SimpleClick: Interactive Image Segmentation with Simple Vision Transformers

Qin Liu, Zhenlin Xu, Gedas Bertasius, Marc Niethammer
University of North Carolina at Chapel Hill
ICCV 2023

Introduction

- ❖ Goal of interactive segmentation:

  Obtain high-quality pixel-level annotations with limited user interaction

- ❖ Interaction types:

  bounding boxes polygons, clicks , scribbles,and their combinations

- ❖ click-based->simplicity and well-established training and evaluation protocols
- ❖ hierarchical backbone as predominant architecture for current methods

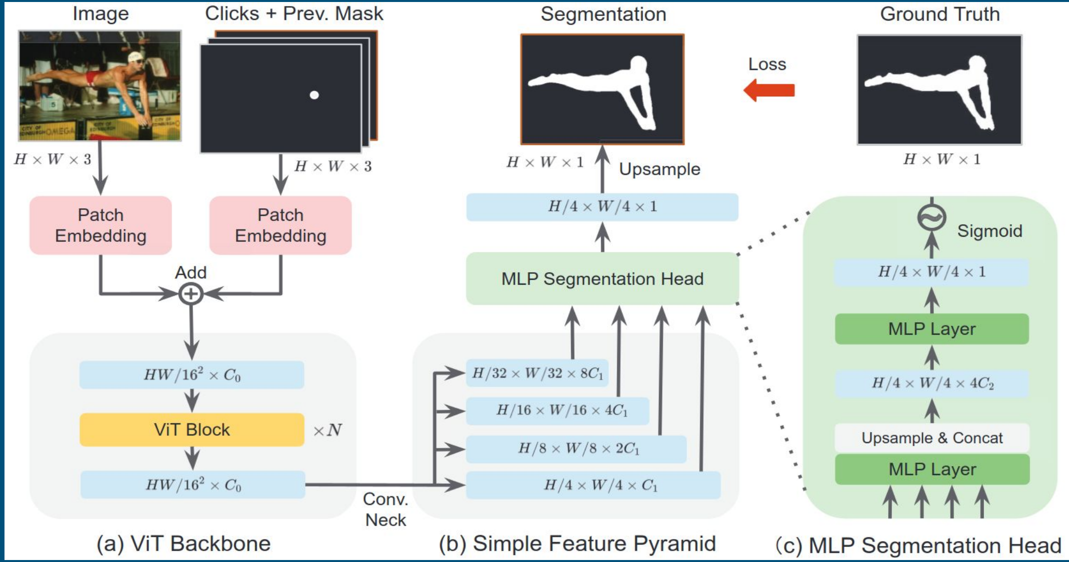
Method

Method includes 3 main modules:

a) a plain ViT backbone

b)a multi-scale simple feature pyramid

c)a light-weight MLP decoder for segmentation



Method

- Adapt a plain-ViT backbone for interactive segmentation with minimal modifications.
  - The plain segmentation backbone
  - Adapt the backbone for interactive segmentation
  - Other modules of Simple Click
  - Training and inference details of method

Method

- ❖ The plain segmentation backbone:
  - ➢ Patch embedding layer
- ❖ ViT backbones:
  - ➢ ViT-B, ViT-L, and ViT-H

Model↓ Module→	ViT Backbone	Conv. Neck	MLP Head
Ours-ViT-B (base)	83.0 (89.3%)	9.0 (9.7%)	0.9 (1.0%)
Ours-ViT-L (large)	290.8 (94.3%)	16.5 (5.3%)	1.1 (0.4%)
Ours-ViT-H (huge)	604.0 (95.7%)	25.8 (4.1%)	1.3 (0.2%)

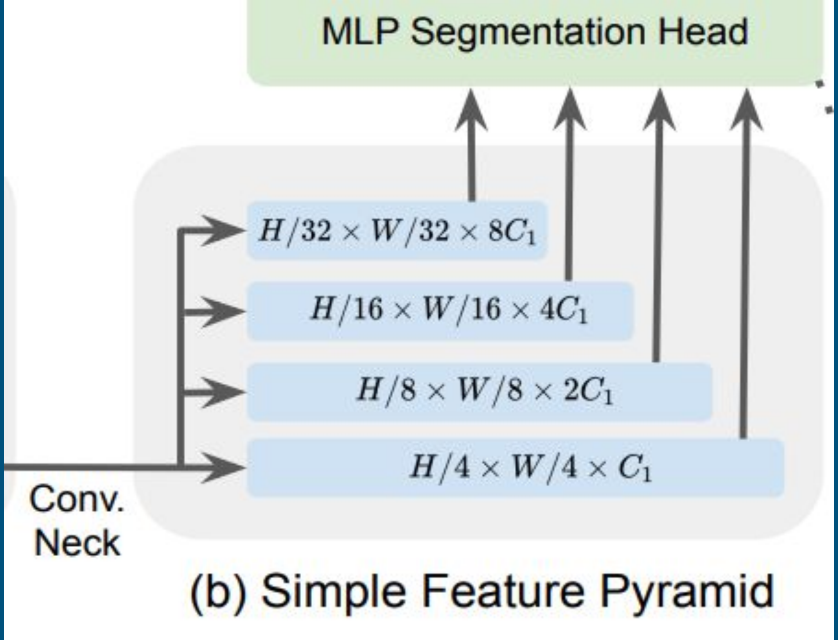
Table 1. **Number of parameters of our models**. The unit is a million. The percentage of parameters is shown in bracket. Most parameters are used by the ViT backbone.

Method

- Clicks Encoding for Segmentation Backbone:
  - Adapt the plain segmentation backbone-> turning user interactions into a form of guidance learned by the network.
  - Encoding each click on a 2-channel masks: + click in foreground -click in background
  - Clicks: human clicks and simulated clicks
  - How to fuse clicks into the backbone?
    - patch-embedding layer symmetric to the one in the backbone
    - concatenate the previous segmentation to the clicks map as an additional channel for better performance(3 channels)
    - two symmetric embedding layers operate on the image and the clicks map.
    - The Inputs ➡ two vector sequences of the same dimension ➡ element-wise addition

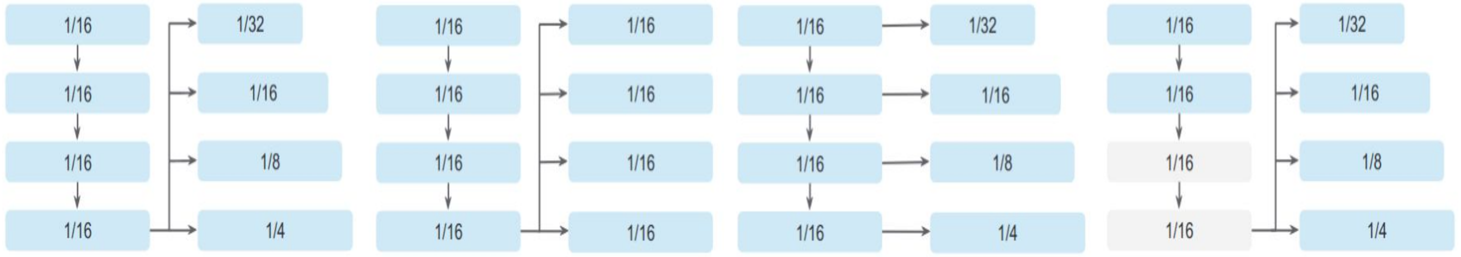
Method

- Other Modules :Simple Feature Pyramid
- hierarchical backbone
  - use FPN to produce feature Pyramid
- Plain backbone
  - use parallel convolutional or deconvolution layers with input be only the last feature map of the backbone to produce feature Pyramid



Method

FP design	frozen ViT	ViT-B SBD	ViT-B Pascal	ViT-L SBD	ViT-L Pascal
(a) simple FP	✓	11.48	6.93	9.75	5.59
(a) simple FP	✗	5.24	2.53	4.46	2.15
(b) single-scale	✗	6.56	2.80	5.53	2.48
(c) parallel	✗	7.21	3.09	6.26	2.79
(d) partial	✗	8.29	4.34	7.51	4.25

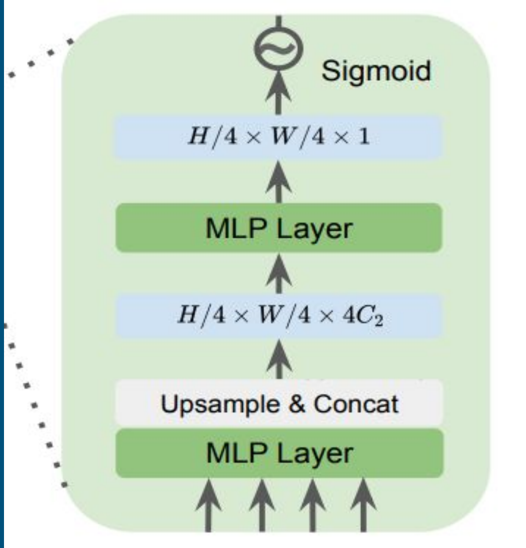


(a) simple feature pyramid (b) single-scale variant (c) parallel variant (d) partial variant

Method

❖ Other Modules :All-MLP Segmentation Head

- ➢ lightweight segmentation head -> MLP layers
- ➢ Input: simple feature pyramid -> a segmentation probability map of scale 1⁄4
  - ■ Transform feature maps to an identical channel dimension ($C2$)
  - ■ Upsample(¼) and concatenation
  - ■ Transform them to single channel feature map after MLP
  - ■ segmentation probability map by Sigmoid
  - ■ Binary segmentation by applying thresold



Method

- Training and Inference Settings:
  - Backbone Pretraining:
    - pretrained as MAEs on ImageNet-1K
    - Pretraing is not done by the authors
    - use the readily available pretrained MAE weights from previous works
- Clicks Simulation and End-to-end Fine-tuning
  - Pipeline:
    - 1.simulate clicks based on the current segmentation and gold standard segmentation:random and iterative strategy(without human-in-the-loop clicks
    - 2.Use segmentation from the previous interaction as an additional input for the backbone
- Normalized focal loss(NFL) for training all models

Method

- Human Evaluation and Automatic Evaluation:
  - Inference modes:
    - 1.automatic evaluation ->quantitative analyses
    - 2.human evaluation -> qualitative assessment

Experiments

- Datasets:10 public datasets including 7 natural image datasets and 3 medical datasets
  - GrabCut: 50 images (50 instances)
  - Berkeley: 96 images (100 instances)
  - DAVIS (Densely Annotated VIdeo Segmentation (DAVIS)): 50 videos; we only use 345 frames
  - Pascal VOC(Visual Object Classes): 1449 images (3427 instances) in the validation set.
  - SBD(Semantic Boundaries Dataset): 8498 training images (20172 instances) and 2857 validation images (6671 instances)
  - COCO(Common Objects in Context) +LVIS (Large Vocabulary Instance Segmentation:LVIS is a newer dataset that complements COCO and focuses on large vocabulary instance segmentation.)(C+L): COCO contains 118K training images (1.2M instances)
  - ssTEM : two image stacks, each contains 20 medical images.
  - BraTS: 369 magnetic resonance image (MRI) volumes; we test on 369 slices
  - OAIZIB: 507 MRI volumes; we test on 150 slices (300 instances)

Experiments

- Evaluation Metrics :
  - Simulation of user clicks:next click will be put at the center of the region with the largest error.
  - Number of Clicks (NoC)as the evaluation metric
  - IoU(Intersection over Union) ->NoC%85and NoC%9
  - average IoU given k clicks (mIoU@k) as an evaluation metric

Experiments

- ❖ Implementation Details:
  - ➢ train models on either SBD or COCO+LVIS with 55 epoch
  - ➢ batch size to 140 for ViT-Base, 72 for ViT-Large, and 32 for ViT-Huge
  - ➢ Train on four NVIDIA GTX A6000 GPUs
  - ➢ data augmentation
  - ➢ ViT backbone was pre trained on images of size224×224 ->finetune on 448 × 448 with non-shifting window attention for better performance

Experiments

- Comparison with Previous Results:

Method	Backbone	GrabCut		Berkeley		SBD		DAVIS		Pascal VOC	
		NoC85	NoC90	NoC85	NoC90	NoC85	NoC90	NoC85	NoC90	NoC85	NoC90
♪ DIOS [48] CVPR16	FCN	-	6.04	-	8.65	-	-	-	12.58	6.88	-
♪ FCA-Net [31] CVPR20	ResNet-101	-	2.08	-	3.92	-	-	-	7.57	2.69	-
♩ LD [27] CVPR18	VGG-19	3.20	4.79	-	-	7.41	10.78	5.05	9.57	-	-
♩ BRS [23] CVPR19	DenseNet	2.60	3.60	-	5.08	6.59	9.78	5.58	8.24	-	-
♩ f-BRS [42] CVPR20	ResNet-101	2.30	2.72	-	4.57	4.81	7.73	5.04	7.41	-	-
♩ RITM [43] Preprint21	HRNet-18	1.76	2.04	1.87	3.22	3.39	5.43	4.94	6.71	2.51	3.03
♩ CDNet [9] ICCV21	ResNet-34	1.86	2.18	1.95	3.27	5.18	7.89	5.00	6.89	3.61	4.51
♩ PseudoClick [34] ECCV22	HRNet-18	1.68	2.04	1.85	3.23	3.38	5.40	4.81	6.57	2.34	2.74
♩ FocalClick [10] CVPR22	HRNet-18s	1.86	2.06	-	3.14	4.30	6.52	4.92	6.48	-	-
♩ FocalClick [10] CVPR22	SegF-B0	1.66	1.90	-	3.14	4.34	6.51	5.02	7.06	-	-
♩ FocusCut [30] CVPR22	ResNet-50	1.60	1.78	1.85†	3.44	3.62	5.66	5.00	6.38	-	-
♩ FocusCut [30] CVPR22	ResNet-101	1.46	1.64	1.81†	3.01	3.40	5.31	4.85	6.22	-	-
♩ Ours	ViT-B	1.40	1.54	1.44	2.46	3.28	5.24	**4.10**	5.48	2.38	2.81
♩ Ours	ViT-L	1.38	1.46	1.40	2.33	2.69	4.46	4.12	5.39	1.95	2.30
♩ Ours	ViT-H	**1.32**	**1.44**	**1.36**	**2.09**	**2.51**	**4.15**	4.20	**5.34**	**1.88**	**2.20**
♫ RITM [43] Preprint21	HRNet-32	1.46	1.56	1.43	2.10	3.59	5.71	4.11	5.34	2.19	2.57
♫ CDNet [9] ICCV21	ResNet-34	1.40	1.52	1.47	2.06	4.30	7.04	4.27	5.56	2.74	3.30
♫ PseudoClick [34] ECCV22	HRNet-32	1.36	1.50	1.40	2.08	3.46	5.54	3.79	5.11	1.94	2.25
♫ FocalClick [10] CVPR22	SegF-B0	1.40	1.66	1.59	2.27	4.56	6.86	4.04	5.49	2.97	3.52
♫ FocalClick [10] CVPR22	SegF-B3	1.44	1.50	1.55	1.92	3.53	5.59	3.61	4.90	2.46	2.88
♫ Ours	ViT-B	1.38	1.48	1.36	1.97	3.43	5.62	3.66	5.06	2.06	2.38
♫ Ours	ViT-L	**1.32**	**1.40**	**1.34**	1.89	2.95	4.89	**3.26**	4.81	**1.72**	**1.96**
♫ Ours	ViT-H	1.38	1.50	1.36	**1.75**	**2.85**	**4.70**	3.41	**4.78**	1.76	1.98

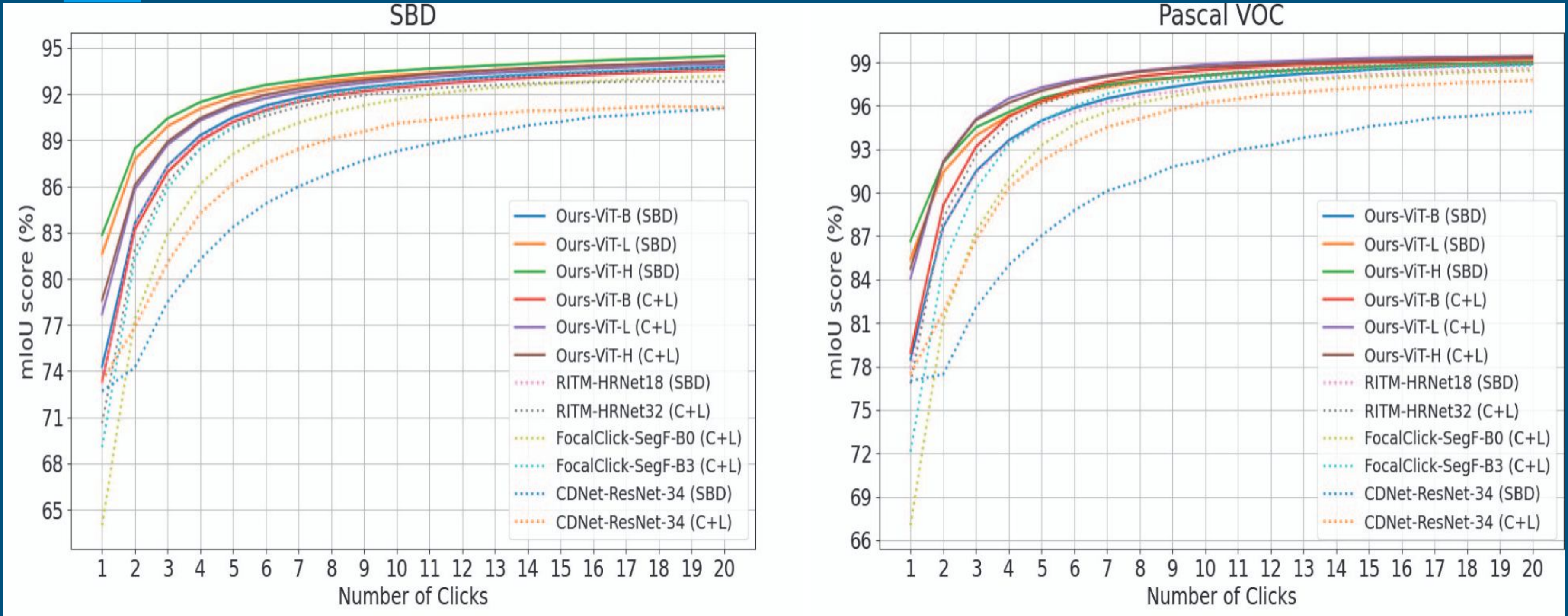
Experiments

❖ Comparison with Previous Results:

(a) the worst case (b) a challenging case (c) five normal cases

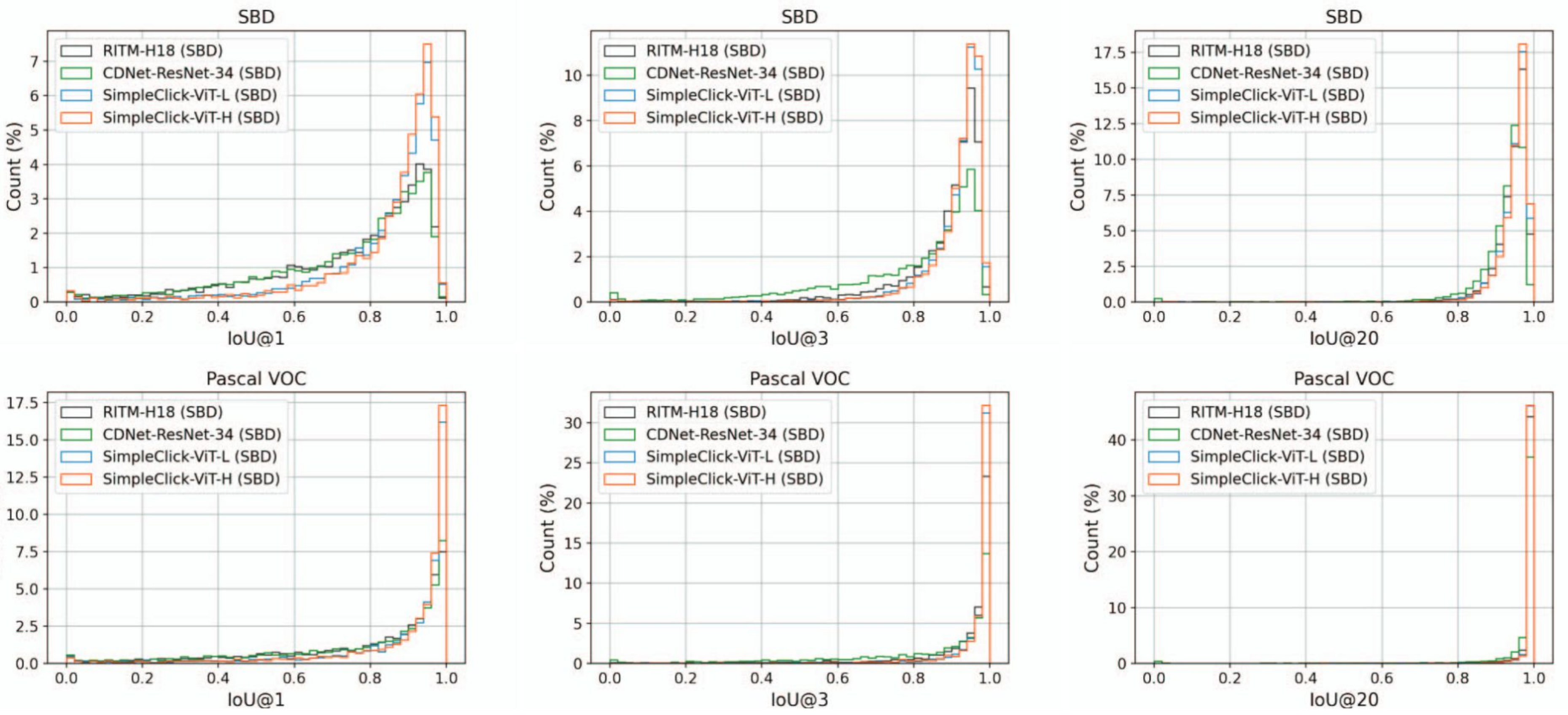
Experiments

❖ Comparison with Previous Results:



Experiments

❖ Comparison with Previous Results:



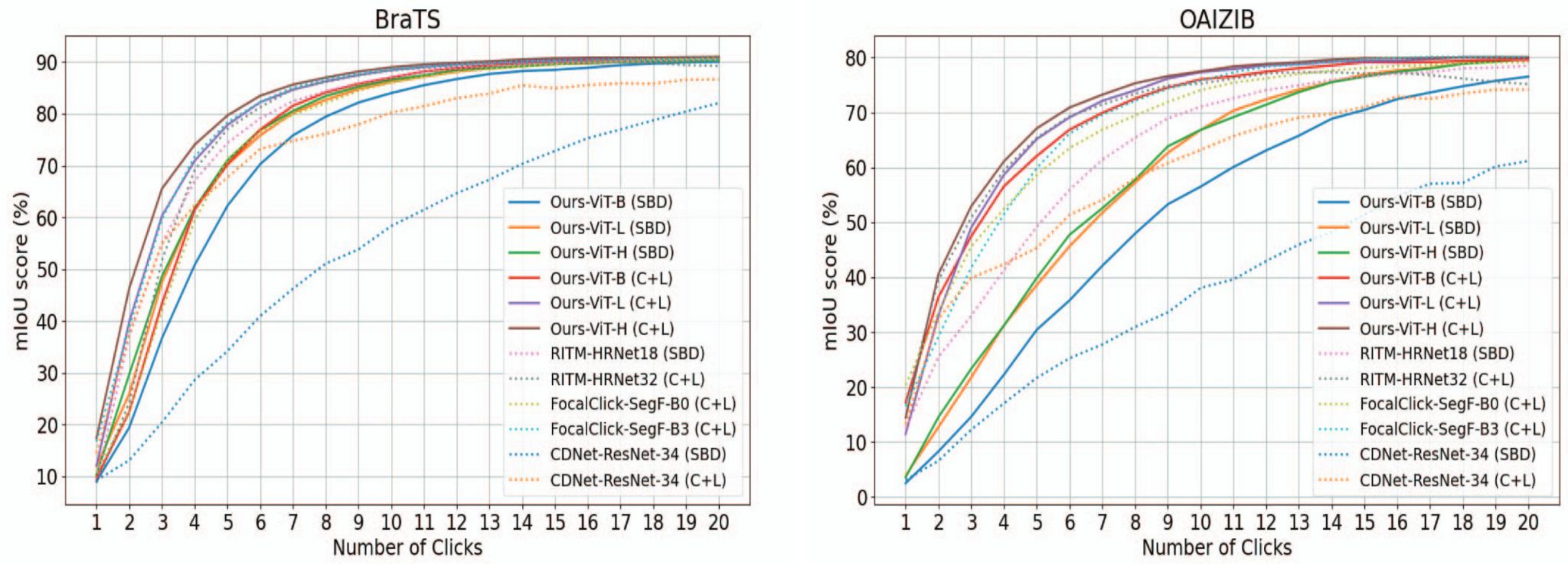
Experiments

- ❖ Out-of-Domain Evaluation on Medical Images:
  - ➢ generalizability of models on 3medical image datasets: ssTEM , BraTS ,and OAIZIB
  - ➢ Models trained on large datasets (i.e. C+L) generalize better than the models trained on smaller datasets (i.e. SBD).
  - ➢ models generalize very well on the three datasets, without finetuning

Model	ssTEM mIoU@10	BraTS mIoU@10 / 20	OAIZIB mIoU@10 / 20
♩ RITM-H18 [43]	93.15	87.05 / 90.47	71.04 / 78.52
♩ CDN-RN34 [9]	66.72	58.34 / 82.07	38.07 / 61.17
♫ RITM-H32 [43]	94.11	88.34 / 89.25	75.27 / 75.18
♫ CDN-RN34 [9]	88.46	80.24 / 86.63	63.19 / 74.21
♫ FC-SF-B0 [10]	92.62	86.02 / 90.74	74.08 / 79.14
♫ FC-SF-B3 [10]	93.61	88.62 / 90.58	75.77 / 80.08
♫ Ours-ViT-B	93.72	86.98 / 90.67	76.05 / 79.61
♫ Ours-ViT-L	**94.34**	88.43 / 90.84	77.34 / 79.97
♫ Ours-ViT-H	94.08	**88.98 / 91.00**	**77.50 / 80.10**

Experiments

❖ Out-of-Domain Evaluation on Medical Images:



Experiments

- Towards Practical Annotation Tool:
  - For low-end devices with limited computational resources->extremely tiny backbone (i.e. ViT-xTiny) for Simple Click (decreases the embedding dimension from 768 to 160 and the number of attention blocks from 12 to 8)
  - trained from scratch but outperforms Focal Click models

Model	Backbone	Pretrained	Params/M	NoC85	NoC90
FocalClick	HRNet-18s-S1	✓	4.22	4.74	7.29
FocalClick	SegFormer-B0-S1	✓	3.72	4.98	7.60
SimpleClick	ViT-xTiny	✗	3.72	4.71	7.09

Experiments

- Towards Practical Annotation Tool:

Backbone	Params/M	FLOPs/G	Mem/G	↓SPC/ms
HR-18s $_{400}$ [43]	4.22	17.94	0.50	54
HR-18 $_{400}$ [43]	10.03	30.99	0.52	56
HR-32 $_{400}$ [43]	30.95	83.12	1.12	86
Swin-B $_{400}$ [33]	87.44	138.21	1.41	36
Swin-L $_{400}$ [33]	195.90	302.78	2.14	44
SegF-B0 $_{256}$ [10]	3.72	3.42	0.10	37
SegF-B3 $_{256}$ [10]	45.66	24.75	0.32	53
ResN-34 $_{384}$ [9]	23.47	113.60	0.25	34
ResN-50 $_{384}$ [30]	40.36	78.82	0.85	331
ResN-101 $_{384}$ [30]	59.35	100.76	0.89	355
Ours-ViT-xT $_{224}$	3.72	2.63	0.17	17
Ours-ViT-xT $_{448}$	3.72	10.52	0.23	29
Ours-ViT-B $_{224}$	96.46	42.44	0.51	34
Ours-ViT-B $_{448}$	96.46	169.78	0.87	54
Ours-ViT-L $_{448}$	322.18	532.87	1.72	86
Ours-ViT-H $_{448}$	659.39	1401.93	3.22	132

Experiments

- Ablation Study:
  - Ablation of the backbone fine-tuning and feature pyramid design
  - freeze the backbone during finetune-ing -> worse performance
  - comparing the default simple feature pyramid design with three variants
    - ablating the multi-scale property in the simple feature pyramid->worse performance
    - The parallel feature pyramid generated by multi-stage feature maps from the backbone does not surpass the simple feature pyramid

FP design	frozen ViT	ViT-B		ViT-L	
		SBD	Pascal	SBD	Pascal
(a) simple FP	✓	11.48	6.93	9.75	5.59
(a) simple FP	✗	5.24	2.53	4.46	2.15
(b) single-scale	✗	6.56	2.80	5.53	2.48
(c) parallel	✗	7.21	3.09	6.26	2.79
(d) partial	✗	8.29	4.34	7.51	4.25

Limitations and Remarks

- ❖ 1- (ViT-H) is much larger than existing models-> unfair comparison.
- ❖ 2- method is not prompt-efficient ->every click update requires recomputing image features.
- ❖ 3-The recent advancements methods use sparse vectors to represent clicks, (more efficient than dense masks)
- ❖ 4- models may fail in some challenging scenarios such as objects with very thin and elongated shapes or cluttered occlusions

Conclusion

- ❖ SimpleClick is a plain-backbone model for interactive image segmentation.
  - ➢ method uses a general-purpose ViT backbone that can benefit from advancements in pretrained ViT models.
  - ➢ With the readily- available MAE weights, SimpleClick achieved state-of- the-art performance on natural images and demonstrated strong generalizability to medical images.
  - ➢ method is simple yet effective, highlighting its suitability as a strong baseline model and a practical annotation tool.

Thanks