TimberVision: A Multi-Task Dataset and Framework for Log-Component Segmentation and Tracking in Autonomous Forestry Operations (Supplementary Material)

Daniel Steininger Julia Simon Andreas Trondl Markus Murschitz
AIT Austrian Institute of Technology (Center for Vision, Automation & Control)
{daniel.steininger,julia.simon,andreas.trondl.fl,markus.murschitz}@ait.ac.at

This supplementary document complements the main paper with additional dataset statistics and presents more detailed results of detection, segmentation and tracking evaluations. Moreover, we illustrate the generalization capacity of our approach to various application domains and conclude with selected corner cases to facilitate a more comprehensive understanding.

A. Extended dataset description and statistics

The following sections include detailed statistics and illustrations regarding the process of dataset creation and its final composition.

A.1. Data acquisition

Fig. 1 shows the ratios of images captured during specific times of day and months. In 141 recording sessions, we captured a wide variety of seasonal aspects as well as lighting and weather conditions across eight different months. Most data was recorded in winter and early spring, as this is a popular time for harvesting timber. As a result, about 9% of total images contain snow. Autumn is currently underrepresented and will be the focus of future data campaigns, although winter conditions without snow show partly similar characteristics. In addition to seasonal changes, TimberVision covers a range of daytime variations. Frequent recording times range from morning to late afternoon, while a smaller percentage was captured in the evening. In total, 16 images show dusk or night scenarios.

We captured images for TimberVision using a total of 10 different sensors, which are listed in Tab. 1 along with their corresponding resolutions and the number of images included in the respective subsets. The additional *OpenSource* subset contains 42 images with resolutions ranging from 672x504 to 3176x2039 pixels. We furthermore recorded image sequences with high scene entropy and log quantities using a DJI Mavic 2 Pro UAV, which were used exclusively for qualitative analysis.

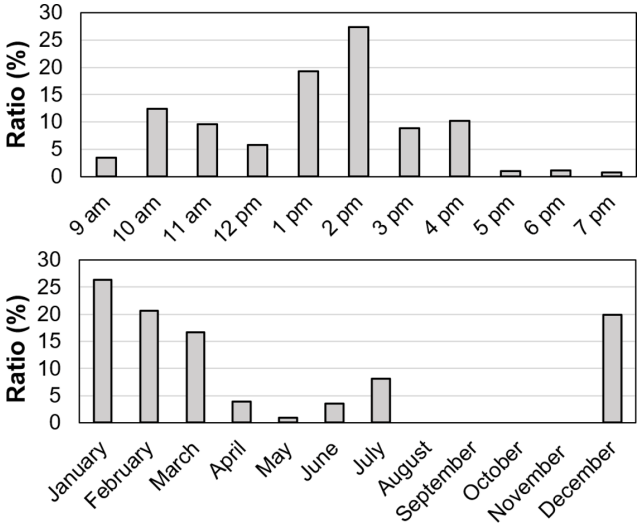

Figure 1. Hourly and monthly recording-time distributions of annotated TimberVision images.

A.2. Annotation concept

To complement the analysis of other state-of-the-art datasets presented in Tab. 2 of the main paper, Fig. 2 compares our annotation concept to the approaches of the two most closely related works. Both provide instance-segmentation masks on the level of trunks, while our dataset includes multiple classes for their individual components. TimberSeg focuses on the detection of cut logs and CanaTree100 on live trees, while we include annotations for both types of trunks, separable by their id range. Furthermore, since the target scenario of TimberSeg is mainly log manipulation, only the top layer of log piles is annotated, while we include all visible trunks in a pile, as visible in the top row. Regarding live trees, not all instances in the far background can reasonably be annotated in dense forest scenarios as visible in the second row of images. Our dataset and CanaTree100 set different thresholds for this purpose,

Sensor	Width	Height	Images	Subset
ZED 2	1280	720	304	Loading, Harvesting, Tracking
Sony Xperia PRO-I	1280	720	217	Tracking
Sony Xperia XZ2 Compact	1500	844	8	Core
Sony Xperia PRO-I	1920	1080	30	Core
Sony Xperia PRO-I	2016	1134	29	Core
Sony Alpha 7S	2120	1192	192	Core
Huawei P20 Lite	2304	1296	70	Core
iPhone 12 Mini	2016	1512	279	Core
Samsung Galaxy S10+	2016	1512	4	Core
Huawei P20 Lite	2048	1536	23	Core
Blackfly BFS-PGE-31S4C-C	2048	1536	10	Loading
Samsung Galaxy S5 Neo	2304	1728	201	Core
Sony Xperia XZ2 Compact	2666	1500	549	Core, Tracking
Sony Alpha 6000	3000	2000	30	Core

Table 1. List of sensors used for data acquisition along with their resolutions, numbers of annotated images and associated subsets.

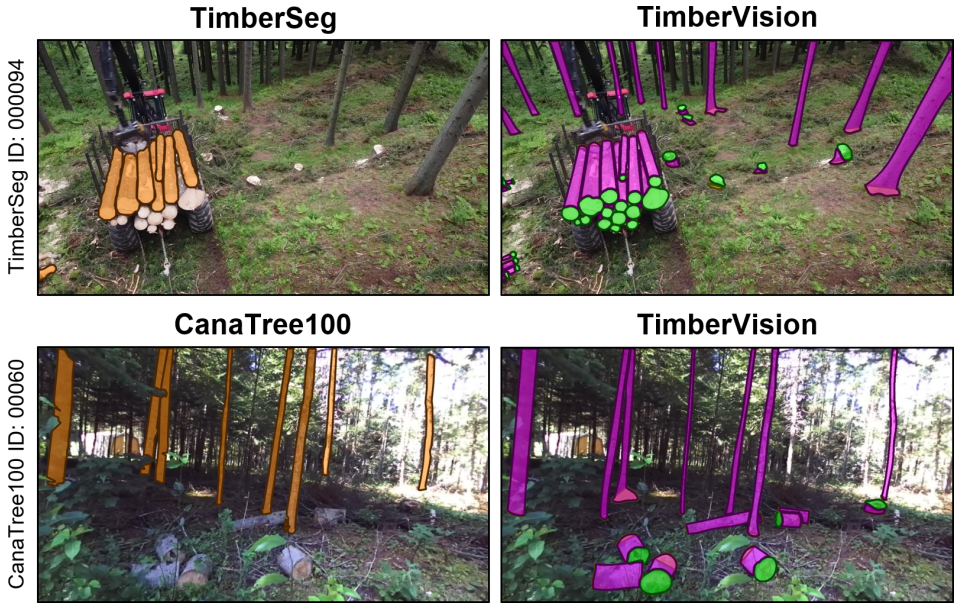

Figure 2. Comparison of annotation schemes in TimberSeg [2], CanaTree100 [3] and TimberVision using original images and corresponding annotations. Orange denotes trunks in single-class annotations, while green and pink identify our *Cut* and *Side* classes, respectively. *Bound* instances overlapping with them are visualized in slightly lighter shades.

which further reduces compatibility, as the same tree may be included in the annotations of one dataset but not the other. Furthermore, the three datasets were recorded at different geographical locations and therefore depict different tree species. Overall, these aspects illustrate the difficulty of comparing the few existing instance-segmentation datasets in the domain of forestry operations or conducting cross-evaluations without substantial limitations or adaptations. On the other hand, the limited compatibility with any existing work shows that our dataset indeed addresses a relevant data gap and represents a valuable complementary addition to the state of the art. To further illustrate our annotation approach, an extended set of representative ground-truth samples is presented in Fig. 3.

A.3. Distribution of scene parameters

For a more detailed analysis of our training setup, the distribution of scene-parameter intensities across our per-session split between training, validation and test data is shown in Tab. 2. It demonstrates that combining the validation and test samples in the evaluation of scene-parameter impact in Fig. 7 of the main paper results in a sufficient

Figure 3. Additional representative examples of semi-automatically generated annotations for instance segmentation of multiple trunk components in the TimberVision dataset.

representation of each intensity.

To provide more insights about our tracking evaluation and the sequences it is based on, we also summarize the distribution of scene parameters across their keyframes in Fig. 4. It shows similar characteristics to the distributions in the overall dataset depicted in Fig. 2 of the main paper and therefore comparable difficulty to the test set used for evaluating detection and fusion performance.

A.4. Instance statistics

As an addition to the dataset analysis, this section provides extended statistics regarding instance characteristics and distributions across the TimberVision dataset. Firstly, Tab. 3 shows the numbers of trunk components included in each subset. Fig. 5 and Fig. 6 show the distributions of their sizes and orientations, respectively. As expected, *Bound* instances are on average the smallest and *Side* instances the largest class, with some of the latter even exceeding image dimensions if the trunk is located diagonally across the

Scene Parameter		Train	Val	Test
Entropy	-	990	202	175
		190	42	64
	+	23	15	21
Quantity	-	634	157	135
		522	91	118
	+	47	11	7
Distance	-	730	164	159
		450	90	92
	+	23	5	9
Irregularity	-	459	107	133
		348	86	63
	+	396	66	64

Table 2. Numbers of images in the training, validation and test splits depicting *Low*, *Mid* and *High* intensities of each annotated scene parameter.

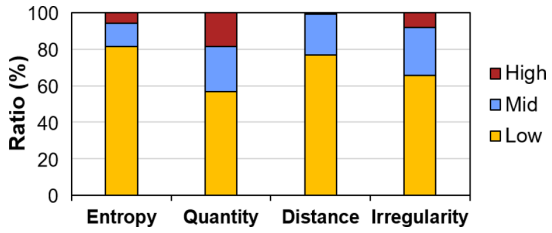

Figure 4. Distribution of scene parameters for annotated keyframes in the *Tracking* subset.

Subset	Cuts	Sides	Bounds
Core	9,535	17,617	6,293
Loading	2,522	3,589	2,671
Harvesting	345	825	550
OpenSource	607	743	284
Tracking	528	1,907	1,273
TimberSeg*	425	922	702
	13,962	**25,603**	**11,773**

Table 3. Detailed statistics of annotated trunk components in all subsets of the TimberVision dataset.

image. Fig. 5 shows the elongated characteristic of *Sides* and *Bounds*, while *Cuts* are closer to square shapes. On the other hand, there is a strong peak regarding orientations of *Cut* instances. Since they are often viewed slightly from the side, they tend to form an upright oval projection in the image plane. *Side* orientations are more uniformly distributed with a slight bias towards completely horizontal or vertical orientations, the latter resulting mainly from upright live trees in contrast to cut trees which are arbitrarily oriented in image space. *Bounds* show a corresponding behaviour, as they form the end points of *Side* instances.

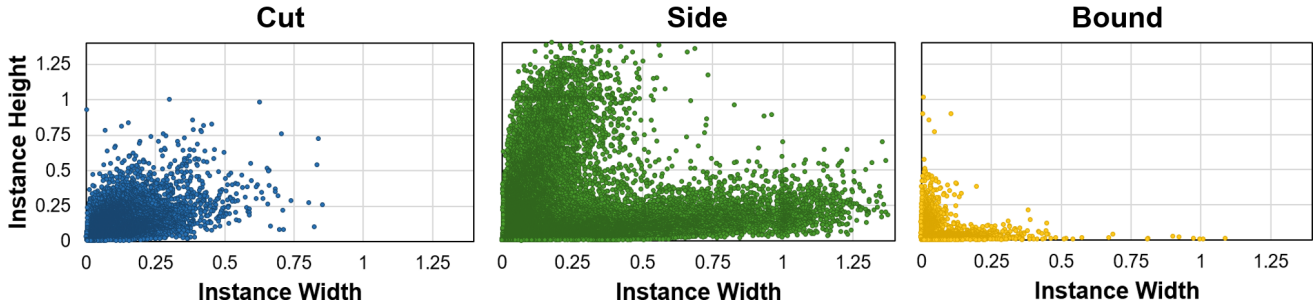

Figure 5. Distribution of instance sizes for each class based on oriented-bounding-box dimensions. Instance width refers to the box side connecting the leftmost corner with the adjacent one in counter-clockwise direction, meaning that a value larger than the corresponding height indicates an orientation below ninety degrees in Fig. 6.

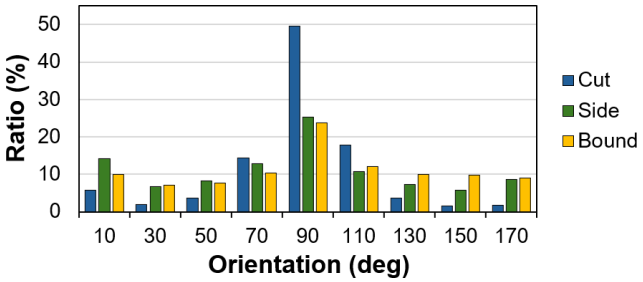

Figure 6. Distribution of instance orientations by class in 20-degree intervals around the given values. The angles are measured between the longest oriented-bounding-box side and the horizontal axis.

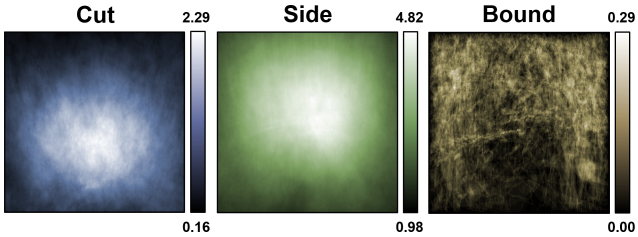

Figure 7. Heat maps illustrating the distribution of instance-segmentation masks within normalized image space for each class. Values are normalized between their respective minima and maxima, which are stated relative to the total numbers of instances for each class (i.e. a white area for the *Side* class indicates that 4.82% of *Side* instances include this position).

Furthermore, the heat maps in Fig. 7 illustrate instance-mask distributions across normalized image space. *Cut* and *Side* instances appear in all positions across the area. However, since many images are captured with hand-held devices from an eye-level perspective, the former tend to be mainly in the lower central region, while the latter are often found in slightly higher positions. *Bound* instances, on the other hand, are clustered along the image edges, as they are usually part of entire visible trunks. The distributions of all classes are largely symmetrical and closely related to realistic application scenarios.

Overall, the statistics show that our data covers a wide range of relevant scene configurations. The characteristics of all classes closely match those to be expected in most target applications.

B. Extended quantitative evaluation

The following tables provide further details regarding the quantitative evaluation of ablation and detection experiments. Tab. 4 shows individual detection results of the subsets comprising *Base* in Tab. 3 of the paper to differentiate model performance for specific types of input data and scenarios. As opposed to the overall evaluation, the validation images are included in addition to the test set to achieve

		Core	Load	Harvest	Open
OOD	**Cut**	83.0	79.5	80.4	80.8
	Side	59.8	66.3	47.1	45.0
ISEG	**Cut**	77.4	67.0	67.7	72.1
	Side	61.2	69.7	45.3	45.4

Table 4. Model performance as $mAP^{50\text{-}95}$ for the classes *Cut* and *Side* for test and validation images of the **Core**, **Load**ing, **Harvest**ing and **Open**Source subsets. Results are reported for *Large* models trained and evaluated on the same image resolution of 1024 pixels.

a representative number of samples for each subset. *Core* constitutes the largest part of the test set and is therefore most closely related to the overall results. The *Cut* class is consistently well detected across all subsets, while *Side* achieves the best results for *Loading* scenarios, but does not perform as well for *Harvesting* and *OpenSource* data. This is consistent with its performance on the *TimberSeg** subset presented in the paper, which commonly features these kinds of scenes. A possible reason for this behaviour, apart from the lower number of samples for harvesting scenarios, might be that the corresponding scenes often contain a high

	FP_{ID} ↓	FN_{ID} ↓	TP_{ID} ↑	MT ↑	PT ↑	ML ↓	Misses ↓	Switches ↓	Frag. ↓
ByteTrack	202	718	1,124	87	42	42	559	68	**51**
Bot-SORT	**181**	696	1,146	87	42	42	560	66	**51**
Optimized	194	**681**	**1,161**	**89**	43	**39**	**534**	**63**	**51**
Opt | 10fps	245	806	1,036	82	**48**	41	606	129	65

Table 5. Additional Clear-Mot [1] and ID [4] metrics for all *Tracking* sequences including 266 keyframes with 1,842 ground-truth annotations. $\mathbf{FP_{ID}}$, $\mathbf{FN_{ID}}$ and $\mathbf{TP_{ID}}$ denote false positives, false negatives and true positives according to the ID metric. **MT**, **PT** and **ML** denote the numbers of mostly tracked, partly tracked and mostly lost objects, respectively.

	Precision	Recall	$mAP^{50\text{-}95}$
Base	84.3	72.9	57.5
TimberSeg	-	50.8	-

Table 6. OBB accuracy for fused trunks on our test set (*Base*) and the original TimberSeg dataset. Since the latter does not include annotations for all trunk instances, only recall is applicable.

number of live trees in the far background, which are inconsistently detected and especially challenging for detection approaches in general.

Furthermore, detailed experimental results of our class ablations for oriented object detection and instance segmentation, which serve as the basis for Fig. 6 in the main paper, are listed in Tab. 7. Additionally, the ablation of both learning tasks is shown in Tab. 8 and forms the basis for Fig. 5 in the main paper.

C. Extended fusion and tracking results

In addition to the fusion results presented in the main paper, Tab. 6 gives an idea of the models' generalization capability using the original TimberSeg dataset. Since fusion results are entire *Trunk* instances, we can use the provided annotations for this class, but still only compare recall, as they do not cover all visible instances (see Fig. 2). The performance drop is consistent with the one for our selected and newly annotated subset presented in Tab. 3 of the main paper. However, according to [2], even models trained and tested only on splits of the 220 TimberSeg images do not yield recalls beyond 65.2% or *mAP* scores beyond 57.5 for the same class, proving the challenging nature of the data.

Fusion results serve as an input for multi-object tracking, for which we list additional MOT metrics in Tab. 5. Fine-tuning experiments on TimberSeg and CanaTree100 with models pre-trained on our TimberVision dataset and MS COCO are illustrated by Fig. 8, which shows *mAP* scores on the validation set for each training epoch of the experiments described in Tab. 5 of the main paper. As discussed, training times can be significantly reduced when using our models as basis for fine-tuning datasets of similar domains.

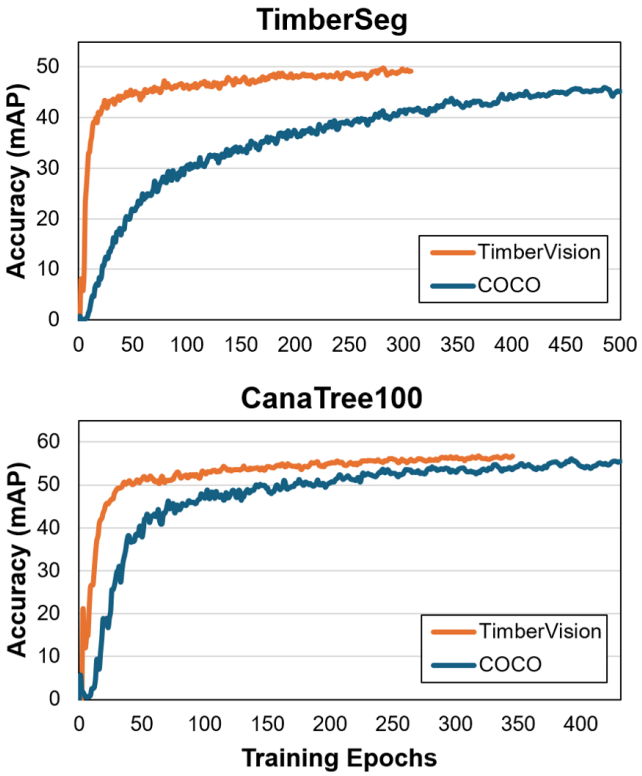

Figure 8. Validation accuracy after each training epoch when fine-tuning on TimberSeg and CanaTree100 with models pre-trained on MS COCO and our TimberVision dataset. Experiments are based on *Large* model architectures with an input size of 1024. *mAP* scores are derived from five-fold validation experiments.

D. Extended qualitative results

To further illustrate our discussion of results and potential application scenarios, we show extended qualitative results, clustered by their subsets of the TimberVision dataset. None of the listed images were included during training or validation. In addition to representative results on the *Core* subset (Fig. 9) and in typical *Loading* and *Harvesting* scenarios (Fig. 10), we demonstrate the generalization potential of our approach on samples of the *TimberSeg* dataset [2] (Fig. 11) and *OpenSource* images (Fig. 12). This is comple-

		Oriented Object Detection						Instance Segmentation					
	Size	**C**	**S**	**B**	**C**	**S**	**T**	$\mathbf{C_{Box}}$	$\mathbf{S_{Box}}$	$\mathbf{C_{Mask}}$	$\mathbf{S_{Mask}}$	$\mathbf{T_{Box}}$	$\mathbf{T_{Mask}}$
n	**768**	76.5	22.6	49.7	76.7	49.1	55.9	75.5	55.5	69.8	49.0	64.6	59.1
	1024	77.8	22.3	49.3	77.7	50.1	56.8	77.5	56.7	72.8	50.3	64.7	59.6
m	**768**	79.8	25.7	54.7	79.8	54.7	61.2	78.7	62.0	72.9	56.5	69.2	65.4
	1024	80.8	27.0	55.9	81.1	56.0	62.7	80.0	62.0	75.5	56.6	70.5	66.5
x	**768**	80.8	28.7	56.6	81.0	56.6	63.8	79.5	63.1	73.9	58.7	70.6	67.0
	1024	**81.9**	**30.5**	**58.4**	**82.1**	**58.4**	**65.3**	**80.7**	**64.1**	**75.8**	**59.3**	**71.5**	**68.3**

Table 7. Complete results of class ablation experiments for the model capacities *Nano* (**n**), *Medium* (**m**) and *X-Large* (**x**) and different image sizes for the classes ***C**ut*, ***S**ide*, ***B**ound* and ***T**runk*. All scores are given as $mAP^{50\text{-}95}$ on the test set. In the case of instance segmentation, scores are reported separately for the *Box* and *Mask* stages.

		Oriented Object Detection				Instance Segmentation						
	Size	**C**	**S**	**⌀**	**t**	$\mathbf{C_{Box}}$	$\mathbf{S_{Box}}$	$\mathbf{⌀_{Box}}$	$\mathbf{C_{Mask}}$	$\mathbf{S_{Mask}}$	$\mathbf{⌀_{Mask}}$	**t**
n	**640**	74.7	47.5	61.1	1.8	73.7	53.8	63.8	67.2	47.2	57.2	1.9
	768	76.7	49.1	62.9	2.0	75.5	55.5	65.5	69.8	49.0	59.4	2.3
	896	77.5	49.6	63.6	2.2	76.3	56.1	66.2	71.2	50.1	60.7	2.6
	1024	77.7	50.1	63.9	2.6	77.5	56.7	67.1	72.8	50.3	61.6	3.2
	1152	78.5	51.0	64.8	3.1	78.4	56.5	67.5	73.6	50.6	62.1	3.9
s	**640**	77.2	50.6	63.9	2.5	76.3	57.1	66.7	69.6	50.7	60.2	3.0
	768	77.9	51.9	64.9	3.4	77.5	58.2	67.9	71.4	52.3	61.9	4.2
	896	79.1	52.7	65.9	4.3	78.5	58.6	68.6	73.1	52.8	63.0	5.3
	1024	79.5	53.0	66.3	5.3	79.1	58.7	68.9	74.1	53.0	63.6	6.7
	1152	79.9	53.4	66.7	6.5	79.3	59.2	69.3	75.1	53.6	64.4	8.2
m	**640**	78.4	53.8	66.1	4.8	77.7	60.1	68.9	71.3	54.6	63.0	5.7
	768	79.8	54.7	67.3	6.9	78.7	62.0	70.4	72.9	56.5	64.7	8.3
	896	80.2	55.7	68.0	9.2	79.7	61.8	70.8	74.4	56.9	65.7	11.1
	1024	81.1	56.0	68.6	11.8	80.0	62.0	71.0	75.5	56.6	66.1	14.3
	1152	81.4	55.9	68.7	14.4	80.0	62.4	71.2	75.8	56.9	66.4	17.4
l	**640**	79.7	55.4	67.6	7.9	78.4	62.0	70.2	71.6	56.8	64.2	9.4
	768	80.5	56.1	68.3	11.4	79.4	63.6	71.5	73.6	58.3	66.0	13.6
	896	81.3	57.4	69.4	15.4	80.2	63.4	71.8	75.1	58.6	66.9	18.3
	1024	81.8	58.0	69.9	19.7	80.4	**64.2**	72.3	76.0	58.9	67.5	23.3
	1152	82.2	57.9	70.1	24.2	80.2	64.1	72.2	75.6	58.5	67.1	28.8
x	**640**	80.0	55.5	67.8	12.2	78.4	62.3	70.4	71.6	56.8	64.2	14.5
	768	81.0	56.6	68.8	17.8	79.5	63.1	71.3	73.9	58.7	66.3	21.2
	896	82.0	58.2	70.1	24.1	80.1	63.9	72.0	74.7	**59.4**	67.1	28.2
	1024	82.1	58.4	70.3	30.6	80.7	64.1	72.4	75.8	59.3	67.6	36.1
	1152	**82.3**	**59.2**	**70.8**	37.6	**80.9**	64.1	**72.5**	**76.4**	59.2	**67.8**	44.1

Table 8. Complete results of oriented-object-detection and instance-segmentation experiments for the model capacities *Nano* (**n**), *Small* (**s**), *Medium* (**m**), *Large* (**l**) and *X-Large* (**x**) and different input sizes. The $mAP^{50\text{-}95}$ scores on the test set are listed for the classes ***C**ut* and ***S**ide* as well as their average (**⌀**) along with mean inference time (**t**) in milliseconds. In the case of instance segmentation, scores are listed for the *Box* and *Mask* stages separately.

mented by images from all subsets and TimberSeg showing selected corner cases and limitations (Fig. 13) to identify challenging scenarios and potentials for improvement. As discussed in the main paper, especially images in low-light conditions and trunks triggering multiple detections due to large occlusions need further investigation during future iterations of the dataset.

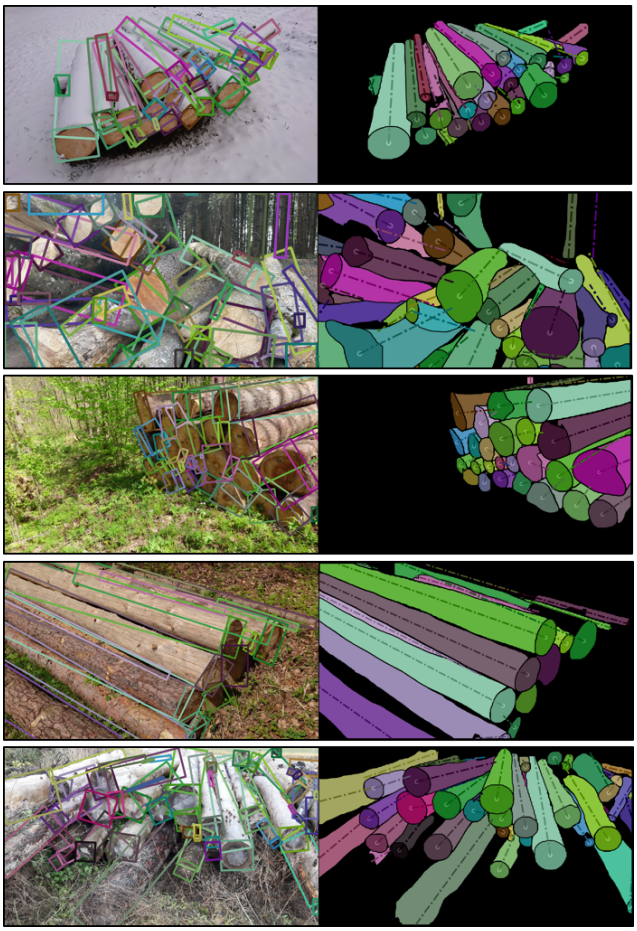

Figure 9. Additional qualitative results on the test split of the *Core* subset recorded in forests and other outdoor locations.

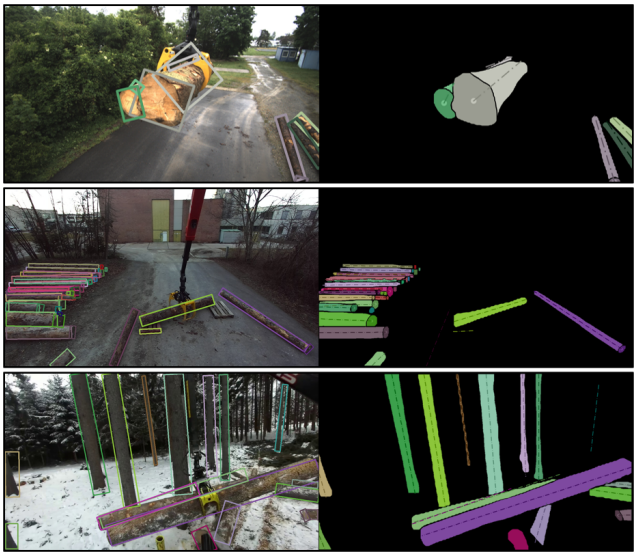

Figure 10. Additional qualitative results on the test splits of the *Loading* and *Harvesting* subsets depicting realistic application scenarios.

Figure 11. Additional qualitative results on the TimberSeg dataset [2] demonstrating the generalization capability of our approach.

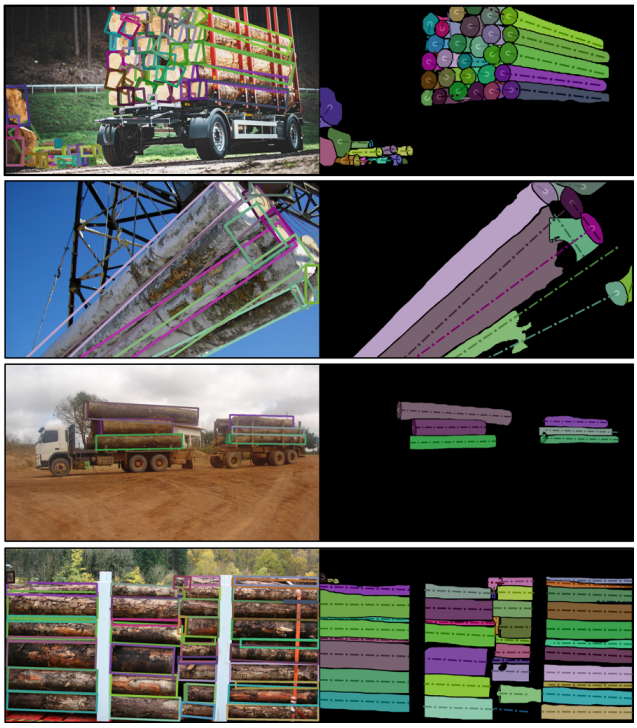

Figure 12. Additional qualitative results on the test split of the *OpenSource* subset with complementary scenarios to the main data from public sources.

Figure 13. Additional qualitative results showing limitations on our test set and the TimberSeg dataset [2].

References

[1] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 5

[2] Jean-Michel Fortin, Olivier Gamache, Vincent Grondin, François Pomerleau, and Philippe Giguère. Instance segmentation for autonomous log grasping in forestry operations. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 6064–6071. IEEE, 2022. 2, 5, 7, 8

[3] Vincent Grondin, Jean-Michel Fortin, François Pomerleau, and Philippe Giguère. Tree detection and diameter estimation based on deep learning. *Forestry*, 96(2):264–276, 2023. 2

[4] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 17–35. Springer, 2016. 5