NEURAL INFORMATION PROCESSING SYSTEMS

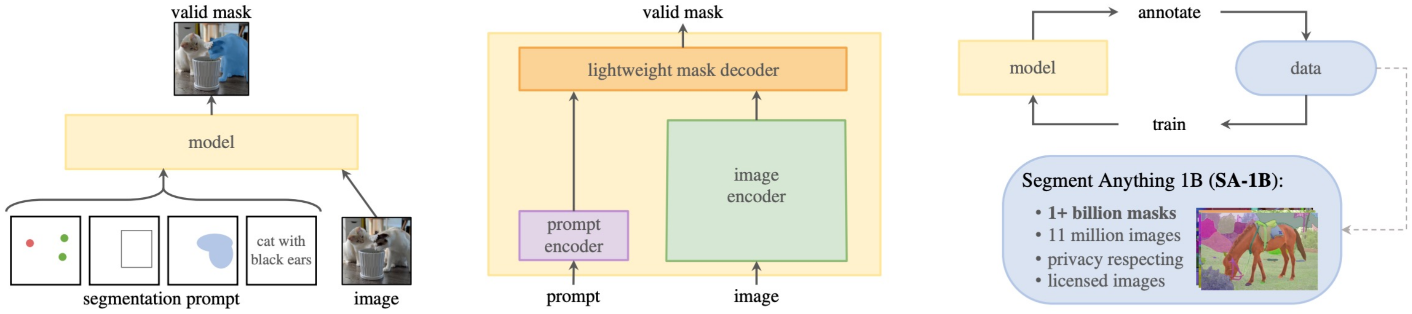
Segment Anything in 3D with NeRFs

Jiazhong Cen[1,2*], Zanwei Zhou[1*], Jiemin Fang[2,3], Chen Yang[1], Wei Shen[1],
Lingxi Xie[2], Dongsheng Jiang[2], Xiaopeng Zhang[2], Qi Tian[2]

[1]Institute of AI, SJTU [2]Huawei Inc. [3]School of EIC, HUST

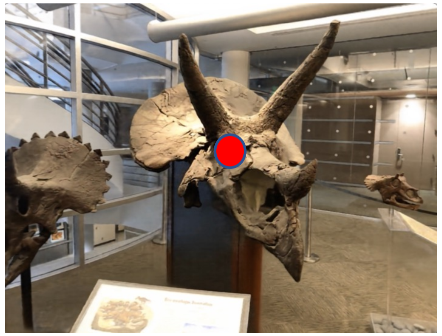
Background: Segment Anything Model (SAM)

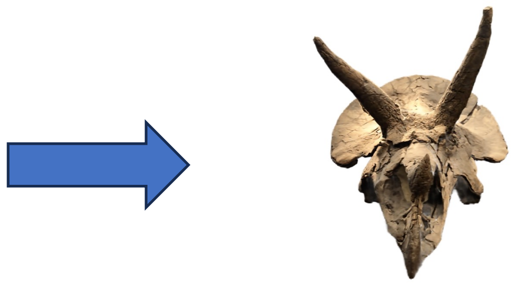
- An impressive 2D segmentation foundation model



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

- How to build a 3D segmentation foundation model?
  - Lack of 3D data
  - Rich structure priors

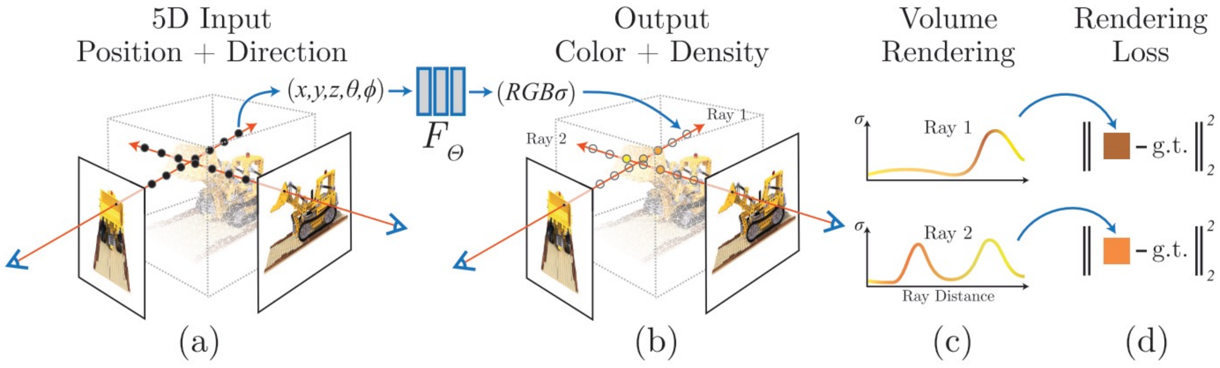




Preliminaries

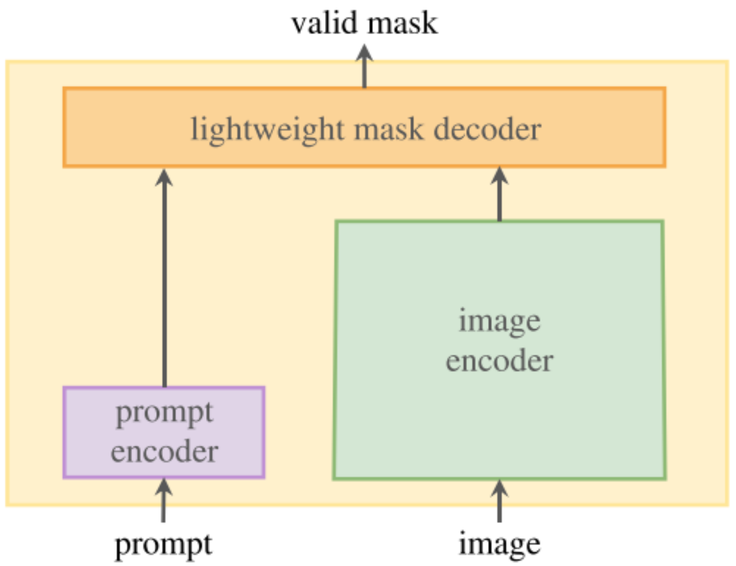
- Neural Radiance Fields (NeRFs)
  - A kind of **3D representation**
  - Differentiable volumetric rendering:

$$\mathbf{I}_{\boldsymbol{\theta}}(\mathbf{r}) = \int_{t_n}^{t_f} \omega(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})\mathrm{d}t$$

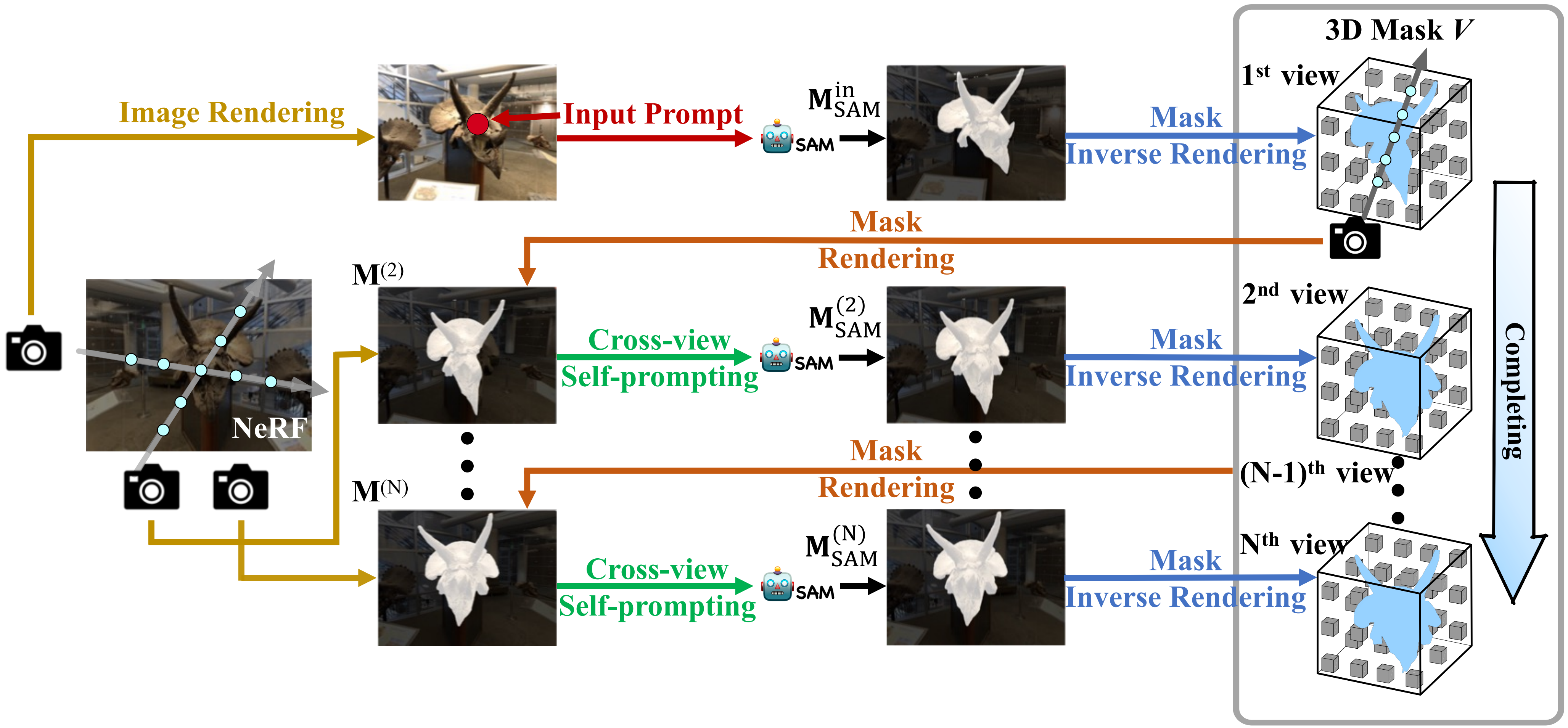


- Segment Anything Model (SAM)
  - Image encoder + prompt encoder + mask decoder
  - **Input an image and prompts, output masks**

$$\mathbf{M}_{\text{SAM}} = s(\mathbf{I}, \mathcal{P})$$



Overall Pipeline



Mask Inverse Rendering

- **Mask grids rendering**:

$$\mathbf{M}(\mathbf{r}) = \int_{t_n}^{t_f} \omega(\mathbf{r}(t))\mathbf{V}(\mathbf{r}(t))\mathrm{dt}$$

  - Weights are from the pretrained NeRF

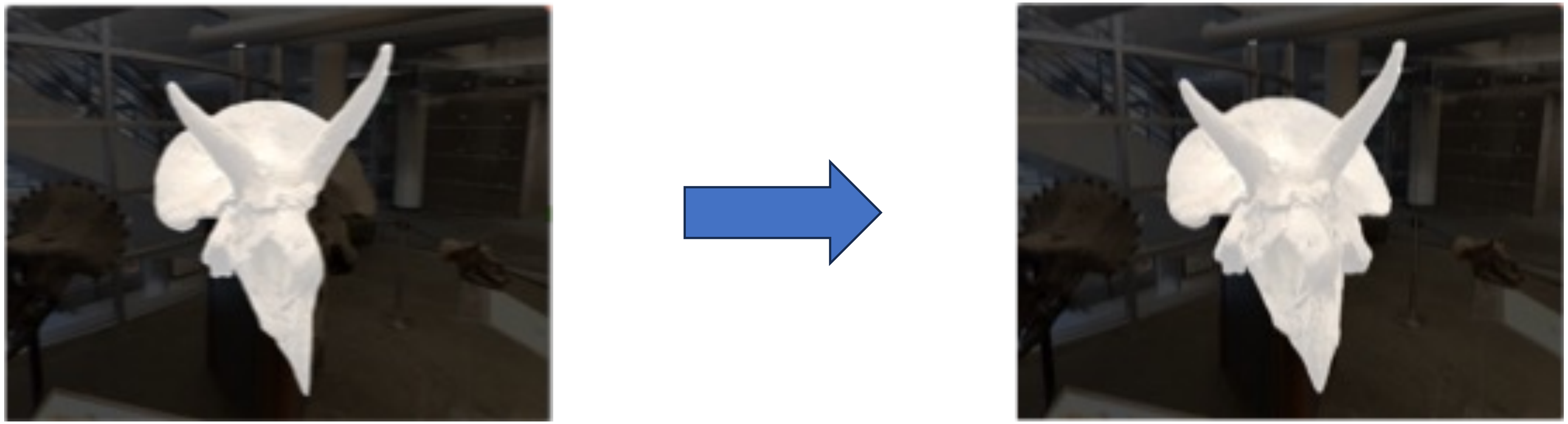
- **Projecting a 2D mask onto the mask grids** is equivalent to assigning mask confidence scores according to the weights
  - Can be solved by gradient descent

$$\mathcal{L}_{\texttt{proj}} = -\sum_{\mathbf{r}\in\mathcal{R}(\mathbf{I})} \mathbf{M}_{\texttt{SAM}}(\mathbf{r}) \cdot \mathbf{M}(\mathbf{r})$$

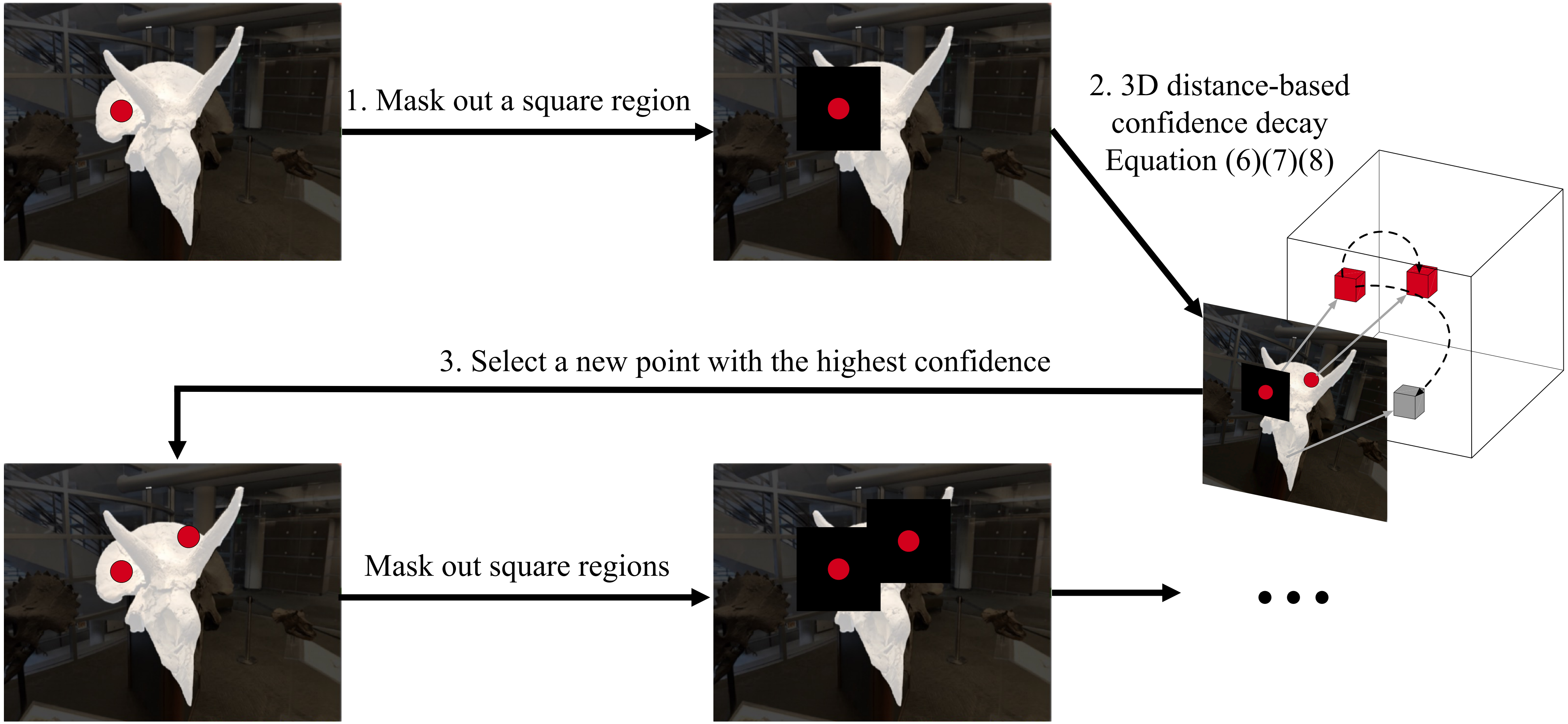
  - SAM is not always correct

$$\mathcal{L}_{\texttt{proj}} = -\sum_{\mathbf{r}\in\mathcal{R}(\mathbf{I})} \mathbf{M}_{\texttt{SAM}}(\mathbf{r}) \cdot \mathbf{M}(\mathbf{r}) + \lambda \sum_{\mathbf{r}\in\mathcal{R}(\mathbf{I})} (1 - \mathbf{M}_{\texttt{SAM}}(\mathbf{r})) \cdot \mathbf{M}(\mathbf{r})$$

Cross-view Self-prompting



Cross-view Self-prompting



Quantitative Results

Table 1: Quantitative results on NVOS.

Method	mIoU (%)	mAcc (%)
Graph-cut (3D) [48, 47]	39.4	73.6
NVOS [47]	70.1	92.0
ISRF [15]	83.8	96.4
SA3D (ours)	**90.3**	**98.2**

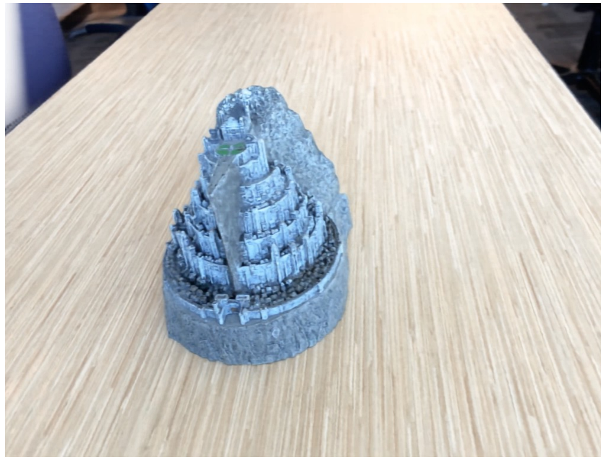
Table 2: Quantitative results on the SPIn-NeRF dataset.

Scenes	Single view		MVSeg [39]		SA3D (ours)	
	IoU (%)	Acc (%)	IoU (%)	Acc (%)	IoU (%)	Acc (%)
Orchids	79.4	96.0	**92.7**	98.8	83.6	96.9
Leaves	78.7	98.6	94.9	99.7	**97.2**	99.9
Fern	95.2	99.3	94.3	99.2	**97.1**	99.6
Room	73.4	96.5	**95.6**	99.4	88.2	98.3
Horns	85.3	97.1	92.8	98.7	**94.5**	99.0
Fortress	94.1	99.1	97.7	99.7	**98.3**	99.8
Fork	69.4	98.5	87.9	99.5	**89.4**	99.6
Pinecone	57.0	92.5	**93.4**	99.2	92.9	99.1
Truck	37.9	77.9	85.2	95.1	**90.8**	96.7
Lego	76.0	99.1	74.9	99.2	**92.2**	99.8
mean	74.6	95.5	90.9	98.9	**92.4**	98.9

Table 3: Quantitative results on Replica (mIoU).

Scenes	office0	office1	office2	office3	office4	room0	room1	room2	mean
Single view	68.7	56.5	68.4	62.2	57.0	55.4	53.8	56.7	59.8
MVSeg [39]	31.4	40.4	30.4	30.5	25.4	31.1	40.7	29.2	32.4
SA3D (ours)	**84.4**	**77.0**	**88.9**	**84.4**	**82.6**	**77.6**	**79.8**	**89.2**	**83.0**

Qualitative Results



Qualitative Results

More Analysis

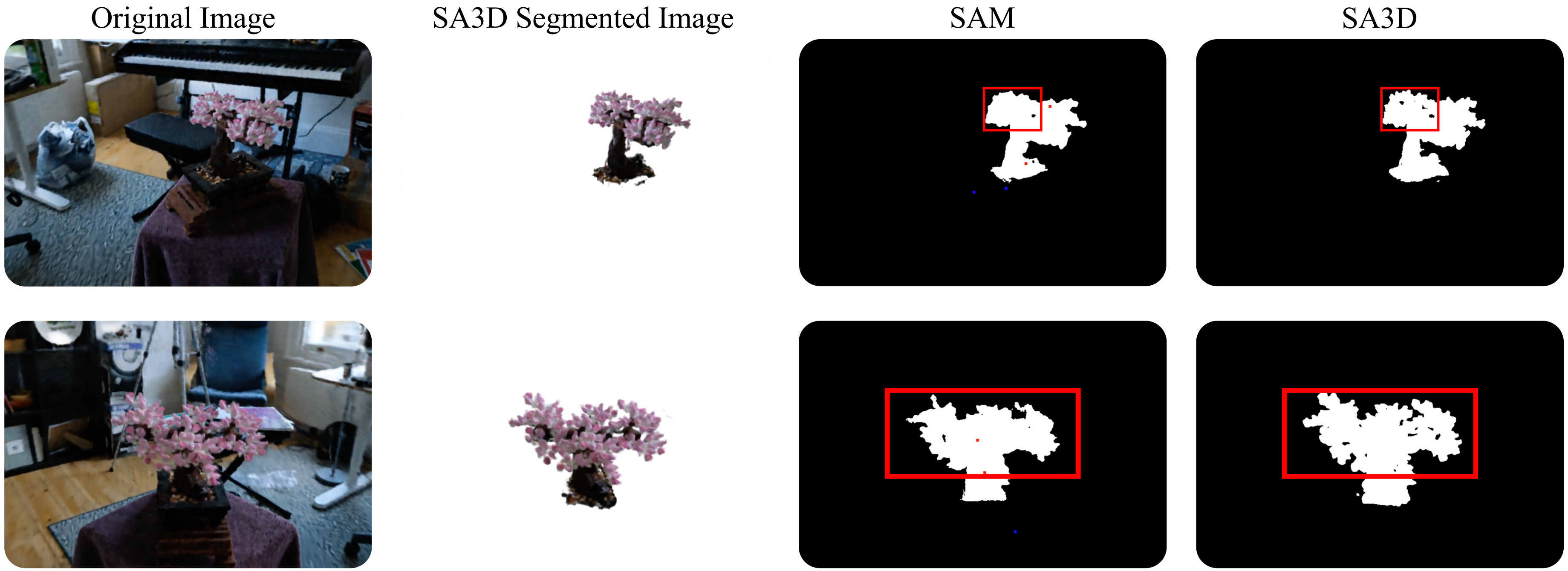
- Different 2D segmentation model (evaluated on the NVOS dataset)

SEEM [75]		SimpleClick [32]		RITM [49]		FocalClick [5]	
mIoU (%)	mAcc (%)	mIoU (%)	mAcc (%)	mIoU (%)	mAcc (%)	mIoU (%)	mAcc (%)
86.0	97.0	87.7	97.8	81.2	96.3	88.9	98.1

  - SEEM: segmentation foundation model
  - SimpleClick, RITM, FocalClick: interactive segmentation models
  - **SA3D can generalize to different models** if they can steadily address promptable segmentation across multiple views

More Analysis

- **NeRF helps SAM**



- The depth information provided by NeRFs can help to generate more precise 2D segmentation results

More Analysis

- The effect of occlusion

  - Occluded but seen in other views
  - Never seen in any view: maybe can use diffusion models for repairing

Summary

- We present SA3D, a novel framework for lifting 2D segmentation models to 3D
- We demonstrate the effectiveness of SA3D on various datasets
- Comprehensive experiments are conducted to analyze the characteristic of SA3D

Thanks for listening!

Segment Anything in 3D with NeRFs

Jiazhong Cen[1,2*], Zanwei Zhou[1*], Jiemin Fang[2,3], Chen Yang[1], Wei Shen[1],
Lingxi Xie[2], Dongsheng Jiang[2], Xiaopeng Zhang[2], Qi Tian[2]

[1]Institute of AI, SJTU [2]Huawei Inc. [3]School of EIC, HUST