Appendix of "GraCo: Granularity-Controllable Interactive Segmentation"

1. Limitations

In this work, we introduce **Gra**nularity-**Co**ntrollable interactive segmentation (**GraCo**) that allows users to control the segmentation granularity to resolve ambiguity. Although we develop a novel and flexible paradigm and achieve inspiring results, the proposed method still has some limitations: (i). Due to the randomness in the interaction signals generated by the multi-granularity loop simulation in the any-granularity mask generator, which causes the object-level pre-trained IS model to generate semantically inconsistent parts or noisy boundaries, providing inaccurate granularity-controllability guidance. (ii). Considering the variance in the computational cost of running the mask engine at different granularities, we choose to generate proposals offline to improve the efficiency of parallel computing. As a result, there is a trade-off between storage space and granularity abundance. The online fine-tuning paradigm of granularity-controllability is a future exploration to overcome this limitation.

2. Additional Experiments and Analysis

2.1. IoU@1 Analysis

Considering that the segmentation mask after the first click directly affects the user experience, we evaluate the IoU@1 of the IS methods. As shown in Table A, we evaluate the IoU@1 of SimpleClick [6], SAM [5] and our GraCo. For SimpleClick, we report the results of the pre-trained model and the model fine-tuned with part annotations. From the results, we conclude that fine-tuning with part annotations leads to a significant decrease in IoU@1 on object-level benchmarks. In contrast, the results on part-level benchmarks are effectively improved, indicating that the model tends to perform fine-grained part segmentation after fine-tuning. For SAM, we present the results for single-output and multi-output (default 3) respectively. We observe that SAM exhibits excellent performance. Specifically, the first click performance of SAM is significantly superior than SimpleClick, especially when selecting the optimal mask from multiple outputs for each instance. Moreover, the IoU@1 obtained by multi-output outperforms single-output considerably, as denoted by the green-highlighted increment. This enhances SAM's user experience. For our GraCo, we present the results of fine-tuning with part annotations and AGG-generated mask proposals respectively. We observe that GraCo w/ AGG is superior than GraCo w/ GT. We argue that this is because AGG generates a wealth of mask proposals to cover a wider range of granularity. Our GraCo achieves comparable first click performance to SAM on all benchmarks at a low cost.

2.2. More Ablations

Proposal Sampling. We also conduct an ablation study on the proposal sampling. We compare the performance of uniform sampling to inverse-proportional sampling with identical mask proposals (*cf.* Table B). The results show that the inverse-proportional sampling method achieves a superior performance on all benchmarks, which indicates that the method enables the IS model to learn uniformly from any-granularity proposals in GCL.

LoRA. We supplement the ablation study on LoRA, as shown in Table C. We employ identical AGG-generated mask proposals to train our GraCo equipped with ViT-B as backbone. We set the LoRA rank as 4, 8, 16, 32, respectively, and evaluate the performance on both levels of benchmarks. Based on the results, we conclude that the performance of GraCo is not sensitive to the LoRA rank.

Granularity Definition. We evaluate the performance of the two definitions on part-level benchmarks, which indicates that employing only scale granularity leads to a slight decrease (*cf.* Table D). This demonstrates the necessity of the two types of granularity for definition.

3. Dataset Description

We evaluate both object-level and part-level benchmarks to demonstrate the performance of the IS model in multi-granularity scenarios. The details of these datasets are described as follows.

- **GrabCut** [9]. The dataset contains 50 images, each containing a single instance.
- **Berkeley** [7]. The dataset contains 96 images with 100 instances and some of them are more challenging for segmentation.
- **SBD** [3]. The dataset contains 2,857 images with 6,671 challenging instances for evaluation and not be used for training.
- **DAVIS** [8]. The dataset contains 50 high-quality videos and we use 345 frames for evaluation.
- **PascalPart** [1]. The dataset provides part annotations of 20 Pascal VOC [2] classes, a total of 193 part categories. As PascalPart contains a large number of parts, we randomly select 5 out of 16 classes (excluding boat, chair, dining table, and sofa which do not have part annotations) to reduce the computational cost of conducting interactive simulations during evaluation. The selected classes are train, bicycle, cow, aeroplane, and bus in experiments.

Method	Backbone	GrabCut	Berkeley	SBD	DAVIS	PascalPart	PartImageNet
SimpleClick [6]	ViT-B	0.90	0.85	0.74	0.76	0.17	0.30
SimpleClick¶ [6]	ViT-B	0.47 (↓ **0.43**)	0.43 (↓ **0.42**)	0.42 (↓ **0.32**)	0.31 (↓ **0.45**)	0.48 (↑ **0.31**)	0.49 (↑ **0.19**)
SimpleClick [6]	ViT-L	0.91	0.84	0.82	0.78	0.18	0.30
SimpleClick¶ [6]	ViT-L	0.48 (↓ **0.43**)	0.46 (↓ **0.38**)	0.46 (↓ **0.36**)	0.38 (↓ **0.40**)	0.53 (↑ **0.35**)	0.54 (↑ **0.24**)
SAM [5]	ViT-B	0.55	0.56	0.45	0.41	0.43	0.42
SAM⋆ [5]	ViT-B	0.90 (↑ **0.35**)	0.88 (↑ **0.32**)	0.75 (↑ **0.30**)	0.74 (↑ **0.33**)	0.57 (↑ **0.14**)	0.55 (↑ **0.13**)
SAM [5]	ViT-L	0.61	0.61	0.50	0.45	0.44	0.42
SAM⋆ [5]	ViT-L	0.94 (↑ **0.33**)	0.90 (↑ **0.29**)	0.80 (↑ **0.30**)	0.78 (↑ **0.33**)	0.57 (↑ **0.13**)	0.56 (↑ **0.14**)
GraCo w/ GT	ViT-B	0.86	0.80	0.66	0.62	0.52	0.53
GraCo w/ AGG	ViT-B	0.89 (↑ **0.03**)	0.84 (↑ **0.04**)	0.72 (↑ **0.06**)	0.70 (↑ **0.08**)	0.53 (↑ **0.01**)	0.55 (↑ **0.02**)
GraCo w/ GT	ViT-L	0.81	0.76	0.66	0.56	0.56	0.55
GraCo w/ AGG	ViT-L	0.93 (↑ **0.12**)	0.89 (↑ **0.13**)	0.81 (↑ **0.15**)	0.75 (↑ **0.19**)	0.55 (↓ **0.01**)	0.58 (↑ **0.03**)

Table A. **IoU@1 Analysis on both object and part level benchmarks.** ¶ represents fine-tuning the model utilizing the part annotation, and ⋆ represents selecting the best matching result from multiple predictions. SimpleClick [6] and our GraCo are trained on SBD [3] and SAM are trained on SA-1B [5]. SimpleClick and SAM are from official models and use specific data pre-processing pipeline.

Sampling	GrabCut			Berkeley			SBD			PascalPart	
	NoC@85↓	NoC@90↓	IoU@1↑	NoC@85↓	NoC@90↓	IoU@1↑	NoC@85↓	NoC@90↓	IoU@1↑	NoC@85↓	IoU@1↑
Uniform	1.46	1.52	0.86	1.41	2.29	0.83	3.49	4.93	0.70	6.44	0.52
Inverse-prop.	**1.34**	**1.46**	**0.89**	**1.37**	**2.21**	**0.84**	**3.44**	**4.89**	**0.72**	**6.38**	**0.53**

Table B. **Results of ablation study on proposal sampling.**

LoRA Rank	GrabCut		Berkeley		SBD		DAVIS		PascalPart	PartImageNet
	NoC@85	NoC@90	NoC@85	NoC@90	NoC@85	NoC@90	NoC@85	NoC@90	NoC@85	NoC@85
4	1.36	1.48	1.43	2.25	3.48	4.93	4.62	5.84	6.47	6.03
8	1.34	1.46	**1.37**	**2.21**	3.44	4.89	4.44	5.72	**6.38**	**6.01**
16	**1.32**	**1.44**	1.40	2.23	3.45	4.90	4.68	5.85	6.39	6.03
32	1.32	1.44	1.37	2.24	**3.40**	**4.85**	**4.41**	**5.70**	6.42	6.03

Table C. **Ablation study on LoRA.** We train our GraCo on the same AGG-generated proposals with different ranks of the LoRA. We utilize ViT-B as the backbone. **Bold** indicates the best performance and underlined the second best.

Granularity Definition	PascalPart		PartImageNet	
	NoC@85↓	IoU@1↑	NoC@85↓	IoU@1↑
Scale-only	6.43	0.52	6.08	0.54
Scale & Semantic	**6.38**	**0.53**	**6.01**	**0.55**

Table D. **Results of ablation study on granularity definition.**

- **PartImageNet** [4]. The dataset groups 158 classes from ImageNet [10] into 11 super-categories and provides a total of 40 part categories, which is a large, high-quality dataset for part segmentation, offering part-level annotations on a broad range of classes, including non-rigid, articulated objects. We use the validation set of PartImageNet to evaluate the performance of IS model at the part-level, which includes 1206 images and 5626 parts.
- **SA-1B** [5]. The dataset consists of 11M high-resolution (3300×4950 pixels on average), diverse, and licensed images and 1.1B high-quality segmentation masks. To alleviate storage pressure, released images are downsampled and their shortest side is set to 1500 pixels. We use the first 1000 images to evaluate the performance of different methods.

4. Additional Qualitative Results

We supplement more examples to demonstrate the granularity controllability and excellent segmentation performance of our GraCo in multi-granularity scenarios, *cf.* Figure A. For complex scenarios, our GraCo allows the user to select the appropriate granularity to generate the required mask. Furthermore, our GraCo facilitates precise control over the expansion of segmentation masks through multiple positive clicks by applying a small granularity. This advantage effectively overcomes the limitations of current object-level IS methods (*e.g.*, SimpleClick [6]) when dealing with tiny or detached components. We also demonstrate the qualitative results of the proposed GraCo on four object-level benchmarks with a fixed input granularity of 1.0, *cf.* Figure B. Our GraCo achieves impressive qualitative results.

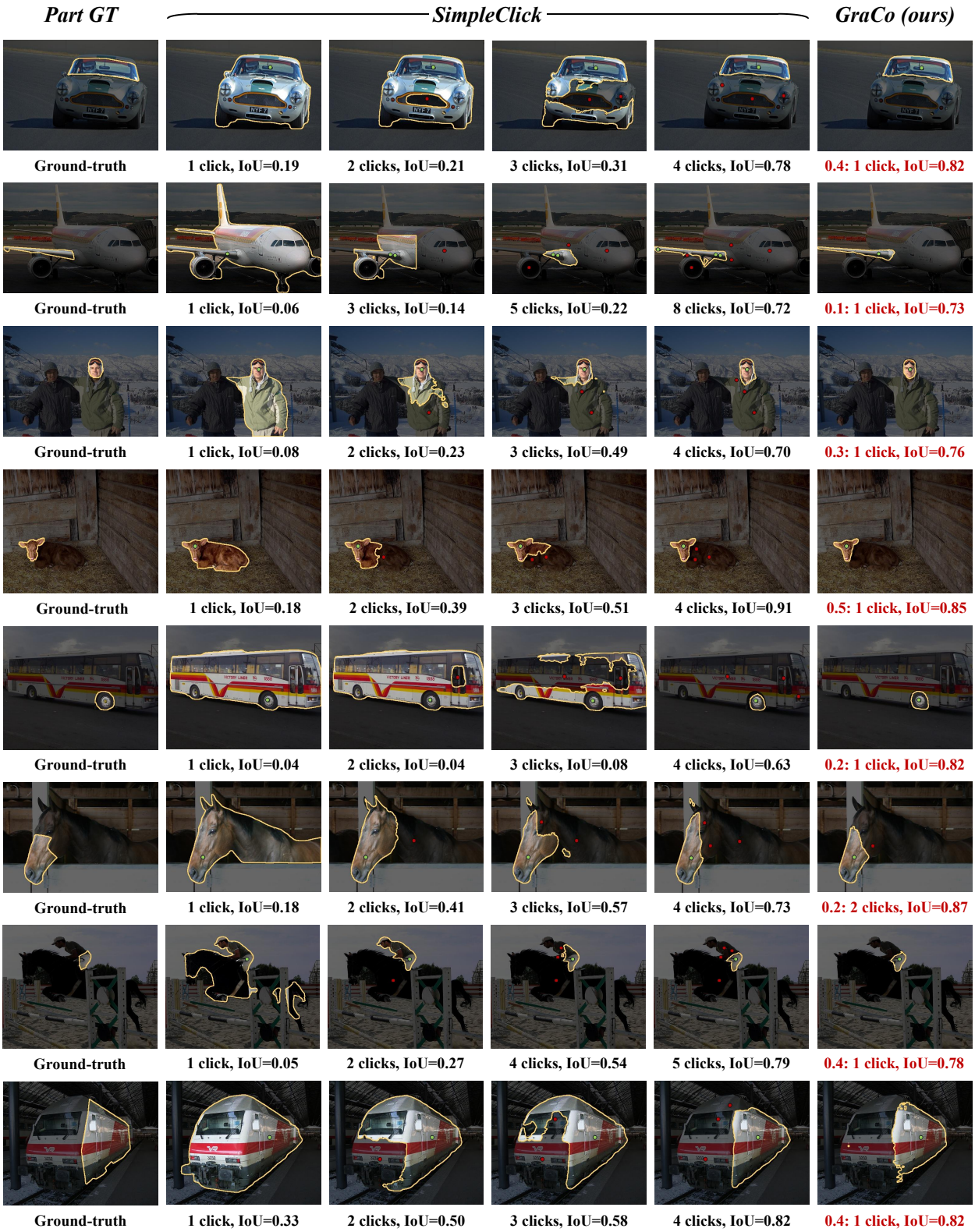

Figure A. **More visualization examples of interactive segmentation on part GT using SimpleClick** [6] **and our GraCo.** The proposed method satisfies the user's requirements with just one or two clicks.

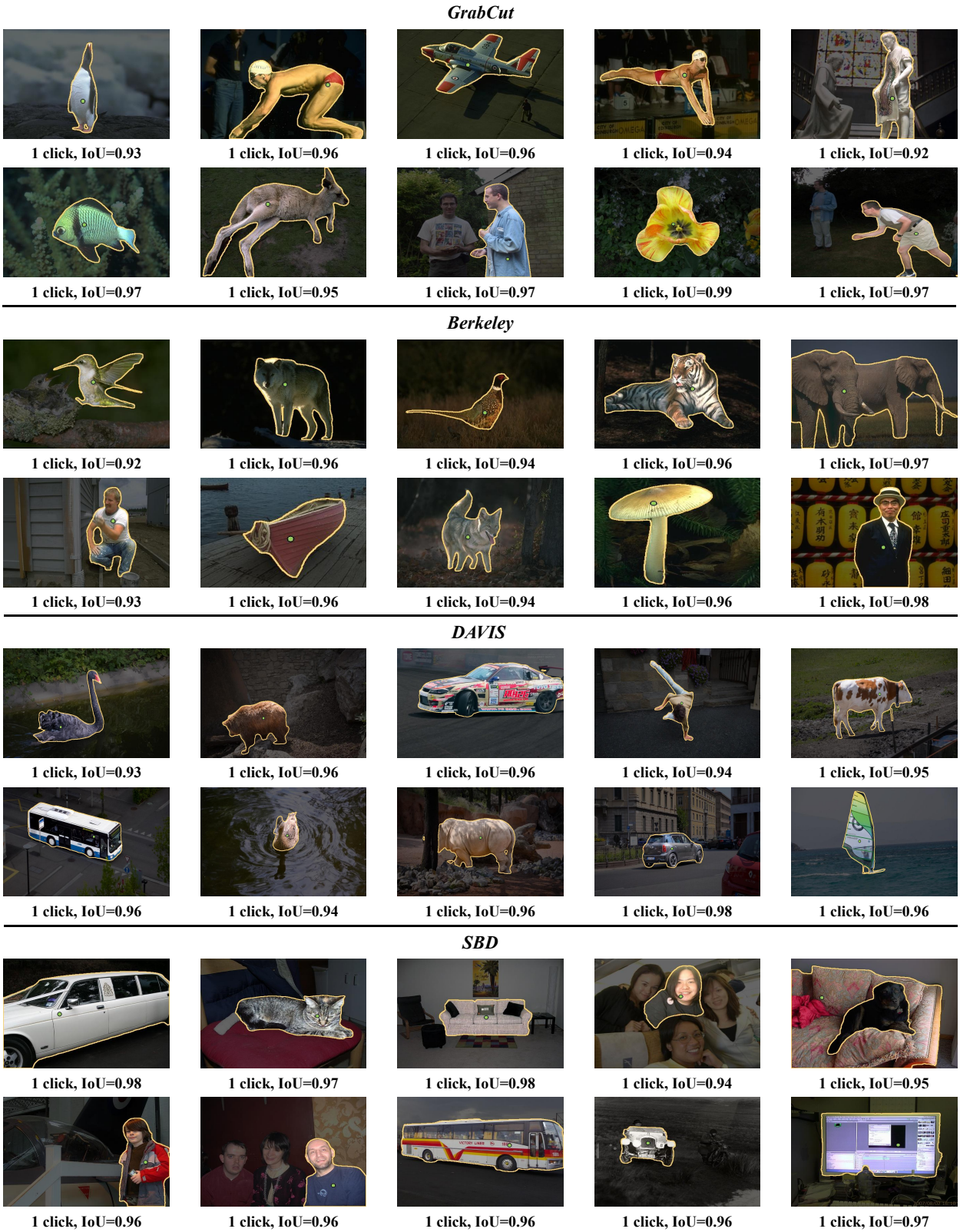

Figure B. **Visualization on four object-level benchmarks.** Note that the input granularity of GraCo is fixed to 1.0.

References

[1] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1971–1978, 2014. 1

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010. 1

[3] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 991–998. IEEE, 2011. 1, 2

[4] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022. 2

[5] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 2

[6] Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22290–22300, 2023. 1, 2, 3

[7] Kevin McGuinness and Noel E O'connor. A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, 43(2):434–444, 2010. 1

[8] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016. 1

[9] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3): 309–314, 2004. 1

[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 2