AUTOMATIC OBJECT SEGMENTATION TO SUPPORT CRISIS MANAGEMENT OF LARGE-SCALE EVENTS

S. M. Azimi[1,*], R. Kiefl[2], V. Gstaiger[1], R. Bahmanyar[1], N. Merkle[1], C. Henry[1], D. Rosenbaum[1], F. Kurz[1]

[1] Remote Sensing Technology Institute, German Aerospace Center (DLR), Oberpfaffenhofen, Germany
{seyedmajid.azimi; veronika.gstaiger; reza.bahmanya; nina.merkle; corentin.henry; dominik.rosenbaum; franz.kurz}@dlr.de
[2] German Remote Sensing Data Center, German Aerospace Center (DLR), Oberpfaffenhofen, Germany
ralph.kiefl@dlr.de

Commission II,WG 6

KEY WORDS: Crisis Management, Segmentation, Aerial Imagery, Large-scale Events, Machine Learning

ABSTRACT:

The management of large-scale events with a widely distributed camping area is a special challenge for organisers and security forces and requires both comprehensive preparation and attentive monitoring to ensure the safety of the participants. Crucial to this is the availability of up-to-date situational information, e.g. from remote sensing data. In particular, information on the number and distribution of people is important in the event of a crisis in order to be able to react quickly and effectively manage the corresponding rescue and supply logistics. One way to estimate the number of persons especially at night is to classify the type and size of objects such as tents and vehicles on site and to distinguish between objects with and without a sleeping function. In order to make this information available in a timely manner, an automated situation assessment is required. In this work, we have prepared the first high-quality dataset in order to address the aforementioned challenge which contains aerial images over a large-scale festival of different dates. We investigate the feasibility of this task using Convolutional Neural Networks for instance-wise semantic segmentation and carry out several experiments using the Mask-RCNN algorithm and evaluate the results. Results are promising and indicate the possibility of function-based tent classification as a proof-of-concept. The results and thereof discussions can pave the way for future developments and investigations.

1. INTRODUCTION

Large-scale events with widely distributed parking and camping areas represent a particular challenge for event and crisis management and require extensive preparation and constant monitoring to guarantee the safety of participants. Injuries and deaths occur repeatedly at large gatherings of people and for years research has been conducted into the causes of accidents and ways of avoiding them in order to make large events safer (Fruin, 1993, Helbing et al., 2000). In order to prevent situations of danger or damage at large events and to be able to act quickly and effectively in an emergency, decision-makers need information with spatial reference for a situation picture that is as close to reality as possible during the event. Due to the increasing availability of high resolution remote sensing data and the growing awareness of the possibility of deriving area-wide information from it, this is more and more being integrated into disaster management procedures (Aina, Bello, 2014, Römer et al., 2016). In the event of an emergency, it must be ensured that rescue routes are wide enough and, above all, free of any objects that would obstruct the passability of the emergency services and that participants can leave the site at any time. An important aspect is therefore the information on both the number of event participants and their distribution on the event site. In general, this information is also essential for the installation of infrastructures such as waste disposal or the supply of food and drinking water.

There are existing works (Meynberg et al., 2016, Bahmanyar et al., 2019) on crowd analysis and the measurement of their den-

Figure 1. Illustration of a sample result from DLR-AerialTent dataset with pixel- and instance-wise segmentation using aerial imagery of a music festival in 2013 in Germany with 9cm/px GSD. The red outlined area represents a sample from the test set.

*Corresponding author

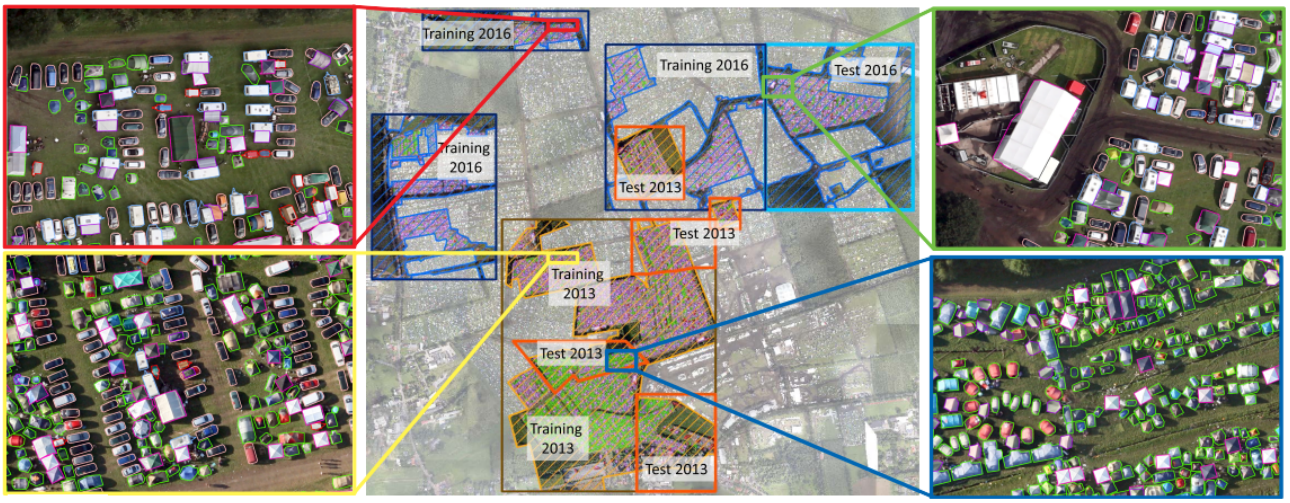

Figure 2. Illustration of the DLR-AerialTent dataset with overlaid annotations for 10 semantic classes including different types of tents, vehicles and infrastructure elements. The images were acquired during a music festival in Germany. The Ground Sampling Distance (GSD) of the images is 9cm/px and 10cm/px respectively. The image in the center shows the festival area and the selected training and test areas of the 2013 and 2016 datasets. Color codes: tent: sleep function, small vehicle/transporter, trailer, truck/bus, camper/caravan, pavilion/large tent: assembly and supply function, awning, tarpaulin, inflatable pool, infrastructure, other objects ("clutter").

sity on festival sites; however, the situation at night has not yet been investigated. In order to estimate the distribution and number of people on the festival site as accurately as possible, we propose the approach of analyzing the sleeping facilities on the different sites. In particular, the type and size of tents, vehicles and similar objects has to be determined and a distinction according to their function has to be made, such as those with and without a sleeping function. In order to make this information available in a timely manner, an automated situation assessment is required as a manual evaluation of larger areas would be too time-consuming.

In recent years, an end-to-end monitoring system has been developed, improved and tested under real world conditions and was successfully demonstrated at several large scale events (Römer et al., 2016). It aims to support the management of events and authorities in charge of security and rescuing effort by recording and providing optical aerial imagery and relevant derived information. Examples include overviews of the current traffic situation and the occupancy of parking and camping areas. This system consists of a chain of loosely coupled components. It includes an optical camera system (Kurz et al., 2014), software and hardware for pre-processing and analysing data on board, a down-link for data transmission in near real-time, additional ground-based components for information extraction (Römer et al., 2014, Kersten, 2014) as well as modules for provision and interactive visualisation of situational information based on web services (Römer et al., 2016).
To prepare future advancements of the image analysis components of such a processing chain, this study focuses on the detection and feature-based classification of vehicles, tents, and similar objects.

2. RELATED WORKS

In recent years, deep learning methods have shown promising object detection and instance-wise segmentation results for ground imagery and outperformed the traditional methods. The enhanced performance owe its rapid promotion to a large extent to large-scale datasets such as ImageNet (Deng et al., 2009), Pascal VOC (Everingham et al., 2010) and MS-COCO (Lin et al., 2014). However, as for aerial imagery, similar datasets are scarce, which has slowed down the development of such methods. Furthermore, the existing aerial image datasets for semantic segmentation are either limited to a few individual classes such as roads and building boundaries in the INRIA (Maggiori et al., 2017), Massachusetts (Mnih, 2013), SpaceNet (Van Etten et al., 2018), and DeepGlobe (Demir et al., 2018) datasets, or provide very coarse classes in the ISPRS Vaihingen and Potsdam (Cramer, 2010) datasets. For object detection and instance-wise segmentation on the other hand, multi-class object detection plays a major role in remote sensing applications and several datasets are public available for these tasks. Example aerial image datasets in this area are iSAID (Waqas Zamir et al., 2019), DOTA (Xia et al., 2017), TAS (Heitz, Koller, 2008), VEDAI (Razakarivony, Jurie, 2016), COWO (Mundhenk et al., 2016), DLR-3K-Munich-Vehicle (Liu, Mattyus, 2015), and UCAS-AOD (Zhu et al., 2015). These datasets were generated either for general purposes or particular applications. However, to the best of our knowledge, none of them tackles the tent classification in large events with campsites. To address this limitation, we propose a new aerial image dataset with detailed annotations, the so-called "DLR-AerialTent" (see Figure 2).

To investigate the feasibility of instance-wise segmentation for function-based tent classification, we apply, among others, a well-established variant of the Region-based Convolutional Neural Network (RCNN) algorithm (Girshick et al., 2014), the so-called Mask-RCNN (He et al., 2017), as our baseline. As the other RCNN variants, Fast-RCNN (Girshick, 2015) augments the detection performance of RCNN by the minimization of the region proposal regression and classification losses simultaneously. Faster-RCNN (Ren et al., 2015) improves the localization accuracy of Fast-RCNN by deploying a region proposal network (RPN) for learning the region proposals. Faster-RCNN can be further improved by multi-scale training and testing to learn the feature maps in multiple levels. However, this increases the memory usage and the inference time. Alternatively, image pyramids or Feature Pyramid Networks (FPNs) (Pinheiro et al., 2016, Honari et al., 2016, Ghiasi, Fowlkes, 2016, Newell et al., 2016, Lin et al., 2017) can be utilized to improve the per-

Figure 3. Samples of objects of interests in the DLR-AerialTent dataset which are commonly present in large scale events and camps.

class name	number of instances from 2013		number of instances from 2016		total number per class	
	training	test	training	test	training	test
tent (sleeping function)	7474	3344	1507	822	8981	4166
small vehicle/transporter	2656	1065	1434	743	4090	1808
trailer	173	88	1434	743	324	147
truck/bus	22	17	22	14	44	31
camper/caravan	169	55	519	197	688	252
pavilion/large tent	1344	501	540	281	1884	782
awning, tarpaulin	492	221	358	144	850	365
inflatable pool	31	17	13	5	44	22
infrastructure	155	66	170	64	325	130
other objects (“clutter”)	172	146	303	68	475	214
total number of instances	12688	5520	5017	2397	17705	7917

Table 1. Overview of the ten classes contained in the dataset and their instance numbers.

formance in different scales at a marginal extra cost. Rotated region proposals (Liu et al., 2017) improve the localization of the oriented bounding box (OBB) tasks by predicting object orientations using single shot detector (SSD) (Liu et al., 2016).
For instance-wise segmentation, a new method has been proposed which applies adaptive weighted pooling and discriminative Region of Interest (RoI)-pooling in a two-stage process together with a RPN (Cao et al., 2020). In addition, ISDNet (Garg et al., 2020) has been developed which applies atrous spatial pyramid pooling (ASPP) module from the DeepLabv3+ (Chen et al., 2018) algorithm in the Mask-RCNN and Cascaded-RCNN manner. In this paper, we are providing a new aerial dataset for instance-wise segmentation with highly accurate annotations and fine-grained classes for camp-relevant objects to promote the development of models for previously unsupported tasks, such as accommodation-wise event monitoring. Additionally, we are carrying out first evaluations of one of the well-established instance-wise segmentation algorithms.

3. DATASET

This study is based on true color aerial images taken over a festival in Germany in early August 2013 and 2016. The images were acquired by a camera-array sensor system mounted on a helicopter, which provides high flexibility for airborne monitoring and is usually available to rescue and security related authorities and organizations (Kurz et al., 2014). The images cover an area of 3.44 km^2 and were acquired at a flight height of around 1000 m above ground, which results in a ground sampling distance of 9 cm and 10 cm, respectively. Note that a part of the aerial images acquired in 2013 were already described in (Römer et al., 2016). We prepared a dataset called “DLR-AerialTent” with images from the years 2013 and 2016 and split it into training and test sets as shown in Figure 2. It is composed of the following 10 semantic classes: 1) tents (with sleeping function), 2) small vehicle / transporter, 3) trailer, 4) truck/bus, 5) camper/caravan, 6) pavilion/large tent (assembly and supply function), 7) awning, tarpaulin, 8) inflatable pool, 9) infrastructure and 10) other objects (“clutter”). This classification is based on experiences with large events gained over the past 10 years. It takes into account the most common and, for our research question, most important classes of objects found in parking and camping areas at festivals and similar large scale events in Germany, and should be considered as a first proposal for such a dataset. Figure 3 shows some samples of the different classes and Table 1 provides an overview of the classes and the number of instances contained in each class.

In total, 25622 objects have been manually derived and labeled by experts from which 17705 (69.10%) are in the training set and 7917 (30.90%) are in the test set. From the 2013 dataset, there are 12688 (69.7%) and 5520 (30.3%) objects in the training and test sets, respectively. As for 2016, there are 5017 (67.7%) and 2397 (32.3%) objects divided into training and test sets, respectively. Area-wise coverage speaking, 147430.7 m^2 (68.1%) are in the training set and 69142.8 m^2 (31.9%) are included in the test set from which 97343.2 m^2 (68.1%) and 45595.1 m^2 (31.9%) are from 2013, 50087.4 m^2 (68.0%) and 23547.7 m^2 (32.0%) from the 2016 festival are divided into training and test sets, respectively.

4. METHOD

At the beginning of this research work we would like to find out if it is possible to detect and distinguish tents based on their function. For this reason, we apply a pixel-wise semantic segmentation on a small dataset and focus on identifying and localizing tents with sleeping function, pavilion/large tents and vehicles. First, we annotate a part of the aerial image of 2013 to serve as training set and then, we test our methods on the rest of the image. A two stream pixel-wise semantic segmentation algorithm is used which considers large and small scale

objects to combine shallow features from high spatial resolution inputs and rich features from low spatial resolution inputs as described in (d'Angelo et al., 2019). The first segmentation results are visible in the left red outlined sample in Figure 1. After achieving these promising results, we set our goal to identify further types of tents, vehicles and other artificial structures such as infrastructure elements.

In order to localize objects more accurately and to be be able to count each instance object, each object of interest has to be identified separately regardless of having shared border with another object having the same class. Therefore, we decide to analyse the images using the instance-wise segmentation approach. We choose Mask-RCNN as the base-line which is a well-established deep neural network aiming to resolve instance segmentation problems in computer vision. Specifically speaking, it separates different instances of objects in an image by providing object bounding boxes, classes and masks as three heads. Mask-RCNN is a extension of Faster-RCNN for instance-wise segmentation. Similar to the Faster-RCNN, there are also two stages in Mask-RCNN. First, it generates region proposals for possible existing object regions, and second, it predicts the object class, refines its bounding box and generates a polygon mask in the pixel level. Both stages are added downstream of the backbone network to extract high-level features, which can be either single-scale or multi-scale. In other words, to adapt Faster-RCNN to the instance-wise instance segmentation, Mask-RCNN contains two heads: One head for box object detection and another for instance mask segmentation, which are trained end-to-end.

In contrast to the majority of the recent systems, where classification is dependent on mask predictions, Mask-RCNN outputs single binary mask for each RoI. During training, a multi-task loss is utilized on each selected RoI $L_{as} = L_{cls}+L_{box}+L_{mask}$. The classification L_{cls} and bounding-box L_{box} loss functions are identical to Faster-RCNN. Considering K classes in total, the mask branch yields a Km^2-dimensional output per RoI encoding K binary masks of $m \times m$ resolution. A per-pixel sigmoid is applied to this, and we define L_{mask} as the average binary cross-entropy loss function. Therefore, for a sampled RoI connected with ground-truth class k, L_{mask} is applied only on the $k-th$ mask *i.e.*, other mask outputs do not affect the loss. There are several sub-modules in the algorithm in the case of multi-scale backbone such as FPN, RPN, region of interest network (ROI), non-maximum suppression (NMS) and the mask head. As in the RPN module, we minimize the multi-task loss

$$L\left(\{p_i\},\{t_i\}\right) = \frac{1}{N_{obj}} \sum_i L_{obj}(p_i, p_i^*) \quad (1)$$

$$+ \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}\left(t_i, t_i^*\right), \quad (2)$$

where for an anchor i in a mini-batch, p_i is the predicted probability of an object existence and p_i^* is the ground-truth binary label. For the classification (object/not-object), the log-loss $L_{obj}(p_i, p_i^*) = -p_i^* \log p_i$ is applied, while we employ the smooth l_1 loss function

$$L_{reg}(t_i, t_i^*) = l_1^{\text{smooth}}(t_i - t_i^*) \quad \text{with} \quad (3)$$

$$l_1^{\text{smooth}}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \quad (4)$$

for the bounding box regression. Here,

$$t_{xi} = (x_i - x_{i,a})/w_a, \qquad t_{yi} = (y_i - y_{i,a})/h_a \quad (5)$$

$$t_{xi}^* = (x_i^* - x_{i,a})/w_a, \qquad t_{yi}^* = (y_i^* - y_{i,a})/h_a \quad (6)$$

are the coordinates of the predicted and ground-truth anchors with x_i, $x_{i,a}$, and x_i^* indicates the predicted, anchor, and ground-truth respectively (the same also goes for y); and w_a and h_a are the anchor width and height. N_{obj} and N_{reg} normalize hyper-parameters (the mini-batch size as well as the number of anchor locations); and λ denotes the balancing hyper-parameter between the two loss functions, which is set to 10.

In the module of ROI, each chosen region proposal is regressed and classified simultaneously.

$$L(p, u, t^u, v) = L_{cls}(p, u) + \lambda[u \geq 1] L_{loc-HBB}(t^u, v) \quad (7)$$

where horizontal bounding box (HBB) and $L_{cls}(p, u) = -u \log p$. u is the true class and p is the discrete probability distribution of the predicted classes which is defined over $K+1$ categories as $p = (p_0, ..., p_K)$ where "1" is for the background. In contrast to Faster-RCNN, Mask-RCNN uses ROIAlign instead of ROIPool to improve localization performance of each

$L_{loc-HBB}(t^u, v)$ROI. is defined similar to the L_{reg} in which $\{xmin, ymin, w, h\}$ (the upper-left coordinates, width and height) of t^u and v for the corresponding HBB coordinates are computed.

In the case of classification of an object as background, $[u \geq 1]$ ignores the offset regression. The balancing hyper-parameter λ is also set to 1 in this case. The same region proposal is fed to the mask-head, which outputs the boundary mask for the object inside of the region proposal. It is accepted as final output, if the region proposal is classified with a class except background. To obtain the final detections, as the final post-processing, we deploy NMS in which overlaps among detections is computed to choose the best localized region and to omit redundant regions.

5. EXPERIMENTAL SETUP

We have carried out the experiments using two Titan XP GPUs and the Detectron[1] framework based on Caffe2. We trained algorithms for 5000, 10000, 20000, and 30000 iterations denoted in the result tables as 1x, 2x, 3x, and 4x. For the training, we used the learning rate of 0.02 with a scheduled learning rate procedure of 60% and 80% of the total iteration with the gamma of 0.1. As the backbone networks, we used ResNet-50, ResNet-101 (He et al., 2016), and ResNeXt-101 (Xie et al., 2017). The ResNeXt backbones are trained with the cardinalities of 32 and 64 with the bottleneck widths of 8d and 4d, respectively. In addition, the features of the last convolution layer of the 4-th stage of the backbones (C4) as well as the FPN features are used as inputs for the three heads. Using FPN after the backbone network allows images to be processed at multiple feature scales, which should improve the performance on small objects significantly as they are usually lost in the output of high-level features.

The head resolution of Mask-RCNN is 28 and uses RoIAlign for aligning the region proposals. RoI batch size for each image is 512 and the image-wise batch size is 1. Moreover, we use

[1] `https://github.com/facebookresearch/detectron`

Table 2. Comparison of the baselines for instance segmentation. (+) Trained with augmented training data. (*) augmented testing.

Backbone	Feature Pyramid	mAP^{Box}	mAP^{Box}_{InsW}	Per Category $AP^{mask}_{50,95}$									
				Tent	Vehicle	Trailer	Truck	Caravan	Pavilion	Awning	Pool	Infrastructure	Clutter
ResNet-50	$C4^{1x}$	26.5	44.46	41.2	60.6	15.9	7.4	51.1	61.8	15.9	5.0	1.6	4.1
ResNet-50	$C4^{2x}$	26.2	44.53	41.3	60.5	16.0	3.6	50.6	62.3	15.5	4.6	1.7	5.5
ResNet-50	FPN^{1x}	29.0	45.07	43.1	53.3	14.3	9.3	54.4	67.8	24.6	8.8	5.9	8.7
ResNet-50	FPN^{2x}	28.0	43.97	41.5	53.2	10.6	10.1	52.0	67.8	23.1	8.5	3.9	8.9
ResNet-101	FPN^{1x}	29.1	46.00	43.9	55.6	15.7	8.7	55.7	68.3	21.9	6.9	5.1	9.1
ResNet-101	FPN^{2x}	25.8	41.63	40.9	46.8	8.3	6.0	50.7	64.8	20.1	8.0	4.3	7.9
ResNeXt-101_32x8d	FPN^{1x}	29.5	46.61	44.0	58.1	13.0	8.0	55.2	67.4	23.2	6.7	9.8	9.2
ResNeXt-101_32x8d	FPN^{2x}	28.4	45.39	43.7	53.3	8.0	4.5	55.5	68.5	23.5	9.3	7.5	10.6
ResNeXt-101_32x8d*	FPN^{1x}	31.1	47.41	44.3	58.2	12.2	8.8	57.8	69.7	27.2	7.8	12.3	12.9
ResNeXt-101_32x8d*	FPN^{2x}	32.0	46.98	42.8	59.3	12.2	10.9	61.2	68.8	27.3	9.2	14.3	14.2
ResNeXt-101_32x8d*	FPN^{3x}	31.7	47.70	43.7	60.5	11.4	9.2	59.7	69.9	26.5	9.2	14.0	13.1
ResNeXt-101_32x8d*+	FPN^{3x}	**36.5**	53.93	49.4	68.6	**21.3**	6.1	**65.6**	**74.8**	**34.9**	9.2	**20.3**	**15.1**
ResNeXt-101_32x8d*+	FPN^{4x}	36.2	**54.04**	**49.9**	**69.5**	19.4	8.4	65.3	74.5	29.5	**12.7**	19.5	13.6
ResNeXt-101_64x4d	FPN^{1x}	29.1	45.20	42.3	56.1	12.4	**11.9**	55.3	67.5	22.7	6.6	7.9	8.6
ResNeXt-101_64x4d	FPN^{2x}	29.2	45.12	42.6	55.0	14.7	10.4	53.8	68.4	20.8	8.4	9.3	8.3

Table 3. Comparison of the baselines for instance segmentation. (+) Trained with augmented training data. (*) augmented testing.

Backbone	Feature Pyramid	mAP^{Box}	AP^{Box}_{50}	AP^{Box}_{75}	AP^{Box}_{s}	AP^{Box}_{m}	AP^{Box}_{l}
ResNet-50	$C4^{1x}$	26.47	43.04	31.74	10.38	39.27	43.39
ResNet-50	$C4^{2x}$	26.16	42.27	30.94	10.45	32.05	45.60
ResNet-50	FPN^{1x}	29.03	40.50	34.74	10.03	46.08	52.16
ResNet-50	FPN^{2x}	27.97	39.40	33.21	8.63	45.02	45.86
ResNet-101	FPN^{1x}	29.09	41.16	34.70	10.46	46.16	48.71
ResNet-101	FPN^{2x}	25.79	38.13	30.20	6.95	41.68	48.28
ResNeXt-101_32x8d	FPN^{1x}	29.46	43.00	33.82	10.54	46.05	48.90
ResNeXt-101_32x8d	FPN^{2x}	28.45	41.63	33.13	9.44	44.05	50.20
ResNeXt-101_32x8d*	FPN^{1x}	31.13	45.00	36.07	10.28	48.30	50.67
ResNeXt-101_32x8d*	FPN^{2x}	32.03	46.02	37.45	10.83	48.63	54.13
ResNeXt-101_32x8d*	FPN^{3x}	31.72	44.69	36.60	10.27	49.60	54.77
ResNeXt-101_32x8d*+	FPN^{3x}	**36.53**	**50.57**	**43.10**	**15.58**	**55.22**	**52.70**
ResNeXt-101_32x8d*+	FPN^{4x}	36.25	49.47	42.99	15.28	54.75	51.85
ResNeXt-101_64x4d	FPN^{1x}	29.12	41.98	34.25	8.95	41.64	51.66
ResNeXt-101_64x4d	FPN^{2x}	29.15	41.83	34.46	10.15	45.78	50.81

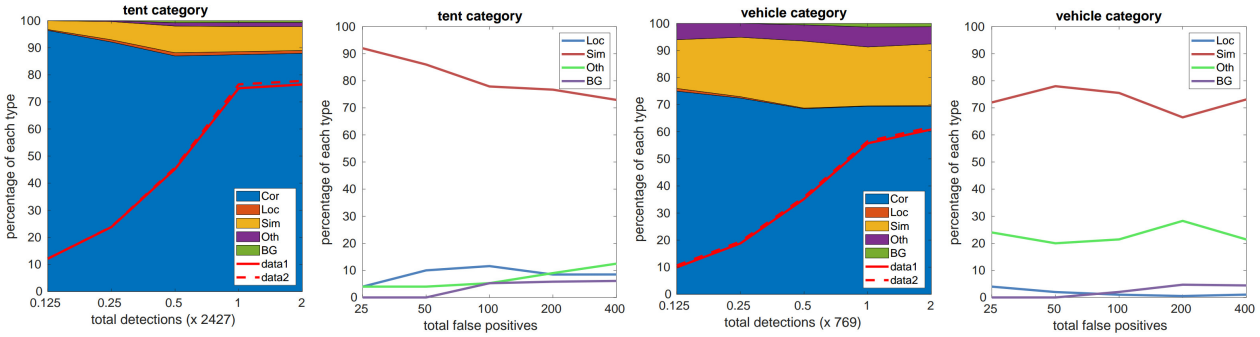

Figure 4. Performance visualization for the best Mask-RCNN setting on tent and vehicle categories from DLR-AerialTent test set.The first and third diagram from the left show the cumulative fraction of detections which were classified correctly (Cor), or represent false positive classifications due to poor localization (Loc), or due to confusion with similar (Sim), or with other (Oth) categories, or with the background (BG). The solid red line indicates the change of recall with the strong criteria of 0.5 (jaccard overlap) by increasing detection numbers. The dashed red line reflects the weak criteria of 0.1 (jaccard overlap). The diagrams on the right side indicate the distribution of top-ranked false positive factors.

union and soft-average for the bounding box and masks heuristically with horizontal flipping at the test time. For data augmentation at the train phase, we crop the images with the size of 1024×1024 pixels and with rotations of 0°, 90°, 180° and 270°. This results in 3280 test and 73 train samples, respectively.

We employ mean Average Precision (mAP) as evaluation metric, similar to the evaluation of the MS-COCO dataset. For bounding box and segmentation mask detections, APs are computed based on Intersection over Union (IoU) with 50%, 75%, and 95% intersection rates. Furthermore, since the dataset is heavily skewed and unbalanced, we calculate the instance-weighted mAP (mAP_{insW}).

6. RESULTS AND DISCUSSION

Table 2 and Table 3 represent a baseline comparison for the instance-wise segmentation task. According to the results, ResNeXt-101 with cardinality = 32 and bottleneck width = 8d after $3\times$ training with an augmentation in the training and test sets outperforms the other configurations. It achieves mAPs of 36.5% on the instance-wise segmentation task. This configuration also achieves the best mAP_{insW} (54.04%) with $4\times$ training. Moreover, according to Table 2, almost all configurations perform poorly for the infrastructure, inflatable pool, and

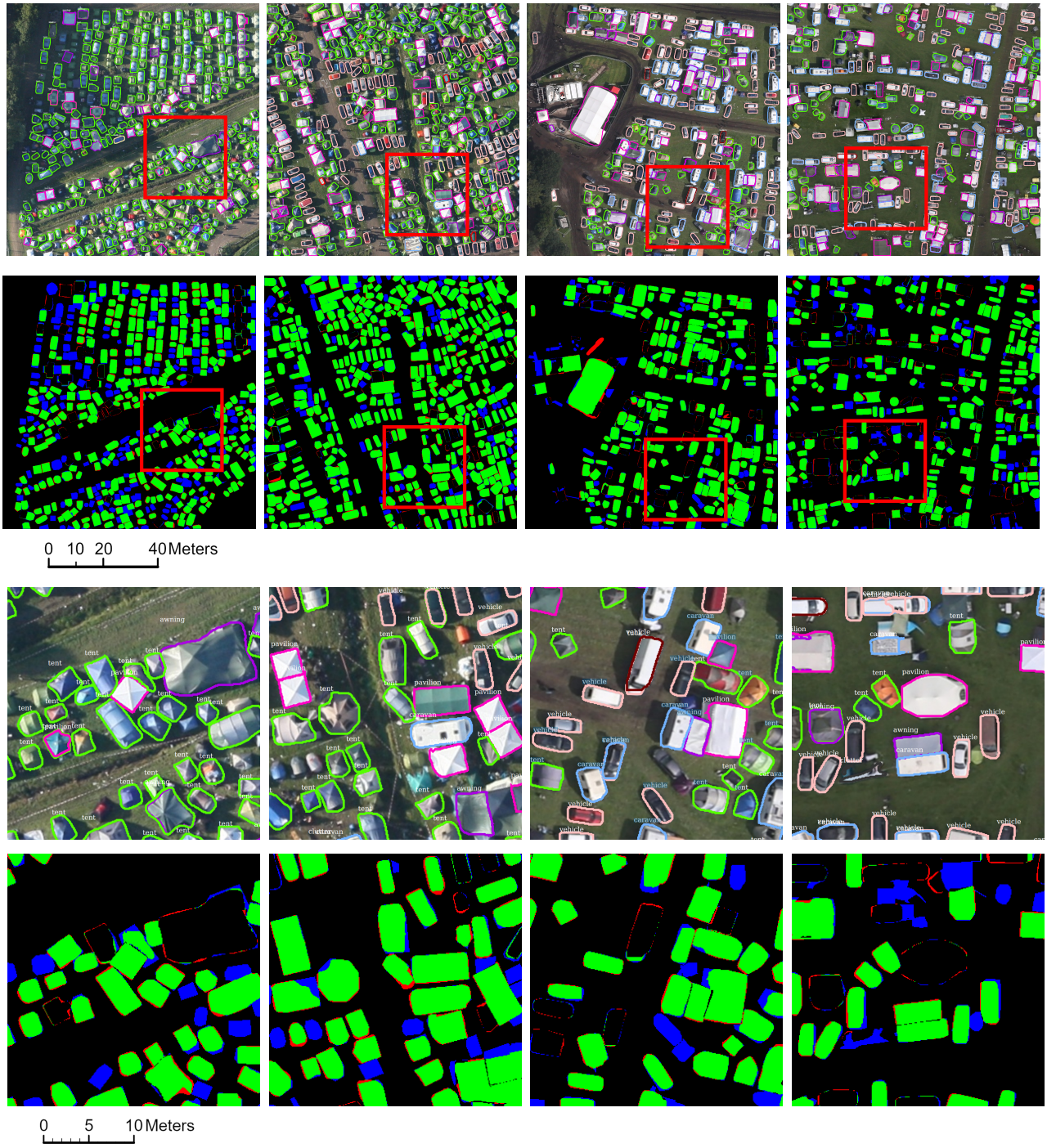

Figure 5. Samples of visual outputs for mask segmentation and confusion in the DLR-AerialTent test set. Color codes for the first and third row: tent: sleep function, small vehicle/transporter, trailer, truck/bus, camper/caravan, pavilion/large tent: assembly and supply function, awning, tarpaulin, inflatable pool, infrastructure, other objects ("clutter"). Confusion color codes for second and fourth: true positive, false positive (wrong class) and false negative (object not detected).

truck/bus classes. This could be expected due to the small number of available samples for these classes (see Table 1). Results show that more training iteration improves the performance for the inflatable pool class. However, it decreases the performance for the other classes due to overfitting. They also show that, despite their large diversities, tents with sleeping functions can be distinguished from similar objects classes such as large tents, pavilions, awnings, tarpaulins, and sun sails with a high accuracy. In addition, it can be seen that camping vehicles with a sleeping function can be distinguished from other vehicle classes with a relatively high level of confidence. In order to better analyse the correlation of the performance with the object sizes, in Table 3, we show the average precision for large (AP_l^{Box}), medium (AP_m^{Box}) and small (AP_s^{Box}) objects. According to the results, small objects are harder to be detected and segmented in comparison to the larger ones. This is due to their smaller number of samples in our dataset as well as their complex features resembling those of the other classes. This can be confirmed by analysing false positives in Figure 4. This figure demonstrates the performance for the best Mask-RCNN

configuration on the tent and vehicle categories of the DLR-AerialTent test set. The diagrams on the left side show the cumulative fraction of detections which were classified correctly (Cor), or represent false positive classifications due to poor localization (Loc), or due to confusion with similar (Sim), or with other (Oth) categories, or with the background (BG). The solid red line indicates the change of recall with the strong criteria of 0.5 (jaccard overlap) by increasing detection numbers. The dashed red line reflects the weak criteria of 0.1 (jaccard overlap). The diagrams on the right side indicate the distribution of top-ranked false positive factors.

We have carried out such analysis for all classes; however, for the sake of space, we merge the tent, pavilion, large-tents and awning classes and the small-vehicle, caravan, camper, trailer and truck/bus classes. In both cases, similarity and confusion with objects from other classes can be considered as the main reason for the false positives. Figure 5 shows also some examples of the visual output for the mask segmentation in the DLR-AerialTent test set.

7. CONCLUSION AND FUTURE WORKS

In this paper, we present a proof-of-concept that it is feasible to distinguish tents based on their functionalities on camp sites. We introduce the first dataset for this application, which we use to train an instance-wise segmentation algorithm of Mask-RCNN with multiple configuration. The results show promising outputs for the most important categories despite low performance for a few classes. From the operational point of view, results of this study can support future developments and improve monitoring systems for area occupancy and passability of rescue routes during large-scale events. With the help of the object classes, the number of people and their distribution can be estimated by assigning specific, empirically determined values to the classes. This step, as well as the evaluation of the results, will follow this study. Additionally, we will investigate more recent network architectures and will work on developing dedicated algorithms for this task to achieve better performance. An expansion to analyse data of temporary refugee camps, as well as the use of satellite data, is being considered.

REFERENCES

Aina, Y., Bello, O., 2014. Satellite Remote Sensing as a Tool in Disaster Management and Sustainable Development: Towards a Synergistic Approach. *Procedia - Social and Behavioral Sciences*, 120, 365-373.

Bahmanyar, R., Vig, E., Reinartz, P., 2019. Mrcnet: Crowd counting and density map estimation in aerial and ground imagery. *BMVC Workshop on Object Detection and Recognition for Security Screening*, 1–12.

Cao, J., Cholakkal, H., Anwer, R. M., Khan, F. S., Pang, Y., Shao, L., 2020. D2det: Towards high quality object detection and instance segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834–848.

Cramer, M., 2010. The DGPF-Test on Digital Airborne Camera Evaluation - Overview and Test Design. *Photogrammetrie Fernerkundung Geoinformation*, 2010, 73-82.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raska, R., 2018. Deepglobe 2018: A challenge to parse the earth through satellite images. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 172–17209.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 248–255.

d'Angelo, P., Cerra, D., Azimi, S. M., Merkle, N., Tian, J., Auer, S., Pato, M., de los Reyes, R., Zhuo, X., Bittner, K. et al., 2019. 3d semantic segmentation from multi-view optical satellite images. *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 5053–5056.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., Zisserman, A., 2010. The Pascal visual object classes (VOC) challenge. *International journal of computer vision*, 88(2), 303–338.

Fruin, J., 1993. The causes and prevention of crowd disasters. *Engineering for Crowd Safety*, 99-108.

Garg, P., Chakravarthy, A. S., Mandal, M., Narang, P., Chamola, V., Guizani, M., 2020. Isdnet: Ai-enabled instance segmentation of aerial scenes for smart cities. *ACM Trans. Internet Technol*, 1(1).

Ghiasi, G., Fowlkes, C. C., 2016. Laplacian Pyramid Reconstruction and Refinement for Semantic Segmentation. *ECCV*.

Girshick, R., 2015. Fast R-CNN. *CVPR*.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR*.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. *ICCV*, 2961-2969.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. *CVPR*, 770-778.

Heitz, G., Koller, D., 2008. Learning spatial context: Using stuff to find things. D. Forsyth, P. Torr, A. Zisserman (eds), *ECCV*, Springer Berlin Heidelberg.

Helbing, D., Farkas, I., Vicsek, T., 2000. Simulating dynamical features of escape panic. *Nature*, 407, 487-490.

Honari, S., Yosinski, J., Vincent, P., Pal, C., 2016. Recombinator Networks: Learning Coarse-to-Fine Feature Aggregation. *CVPR*.

Kersten, J., 2014. Simultaneous feature selection and Gaussian mixture model estimation for supervised classification problems. *Pattern Recognition*, 47.

Kurz, F., Rosenbaum, D., Meynberg, O., Mattyus, G., 2014. Real-time mapping from a helicopter with a new optical sensor system. E. Seyfert, E. Gülch, C. Heipke, J. Schiewe, M. Sester (eds), *34. Wissenschaftlich-Technische Jahrestagung der DGPF*, 23, DFG, 1–8. Beitrag-Nr. 196.

Lin, T., Dollár, P., Girshick, R. B., He, K., Hariharan, B., Belongie, S. J., 2017. Feature Pyramid Networks for Object Detection. *CVPR*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft coco: Common objects in context. *European conference on computer vision*, Springer, 740–755.

Liu, K., Mattyus, G., 2015. Fast Multiclass Vehicle Detection on Aerial Images. *GRSL*.

Liu, L., Pan, Z., Lei, B., 2017. Learning a Rotation Invariant Detector with Rotatable Bounding Box. *arXiv preprint arXiv:1711.09405*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C., Berg, A. C., 2016. SSD: Single Shot MultiBox Detector. *ECCV*.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE.

Meynberg, O., Cui, S., Reinartz, P., 2016. Detection of High-Density Crowds in Aerial Images Using Texture Classification. *Remote Sensing*, 8(6), 1–17.

Mnih, V., 2013. Machine Learning for Aerial Image Labeling. PhD thesis, University of Toronto.

Mundhenk, T. N., Konjevod, G., Sakla, W. A., Boakye, K., 2016. A large contextual dataset for classification, detection and counting of cars with deep learning. B. Leibe, J. Matas, N. Sebe, M. Welling (eds), *ECCV*.

Newell, A., Yang, K., Deng, J., 2016. Stacked Hourglass Networks for Human Pose Estimation. *ECCV*.

Pinheiro, P. H. O., Lin, T., Collobert, R., Dollár, P., 2016. Learning to Refine Object Segments. *ECCV*.

Razakarivony, S., Jurie, F., 2016. Vehicle Detection in Aerial Imagery: A small target detection benchmark. *JVCIR*.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *NIPS*.

Römer, H., Kersten, J., Kiefl, R., Plattner, S., Mager, A., Voigt, S., 2014. Airborne near real-time monitoring of assembly and parking areas in case of large scale public events and natural disasters. *International Journal of Geographical Information Science*, 28(4), 682–699.

Römer, H., Kiefl, R., Henkel, F., Wenxi, C., Nippold, R., Kurz, F., Kippnich, U., 2016. Using airborne remote sensing to increase situational awareness in civil protection and humanitarian relief-the importance of user involvement. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, XLI-B8, 1363-1370.

Van Etten, A., Lindenbaum, D., Bacastow, T. M., 2018. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*.

Waqas Zamir, S., Arora, A., Gupta, A., Khan, S., Sun, G., Shahbaz Khan, F., Zhu, F., Shao, L., Xia, G.-S., Bai, X., 2019. isaid: A large-scale dataset for instance segmentation in aerial images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 28–37.

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S. J., Luo, J., Datcu, M., Pelillo, M., Zhang, L., 2017. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. *CVPR*.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. *CVPR*.

Zhu, H., Chen, X., Dai, W., Fu, K., Ye, Q., Ji, X., 2015. Orientation robust object detection in aerial images using deep convolutional neural network. *ICIP*.