Semantic segmentation of water bodies in very high-resolution satellite and aerial images

Marc Wieland [1] *, Sandro Martinis [1], Ralph Kiefl [1], Veronika Gstaiger [2]

[1] German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Oberpfaffenhofen, D-82234 Wessling, Germany

[2] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, D-82234 Wessling, Germany

* Correspondence: marc.wieland@dlr.de

Abstract

This study evaluates the performance of convolutional neural networks for semantic segmentation of water bodies in very high-resolution satellite and aerial images from multiple sensors with particular focus on flood emergency response applications. Different model architectures (U-Net and DeepLab-V3+) are combined with encoder backbones (MobileNet-V3, ResNet-50 and EfficientNet-B4) and tested for their ability to delineate inundated areas under varying environmental conditions and data availability scenarios. An unprecedented reference dataset of 1,120 globally sampled images with quality checked binary water masks is introduced and used to train, validate and test the models for water body segmentation. Furthermore, independent test datasets are developed to test the generalization ability of the trained models across regions, sensors (IKONOS, GeoEye-1, WorldView-2, WorldView-3 and four different airborne camera systems) and tasks (normal water and flood water segmentation). Results indicate that across all tested scenarios a U-Net model with Mobilenet-V3 backbone pre-trained on ImageNet performs best. While using R-G-B image bands performs well, adding the near infrared band (if available) slightly improves prediction results. Similarly, adding slope information from an independent digital elevation model increases accuracies. Train-time augmentation and contrast enhancement could improve transferability across sensors and in particular between satellite and aerial images. Moreover, adding noisy training data from freely available online resources could further improve performance with minimal annotation effort.

Keywords

Convolutional Neural Networks; Semantic segmentation; Water; Emergency response; Rapid mapping

1 Introduction

Very high-resolution optical satellite and aerial images (in the following defined as having a ground sampling distance < 4 m and spectral bands in the visible and near infrared part of the electromagnetic spectrum) are frequently requested during rapid mapping activations to gain situational awareness about large-scale flood disasters (Martinis et al., 2017; Voigt et al., 2016). Near-real time information about (flood) water extent and affected infrastructure are amongst the most critical components that can be delivered by this kind of images to support prioritization of response actions. The tasking capability of some of the satellite imaging platforms and the fact that airborne sensors can acquire images at any (day) time and below cloud-coverage makes them particularly favourable for time-critical emergency response applications. These can provide a valuable supplement to systematically acquiring satellites with Synthetic Aperture Radar (SAR) or multi-spectral sensors, like Sentinel-1, Sentinel-2 or Landsat. These satellites acquire images at lower spatial resolution (~10-30 m) and on a regular schedule usually every couple of days depending on the area of interest. Hence, very high-resolution optical satellite and aerial images help to fill temporal acquisition gaps and provide more spatial details of an ongoing disaster situation.

To assure that geo-information products have the highest possible spatial and temporal resolutions and information content, it is crucial that algorithms used for remote sensing image-based emergency mapping are able to simultaneously use data from a variety of acquisition platforms and sensors. Automated water segmentation in this kind of images, however, is a challenging task because of significant variations in spectral reflectance characteristics, size and shape of water bodies. Compared to normal water, flood water for example may be characterized by higher contents of sediment and debris. Vegetation, infrastructure, boats or other vehicles may further interact with the water surface or sub-surface, and also lighting conditions (e.g., sun reflectance on water surface or shadows by buildings or vegetation) may cause high intra and inter water body variability. The fractal geometry of the land-water border further increases the complexity of the segmentation task, since the definition of a hard class boundary may be subjective depending on the image resolution. Moreover, bottom features may be visible through the water column, especially in clear shallow water bodies, which further increases the complexity of automated water segmentation. As a result, common operational procedures for remote sensing image-based emergency flood mapping rely largely on manual image interpretation.

Traditional methods for the segmentation of water bodies mainly focus on thresholding water indices, such as the Normalized Difference Water Index (NDWI) (McFeeters, 1996). These methods exclusively

rely on spectral information and require at least one or more image bands in the near- or short-wave infrared, which negatively impacts their transferability to other sensors. When applied to very high-resolution images such methods, moreover, show a lack of generalization ability, because the more spatial details are visible, the higher the spectral variability within the water class becomes. To overcome some of the limitations of simple thresholding methods, supervised machine learning methods such as Support Vector Machine (SVM) that learn a decision boundary between classes from labelled training samples and based on hand-crafted features have been applied for water mapping (Ireland et al., 2015). In recent years, deep learning approaches have shown superior performance on various image analysis tasks across disciplines. Convolutional Neural Networks (CNNs) are widely used for semantic segmentation because of their ability to implement nonlinear decision functions and to learn features directly from raw images by combining convolutional and pooling layers (Ball et al., 2017).

Current semantic segmentation algorithms, such as U-Net (Ronneberger et al., 2015), PSP-Net (Zhao et al., 2017) or DeepLab-V3+ (Chen et al., 2017) follow an encoder-decoder architecture, which consists of a down-sampling path to learn dense image features and capture context (encoder) as well as a symmetric up-sampling path that learns an optimal interpolation of the features to recover the original image resolution (decoder). U-Net uses skip connections for feature propagation, whereas PSP-Net and DeepLab-V3+ use spatial pyramid pooling to get contextual features at multiple scales. Dilated convolutions are utilized by the latter to capture more contextual information without increasing the number of trainable parameters. Different encoder networks have been proposed in literature, amongst which the most prominent are VGG-16 (Simonyan and Zisserman, 2015) and ResNet (He et al., 2015a). These networks are originally intended for image classification, but can be used as feature extractors for semantic segmentation. Accuracy improvements between different network versions and architectures have largely been achieved by scaling in depth (adding more layers), width (increasing spatial size of layers) or resolution (adding more channels per layer). This has mostly been done by manual tuning which, given the large number of parameters involved, makes finding an optimal performance point difficult and resource consuming. Tan and Le (2019) showed that scaling all three dimensions uniformly with a fixed set of scaling factors can improve model performance while keeping resource consumption minimal (Tan and Le, 2019). They used neural architecture search to develop a novel baseline model and apply compound scaling to optimize network depth, width, and resolution simultaneously. The main building blocks of their EfficientNet models are mobile inverted bottleneck blocks (MBConv) with squeeze-and-excitation optimization (Hu et al., 2020). EfficientNet models outperform other CNNs on ImageNet (Deng et al., 2009) and several transfer learning benchmark datasets while consistently being

smaller and thus more efficient than previous models. Howard et al. (2019) also use hardware-aware neural architecture search to develop Mobilenet-V3, which is particularly light-weight and optimized to run on limited hardware resources like mobile phone CPUs. Despite their promising characteristics, Mobilenet-V3 and EfficientNet encoders are not (yet) widely used in remote sensing studies. Other encoders that have been tested in remote sensing applications include but are not limited to ResNext (Xie et al., 2017), PSPNet (Yuan et al., 2022) and ShuffleNet (Gomes et al., 2021). A new branch of networks that gain popularity recently are transformer networks. Cao et al. (2021) for example report superior performance of their U-Net-like transformer for medical image segmentation compared to fully convolutional networks. First studies that apply transformer networks on remote sensing images also show promising results (Ding et al., 2022; Gu et al., 2022; Xu et al., 2021).

Since reference data in remote sensing is expensive to generate and satellite images are often restricted by conservative licenses that prevent re-distribution of purchased imagery, only few publicly available benchmark datasets exist for semantic segmentation of very high-resolution satellite and aerial images, namely ISPRS (Rottensteiner et al., 2013), Dstl (Iglovikov et al., 2017) and Deepglobe (Demir et al., 2018). Of these only the Dstl Kaggle Challenge and Deepglobe Land Cover Classification Challenge datasets contain water classes, whereas only the latter covers samples outside of urban areas. Castillo-Navarro et al. (2021) created the MiniFrance dataset that is specifically designed for semi-supervised segmentation of land-use. It contains over 2,000 labelled and unlabelled aerial images across France, with the water class covering approximately 1 % of the dataset. Due to their general focus on land-cover / land-use classification, the available datasets can cover only limited variations in spectral reflectance characteristics, size and shape of water bodies and none of them considers a flood water class. Recently, several datasets for natural disaster applications have been released (Bonafilia et al., 2020; Gupta et al., 2019). Hänsch et al. (2022) present the SpaceNet 8 dataset for the detection of flooded roads and buildings. Rahnemoonfar et al. (2021) introduce the FloodNet dataset, which is composed of partially annotated UAV images after Hurricane Harvey and aims at post flood scene understanding. The semantic segmentation labels include amongst others a water class and several classes of flooded infrastructure. The dataset is, however, from a single location and covers on a relatively small amount of labelled data. The available disaster related datasets are often limited in geographical coverage and / or have a strong focus on infrastructure.

To artificially increase the number of training samples in limited data scenarios, data augmentation is increasingly being used in remote sensing. Also transfer learning (Pan and Yang, 2010) is an approach to ease the problem of scarce training data. CNNs trained on large-scale natural image datasets (e.g.,

ImageNet) have successfully been transferred to segment remote sensing images, either by means of fine-tuning the network with data from the target domain or by using the pre-trained network directly as feature extractor. However, in particular deep models with a large number of trainable parameters may require large amounts of labelled samples of the target domain to avoid over-fitting. Semi-supervised learning, which jointly uses labelled and unlabelled data, provides another promising approach to improve segmentation when training data is limited (Li et al., 2019). Few studies also experiment with training on noisy labels. Mnih and Hinton (2012) and Kaiser et al. (2017) report that satisfying performance can be obtained with significantly less manual annotation effort, by exploiting noisy large-scale training data from independent maps or data from crowd-sourcing platforms like OpenStreetMap (Haklay, 2010).

While most deep learning studies that use very high-resolution optical satellite and aerial images focus on building segmentation (Neupane et al., 2021), road delineation (Mattyus et al., 2015) or object detection (Azimi et al., 2021), relatively few studies exist for water body segmentation (Huang et al., 2018). Yang et al. (2020) propose a model based on mask R-CNN to detect and segment water bodies in GaoFen-2 images. They train their models on Google Earth image tiles and GaoFen-2 images and test on a hold-out dataset of GaoFen-2 images. Their results indicate superior performance of a ResNet-50-based model compared to a more complex ResNet-100-based model. Guo et al. (2020) introduce a multi-scale water extraction CNN for GaoFen-1 images and compare their results to U-Net and DeepLab-V3+ models. Chen et al. (2018) apply a self-adaptive pooling CNN on clusters of image super-pixels to extract urban water bodies in GaoFen-2 and Ziyuan-3 images. They train and test their method on images of Beijing, Chengdu and Tianjin and report superior accuracy compared to SVM and water index thresholding. Gebrehiwot et al. (2019) investigate the potential of a VGG-based fully convolutional network to extract flooded areas from Unmanned Aerial Vehicle (UAV) images acquired over three study areas in North Carolina, United States after Hurricane Matthew. Their results support conclusions of studies from other disciplines that adapting pre-trained models and fine-tuning them can lead to superior performance, especially when domain-specific training data are scarce. Duan and Hu (2020) compare a multi-scale refinement network with U-Net, SegNet and DeepLab-V3+ networks for water segmentation in WorldView-3 and GaoFen-2 satellite images. They report superior performance of the refinement network and DeepLab-V3+ compared to U-Net and SegNet. Feng et al. (2019) propose a modified U-Net architecture and combine it with a super-pixel segmentation method to refine details of the land-water borders. They test their method on four images acquired by WorldView-2 and GaoFen-2 satellites.

Existing studies may achieve encouraging results for experimental setups, but largely focus on limited

geographical coverage, single sensors and propose complex solutions that may not scale well to time-critical emergency response applications. Further research is required to train water segmentation models that are applicable to different satellite sensors and images across varying atmospheric conditions and scene properties. The objective of this study is to evaluate the performance of CNNs for semantic segmentation of surface water bodies in very high-resolution satellite and aerial images. We render this as a binary segmentation task with classes "land" and "water". Depending on whether a satellite image has been acquired during normal hydrological conditions or during a flood event, we distinguish between scenes that primarily depict "normal water" or "flood water".

Most promising architectures (U-Net and DeepLab-V3+) are combined with different encoders (MobileNet-V3, ResNet-50 and EfficientNet-B4) and tested for their ability to delineate inundated areas under varying environmental conditions and data availability scenarios. For each combination of architecture and encoder we test different weight initializations and evaluate the performance of the trained models to (I) segment normal water in images of the same sensors, (II) segment normal water in images of other sensors, and (III) segment flood water in images of other sensors. We introduce an unprecedented reference dataset of 1,120 images with quality checked binary water masks. We also develop an independent test dataset for four flood events to evaluate the generalization ability of the trained models across regions, sensors and tasks. Compared to other studies in this direction, we consider sensors from multiple sensors and acquisition platforms (IKONOS, GeoEye-1, WorldView-2, WorldView-3, four different airborne camera systems and images from a Web Map Service of Mapbox ("Mapbox Satellite," 2021)) and train, validate and test considered networks on a globally sampled dataset that covers variations in environmental and atmospheric conditions as well as spectral reflectance characteristics, size and shape of water bodies.

2 Data

We compile a reference dataset based on 1,120 globally distributed very high-resolution satellite and aerial images to train, test and validate the water segmentation models. For the dataset to be representative for a large variety of climatic, atmospheric, and land-cover conditions, we combine several global environmental layers into a unified stratification layer. Each stratum is aligned to a global grid with approximately 1.5 x 1.5 km spacing, contains water bodies (Pekel et al., 2016) and represents a unique combination of biome (Olson et al., 2001) and pre-dominant land-cover (Buchhorn et al., 2020). Samples are selected by means of a stratified random sampling procedure (Figure 1). Data availability constraints are considered, acquisition times cover different seasons and the maximum cloud-cover percentage for

image acquisitions is set to 5 %. For the purpose of this study, we acquire Internal Data Access (IDA) to archived IKONOS images received by the satellite down-link for Europe and Middle-East between 2000 and 2009 (Schreier et al., 2008). For each sample we acquire the respective images, orthorectify them, convert digital numbers to top of canopy reflectance, pan-sharpen them to 0.8 m spatial resolution, stack the Red (R), Green (G), Blue (B) and Near-Infrared (NIR) image bands together and create a 2,048 x 2,048 pixels subset. Thematic masks are delineated into classes "water" and "land" based on semi-automated image analysis. We first apply a simple NDWI thresholding to extract an initial water mask, which we iteratively improve by manual adjustments and quality controls by several experienced operators (Figure 2). Due to the limited geographical coverage of the IDA IKONOS archive, we complement the reference dataset with freely available optical images from a Web Map Service of Mapbox ("Mapbox Satellite," 2021). Images are acquired as geo-referenced R-G-B composites at zoom levels 17 and 18, which approximately equals a ground sampling distance of 0.6 to 1.2 m depending on the latitude of the sample location. Data at these zoom levels are a combination of satellite and aerial images from Maxar and various open national aerial image archives with JPG compression, reduced radiometric resolution and enhanced contrast compared to the original images. We download thematic water masks for each of these samples from OpenStreetMap ("OpenStreetMap," 2021) and iteratively improve them by manual quality control and adjustments by experienced operators. Images and masks are shuffled and divided into training (50 %), validation (25 %) and testing (25 %) datasets, which are then split into non-overlapping tiles with 256 x 256 pixels size. The final dataset covers approximately 90,000 tiles and 2,800 km².

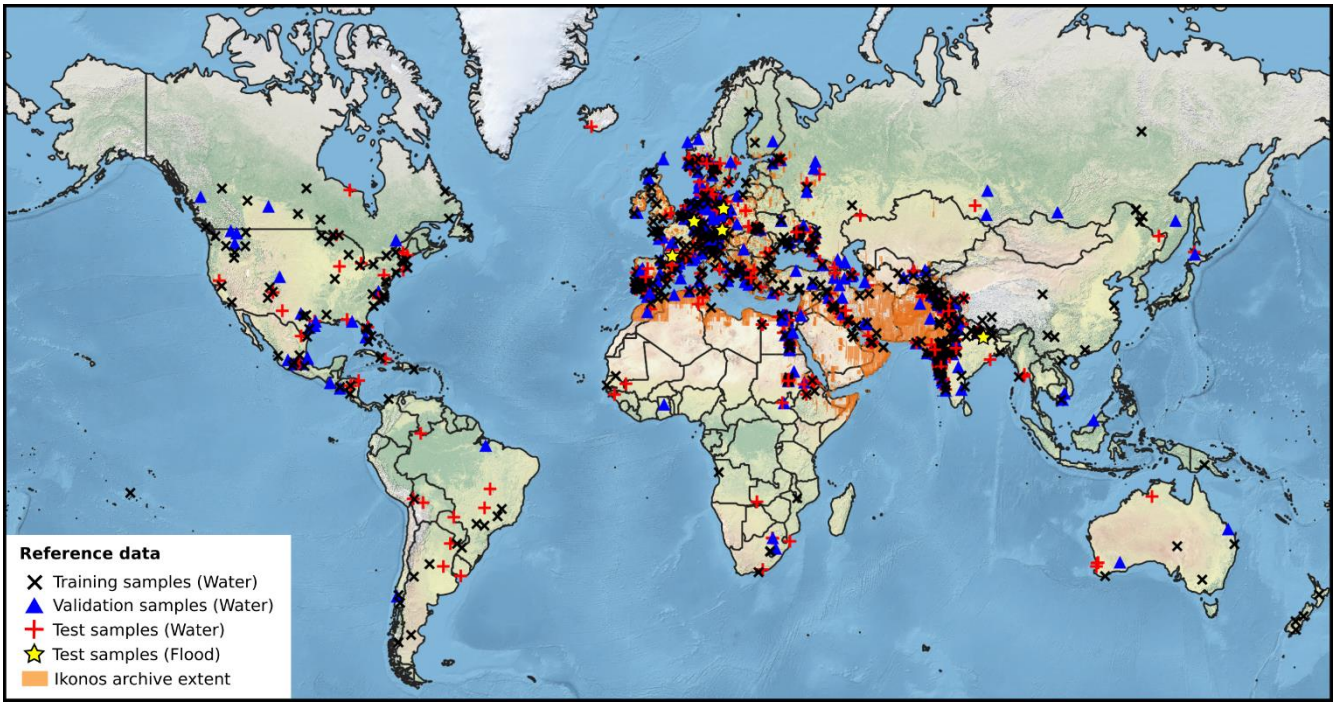

Figure 1: Spatial distribution of the reference data samples used for training, validation and testing of the water segmentation method; Independent test samples for flood water segmentation from multiple sensors (GeoEye-1, WorldView-2, WorldView-3 and four different airborne camera systems).

To independently test the performance of trained models on flood water scenes and to consider additional platforms and sensors, we select six locations of four flood events (Dresden Germany 2013, Landes France 2019, Bihar India 2020, Minden, Brück and Altenahr Germany 2021), for which normal water (pre- or post-event) and flood water (co-event) satellite or aerial images are available. We use normal water images to test transfer of models to other sensors for water segmentation. To test model transfer to other sensors and tasks (flood water segmentation), we use flood images. For each test case we manually delineate water extents by means of a standard rapid mapping workflow and make sure to cross-check results between different experienced operators (Long et al., 2021). Cloud and cloud shadow pixels are manually masked in both reference and predicted masks and excluded from accuracy assessments. Table 1 shows an overview of the available images for each case study. The dataset covers approximately 15,000 tiles (256 x 256 pixels) and 204 km².

Each sample that we consider in this study is further complemented with slope information derived from a Digital Elevation Model (DEM) raster of the freely and globally available Copernicus DEM (Fahrland et al., 2020). The slope raster is up-sampled from its original 30 m resolution to align with the respective satellite and aerial images.

Table 1: Overview of normal water and flood images for case studies with information about acquisition dates,

platform, sensor, spatial resolution and available spectral bands.

Case study		Normal water images			Flood water images		
		Acquisition	**Sensor Resolution Spectral bands**	**Coverage (km²)**	**Acquisition**	**Sensor Resolution Spectral bands**	**Coverage (km²)**
Satellite	**Landes, France** ***Riverine floods***	2019-09-11	GeoEye-1 0.50 m R-G-B-NIR	35	2019-12-19	GeoEye-1 0.50 m R-G-B-NIR	35
	Bihar, India ***Monsoon rains***	2020-02-11	WorldView-2 0.50 m R-G-B-NIR	60	2020-08-24	GeoEye-1 0.50 m R-G-B-NIR	60
	Brück, Germany ***Riverine floods***	2021-09-07	GeoEye-1 0.50 m R-G-B-NIR	4	2021-07-21	Worldview-3 0.50 m R-G-B-NIR	4
Aerial	**Altenahr, Germany** ***Riverine floods***	2021-10-23	Zenmuse H20T 0.07 m R-G-B	1	2021-07-20	DLR MACS 0.20 m R-G-B	1
	Minden, Germany ***Riverine floods***	2019-06-28	DOP20 0.20 m R-G-B	1	2021-07-16	DLR 4K 0.07 m R-G-B	1
	Dresden, Germany ***Riverine floods***	2011-05-31	DOP20 0.20 m R-G-B	1	2013-06-08	DLR 3K 0.25 m R-G-B	1

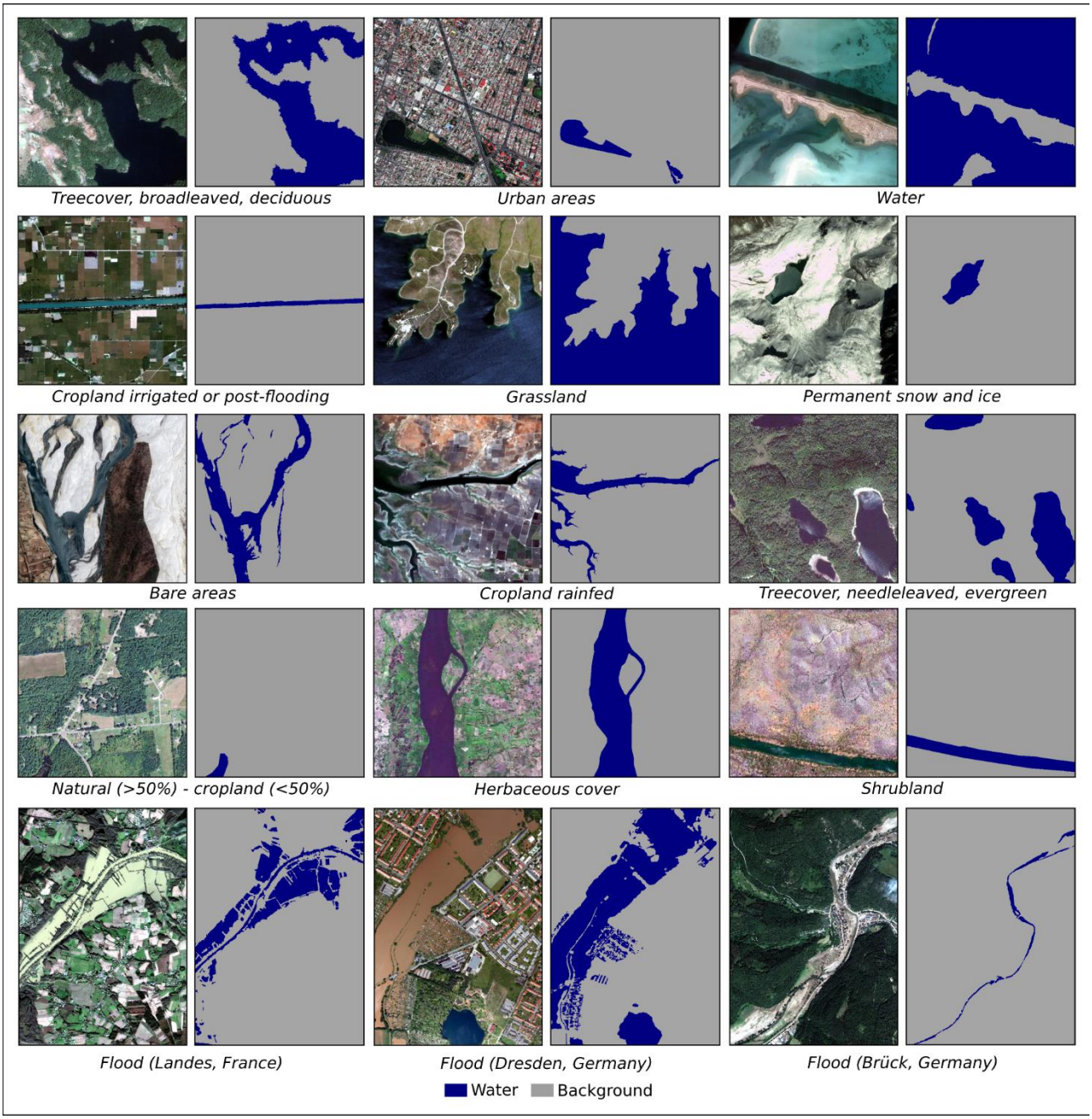

Figure 2: Examples of reference data samples with pairs of image and corresponding water mask for different land-cover settings. Each sample covers an area of approximately 1.3 km². The last row shows flood water images of the independent test dataset. The area covered by the flood water samples varies and is depicted in Table 1.

3 Method

In a series of experiments, we aim to answer the following specific research questions.

1. Baseline model (BM): Which combination of architecture and encoder performs best across varying data availability scenarios?

2. Input bands (IB): Can additional spectral band and independent slope information improve the performance?
3. Data augmentation (DA): Can different data augmentation techniques improve the generalization ability?
4. Pre-training (PT): What influence do different pre-trained weights have on the performance?
5. Transfer model – satellite to aerial (TM): Can a model trained on satellite images be transferred to contrast enhanced aerial images with reduced radiometric resolution? Can noisy training data improve the performance?

In the following we describe the setup for each experiment and general considerations that apply for all experiments.

Baseline model (BM)

Several studies have shown that the U-Net architecture (Ronneberger et al., 2015) is able to deliver state-of-the-art results in water segmentation tasks in high-resolution satellite images while keeping computational complexity low (Wieland and Martinis, 2019). For comparison we choose the DeepLabV3+ architecture (Chen et al., 2017), which extracts information on multiple scales simultaneously with atrous spatial pyramid pooling using diluted convolution kernels with different rates. Depth-wise separable convolution is applied to decrease computation complexity. The hypothesis is, that these architectural differences may allow the model to have a higher level of context-awareness, which may positively impact on the segmentation performance. In combination with these architectures we compare promising encoders, namely MobileNet-V3, ResNet-50 and EfficientNet-B4. We selected these for our experiments since they show a good trade-off between number of model parameters and ImageNet Top-1 accuracy (Tan and Le, 2019).

Input bands (IB)

Water shows low reflectance in the NIR wavelengths as it absorbs more energy, while non-water generally has a higher reflectance. Thus, a high contrast in reflectance values between water and non-water landcover classes is particularly dominant in the NIR spectral band compared to the visible R, G and B bands. To this regard, we test the influence of the NIR spectral band on the water segmentation performance. Additionally, we also consider slope information derived from the freely and globally available Copernicus DEM. For this purpose, we compute slope in percent at the 30 m native spatial resolution of the DEM and resample it to match the respective sample image resolution.

Data augmentation (DA)

Data augmentation provides a way to learn invariance to changes in the augmented domains beyond what is present in the raw training images. Remote sensing image properties are affected by changes in atmospheric conditions, land-use / land-cover, seasonality and other scene and image properties such as sun elevation or radiometric resolution. Therefore, even large remote sensing training datasets may not cover all eventualities that may occur in real-world applications. To this regard, we test the influence of different data augmentation techniques on the segmentation performance. In particular, we apply training- and test-time augmentations. For this experiment the training dataset is augmented with random contrast, brightness, scale and rotation. Factors are randomly applied within predefined ranges to contrast [-0.1, 0.1], brightness [-0.1, 0.1] and scale [0.9, 1.1]. Rotation is performed in steps of 90 degrees. All augmentations are applied with equal probability. During test-time we predict on the original input image and five randomly augmented versions of it. The final prediction is based on the averaged class probabilities.

Pre-training (PT)

Starting from pre-trained weights can improve performance, even if these have been trained on different image types, because low-level features that are being learned in early network layers are similar across image domains (Kaiser et al., 2017). In this experiment we compare the influence of different pre-trained weights on the segmentation performance. We initialize weights either randomly as described in He et al. (2015b) or we use pre-trained weights that were previously learned on ImageNet (Deng et al., 2009) or on a global reference dataset for water segmentation in Sentinel-2 satellite images (Wieland et al., in preparation). Since ImageNet weights only cover R-G-B bands, we initialize the weights of additional image bands (N and slope) randomly. In case of Sentinel-2 weights, we pre-train the model specifically on the same spectral bands and slope information. The assumption here is that while in general pre-training has been reported to boost performance, application and domain specific pre-training on Sentinel-2 satellite images may be superior compared to pre-training on ImageNet (despite the lower 10 m spatial resolution of Sentinel-2).

Transfer model – satellite to aerial (TM)

In this experiment we investigate if a model trained on satellite images can be transferred to aerial images with reduced radiometric resolution and enhanced contrast. We also aim at answering the question, whether noisy training data can improve the generalization ability across images acquired by different sensors and acquisition platforms. Specifically, we test the influence of different training approaches and

compare the performance of models trained solely on the high-quality IKONOS dataset or the noisy Mapbox dataset. We further evaluate the influence of contrast enhancement on the performance. In this case, we apply a percentage linear contrast stretch between the 2^{nd} and 98^{th} percentiles separately for each spectral band. Slope information is not stretched.

General considerations

For all experiments we evaluate the performance of the trained models under consideration of varying data availability across three test scenarios.

- **Test scenario I**: Segment normal water in images of the same sensors. This scenario applies the trained models on images of the same sensors and for the same task that they have been trained for. Test data are the test split of our global reference dataset.

- **Test scenario II**: Segment normal water in images of other sensors. This scenario applies the trained models on images of other sensors but for the same task that they have been trained for. Test data are independent images taken during normal water conditions over locations and by sensors that were not used for training.

- **Test scenario III**: Segment flood water in images of other sensors. This scenario applies the trained models on images of other sensors and for another task that they have been trained for. Test data are independent co-flood images. The target class is flood water, which due to its largely differing spectral characteristics and appearance in the images can be considered as a different task than normal water segmentation. Images are acquired over the same locations as in test scenario II to minimize the effect of the background class on the prediction results and hence isolate the influence of different sensors and water conditions.

We use reference data acquired by satellites with known preprocessing, radiometry and additional NIR spectral band as basis for all experiments. Only the transfer experiment (TM) makes use of the Mapbox and aerial image reference dataset, with reduced radiometric resolution, enhanced contrast and for which a NIR spectral band is not available. The input image feature space in all experiments is standardized to zero mean and unit variance with mean and standard deviation being computed on the training dataset and applied to the validation and testing datasets. The training set is shuffled once between every epoch.

Initial learning rate and weight decay hyperparameters are optimized for each network setup specifically using a tree-structured Parzen estimator with pruning (Akiba et al., 2019) over 15 trials (for five epochs each) while maximizing the *IoU* score on the validation dataset. As search space we define log uniform

distributions (from $1e^{-5}$ to $1e^{-1}$) for the initial learning rate and weight decay. A weighted combination of cross-entropy (L_{CE}) and Lovász loss ($L_{Lovász}$) is used for a direct optimization of the *IoU* score (Rakhlin et al., 2018) during backpropagation (Equation 1).

$$L = (1 - \alpha_{Loss}) * L_{CE} + \alpha_{Loss} * L_{Lovász} \quad \text{Equation 1}$$

With α_{Loss} being the weighting coefficient to weight the two loss functions. Following the results of (Helleis et al., 2022) we set $\alpha_{Loss} = 0.5$. To further account for class imbalance, we apply weights on the positive class which are computed on the class distribution of the training dataset. We step-wise reduce the learning rate by a factor of 0.1 if no improvement is seen for three epochs and early stopping is applied in case of no improvement for nine consecutive epochs. For the optimization algorithm we select Adam (Kingma and Lei, 2015). For model evaluation we track cross-entropy Lovász loss, *IoU* score, Precision (*Prec*) and Recall (*Rec*). In order to account for the focus on applications in emergency response, where computation time is a critical performance criterion, we also report model *throughput* measured in megapixel per second (mp/s). Measurements per experiment are averaged across five prediction runs on 5,000 tiles with shape (256, 256, 3). We use Pytorch as deep learning framework and train the networks in batches of 24 until convergence on two NVIDIA Tesla K80 GPUs.

Since all CNNs make predictions on local windows, higher prediction errors towards the image borders may be observed. Therefore, during inference we expand the input image with mirror-padding, split it into overlapping tiles, run the predictions over batches of tiles, blend the prediction tiles to reconstruct the expanded input image's x-y-shape and un-pad the resulting prediction image. We use a tapered cosine window function to weight pixels when blending overlapping tiles together (Wieland and Martinis, 2019).

To assess the segmentation performance, standard accuracy metrics Intersection over Union score (*IoU*), Precision (*Prec*) and Recall (*Rec*) are reported for a threshold of 0.5 on the prediction probabilities. We do not consider cross-validation due to the computational overhead of running each experiment multiple times. However, to ensure that results are reproducible and comparable on the same system we fix random seeds for all components of the experimental setup that involve randomness (e.g., shuffling of datasets, noisy hidden layers, weight initialization, etc.). We also enforce deterministic GPU floating point calculations.

4 Results

Baseline model (BM)

Table 2 shows the results of different decoder-encoder combinations on the three test scenarios. Models have been trained on IKONOS R-G-B satellite images. The results on test scenario I, which compares against a test split of the reference data (same sensor, same task), show that models based on U-Net decoder perform better than Deeplab-V3+ models. This accounts for their accuracy (0.73 mean *IoU* across all U-Net models compared to 0.72 mean *IoU* for Deeplab-V3+) as well as inference speed (18.80 mean *throughput* for U-Net compared to 5.03 mean *throughput* for Deeplab-V3+). Similarly, Mobilenet-V3 encoder models provide better results than ResNet-50 and EfficientNet-B4 (0.74 mean *IoU* and 20.5 mean *throughput* for Mobilenet-V3 compared to 0.71 mean *IoU* and 6.91 mean *throughput* for EfficientNet-B4, 0.73 mean *IoU* and 8.30 mean *throughput* for ResNet-50).

Test scenario II compares model results against an independent test dataset of water scenes from satellite sensors that have not been used during training (different sensors, same task). All models show a clear performance decrease when transferred to other sensors (0.31 mean drop in *IoU* across all models). In this scenario models based on U-Net decoder perform better than Deeplab-V3+ models (0.41 mean *IoU* for U-Net compared to 0.40 mean *IoU* for Deeplab-V3+). Mobilenet-V3 encoder models provide better results than ResNet-50 and EfficientNet-B4 models (0.45 mean *IoU* for Mobilenet-V3 compared to 0.41 mean *IoU* for EfficientNet-B4 and 0.37 mean *IoU* for ResNet-50).

Table 2: Results for different baseline models trained on IKONOS R-G-B satellite images. Test scenario I compares against a test split of the reference data (same sensors, same task); Test scenario II against an independent test dataset of water scenes from different sensors (different sensors, same task); Test scenario III against an independent test dataset of flood scenes from different sensors (different sensors, different task).

ID	Decoder	Encoder backbone	Throughput (mp/s)	Test scenario I			Test scenario II			Test scenario III		
				IoU	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*
BM-0	**U-Net**	**Mobilenet-V3**	**32.69**	**0.74**	**0.85**	**0.82**	**0.45**	**0.47**	**0.85**	**0.45**	**0.48**	**0.78**
BM-1	U-Net	ResNet-50	13.20	0.73	0.85	0.81	0.38	0.39	0.76	0.41	0.43	0.78
BM-2	U-Net	EfficientNet-B4	10.58	0.71	0.89	0.77	0.41	0.41	0.80	0.42	0.45	0.78
BM-3	DL-V3+	Mobilenet-V3	8.40	0.73	0.85	0.82	0.45	0.46	0.86	0.42	0.44	0.80
BM-4	DL-V3+	ResNet-50	3.36	0.72	0.84	0.81	0.35	0.40	0.73	0.36	0.37	0.77
BM-5	DL-V3+	EfficientNet-B4	3.23	0.70	0.81	0.79	0.40	0.45	0.72	0.42	0.46	0.77

Test scenario III compares model results against an independent test dataset of flood scenes from sensors that have not been used during training (different sensors, different task). A general performance decrease can be observed compared to test scenario I. Similar to the previous tests, also in this scenario models based on U-Net perform better than Deeplab-V3+ models (0.43 mean *IoU* for U-Net compared to 0.40 mean *IoU* for Deeplab-V3+). Mobilenet-V3 encoder models outperform ResNet-50 and EfficientNet-B4 encoder models (0.44 mean *IoU* for Mobilenet-V3 compared to 0.42 mean *IoU* for EfficientNet-B4 and 0.39 mean *IoU* for ResNet-50).

In summary, the U-Net Mobilenet-V3 model provides best results across all test scenarios while being small and fast during inference compared to the other models.

Input bands (IB)

Based on the results of the initial experiment, we use a U-Net model with Mobilenet-V3 encoder and train it on IKONOS satellite images with varying input bands. From Table 3 it can be seen that compared to the baseline (IB-0) adding the NIR spectral band (IB-1) clearly improves the segmentation results across all test scenarios by 0.07 *IoU* for test scenario I, 0.17 *IoU* for scenario II and 0.32 *IoU* for scenario III. Also adding slope information (IB-2) improves the baseline R-G-B model by 0.02 *IoU* for scenario I, 0.03 *IoU* for scenario II and 0.02 *IoU* for scenario III. The combined R-G-B with NIR and Slope model (IB-3) shows best performance and exceeds the baseline by 0.07 *IoU* for scenario I, 0.20 *IoU* for scenario II and 0.35 *IoU* for scenario III.

Table 3: Results of a U-Net model with Mobilenet-V3 encoder trained on IKONOS satellite images with varying input bands.

ID	Input bands	Throughput (mp/s)	Test scenario I			Test scenario II			Test scenario III		
			IoU	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*
IB-0	R-G-B	32.69	0.74	0.85	0.82	0.45	0.47	0.85	0.45	0.48	0.78
IB-1	R-G-B-NIR	32.74	0.81	0.92	0.87	0.62	0.79	0.78	0.77	0.83	0.93
IB-2	R-G-B-Slope	32.98	0.76	0.89	0.82	0.48	0.55	0.78	0.47	0.49	0.80
IB-3	**R-G-B-NIR-Slope**	**33.00**	**0.81**	**0.92**	**0.87**	**0.65**	**0.84**	**0.73**	**0.80**	**0.87**	**0.91**

Data augmentation (DA)

Table 4 depicts the influence of different data augmentation techniques on the segmentation results for the three test scenarios. Based on the results of previous experiments, we use a U-Net model with

Mobilenet-V3 encoder trained on IKONOS R-G-B-NIR-Slope satellite images. DA-0 uses the training dataset as is without augmentation. The results indicate that when augmentation (DA-1) is applied to the training dataset, better performance can be achieved on all test scenarios compared to using the unmodified training data. The effects of augmentation are, however, minor. DA-2 applies test-time augmentation. For test scenarios I and III, the results are comparable to DA-1. On test scenario II test-time augmentation shows negative effects on the accuracy. Since every tile needs to be augmented and predicted multiple times before the final prediction can be assigned, inference times of this approach are longer compared to DA-2.

Table 4: Results of a U-Net model with Mobilenet-V3 encoder trained on IKONOS R-G-B-NIR-Slope satellite images for different data augmentation techniques.

ID	**AUG** Train-time	**AUG** Test-time	**Test scenario I**			**Test scenario II**			**Test scenario III**		
			IoU	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*
DA-0	False	False	0.81	0.92	0.87	0.65	0.84	0.73	0.80	0.87	0.91
DA-1	**True**	**False**	**0.82**	**0.92**	**0.87**	**0.66**	**0.92**	**0.75**	**0.82**	**0.90**	**0.91**
DA-2	True	True	0.82	0.92	0.87	0.62	0.93	0.66	0.82	0.90	0.91

Pre-training (PT)

Table 5 shows the results of a U-Net model with Mobilenet-V3 encoder trained on IKONOS R-G-B-NIR-Slope satellite images with train-time augmentation under consideration of different pre-trained weights. When training and testing on data of the same sensor and task (test scenario I) no improvement can be observed by using pre-trained weights. However, the generalization ability of the model improves with higher *IoU* being measured on test scenarios II and III when ImageNet weights are used to initialize R-G-B bands. Initializing weights for all bands (R-G-B-NIR-Slope) by pre-training on Sentinel-2 images seems to have a negative impact on model transferability to scenarios II and III.

Table 5: Results of a U-Net model with Mobilenet-V3 encoder trained on IKONOS R-G-B-NIR-Slope satellite images with train-time augmentation under consideration of different pre-trained weights.

ID	**Weights**	**Test scenario I**			**Test scenario II**			**Test scenario III**		
		IoU	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*
PT-0	Random	0.82	0.92	0.87	0.66	0.92	0.75	0.82	0.90	0.91
PT-1	**ImageNet (R-G-B) Random (NIR-Slope)**	**0.82**	**0.90**	**0.89**	**0.68**	**0.92**	**0.74**	**0.84**	**0.89**	**0.93**

PT-2	Sentinel-2	0.82	0.91	0.88	0.63	0.75	0.80	0.76	0.82	0.92

Transfer model – satellite to aerial (TM)

Table 6 tests the transferability of a U-Net model with Mobilenet-V3 encoder between satellite and aerial images for different training setups. Since the aerial images used in this study do not have a NIR spectral band available, we train and test with R-G-B-Slope images. Training and testing on satellite images (TM-0) performs significantly better than testing the same model on aerial images for all three scenarios. Training on Mapbox images (TM-1 and TM-2) performs better on aerial images than on satellite images. The performance difference between aerial and satellite images is, however, much smaller compared to TM-0. Training Mapbox with pre-trained weights from satellite images (TM-2) shows better performance than using ImageNet weights. The best overall performance, however, can be reached when a model is trained on a joined training dataset of contrast enhanced satellite images and Mapbox images (TM-3). This model shows the best performance on both satellite and aerial images across all test scenarios.

Table 6: Transferability of a U-Net model with Mobilenet-V3 encoder between satellite and aerial images for different training setups.

ID	**Training**	**Testing**	**Test scenario I**			**Test scenario II**			**Test scenario III**		
			IoU	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*
TM-0	Satellite R-G-B-Slope, pretrained on ImageNet, train-time augment.	Aerial	0.28	0.31	0.81	0.10	0.12	0.88	0.15	0.25	0.31
		Satellite	0.77	0.90	0.82	0.63	0.76	0.79	0.67	0.78	0.81
TM-1	Mapbox R-G-B-Slope, pretrained on ImageNet, train-time augment.	Aerial	0.70	0.83	0.79	0.54	0.82	0.62	0.64	0.88	0.69
		Satellite	0.67	0.76	0.84	0.54	0.60	0.79	0.62	0.72	0.84
TM-2	Mapbox R-G-B-Slope, pretrained on TM-0 train-time augment.	Aerial	0.72	0.84	0.80	0.56	0.89	0.57	0.62	0.74	0.83
		Satellite	0.69	0.84	0.78	0.55	0.54	0.91	0.64	0.69	0.89
TM-3	**TM-0 and TM-1 trained together with contrast enhancement**	**Aerial**	**0.72**	**0.84**	**0.80**	**0.78**	**0.82**	**0.82**	**0.70**	**0.86**	**0.74**
		Satellite	**0.79**	**0.89**	**0.87**	**0.70**	**0.78**	**0.89**	**0.82**	**0.89**	**0.90**

Since contrast enhancement of the satellite images seems to have a positive effect on the segmentation performance, we also applied this additional pre-processing step to the NIR spectral band and trained a model on contrast enhanced R-G-B-NIR bands and slope information. Table 7 indicates that when training and testing on data of the same sensor, the effect of contrast enhancement is not seen and

performances on test scenario I are equal between PT-1 and PT-1 Contrast. However, the model trained on contrast enhanced data (PT-1 Contrast) transfers better to test scenarios II and III than the same model trained on the original top of canopy reflectance values (PT-1).

Table 7: Influence of contrast enhancement on best performing U-Net model with Mobilenet-V3 encoder with pre-trained ImageNet weights and train-time augmentation for R-G-B-NIR-Slope satellite images (PT-1).

ID	Weights	Test scenario I			Test scenario II			Test scenario III		
		IoU	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*
PT-1	ImageNet (R-G-B) Random (NIR-Slope)	0.82	0.90	0.89	0.68	0.92	0.74	0.84	0.89	0.93
PT-1 Contrast	**ImageNet (R-G-B) Random (NIR-Slope) Contrast enhancement**	**0.82**	**0.90**	**0.89**	**0.72**	**0.82**	**0.86**	**0.85**	**0.92**	**0.92**

To further evaluate the usefulness of our results for flood mapping applications we also tested the influence of adding samples of the target (flood water) domain to the training data. Due to the limited availability of flood water samples, we divided the six available flood images from test scenario III into training and test splits. The best-performing model for aerial and satellite images (TM-3) is then retrained including data from the target (flood water) domain. Specifically, we add two of the flood images to the training data and use the remaining four flood images as independent test data. Table 8 provides an overview of the data scenarios and compares the results for simply applying the TM-3 model (trained without images of the target domain) with TM-3 Flood (trained including images of the target domain). It can be seen that adding few flood water samples during training could slightly improve the results on test images that show flood situations. An increase in *Rec* of 0.02 indicates that the improvement is mainly related to more flood water pixels being correctly detected.

Table 8: Influence of retraining TM-3 with limited training data of the target (flood water) domain.

ID	Training	Testing	*IoU*	*Prec*	*Rec*
TM-3	TM-3 without flood water images	Landes, France (GeoEye-1), Brück, Germany (WorldView-3), Altenahr, Germany (DLR MACS), Dresden, Germany (DLR 3K)	0.73	0.89	0.80
TM-3 Flood	**TM-3 with flood water images from Bihar, India (GeoEye-1) and Minden, Germany (DLR 4K)**	**Landes, France (GeoEye-1), Brück, Germany (WorldView-3), Altenahr, Germany (DLR MACS), Dresden, Germany (DLR 3K)**	**0.74**	**0.89**	**0.82**

Finally, we compared the influence of image resolution on the segmentation performance. All previous experiments used the original resolution of the test images. As can be seen from Table 9, resampling of the test images leads to worse performance on all test scenarios. Predicting on the original image resolution, despite varying spatial resolution across the test images seems to be the best choice. This is further underlined by the observation that the closer the resampled resolution is to the average original image resolution (0.8 m for scenario I and 0.5 m for scenarios II and III) the better the results.

Table 9: Influence of image spatial resolution on best performing U-Net model with Mobilenet-V3 encoder, pre-trained ImageNet weights and train-time augmentation for contrast enhanced R-G-B-NIR-Slope satellite images (PT-1 Contrast).

ID	Resolution (m)	Test scenario I			Test scenario II			Test scenario III		
		IoU	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*	*IoU*	*Prec*	*Rec*
RES-0	Varying original	**0.82**	**0.90**	**0.89**	**0.72**	**0.82**	**0.86**	**0.85**	**0.92**	**0.92**
RES-1	0.25	0.76	0.81	0.91	0.68	0.79	0.83	0.80	0.85	0.90
RES-2	0.50	0.80	0.87	0.89	0.72	0.84	0.83	0.84	0.90	0.92
RES-3	0.75	0.82	0.89	0.90	0.62	0.73	0.76	0.79	0.91	0.82
RES-4	1.00	0.80	0.89	0.87	0.56	0.70	0.70	0.76	0.90	0.80

Summary of results for best performing model

Figure 3 shows the results of the best-performing U-Net Mobilenet-V3 model with pre-trained ImageNet weights trained on contrast enhanced and augmented R-G-B-Slope IKONOS and Mapbox data (TM-3) on test scenario I grouped by predominant landcover of the test samples (aerial and satellite test datasets combined). Overall good performance across all landcover types can be observed with median values for all metrics above 0.75. Performance over the majority of landcover types, moreover, shows relatively small variation. Few landcover types, however, highlight a larger spread of the performance with lower quartiles for *IoU* being below 0.70. This becomes evident in particular for bare areas (class 1), herbaceous cover (class 5) and urban areas (class 16). A closer look at *Prec* and *Rec* reveals that largely *Prec* is affected for water bodies located in bare areas, which indicates a tendency to overestimate water in these areas. Herbaceous cover and urban areas show a larger spread of *Rec* values, which indicates a tendency to underestimate water.

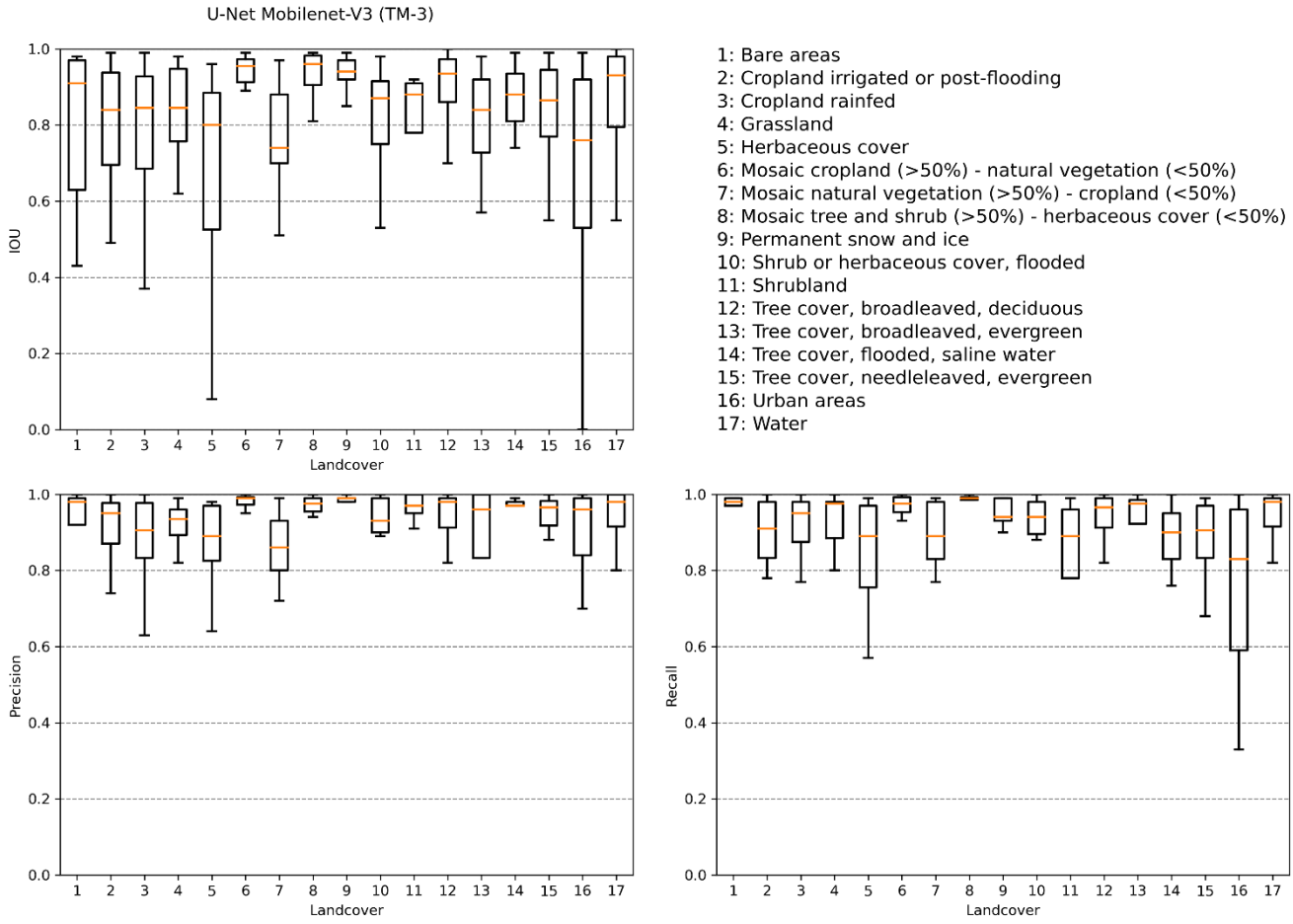

Figure 3: Results of best-performing U-Net Mobilenet-V3 model with pre-trained ImageNet weights trained on contrast enhanced and augmented R-G-B-Slope IKONOS and Mapbox images (TM-3) on test scenario I (aerial and satellite test datasets combined) grouped by predominant landcover.

Figure 4 shows results of the best-performing overall model (TM-3) for selected scenario I test images per landcover. All water predictions show good performance and match well the reference masks. Some minor issues with false positives and negatives can be observed for example in water look-alike areas (e.g., shadows or moist soil) or where a clear border between water and land is difficult to define even by visual image interpretation. The "cropland irrigated or post-flooding" sample (zoom A) depicts a complex scene with a mixture of different shaped and coloured water basins and canals. Despite the large variety of water appearance in the scene, the overall accuracy of the predicted water mask is good with an *IoU* of 0.77, *Prec* and *Rec* above 0.85. In the "urban areas" sample (zoom B) it can be seen that narrow water canals of central Venice in between dense built-up areas are not sufficiently segmented thus leading to false negatives and a lower *Rec*. Similar effects can be observed in other urban samples (see also Figure 3). Despite the very good model performance on the "cropland (>50%) – natural (<50%)" sample (zoom C), it is a typical example for a land-water border that even by visual image interpretation is not

clearly identifiable. Especially shallow water areas and muddy river banks in the lower left corner make it difficult to identify a clear land-water border.

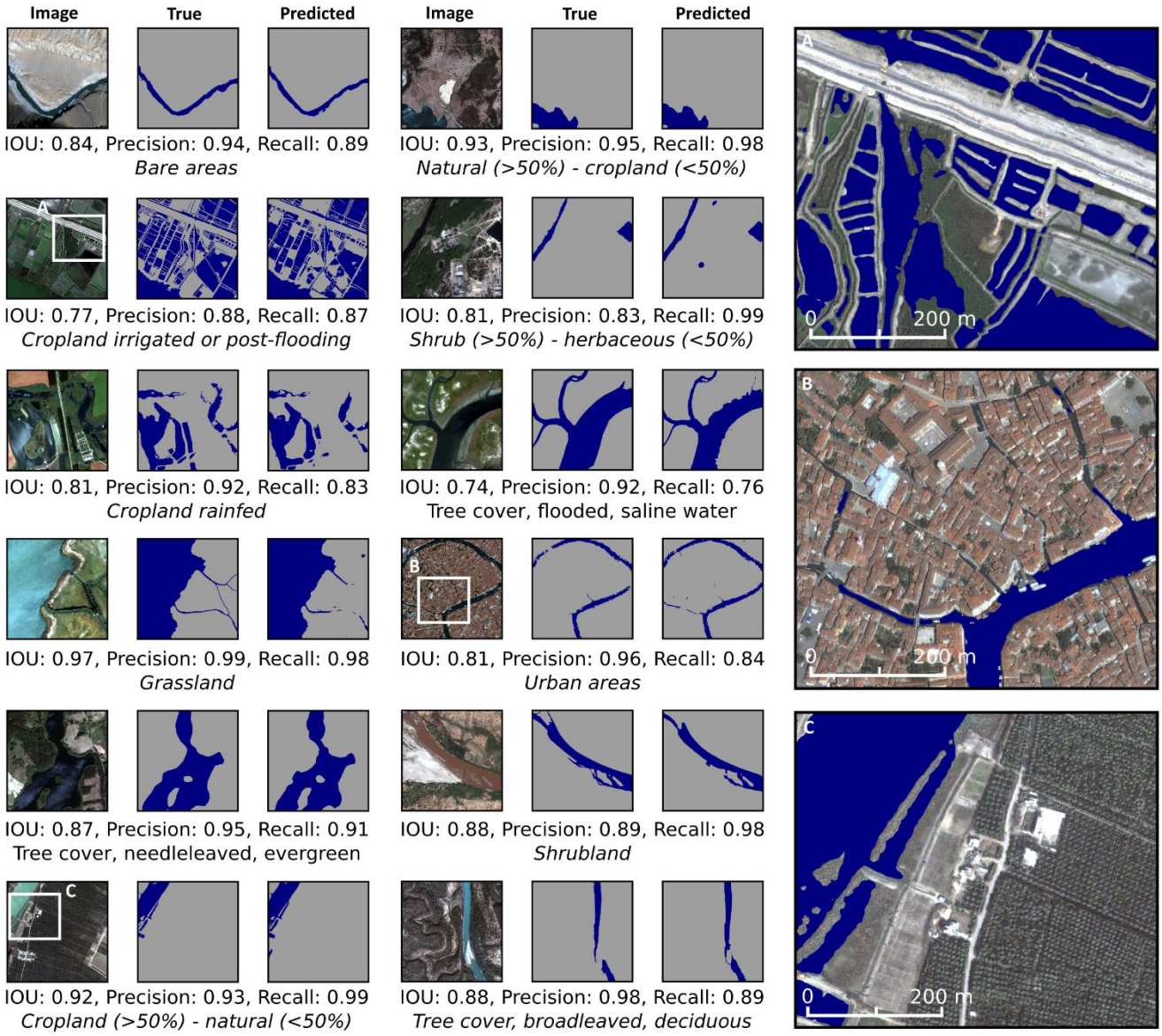

Figure 4: Results of best-performing model (TM-3) for randomly selected scenario I test images per landcover. Water is outlined in blue, background in grey. Each sample covers an area of approximately 1.3 km².

Figures 5 shows results of the U-Net Mobilenet-V3 model with pre-trained ImageNet weights trained on contrast enhanced and augmented IKONOS and Mapbox data (TM-3) for scenario II and III satellite test images. The flood water image of Landes acquired by the GeoEye-1 satellite highlights the overall good performance of the model across water bodies with heterogeneous appearance. False negatives are mainly observed in water with high sediment content and along fuzzy land-water borders with partially overlapping vegetation. The normal water WorldView-2 scene of Bihar (zoom B) indicates good

performance with only few false positives, despite some visible building shadows. Water bodies inside the city are segmented with only minor misclassifications even in muddy areas along the main river.

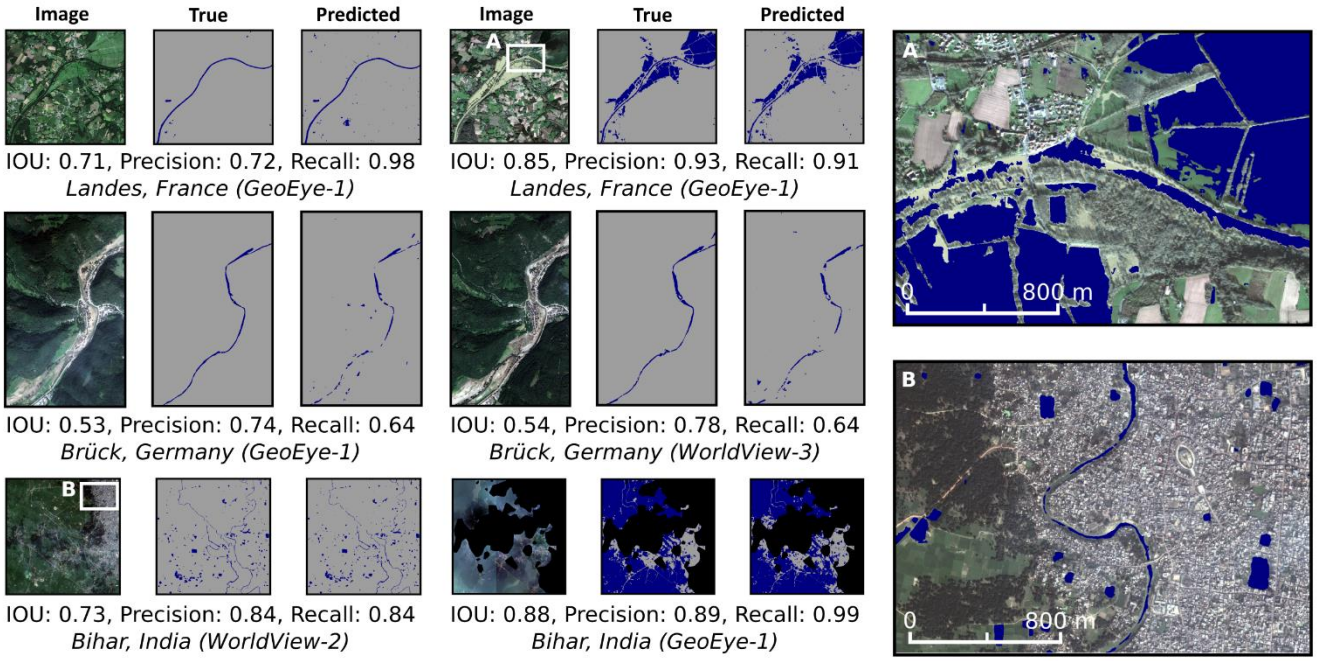

Figure 5: Results of best-performing model (TM-3) for scenario II (left) and III (right) satellite test images. Water is outlined in blue, background in grey and no-data areas in black. The area covered by each image varies and is depicted in Table 1.

Figure 6 shows results of the U-Net Mobilenet-V3 model with pre-trained ImageNet weights trained on contrast enhanced and augmented IKONOS and Mapbox data (TM-3) for scenario II and III aerial test images. The normal water Zenmuse H20T image acquired by UAV of Altenahr (zoom A) shows problems with false negatives related to shallow transparent water that may cause confusion with the sandy river bottom. This is reflected in the low *Rec* values for this scene. The DLR 3K aerial flood water scene of Dresden (zoom B) shows low *Rec* values and has problems with false negatives that are likely related to the large amount of sediment in the water in combination with partially inundated vegetation and allotment gardens.

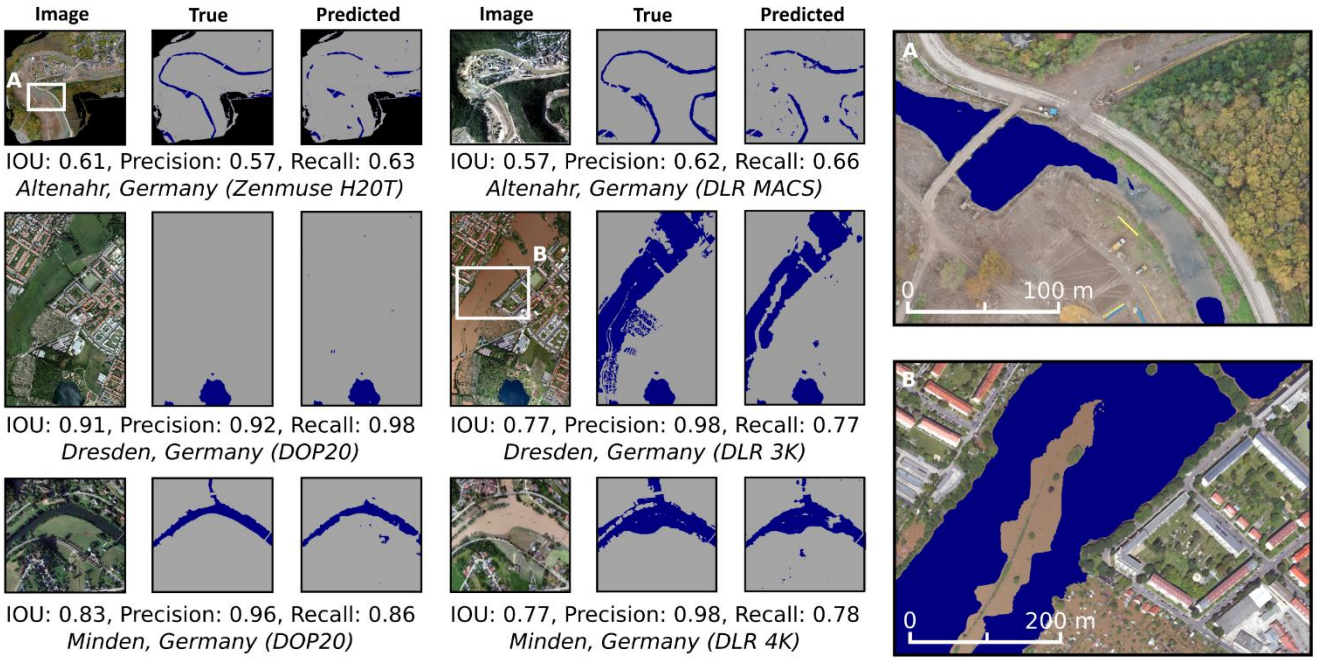

Figure 6: Results of best-performing model (TM-3) for scenario II (left) and III (right) aerial test images. Water is outlined in blue, background in grey and no-data areas in black. The area covered by each image varies and is depicted in Table 1.

5 Discussion

The objective of this study is to train a model for semantic segmentation of normal water and flood water bodies in very high-resolution satellite and aerial images. Performance evaluation of CNN architectures and encoders indicate superior performance of a U-Net Mobilenet-V3 model when considering *IoU*, *Prec*, *Rec* and *throughput* (Table 2). Compared to other studies, we focus specifically on multi-sensor generalization ability, global applicability, simplicity and processing speed to produce timely, accurate and relevant information under varying data availability scenarios. Our work is thus targeted towards operational use in rapid mapping activities to support situational awareness in flood emergency response. Similar to Yang et al. (2020) we can report good performance for all tested models with a ResNet-50 encoder. Following their conclusions, we have not tested the more complex ResNet-100, which performed worse in their experiments. Instead we focussed on the more recently developed Mobilenet-V3 and EfficientNet-B4, which showed comparable results than ResNet-50 at higher *throughput*. Contrary to the findings of Duan and Hu (2020), who compare a multi-scale refinement network with U-Net, SegNet and DeepLab-V3+ networks for water segmentation in WorldView-3 and GaoFen-2 satellite images, our experiments indicate that U-Net-based models outperform models based on the more complex DeepLab-V3+ architecture both in terms of accuracy metrics and *throughput*. While the

549 difference in *IoU* between the two architectures is minimal in case of test scenario I (same sensor, same
550 task), it becomes more evident in test scenarios II (different sensor, same task) and III (different sensor,
551 different task). Further performance improvements for DeepLab-V3+ could be achieved by longer
552 training without early stopping. More computational resources would be required for a more thorough
553 comparison of the architectures, since deeper architectures like DeepLab-V3+ commonly take longer to
554 converge and hence their training in our experiments may be sub-optimal given the relatively short
555 patience before learning rate reduction and early stopping. Independent on training times, the main
556 decision criterion in this study to select a U-Net-based model has been the significantly higher throughput
557 during inference.

558 Duan and Hu (2020) use the Deepglobe dataset (Demir et al., 2018) and expand it with a custom dataset
559 of GaoFen-2 images over China. We found the Deepglobe dataset not to be suitable for our study, due to
560 limited geographical coverage and relatively poor annotation quality for the water class. Therefore, we
561 acquired a comprehensive multi-sensor reference dataset that we used to train, validate and test the water
562 segmentation. The reference dataset has been sampled globally to account for a large variety of climatic,
563 atmospheric and land-cover conditions. To overcome the limited geographical coverage of the IKONOS
564 data archive that we had access to, we complement the reference dataset with freely available optical
565 images from a Web Map Service and water masks from OpenStreetMap. Nevertheless, even the
566 combined reference dataset shows a geographical bias towards Europe, Middle East and India (Figure
567 1). Also, the acquired independent test datasets for other sensors (test scenario II) and tasks (test scenario
568 III) show a geographical bias towards Europe and India. Similar to the reference dataset (test scenario I),
569 this is caused by limited availability of suitable test images from other locations. The aim in this study
570 was to compile test data from different sensors covering the same location during normal and flood water
571 situations. This constraint strongly reduced the number of available and accessible samples. The dataset
572 should be further extended and diversified in future to improve the significance of the results regarding
573 model transferability.

574 To allow for a high degree of transferability across sensors, we focussed on the use of R-G-B image
575 bands, which are acquired by almost all available optical sensors independent on the acquisition platform.
576 For sensor-specific applications, however, including additional spectral bands like the NIR band into the
577 input feature space has clearly a positive effect on the performance (Table 3 and Table 7). Additionally,
578 slope information derived from a DEM can further improve results. Since high-quality DEMs like the
579 Copernicus DEM are freely available at global coverage, slope information can be added independently
580 for all sensors.

Sample water masks are derived by support of manual interpretation and digitization of image tiles, which means that the results are to be regarded as relative to the performance of a human analyst and not to an absolute ground-truth. Especially the fractal geometry of the border between land and water may lead to label inconsistencies when trying to define a hard class boundary. Their manual delineation is at least partially a matter of subjective interpretation and may introduce a bias in the reference dataset and hence into the performance results. In this context, it also needs to be considered which level of detail and resolution is actually required by an application. Wetland monitoring for example may have different requirements compared to flood emergency response. In this study we focus on emergency response and have followed as much as possible best-practises of rapid mapping procedures as being conducted at the Centre for Satellite-based Crisis Information (ZKI) of the German Aerospace Center (DLR) (Lechner and Gähler, 2017).

Results show that more efforts should be spend on preparing high-quality training data and dedicate more research towards reference data preparation and its incremental improvement. The target domain is affected by varying scene and image properties that are dependent on sensor characteristics (e.g., radiometry, spatial resolution), atmospheric conditions, land-use / land-cover of the background class and appearance of the water class. Despite our aim to cover such variations as part of the reference dataset, we decided to apply image augmentation to artificially increase the training sample size and to cover a larger range of conditions that may occur during inference in real-world applications. The need for augmentation has been reinforced by the outcomes of an experiment to test the influence of different levels and combinations of data augmentation on the segmentation results (Table 4). Although we are not specifically isolating the effects of single augmentations (contrast, brightness, scale and rotation) on the performance, it can be noted that scaling images during augmentation seems to improve the performance on images of a fraction of the original resolution. Test-time augmentation as well as resampling images before prediction (Table 9) did not improve results. Predicting on the original image and resolution seems to be the best choice, despite varying spatial resolutions across the test images.

Similar to what is reported by Gebrehiwot et al. (2019) for the task of extracting flooded areas from UAV images, our results confirm that adapting pre-trained models is highly beneficial and can lead to superior performance compared to training with randomly initialized weights. While He et al. (2018) challenge the commonly used technique of pre-training on ImageNet and fine-tuning on custom datasets, we can report that for our use case pre-training on not remote sensing specific image databases like ImageNet helped to improve the results. Pre-training on a domain-specific water dataset of lower resolution Sentinel-2 images did not improve performance on tests scenario I and even produced lower accuracies

when transferred on test scenario II and III compared to a model with random weight initialization (Table 5). A possible explanation for this behaviour could be that pre-training on Sentinel-2 emphasizes the importance of spectral intensity values for the final model, which could hamper the transferability to other sensors and tasks. In contrast, models pre-trained on ImageNet are known to be biased towards texture, which may increase the robustness towards a range of image distortions and thus support model transferability (Geirhos et al., 2019). This assumption is further supported by the observation that applying percentage linear contrast stretch improved model accuracies on test scenarios II and III (Table 6 and Table 7). Similar to the findings of Kaiser et al. (2017)), our results indicate that by adding noisy large-scale training data (in our case from Mapbox images with OpenStreetMap annotations) performance improvements can be achieved with relatively little manual annotation effort.

Moreover, we could show that adding a limited number of training samples from the target (flood water) domain, could further improve results on test scenario III (Table 8). This is particularly relevant for applications targeted towards flood disaster response. More flood samples are, however, required to further test and verify the effectiveness of this approach. Further approaches to improve model accuracy, especially on test scenarios II and III, include fine-tuning on limited amounts of domain-specific training data (e.g., from flood events and / or other sensors), multi-source domain adaptation by data standardization (Tasar et al., 2020) or few-shot learning (Rußwurm et al., 2020).

6 Conclusions

In this study, we combined different architectures (U-Net and DeepLab-V3+) with various encoder backbones (Mobilenet-V3, ResNet-50 and EfficientNet-B4) and tested their ability to delineate normal and flood water under varying environmental conditions and data availability scenarios. We introduced a reference dataset of 1,120 globally sampled images with quality checked binary water masks to train, validate and test the CNN models. Compared to previous studies, we considered a large variety of available satellite and aerial sensors as input, namely IKONOS, GeoEye-1, WorldView-2, WorldView-3 and four different airborne cameras. Across several experiments we identified the superior performance of a U-Net Mobilenet-V3 model with initial weights being pre-trained on ImageNet. The best-performing model indicates good generalization ability across sensors and varying environmental conditions. In this context, not only the combination of architecture, encoder and weight initialization is relevant, but also the way training data are augmented. While a model trained with R-G-B-NIR-Slope input feature space produces the highest accuracies on all test scenarios (PT-1 Contrast), models without the NIR spectral band can achieve comparable results (TM-3). The benefit of a model that requires only R-G-B-Slope

input features is that it is applicable across a larger spectrum of sensors (e.g., none of the aerial images used in this study had an additional NIR spectral band available). Contrast enhancement and noisy training data from Mapbox and OpenStreetMap could improve performance on all test scenarios. By successfully applying the model to four exemplary flood events, we could highlight the usefulness of this work for rapid mapping activities to support situational awareness in emergency response. In particular, we could show that it is possible to train a model that is able to cope with highly diverse data availability scenarios in disaster situations. Our experiments further indicate that in order to improve flood mapping results it is beneficial to add even limited amounts of flood water samples during model training. More work is, however, required to further adapt to specific scene properties that are common during floods and that have caused misclassifications in the test images. This includes high degrees of sediment in the water, strong sun-light reflections on the water surface and presence of cloud and terrain shadows. Ongoing and future works focus on acquiring more samples for fine-tuning models with domain-specific training data from flood events and applying the method on large scale datasets of past floods (e.g., flood in Germany 2013 and 2021). Finally, the proposed method will support complementation of existing systematic water and flood monitoring services based on Sentinel-1 (Helleis et al., 2022; Twele et al., 2016) and Sentinel-2 (Wieland and Martinis, 2019) with an ad-hoc component for on-demand very high-resolution optical satellite and aerial imagery.

Funding: This work was supported by the German Federal Ministry of Education and Research (BMBF) as part of the AIFER project [Grant No. 13N15525] and DLR-internal funding.

Conflicts of interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. arXiv:1907.10902 [cs, stat].

Azimi, S.M., Kiefl, R., Gstaiger, V., Bahmanyar, R., Merkle, N., Henry, C., Rosenbaum, D., Kurz, F., 2021. Automatic object segmentation to support crisis management of large-scale events. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLIII-B2-2021, 433–440. https://doi.org/10.5194/isprs-archives-XLIII-B2-2021-433-2021

Ball, J.E., Anderson, D.T., Sr, C.S.C., 2017. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. JARS 11, 042609. https://doi.org/10.1117/1.JRS.11.042609

Bonafilia, D., Tellman, B., Anderson, T., Issenberg, E., 2020. Sen1Floods11: a georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 835–845.

https://doi.org/10.1109/CVPRW50498.2020.00113
Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L., Smets, B., 2020. Copernicus Global Land Cover Layers—Collection 2. Remote Sensing 12, 1044. https://doi.org/10.3390/rs12061044
Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2021. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. arXiv:2105.05537v1 [eess.IV].
Castillo-Navarro, J., Le Saux, B., Boulch, A., Audebert, N., Lefèvre, S., 2021. Semi-supervised semantic segmentation in Earth Observation: the MiniFrance suite, dataset analysis and multi-task network study. Mach Learn. https://doi.org/10.1007/s10994-020-05943-y
Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv:1706.05587 [cs].
Chen, Y., Fan, R., Yang, X., Wang, J., Latif, A., 2018. Extraction of Urban Water Bodies from High-Resolution Remote-Sensing Imagery Using Deep Learning. Water 10, 585. https://doi.org/10.3390/w10050585
Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, 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) 172–17209. https://doi.org/10.1109/CVPRW.2018.00031
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848
Ding, L., Lin, D., Lin, S., Zhang, J., Cui, X., Wang, Y., Tang, H., Bruzzone, L., 2022. Looking Outside the Window: Wide-Context Transformer for the Semantic Segmentation of High-Resolution Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing 60, 1–13. https://doi.org/10.1109/TGRS.2022.3168697
Duan, L., Hu, X., 2020. Multiscale Refinement Network for Water-Body Segmentation in High-Resolution Satellite Imagery. IEEE Geosci. Remote Sensing Lett. 17, 686–690. https://doi.org/10.1109/LGRS.2019.2926412
Fahrland, E., Jacob, P., Schrader, H., Kahabka, H., 2020. Copernicus Digital Elevation Model Product Handbook (No. AO/1-9422/18/I-LG), Product handbook. Airbus.
Feng, W., Sui, H., Huang, W., Xu, C., An, K., 2019. Water Body Extraction From Very High-Resolution Remote Sensing Imagery Using Deep U-Net and a Superpixel-Based Conditional Random Field Model. IEEE Geosci. Remote Sensing Lett. 16, 618–622. https://doi.org/10.1109/LGRS.2018.2879492
Gebrehiwot, A., Hashemi-Beni, L., Thompson, G., Kordjamshidi, P., Langan, T., 2019. Deep Convolutional Neural Network for Flood Extent Mapping Using Unmanned Aerial Vehicles Data. Sensors 19, 1486. https://doi.org/10.3390/s19071486
Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W., 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arxiv.org/abs/1811.12231 [cs, q-bio, stat].
Gomes, R., Rozario, P., Adhikari, N., 2021. Deep Learning optimization in remote sensing image segmentation using dilated convolutions and ShuffleNet, in: 2021 IEEE International Conference on Electro Information Technology (EIT). pp. 244–249. https://doi.org/10.1109/EIT51626.2021.9491910
Gu, X., Li, S., Ren, S., Zheng, H., Fan, C., Xu, H., 2022. Adaptive enhanced swin transformer with U-net for remote sensing image segmentation. Computers and Electrical Engineering 102, 108223. https://doi.org/10.1016/j.compeleceng.2022.108223
Guo, H., He, G., Jiang, W., Yin, R., Yan, L., Leng, W., 2020. A Multi-Scale Water Extraction

Convolutional Neural Network (MWEN) Method for GaoFen-1 Remote Sensing Images. IJGI 9, 189. https://doi.org/10.3390/ijgi9040189
Gupta, R., Goodman, B., Patel, N., Hosfelt, R., Sajeev, S., Heim, E., Doshi, J., Lucas, K., Choset, H., Gaston, M., 2019. Creating xBD: A Dataset for Assessing Building Damage from Satellite Imagery. arXiv:1911.09296 [cs.CV].
Haklay, M., 2010. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. Environment and Planning B: Planning and Design 37, 682–703. https://doi.org/10.1068/b35097
Hänsch, R., Arndt, J., Lunga, D., Gibb, M., Pedelose, T., Boedihardjo, A., Petrie, D., Bacastow, T.M., 2022. SpaceNet 8 - The Detection of Flooded Roads and Buildings. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).
He, K., Girshick, R., Dollár, P., 2018. Rethinking ImageNet Pre-training. arXiv:1811.08883v1 [cs.CV].
He, K., Zhang, X., Ren, S., Sun, J., 2015a. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs].
He, K., Zhang, X., Ren, S., Sun, J., 2015b. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. arXiv:1502.01852 [cs].
Helleis, M., Wieland, M., Krullikowski, C., Martinis, S., Plank, S., 2022. Sentinel-1-based water and flood mapping: benchmarking convolutional neural networks against an operational rule-based processing chain. Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15, 2023–2036.
Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.-C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., Le, Q., 2019a. Searching for MobileNetV3, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Seoul, Korea (South), pp. 1314–1324. https://doi.org/10.1109/ICCV.2019.00140
Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E., 2020. Squeeze-and-Excitation Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372
Huang, C., Chen, Y., Zhang, S., Wu, J., 2018. Detecting, Extracting, and Monitoring Surface Water From Space Using Optical Sensors: A Review. Rev. Geophys. 56, 333–360. https://doi.org/10.1029/2018RG000598
Iglovikov, V., Mushinskiy, S., Osin, V., 2017. Satellite Imagery Feature Detection using Deep Convolutional Neural Network: A Kaggle Competition. arXiv:1706.06169 [cs].
Ireland, G., Volpi, M., Petropoulos, G., 2015. Examining the Capability of Supervised Machine Learning Classifiers in Extracting Flooded Areas from Landsat TM Imagery: A Case Study from a Mediterranean Flood. Remote Sensing 7, 3372–3399. https://doi.org/10.3390/rs70303372
Kaiser, P., Wegner, J.D., Lucchi, A., Jaggi, M., Hofmann, T., Schindler, K., 2017. Learning Aerial Image Segmentation From Online Maps. IEEE Transactions on Geoscience and Remote Sensing 55, 6054–6068. https://doi.org/10.1109/TGRS.2017.2719738
Kingma, D.P., Lei, J., 2015. Adam: A Method for Stochastic Optimization. ArXiv:1412.6980v9 15.
Lechner, K., Gähler, M., 2017. Earth observation based crisis information — Emergency mapping services and recent operational developments, in: 2017 4th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM). pp. 1–7. https://doi.org/10.1109/ICT-DM.2017.8275682
Li, Y., Martinis, S., Wieland, M., 2019. Urban flood mapping with an active self-learning convolutional neural network based on TerraSAR-X intensity and interferometric coherence. ISPRS Journal of Photogrammetry and Remote Sensing 152, 178–191. https://doi.org/10.1016/j.isprsjprs.2019.04.014

Long, Y., Xia, G.-S., Li, S., Yang, W., Yang, M.Y., Zhu, X.X., Zhang, L., Li, D., 2021. On Creating Benchmark Dataset for Aerial Image Interpretation: Reviews, Guidances, and Million-AID. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14, 4205–4230. https://doi.org/10.1109/JSTARS.2021.3070368

Mapbox Satellite: global base map & satellite imagery [WWW Document], 2021. URL https://www.mapbox.com/maps/satellite (accessed 11.23.21).

Martinis, S., Twele, A., Plank, S., Zwenzner, H., Danzeglocke, J., Strunz, G., Lüttenberg, H.-P., Dech, S., 2017. The International Charter 'Space and Major Disasters': DLR's Contributions to Emergency Response Worldwide. PFG 85, 317–325. https://doi.org/10.1007/s41064-017-0032-1

Mattyus, G., Wang, S., Fidler, S., Urtasun, R., 2015. Enhancing Road Maps by Parsing Aerial Images Around the World, in: 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, Santiago, pp. 1689–1697. https://doi.org/10.1109/ICCV.2015.197

McFeeters, S.K., 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. International Journal of Remote Sensing 17, 1425–1432. https://doi.org/10.1080/01431169608948714

Mnih, V., Hinton, G., 2012. Learning to Label Aerial Images from Noisy Data, in: Proceedings of the 29 Th International Conference on Machine Learning. Edinburgh, p. 8.

Neupane, B., Horanont, T., Aryal, J., 2021. Deep Learning-Based Semantic Segmentation of Urban Features in Satellite Images: A Review and Meta-Analysis. Remote Sensing 13, 808. https://doi.org/10.3390/rs13040808

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., D'amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C., Loucks, C.J., Allnutt, T.F., Ricketts, T.H., Kura, Y., Lamoreux, J.F., Wettengel, W.W., Hedao, P., Kassem, K.R., 2001. Terrestrial Ecoregions of the World: A New Map of Life on Earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. BioScience 51, 933–938. https://doi.org/10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2

OpenStreetMap [WWW Document], 2021. . OpenStreetMap. URL https://www.openstreetmap.org/ (accessed 11.23.21).

Pan, S.J., Yang, Q., 2010. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering 22, 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Pekel, J.-F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature 540, 418–422. https://doi.org/10.1038/nature20584

Rahnemoonfar, M., Chowdhury, T., Sarkar, A., Varshney, D., Yari, M., Murphy, R.R., 2021. FloodNet: A High Resolution Aerial Imagery Dataset for Post Flood Scene Understanding. IEEE Access 9, 89644–89654. https://doi.org/10.1109/ACCESS.2021.3090981

Rakhlin, A., Davydow, A., Nikolenko, S., 2018. Land Cover Classification from Satellite Imagery with U-Net and Lovász-Softmax Loss, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, Salt Lake City, UT, USA, pp. 257–2574. https://doi.org/10.1109/CVPRW.2018.00048

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J.D., 2013. ISPRS Test Project on Urban Classification and 3D Building Reconstruction (No. 12).

Rußwurm, M., Wang, S., Korner, M., Lobell, D., 2020. Meta-Learning for Few-Shot Land Cover Classification, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition

Workshops (CVPRW). IEEE, Seattle, WA, USA, pp. 788–796. https://doi.org/10.1109/CVPRW50498.2020.00108
Schreier, G., Dech, S., Diedrich, E., Maass, H., Mikusch, E., 2008. Earth observation data payload ground segments at DLR for GMES. Acta Astronautica, Touching Humanity - Space for Improving Quality of Life. Selected Proceedings of the 58th International Astronautical Federation Congress, Hyderabad, India, 24-28 September 2007 63, 146–155. https://doi.org/10.1016/j.actaastro.2007.12.010
Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs].
Tan, M., Le, Q.V., 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv:1905.11946 [cs, stat].
Tasar, O., Tarabalka, Y., Giros, A., Alliez, P., Clerc, S., 2020. StandardGAN: Multi-source Domain Adaptation for Semantic Segmentation of Very High Resolution Satellite Images by Data Standardization, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, Seattle, WA, USA, pp. 747–756. https://doi.org/10.1109/CVPRW50498.2020.00104
Twele, A., Cao, W., Plank, S., Martinis, S., 2016. Sentinel-1-based flood mapping: a fully automated processing chain. International Journal of Remote Sensing 37, 2990–3004. https://doi.org/10.1080/01431161.2016.1192304
Voigt, S., Giulio-Tonolo, F., Lyons, J., Kučera, J., Jones, B., Schneiderhan, T., Platzeck, G., Kaku, K., Hazarika, M.K., Czaran, L., Li, S., Pedersen, W., James, G.K., Proy, C., Muthike, D.M., Bequignon, J., Guha-Sapir, D., 2016. Global trends in satellite-based emergency mapping. Science 353, 247–252. https://doi.org/10.1126/science.aad8728
Wieland, M., Helleis, M., Fichtner, F., Krullikowski, C., Martinis, S., Plank, S., Motagh, M., in preparation. S1S2-Water: A global dataset for semantic segmentation of water bodies from Sentinel-1 and Sentinel-2 data.
Wieland, M., Martinis, S., 2019. A modular processing chain for automated flood monitoring from multi-spectral satellite data. Remote Sensing 11, 2330.
Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated Residual Transformations for Deep Neural Networks. https://doi.org/10.48550/arXiv.1611.05431
Xu, Z., Zhang, W., Zhang, T., Yang, Z., Li, J., 2021. Efficient Transformer for Remote Sensing Image Segmentation. Remote Sensing 13, 3585. https://doi.org/10.3390/rs13183585
Yang, F., Feng, T., Xu, G., Chen, Y., 2020. Applied method for water-body segmentation based on mask R-CNN. J. Appl. Rem. Sens. 14, 1. https://doi.org/10.1117/1.JRS.14.014502
Yuan, W., Wang, J., Xu, W., 2022. Shift Pooling PSPNet: Rethinking PSPNet for Building Extraction in Remote Sensing Images from Entire Local Feature Pooling. Remote Sensing 14, 4889. https://doi.org/10.3390/rs14194889
Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network. arXiv:1612.01105 [cs].