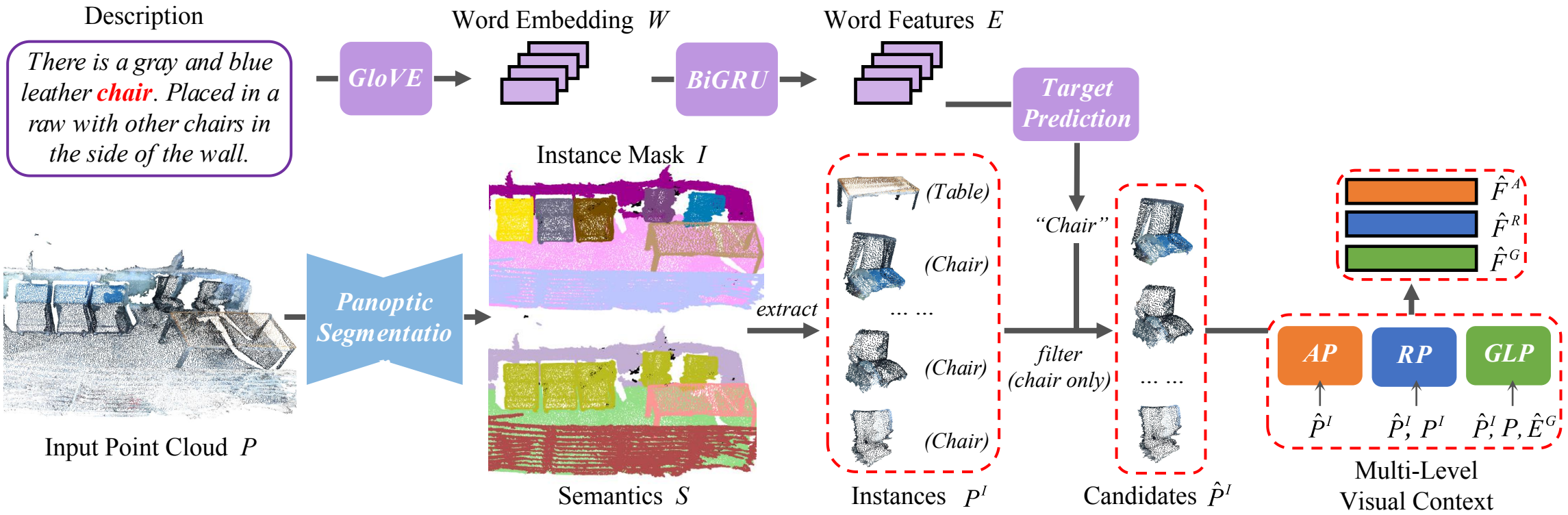
Eliminating irrelative instances by the target category (inferred by language).



Method

InstanceRefer Architecture:

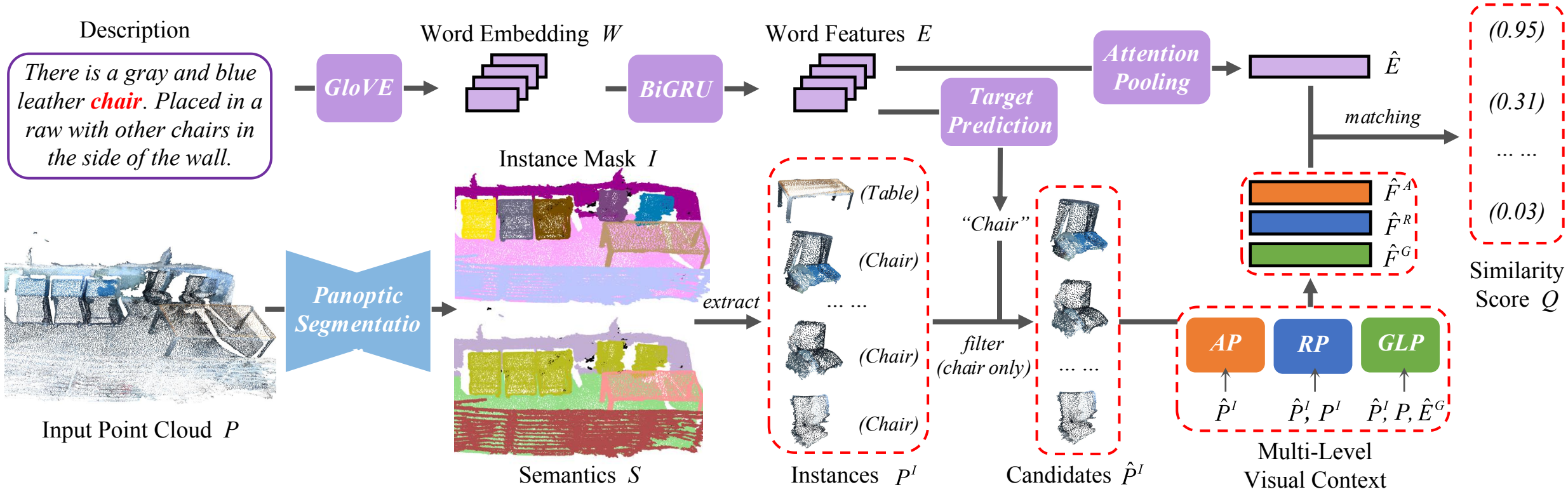
Generating visual feature of each candidate by multi-level referring (three novel modules are proposed).



Method

InstanceRefer Architecture:

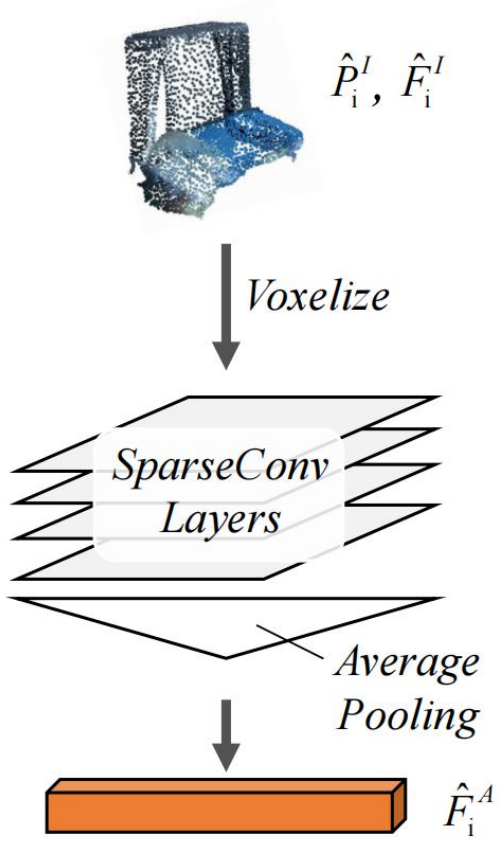
Scoring each candidate matching language and visual features (the candidate with the largest score will be regarded as output).



Method

Specific Modules:

(a) Attribute Perception (AP) Module.

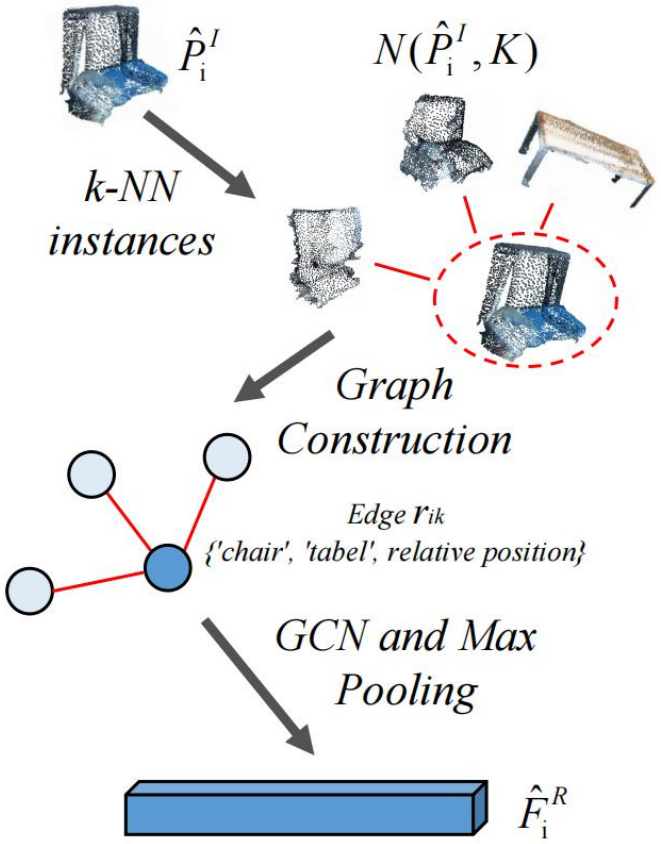


- *It construct a four-layer Sparse Convolution (SparseConv) as the feature extractor;*
- *After an average pooling, the global attribute perception feature is obtained.*

Method

Specific Modules:

(b) Relation Perception (RP) Module.



- *It uses k-nearest neighbors to construct a graph, where nodes' features are their semantics obtained by panoptic segmentation and edges are consisted of their semantics and relative position;*
- *Dynamic graph convolution network (DGCNN) is exploited to update the node's feature*

$$r_{ik} = \texttt{MLP}([\mathcal{C}(\hat{P}_i^I) - \mathcal{C}(P_k^I); S_i^I; S_k^I])$$

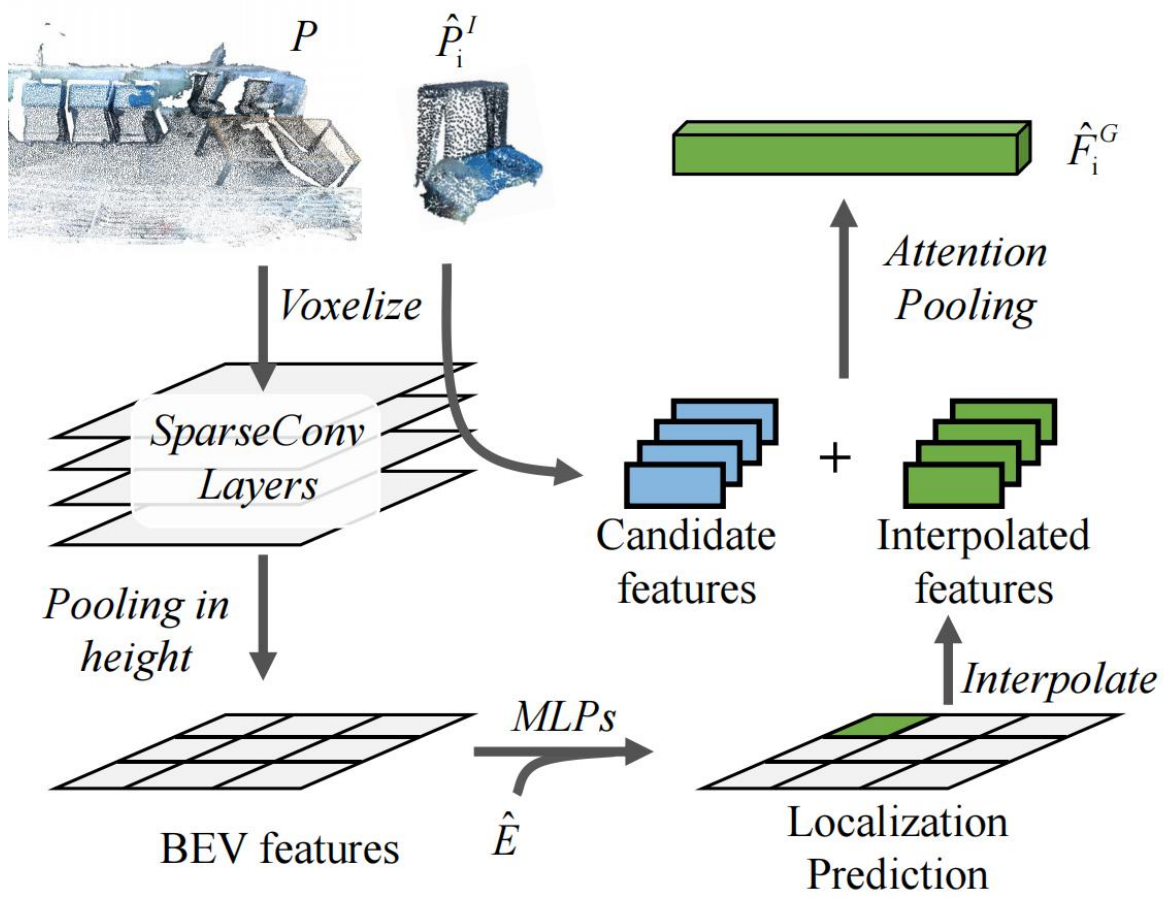
$$h_{ik} = \texttt{MLP}([P_k^I; S_k^I]),\ \forall\ P_k^I \in \mathcal{N}(\hat{P}_i^I, K)$$

$$\hat{F}_i^R = \texttt{MaxPool}(\{r_{ik} \odot h_{ik}\}_{k=1}^K)$$

Method

Specific Modules:

(c) Global Localization Perception (GLP) Module.



- *It uses SparseConv layers with height-pooling to generate a 3 ×3 bird-eyes-view (BEV) plane;*
- *By combining language feature, it predicts which grid the target object is located in;*
- *It interpolates probabilities and generates the global perception features by merging features from AP module.*

Method

Specific Modules:

(d) Matching Module

- A naive version by using Cosine similarity;
- An enhance version by using modular co-attention from MCAN [1].

(e) Contrastive Objective

$$L_{\text{mat}} = -\log \frac{\sum_{i=1}^{L} \exp(Q_i^+)}{\sum_{i=1}^{L} \exp(Q_i^+) + \sum_{i=L+1}^{M} \exp(Q_i^-)}$$

where Q+ and Q− denote the scores of positive and negative pairs.

[1] Deep modular co-attention networks for visual question answering

Results

ScanRefer:

Table 1. Comparison of localization results. TGNN replaces the original GRU layers with pre-trained BERT to extract language features. Our method follows TGNN's strategy of only taking coordinates (Geo) and color information (RGB) as input, while results of ScanRefer on benchmark are obtained by using additional normals (Nor) and multi-view features from a pre-trained 2D feature extractor. Scores for the test set are obtained from the online evaluation. Only the published methods are compared. Accessed on March 18, 2021.

Method	Input	Unique		Multiple		Overall	
		Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
Validation results							
SCRC [9]	RGB image	24.03	9.22	17.77	5.97	18.70	6.45
One-stage [34]	RGB image	29.32	22.82	18.72	6.49	20.38	9.04
ScanRefer [2]	Geo + RGB	65.00	43.31	30.63	19.75	37.30	24.32
TGNN [10]	Geo + RGB	64.50	53.01	27.01	21.88	34.29	27.92
TGNN[10]+BERT [5]	Geo + RGB	68.61	56.80	29.84	23.18	37.37	29.70
IntanceRefer (Ours)	Geo + RGB	**77.45**	**66.83**	**31.27**	**24.77**	**40.23**	**32.93**
Test results (ScanRefer benchmark)							
ScanRefer [2]	Geo+Nor+Multiview	68.59	43.53	34.88	20.97	42.44	26.03
TGNN [10]	Geo + RGB	62.40	53.30	28.20	21.30	35.90	28.50
TGNN [10]+BERT [5]	Geo + RGB	68.34	58.94	33.12	25.26	41.02	32.81
IntanceRefer (Ours)	Geo + RGB	**77.82**	**66.69**	**34.57**	**26.88**	**44.27**	**35.80**

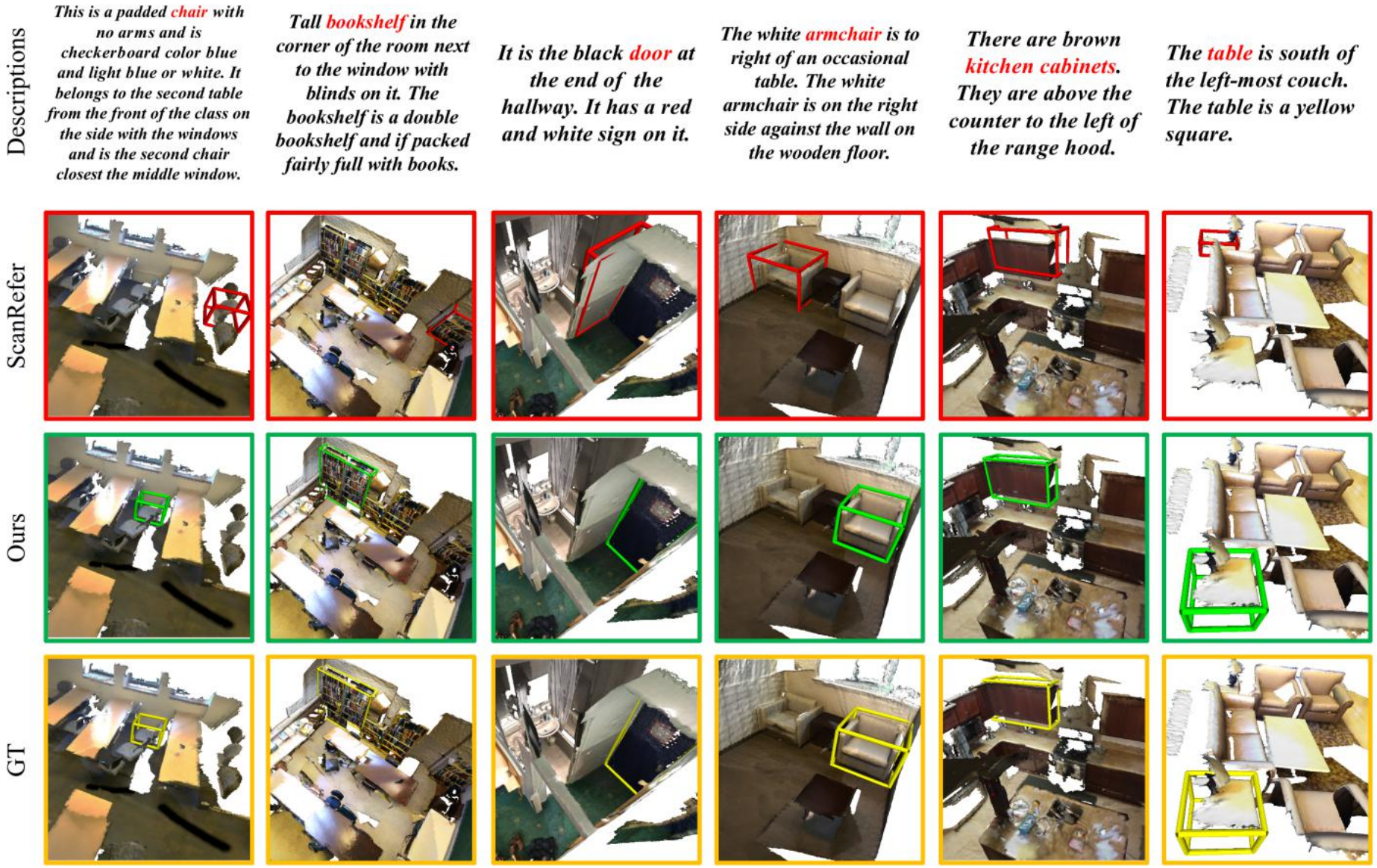
Results

ScanRefer Benchmark

This table lists the benchmark results for the ScanRefer Localization Benchmark scenario.

Method	Info	Unique acc@0.25IoU	Unique acc@0.5IoU	Multiple acc@0.25IoU	Multiple acc@0.5IoU	Overall acc@0.25IoU	Overall acc@0.5IoU
InstanceRefer	P	**0.7782** 1	**0.6669** 1	0.3457 4	0.2688 2	0.4427 3	**0.3580** 1
Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Zhen Li*, Shuguang Cui: InstanceRefer: Cooperative Holistic Understanding for Visual Grounding on Point Clouds through Instance Multi-level Contextual Referring. arXiv preprint							
DetrRefer&Trick&Aug		0.7576 2	0.5515 4	**0.4224** 1	**0.2933** 1	**0.4976** 1	0.3512 2
PointGroup_MCAN		0.7510 3	0.6397 2	0.3271 6	0.2535 3	0.4222 5	0.3401 3
TGNN		0.6834 6	0.5894 3	0.3312 5	0.2526 4	0.4102 6	0.3281 4
Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, Tyng-Luh Liu: Text-Guided Graph Neural Network for Referring 3D Instance Segmentation. AAAI 2021							
SRGA		0.7494 4	0.5128 5	0.3631 2	0.2218 5	0.4497 2	0.2871 5
ScanRefer	P	0.6859 5	0.4353 6	0.3488 3	0.2097 6	0.4244 4	0.2603 6
Dave Zhenyu Chen, Angel X. Chang, Matthias Nießner: ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language. 16th European Conference on Computer Vision (ECCV), 2020							
ScanRefer Baseline		0.6422 7	0.4196 7	0.3090 7	0.1832 7	0.3837 7	0.2362 7

Results


Descriptions
This is a padded chair with no arms and is checkerboard color blue and light blue or white. It belongs to the second table from the front of the class on the side with the windows and is the second chair closest the middle window.
Tall bookshelf in the corner of the room next to the window with blinds on it. The bookshelf is a double bookshelf and if packed fairly full with books.
It is the black door at the end of the hallway. It has a red and white sign on it.
The white armchair is to right of an occasional table. The white armchair is on the right side against the wall on the wooden floor.
There are brown kitchen cabinets. They are above the counter to the left of the range hood.
The table is south of the left-most couch. The table is a yellow square.
ScanRefer
Ours
GT

Results

Nr3D/Sr3D:

Table 2. Comparison of referring object identification on Nr3D and Sr3D datasets. Here 'easy' and 'hard' determined by whether there are more than two instances of the same object class in the scene. 'view-dependent' and 'view-independent' determined by whether the referring expression depending on camera view.

Dataset	Method	Easy	Hard	View-dep.	View-indep.	Overall
Nr3D	ReferIt3DNet [1]	$43.6\% \pm 0.8\%$	$27.9\% \pm 0.7\%$	$32.5\% \pm 0.7\%$	$37.1\% \pm 0.8\%$	$35.6\% \pm 0.7\%$
	TGNN [10]	$44.2\% \pm 0.4\%$	$30.6\% \pm 0.2\%$	$\mathbf{35.8\% \pm 0.2\%}$	$38.0\% \pm 0.3\%$	$37.3\% \pm 0.3\%$
	IntanceRefer (Ours)	$\mathbf{46.0\% \pm 0.5\%}$	$\mathbf{31.8\% \pm 0.4\%}$	$34.5\% \pm 0.6\%$	$\mathbf{41.9\% \pm 0.4\%}$	$\mathbf{38.8\% \pm 0.4\%}$
Sr3D	ReferIt3DNet [1]	$44.7\% \pm 0.1\%$	$31.5\% \pm 0.4\%$	$39.2\% \pm 1.0\%$	$40.8\% \pm 0.1\%$	$40.8\% \pm 0.2\%$
	TGNN [10]	$48.5\% \pm 0.2\%$	$36.9\% \pm 0.5\%$	$\mathbf{45.8\% \pm 1.1\%}$	$45.0\% \pm 0.2\%$	$45.0\% \pm 0.2\%$
	IntanceRefer (Ours)	$\mathbf{51.1\% \pm 0.2\%}$	$\mathbf{40.5\% \pm 0.3\%}$	$45.4\% \pm 0.9\%$	$\mathbf{48.1\% \pm 0.3\%}$	$\mathbf{48.0\% \pm 0.3\%}$

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

InstanceRefer: Cooperative Holistic Understanding for Visual Grounding on Point Clouds through Instance Multi-level Contextual Referring

Thanks for watching !

Zhihao Yuan [1,†], Xu Yan [1,†], Yinghong Liao [1], Ruimao Zhang [1]
Sheng Wang [2], Zhen Li [1,], and Shuguang Cui [1]*

[1] The Chinese University of Hong Kong (Shenzhen),
Shenzhen Research Institute of Big Data
[2] CryoEM Center, Southern University of Science and Technology

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

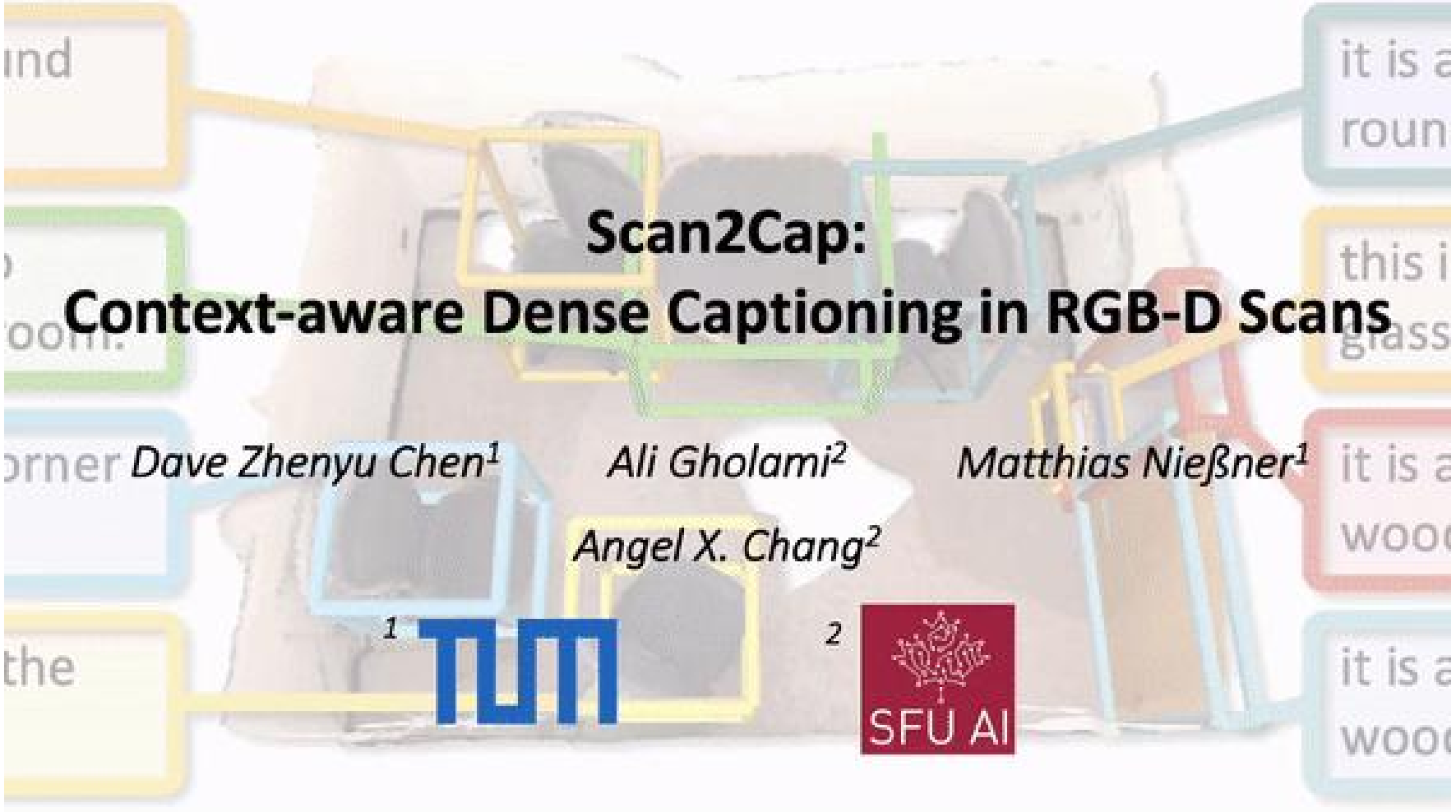
X-Trans2Cap: Cross-Modal Knowledge Transfer using Transformer for 3D Dense Captioning

Zhihao Yuan[1,†], Xu Yan[1,†], Yinghong Liao[1], Yao Guo[2], Guanbin Li[3], Shuguang Cui[1], Zhen Li[1,]*

[1] The Chinese University of Hong Kong (Shenzhen),
The Future Network of Intelligence Institute,
Shenzhen Research Institute of Big Data,
[2] Shanghai Jiao Tong University, [3] Sun Yat-sen University

Background

Task Description (3D Dense Captioning)

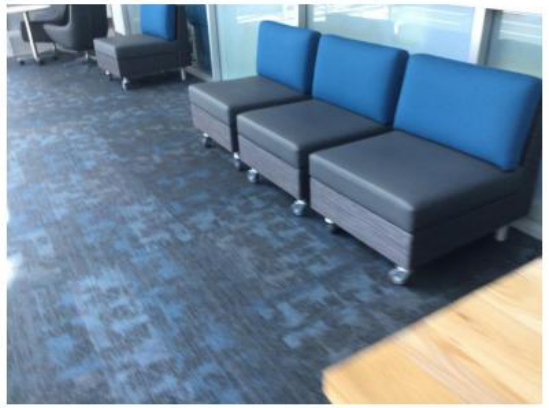


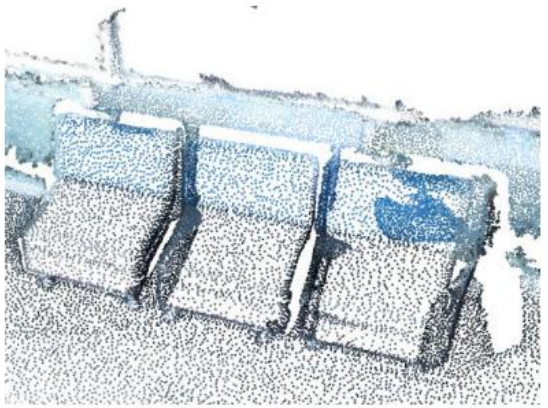
Scan2Cap: Context-aware Dense Captioning in RGB-D Scans Dave

Background

Limitations

- The object representations in Scan2Cap are defective since they are solely learned from sparse 3D point clouds, thus failing to provide strong texture and color information compared with the ones generated from 2D images.
- It requires the extra 2D input in both training and inference phases. However, the extra 2D information is usually computation intensive and unavailable during inference.

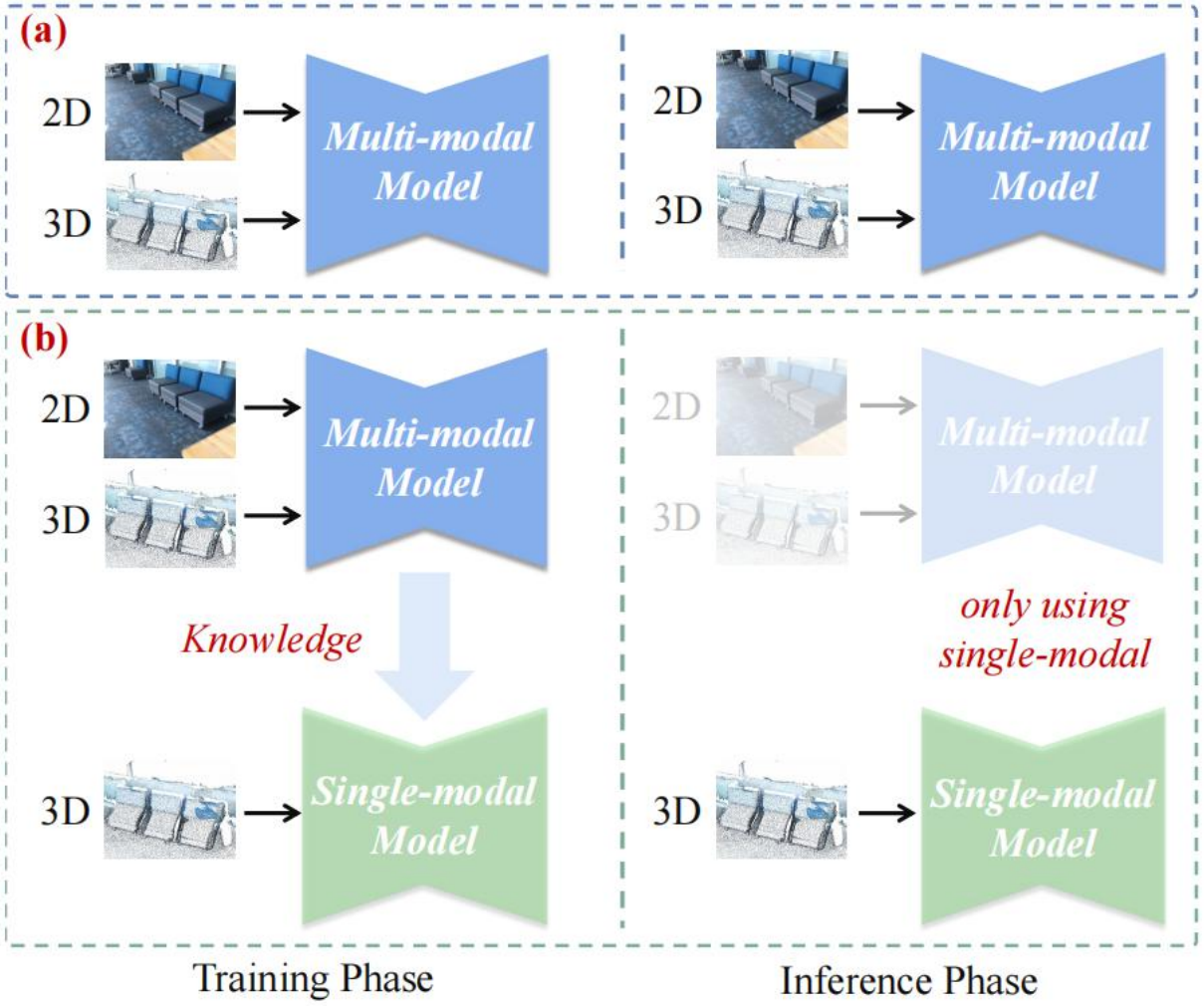




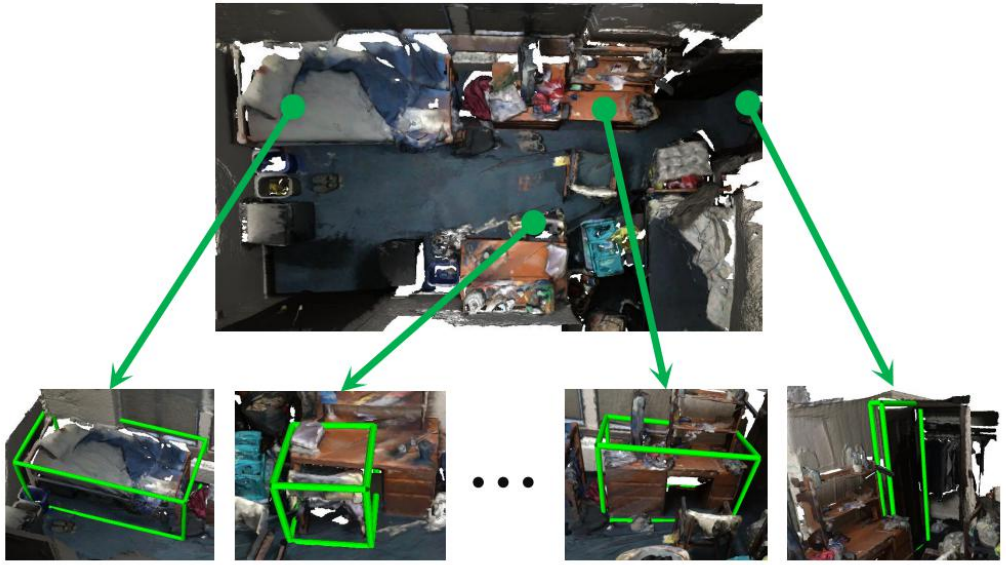
X-Trans2Cap

Motivation

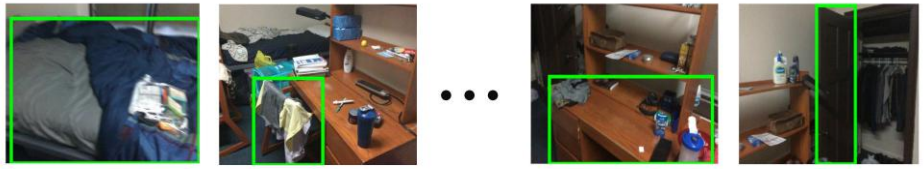
- We propose a Cross-Modal Knowledge Transfer framework on 3D dense captioning task.
- During the training phase, the teacher network exploits auxiliary 2D modality and guides the student network that only takes point clouds as input through the feature consistency constraints.
- A more faithful caption can be generated only using point clouds during the inference.



2D and 3D Inputs

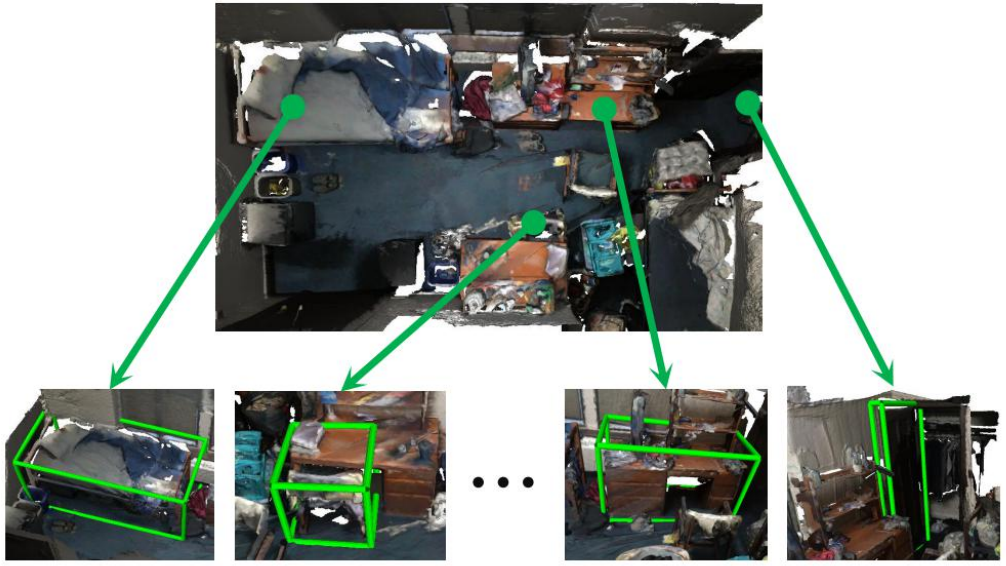


3D Proposals

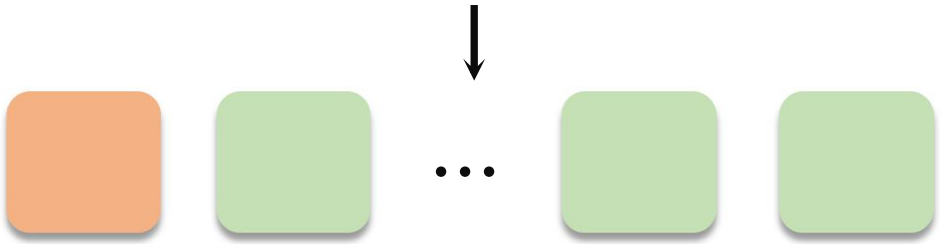


2D Proposals

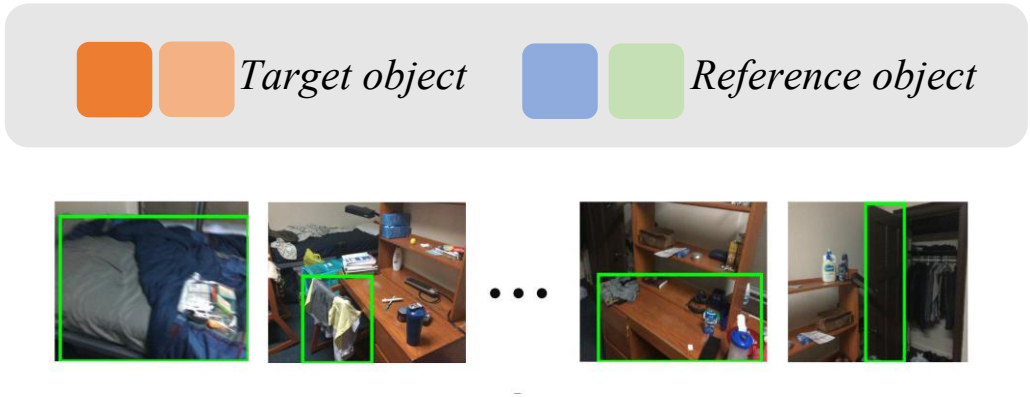
2D and 3D Inputs



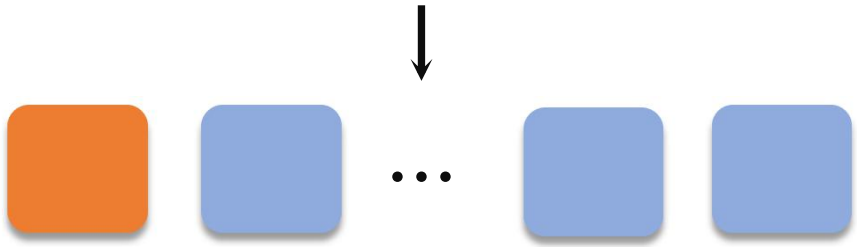
3D Proposals



3D Modal Inputs



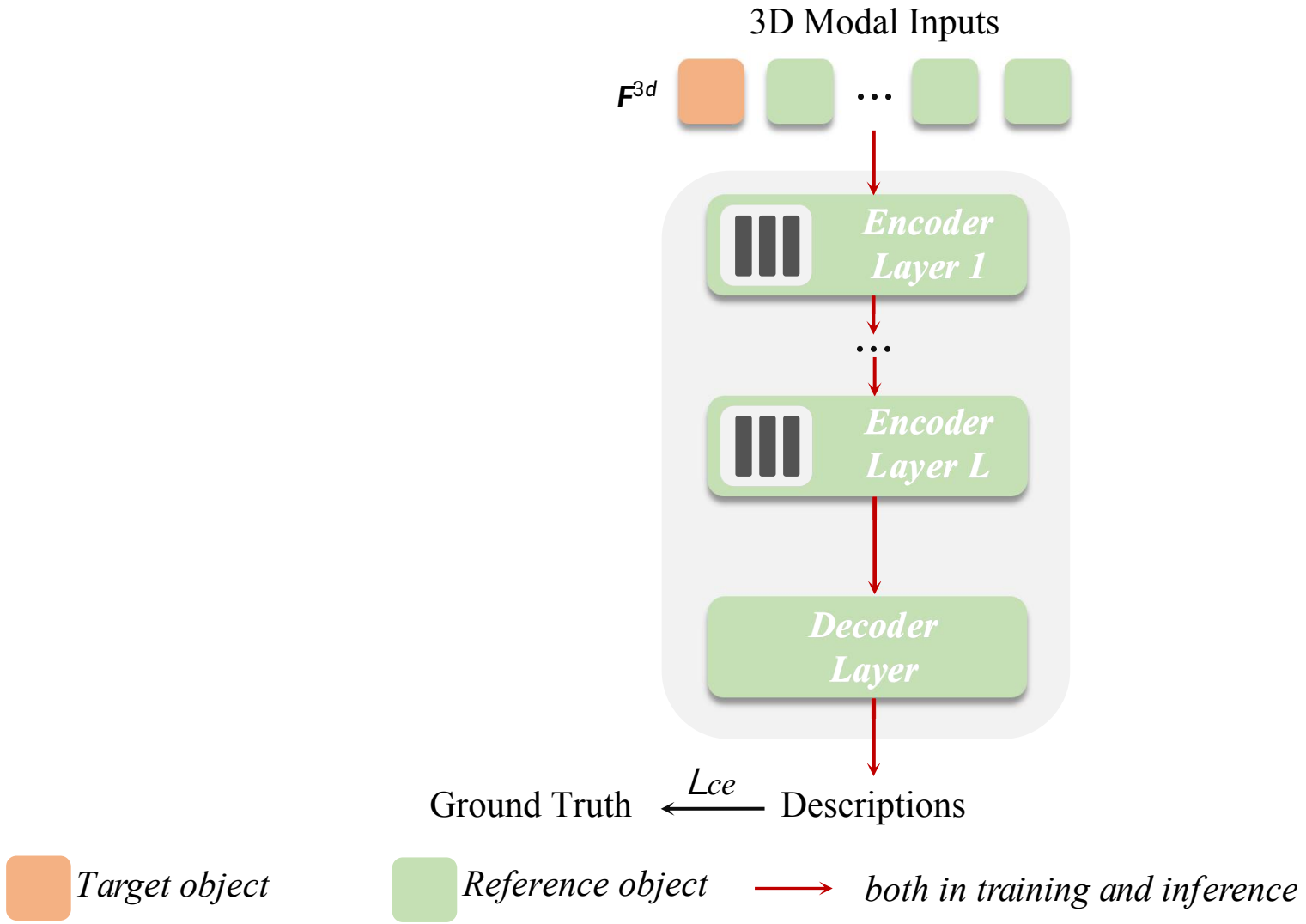
2D Proposals



Multi-Modal Inputs

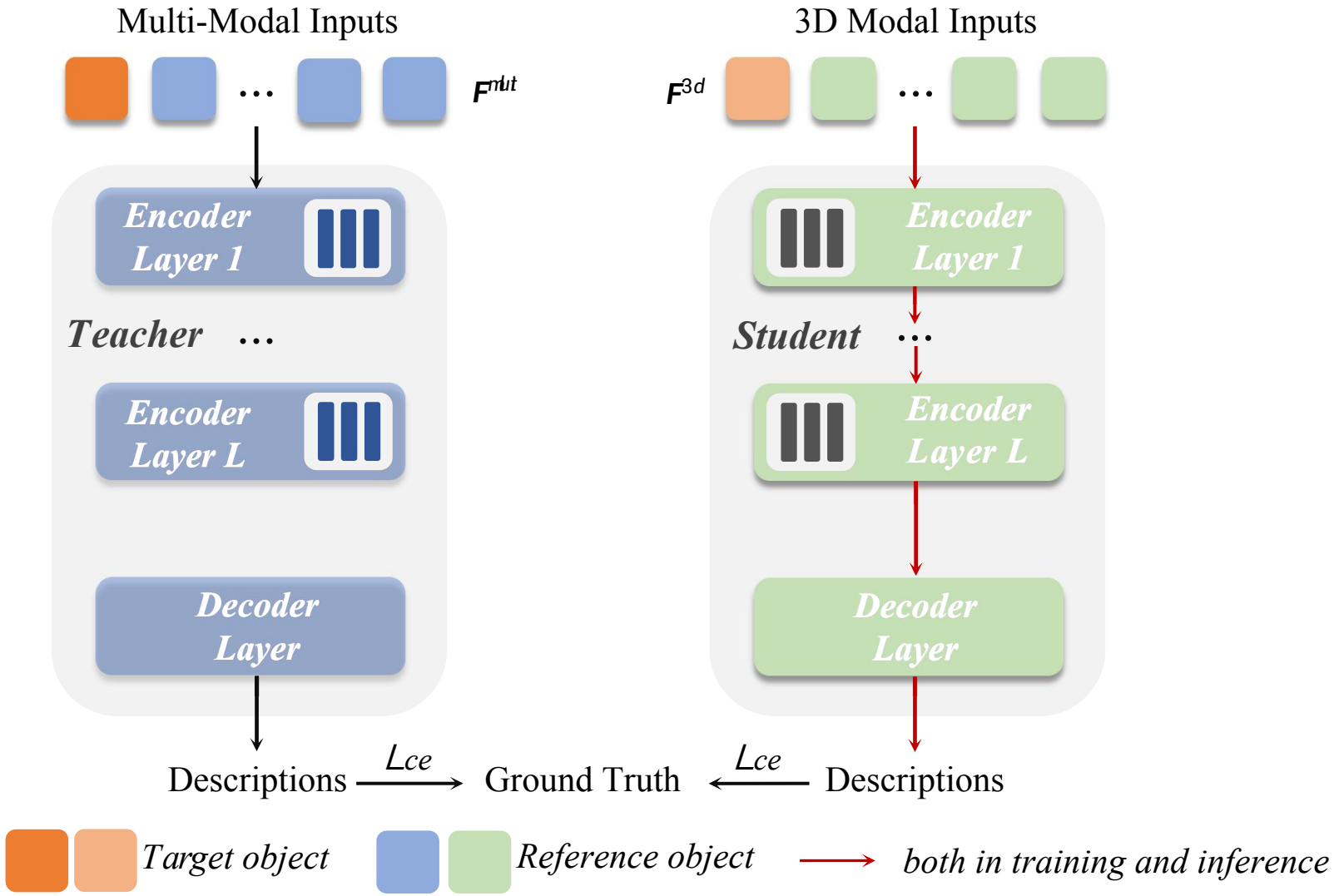
X-Trans2Cap

Architecture



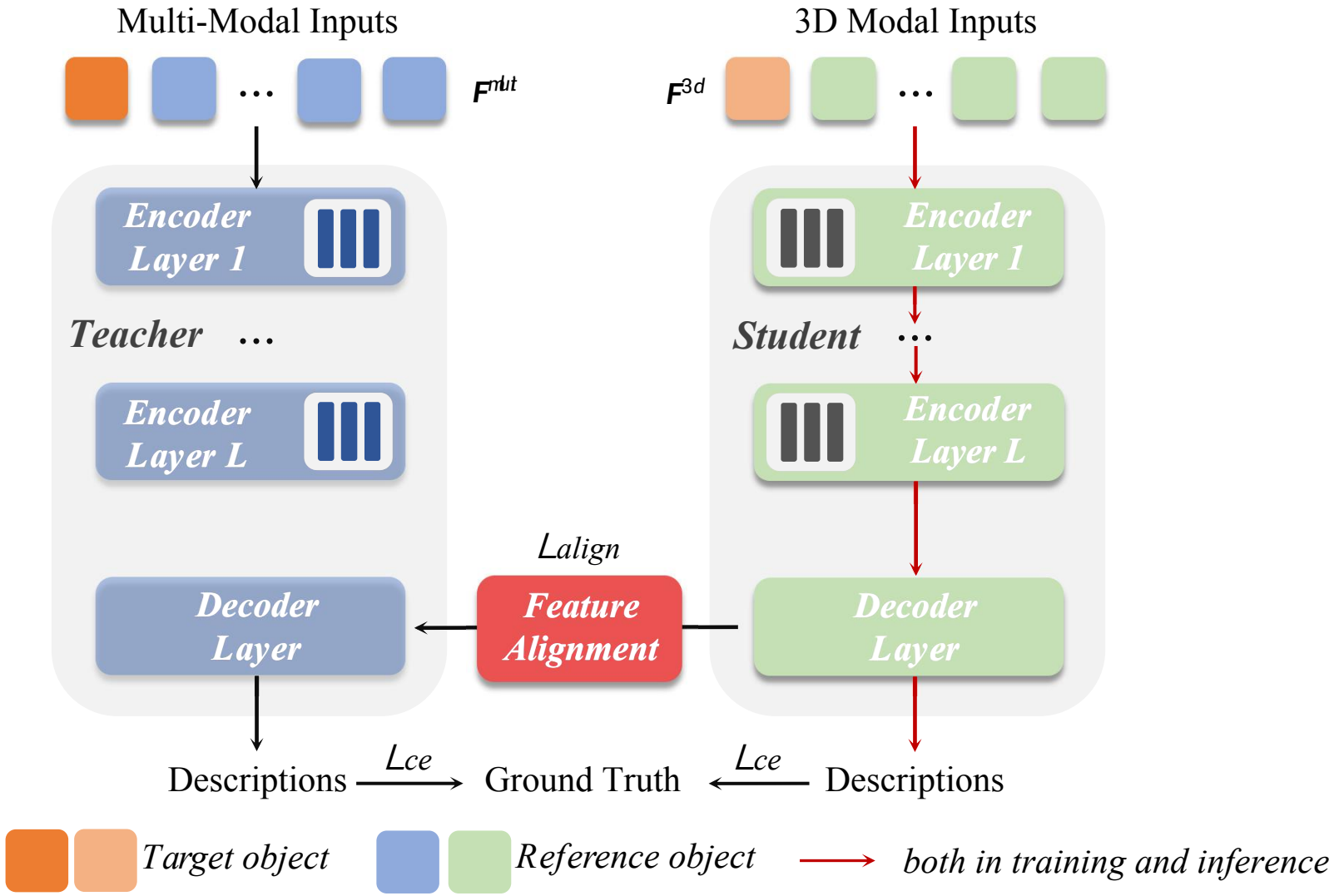
X-Trans2Cap

Architecture



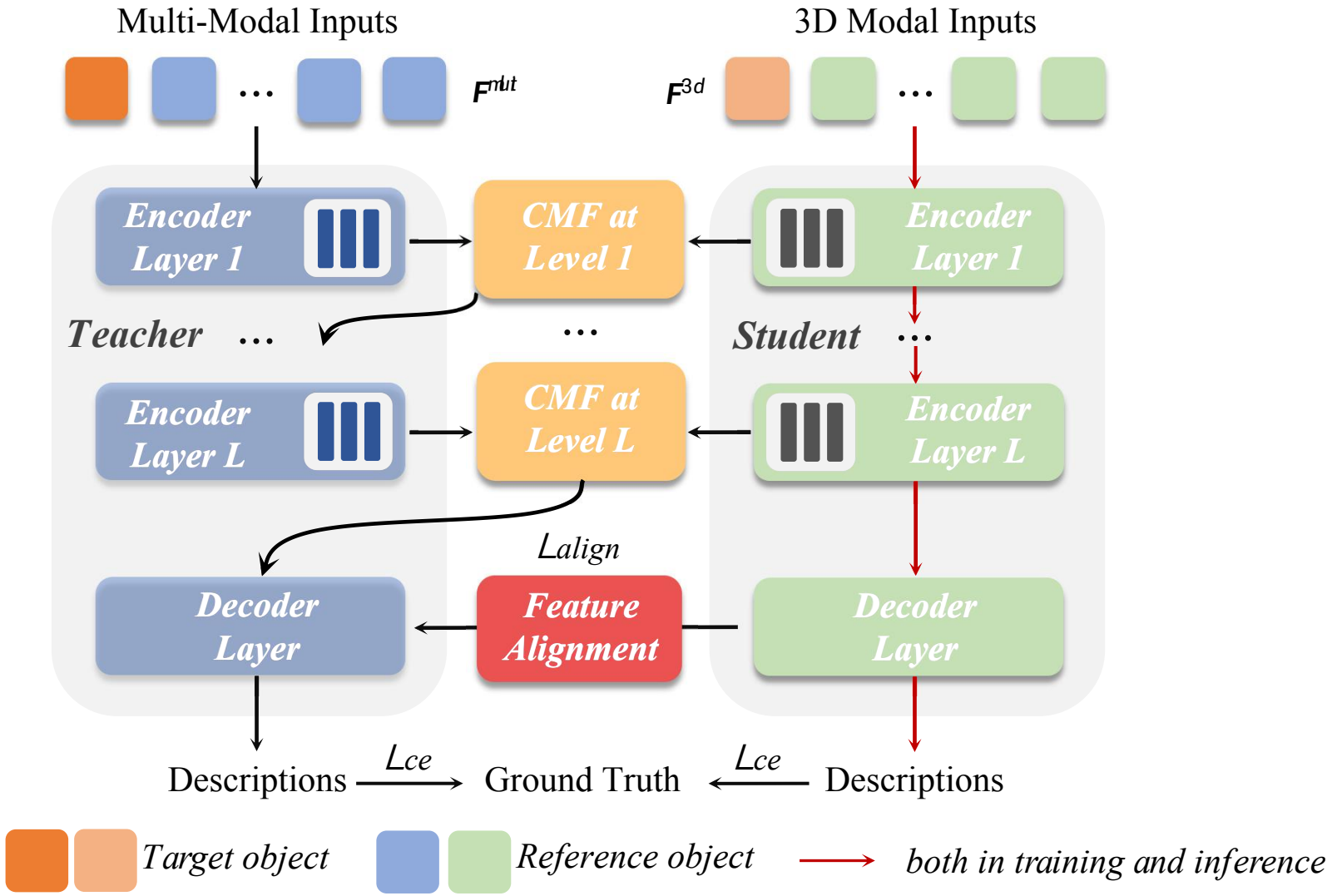
X-Trans2Cap

Architecture

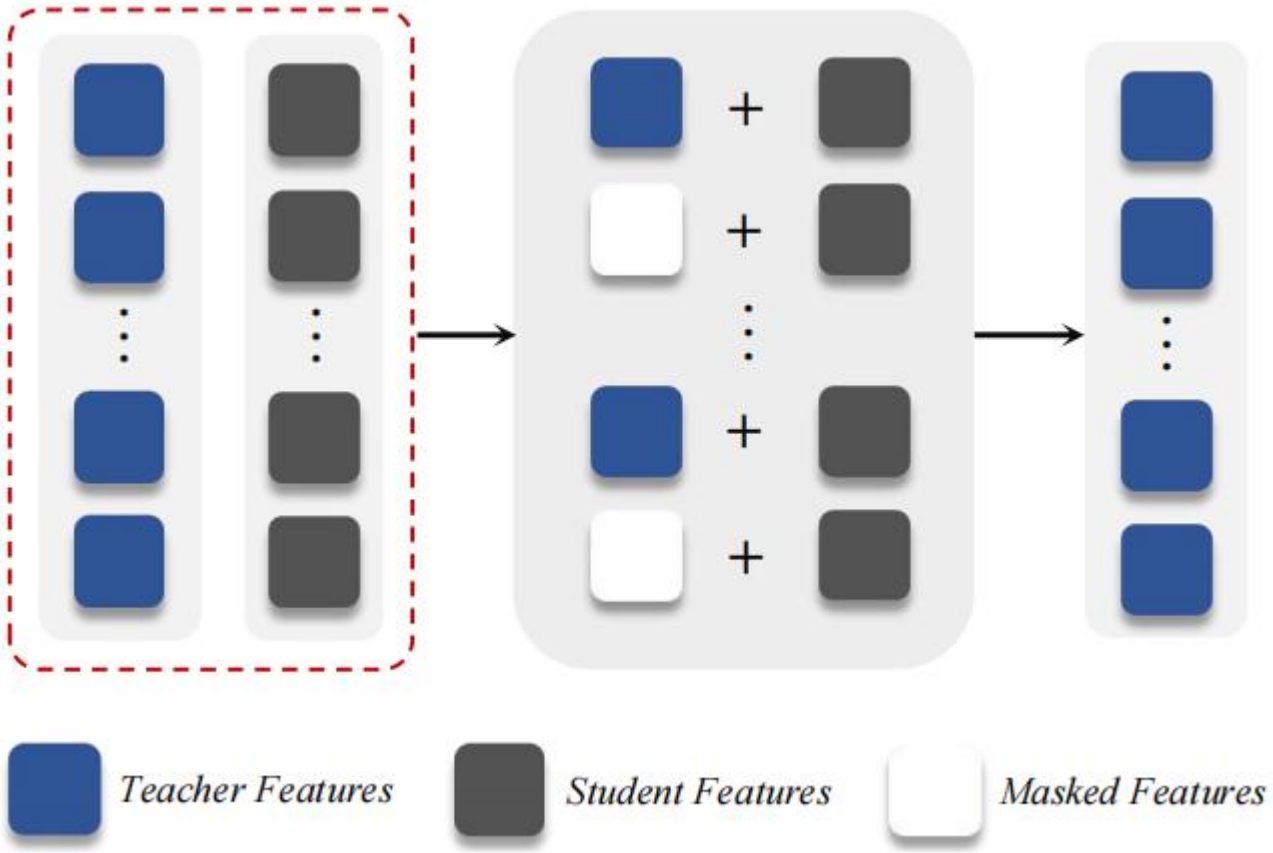


X-Trans2Cap

Architecture



Cross-Modal Fusion (CMF) Module



Experiments

3D Dense Captioning with Gound Truth Proposals (Nr3D and ScanRefer)

Method	Extra 2D	ScanRefer				Nr3D			
		C	B-4	M	R	C	B-4	M	R
Scan2Cap [9]	✗	65.79	38.54	28.81	61.93	63.36	32.07	28.92	64.56
Scan2Cap (Inst)	✗	64.44	36.89	28.42	60.42	61.89	32.02	28.88	64.17
TransCap	✗	75.75	42.06	28.82	62.62	70.60	35.99	29.04	66.00
$\mathcal{X}$-Trans2Cap	✗	87.09	44.12	30.67	64.37	80.02	37.90	30.48	67.64
$\mathcal{X}$-Trans2Cap (C)	✗	**89.46**	**44.46**	**30.71**	**64.55**	**81.44**	**39.08**	**30.79**	**68.15**
Scan2Cap [9]	✓	67.95	41.49	29.23	63.66	64.13	32.98	29.75	65.24
Scan2Cap (Inst)	✓	70.04	41.57	29.67	64.10	64.00	33.19	29.53	65.29
TransCap	✓	88.72	44.24	30.95	64.70	77.55	37.25	30.63	67.43
$\mathcal{X}$-Trans2Cap	✓	89.73	44.25	31.00	64.50	85.38	39.52	31.23	68.18
$\mathcal{X}$-Trans2Cap (C)	✓	**106.11**	**49.07**	**32.25**	**65.54**	**85.40**	**40.51**	**31.36**	**68.84**

Experiments

3D Dense Captioning with Detection Proposals (Nr3D and ScanRefer)

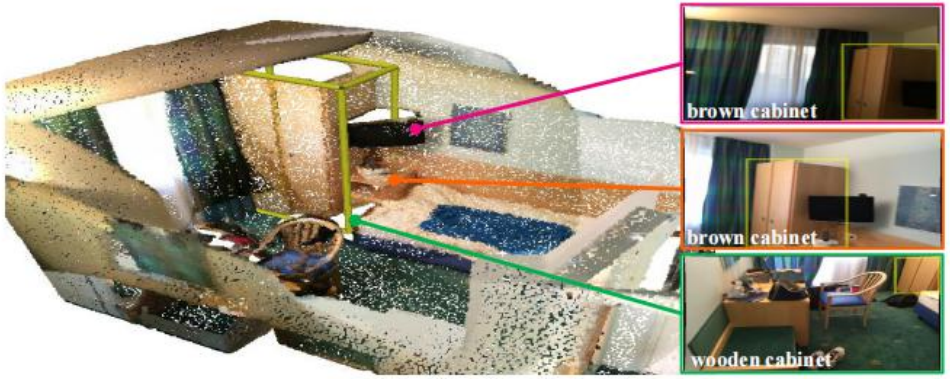
Method	Extra 2D	Proposals	C@0.25	B-4@0.25	M@0.25	R@0.25	C@0.5	B-4@0.5	M@0.5	R@0.5	mAP@0.5
Scan2cap [9]	✗	VoteNet	50.71	33.01	25.47	53.60	33.53	21.58	21.04	43.03	32.46
TransCap	✗	VoteNet	55.36	32.46	25.64	53.19	40.08	22.86	21.72	44.04	33.34
$\mathcal{X}$-Trans2Cap	✗	VoteNet	**58.81**	**34.17**	**25.81**	**54.10**	**41.52**	**23.83**	**21.90**	**44.97**	**34.68**
2D-3D Proj. [50]	✓	Mask R-CNN	18.29	10.27	16.67	33.63	8.31	2.31	12.54	25.93	10.50
3D-2D Proj. [50]	✓	VoteNet	19.73	17.86	19.83	40.68	11.47	8.56	15.73	31.65	31.83
Scan2cap [9]	✓	VoteNet	56.82	34.18	26.29	**55.27**	39.08	23.32	21.97	44.78	32.21
TransCap	✓	VoteNet	60.04	35.04	26.27	54.46	43.12	24.25	22.15	44.72	34.34
$\mathcal{X}$-Trans2Cap	✓	VoteNet	**61.83**	**35.65**	**26.61**	54.70	**43.87**	**25.05**	**22.46**	**45.28**	**35.31**

Method	Extra 2D	Proposals	C@0.25	B-4@0.25	M@0.25	R@0.25	C@0.5	B-4@0.5	M@0.5	R@0.5	mAP@0.5
Scan2cap	✗	VoteNet	41.76	24.12	24.98	55.79	23.70	14.88	20.95	47.50	32.17
TransCap	✗	VoteNet	44.32	25.63	25.25	55.69	27.24	17.76	21.60	49.16	34.09
$\mathcal{X}$-Trans2Cap	✗	VoteNet	**47.26**	**27.38**	**25.45**	**56.28**	**30.96**	**18.70**	**22.15**	**49.92**	**34.13**
3D-2D Proj.	✓	VoteNet	8.57	8.49	18.83	44.95	3.93	4.21	16.68	41.24	31.83
Scan2cap	✓	VoteNet	42.24	24.43	25.07	55.88	24.10	15.01	21.01	47.95	32.21
TransCap	✓	VoteNet	45.06	25.79	25.22	55.55	33.45	19.09	22.24	50.00	33.71
$\mathcal{X}$-Trans2Cap	✓	VoteNet	**51.43**	**27.62**	**25.75**	**56.46**	**33.62**	**19.29**	**22.27**	**50.00**	**34.38**

Experiments

Visualization

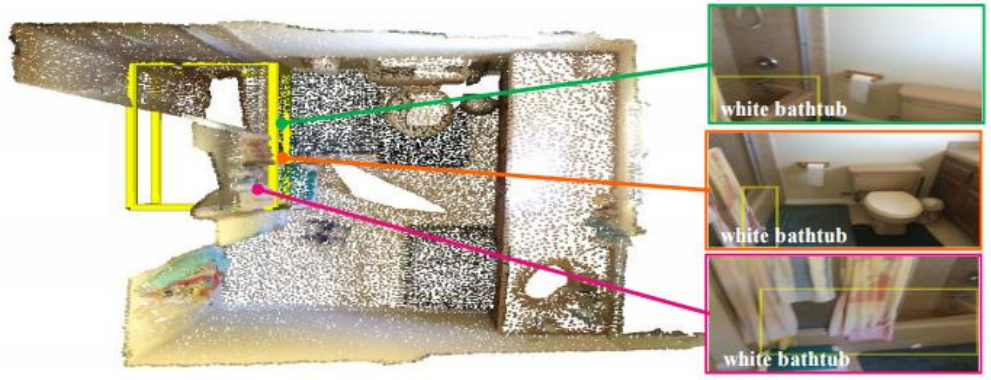
(a) Oracle DC



Scan2Cap: *This is a white cabinet. It is to the right of the bed.*

X-TransCap: *This is a brown wooden cabinet. It is to the left of the bed.*

Ground Truth: *This cabinet is called a wardrobe. It is tall and wooden. It is between the window and the bed.*

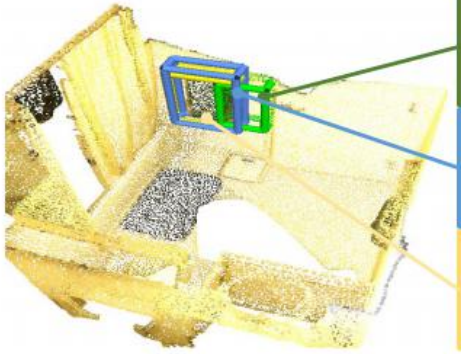


Scan2Cap: *The bathtub is brown . It is to the right of the toilet.*

X-TransCap: *This is a white bathtub. It is to the left of the toilet.*

Ground Truth: *This is a white bathtub. It is to the left of the toilet.*

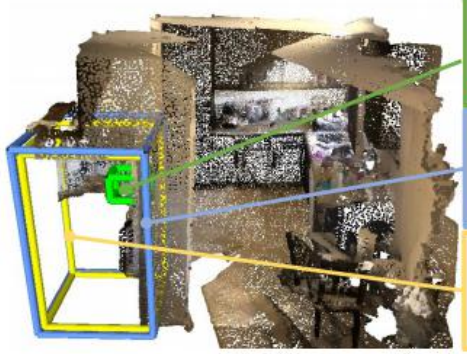
(b) Scan DC



Scan2Cap
This is a white towel. It is to the left of another towel.

X-TransCap
This is a green towel. It is hanging on the wall.

Ground Truth
A dark green towel. It is hanged in a rod that is attached to the wall.



Scan2Cap
This is a white refrigerator. It is to the right of the refrigerator.

X-TransCap
This is a white refrigerator. It is to the left of the stove.

Ground Truth
This is a white refrigerator. It is to the left of the stove.

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

X-Trans2Cap: Cross-Modal Knowledge Transfer using Transformer for 3D Dense Captioning

Thanks for watching!

Zhihao Yuan[1,†], Xu Yan[1,†], Yinghong Liao[1], Yao Guo[2], Guanbin Li[3], Shuguang Cui[1], Zhen Li[1,]*

[1] The Chinese University of Hong Kong (Shenzhen), The Future Network of Intelligence Institute, Shenzhen Research Institute of Big Data,
[2] Shanghai Jiao Tong University, [3] Sun Yat-sen University

目录

说话人脸

任务介绍

➢ 目标：

- 给定文本或语音作为驱动信息，同时给定人脸图片或视频提供人物信息，目标是生成人脸视频其嘴型与文本或语音内容保持一致；

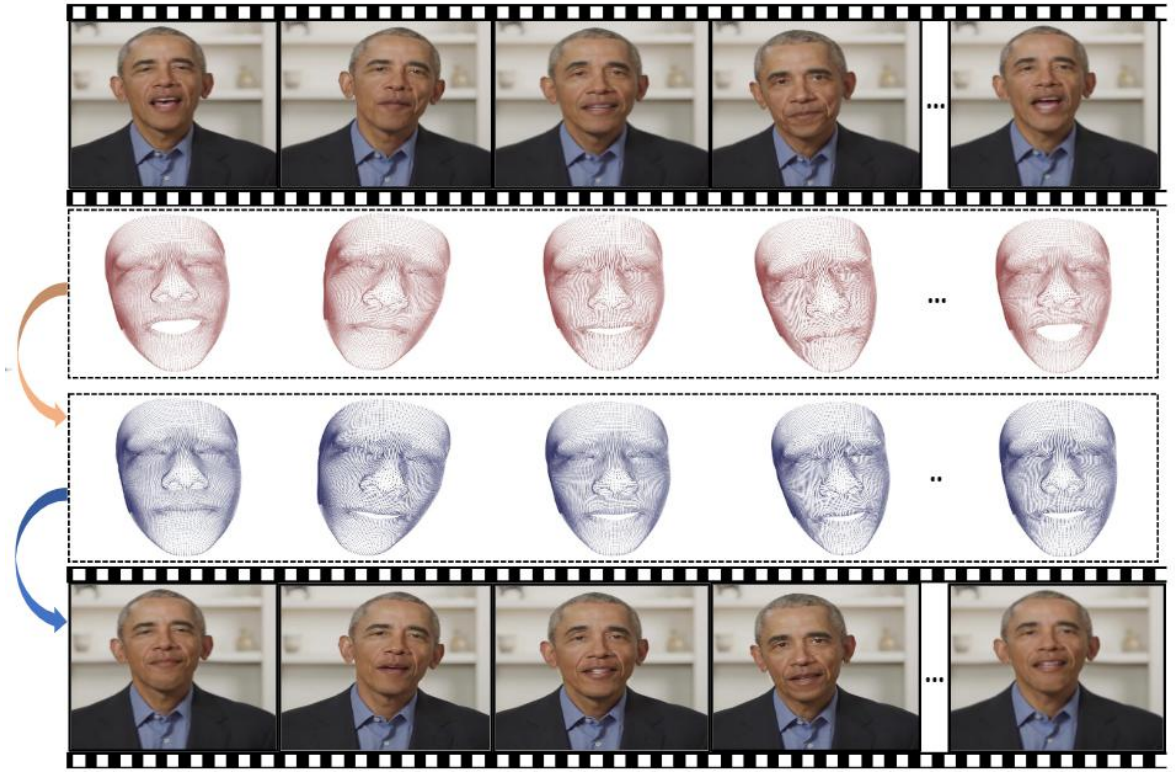
➢ 挑战：

- 跨模态学习任务，语音/文本模态到图像模态的映射，需要设计多模态的特征提取器以及跨模态交互学习；
- 人类视觉系统对生成视频图像质量和语音—嘴型同步质量比较敏感，生成高质量的说话人脸视频有挑战；

说话人脸

已有成果

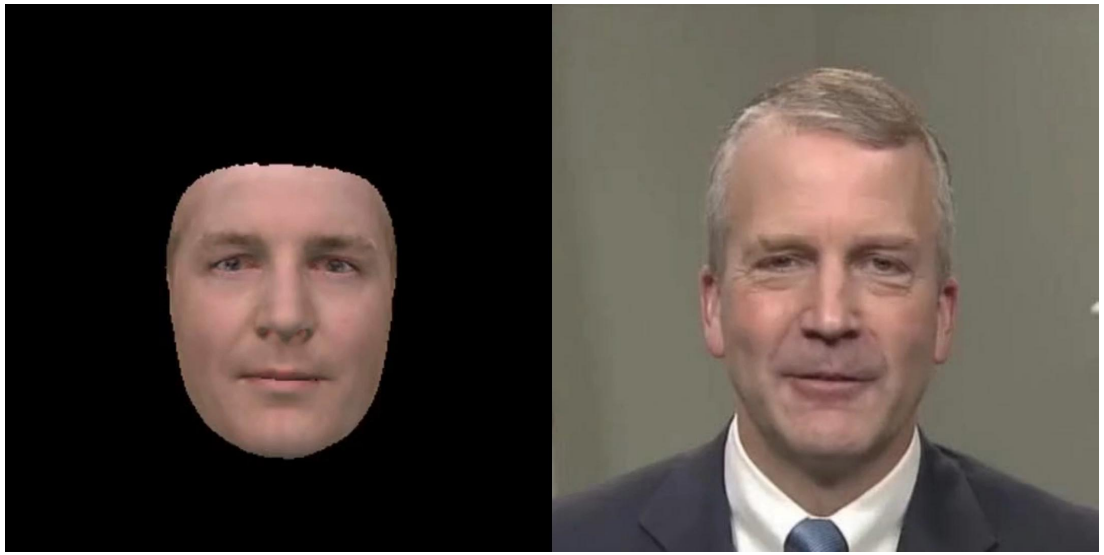
- 时序3DMM方案（利用点云解析相关算法预测脸部关键点）
- 利用精细3D人脸顶点进行语音到嘴型的显式监督，并考虑长时时序信息，得到稳定人脸视频；



点云解析驱动的高清说话人脸

生成结果

3D animation　　Blend result　　Generated result



点云解析驱动的高清说话人脸

不同语言生成结果

中文

德语

点云解析驱动的高清说话人脸

不同语言生成结果

粤语

英文

目录

ArSDM: Colonoscopy Images Synthesis with Adaptive Refinement Diffusion Models

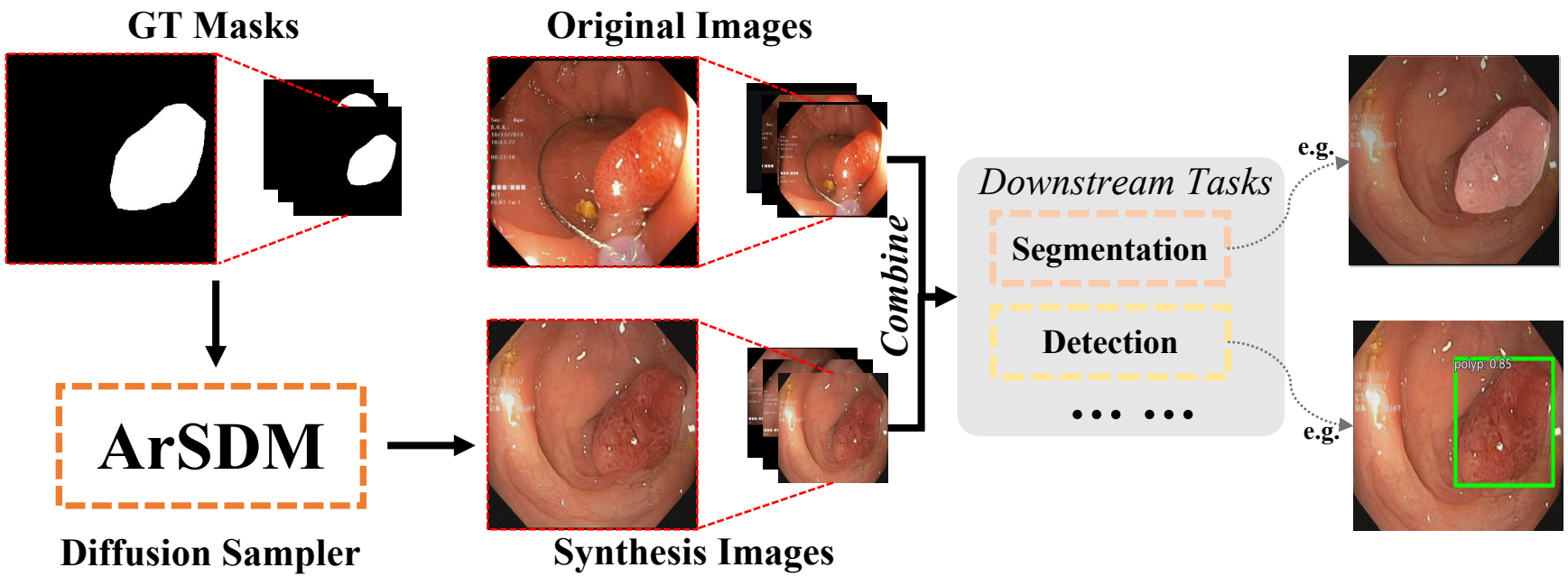
Background

((1)) Colonoscopy analysis, particularly automatic polyp segmentation and detection are essential for assisting clinical diagnosis and treatment, while the scarcity of annotated data limits the effectiveness and generalization of existing models.

((2)) The quality of generated data by GANs or other data augmentation methods is poor.

((3)) Diffusion models have demonstrated remarkable progress in generating multiple modalities of medical data (CT, MRI, …).

Overview of the Pipeline

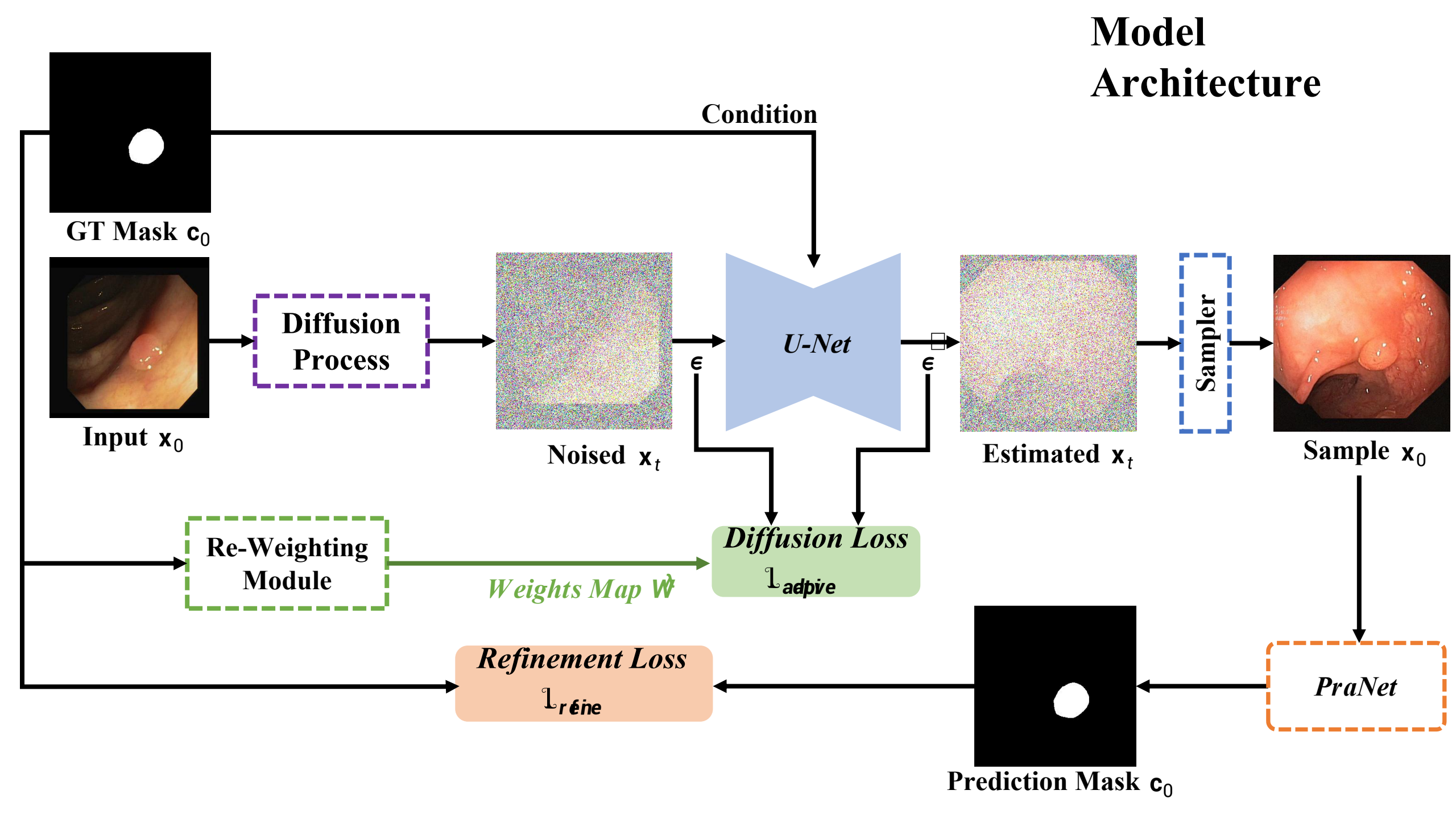


Pipeline

(1) Train a semantic diffusion model (our ArSDM).

(2) For each mask in the training set, sample a synthesized image, the synthesized dataset has the same number of image-mask pairs as the original dataset.

(3) Combine the original diffusion training set with the synthesized dataset for training polyp segmentation and detection models


Model Architecture
Condition
GT Mask c_0
Input x_0
Diffusion Process
Noised x_t
U-Net
ϵ
ϵ
Estimated x_t
Sampler
Sample x_0
Re-Weighting Module
Weights Map w
Diffusion Loss l_{adpive}
Refinement Loss l_{refine}
Prediction Mask c_0
PraNet

Model Architecture

- **Mask Conditioning**
  - Using the segmentation masks as conditions, similar to semantic masks but have only two categories: foreground (polyp) and background (intestine wall)
  - The conditional U-Net model is the same as SDM (Semantic Image Synthesis via Diffusion Models https://arxiv.org/abs/2207.00050)
- **Adaptive Loss Function**
  - Based on l_1 loss, define a pixel-wise weights matrix that vests different weights according to the size ratio of the polyp over the background.
  - For coding, it is convenient to use the pixel values of the segmentation mask (0,1).

propose an adaptive loss function that vests different weights according to the size ratio of the polyp over the background. Specifically, we define a pixel-wise weights matrix $W^\lambda \in \mathbb{R}^{H\times W}$ with each entry $w^\lambda_{i,j}$ to be:

$$w^\lambda_{i,j} = \begin{cases} 1-r & , p=1 \\ r & , p=0 \end{cases}, \qquad r = \frac{\#(p=1)}{H\times W} \tag{7}$$

where $p=1$ means the pixel p at (h,w) belongs to the polyp region and $p=0$ means it belongs to the background region. Thus, the loss function becomes:

$$\mathcal{L}_{\text{adaptive}} = \mathbb{E}_{t,\mathbf{x}_t,\mathbf{c}_0,\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[W^\lambda \cdot \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c}_0)\|^2\right] \tag{8}$$

Model Architecture

- **Mask Conditioning**
- **Adaptive Loss Function**
- **Refinement**
  - Using a pre-trained segmentation model to fine-tune the diffusion model, in which the U-Net parameters are updated while the segmentation model parameters are fixed.
  - For each time-step t, we need to sample an image, which is time-consuming.

Algorithm 1: One training iteration of ArSDM

Input: $t \sim \text{Uniform}(\{1, ..., T\})$, $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, $\mathbf{c}_0$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

Output: $\tilde{\boldsymbol{\epsilon}}$, $\tilde{\mathbf{c}}_0$

$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$; $\tilde{\mathbf{x}}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c}_0)$

for $i = t, ..., 1$ **do**

  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $i > 1$, else $\mathbf{z} = \mathbf{0}$; $\tilde{\mathbf{x}}_{i-1} = \frac{1}{\sqrt{\alpha_i}}\left(\tilde{\mathbf{x}}_i - \frac{1-\alpha_i}{\sqrt{1-\bar{\alpha}_i}}\boldsymbol{\epsilon}_\theta(\tilde{\mathbf{x}}_i, i, \mathbf{c}_0)\right) + \sigma_i\mathbf{z}$

end for

$\tilde{\mathbf{c}}_0 = \mathcal{P}(\tilde{\mathbf{x}}_0)$

Take gradient descent step on $\nabla_\theta \mathcal{L}_{\text{total}}$

Experimental Settings

Diffusion Training

- Training Set: Kvasir + CVC-ClinicDB (1450 image-mask pairs)
- Image Size: Padding to have the same height and width and then resize to 384×384
- Duration:
  - with Refinement: around one-half NVIDIA A100 days (80GB Memory)
  - w/o refinement: around one A100 day.

Diffusion Sampling

- DDIM sampler with $T = 200$
- Random noise as input and mask as a condition

Comparison Results

Polyp Segmentation

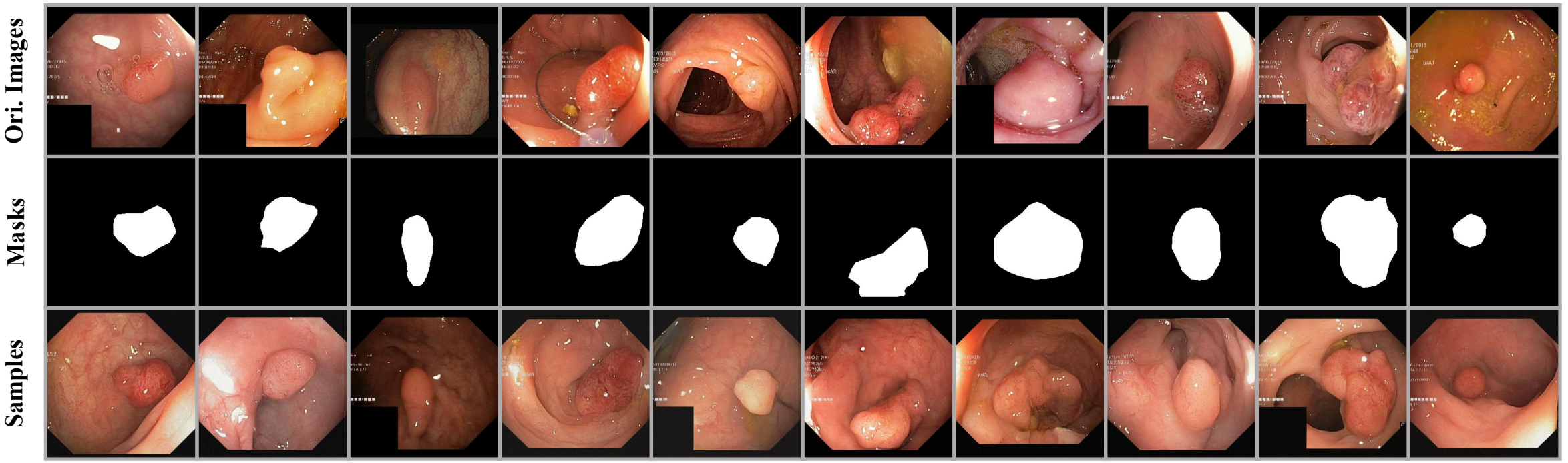
Methods	EndoScene		ClinicDB		Kvasir		ColonDB		ETIS		Overall		
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	
PraNet	87.1	79.7	89.9	84.9	89.8	84.0	70.9	64.0	62.8	56.7	74.0	67.5	
+LDM	83.7	76.9	88.2	83.5	88.4	83.0	62.6	56.0	56.2	50.3	67.8	61.7	
+SDM	**89.9**	**83.2**	89.2	83.7	88.4	82.6	74.2	66.5	66.4	60.3	76.4	69.6	
+**Ours**	89.7	82.7	**93.3**	**88.5**	**89.9**	**84.5**	**76.1**	**68.9**	**75.5**	**68.1**	**80.0**	**73.2**	+6.0%, +5.7%
SANet	88.8	81.5	**91.6**	85.9	90.4	84.7	75.3	67.0	75.0	65.4	79.4	71.4	
+LDM	72.7	60.5	88.8	82.8	88.7	82.7	64.3	55.4	58.0	49.2	68.3	59.8	
+SDM	**90.2**	83.0	89.9	84.1	90.9	85.4	77.6	69.3	74.7	66.8	80.4	72.9	
+**Ours**	**90.2**	**83.2**	91.4	**86.1**	**91.1**	**85.6**	**77.7**	**70.0**	**78.0**	**69.5**	**81.5**	**74.1**	+2.1%, +2.7%
PVT	**90.0**	**83.3**	93.7	88.9	**91.7**	**86.4**	80.8	72.7	78.7	70.6	83.3	76.0	
+LDM	88.2	81.2	92.3	87.1	91.2	85.7	78.7	70.4	78.0	69.6	81.9	74.2	
+SDM	88.8	81.7	**93.9**	**89.2**	91.2	86.1	81.3	73.5	78.7	71.1	83.4	76.3	
+**Ours**	88.2	81.2	92.2	87.5	91.5	86.3	**81.7**	**73.8**	**80.6**	**72.9**	**84.0**	**76.7**	+0.7%, +0.7%

Comparison Results

Polyp Detection

Methods	EndoScene		ClinicDB		Kvasir		ColonDB		ETIS		Overall		
	AP	F1	AP	F1	AP	F1	AP	F1	AP	F1	AP	F1	
Center.	86.9	**91.4**	84.7	89.2	75.6	81.4	62.2	72.3	62.7	70.1	56.6	76.0	
+LDM	84.1	84.4	**90.4**	89.9	81.3	81.8	73.4	74.5	65.2	71.7	62.0	76.9	
+SDM	**87.8**	86.9	88.7	**91.0**	77.0	82.8	71.8	78.1	68.2	72.6	61.8	79.1	
+**Ours**	85.0	89.1	86.1	90.8	**77.3**	**84.7**	**74.2**	**80.2**	**68.7**	**75.6**	**65.7**	**81.3**	+9.1%, +5.3%
Sparse.	89.9	87.8	81.4	86.4	75.6	80.2	78.2	73.2	63.8	62.4	63.7	73.2	
+LDM	87.4	76.3	**95.0**	**93.5**	81.5	58.8	80.0	71.0	64.4	54.3	65.3	66.3	
+SDM	**94.5**	**90.5**	88.7	86.5	79.0	80.5	**81.4**	76.8	67.8	67.1	65.2	76.7	
+**Ours**	92.8	86.2	92.2	90.6	**81.6**	**82.3**	80.1	**79.8**	**72.4**	**70.4**	**66.4**	**79.0**	+2.7%, +5.8%
Deform.	**98.1**	**94.4**	89.7	89.9	**80.2**	74.4	**82.2**	75.5	65.3	54.7	64.5	71.8	
+LDM	94.6	90.5	91.6	89.5	79.3	73.4	78.0	73.2	69.0	64.0	63.4	73.3	
+SDM	96.0	90.6	90.3	91.2	82.2	78.9	80.1	75.1	67.5	66.7	65.1	75.8	
+**Ours**	94.7	94.3	**92.3**	**92.0**	80.0	**80.3**	81.4	**77.3**	**74.1**	**69.3**	**67.9**	**77.9**	+3.4%, +6.1%

Visualization



Colonscopy Video Generation with Diffusion Models

PVDM **用sky-time-lapse训练autoencoder, from scratch**

Dataset: Sky Time-lapse
997 段视频; 总共 1,172,641 帧

Training: 1 V100, 1 day

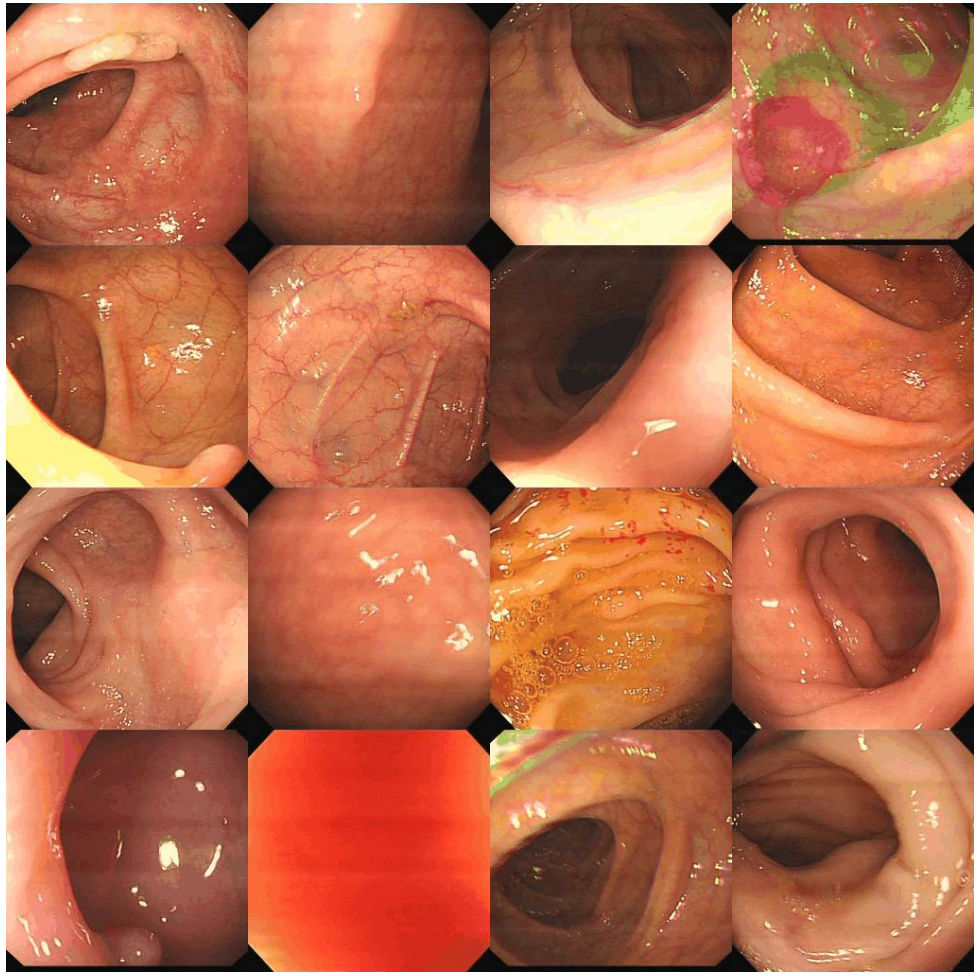
Inputs

Reconstructions

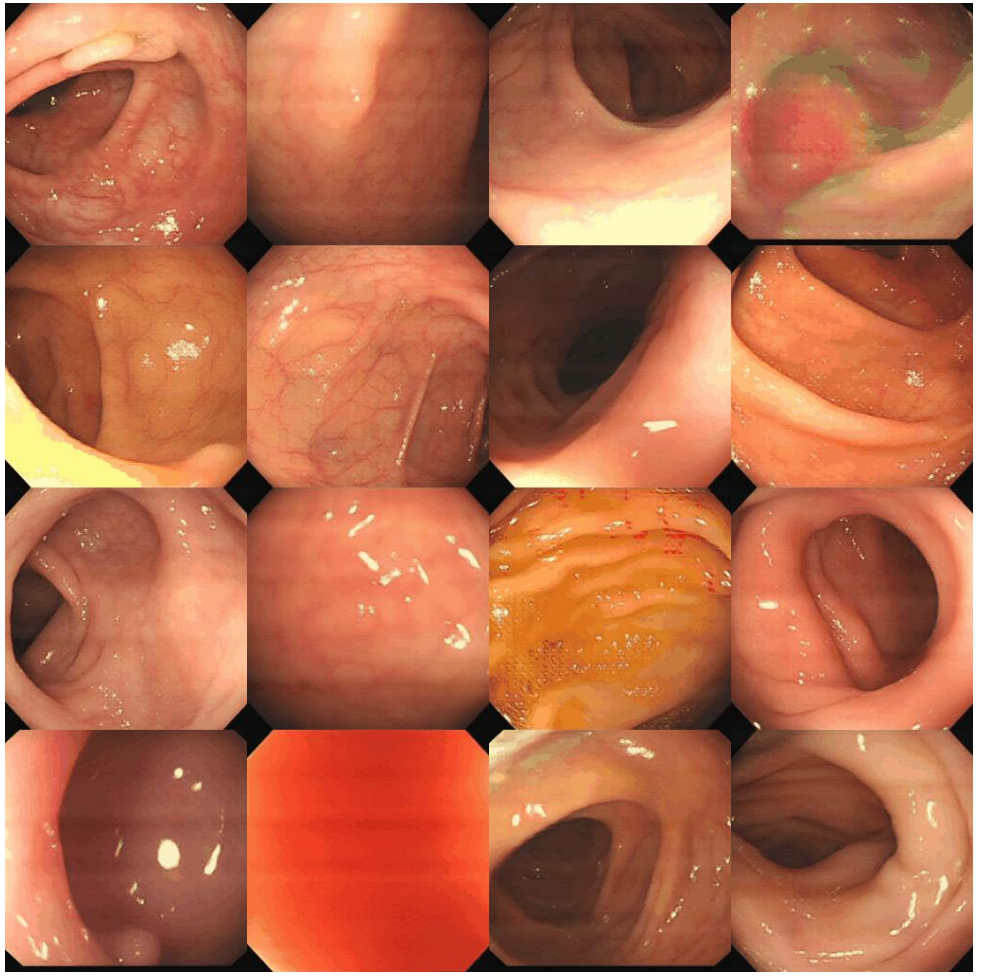
PVDM 用LDPolyp训练autoencoder, 加载sky-time-lapse的权重

Dataset: LDPolyp
100 段视频; 总共 24,789 帧

Training: 1 V100, 1.6 days



Inputs



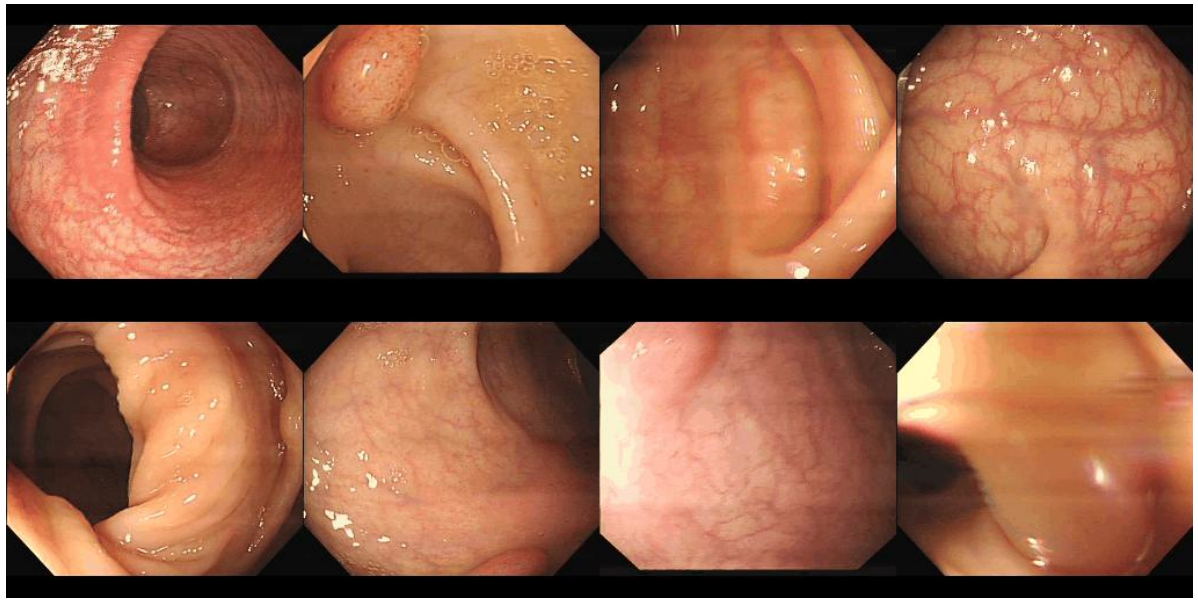
Reconstructions

LVDM

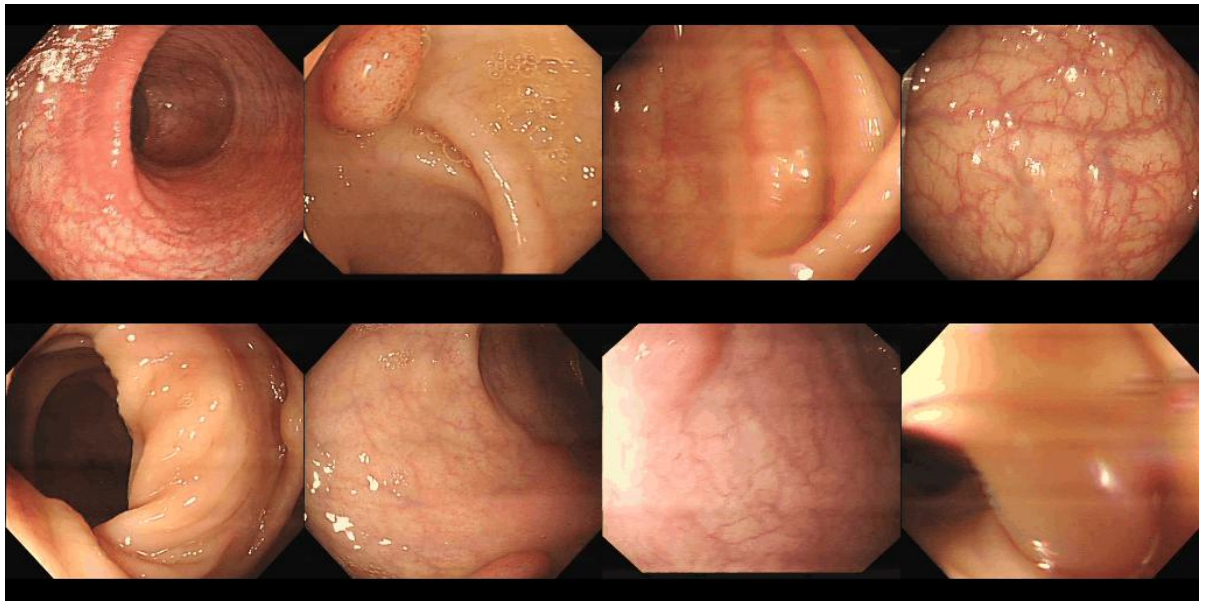
用LDPolyp训练 2D autoencoder, 加载ImageNet训练权重

Dataset: LDPolyp
100 段视频; 总共 24,789 帧

Training: 1 V100, 1 day



Inputs



Reconstructions

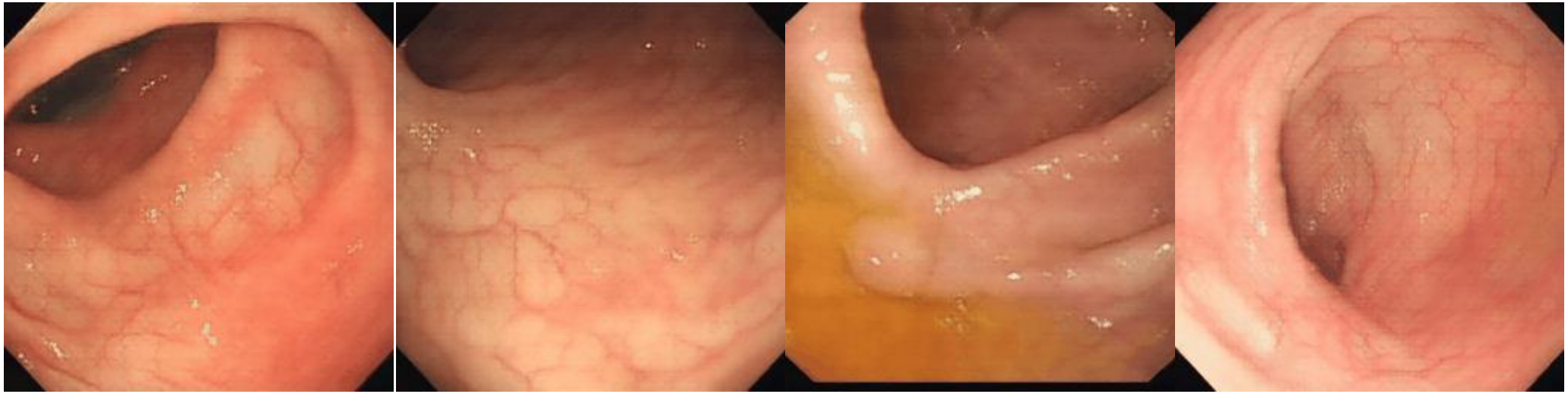
LVDM-2 **用LDPolyp训练 unconditional diffusion model**

(LVDM 5.1 release codes)

Dataset: LDPolyp
100 段视频; 总共 24,789 帧

Training: 1 80GB A100, 1 day

Details: 3D-Unet 添加部分Attention layer 作attention
参数过多，batch_size 2 -> 48 GB 显存



Samples

√

下一步启示

- 进一步优化多模态在3D场景的解析与生成
- 结合video diffusion来强化说话人脸的效果
- 结合condition mask来进行医疗图像场景的video diffusion生成