RepViT-SAM: Towards Real-Time Segmenting Anything

Ao Wang[1]   Hui Chen[1*]   Zijia Lin[1]   Jungong Han[2]   Guiguang Ding[1*]

[1]Tsinghua University   [2] The University of Sheffield

wa22@mails.tsinghua.edu.cn   huichen@mail.tsinghua.edu.cn   linzijia07@tsinghua.org.cn

jungonghan77@gmail.com   dinggg@tsinghua.edu.cn

Abstract

Segment Anything Model (SAM) [13] has shown impressive zero-shot transfer performance for various computer vision tasks recently [3, 9, 19, 26, 28]. However, its heavy computation costs remain daunting for practical applications. MobileSAM [27] proposes to replace the heavyweight image encoder in SAM with TinyViT [27] by employing distillation, which results in a significant reduction in computational requirements. However, its deployment on resource-constrained mobile devices still encounters challenges due to the substantial memory and computational overhead caused by self-attention mechanisms. Recently, RepViT [21] achieves the state-of-the-art performance and latency trade-off on mobile devices by incorporating efficient architectural designs of ViTs into CNNs. Here, to achieve real-time segmenting anything on mobile devices, following [27], we replace the heavyweight image encoder in SAM with RepViT model, ending up with the RepViT-SAM model. Extensive experiments show that RepViT-SAM can enjoy significantly better zero-shot transfer capability than MobileSAM [27], along with nearly $10\times$ faster inference speed. The code and models are available at https://github.com/THU-MIG/RepViT.

1. Methodology

SAM [13] is composed of a heavyweight ViT-based image encoder and a lightweight prompt-guided mask decoder. Its huge image encoder accounts for the majority of the inference time overhead. Therefore, MobileSAM [27] suggests to replace the default ViT-H [6] image encoder in SAM with a lightweight one, *i.e.*, TinyViT [24]. TinyViT consists of four stages which gradually reduce the resolution. The initial stage of TinyViT is composed of convolution blocks utilizing inverted residual blocks [20]. To downsample the resolution at the outset of the model, two convolution blocks with a stride of 2 are employed. Similarly, the

---

*Corresponding author.

Table 1. Comparison between RepViT-SAM and others in terms of latency. The latency (ms) is measured with the standard resolution [7] of 1024×1024 on iPhone 12 and Macbook M1 Pro by Core ML Tools. OOM means out of memory.

Platform	Image encoder			Mask decoder
	RepViT-SAM	MobileSAM [27]	ViT-B-SAM [13]	
iPhone	**48.9**	OOM	OOM	11.6
Macbook	**44.8**	482.2	6249.5	11.8

convolution blocks with a stride of 2 is adopted to perform spatial downsampling between adjacent stages. In order to align the final resolution of TinyViT with that of the ViT-H image encoder of the original SAM, MobileSAM sets the stride of the last downsampling convolution in the TinyViT to 1. Besides, MobileSAM presents the decoupled distillation strategy to efficiently train the lightweight image encoder, in which the TinyViT model is directly distilled from the ViT-H in the original SAM without the prompt-guided mask decoder. Despite significantly reducing the computational requirements of segmenting anything, the deployment of MobileSAM on mobile devices still poses considerable challenges. As shown in Table 1, MobileSAM fails to run on an iPhone 12 due to its substantial memory footprint. Besides, on a Macbook, its inference time for processing a single image is 494ms, indicating significant room for improvement.

Recently, RepViT [21] demonstrates the state-of-the-art performance and latency trade-off on mobile devices by revisiting the efficient designs of CNNs from ViT perspective. RepViT employs the early convolutions [25] as the stem, *i.e.*, two convolutions with a stride of 2 to perform $4\times$ downsampling for the input. It adopts the RepViT block, which is composed of structural reparameterized depthwise convolutions [4, 5] and feed forward module. A deep downsample module is employed between adjacent stages, which leverage a depthwise convolution with a stride of 2 and a pointwise convolution to perform spatial downsampling and channel dimension modulation, respectively. Besides, the squeeze-and-excitation [8] layers is utilized for

1

Table 2. Comparison results on zero-shot edge detection. Bold indicates the best, and underline indicates the second best.

Model	zero-shot edge detection		
	ODS	OIS	AP
ViT-H-SAM [13]	**.768**	**.786**	**.794**
ViT-B-SAM [13]	.743	.764	.726
MobileSAM [27]	.756	.768	.746
RepViT-SAM	.764	**.786**	.773

Table 3. Comparison results on zero-shot instance segmentation and segmentation in the wild benchmark (SegInW). Bold indicates the best, and underline indicates the second best.

Model	zero-shot instance segmentation				SegInW
	AP	AP^S	AP^M	AP^L	Mean AP
ViT-H-SAM [13]	**46.8**	**31.8**	**51.0**	**63.6**	**48.7**
ViT-B-SAM [13]	42.5	29.8	47.0	56.8	44.8
MobileSAM [27]	42.7	27.0	46.5	61.1	43.9
RepViT-SAM	44.4	29.1	48.6	61.4	46.1

Table 4. Comparison results on zero-shot video object segmentation (z.s. VOS) and video instance segmentation (z.s. VIS). Bold indicates the best, and underline indicates the second best.

Model	z.s. VOS			z.s. VIS
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	AR100
ViT-H-SAM [13]	**77.4**	**74.6**	**80.2**	**28.8**
ViT-B-SAM [13]	71.3	68.5	74.1	19.1
MobileSAM [27]	71.1	68.6	73.6	22.7
RepViT-SAM	73.5	71.0	76.1	25.3

all stages in a cross-block manner. RepViT shows substantial advantages in terms of latency in high-resolution vision tasks [21], due to its pure convolutional architecture. As shown in Table 1, after replacing the ViT-H image encoder with the RepViT-M2.3 model, RepViT-SAM exhibits a significant reduction in latency compared with others. On iPhone 12, RepViT-SAM can perform model inference smoothly. Besides, on Macbook, RepViT-SAM is nearly 10× faster than MobileSAM.

Following [27], we train the RepViT-SAM by directly distilling the image encoder, *i.e.*, RepViT-M2.3, from the ViT-H in the original SAM [13], leveraging a simple MSE loss. Like [27], the stride of 2 in the last downsampling depthwise convolution in RepViT is set to 1 for making the output resolution compatible with the prompt-guided mask decoder in the original SAM [13].

2. Experiments

2.1. Implementation details

RepViT-SAM is trained for 8 epochs under the same setting as [27]. Like MobileSAM [27], we use only 1% data in the SAM-1B dataset [13]. To expedite the training process, we precompute and save the image embeddings from the ViT-H image encoder prior to the distillation phase, which eliminates the need to run the forward process of ViT-H during distillation, like [27]. We evaluate the performance of RepViT-SAM on zero-shot edge detection on BSDS500 [1, 17], zero-shot instance segmentation using COCO [14], and segmentation in the wild benchmark [29] (SegInW), zero-shot video object/instance segmentation using DAVIS 2017 [18]/UVO v1.0 [23], zero-shot salient object segmentation using DUTS [22], and zero-shot anomaly detection using MVTec-AD [2], following [3, 9, 12, 13, 19].

2.2. Zero-shot edge detection

Following [13], we utilize a 16×16 regular grid of foreground points as prompts for the model. We apply NMS to remove redundant masks. Additionally, a Sobel filter is employed to the unthreshold probability maps of the masks, along with standard lightweight postprocessing, including edge NMS, as described in [13]. As shown in Table 2, our

RepViT-SAM outperforms MobileSAM and ViT-B-SAM on all metrics. Compared with ViT-H-SAM, which is the largest SAM model with over 615M parameters, our small RepViT-SAM can obtain comparable performance in terms of ODS and OIS.

2.3. Zero-shot instance segmentation

We leverage the state-of-the-art H-Deformable-DETR [10] with Swin-L [16] as the object detector and prompt the model with its output boxes, like [12, 13]. As shown in Table 3, our RepViT-SAM significantly outperforms Mobile-SAM and ViT-B-SAM by 1.7 and 1.9 AP, respectively. Besides, it can be observed that the models obtained by the decoupled distillation strategy, *i.e.*, MobileSAM and RepViT-SAM, typically outperform the model obtained by the coupled distillation, *i.e.*, ViT-B-SAM, on large objects, but exhibits comparatively poorer performance on small objects. Such phenomenon may be attributed to the fact that the decoupled distillation strategy enables more transfer of macro-level visual features but falls short in fine-grained features.

2.4. Segmentation in the wild benchmark

We follow [12] to equip the Grounding-DINO [15] as box prompts to evaluate the model on the zero-shot track. As shown in Table 3, our RepViT-SAM outperforms Mobile-SAM and ViT-B-SAM by 2.2 and 1.3 mean AP, respectively. It well indicates the strong transferability of RepViT-SAM to diverse downstream segmentation tasks.

Table 5. Comparison results on zero-shot salient object segmentation (z.s. s.o.s.) and zero-shot anomaly detection (z.s. a.d.). Bold indicates the best, and underline indicates the second best.

Model	z.s. s.o.s.	z.s. a.d.
	$\mathcal{M} \downarrow$	$\mathcal{F}_p$
ViT-H-SAM [13]	**0.046**	<u>37.65</u>
ViT-B-SAM [13]	0.121	36.62
MobileSAM [27]	0.147	36.44
RepViT-SAM	<u>0.066</u>	**37.96**

2.5. Zero-shot video object/instance segmentation

We leverage SAM-PT [19] to evaluate the performance of RepViT-SAM on the validation splits of DAVIS 2017 [18] and UVO [23] datasets. We simply replace the original SAM in SAM-PT with our RepViT-SAM, using the Co-Tracker [11] as the point tracker without the reinitialization strategy. As shown in Table 4, on DAVIS 2017, our RepViT-SAM obtains a mean $\mathcal{J}\&\mathcal{F}$ score of 73.5, significantly outperforming MobileSAM and ViT-B-SAM by 2.4 and 2.2, respectively. Besides, on UVO, RepViT-SAM surpasses MobileSAM and ViT-B-SAM with considerable margins of 2.6 and 6.2, respectively. These results show the superiority of RepViT-SAM compared with other lightweight ones is not limited to image segmentation tasks but can also be extended to video segmentation tasks.

2.6. Zero-shot salient object segmentation

We follow [9] to first leverage RepViT-SAM to generate N potential objects masks for a given image. Subsequently, we select the most appropriate mask by evaluating its alignment with the ground truth. The intersection over union (IoU) score is adopted as the alignment metric. Following [9], we report mean absolute error (MAE) score. Note that a lower MAE indicates better segmentation results. As shown in Table 5, our RepViT-SAM shows its superiority compared with MobileSAM and ViT-B-SAM in terms of MAE score.

2.7. Zero-shot anomaly detection

We leverage SAA+ [3] to evaluate the performance of RepViT-SAM on the MVTec-AD benchmark dataset. We simply replace the original SAM in SAA+ with our RepViT-SAM. We follow [3] to report max-F1-pixel ($\mathcal{F}_p$), which quantifies the F1-score for pixel-wise segmentation at the optimal threshold. As shown in Table 5, our RepViT-SAM achieves the best performance in terms of $\mathcal{F}_p$, even outperforming ViT-H-SAM. It indicates that for certain specific downstream task, the decoupled distilled model can achieve performance than the original SAM, along with much lower latency, showing the promising applications of RepViT-SAM.

3. Visualization

In Figure 1, we showcase the predicted masks of SAM [13], MobileSAM [27], and RepViT-SAM with point and box prompts. It shows that RepViT-SAM can generate high-quality mask predictions similar to SAM across various scenarios, where MobileSAM may encounter challenges in accurately predicting masks. Furthermore, we present qualitative comparison results for zero-shot edge detection in Figure 2. RepViT-SAM effectively produces reasonable edge maps, achieving performance comparable to SAM. In contrast, MobileSAM may struggle to generate accurate edge maps in regions that contain intricate details.

4. Conclusion

In summary, our RepViT-SAM demonstrates outstanding efficiency while maintaining impressive transfer performance for various downstream tasks. We hope that our RepViT-SAM model can serve as a robust baseline for SAM in edge deployments, showcasing its potential for practical applications.

References

[1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010. 2

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 2

[3] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023. 1, 2, 3

[4] Xiangxiang Chu, Liang Li, and Bo Zhang. Make repvgg greater again: A quantization-aware approach. *arXiv preprint arXiv:2212.01593*, 2022. 1

[5] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13733–13742, 2021. 1

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[7] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF con-*
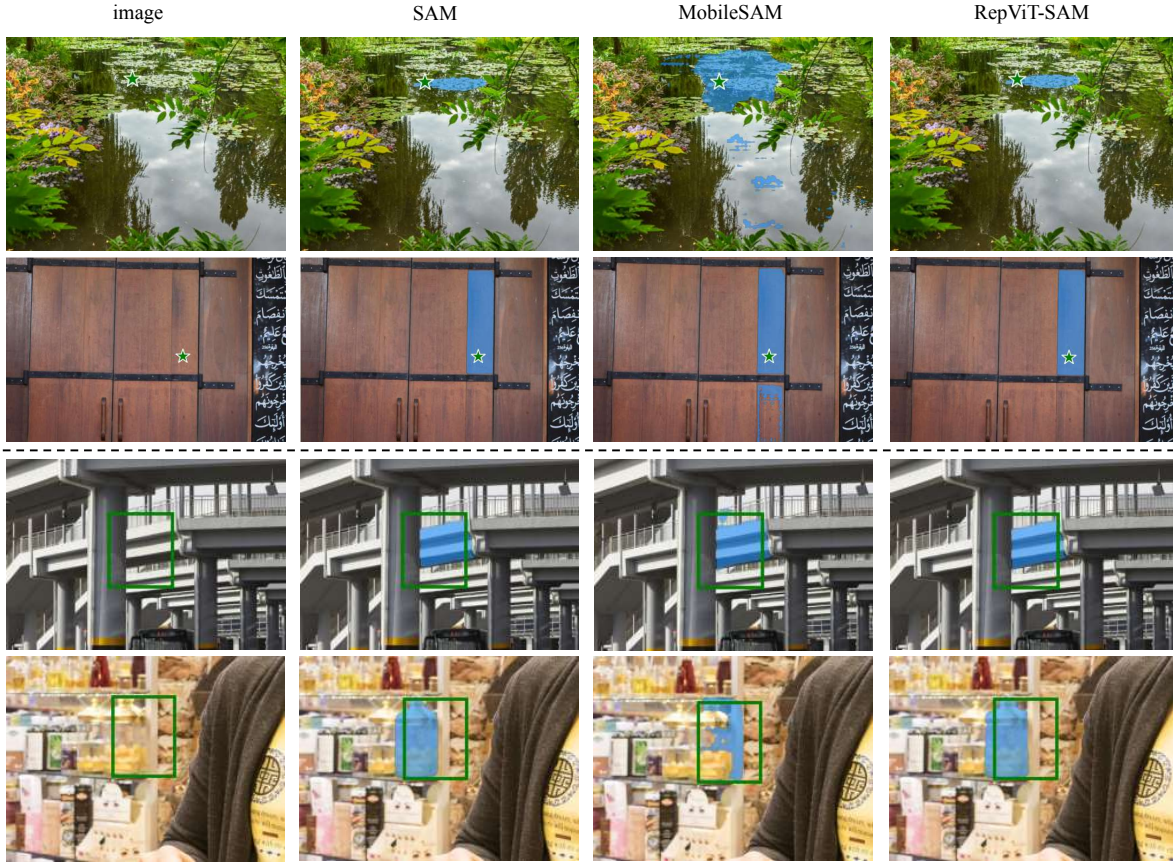
Figure 1. Mask predictions of SAM, MobileSAM, and RepViT-SAM with point prompts (top) and box prompts (bottom).
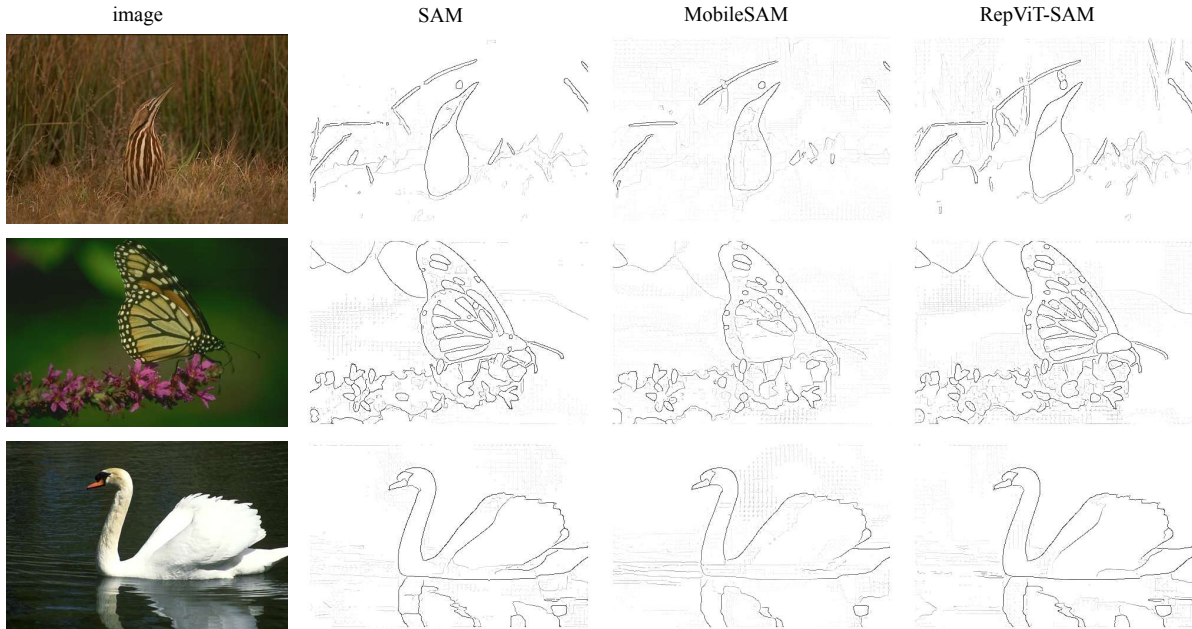

Figure 2. Visualization results of SAM, MobileSAM, and RepViT-SAM for zero-shot edge detection on BSDS500.

ference on computer vision and pattern recognition, pages 2918–2928, 2021. 1

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1

[9] Wei Ji, Jingjing Li, Qi Bi, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *arXiv preprint arXiv:2304.05750*, 2023. 1, 2, 3

[10] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19702–19712, 2023. 2

[11] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-tracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023. 3

[12] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. *arXiv preprint arXiv:2306.01567*, 2023. 2

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 2, 3

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[15] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[17] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, pages 416–423. IEEE, 2001. 2

[18] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 2, 3

[19] Frano Rajič, Lei Ke, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Segment anything meets point tracking. *arXiv preprint arXiv:2307.01197*, 2023. 1, 2, 3

[20] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the*

[21] Ao Wang, Hui Chen, Zijia Lin, Hengjun Pu, and Guiguang Ding. Repvit: Revisiting mobile cnn from vit perspective. *arXiv preprint arXiv:2307.09283*, 2023. 1, 2

[22] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 136–145, 2017. 2

[23] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10776–10785, 2021. 2, 3

[24] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *European Conference on Computer Vision*, pages 68–85. Springer, 2022. 1

[25] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34:30392–30400, 2021. 1

[26] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. *arXiv preprint arXiv:2312.00863*, 2023. 1

[27] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 1, 2, 3

[28] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023. 1

[29] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. 2