

WSESeg: introducing a dataset for the segmentation of winter sports equipment with a baseline for interactive segmentation

Robin Schön, Daniel Kienzle, Rainer Lienhart

Angaben zur Veröffentlichung / Publication details:

Schön, Robin, Daniel Kienzle, and Rainer Lienhart. 2024. "WSESeg: introducing a dataset for the segmentation of winter sports equipment with a baseline for interactive segmentation." In 2024 21st International Conference on Content-based Multimedia Indexing (CBMI), 18-20 September 2024, Reykjavik, Iceland, 1-7. Piscataway, NJ: IEEE.
<https://doi.org/10.1109/CBMI62980.2024.10859243>.

Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:

Deutsches Urheberrecht

Weitere Informationen finden Sie unter: / For more information see:

<https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publiz/>



WSESeg: Introducing a Dataset for the Segmentation of Winter Sports Equipment with a Baseline for Interactive Segmentation

Robin Schön

Fakultät für Angewandte Informatik
University of Augsburg
Augsburg, Germany
robin.schoen@uni-a.de

Daniel Kienzle

Fakultät für Angewandte Informatik
University of Augsburg
Augsburg, Germany
daniel.kienzle@uni-a.de

Rainer Lienhart

Fakultät für Angewandte Informatik
University of Augsburg
Augsburg, Germany
rainer.lienhart@uni-a.de

Abstract—In this paper we introduce a new dataset containing instance segmentation masks for ten different categories of winter sports equipment, called WSESeg (Winter Sports Equipment Segmentation)¹. Furthermore, we carry out interactive segmentation experiments on said dataset to explore possibilities for efficient further labeling. The SAM and HQ-SAM models are conceptualized as foundation models for performing user guided segmentation. In order to measure their claimed generalization capability we evaluate them on WSESeg. Since interactive segmentation offers the benefit of creating easily exploitable ground truth data during test-time, we are going to test various online adaptation methods for the purpose of exploring potentials for improvements without having to fine-tune the models explicitly. Our experiments show that our adaptation methods drastically reduce the Failure Rate (FR) and Number of Clicks (NoC) metrics, which generally leads faster to better interactive segmentation results.

Index Terms—interactive segmentation, instance segmentation, sports, winter sports, ski, snowboards, winter sports equipment, dataset

I. INTRODUCTION

There is an interest in the exact analysis of the pose and limb positions of a human being depicted in an image or video. The corresponding computer vision tasks are human pose estimation [1] and body part segmentation [2], [3], where progress mainly benefits the analysis of sports-related image and video data. In some cases deep learning applications can be applied for the post hoc analysis of the athletic performance with the aim of finding room for improvement. For this purpose it is necessary to detect the exact position of each limb of the body.

The authors of [4], [5] leverage segmentation masks for body parts of human athletes as well as equipment to train a network capable of localizing any desired keypoint on the human and their equipment. There is a wide availability of datasets containing human poses in a skeletal form and body part segmentation masks for human limbs. However, data concerning the equipment used in the sports is relatively scarce, which lead to the authors resorting to pseudo labels.

This insufficient availability of data particularly concerns the winter sports domain, which is why we created a dataset containing segmentation masks for various types of winter sports equipment worn or used by the respective athletes.

Most available datasets containing segmentation masks only provide these types of annotations for general consumer images. Often, there is a lack of segmentation masks for rare applications that considerably differ from the domain of general consumer images. In these cases, there will be a need to annotate ground truth segmentation masks for the new purpose. However, directly annotating these masks generally constitutes a considerable amount of work. This is especially the case when the segmentation masks are annotated by the means of drawing polygons (as in the case of COCO [6]) and the annotator has to draw masks for fine structures. In order to alleviate this problem, there has been a considerable amount of research dedicated to the development of systems that are able to infer a segmentation mask from simple, quickly providable user guidance. In the most cases, this guidance amounts to clicks or scribbles on the foreground or background to indicate the position of the object.

Such interactive segmentation systems should be applicable to a wide range of data. The authors of [7] present the *Segment Anything Model (SAM)* as a foundation model that aims at generalizing to unseen domains or types of objects. The authors of [8] present HQ-SAM, which is a slight modification of SAM. It is geared towards producing segmentation masks with a higher quality and endowing the model with the capacity to segment finer structures. For this purpose, HQ-SAM was additionally fine-tuned with the high-quality human annotated masks from HQSeg-44k. In contrast, SAM has only been trained on automatically generated masks. Despite the extensive training, foundation models such as SAM and HQ-SAM often face issues when exposed to domains that do not resemble their training data [9]. In order to rectify these limitations, we will explore the applicability of test-time adaptation (TTA) methods when adapting SAM and HQ-SAM to this winter sports dataset. Our experiments will show that there is a considerable room for improvement when applying

¹Available at <https://github.com/Schorob/wseseg>.

TTA. Our contributions can be summarized as follows:

- We provide a dataset which contains high-quality instance masks for 10 types of winter sports equipment. We call this dataset WSESeg.²
- We explore the usability of SAM and HQ-SAM for the usage as interactive segmentation systems on our winter sports dataset. We measure their performance in terms of the Failure Rate and Number of Clicks metrics.
- We compare various schemes for the test-time adaptation of interactive segmentation models to boost performance compared to the standard versions of SAM and HQ-SAM.

II. THE WINTER SPORTS EQUIPMENT SEGMENTATION (WSESEG) DATASET

There has already been a considerable amount of literature and datasets discussing the importance of being able to segment the body parts, clothing and even certain types of equipment a person is wearing. The authors of [2], [3] have published datasets for the task of segmenting all parts of a human body, including clothing items. In addition to that, the usage of segmentation masks of sports equipment is discussed in [4], [5], [10]. The authors of [11] demonstrate a viable use case for segmentation masks of sports equipment, by segmenting swords used in fencing.

With the aim of contributing to this line of research, we publish a novel dataset containing instance segmentation masks of winter sports equipment. The published dataset is called WSESeg (Winter Sports Equipment Segmentation). The datasets contains ten classes. This corresponds to nine different types of object. Skis occur in the form of two classes: One class specifically contains skis in the context of ski jumping, while the other class contains skis used in other contexts.

An overview over the classes as well as the amount of images and masks that can be found in each class are given in Table I. Most of the images have been collected from Flickr, by automatically downloading all search results corresponding to a certain query (one query for each class). Afterwards the resulting set of images has been manually filtered to only contain images that actually display the desired type of object. The only exception is provided by the class for skis in the context of ski jumping, where the images originate from the first video of the YouTube Skijump Dataset [5]. The chosen frames are exactly the ones indicated by the authors. In order to provide high quality annotations in the form of fine-grained masks, we used an existing interactive segmentation system [12].

In addition to this, any form of segmentation masks for rarely occurring objects constitutes a viable way of evaluating interactive segmentation systems, as the authors of [7] demonstrate. SAM [7] and HQ-SAM [8] have been trained for the usage of general consumer images. More specific types of items, such as winter sports equipment in a scenery containing a considerable amount of snow, are rare in such image sets. We

²Dataset, benchmark and interactive segmentation baseline will be released on Github right before publication.

TABLE I
THE NUMBER OF MASKS AND IMAGES IN EACH CLASS OF THE NOVEL WSESEG DATASET.

| Class Name | # Masks | # Images |
|-------------------|---------|----------|
| Ski (Jump) | 498 | 249 |
| Ski (Misc) | 601 | 245 |
| Bobsleighs | 620 | 572 |
| Curling Brooms | 656 | 284 |
| Curling Stones | 983 | 285 |
| Ski Goggles | 599 | 501 |
| Ski Helmets | 684 | 555 |
| Slalom Gate Poles | 1034 | 507 |
| Snowboards | 650 | 491 |
| Snow Kites | 1127 | 532 |
| Total | 7452 | 4221 |

are going to evaluate the viability of these systems for the task of interactive segmentation on the rare domains provided by our WSESeg dataset. Furthermore, we are going to test various tactics geared towards improving the systems performance during usage time. These methods are described in Section IV. Another reason why our data may be challenging for a model lies in the great variety of object sizes. Bobsleighs constitute the class with the largest objects covering 13.05 % of the area of the respective images, while the ski of skijumpers cover only 0.78 % of the images. Although the COCO dataset (see [6]) also offers masks for skis, they only provide the mask annotations in the form of rather coarse polygons.

A selection of sample images can be seen in Fig.1.

III. RELATED WORK

A. Segmentation of Instances in Sports

The authors of [13] propose a system for the tasks of player instance segmentation and ball localization. Their paper points out the importance of segmenting sports equipment. In [14] we find a player instance segmentation system for the DeepSportsRadar dataset. The authors of [15] develop a real time system for semantic segmentation. The players are viewed as a single mask. [11] provides us with another perspective of why the tracking of sports gear is important. Therein the authors present a system for tracking the sword during fencing. In addition to performing instance segmentation on the sword itself, the tip of the sword is tracked in the form of a keypoint.

B. Interactive Segmentation

Interactive segmentation deals with methods to segment objects in images with the help of repeated user interaction (see [16]–[18]). In many cases this interaction takes the form of clicks provided by the user (see [19], [20] and [21]). The authors of [7] present the segment anything model (SAM), which has been trained on SA-1B, a dataset containing 1.1B masks for 11M images. The authors publish the trained model weights with the aim of providing a foundation model for a task they call promptable segmentation. This task constitutes a generalization of interactive segmentation. In order to prove the generality of their method, they perform evaluations on datasets containing rare objects. [8] discusses an extension of SAM, called HQ-SAM, that has been fine tuned with the

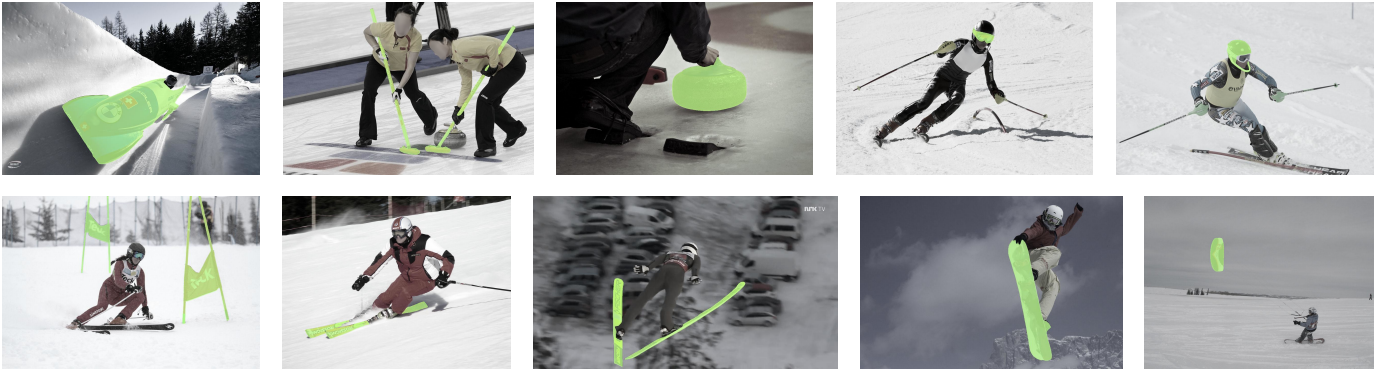


Fig. 1. Sample images from the WSESeg dataset. The object types being segmented are the following. Upper row: Bobsleigh, Curling Broom, Curling Stone, Ski Goggles, Ski Helmet. Lower Row: Slalom Gate Poles, Skis (Misc), Skis (Skijump), Snowboard, Snowkite. The saturation of the images has been decreased for better visibility.

HQSeg-44k dataset, a dataset containing high quality human annotated segmentation masks. Interactive segmentation provides the benefit of generating ground truth annotations during test-time. In [22], [23], the clicks annotated by the authors are employed as labels for single pixels to further optimize the model whilst being in use. The authors of [24] perform online optimization based on past annotations. [25] briefly mentions the usage of intermediate masks without going into further detail.

IV. ONLINE ADAPTATION METHODS

Whenever we use an interactive segmentation model, we can assume the model to be applied to more than one image in the usage domain. In addition to this, interactive segmentation has the property of generating high quality masks for objects without any previously existing ground truth, while the employed system is currently in use. These masks are created to be used as ground truth in the future, and can thus be exploited to progressively adapt the system to the domain to which it is being applied. In addition to this, we obtain the ground truth for a single pixel after each click, allowing us to directly adapt to the current image. In order to ameliorate the user experience when using interactive segmentation systems, we are going to explore various existing techniques for the purpose of adapting pretrained models in an online fashion. By this, we mean that the model does not require any form of fine-tuning before being used. We are first going to provide a quick summary of the problem of interactive segmentation. Afterwards, we are going to present the various possible adaptation methods that can be applied to the foundation models.

A. Interactive Segmentation

Interactive Segmentation methods are usually conceptualized for the purpose of segmenting the surface of an object in an image by giving the system some form of user guidance [19]. We are only looking at the case where the guidance is provided in the form of clicks/coordinates on the object’s surface or the background, respectively. This constitutes a form of ground truth for single pixels which will help the system with the segmentation of the desired object. Interactive

segmentation starts out with an image $\mathbf{x}_{\text{img}} \in \mathbb{R}^{H \times W \times 3}$. Our goal is to create a segmentation mask $\mathbf{m} \in \{0, 1\}^{H \times W}$ for a particular object in the image.

For this, we train a network Φ_{IntSeg} to predict a segmentation mask from the clicked coordinates and the images. In our case, we are going to look at a scenario in which Φ_{IntSeg} is also given a preexisting, potentially faulty segmentation mask we want to improve. Whenever we have no such mask, we give the network a mask purely consisting of zeros. This is the case whenever the user has just made the first click. The network Φ_{IntSeg} is going to be applied in an iterative fashion in order to progressively improve the mask with each interaction (i.e. with each click). Let τ be the index of the current interactive step. The user inspects the currently estimated mask $\mathbf{m}_{\tau-1}$ and places a click on a region that isn’t yet correctly labeled. These clicks take the form (i_τ, j_τ, l_τ) with (i_τ, j_τ) being the coordinate on the image and l_τ being the label. This label either indicates the pixel as background ($l_\tau = -$) or foreground ($l_\tau = +$). In application scenarios, the label is usually indicated by using the right or left mouse button, respectively. Together with this new click, the notation $\mathbf{p}_{1:\tau} = \{\mathbf{p}_1, \dots, \mathbf{p}_\tau\}$ represents all so far accumulated clicks. The interactive segmentation network then predicts a corrected mask $\mathbf{m}_\tau = \Phi_{\text{IntSeg}}(\mathbf{x}_{\text{img}}, \mathbf{p}_{1:\tau}, \mathbf{m}_{\tau-1})$. These steps are repeated until the user judges the mask quality to be sufficient. Since this judgement is subjective, the mask may still contain incorrectly annotated areas. We will call this resulting mask $\mathbf{m}^{\text{Result}}$.

We are going to look at two interactive segmentation models: SAM [7] and HQ-SAM [8]. We deem these models to be of elevated interest, because they are considered as *foundation models* by the authors. This characterization specifically implies the ambition of creating a model that has been trained with such an enormous amount of training data, that it can be successfully applied to any arbitrary domain without having to be fine-tuned. The architectures of these two models can be divided into three parts:

- The **image encoder** is a vision transformer (see [26]) that receives the image and outputs feature maps describing

the input image.

- The **prompt encoder** encodes the clicks and the previous masks into an internal feature representation.
- The **mask decoder** predicts the mask from the encoded prompts and the extracted image features.

This type of task division allows for an increased processing speed during interaction. The image only has to be processed once in order to extract the features. During the interaction, the generation of a predicted mask only requires the repeated execution of the prompt encoder and the mask decoder. As long as these two networks are sufficiently lightweight, the user can adjust the mask in real time. It should be noted that SAM has been designed for the more general task of promptable segmentation, which also involves bounding boxes or text as user guidance. In our experiments, however, we only look at click-based interactive segmentation. SAM has been trained on the dataset SA-1B, which contains 1.1B segmentation masks in 11M different images. This dataset has been published together with SAM and only contains automatically generated masks. The authors of HQ-SAM argue that purely training a model on automatically generated labels might lower the performance due to low-quality masks as the ground truth for the dataset. They propose a slight modification of the SAM architecture which they call HQ-SAM. In addition to the large pretraining of SAM, HQ-SAM is fine-tuned on a dataset called HQSeg-44k, which contains high quality human-annotated masks.

B. Click Adaptation (CA) and Click Mask (CM)

Right after the placement of a click in the image, we obtain the ground truth information for a single pixel. Over the course of τ interactions on said image, we accumulate progressively more annotated pixels $\mathbf{p}_{1:\tau}$. We can use these pixel as pseudo labels to adapt our model directly to the image (see [22], [23]). This is done by creating a sparse mask which indicates that only some pixels are annotated by labeling the pixels without any label with the class -1 :

$$\mathbf{m}_{\tau,i,j}^{\text{Sparse}} = \begin{cases} 1, & \text{if } (i, j, +) \in \mathbf{p}_{1:\tau} \\ 0, & \text{if } (i, j, -) \in \mathbf{p}_{1:\tau} \\ -1, & \text{otherwise} \end{cases} \quad (1)$$

We can then employ this mask to directly compute and minimize the sparse binary cross-entropy

$$\mathcal{L}_{\text{Sparse}}(\mathbf{m}_{\tau}^{\text{Sparse}}, \mathbf{m}_{\tau}) = \frac{\sum_{i,j} \mathbf{1}_{\mathbf{m}_{\tau,i,j}^{\text{Sparse}}=1} \mathcal{L}_{\text{BCE}}(\mathbf{m}_{\tau,i,j}^{\text{Sparse}}, \mathbf{m}_{\tau,i,j})}{\sum_{i,j} \mathbf{1}_{\mathbf{m}_{\tau,i,j}^{\text{Sparse}}=1}} + \frac{\sum_{i,j} \mathbf{1}_{\mathbf{m}_{\tau,i,j}^{\text{Sparse}}=0} \mathcal{L}_{\text{BCE}}(\mathbf{m}_{\tau,i,j}^{\text{Sparse}}, \mathbf{m}_{\tau,i,j})}{\sum_{x,y} \mathbf{1}_{\mathbf{m}_{\tau,x,y}=0}} \quad (2)$$

with \mathbf{m}_{τ} being the currently estimated mask. We use this loss to partially optimize the network parameters. This optimization poses a form of intentional overfitting to the current image. We will call this **click adaptation (CA)**. In our experiments (see Tables II and III) we look at the options to either reverse

this overfitting (marked with an **R**) after each image, or simply carry on without reversing (marked with a **C**).

We also have the possibility of using the sparse mask to carry out an optimization after each image, in order to adapt the model to the images domain as a whole. We refer to this optimization as **click mask (CM)**. In Table II and Table III we use a **checkmark** (\checkmark) whenever this technique is employed.

C. Using the Resulting Mask (RM) as Ground Truth

The purpose of interactive segmentation is the creation of high quality object masks by a user. Thus we obtain a mask for a single object, once the user is done annotating. We can use the resulting mask as a pseudo label to adapt our model to the usage domain. Since the quality of the mask depends on what the user regards as sufficient, there still may be erroneous areas on the mask. In order to allow the loss function to ignore these areas we tried two different methods of discarding potentially faulty mask pixels. The first approach is based on the assumption, that the erroneous areas of the resulting mask are most likely to be found close to the borders between foreground and background. We will therefore erode the background and foreground areas and ignore the eroded parts of the mask. We carry out an iterative form of erosion. Let \mathbf{m} be a mask. Then we define k -fold iterative erosion to be

$$\gamma^0(\mathbf{m}) = \mathbf{m}, \quad (3)$$

$$\gamma^{k+1}(\mathbf{m}) = \gamma^k(\mathbf{m}) \ominus \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}. \quad (4)$$

This erosion operation is applied once to the background mask and once to the foreground mask. The results are then merged again into a new mask, where the eroded areas are set to the ignore label (-1).

The second approach is to use confidence based pseudo-label filtering. The mask generated by the interactive segmentation model after the last user interaction can be seen as a probability map $\mathbf{m}^{\text{Result}, \text{p}}$. For each pixel, probabilities closer to 0 indicate a higher likelihood of the pixel belonging to the background and probabilities closer to 1 a higher likelihood of belonging to the foreground. The value 0.5 then corresponds exactly to the decision border between foreground and background. We define a confidence threshold $\delta \in (0, 0.5)$ which indicates how far the probability should be away from the decision border 0.5 in order to be used as a label during optimization. Only pixels whose probability values have a distance of at least δ from 0.5 are considered during loss computation (i.e. $|\mathbf{m}_{i,j}^{\text{Result}, \text{p}} - 0.5| \geq \delta$). In Tables II and III the column **RM** indicates in what way we use the result mask during optimization. **E** stand for label filtering by erosion, **CT** stands for the application of a confidence threshold and **U** for using an unfiltered result mask.

TABLE II

THE RESULTS ON THE WSESEG DATASET FOR SAM. NOC MEANS THE NOC₂₀@85 METRIC AND FR IS THE FR₂₀@85, DESCRIBING THE NUMBER OF OBJECTS THAT COULD NOT BE SEGMENTED AFTER 20 CLICKS. FOR BOTH METRICS, A SMALLER VALUE INDICATES A BETTER PERFORMANCE. AN EXPLANATION OF THE CONFIGURATIONS CAN BE FOUND IN SECTION IV: CLICK ADAPTATION (CA) CAN EITHER BE CARRIED OUT WITH (R)ESETS OR (C)ONTINUOUSLY. THE RESULTING MASK (RM) CAN EITHER BE USED (U)NTREATED, PRUNED WITH (E)ROSION OR CONFIDENCE THRESHOLDING (CT). WE MAY ALSO USE ALL ACCUMULATED CLICKS AS A CLICK MASK (CM).

| Configuration | | | Bobsleigh | | Curl. Stone | | Ski Helmet | | Snow Kite | | Ski (Jump) | |
|---------------|--------------|----|---------------|--------------|---------------|--------------|-------------------|--------------|--------------|--------------|---------------|--------------|
| CA | RM | CM | NoC | FR | NoC | FR | NoC | FR | NoC | FR | NoC | FR |
| | | | 3.379 | 3.39 | 3.586 | 9.56 | 9.050 | 33.04 | 5.972 | 23.96 | 16.050 | 74.50 |
| R | E | ✓ | 3.502 | 3.06 | 3.617 | 9.36 | 8.173 | 26.17 | 5.996 | 23.96 | 15.323 | 68.67 |
| R | CT | ✓ | 3.511 | 3.06 | 3.641 | 9.66 | 8.151 | 25.88 | 5.983 | 23.60 | 15.229 | 68.88 |
| R | | | 3.415 | 3.06 | 3.481 | 8.75 | 7.972 | 25.29 | 5.966 | 23.78 | 15.329 | 68.67 |
| C | | | 3.450 | 3.06 | 3.452 | 8.65 | 8.044 | 25.00 | 5.926 | 23.51 | 15.299 | 68.88 |
| | E | | 3.440 | 3.06 | 3.579 | 9.26 | 8.792 | 31.14 | 5.987 | 24.05 | 15.763 | 72.89 |
| | CT | | 3.445 | 2.90 | 3.574 | 9.16 | 8.542 | 29.24 | 5.981 | 23.87 | 15.707 | 72.29 |
| | | ✓ | 3.384 | 2.74 | 3.585 | 9.36 | 8.692 | 30.12 | 6.004 | 24.13 | 15.677 | 71.69 |
| R | | ✓ | 3.461 | 3.06 | 3.452 | 8.55 | 7.985 | 25.15 | 5.955 | 23.69 | 15.337 | 68.67 |
| R | U | ✓ | 3.495 | 3.23 | 3.617 | 9.26 | 8.218 | 27.34 | 6.008 | 23.69 | 15.251 | 68.67 |
| Curl. Broom | | | Ski Goggles | | Ski (Misc) | | Slalom Gate Poles | | Snowboards | | Average | |
| NoC | FR | | NoC | FR | NoC | FR | NoC | FR | NoC | FR | NoC | FR |
| 13.855 | 62.20 | | 10.942 | 44.07 | 12.153 | 52.91 | 6.649 | 22.92 | 6.678 | 23.85 | 8.831 | 35.03 |
| 13.213 | 57.16 | | 10.441 | 39.90 | 11.589 | 49.08 | 6.349 | 20.12 | 6.522 | 22.62 | 8.472 | 32.01 |
| 13.133 | 57.01 | | 10.501 | 40.90 | 11.554 | 48.59 | 6.402 | 20.21 | 6.648 | 24.00 | 8.475 | 32.18 |
| 13.076 | 55.49 | | 10.508 | 40.07 | 11.434 | 47.42 | 6.119 | 18.67 | 6.686 | 23.85 | 8.399 | 31.50 |
| 13.108 | 56.25 | | 10.558 | 40.40 | 11.720 | 49.75 | 6.346 | 20.50 | 6.605 | 23.23 | 8.451 | 31.92 |
| 13.640 | 60.37 | | 10.741 | 42.57 | 11.854 | 50.75 | 6.565 | 22.15 | 6.734 | 24.15 | 8.710 | 34.04 |
| 13.555 | 59.60 | | 10.618 | 41.74 | 11.927 | 51.75 | 6.640 | 22.15 | 6.694 | 23.85 | 8.668 | 33.65 |
| 13.773 | 60.82 | | 10.920 | 43.91 | 11.880 | 51.08 | 6.742 | 22.82 | 6.669 | 23.85 | 8.732 | 34.05 |
| 13.189 | 56.25 | | 10.372 | 39.57 | 11.399 | 47.75 | 6.359 | 20.02 | 6.592 | 23.69 | 8.410 | 31.64 |
| 13.392 | 58.23 | | 10.466 | 40.23 | 11.654 | 48.75 | 6.465 | 20.89 | 6.672 | 24.00 | 8.524 | 32.43 |

V. EXPERIMENTS

A. Evaluation Details

In order to carry out an automatic evaluation, we follow common practice (see [19]) to simulate a plausible approximation to the behaviour of a human user. Note that the act of evaluating our system requires the availability of the ground truth for the images we evaluate on. Each click is generated in the following way:

- 1) Compare the predicted mask with the ground truth to obtain segmentation masks covering the false positive and false negative areas.
- 2) Compute \mathcal{D}_{FP} and \mathcal{D}_{FN} , which are the distance transforms (see [27]) of the false positive and false negative area respectively.
- 3) Find the coordinates $\mathbf{p}_{\max, FP}$ and $\mathbf{p}_{\max, FN}$ at which you can find the maximum values of the respective distance transforms. The new click will be placed at one of these two coordinates.
- 4) If $\mathcal{D}_{FP}[\mathbf{p}_{\max, FP}] > \mathcal{D}_{FN}[\mathbf{p}_{\max, FN}]$, place the new click at $\mathbf{p}_{\max, FP}$ and give it a background (−) label. Otherwise, place the new click at $\mathbf{p}_{\max, FN}$ and give it a foreground (+) label.

In order to ensure efficient processing by the model, we only optimize the mask decoder. In this way, the large encoder only has to be executed once. We utilize the SAM and HQ-SAM architectures with a ViT-b backbone. We discovered appropriate values for hyperparameters in preliminary experiments: For adaptation, we use the Adam optimizer [28] with a learning rate of $5 \cdot 10^{-8}$ for SAM and 10^{-6} for HQ-SAM. We found

$k = 5$ to be the best performing number of iterations for iterative erosion, and $\delta = 0.45$ to be the best performing confidence threshold.

Since our experiments are going to measure the usability of SAM and HQ-SAM as interactive segmentation systems, we are going to measure the NoC₂₀@85 and FR₂₀@85 metrics, which are computed as follows: Let \mathbf{m}_{GT} and \mathbf{m}_τ be the ground truth mask and the currently estimated mask, respectively. We assess the mask quality by $\text{IoU}(\mathbf{m}_{GT}, \mathbf{m}_\tau) = \frac{\mathbf{m}_{GT} \cap \mathbf{m}_\tau}{\mathbf{m}_{GT} \cup \mathbf{m}_\tau}$. For both metrics, we have a certain IoU threshold (in our case 85). In order for a mask to be regarded as sufficiently well annotated, the overlap with a preexisting ground truth has to exceed this IoU threshold. In the domain of interactive segmentation, NoC₂₀@85 measures the number of click interactions necessary to reach this threshold. This number of interactions is capped to 20 foreground / background clicks. This is also the same concept used for the failure rate FR₂₀@85: If the predicted mask does not reach an IoU of at least 85 with the ground truth after 20 clicks, we consider the system to have failed annotating this object. The FR₂₀@85 measures this failure rate. Out of the two metrics we regard the failure rate as slightly more important, since it captures a form of inapplicability of the model.

B. Results

We are first going to look at the performance of our method when being applied to the SAM model. For this, we use the NoC₂₀@85 and FR₂₀@85 metric as a measure of performance. The results for this setting can be seen in Table II. On both metrics the best average result is achieved by click adaptation

TABLE III

THE RESULTS ON THE WSESEG DATASET FOR HQ-SAM. NOC MEANS THE NOC₂₀@85 METRIC AND FR IS THE FR₂₀@85, DESCRIBING THE NUMBER OF OBJECTS THAT COULD NOT BE SEGMENTED AFTER 20 CLICKS. FOR BOTH METRICS, A SMALLER VALUE INDICATES A BETTER PERFORMANCE. AN EXPLANATION OF THE CONFIGURATIONS CAN BE FOUND IN SECTION IV: CLICK ADAPTATION (CA) CAN EITHER BE CARRIED OUT WITH (R)ESETS OR (C)ONTINUOUSLY. THE RESULTING MASK (RM) CAN EITHER BE USED (U)NTREATED, PRUNED WITH (E)ROSION OR CONFIDENCE THRESHOLDING (CT). WE MAY ALSO USE ALL ACCUMULATED CLICKS AS A CLICK MASK (CM).

| Configuration | | | Bobsleigh | | Curl. Stone | | Ski Helmet | | Snow Kite | | Ski (Jump) | |
|---------------|--------------|----|---------------|--------------|---------------|--------------|-------------------|--------------|--------------|--------------|---------------|--------------|
| CA | RM | CM | NoC | FR | NoC | FR | NoC | FR | NoC | FR | NoC | FR |
| | | | 8.961 | 26.61 | 10.897 | 43.34 | 18.785 | 91.23 | 8.335 | 34.69 | 19.189 | 94.78 |
| R | E | ✓ | 5.219 | 8.55 | 8.531 | 27.67 | 14.208 | 53.51 | 8.363 | 34.34 | 18.926 | 91.16 |
| R | CT | ✓ | 8.040 | 21.61 | 11.171 | 44.86 | 18.443 | 88.30 | 8.618 | 35.94 | 19.155 | 94.38 |
| R | | | 6.961 | 11.29 | 4.802 | 12.61 | 9.088 | 27.49 | 7.738 | 31.32 | 17.394 | 79.72 |
| C | | | 7.276 | 14.03 | 4.583 | 12.31 | 8.873 | 26.46 | 8.118 | 33.27 | 16.578 | 74.70 |
| | E | | 6.284 | 12.42 | 9.627 | 36.22 | 18.864 | 91.81 | 8.192 | 34.07 | 19.442 | 96.18 |
| | CT | | 8.387 | 22.58 | 11.151 | 44.66 | 18.794 | 91.23 | 8.512 | 35.58 | 19.195 | 94.78 |
| | | ✓ | 17.248 | 80.81 | 15.260 | 68.67 | 19.067 | 92.98 | 9.594 | 40.20 | 19.038 | 93.78 |
| R | | ✓ | 13.229 | 44.19 | 7.411 | 20.55 | 9.418 | 28.65 | 9.447 | 39.22 | 16.247 | 73.09 |
| R | U | ✓ | 5.918 | 10.81 | 9.072 | 28.38 | 14.361 | 54.09 | 8.038 | 33.81 | 19.062 | 92.97 |
| Curl. Broom | | | Ski Goggles | | Ski (Misc) | | Slalom Gate Poles | | Snowboards | | Average | |
| NoC | FR | | NoC | FR | NoC | FR | NoC | FR | NoC | FR | NoC | FR |
| 18.605 | 90.70 | | 17.424 | 80.80 | 17.504 | 84.36 | 13.480 | 60.54 | 11.215 | 49.38 | 14.440 | 65.64 |
| 18.460 | 87.96 | | 14.154 | 59.10 | 14.423 | 60.90 | 14.339 | 57.93 | 10.743 | 45.23 | 12.737 | 52.63 |
| 18.957 | 92.99 | | 17.691 | 82.30 | 17.571 | 85.02 | 14.276 | 64.80 | 11.843 | 53.08 | 14.576 | 66.33 |
| 14.329 | 60.37 | | 11.456 | 39.40 | 12.659 | 50.75 | 9.511 | 32.11 | 7.791 | 28.31 | 10.173 | 37.34 |
| 14.220 | 59.91 | | 11.456 | 39.73 | 12.664 | 50.58 | 9.952 | 34.43 | 7.637 | 27.38 | 10.136 | 37.28 |
| 19.041 | 93.45 | | 17.678 | 82.64 | 17.539 | 85.02 | 15.103 | 67.89 | 11.251 | 49.38 | 14.302 | 64.91 |
| 18.587 | 90.55 | | 17.372 | 80.63 | 17.567 | 84.86 | 13.446 | 60.06 | 11.257 | 49.85 | 14.427 | 65.48 |
| 18.991 | 93.60 | | 18.526 | 87.15 | 18.268 | 89.35 | 15.484 | 71.76 | 12.720 | 58.15 | 16.420 | 77.64 |
| 15.970 | 71.49 | | 12.137 | 43.24 | 15.068 | 65.89 | 8.809 | 30.27 | 8.914 | 33.69 | 11.665 | 45.03 |
| 18.672 | 89.79 | | 16.634 | 72.12 | 16.128 | 74.88 | 16.479 | 75.73 | 10.865 | 46.31 | 13.523 | 57.89 |

with resets, without using any form of dense mask. Here the FR is reduced from 35.02 to 31.50 while the NoC is reduced from 8.831 to 8.399. This is however not the case for all classes: The lowest FR and NoC on the snowboards class is achieved by a mixture of click adaptation with resets and using the eroded result masks, which incurs a reduction of the FR to 22.62 percentage points and the NoC to 6.522 clicks. On the bobsleigh class we reach the lowest FR only using the click mask. Pruning the result mask incurs an improvement with both strategies. Using erosion reduces the failure rate by 0.99 percentage points while using confidence thresholding reduces the FR by 1.38 percentage points. If we combine these methods with the click mask and click adaptation with resets, the FR is reduced even further by 3.02 percentage points for erosion and 2.85 percentage points for confidence thresholding.

We also test our method on a second type of model: HQ-SAM, which is a slightly altered extension of SAM that has been fine-tuned on the HQSeg-44K dataset. The authors of [8], which proposed HQ-SAM, criticized the lack of fine grained quality in the masks predicted by SAM. The results obtained by HQ-SAM can be seen in Table III. The HQSeg-44k dataset provides high-quality human-made annotations, in contrast to SA-1B, in which the annotations have been automatically generated. While this fine-tuning results in a general improvement of mask quality, it also specializes the architecture to a vastly smaller variety of data. This turns out to be detrimental when applying the resulting model to new domains, such as winter sports equipment. This can especially be seen by the drastically increased average failure rate of

65.64 for HQ-SAM vs. 35.03 for SAM. Here, click adaptation without resets incurs the biggest improvement, reducing the failure rate from 65.64 to 37.28 on and the NoC from 14.440 to 10.136. For click adaptation with resets, the failure rate is reduced to 37.34. On the class of bobsleighs we can see a deviation from this trend, where the combination of click adaptation with resets, using the click mask and the eroded result mask reduces the failure rate from 26.61 to 8.55 and the NoC from 8.961 to 5.219. We can observe a similar behavior on the slalom gate pole class, where the best results are achieved by a combination of the click mask with click adaptation with resets. While there is no single best method for all types of objects, we can reach notable improvements employing adaptation methods on average.

VI. CONCLUSION

In this paper we presented a new dataset containing instance segmentation masks on winter sports equipment of ten different classes. With the publication of this dataset we aim at supporting research in the direction of applying computer vision for sports analysis and human body part segmentation. Since our dataset contains rare classes, it provides a viable way of testing foundation models for interactive segmentation. We test the performance of two foundation models, SAM and HQ-SAM. In order to look at ways of improving the user experience regarding such models, we also test various methods to adapt said foundation models to the winter sports domain during usage. Despite being conceptualized as foundation models, their performance can be improved using test-time adaptation methods.

REFERENCES

- [1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [2] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, Terence Sim, Shuicheng Yan, and Jiashi Feng. Multiple-human parsing in the wild. *arXiv: Computer Vision and Pattern Recognition*, 2017.
- [3] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network.
- [4] Katja Ludwig, Julian Lorenz, Robin Schön, and Rainer Lienhart. All keypoints you need: Detecting arbitrary keypoints on the body of triple, high, and long jump athletes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5178–5186, 2023.
- [5] Katja Ludwig, Daniel Kienzle, Julian Lorenz, and Rainer Lienhart. Detecting arbitrary keypoints on limbs and skis with sparse partly correct segmentation masks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 461–470, 2023.
- [6] 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.
- [7] 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.
- [8] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023.
- [9] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Shangzhan Zhang, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything?—sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. *arXiv preprint arXiv:2304.09148*, 2023.
- [10] Maxime Istasse, Vladimir Somers, Pratheeban Elancheliyan, Jaydeep De, and Davide Zambrano. Deepsporadar-v2: A multi-sport computer vision dataset for sport understandings. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports, MMSports '23*, page 23–29, New York, NY, USA, 2023. Association for Computing Machinery.
- [11] Takehiro Sawahata, Alessandro Moro, Sarthak Pathak, and Kazunori Umeda. Instance segmentation-based markerless tracking of fencing sword tips. In *2024 IEEE/SICE International Symposium on System Integration (SII)*, pages 472–477. IEEE, 2024.
- [12] 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.
- [13] Seyed Abolfazl Ghasemzadeh, Gabriel Van Zandycke, Maxime Istasse, Niels Sayez, Amirafshar Moshtaghpour, and Christophe De Vleeschouwer. Deepsporlab: a unified framework for ball detection, player instance segmentation and pose estimation in team sports scenes. *arXiv preprint arXiv:2112.00627*, 2021.
- [14] Fang Gao, Wenjie Wu, Yan Jin, Lei Shi, and Shengheng Ma. A sparse attention pipeline for deepsporadar basketball player instance segmentation challenge. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, pages 131–135, 2023.
- [15] Anthony Cioppa, Adrien Deliege, Maxime Istasse, Christophe De Vleeschouwer, and Marc Van Droogenbroeck. Arthus: Adaptive real-time human segmentation in sports through online distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
- [16] Zhuwen Li, Qifeng Chen, and Vladlen Koltun. Interactive image segmentation with latent diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 577–585, 2018.
- [17] Xi Chen, Zhiyan Zhao, Feiwu Yu, Yilei Zhang, and Manni Duan. Conditional diffusion for interactive segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7345–7354, 2021.
- [18] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *CVPR*, 2022.
- [19] 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.
- [20] Konstantin Sofiiuk, Ilya Petrov, Olga Barinova, and Anton Konushin. f-brs: Rethinking backpropagating refinement for interactive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8623–8632, 2020.
- [21] Won Dong Jang and Chang Su Kim. Interactive image segmentation via backpropagating refinement scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5297–5306, 2019.
- [22] Theodora Kontogianni, Michael Gygli, Jasper Uijlings, and Vittorio Ferrari. Continuous adaptation for interactive object segmentation by learning from corrections. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 579–596. Springer, 2020.
- [23] Qingxuan Shi, Yihang Li, Huijun Di, and Enyi Wu. Self-supervised interactive image segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [24] Zheng Lin, Zhao Zhang, Zi-Yue Zhu, Deng-Ping Fan, and Xia-Lei Liu. Sequential interactive image segmentation. *Computational Visual Media*, 9(4):753–765, 2023.
- [25] Yuying Hao, Yi Liu, Juncai Peng, Haoyi Xiong, Guowei Chen, Shiyu Tang, Zeyu Chen, and Baohua Lai. Rais: Robust and accurate interactive segmentation via continual learning. *arXiv preprint arXiv:2210.10984*, 2022.
- [26] 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.
- [27] Pedro F Felzenszwalb and Daniel P Huttenlocher. Distance transforms of sampled functions. *Theory of computing*, 8(1):415–428, 2012.
- [28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.