Airborne-Shadow: Towards Fine-Grained Shadow Detection in Aerial Imagery

Seyed Majid Azimi[1][0000−0002−6084−2272] and
Reza Bahmanyar[1][0000−0002−6999−714X]

German Aerospace Center, Remote Sensing Technology Institute, Germany
https://www.dlr.de/eoc/en
{seyedmajid.azimi,reza.bahmanyar}@dlr.de

Abstract. Shadow detection is the first step in the process of shadow removal, which improves the understanding of complex urban scenes in aerial imagery for applications such as autonomous driving, infrastructure monitoring, and mapping. However, the limited annotation in existing datasets hinders the effectiveness of semantic segmentation and the ability of shadow removal algorithms to meet the fine-grained requirements of real-world applications. To address this problem, we present Airborne-Shadow (ASD), a meticulously annotated dataset for shadow detection in aerial imagery. Unlike existing datasets, ASD includes annotations for both heavy and light shadows, covering various structures ranging from buildings and bridges to smaller details such as poles and fences. Therefore, we define shadow detection tasks for multi-class, single class, and merging two classes. Extensive experiments show the challenges that state-of-the-art semantic segmentation and shadow detection algorithms face in handling different shadow sizes, scales, and fine details, while still achieving comparable results to conventional methods. We make the ASD dataset publicly available to encourage progress in shadow detection.

Keywords: Shadow detection · Aerial imagery · Benchmark dataset.

1 Introduction

Shadows are common in natural images taken from ground, aerial, and satellite imagery. When a light source is blocked by an object, as a result a shadow is created, causing colors to appear darker and textures to be less detailed. Therefore, shadow influences almost every Artificial Intelligence (AI) algorithms in the processing of image features. Since shadow can add information to images, such as the geometrical properties of the objects within the images, most of the existing AI-based computer vision and image processing algorithms work more effectively on shadow-free images. For example, detecting an object is certainly easier if the illumination conditions remain constant over the object surface and within different images.

In this paper, we focus on shadow detection in high-resolution aerial images, which can lead to the development of more shadow-adapted AI algorithms that

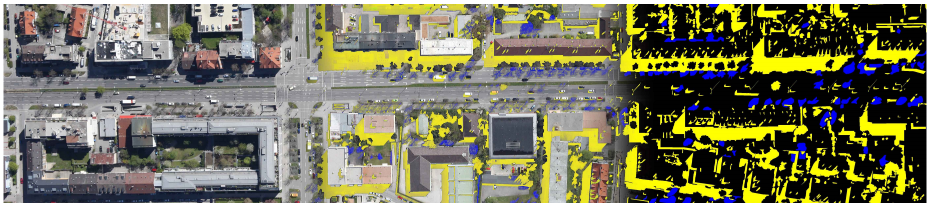


Fig. 1. A cropped section of an aerial image from the ASD dataset with 🟨 heavy and 🟦 light shadow annotations, covering 5.7 km^2 of Munich, Germany.

effectively handle shadowed areas for various applications such as semantic segmentation and object detection, where the existing methods often fail in the areas covered by the shadows of objects such as buildings and trees [1,15]. In the past few years, various data-driven methods have been proposed for shadow detection in ground imagery owing to the various shadow detection datasets. Although the authors tried to cover various scenes and object classes, these datasets mostly contain images from ground perspective. Apart from the Wide Area Motion Imagery (WAMI) [31] dataset introduced recently, existing shadow detection datasets offer only a limited number of aerial images, which inadequately represent the real-world challenges of aerial imagery. As a result, the number of data-driven shadow detection methods for aerial imagery is limited.

To address this shortage, we present ASD, a novel and meticulously annotated aerial shadow detection dataset. To the best of our knowledge, it is the first dataset of its kind to include extensive and thorough manual annotation of natural shadows and to include two distinct shadow classes, heavy and light, providing a new level of detail. Figure 1 shows a section of a large aerial image with its shadow annotations overlaid. The large number of objects as well as their diverse sizes and shapes introduce significant challenge to the shadow annotation in aerial images. In order to assure the quality and validity of the annotations, a rigorous three-step review process involving remote sensing experts was conducted. The annotations were refined based on their feedback. To evaluate the impact of our dataset on the performance of data-driven shadow detection methods, we perform extensive evaluations by training and testing various semantic segmentation and dedicated shadow detection methods. The results highlight the persistent challenges that existing approaches face in accurately detecting and extracting the edges of very small shadows. Furthermore, we assess the generalizability of the trained models on our dataset to other shadow detection datasets. The results show that the model trained on ASD performs well on the aerial part of the other shadow detection datasets, while it has difficulties when applied to ground images, indicating the presence of different challenges in the two domains.

2 Shadow Detection Datasets

Over the past decade, numerous shadow detection and removal datasets have been introduced. In this paper, we present an overview of various shadow detection datasets, with a focus on the publicly available ones with aerial images.

The first shadow detection dataset is the dataset of the University of Central Florida (UCF) introduced by Zhu *et al.* [44]. It has 245 images and their shadow masks of size 257×257 pixels, selected from Overhead Imagery Research Dataset (OIRDS) [29]. The authors created the shadow masks through a manual image annotation. Most of the images in UCF are scenes with dark shadows and dark albedo objects. Guo *et al.* [9] introduced the University of Illinois at Urbana-Champaign (UIUC) shadow detection and removal dataset. This dataset is composed of 108 RGB natural scene shadow images as well as their corresponding shadow-free images and shadow masks. The authors took two images of a scene after manipulating the shadows by blocking the direct light source (to have shadow in the whole scene) or by putting a shadow into the scene. In this dataset, the shadow masks were generated automatically by thresholding the ratio between the shadow and shadow-free images. The authors claim that this approach is more accurate than manual annotation. In this dataset, a large number of images contain close shots of objects. Vicente *et al.* [33] introduced Stony Brook University (SBU) shadow detection dataset with 4,727 images. The authors collected a quarter of the images from the MS COCO dataset [20] and the rest from the web. The images include aerial, landscape, close range, and selfie images. The annotations were performed through a lazy-labeling procedure in which shadows were segmented by a geodesic convexity image segmentation and then were refined manually.

Wang *et al.* [35] generated Image Shadow Triplets Dataset (ISTD), the first benchmark dataset for simultaneous evaluations of shadow detection and removal. This dataset contains 1,870 image triplets including shadow image, shadow mask, and shadow-free image. Each shadow and shadow-free image pair was generated in a fixed exposure setup by inserting and removing an object in the scene. In order to have diverse scenes and shadow shapes, the authors considered 135 different ground materials and objects with various shapes. Hu *et al.* [12] introduced Chinese University of Hong Kong (CUHK) dataset for shadow detection in complex real-world scenarios with 10,500 shadow images and their manually labeled ground-truth masks. The shadow images were collected from a web-search, Google MAP, and three different datasets including the ADE20K [43], KITTI [6], and USR [11] datasets. Therefore, CUHK contains shadows in diverse scenes such as cities, buildings, satellite images, and roads with shadows cast by objects on themselves and the other objects. To leverage instance shadow detection, Wang *et al.* [36] generated Shadow-OBject Association (SOBA) dataset which include 3,623 pairs of shadow and object instances in 1,000 images collected through a web search and from the ADE20K, SBU, ISTD, and MS COCO datasets. Yücel *et al.* [40] presented the Patch Isolation Triplets with Shadow Augmentations (PITSA) dataset, which contains 172k triplets derived from 20k unique shadow-free ground images. The dataset was created using a pipeline

Table 1. Statistics of the shadow detection datasets.

Datasets	# Imgs	Img Type	Shadow Casting	Aerial Imgs								Year
				# Imgs	Avg. Size	Shadow Quantity (px)			# Shadow Instances			
						Total	Mean	Std.	Total	Mean	Std.	
UCF [44]	245	Ground, Aerial	Natural	74	433×594	0.6M	2.1k	4.2k	275	3.72	1.86	2010
UIUC [9]	108	Ground	Artificial	0	-	-	-	-	-	-	-	2012
SBU [33]	4,727	Ground, Aerial	Natural	153	447×557	3.1M	2.7k	7.1k	1,088	7.11	5.21	2016
ISTD [35]	1,870	Ground	Artificial	0	640×480	-	-	-	-	-	-	2018
SOBA [36]	1,000	Ground	Natural, Artificial	0	-	-	-	-	-	-	-	2020
CUHK [12]	10,500	Ground, Aerial	Natural	311	455×767	23.8M	5.6k	17.7k	4,242	13.64	7.51	2021
WAMI [31]	137,180	Aerial	Artificial	137,180	660×440	-	-	-	-	-	-	2021
PITSA [40]	172,539	Ground	Artificial	0	-	-	-	-	-	-	-	2023
ASD	1,408	Aerial	Natural	1,408	512×512	72.7M	51.6k	26.1k	67,781	48.14	25.36	2023

designed for generating large shadow detection and removal datasets, focusing on shadow removal. Shadows were superimposed using a shadow library.

Table 1 shows statistics of the reviewed shadow detection datasets. Among these datasets, UCF, SBU, and CUHK contain a few aerial shadow images; however, the number of samples and their diversity are very limited. Due to their coverage, aerial images usually contain large number of objects with diverse shapes and sizes which makes them different from the terrestrial images. Therefore, in order to develop efficient aerial shadow detection algorithms for real-world applications, large aerial image shadow datasets are crucial. The existing datasets are not appropriate for training Deep Learning (DL)-based algorithms for real-world aerial shadow detection applications. Dealing with this shortcoming, Ufuktepe *et al.* [31] introduced WAMI dataset containing 137k aerial images, which is the largest shadow detection dataset for aerial imagery. The shadows in the dataset were generated and superimposed using a 3D scene model approach, eliminating the need for tedious manual annotation. However, the generated shadows may contain imperfections due to inaccuracies in the 3D model and the superposition process. In addition, this dataset is currently not publicly available.

3 Airborne-Shadow Dataset

To address the limitations of existing datasets for shadow detection in aerial imagery and to promote the development of efficient and effective data-driven shadow detection and removal algorithms, we present the ASD dataset. It consists of 1,408 non-overlapping RGB images with dimensions of 512 × 512 pixels. These images were derived by splitting the 16 large aerial images (5616 × 3744 pixels) from the publicly available SkyScapes dataset[1][1]. The images were captured using the German Aerospace Center (DLR)'s 3K camera system (three DSLR cameras mounted on an airborne platform) during a helicopter flight over Munich, Germany in 2012. The images are nadir looking and have been taken from an altitude of 1000 m where their average Ground Sampling Distance (GSD) is 13 cm/pixel. We split the dataset into training and test sets according to the train-test split of the SkyScapes dataset, where ten large images are assigned to train and six images to the test set.

[1] https://eoc-datasets.dlr.de

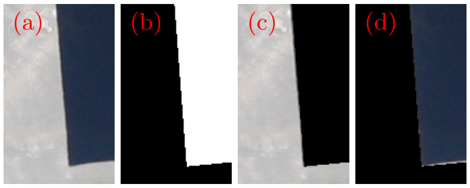


Fig. 2. Example of imperfection in shadow annotation at the shadow borders. Original image (a), binary shadow mask (b), shadow-free region (c), shadowed region in which some non-shadowed pixels are still present (d).

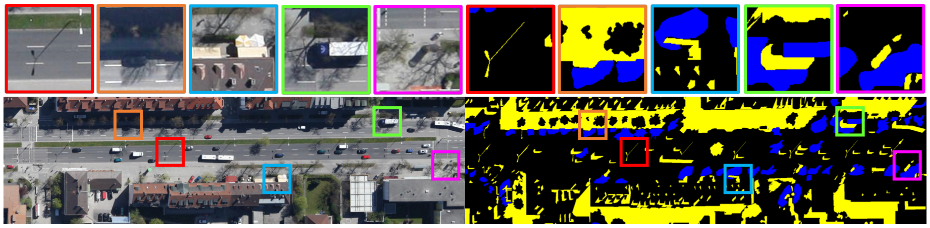


Fig. 3. A scene from the ASD dataset with its overlaid annotation for 🟨 heavy and 🟦 light classes and sample zoomed areas.

3.1 Shadow Annotation

We manually annotated the shadowed areas by 2D polygons and classified them into light and heavy shadows. According to our annotation guidelines, shadows caused by objects that allow sunlight to pass through (such as tree crowns) are classified as light shadows. On the other hand, shadows caused by objects that completely block sunlight are classified as heavy shadows. Furthermore, if an area is partially illuminated by direct sunlight, it is called a light shadow, while if the entire area is illuminated by indirect sunlight reflected from other objects, it is called a heavy shadow. The annotation resulted in 67,781 instances with 72,658,346 annotated pixels, of which about 17.5% are assigned to the light shadow class and the rest to the heavy shadow class. Table 1 represents the statistics of the ASD and the other shadow detection datasets.

Annotating shadows in aerial imagery is challenging due to the large number of objects in each image and the wide range of shadow sizes and shapes. In our dataset, we aimed to provide shadow annotations with exceptional precision, which posed additional challenges, particularly when delineating shadow boundaries and shadows cast by small objects from an aerial perspective (e.g., lamps and poles). In addition, distinguishing between shadowed and non-shadowed pixels, especially along shadow edges, can be a complex task. Figure 2 illustrates an imperfection in the annotation of shadow borders. As another example, for objects such as trees, it is usually difficult to decide for shadow pixels in their border with the tree branches and leaves. In addition, it is not always easy to separate the shadows of densely packed objects, such as trees in a forest. Also, the heavy shadows are annotated more precisely than the light shadows because their boundaries were much clearer. For the light shadows, the annotations are made by drawing a polygon around the approximate boundary of the shadows. Our annotation process overcomes the challenges of shadow annotation, enabling high accuracy and consistency while reducing potential annotation errors. Validation by remote sensing experts at multiple levels ensures the quality and validity of annotations, resulting in refined annotations. Figure 3 shows an example scene from the ASD dataset along with its corresponding annotation. Zoomed areas highlight the precision of the annotations in capturing fine details.

3.2 Comparison to the Other Datasets

Except for the WAMI dataset, existing shadow detection datasets are primarily general-purpose and thus contain either no or only a limited number of aerial images. In addition, these datasets typically consist of images obtained through web searches or from publicly available sources such as OIRDS and Google Maps. Figure 4 shows some examples from the UCF, SBU and CUHK datasets. As can be seen in the examples, the GSDs, viewing angles, illumination conditions, and scene types of the images in SBU are very diverse. Given the limited number of samples in this dataset, this sample heterogeneity may prevent the algorithms from learning different features correctly. Regarding the CUHK dataset, since the images are taken from Google Maps as snapshots of the 3D reconstruction of the environment, they do not represent the true structures and reconstruction distortions are evident in the images. As a consequence, the shadows are also not real shadows. Therefore, a model trained only on these images may not be applicable to real-world scenarios.

In contrast, the images in the ASD dataset were acquired during an aerial campaign specifically designed for urban monitoring, which closely resemble real-world scenarios. In addition, the high quality and resolution of the images in our dataset allow algorithms to learn shadow features from objects of different sizes. The ASD dataset was developed to demonstrate practical applications of aerial image shadow datasets. Careful selection of images from the same flight campaign ensures a diverse range of shadow samples while maintaining consistency in parameters such as illumination, weather conditions, viewing angle, and GSD. Compared to the WAMI dataset, the manual annotation in ASD ensures precise annotation of small objects and shadow edges, which may be prone to errors in the automatic annotation of WAMI due to imperfections in 3D reconstruction and overlay techniques. In addition, ASD includes two shadow classes, which improves fine-grained shadow detection in aerial imagery. Furthermore, by building ASD on top of the SkyScapes fine-grained semantic segmentation dataset, it provides opportunities to explore shadow detection and removal on different objects and surfaces, as well as to investigate the impact of shadow removal on the segmentation of different semantic categories. In contrast to the WAMI dataset, our dataset will be made publicly available, fostering advancements in the development of shadow detection algorithms.

We compare the statistics of our dataset with publicly available shadow datasets containing aerial imagery. As shown in Table 1, the total number of annotated shadow instances in ASD is about 16 times higher than that of CUHK.

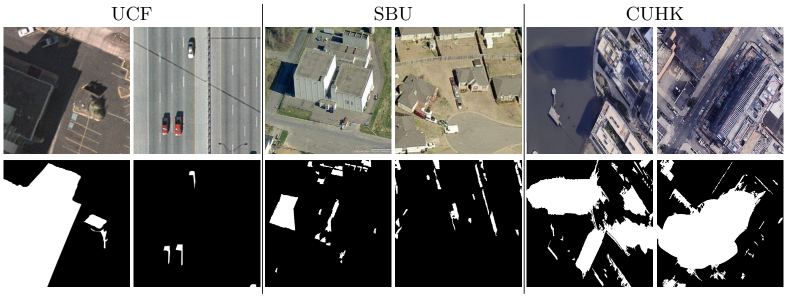
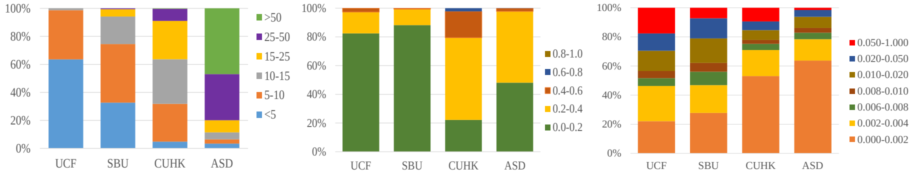


Fig. 4. Sample aerial images with their corresponding masks from the UCF, SBU, and CUHK shadow datasets.



(a) (b) (c)

Fig. 5. Statistical properties of the aerial set of the UCF, SBU, CUHK, and ASD datasets. The number of annotated shadow instances per image (a), shadowed fraction of images (b), and the size of annotated shadow instances relative to image sizes (c).

Also, according to the diagrams in Figure 5, ASD contains a larger number of shadow instances, especially at smaller sizes, than the other datasets. According to Figure 5-(a) most of the images in our dataset contains more than 25 instances in contrast to the other datsets which rarely have such images. Moreover, Figure 5-(c) indicates that most of the annotations in our dataset are tiny shadows which could be caused by the fine looking objects in aerial images. This also shows the high quality and resolution of our images so that even such fine annotations could be provided. In addition, similar to the other shadow detection datasets, the number of shadowed and non-shadowed pixels is not balanced in our dataset. However, according to Figure 5-(b) about half of the images in ASD are covered by 20% to 40% shadow pixels.

4 Shadow Detection Methods

Previous works in the field of shadow detection have primarily focused on the extraction of shadows using engineered feature descriptors, illumination models and physical properties of shadows [24,44,17,14,8,9,34,30]. The classification of shadowed pixels was often performed using algorithms such as Support Vector Machine (SVM) and decision tree. In the field of remote sensing, threshold-based methods have been proposed with promising results in [7], using Gram-Schmidt orthogonalization in the LAB color space. In addition, an image index has been developed in [22] that incorporates all available bands of multispectral images. Furthermore, the use of Principal Component Analysis (PCA) to detect shadows in multispectral satellite images was introduced in [5]. These methods are not specifically designed for shadow detection. Therefore, their results are not directly comparable with state-of-the-art shadow detection methods.

Inspired by the successful applications of Deep Neural Network (DNN)s in various image processing and computer vision tasks, a number of recent works have proposed to exploit the ability of DNNs to automatically learn relevant features for shadow detection. A structured Convolutional Neural Network (CNN) framework was used in [18] to predict the local structure of shadow edges, which improves the accuracy and local consistency of pixel classification. A CNN structure was proposed in [33] for patch-level shadow detection, which refines the detected shadow patches based on image-level semantics. A method called scGAN was proposed in [23], which uses a stacked Conditional Generative Adversarial Networks (CGAN) with a sensitivity parameter to both control the sensitivity of the generator and weight the relative importance of the shadow and non-shadow classes. The BDRAR method, introduced in [45], utilizes a bidirectional Feature Pyramid Network (FPN) with Recurrent Attention Residual (RAR) modules. It extracts feature maps at different resolutions using a CNN, enabling the capture of both shadow details and shadow semantics. Another method, called Direction-aware Spatial Context (DSC), proposed in [13], incorporates an attention mechanism in a spatial Recurrent Neural Network (RNN) with a newly introduced DSC module to learn spatial contexts of images and shadows. Despite its performance, DSC is unable to outperform BDRAR. The FDRNet method proposed in [46] uses a feature decomposition and reweighting scheme to mitigate intensity bias. It separates features into intensity-variant and -invariant components and reweights these two types of features to redistribute attention and balance their evaluation. The Stacked Conditional Generative Adversarial Network (ST-CGAN) method, introduced in [35], enables joint detection and removal of shadows in an end-to-end manner. It utilizes two stacked CGANs, with the first generator producing shadow detection masks and the second generator removing shadows based on the generated masks. A context preserver CNN called CPAdv-Net was proposed in [21], which is based on the U-Net [27] structure and trained by adversarial images. In [12], a fast shadow detection method called FSDNet was proposed. It utilizes the MobileNet-V2 [28] architecture and a novel detail enhancement module. FSDNet achieves competitive results with state-of-the-art methods in a shorter time.

5 Evaluation Metrics

For the evaluations, we use commonly used metrics including mean Intersection over Union (IoU), Dice similarity coefficient, and Balanced Error Rate (BER). In the following equations, n_{ij} is the number of pixels of class i predicted as class j and n_{cl} is the number of classes, with $t_i = \sum_j n_{ij}$ representing the total number of pixels of class i. TP, TN, FP, and FN denote the number of true

Table 2. Benchmark on our dataset for multiple semantic segmentation and dedicated shadow detection methods for different class setups evaluated by IoU↑, Dice↑, BER↓, and Class IoU↑. B, H, L, and M denote background, heavy shadow, light shadow, and merged classes. The values are given in percent. In red and blue are the best and second best results. IoU is in %.

Method	**Heavy and Light**			**Heavy**				**Merged**			
	IoU	Dice	Class IoU [B,H,L]	IoU	BER	Dice	Class IoU [B,H]	IoU	BER	Dice	Class IoU [B,M]
AdapNet-Incep.V4 [32]	66.59	78.17	[89.36, 69.30, 41.12]	80.27	11.64	88.65	[91.23, 69.32]	79.02	12.91	87.91	[89.23, 68.80]
BiSeNet-Res.152 [39]	69.15	80.51	[89.70, 70.16, 47.60]	80.80	11.43	88.99	[91.52, 70.08]	80.02	12.24	88.58	[89.78, 70.27]
DeepLabv3 - Res.152 [2]	65.28	77.43	[88.24, 64.99, 42.62]	78.26	13.43	87.28	[90.38, 66.15]	77.20	14.24	86.69	[88.26, 66.13]
DeepLabv3+ - Res.152 [3]	68.45	79.81	[89.80, 70.62, 44.94]	81.22	11.19	89.27	[91.73, 70.71]	80.30	11.89	88.76	[89.87, 70.72]
DeepLabv3+ - Xcep.65 [3]	**71.00**	**81.88**	[90.67, 72.59, **49.72**]	**82.57**	**10.48**	**90.13**	[92.41, **72.73**]	81.84	**10.75**	89.76	[90.67, 73.02]
DenseASPP-Res.50 [38]	65.36	77.03	[88.76, 69.10, 38.23]	79.85	10.65	88.39	[90.67, 69.04]	78.68	12.89	87.70	[88.95, 68.41]
Encoder-Decoder-Skip	69.76	80.86	[90.23, 71.95, 47.10]	81.70	10.98	89.57	[91.98, 71.41]	80.37	11.78	88.81	[89.89, 70.85]
FC-DenseNet65	70.02	80.99	[90.62, **72.77**, 46.67]	82.02	12.04	89.76	[**92.45**, 71.59]	**81.88**	**10.75**	89.78	[90.70, **73.06**]
FRRN-A-Incep.V4 [26]	68.98	80.15	[90.24, 71.70, 45.00]	81.81	11.37	89.64	[92.15, 71.47]	80.92	11.47	89.16	[90.21, 71.63]
FRRN-B [26]	69.17	80.31	[90.25, 71.88, 45.37]	81.72	11.07	89.59	[92.02, 71.42]	80.78	11.85	89.07	[90.23, 71.33]
GCN-Res101 [25]	69.47	80.60	[90.31, 71.76, 46.34]	81.57	11.08	89.49	[91.93, 71.21]	80.84	11.89	89.10	[90.29, 71.40]
InternImage [37]	**71.39**	**82.16**	[**91.00**, **73.37**, **49.79**]	**82.75**	**10.15**	**90.35**	[**92.45**, **73.04**]	**82.51**	**9.95**	**90.02**	[**91.00**, **74.03**]
MobileUNet-Skip-Incep.V4 [10]	68.14	79.46	[89.87, 70.98, 43.56]	81.35	11.64	89.34	[91.93, 70.77]	80.38	11.55	88.82	[89.82, 70.95]
PSPNet-Res.152 [42]	68.13	79.50	[89.70, 70.77, 43.94]	81.12	11.65	89.19	[91.78, 70.45]	80.17	11.76	88.68	[89.72, 70.61]
RefineNet-Res.152 [19]	69.05	80.32	[89.85, 71.11, 46.21]	81.46	10.92	89.42	[91.82, 71.10]	80.51	11.20	88.91	[89.80, 71.22]
BDRAR [45]	68.01	79.33	[89.91, 71.05, 43.09]	81.84	10.99	89.69	[92.08, 71.60]	80.63	12.02	88.91	[90.05, 71.21]
FDRNet [46]	69.87	80.88	[**90.69**, 72.61, 46.33]	**82.46**	11.09	90.08	[92.42, 72.49]	81.68	10.95	**90.14**	[**90.74**, 72.98]

positives, true negatives, false positives, and false negatives, respectively. P and T refer to prediction and ground truth, respectively. For IoU and Dice, higher values indicate better results, while for BER, lower values show better results.

$$MeanIoU = \frac{1}{n_{cl}} \sum_i \frac{n_{i,i}}{t_i + \sum_j n_{j,i} - n_{i,i}},$$

$$Dice = \frac{2 \mid P \cap T \mid}{\mid P \mid + \mid T \mid},$$

$$BER = 1 - \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right).$$

6 Results and Discussion

In this section, we evaluate the performance of several DL-based semantic segmentation and shadow detection methods on the ASD dataset. We train and test them on the training and test sets of the dataset for three tasks: heavy and light shadow classes, heavy shadow class only, and merging the two classes.

Among the existing methods, we select methods with publicly available training source code, including the state-of-the-art InternImage [37] together with AdaptNet [32], BiSeNet [39], DeepLab [2], DeepLabv3+ [3], DenseASPP [38], Context-Encoding [41], FC-DenseNet [16], FRRN [26], GCN [25], MobileUNet [27,10], PSPNet [42], and RefineNet [19] for semantic segmentation, and BDRAR [45] and FDRNet [46] for dedicated shadow detection methods. Moreover, for semantic segmentation, we conduct experiments with multiple available variants and different backbones of each method. The best result for each method is listed in Table 2. We refer the reader to the supplementary material for the full set of experimental results.

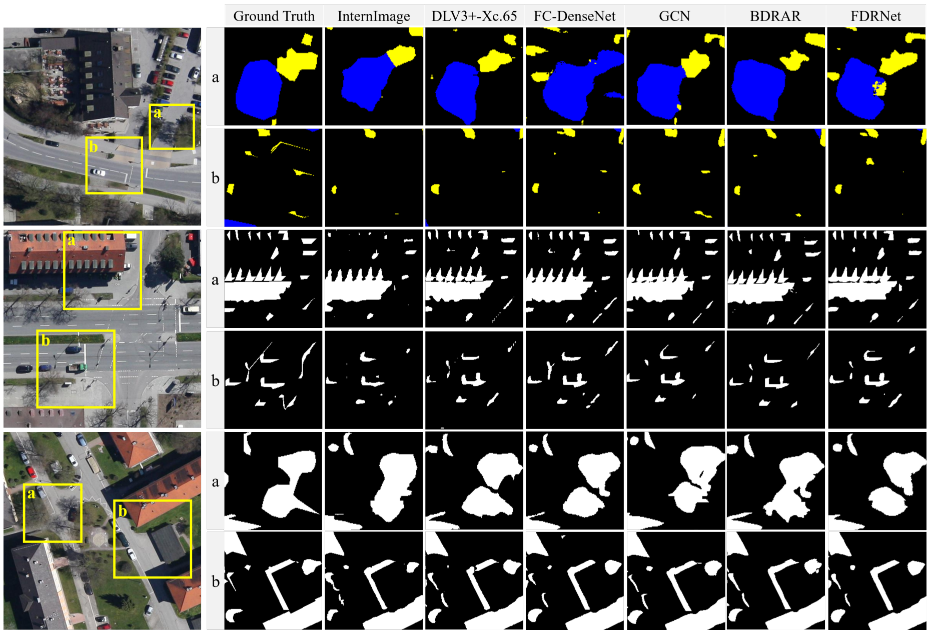


Fig. 6. Shadow segmentation results by six different methods for two class (first row), only heavy (2nd row), and the merged classes (3rd row). 🟨 heavy and 🟦 light shadows.

For our experiments, we crop the images into 512 × 512px patches. The reason is the original size of images is 21 MP which does not fit into the GPU memory. We use Titan XP and Quadro P6000 GPUs for training of the semantic segmentation algorithms and A100 GPU for training the InternImage and shadow detection algorithms. Regarding data augmentation, we apply both horizontal and vertical flipping, as well as 50% overlap between neighboring crops. During inference, we apply 10% overlap to alleviate the lower performance at boundary regions. The learning rate is 0.0001 with the batch size of 1. We train the algorithms for 60 epochs to make the comparison fair with all algorithms converged until this step instead of using early stopping and learning-rate scheduling techniques. In total, there are 8820 training crops. For training InernImage[2], BDRAR[3] and FDRNet[4], we use the original implementations.

As shown in Table 2, InternImage achieves the highest performance of 71.29 IoU and 10.15, 9.95 BER in ASD-two-classes, heavy(H) and -merged(M) respectively, closely followed by DeepLabv3+-Xcep.65 as the second best method for most of the evaluations. Notably, the results of InternImage, currently the state-of-the-art semantic segmentation method on the CityScapes dataset [4], are not significantly better than many of the earlier methods, indicating that the challenges in ASD are still affecting newer methods. Figure 6 demonstrates the

[2] https://github.com/OpenGVLab/InternImage (accessed June 5 2023)
[3] https://github.com/zijundeng/BDRAR (accessed June 5 2023)
[4] https://github.com/rayleizhu/FDRNet (accessed June 5 2023)

qualitative performance of selected algorithms. InternImage and DeepLabv3+-Xcep.65 excel at extracting large shadow instances, while FC-DenseNet and GCN perform better on small shadow instances such as poles. In addition, FDRNet and BDRAR have comparable performance. Overall, the methods perform better with heavy shadows and face more challenges with light shadows. Light shadows often have lower contrast and can be more difficult to accurately detect than heavy shadows, which tend to have more distinct boundaries and higher contrast. Therefore, methods that focus on edge refinement and enhancing contrast in light shadow regions could potentially improve the overall performance of shadow detection algorithms in real-world scenarios. Still almost all algorithms have a major difficulty in extracting shadows of tiny objects e.g., poles which is important in pole detection algorithms as poles normally appear as one point in ortho aerial images. With further investigation, as expected we notice that algorithms are under performing in edge areas. We contemplate that one of the reasons for the slight better performance of InternImage is due to the more extended extracted shadowed areas leading to higher quantitative performance especially in IoU. Overall, ASD shows that there is a significant challenge still remaining in the shadow detection task in aerial imagery which we hope this dataset could support further developments to shorten this gap.

Cross Validation To assess the generalizability of the models trained on our dataset, we trained DeepLabv3+-Xcep.65 on ASD and tested it on SBU, CUHK, and ISTD datasets. Similarly, we tested models trained on these three datasets on ASD. The quantitative results are presented in Table 3, and Figure 7 illustrates some qualitative results. The models trained on the heavy and merged classes of ASD performs significantly better on the aerial images of the SBU and CUHK datasets than on the entire dataset, which includes both aerial and ground images. Furthermore, the results on the ISTD dataset clearly show that the model trained on the aerial images does not generalize well to the ground images. This highlights the different challenges in different domains of shadow detection. BDRAR and FDRNet achieve 3.64 and 3.04 in BER on SBU dataset respectively while the trained algorithms on ASD-H and -M achieve 8.04. The reason is the significant change in aerial images of SBU such as oblique ones shown in Figure 7. This justification is confirmed on CUHK dataset in which trained algorithms on ASD-H and -M achieve 11.79 versus the trained one on CUHK yielding 13.34 while tested on the aerial images in CUHK dataset i.e., CUHK-aerial(A). Interestingly, DeepLabv3+-Xcep.65 achieves 4.90 in BER when trained on SBU. We will further investigate the lower performance of BDRAR and FDRNet compared to DeepLabv3+-Xcep.65 when tested on ASD in the future. ISDT dataset has no aerial images, however, we can see that trained algorithm on ASD-H or -M can achieve better BER on ISDT with 25.49 and 24.67 than trained ones on ISDT, tested on ASD-H and -M with 27.16 and 26.50. The same phenomena occurs in the SBU and CUHK datasets. Therefore, ASD has a better generalization capability than the aerial images contained in SBU and CUHK dataset. Furthermore, the model trained on ASD performs significantly better on the

Table 3. Evaluation of the generalizability of models trained on ASD and tested on SBU, ISTD, and CUHK datasets, and vice versa, using the DeepLabv3+-Xcep.65 network. H and M refer to the heavy shadow and merged classes in ASD, and A denotes the aerial parts of the SBU and CUHK dataset. The results are in percent, with the best and second best results marked in red and blue, respectively. IoU is in %.

ASD vs. SBU					ASD vs. CUHK					ASD vs. ISTD				
Train	**Test**	**IoU↑**	**BER↓**	**Dice↑**	**Train**	**Test**	**IoU↑**	**BER↓**	**Dice↑**	**Train**	**Test**	**IoU↑**	**BER↓**	**Dice↑**
ASD-H	SBU	65.95	13.70	77.56	ASD-H	CUHK	44.22	22.56	59.07	ASD-H	ISTD	53.09	25.49	63.39
ASD-M	SBU	67.06	14.26	78.57	ASD-M	CUHK	45.25	22.82	60.27	ASD-M	ISTD	54.43	24.67	65.58
ASD-H	SBU-A	85.03	8.04	91.52	ASD-H	CUHK-A	70.27	11.79	81.80	ISTD	ISTD	91.47	6.62	95.46
ASD-M	SBU-A	85.29	8.72	91.69	ASD-M	CUHK-A	75.57	12.68	82.82	ISTD	ASD-H	61.21	27.16	74.46
SBU	SBU	90.17	4.90	94.73	CUHK	CUHK	83.55	9.07	91.03	ISTD	ASD-M	60.44	26.50	74.03
SBU	SBU-A	84.09	7.17	90.90	CUHK	CUHK-A	78.61	13.34	87.84					
SBU	ASD-H	69.77	15.55	80.96	CUHK	ASD-H	64.01	24.72	77.25					
SBU	ASD-M	64.41	17.29	76.80	CUHK	ASD-M	64.81	23.63	78.02					

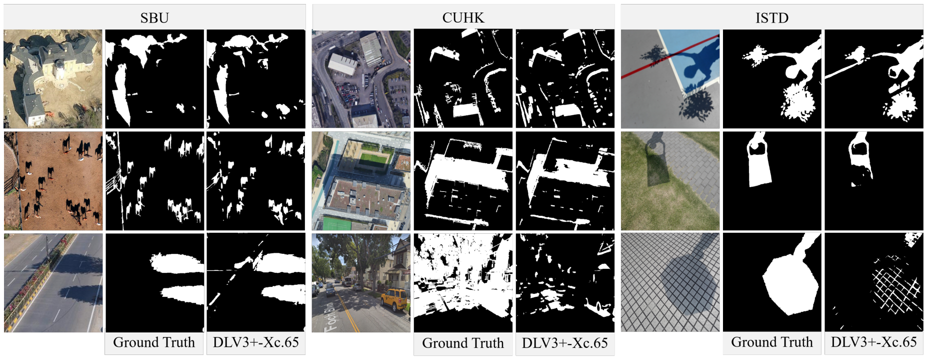


Fig. 7. Shadow segmentation results for DeepLabv3+-Xcep.65 trained on ASD (merged classes) and tested on SBU, CUHK, and ISTD test sets.

SBU and CUHK aerial sets than the model trained on these two datasets and tested on ASD. This indicates that ASD generalizes better than the airborne parts of SBU and CUHK.

7 Conclusion and future works

In this paper, we present ASD, the first and largest publicly available dataset dedicated to fine-grained shadow detection in aerial imagery, including both heavy and light shadow classes. By evaluating and benchmarking state-of-the-art methods, we found that current algorithms struggle to achieve high accuracy on this dataset, highlighting the intricate details and complexity of ASD. In addition, the cross-dataset evaluation results show that the model trained on ASD performs well on the aerial set of the other shadow detection datasets, indicating the potential transferability of models trained on ASD to other similar aerial image datasets. Altogether, ASD will advance the development of efficient and effective shadow detection methods, ultimately improving shadow removal and feature extraction for various applications such as HD map creation and autonomous driving.

References

1. Azimi, S.M., Henry, C., Sommer, L., Schumann, A., Vig, E.: Skyscapes fine-grained semantic understanding of aerial scenes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7393–7403 (2019)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic Image Segmentation With Deep Convolutional Nets And Fully Connected CRFs. arXiv preprint arXiv:1412.7062 (2014)
3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 801–818 (2018)
4. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3213–3223 (2016)
5. Dharani, M., Sreenivasulu, G.: Shadow detection using index-based principal component analysis of satellite images. In: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). pp. 182–187. IEEE (2019)
6. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3354–3361 (2012)
7. Guo, J., Yang, F., Tan, H., Lei, B.: Shadow extraction from high-resolution remote sensing images based on gram-schmidt orthogonalization in lab space. In: 3rd International Symposium of Space Optical Instruments and Applications. pp. 321–328. Springer (2017)
8. Guo, R., Dai, Q., Hoiem, D.: Single-image shadow detection and removal using paired regions. In: CVPR 2011. pp. 2033–2040. IEEE (2011)
9. Guo, R., Dai, Q., Hoiem, D.: Paired regions for shadow detection and removal. IEEE transactions on pattern analysis and machine intelligence **35**(12), 2956–2967 (2012)
10. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
11. Hu, X., Jiang, Y., Fu, C.W., Heng, P.A.: Mask-shadowgan: Learning to remove shadows from unpaired data. arXiv preprint arXiv:1903.10683 (2019)
12. Hu, X., Wang, T., Fu, C.W., Jiang, Y., Wang, Q., Heng, P.A.: Revisiting shadow detection: A new benchmark dataset for complex world. IEEE Transactions on Image Processing **30**, 1925–1934 (2021)
13. Hu, X., Zhu, L., Fu, C.W., Qin, J., Heng, P.A.: Direction-aware spatial context features for shadow detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7454–7462 (2018)
14. Huang, X., Hua, G., Tumblin, J., Williams, L.: What characterizes a shadow boundary under the sun and sky? In: 2011 International Conference On Computer Vision. pp. 898–905. IEEE (2011)
15. ISPRS: 2D Semantic Labeling Dataset. `http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html`, [Online; accessed 01-March-2023]
16. Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y.: The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 11–19 (2017)

17. Lalonde, J.F., Efros, A.A., Narasimhan, S.G.: Detecting ground shadows in outdoor consumer photographs. In: European conference on computer vision. pp. 322–335. Springer (2010)
18. Li Shen, Teck Wee Chua, Leman, K.: Shadow optimization from structured deep edge detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2067–2074 (2015)
19. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755 (2014)
21. Mohajerani, S., Saeedi, P.: Shadow detection in single RGB images using a context preserver convolutional neural network trained by multiple adversarial examples. IEEE Transactions on Image Processing **28**(8), 4117–4129 (2019)
22. Mostafa, Y., Abdelhafiz, A.: Accurate shadow detection from high-resolution satellite images. IEEE Geoscience and Remote Sensing Letters **14**(4), 494–498 (2017)
23. Nguyen, V., Vicente, Y., Tomas, F., Zhao, M., Hoai, M., Samaras, D.: Shadow detection with conditional generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4510–4518 (2017)
24. Panagopoulos, A., Wang, C., Samaras, D., Paragios, N.: Illumination estimation and cast shadow detection through a higher-order graphical model. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 673–680 (2011)
25. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters–improve semantic segmentation by global convolutional network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4353–4361 (2017)
26. Pohlen, T., Hermans, A., Mathias, M., Leibe, B.: Full-resolution residual networks for semantic segmentation in street scenes. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3309–3318 (July 2017). https://doi.org/10.1109/CVPR.2017.353
27. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer International Publishing, Cham (2015)
28. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4510–4520 (2018). https://doi.org/10.1109/CVPR.2018.00474
29. Tanner, F., Colder, B., Pullen, C., Heagy, D., Oertel, C., Sallee, P.: Overhead imagery research data set (OIRDS) – an annotated data library and tools to aid in the development of computer vision algorithms (2009)
30. Tian, J., Qi, X., Qu, L., Tang, Y.: New spectrum ratio properties and features for shadow detection. Pattern Recognition **51**, 85 – 96 (2016). https://doi.org/https://doi.org/10.1016/j.patcog.2015.09.006
31. Ufuktepe, D.K., Collins, J., Ufuktepe, E., Fraser, J., Krock, T., Palaniappan, K.: Learning-based shadow detection in aerial imagery using automatic training supervision from 3d point clouds. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 3919–3928 (2021). https://doi.org/10.1109/ICCVW54120.2021.00439
32. Valada, A., Vertens, J., Dhall, A., Burgard, W.: Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In: 2017 IEEE International

Conference on Robotics and Automation (ICRA). pp. 4644–4651 (May 2017). https://doi.org/10.1109/ICRA.2017.7989540

33. Vicente, T.F.Y., Hou, L., Yu, C.P., Hoai, M., Samaras, D.: Large-scale training of shadow detectors with noisily-annotated shadow examples. In: European Conference on Computer Vision. pp. 816–832. Springer (2016)
34. Vicente, Y., Tomas, F., Hoai, M., Samaras, D.: Leave-one-out kernel optimization for shadow detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3388–3396 (2015)
35. Wang, J., Li, X., Yang, J.: Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1788–1797 (2018)
36. Wang, T., Hu, X., Wang, Q., Heng, P.A., Fu, C.W.: Instance shadow detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
37. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al.: Internimage: Exploring large-scale vision foundation models with deformable convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14408–14419 (2023)
38. Yang, M., Yu, K., Zhang, C., Li, Z., Deepmotion, K.Y.: DenseASPP for Semantic Segmentation in Street Scenes. In: CVPR. pp. 3684–3692. Salt Lake City (2018)
39. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 325–341 (September 2018)
40. Yücel, M.K., Dimaridou, V., Manganelli, B., Ozay, M., Drosou, A., Saa-Garriga, A.: LRA&LDRA: Rethinking residual predictions for efficient shadow detection and removal. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4925–4935 (2023)
41. Zhang, H., Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
42. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid Scene Parsing Network. In: CVPR. Honolulu (2017)
43. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5122–5130 (2017)
44. Zhu, J., Samuel, K.G., Masood, S.Z., Tappen, M.F.: Learning to recognize shadows in monochromatic natural images. In: 2010 IEEE Computer Society conference on computer vision and pattern recognition. pp. 223–230. IEEE (2010)
45. Zhu, L., Deng, Z., Hu, X., Fu, C.W., Xu, X., Qin, J., Heng, P.A.: Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 121–136 (2018)
46. Zhu, L., Xu, K., Ke, Z., Lau, R.W.: Mitigating intensity bias in shadow detection via feature decomposition and reweighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4702–4711 (2021)

Airborne-Shadow: Towards Fine-Grained Shadow Detection in Aerial Imagery - Supplementary Materials

Seyed Majid Azimi[1] $^{[0000-0002-6084-2272]}$ and
Reza Bahmanyar[1] $^{[0000-0002-6999-714X]}$

German Aerospace Center, Remote Sensing Technology Institute, Germany
https://www.dlr.de/eoc/en
{seyedmajid.azimi,reza.bahmanyar}@dlr.de

This document presents the comprehensive benchmarking results of several image segmentation networks on the Airborne-Shadow (ASD) dataset. These methods include: Auto-Deeplab [1], DenseASPP [2], BiSeNet [3], Context-Encoding [4], and OcNet [5], PSPNet [6], or stacks of convolutional layers with different dilation rates, as in DeepLab [7], FRRN [8], MobileNet [9], RefineNet [10], Deeplabv3+ [11], AdapNet [12], and FC-DenseNet [13], as well as a custom U-Net-like MobileNet and custom Decoder-Encoder with skip-connections. We also considered shadow detection algorithms such as BDRAR [14] and FDRNET [15].

The segmentation performance is evaluated for two shadow classes: heavy and light shadows (Table 1). Additionally, we analyze the results specifically for the heavy shadow class (Table 2), as well as when both shadow classes are merged (Table 3). Evaluation metrics include IoU, f.w. IoU, Class IoU, BER, f.w. BER, Precision, Recall, and Dice scores. Higher values indicate better algorithm performance, except for the BER and f.w. BER metrics. For the IoU and BER metrics, we also considered their frequency weighted (f.w.) values.

We also provide visualizations of example segmentations for the three tasks in Figure 1, Figure 2, and Figure 3. Each figure shows the results of the best performing variant of the main segmentation network for the respective task.

Figure 4 shows the performance of InternImage on one original test images from the ASD dataset in the heavy class benchmark.

When applied in the real-world scenarios, the Figure 5 and Figure 6 shows the generalization capability of the best performing algorithm trained on ASD when tested on other aerial images, acquired at a different time with different resolution over the city of Karlsruhe as a different area for heavy and merged benchmarks. Figure 7 shows the same performance, but over the city of Wolfratshausen for the two-class shadow benchmark.

Table 1. Benchmark on Airborne-Shadow dataset for the heavy and light shadow classes. B, H, and L denote Background, Heavy shadow, and Light shadow classes. In blue and red are the best and second best results. IoU is in %.

Method	IoU	f.w. IoU	Recall	Precision	Dice	Class IoU [B,H,L]
AdapNet	66.53	83.34	93.79	83.25	78.10	[89.30, 69.45, 40.84]
AdapNet - Incep.V4	66.59	83.37	93.81	83.32	78.17	[89.36, 69.30, 41.12]
BiSeNet - Res.50	68.29	83.71	93.92	83.19	79.78	[89.49, 69.59, 45.80]
BiSeNet - Res.101	68.76	83.97	94.00	83.56	80.14	[89.60, 70.36, 46.31]
BiSeNet - Res.152	69.15	84.06	94.03	83.07	80.51	[89.70, 70.16, 47.60]
DeepLabv3 - Res.50	65.13	81.38	92.95	80.58	77.36	[87.88, 64.70, 42.81]
DeepLabv3 - Res.101	64.72	81.21	92.92	81.49	77.01	[87.85, 64.11, 42.21]
DeepLabv3 - Res.152	65.28	81.70	93.13	81.89	77.43	[88.24, 64.99, 42.62]
DeepLabV3 - Incep.V4	25.40	57.36	83.80	78.81	29.06	[75.68, 00.51, 00.00]
DeepLabv3+ - Res.50	67.96	83.89	93.97	82.49	79.39	[89.62, 70.31, 43.96]
DeepLabv3+ - Res.101	67.92	83.84	93.92	81.70	79.36	[89.57, 70.28, 43.92]
DeepLabv3+ - Res.152	68.45	84.12	94.06	82.32	79.81	[89.80, 70.62, 44.94]
DeepLabv3+ - Xcep.65	**71.00**	**85.38**	94.58	84.66	**81.88**	[90.67, 72.59, **49.72**]
DenseASPP - MobileNetV2	25.41	55.44	81.55	31.98	30.53	[72.23, 03.83, 00.16]
DenseASPP - Res.50	65.36	82.76	93.50	81.58	77.03	[88.76, 69.10, 38.23]
DenseASPP - Res.101	64.72	82.71	93.52	82.14	76.35	[88.83, 68.96, 36.38]
DenseASPP - Res.152	65.31	82.72	93.49	81.11	77.02	[88.81, 68.63, 38.49]
Encoder-Decoder	67.96	83.79	93.94	82.51	79.43	[89.59, 69.88, 44.43]
Encoder-Decoder - Incep.V4	67.53	83.62	93.86	82.20	79.05	[89.47, 69.64, 43.48]
Encoder-Decoder-Skip	69.76	84.81	94.34	83.43	80.86	[90.23, 71.95, 47.10]
Encoder-Decoder-Skip - Incep.V4	69.46	84.75	94.32	83.38	80.61	[90.29, 71.58, 46.50]
FC-DenseNet56	70.02	85.25	94.51	83.01	80.99	[90.62, **72.77**, 46.67]
FC-DenseNet56 - Incep.V4	69.49	85.14	94.49	83.62	80.51	[90.62, 72.48, 45.37]
FC-DenseNet67	69.20	85.16	94.50	84.37	80.19	[90.58, 73.00, 44.02]
FC-DenseNet67 - Incep.V4	69.83	85.22	94.55	84.62	80.82	[90.66, 72.55, 46.27]
FC-DenseNet103	69.63	85.13	94.51	84.55	80.66	[90.60, 72.37, 45.91]
FC-DenseNet103 - Incep.V4	69.59	85.21	94.54	84.96	80.59	[90.64, 72.72, 45.43]
FRRN-A	68.89	84.76	94.36	84.66	80.04	[90.30, 71.98, 44.40]
FRRN-A - Incep.V4	68.98	84.68	94.33	84.75	80.15	[90.24, 71.70, 45.00]
FRRN-B	69.17	84.74	94.34	84.31	80.31	[90.25, 71.88, 45.37]
FRRN-B - Incep.V4	68.63	84.54	94.27	84.40	79.85	[90.15, 71.49, 44.26]
GCN - Res.50	69.29	84.75	94.34	84.36	80.43	[90.24, 71.87, 45.77]
GCN - Res.101	69.47	84.80	94.36	84.39	80.60	[90.31, 71.76, 46.34]
GCN - Res.152	68.70	84.25	94.11	82.88	80.00	[89.86, 70.94, 45.30]
InternImage	**71.39**	**85.43**	**94.61**	**85.44**	**82.16**	[**91.00**, **73.37**, **49.79**]
MobileUNet	65.21	82.31	93.32	80.77	77.07	[88.51, 67.57, 39.54]
MobileUNet-Skip	67.95	84.17	94.10	83.01	79.28	[89.88, 70.96, 43.03]
MobileUNet-Skip - Incep.V4	68.14	84.20	94.11	83.20	79.46	[89.87, 70.98, 43.56]
PSPNet - Res.50	67.80	84.00	94.00	81.92	79.19	[89.72, 70.67, 43.02]
PSPNet - Res.101	67.66	83.89	93.95	81.79	79.06	[89.62, 70.58, 42.77]
PSPNet - Res.152	68.13	84.04	94.02	82.37	79.50	[89.70, 70.77, 43.94]
RefineNet - Res.50	68.86	84.27	94.12	82.82	80.15	[89.85, 71.00, 45.72]
RefineNet - Res.101	68.89	84.34	94.16	83.46	80.16	[89.93, 71.07, 45.66]
RefineNet - Res.152	69.05	84.31	94.12	82.49	80.32	[89.85, 71.11, 46.21]
BDRAR	68.01	84.25	94.19	83.08	79.33	[89.91, 71.05, 43.09]
FDRNet	69.87	85.28	**94.62**	**84.67**	80.88	[**90.69**, 72.61, 46.33]

Table 2. Benchmark on Airborne-Shadow dataset with only the heavy shadow class. B and H denote Background and Heavy shadow classes. In blue and red are the best and second best results. IoU is in %.

Method	IoU	f.w. IoU	BER	f.w. BER	Recall	Precision	Dice	Class IoU [B,H]
AdapNet	80.34	86.93	12.38	9.85	92.84	89.85	88.67	[91.47, 69.20]
AdapNet-Incep.V4	80.27	86.75	11.64	9.26	92.68	88.94	88.65	[91.23, 69.32]
BiSeNet - Res.50	79.96	86.66	12.60	10.03	92.68	89.58	88.42	[91.29, 68.63]
BiSeNet - Res.101	80.26	86.90	12.55	9.99	92.84	89.96	88.62	[91.47, 69.05]
BiSeNet - Res.152	80.80	87.14	11.43	9.10	92.92	89.43	88.99	[91.52, 70.08]
DeepLabv3 - Res.50	77.60	85.05	14.33	11.40	91.71	88.13	86.82	[90.20, 65.00]
DeepLabv3 - Res.101	77.64	85.04	14.06	11.19	91.67	87.87	86.85	[90.14, 65.14]
DeepLabv3 - Res.152	78.26	85.43	13.43	10.69	91.90	88.06	87.28	[90.38, 66.15]
DeepLabV3 - Incep.V4	52.92	66.09	33.20	26.43	77.63	66.03	66.38	[75.18, 30.66]
DeepLabv3+ - Res50	81.07	87.34	11.27	8.97	93.04	89.64	89.17	[91.66, 70.49]
DeepLabv3+ - Res101	81.13	87.36	11.11	8.84	93.05	89.54	89.21	[91.65, 70.61]
DeepLabv3+ - Res152	81.22	87.44	11.19	8.91	93.11	89.75	89.27	[91.73, 70.71]
DeepLabv3+ - Xcep.65	82.57	88.39	10.48	8.34	93.69	**90.79**	**90.13**	[92.41, 72.73]
DenseASPP - MobileNetV2	39.80	63.35	50.00	39.79	79.59	46.30	44.32	[79.59, 00.00]
DenseASPP - Res.50	79.85	86.25	10.65	8.47	92.27	87.53	88.39	[90.67, 69.04]
DenseASPP - Res.101	79.97	86.52	11.66	9.28	92.52	88.57	88.45	[91.03, 68.91]
DenseASPP - Res.152	79.96	86.52	11.73	9.34	92.52	88.62	88.44	[91.04, 68.88]
Encoder-Decoder	80.69	87.08	11.55	9.19	92.89	89.41	88.92	[91.48, 69.90]
Encoder-Decoder - Incep.V4	80.49	86.98	11.96	9.52	92.85	89.58	88.78	[91.46, 69.52]
Encoder-Decoder-Skip	81.70	87.79	10.98	8.74	93.32	90.16	89.57	[91.98, 71.41]
Encoder-Decoder-Skip - Incep.V4	81.80	87.86	10.94	8.71	93.37	90.25	89.64	[92.04, 71.56]
FC-DenseNet56	82.02	88.20	12.04	9.59	93.66	91.96	89.76	[**92.45**, 71.59]
FC-DenseNet56 - Incep.V4	82.41	88.35	11.03	8.78	93.69	91.21	90.02	[92.44, 72.37]
FC-DenseNet67	81.87	87.92	10.98	8.74	93.41	90.40	89.68	[92.10, 71.65]
FC-DenseNet67 - Incep.V4	**82.65**	**88.42**	10.24	8.15	93.69	90.64	90.19	[92.41, **72.89**]
FC-DenseNet103	82.36	88.28	10.82	8.61	93.63	90.89	89.99	[92.36, 72.36]
FC-DenseNet103 - Incep.V4	82.57	88.36	**10.21**	**8.12**	93.65	90.51	90.14	[92.35, 72.80]
FRRN-A	81.81	87.93	11.34	9.02	93.44	90.73	89.64	[92.15, 71.48]
FRRN-A - Incep.V4	81.81	87.93	11.37	9.05	93.44	90.77	89.64	[92.15, 71.47]
FRRN-B	81.72	87.82	11.07	8.81	93.35	90.29	89.59	[92.02, 71.42]
FRRN-B - Incep.V4	81.84	87.90	11.01	8.76	93.40	90.39	89.66	[92.08, 71.59]
GCN - Res.50	81.91	87.95	10.95	8.71	93.43	90.42	89.71	[92.11, 71.71]
GCN - Res.101	81.57	87.70	11.08	8.82	93.27	90.10	89.49	[91.93, 71.21]
GCN - Res.152	81.55	87.74	11.43	9.10	93.32	90.47	89.47	[92.00, 71.09]
InternImage	**82.75**	**88.46**	**10.15**	**8.05**	**93.75**	90.75	**90.35**	[**92.45**, **73.04**]
MobileUNet	79.08	85.90	12.32	9.80	92.14	88.03	87.85	[90.61, 67.55]
MobileUNet-Skip	81.17	87.46	11.59	9.22	93.15	90.12	89.22	[91.80, 70.54]
MobileUNet-Skip - Incep.V4	81.35	87.61	11.64	9.27	93.25	90.43	89.34	[91.93, 70.77]
PSPNet - Res.50	81.05	87.36	11.52	9.17	93.07	89.87	89.15	[91.71, 70.40]
PSPNet - Res.101	81.08	87.34	11.25	8.95	93.04	89.61	89.17	[91.65, 70.50]
PSPNet - Res.152	81.12	87.43	11.65	9.27	93.13	90.12	89.19	[91.78, 70.45]
RefineNet - Res.50	39.80	63.35	50.00	39.79	79.59	89.80	44.32	[79.59, 00.00]
RefineNet - Res.101	81.54	87.64	10.85	8.63	93.22	89.82	89.47	[91.85, 71.23]
RefineNet - Res.152	81.46	87.59	10.92	8.69	93.19	89.79	89.42	[91.82, 71.10]
BDRAR	81.84	87.91	10.99	8.77	93.44	90.28	89.69	[92.08, 71.60]
FDRNet	82.46	88.41	11.09	8.83	**93.74**	**91.26**	90.08	[**92.42**, 72.49]

Table 3. Benchmark on Airborne-Shadow dataset with the merged, heavy and light shadow classes. B and M denote Background and Merged classes. In blue and red are the best and second best results. IoU is in %.

Method	IoU	f.w. IoU	BER	f.w. BER	Recall	Precision	Dice	Class IoU [B,M]
AdapNet	79.01	84.25	12.90	9.76	91.29	88.83	87.91	[89.22, 68.81]
AdapNet - Incep.V4	79.02	84.26	12.91	9.77	91.30	88.85	87.91	[89.23, 68.80]
BiSeNet - Res.50	79.54	84.65	12.50	9.46	91.53	89.12	88.26	[89.49, 69.59]
BiSeNet - Res.101	79.71	84.77	12.35	9.34	91.60	89.18	88.37	[89.57, 69.85]
BiSeNet - Res.152	80.02	85.03	12.24	9.26	91.77	89.49	88.58	[89.78, 70.27]
DeepLabv3 - Res.50	76.79	82.54	14.39	10.89	90.22	87.34	86.42	[87.98, 65.60]
DeepLabv3 - Res.101	76.71	82.45	14.35	10.86	90.16	87.16	86.36	[87.89, 65.52]
DeepLabv3 - Res.152	77.20	82.88	14.24	10.77	90.45	87.76	86.69	[88.26, 66.13]
DeepLabV3 - Incep.V4	37.83	57.25	50.00	37.83	75.66	87.83	43.07	[75.66, 00.00]
DeepLabv3+ - Res.50	79.96	84.91	11.91	9.01	91.66	89.03	88.54	[89.61, 70.30]
DeepLabv3+ - Res.101	79.99	84.94	11.94	9.03	91.68	89.09	88.56	[89.64, 70.33]
DeepLabv3+ - Res.152	80.30	85.21	11.89	8.99	91.86	89.47	88.76	[89.87, 70.72]
DeepLabv3+ - Xcep.65	81.84	86.38	10.75	8.13	92.55	90.30	89.76	[90.67, 73.02]
DenseASPP - MobileNetV2	37.84	57.14	50.05	37.87	75.45	47.01	43.25	[75.44, 00.25]
DenseASPP - Res.50	78.68	83.95	12.89	9.75	91.08	88.34	87.70	[88.95, 68.41]
DenseASPP - Res.101	78.32	83.63	12.93	9.78	90.86	87.88	87.46	[88.66, 67.97]
DenseASPP - Res.152	78.77	83.91	12.28	9.29	91.00	87.83	87.77	[88.79, 68.75]
Encoder-Decoder	79.72	84.76	12.22	9.25	91.58	89.05	88.39	[89.54, 69.91]
Encoder-Decoder - Incep.V4	79.77	84.81	12.22	9.24	91.61	89.12	88.42	[89.58, 69.97]
Encoder-Decoder-Skip	80.37	85.25	11.78	8.91	91.88	89.45	88.81	[89.89, 70.85]
Encoder-Decoder-Skip - Incep.V4	81.12	85.82	11.21	8.48	92.22	89.85	89.30	[90.28, 71.97]
FC-DenseNet56	**81.88**	86.40	10.75	8.13	92.57	90.34	89.78	[90.70, **73.06**]
FC-DenseNet56 - Incep.V4	81.82	86.37	10.87	8.22	92.56	90.40	89.74	[90.69, 72.95]
FC-DenseNet67	81.54	86.22	11.43	8.64	92.50	90.68	89.55	[90.66, 72.42]
FC-DenseNet67 - Incep.V4	81.69	86.28	11.01	8.33	92.51	90.39	89.66	[90.64, 72.74]
FC-DenseNet103	81.51	86.18	11.37	8.60	92.47	90.56	89.53	[90.62, 72.40]
FC-DenseNet103 - Incep.V4	81.43	86.11	11.32	8.56	92.42	90.39	89.49	[90.55, 72.32]
FRRN-A	80.85	85.66	11.65	8.82	92.14	89.97	89.12	[90.22, 71.49]
FRRN-A - Incep.V4	80.92	85.68	11.47	8.68	92.15	89.85	89.16	[90.21, 71.63]
FRRN-B	80.78	85.63	11.85	8.97	92.14	90.11	89.07	[90.23, 71.33]
FRRN-B - Incep.V4	80.86	85.68	11.74	8.88	92.17	90.09	89.12	[90.25, 71.47]
GCN - Res.50	81.03	85.78	11.46	8.67	92.21	90.00	89.23	[90.29, 71.77]
GCN - Res.101	80.84	85.69	11.89	8.99	92.18	90.24	89.10	[90.29, 71.40]
GCN - Res.152	80.44	85.31	11.74	8.88	91.92	89.50	88.85	[89.93, 70.95]
InternImage	**82.51**	**87.34**	**9.95**	**7.53**	**92.78**	**90.49**	**90.02**	[**91.00**, **74.03**]
MobileUNet	77.77	83.28	13.67	10.34	90.68	87.93	87.08	[88.51, 67.03]
MobileUNet-Skip	80.13	85.10	12.09	9.15	91.80	89.47	88.65	[89.81, 70.45]
MobileUNet-Skip - Incep.V4	80.38	85.23	11.55	8.74	91.85	89.22	88.82	[89.82, 70.95]
PSPNet - Res.50	80.10	84.99	11.65	8.81	91.69	88.94	88.64	[89.63, 70.57]
PSPNet - Res.101	80.24	85.13	11.75	8.89	91.80	89.24	88.72	[89.78, 70.70]
PSPNet - Res.152	80.17	85.07	11.76	8.90	91.76	89.15	88.68	[89.72, 70.61]
RefineNet - Res.50	80.10	85.11	12.32	9.32	91.83	89.70	88.62	[89.87, 70.33]
RefineNet - Res.101	80.49	85.31	11.49	8.69	91.90	89.29	88.89	[89.88, 71.10]
RefineNet - Res.152	80.51	85.28	11.20	8.48	91.85	89.03	88.91	[89.80, 71.22]
BDRAR	80.63	84.98	12.02	8.98	92.08	89.85	88.91	[90.05, 71.21]
FDRNet	81.86	**86.45**	**10.95**	**8.32**	**92.65**	**90.51**	**90.14**	[**90.74**, 72.98]

References

[1] C. Liu *et al.*, "Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation," *arXiv preprint arXiv:1901.02985*, 2019.

[2] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Y. Deepmotion, "DenseASPP for Semantic Segmentation in Street Scenes," in *CVPR*, Salt Lake City, 2018, pp. 3684–3692.

[3] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 325–341.

[4] H. Zhang *et al.*, "Context encoding for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] Y. Yuan and J. Wang, "Ocnet: Object context network for scene parsing," *arXiv preprint arXiv:1809.00916*, 2018.

[6] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," in *CVPR*, Honolulu, 2017.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic Image Segmentation With Deep Convolutional Nets And Fully Connected CRFs," *arXiv preprint arXiv:1412.7062*, 2014.

[8] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3309–3318. DOI: `10.1109/CVPR.2017.353`.

[9] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[10] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[12] A. Valada, J. Vertens, A. Dhall, and W. Burgard, "Adapnet: Adaptive semantic segmentation in adverse environmental conditions," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 4644–4651. DOI: `10.1109/ICRA.2017.7989540`.

[13] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 11–19.

[14] L. Zhu *et al.*, "Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 121–136.

[15] L. Zhu, K. Xu, Z. Ke, and R. W. Lau, “Mitigating intensity bias in shadow detection via feature decomposition and reweighting,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4702–4711.

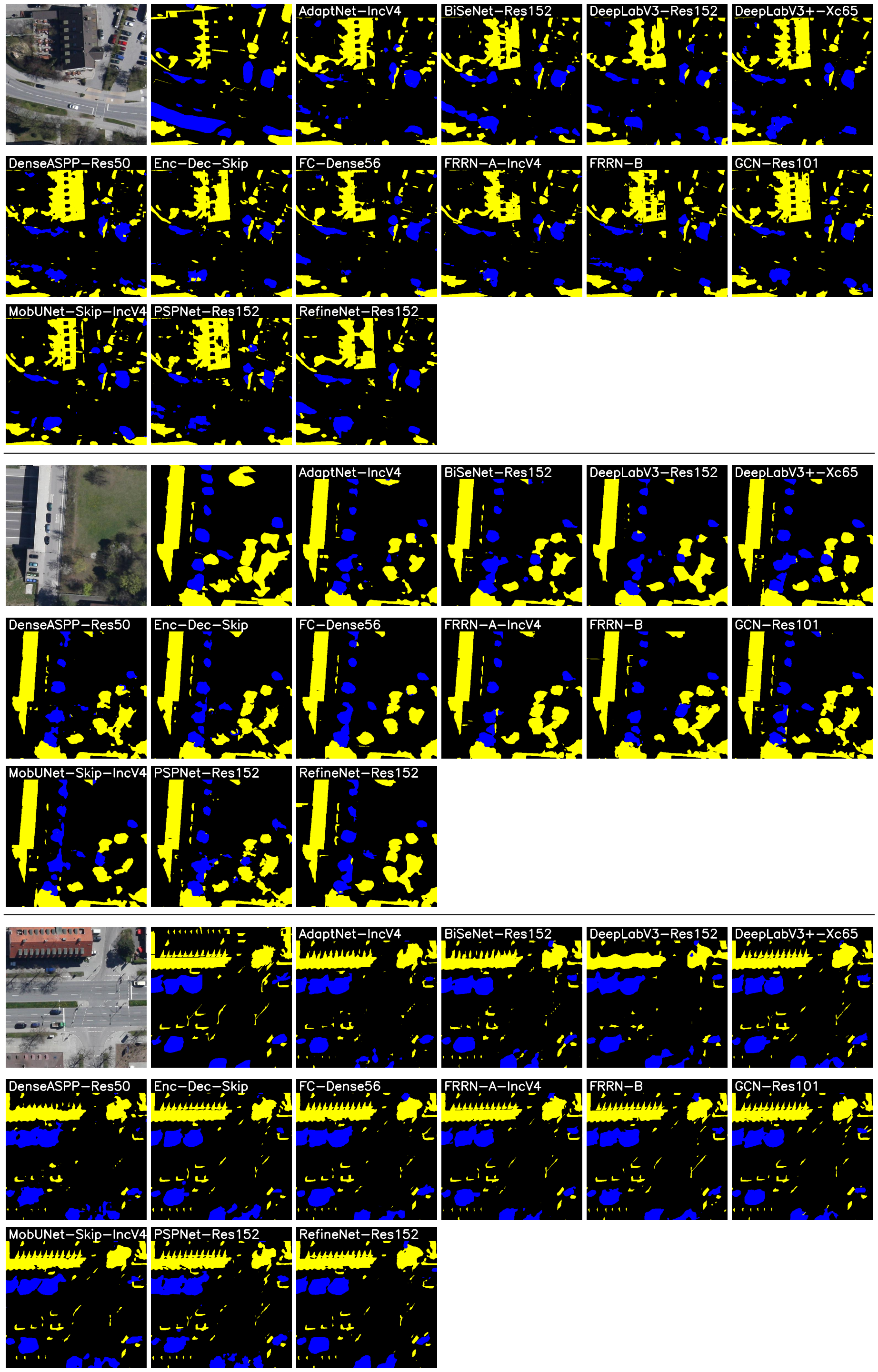


Fig. 1. Three example shadow detections for the 🟨 heavy and 🟦 light shadow classes. Each example includes the original and ground truth images, as well as the results obtained using the best performing variant of the main segmentation networks.

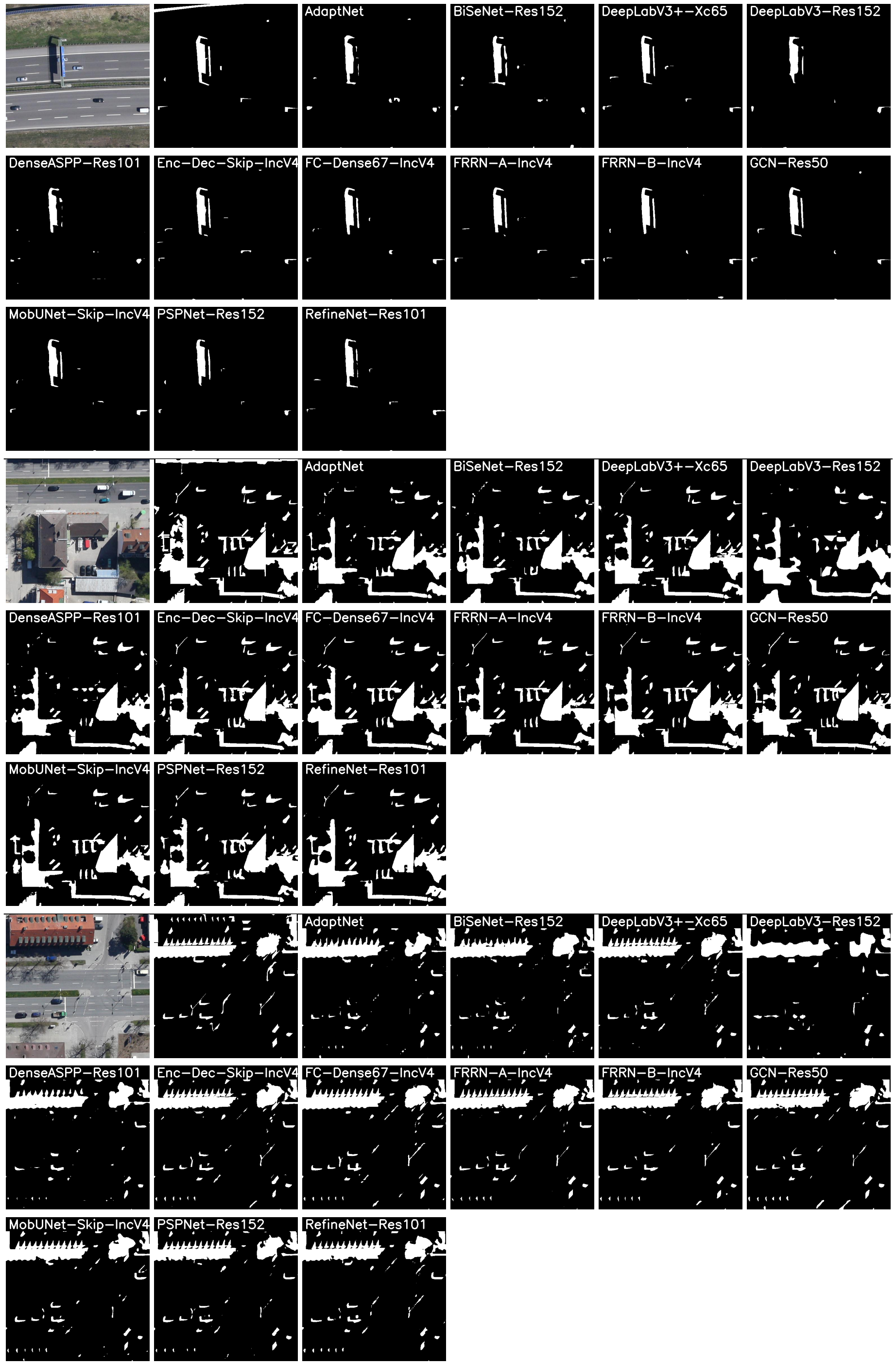


Fig. 2. Three example shadow detections for the heavy shadow class. Each example includes the original and ground truth images, as well as the results obtained using the best performing variant of the main segmentation networks.

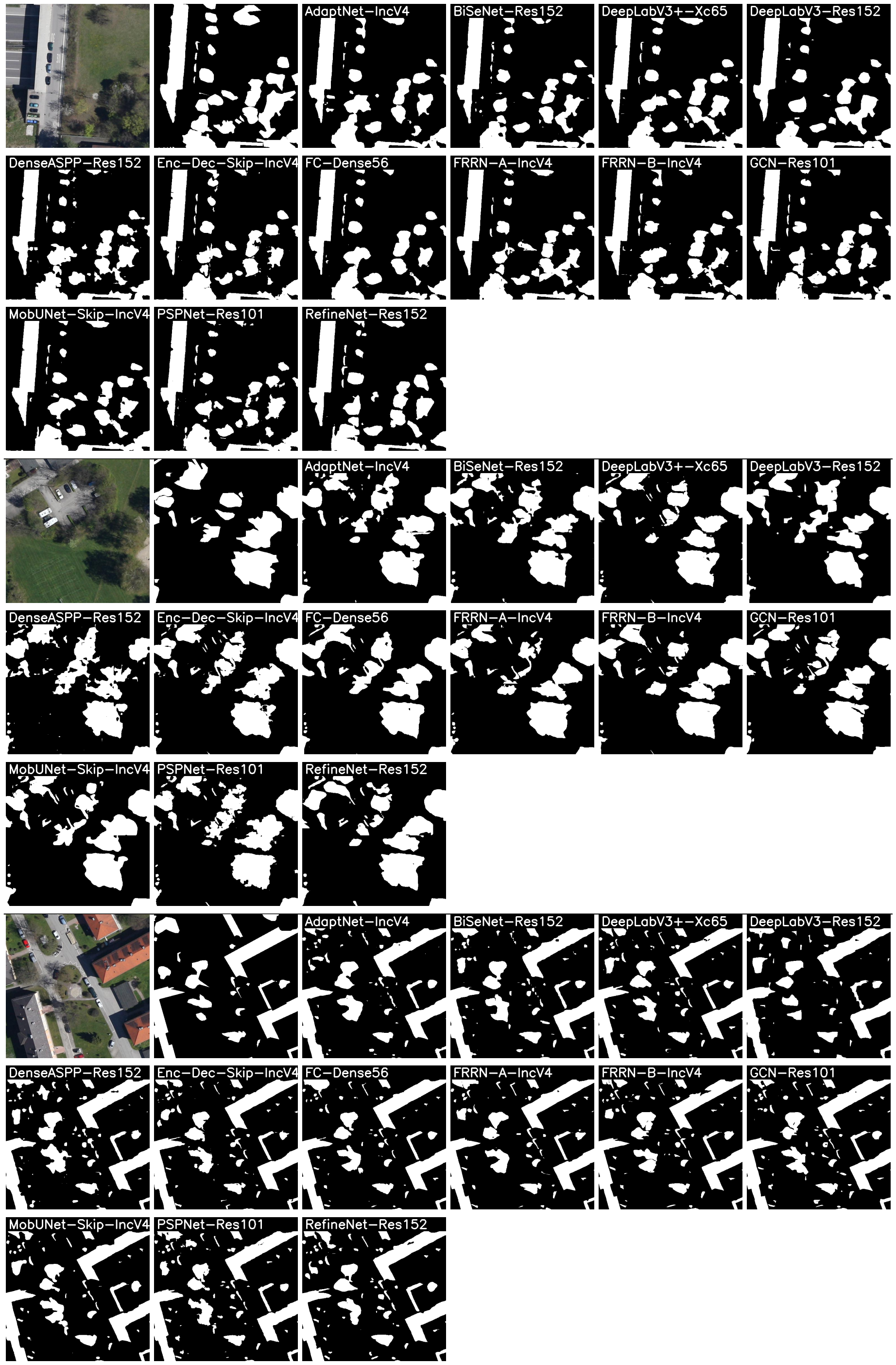


Fig. 3. Three example shadow detections for the merged heavy and light shadow classes. Each example includes the original and ground truth images, as well as the results obtained using the best performing variant of the main segmentation networks.

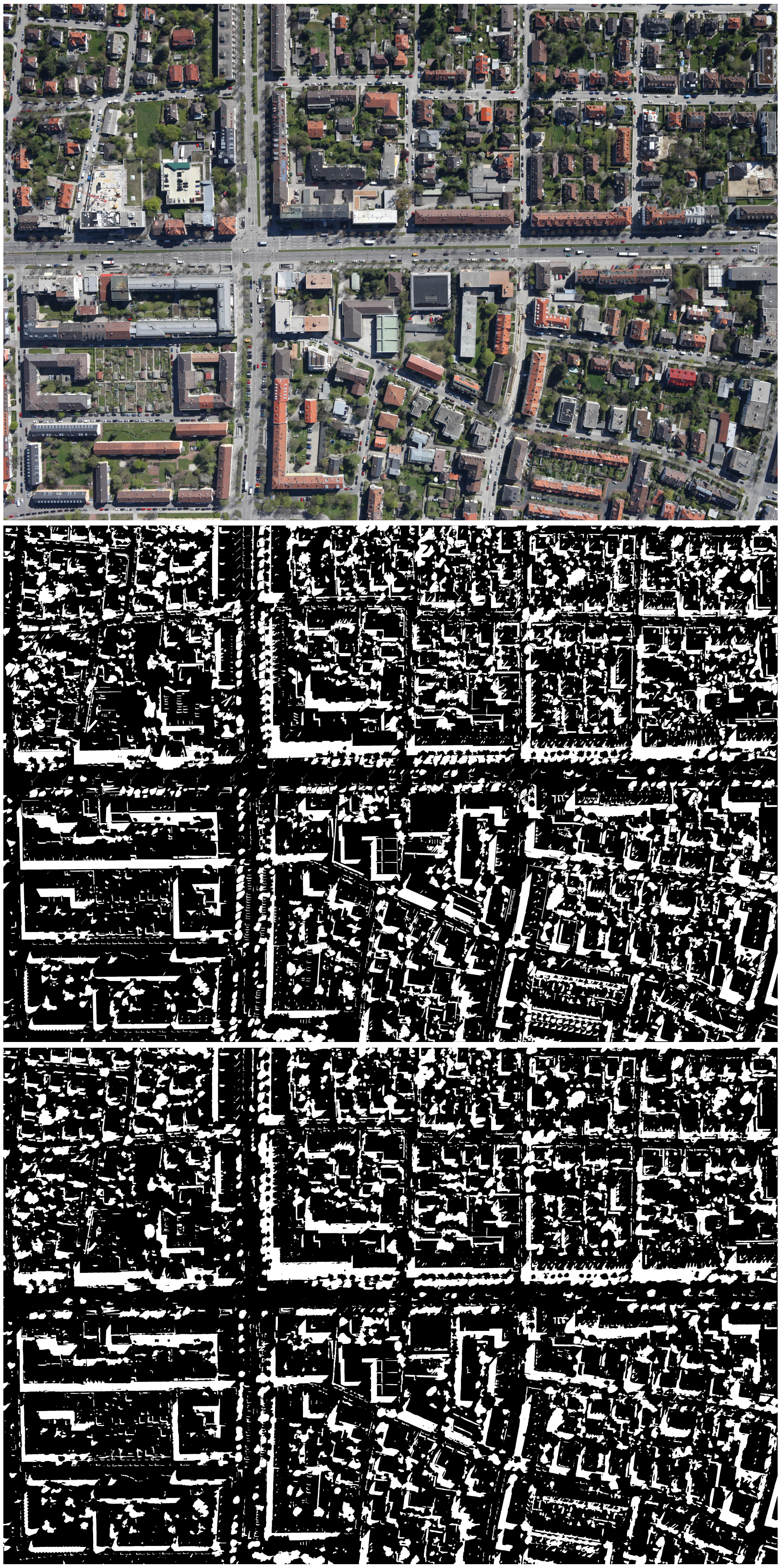


Fig. 4. Performance of the trained InternImage algorithm on ASD dataset (merged) on one sample aerial image from the test set. Top: image, middle: label, bottom: output.

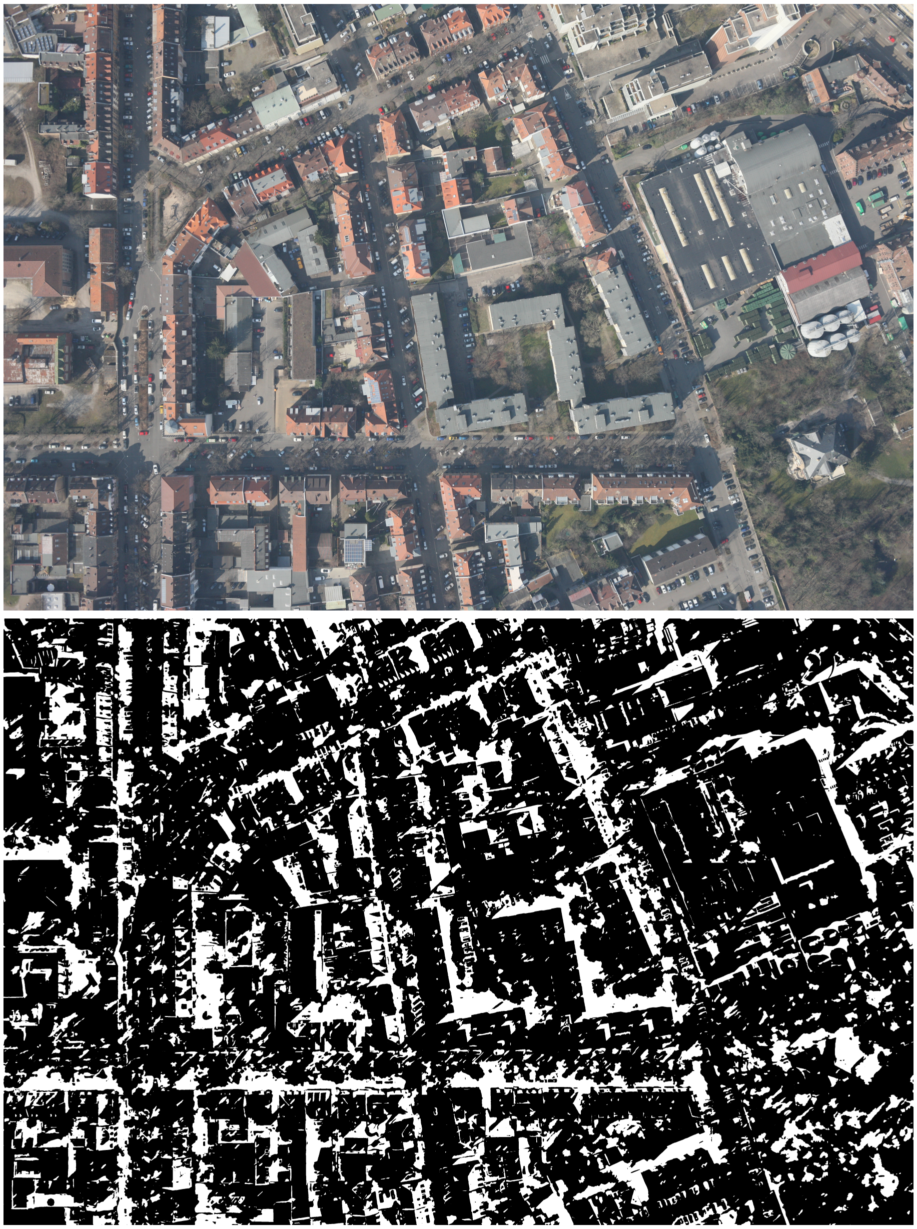


Fig. 5. Generalization capability of the trained InternImage algorithm on ASD dataset (merged) on aerial images over the city of Karlsruhe with different resolution, GSD, acquired over a different city at a different time and altitude.

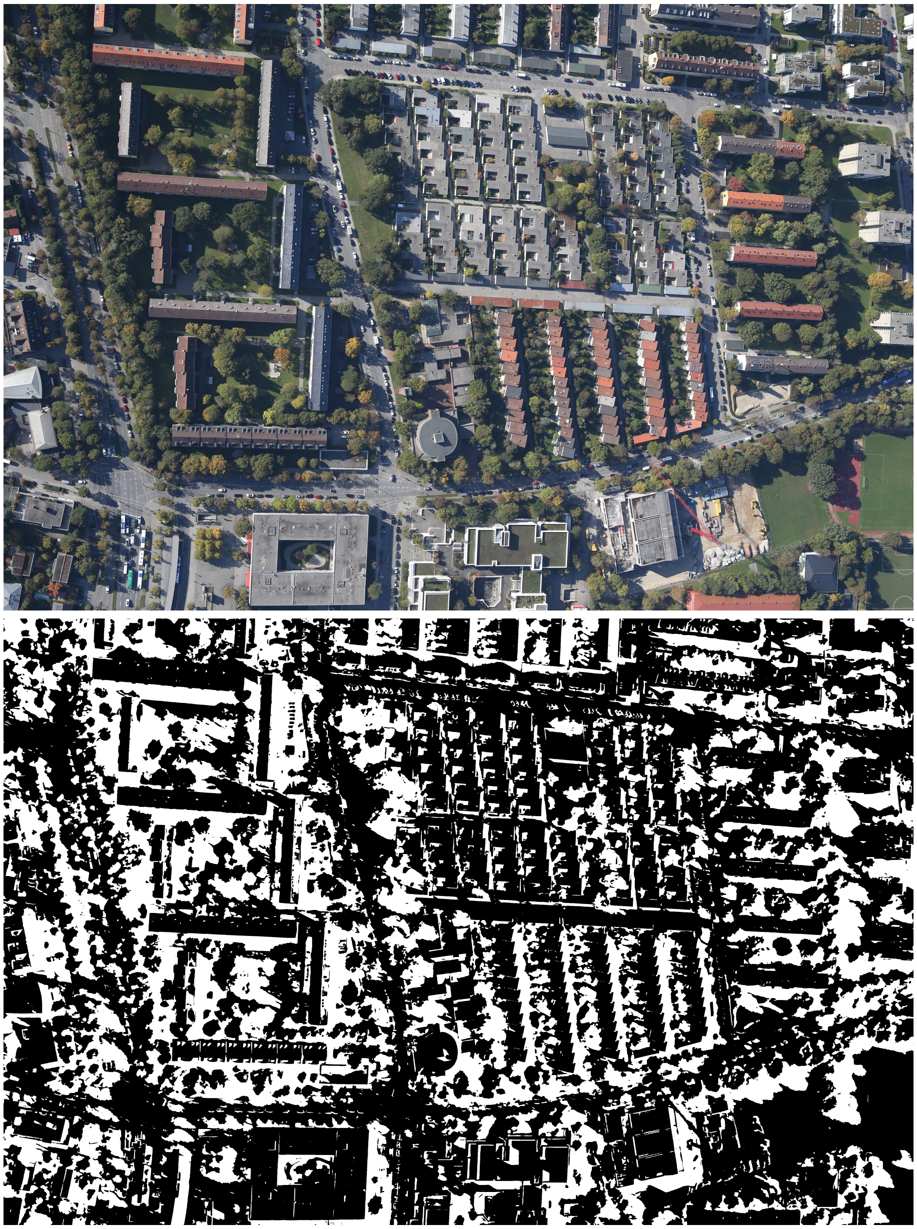


Fig. 6. Generalization capability of the trained InternImage algorithm on ASD dataset (strong) on aerial images over the city of Karlsruhe with different resolution, GSD, acquired over a different city at a different time and altitude.

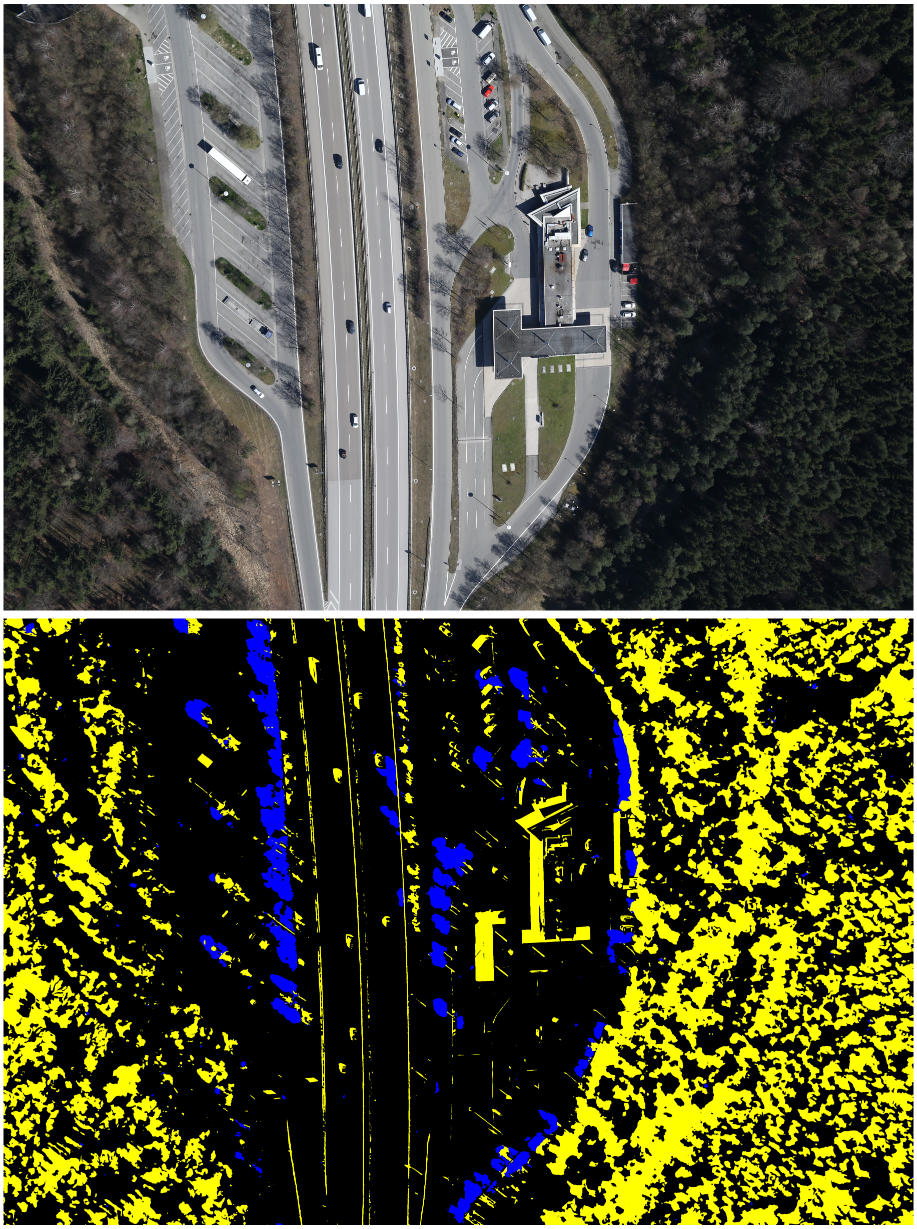


Fig. 7. Generalization capability of the trained InternImage algorithm on ASD dataset (two-classes) on aerial images over the city of Wolfratshausen with different resolution, GSD, acquired over a different city at a different time and altitude.