SkipClick: Combining Quick Responses and Low-Level Features for Interactive Segmentation in Winter Sports Contexts (Supplementary Material)

Robin Schön Julian Lorenz Daniel Kienzle
Rainer Lienhart
University of Augsburg
Germany, 86159 Augsburg, Universitätsstr. 6a
{robin.schoen, julian.lorenz, daniel.kienzle, rainer.lienhart}@uni-a.de

1. Simulating User Clicks

In this section we discuss how we simulate the clicks. More precisely, we want to answer the question: Given a predicted mask $\mathbf{m}_\tau$ and a ground truth mask $\mathbf{m}_{\text{GT}}$, where do we place the next click in order to help our network with improving the mask? We follow common practice and use the method described in [10].

1. For each click we simulate, we compare $\mathbf{m}_\tau$ with $\mathbf{m}_{\text{GT}}$ to obtain the mask of false positives $\mathbf{m}_{\text{FP}}$ and the mask of false negatives $\mathbf{m}_{\text{FN}}$.
2. Afterwards, we compute the euclidean distance transforms (see [4]) of both masks, $\mathcal{D}_{\text{FP}}$ and $\mathcal{D}_{\text{FN}}$.
3. We will then look for the maxima of $\mathcal{D}_{\text{FP}}$ and $\mathcal{D}_{\text{FN}}$. The coordinates of the higher maximum will be the location of the simulated click.
4. Depending on whether this maximum is found in either $\mathcal{D}_{\text{FP}}$ or $\mathcal{D}_{\text{FN}}$, we will label it as a background (-) or foreground (+) click, respectively.

It should be noted that this metric allows for improving the system at the cost of practical usability. If we were to simulate the clicks during training in the exact same way as we do during testing (taking the maximum of the two distance transforms), we would prepare our model to optimally perform under the metric. This can for example be seen in [5, 7]. As [10] however mentions, this inhibits the practical usability of the model, since an actual human would choose other non-optimal click positions. We would see a kind of overfitting to the metric. To make sure the training of our model adheres to practical requirements, we follow common practice [1, 6, 8–10] and use additional random clicks during each training step.

Depth	WSESeg Average	
	NoC@85	NoC@90
2	6.962	9.587
3	**6.311**	9.091
4	6.944	9.163
5	6.689	9.310
6	6.524	**9.023**

Table 1. A comparison of the change in performance for different numbers of ViT blocks. The *depth* does not refer to the backbone, but the additional blocks after mixing the image and prompt features. The NoC is the average over all classes.

2. Changing the Number of Encoder Blocks after Adding the Prompts

The prompt features and the image features in our architecture are fused by multiple transformer encoder blocks. In our standard model we chose four as the number of blocks. In Tab. 1 we compare the performance of the model when altering the number of blocks (the column *Depth*). We cannot observe a clear trend, as a continuous increase of the number of blocks does not necessarily cause an improvement. We even see that reducing the number of blocks to three gives a slightly better performance for a NoC@85 of 6.94 to 6.31, although the best performance depends on the metric, with 6.31 for the NoC@85 and 9.023 for the NoC@90.

3. Qualitative Examples from SHSeg

In Figure 1, we can see qualitative examples from our newly proposed SHSeg (Skiing Human Segmentation) dataset. Our dataset provides 534 masks for skiers on 496 images. The images have been randomly sampled from the SkiTB dataset [2, 3]. A link to the data can be found in our main paper (publication of data upon acceptance).

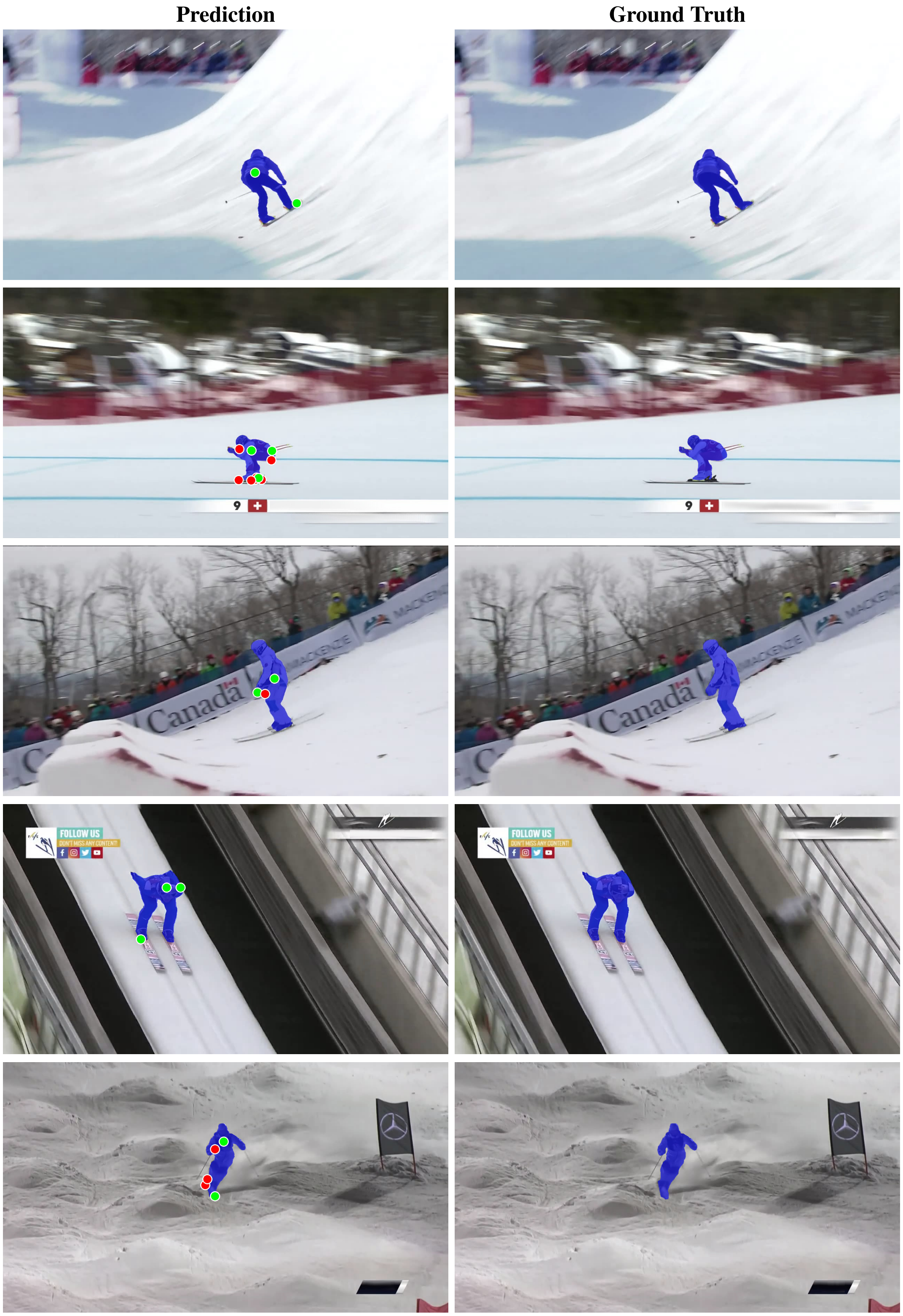

Figure 1. Examples for the masks occurring during the interaction. The *left column* displays the predicted mask along with the clicks. Foreground clicks are *green*, background clicks are *red* and the masks are *blue*. The *right column* displays the corresponding ground truth.

References

[1] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1300–1309, 2022. 1

[2] Matteo Dunnhofer and Christian Micheloni. Visual tracking in camera-switching outdoor sport videos: Benchmark and baselines for skiing. *Computer Vision and Image Understanding*, 243:103978, 2024. 1

[3] Matteo Dunnhofer, Luca Sordi, Niki Martinel, and Christian Micheloni. Tracking skiers from the top to the bottom. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8511–8521, 2024. 1

[4] Pedro F Felzenszwalb and Daniel P Huttenlocher. Distance transforms of sampled functions. *Theory of computing*, 8(1):415–428, 2012. 1

[5] You Huang, Hao Yang, Ke Sun, Shengchuan Zhang, Liujuan Cao, Guannan Jiang, and Rongrong Ji. Interformer: Real-time interactive image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22301–22311, 2023. 1

[6] 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 e-prints*, pages arXiv–2304, 2023. 1

[7] Chaewon Lee, Seon-Ho Lee, and Chang-Su Kim. Mfp: Making full use of probability maps for interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4051–4059, June 2024. 1

[8] 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 (ICCV)*, pages 22290–22300, October 2023. 1

[9] Qin Liu, Meng Zheng, Benjamin Planche, Srikrishna Karanam, Terrence Chen, Marc Niethammer, and Ziyan Wu. Pseudoclick: Interactive image segmentation with click imitation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 728–745, Cham, 2022. Springer Nature Switzerland. 1

[10] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE, 2022. 1