Deep learning-based multi-spectral satellite image segmentation for water body detection

Kunhao Yuan, Xu Zhuang, Gerald Schaefer, Jianxin Feng*, Lin Guan, Hui Fang

***Abstract*—Automated water body detection from satellite imagery is a fundamental stage for urban hydrological studies. In recent years, various deep convolutional neural network (DCNN)-based methods have been proposed to segment remote sensing data collected by conventional RGB or multi-spectral imagery for such studies. However, how to effectively explore the wider spectrum bands of multi-spectral sensors to achieve significantly better performance compared to the use of only RGB bands has been left underexplored. In this paper, we propose a novel deep convolutional neural network model – Multi-Channel Water Body Detection Network (MC-WBDN) – that incorporates three innovative components, a multi-channel fusion module, an Enhanced Atrous Spatial Pyramid Pooling (EASPP) module, and Space-to-Depth (S2D)/Depth-to-Space (D2S) operations, to outperform state-of-the-art DCNN-based water body detection methods. Experimental results convincingly show that our MC-WBDN model achieves remarkable water body detection performance, is more robust to light and weather variations and can better distinguish tiny water bodies compared to other DCNN models.**

***Index Terms*—Multi-spectral remote sensing, water body detection, deep convolutional neural networks, semantic segmentation, feature fusion**

I. Introduction

WATER body detection from remote sensing imagery is of great importance for urban hydrological studies [1]. Urban hydrology has become an emerging research area that allows to improve and manage urban water systems for solving environmental issues caused by rapid urbanisation. It also facilitates timely flood protection planning and water quality control for public safety and health [2]. To achieve an insightful analysis of water systems in cities, automated accurate water body detection is the first and fundamental stage to provide pixel-level identification of water regions [3], [4].

Since its launch in 2015, the Sentinel-2 satellite has provided publicly available multi-spectral imagery that has been widely employed in land-cover applications [5], [6], [7]. It offers one of the most suitable data sources for timely urban hydrological monitoring and analysis due to its near-daily update frequency compared to higher-resolution remote sensing data such as Very High Spatial Resolution (VHR) [8] and Synthetic Aperture Radar (SAR) [4]. Thus, in this paper, we investigate the use of 10 meter resolution multi-spectral data from Sentinel-2 due to its potential for urban hydrological applications that require frequently updated data in their analysis process.

Traditional water body detection methods design handcrafted statistical features extracted from multi-spectral imagery including near infrared (NIR) and short-wave infrared (SWIR). Well-known features include the Normalized Difference Water Index (NDWI) [9], Normalized Difference Moisture Index (NDMI) [10], Modified Normalized Difference Water Index (MNDWI) [11], Automated Water Extraction Index (AWEI) [12], and Pixel Region Index (PRI) [13]. Despite their relatively good performance on well-controlled datasets, they are less useful for water body detection in real-world conditions.

Deep convolutional neural network (DCNN) models have become popular for water body detection in recent years [14], [15], [4], [3], [16]. DCNN-based semantic segmentation networks employed for remote-sensed water detection in urban hydrological applications include fully convolutional networks (FCNs) [17], upsampling pyramid networks (UPNs) [4] and DenseNet [18]. The advantage of these models is that they are able to extract more distinctive feature representations compared to traditional water index features, thus enabling improved water body detection.

Multi-spectral imagery should support further improved water body segmentation compared to using only RGB channels due to the additional information contained in the extra bands that cover a wider part of the electromagnetic spectrum, while the resulting higher-dimensional data can be reduced through appropriate methods [19]. However, in recent work [4], [20] the use of multiple bands did not prove to be very effective. In the Kaggle Satellite Imagery Feature Detection challenge [20], methods using all of the available 20 channels (1 panchromatic channel with pixel resolution of 0.31m, RGB channels (0.31m), 8 multi-spectral bands (1.24m) and 8 short-wave infrared bands (7.5m)) achieved only insignificant improvements compared to those employing RGB bands only. Since the different bands are of different resolutions, NIR and SWIR bands need to be upsampled to the same resolution as the panchromatic/RGB bands, while this interpolation process might compromise the information of the original features.

In this paper, we propose a novel Multi-Channel Water Body Detection Network (MC-WBDN) that exploits the potential of multi-spectral imagery to improve the performance of state-of-the-art DCNN models for water body segmentation. In our model, we use Sentinel-2 RGB, NIR and SWIR bands and design a multi-channel fusion module to deal with the

*Corresponding author.

This work was partially carried out when Kunhao Yuan and Xu Zhuang were with Chengdu UnionBigData Technology Co., Chengdu, China.

Kunhao Yuan, Gerald Schaefer, Lin Guan and Hui Fang are with the Department of Computer Science, Loughborough University, Loughborough, U.K. (e-mail:k.yuan, l.guan, h.fang@lboro.ac.uk.).

Jianxin Feng is with the Information Engineering College, Dalian University, China. (e-mail:fengjianxin863@163.com).

different image resolutions in order to eliminate the above-mentioned upsampling issue. In addition, we introduce a novel Enhanced Atrous Spatial Pyramid Pooling (EASPP) module to extract multi-receptive feature representations and Space-to-Depth (S2D)/Depth-to-Space (D2S) operations to replace the max pooling operation and upsampling process in order to preserve the saliency of the high-dimensional representations. Our experimental results convincingly show that we achieve significant improvements compared to state-of-the-art deep learning methods that employ either RGB or multi-spectral data.

Our contributions in this paper are as follows:

- A multi-channel fusion module is designed to fuse all bands in an end-to-end manner avoiding upsampling operations, so that the learned weights are more effective.
- An Enhanced Atrous Spatial Pyramid Pooling (EASPP) module is designed to extract multi-receptive features from multi-scale levels to obtain an enhanced representation.
- Space-to-Depth (S2D)/Depth-to-Space (D2S) operations are introduced to replace the max pooling and upsampling stages in order to preserve more features for segmentation.
- A comprehensive set of experiments, including an ablation study and a comparison to state-of-the-art methods, are conducted to confirm the effectiveness of our proposed method.
- Our annotated dataset is made publicly available[1] to the research community to allow further work in this area and to support the comparison of different approaches.

The remainder of the paper is organised as follows. Related work is discussed in Section II to provide some background of our proposed model. Section III describes the employed dataset, data augmentation and data pre-processing steps. Section IV then explains our proposed MC-WBDN model in detail, while in Section V experimental results, including an ablation study, are presented to demonstrate its effectiveness. Finally, Section VI concludes the paper and identifies future work.

II. Related work

A. Traditional index-based water detection

Index-based water body detection has been studied extensively since the commercialisation of remote sensing satellites [12], [21]. Various handcrafted features have been designed considering water body characteristics to label pixels into water or non-water categories. [9] proposes NDWI to extract vegetation liquid water based on the green and NIR channels from Landsat imagery, succeeding in suppressing background soil and terrestrial vegetation features by delineating open water features. In [11], the modified NDWI (MNDWI) replaces the green band with the middle infrared (MIR) band to further suppress built-up land noise, vegetation and soil noise, thus enhancing water region segmentation performance. In [12], a dual-coefficient index named AWEI (for Automated Water Extraction Index), is proposed to increase the contrast between water and other dark surfaces, while, more recently, [13], introduces the Pixel Region Index (PRI), a spatial feature index, to exploit the smoothness characteristics of local areas to improve the effectiveness of NDWI.

B. DCNNs for semantic segmentation

Semantic segmentation assigns a class label to each image pixel to support a high level semantic understanding of the image [22]. Traditional machine learning applications rely heavily on pre-defined feature descriptors to achieve pixel-wise classification [23], [24]. Since the introduction of the pioneering DCNN model for this, the fully convolutional network (FCN) [22], many network architectures, e.g., DenseNet [25] and ResNet [26], have been adopted, proposing various innovations such as the re-use of features from previous layers and mapping residuals in a deep model, to yield high segmentation accuracy. In [27], [28], the DeepLabV3 and DeepLabV3+ methods propose a spatial pyramid pooling module between encoder and decoder to take advantage of multi-scale features, while some of the latest methods, including SharpMask [29], U-Net [30], and RefineNet [31], embed hierarchical feature representations extracted from multiple layers of the encoders into their corresponding decoders for better segmentation.

DCNN models have also been deployed in remote sensing applications. Image segmentation of remotely sensed imagery is more challenging due to high intra-class variations and low sensor resolution [32]. To tackle these challenges, several strategies, including hierarchical feature representations [33], multi-modality [34], and fusion schemes [35], [36], [34], have been adopted in recent applications. In [37], a hybrid architecture based on SharpMask and RefineNet achieves the best performance on a 6-band multi-spectral imagery segmentation task due to its diversified feature representation. In [38], three CNN models are ensembled using Monte Carlo dropout uncertainty maps to outperform standard weight averaging for land cover mapping segmentation in urban areas. [39] employs a digital surface model (DSM) to use geometry information in order to improve the FCN network segmentation results of VHR remote sensed images.

C. DCNNs for water body detection

Various standard DCNN models have been adapted in water body detection applications. [17] uses an FCN model to extract the water body of Beijing's metropolitan area from VHR images collected by the GaoFen-2 satellite. In [16], a CRF-refined U-Net is proposed to process VHR images collected from both GaoFen-2 and WorldView-2 satellites. Additional elevation information from SAR images is exploited in [4] and a focal loss function is used to deal with the imbalanced categorical distributions in order to improve the segmentation accuracy of pixels located at boundaries.

III. Data and data preparation

A. Study area and data source

Our research area is Chengdu City and its suburban region (over 15k km^2) in Sichuan Province, China ("Sichuan" literally means "four rivers"). The motivation of our proposed

[1]https://github.com/SCoulY/Sentinel-2-Water-Segmentation

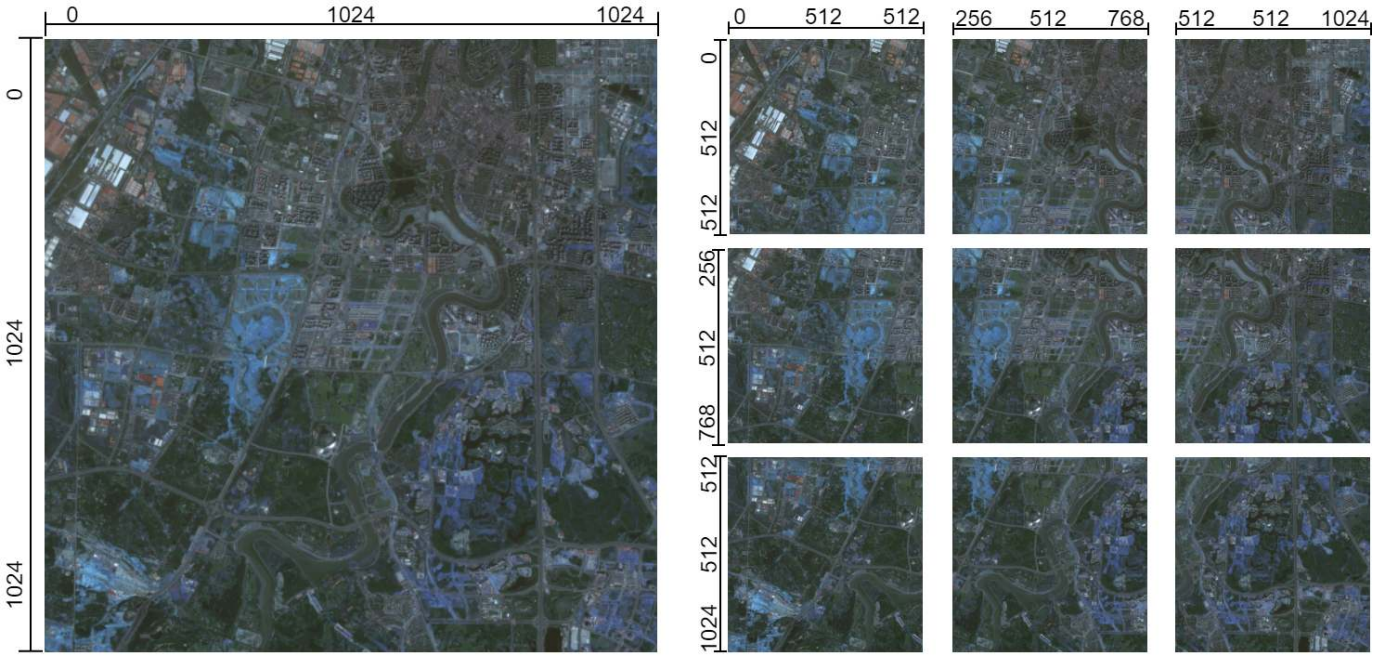

Fig. 1. An 1024×1024 image block (left) and the cropped patches extracted for training (right). The first and the third values on each line segment denote the start and end pixel locations in the original image block while the middle value denotes the length of the line segment.

algorithm is to facilitate timely monitoring and protection of the local water resource by analysing its dynamics at short intervals. Sentinel-2 satellite imagery of Chengdu City is collected for training and testing the proposed model. Details on the Sentinel-2 bands are given in Table I. According to [10], [11], [12], the bands most sensitive to water reflection are green, NIR, and MIR (SWIR in Sentinel-2). Consequently, we select bands 4, 3 and 2 (RGB) together with bands 8 (NIR) and 12 (SWIR) for our approach.

TABLE I
MULTI-BAND INFORMATION OF SENTINEL-2 DATA USED IN THE PAPER.

band	pixel resolution [m]	central wavelength [μm]
1 - Coastal aerosol	60	0.443
2 - Blue	10	0.490
3 - Green	10	0.560
4 - Red	10	0.665
5 - Vegetation Red Edge	20	0.705
6 - Vegetation Red Edge	20	0.740
7 - Vegetation Red Edge	20	0.784
8 - NIR	10	0.842
9 - Water vapour	60	0.945
10 - SWIR-Cirrus	60	1.375
11 - Water vapour	20	1.610
12 - Water vapour	20	2.190

The employed multi-spectral imagery of Chengdu City comprises a 16-bit raster image of size 20976×20982 pixels for R, G, B, and NIR bands (10m resolution), and of size 10488×10491 pixels for the SWIR band (20m resolution). The data used in this paper was retrieved from Sentinel-2 in April 2018. Additionally, we have downloaded two further batches of data, captured in late 2018 and early 2019, respectively, of the same area in order to be able to evaluate robustness to light and cloud variations. Since Chengdu is located in the Sichuan Basin, cloud cover is typically high. Thus, we merged images taken on sunny days across an entire month to create a high-quality near cloud-free dataset. We also performed atmospheric correction on the images using ArcGIS.

B. Data pre-processing and data augmentation

Before model training, data pre-processing and data augmentation steps are applied to enhance the model's effectiveness and speed up the computation. These include:

- *Image splitting:* the raster imagery is split into manageable image blocks in order to avoid large computational and memory requirements as well as to facilitate parallel computation. The input size of our proposed model is 512×512 pixels for NIR and RGB channels and 256×256 pixels for the SWIR channel. Instead of splitting the full multi-spectral image into patches of the required size, we first split it into blocks of 1024×1024 (NIR,RGB)/512×512 (SWIR) pixels. This configuration allows to set different splitting strategies for training and testing purposes. In the training stage, more samples are required to tackle overfitting problems. Therefore, an overlapping split of image blocks is introduced in order to generate more training samples as illustrated in Fig. 1. For testing, patches from randomly sampled non-overlapping blocks are used as input to our proposed model.
- *Cloud filtering and colour normalisation:* based on a preliminary analysis on spectral information for each band, a heuristic threshold of 3000 is used for both NIR and SWIR channels to filter out the remaining cloudy areas with values above the threshold capped. As illustrated in Fig. 2, the data distribution of each wavelength channel generally approximates a Gaussian distribution. Thus, we

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3098678, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

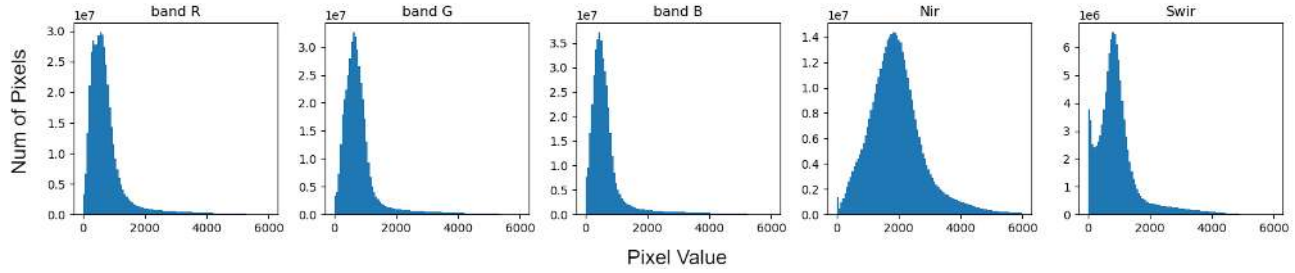

Fig. 2. Pixel intensity distributions of the used bands.

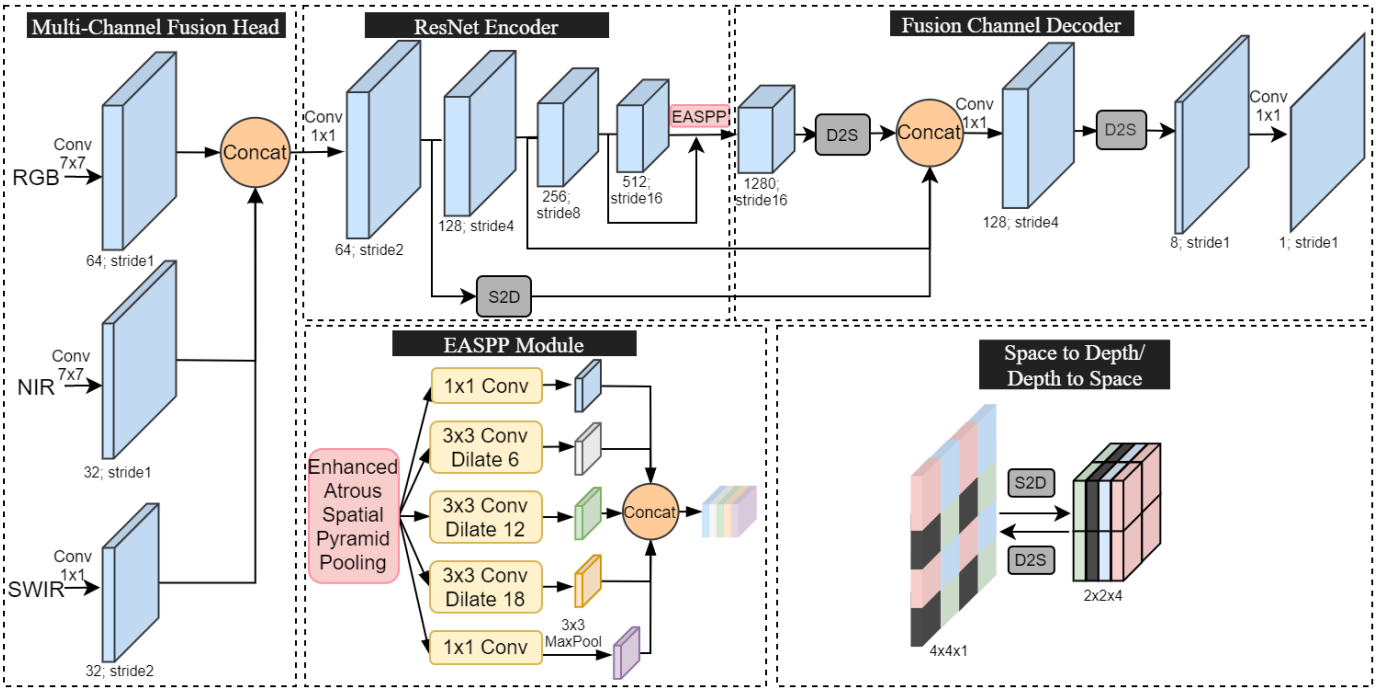

Fig. 3. The proposed MC-WBDN network architecture which adopts the popular encoder-decoder structure for semantic segmentation with a fusion head and works fully end-to-end. The baseline MC-WBDN model replaces S2D and D2S with corresponding pooling subsampling and bilinear upsampling. Residual connections between each convolutional blocks are omitted. The number below each block denotes the channels of feature maps whereas 'stride2' indicates 2 times downsampled resolution compared to original input. The output is a single-channel feature map of the same size as the input.

normalise the intensities X_i in each channel by their mean μ and standard deviation σ as

$$X_i = \frac{X_i - \mu(X_i)}{\sigma(X_i)}. \quad (1)$$

- *Image augmentation:* in addition to the original patches obtained from the splitting stage, we apply the following methods during the training stage for data augmentation: (1) a random horizontal/vertical flip with probability 0.5; (2) a clockwise 90-degree rotation with probability 0.5; and (3) a random HSV colour shift within a small range with probability 0.25.

IV. MC-WBDN MODEL

A. Model Architecture

The architecture of our proposed MC-WBDN model is illustrated in Fig. 3. The RGB channels, NIR channel and SWIR channel form the three input images. These are processed by their corresponding convolution kernels in the multi-channel fusion module. Feature maps of identical size are generated and concatenated in the fusion module and used as input to our backbone encoder-decoder network for pixel-level labelling. The Encoder network is a ResNet-34 model pre-trained on ImageNet [26] and the Fusion Channel Decoder is an enhanced DeepLabV3+ network that uses the fine grained feature maps produced by the EASPP and S2D/D2S modules. The detailed network configuration is given in Table II which lists the kernel width, kernel height, and number of kernels in each convolutional layer, together with the output sizes of the feature maps.

In a classical backbone encoder-decoder architecture, the encoder consecutively downsamples and diversifies the feature representations, while the decoder upsamples and maps them to their correspondent labels. Compared to this, our proposed model has two distinctive traits: (1) we replace all bi-linear upsampling operations in our decoder with Depth-to-Space

TABLE II
LAYER CONFIGURATION OF EMPLOYED NETWORK ARCHITECTURE. BN = BATCH NORMALISATION.

layer	MC-WBDN w/o S2D/D2S	MC-WBDN w/ S2D/D2S	output size		K_{width}	K_{height}	$K_{filters}$	non-linearities
fusion module	RGB		(256,256,64)		7	7	64	BN,Swish
	NIR		(256,256,32)		7	7	32	BN,Swish
	SWIR		(256,256,32)		1	1	32	BN,Swish
Concatenate			(256,256,128)					
Conv2d			(256,256,64)		1	1	64	BN,Swish
ResNet34-Block1			(256,256,64)		(3,3)	(3,3)	$64 \times 2 \times 3$	BN,ReLU
ResNet34-Block2			(128,128,128)		(3,3)	(3,3)	$128 \times 2 \times 4$	BN,ReLU
ResNet34-Block3			(64,64,256)		(3,3)	(3,3)	$256 \times 2 \times 6$	BN,ReLU
ResNet34-Block4			(32,32,512)		(3,3)	(3,3)	$512 \times 2 \times 3$	BN,ReLU
EASPP module			(32,32,1024)		(1,3,3,3,1)	(1,3,3,3,1)	512	BN
Upsample-4x		D2S+S2D	(128,128,1024)	(256,256,128)				
Concatenate			(128,128,1024+128)					
Conv2d			(128,128,128)		1	1	128	BN,Swish
Upsample-4x		D2S	(512,512,128)	(512,512,8)				
Conv2d			(512,512,16)	(512,512,1)	1	1	16	BN,Swish
Conv2d			(512,512,1)		1	1	1	

(D2S) operations. This allows for an improved information exchange between channels which proved effective in SENet [40] and ShuffleNet [41]. Table II highlights the two process pipelines, with and without S2D/D2S operations, while a performance comparison of the two structures is presented in our ablation study. (2) two extra bypasses from lower layers in the encoder are concatenated with dense feature maps given by an Enhanced Atrous Spatial Pyramid Pooling (EASPP) module in order to preserve more fine-grained context.

To ensure numerical stability and non-linear representation, Swish activation functions [42], which are differentiable when dealing with negative gradients and defined as $f(x) = x \cdot sigmoid(x)$, are used in both the fusion head and the decoder, while ReLU activation functions ($f(x) = \max(0, x)$) are employed in the encoder part. This preserves the representative features transferred from the pre-trained deep learning model. For testing, a sliding window prediction mechanism is employed where the central area is kept as the result from sliding a window along the satellite imagery.

In the following, we explain the main features and innovations of our proposed model.

1) Multi-channel fusion head: We fuse RGB channels with NIR and SWIR channels at the very beginning of the processing pipeline. For RGB and NIR channels, which are of the same resolution, we apply 7×7 convolution kernels to enlarge the receptive field and a stride of 2 to align the output size with SWIR, while for the lower resolution SWIR band, we apply 1×1 convolution kernels to densify its feature maps. The three outputs are then concatenated, followed by a 1×1 convolution to yield the channel combination used by the context encoder.

2) EASPP module: To extract distinctive features from multiple receptive fields we introduce an enhancement to Atrous Spatial Pyramid Pooling (ASPP) [36]. In our Enhanced ASPP (EASPP), we apply 1×1 convolution operations followed by a local max pooling to the feature maps from the previous layers, thus avoiding the up-sampling stage in the original ASPP. Intuitively, this modification adds a shortcut from previous layers and makes the trainable weights more effective. Our proposed EASPP module distils dense features from different scales of the input feature maps [36] by individual dilated convolutions [43], [44], [45] at different scales. These scales indicate the various region sizes in the feature maps that can be activated. Benefitting from the hierarchical structure of the receptive fields, the feature pyramid aggregates rich context information from the input. The multi-scale feature pyramid is concatenated and pruned by a 1×1 convolution to produce the output feature maps.

3) Space-to-Depth and Depth-to-Space: In a conventional DCNN pipeline, feature maps are processed with pooling operations in the encoder and upsampling operations in the decoder. These two operations however are sub-optimal since pooling operations discard detailed feature responses while the upsampling operations are non-trainable. Although transposed convolution operations complement upsampling schemes, they significantly increase the parameters in a DCNN [46].

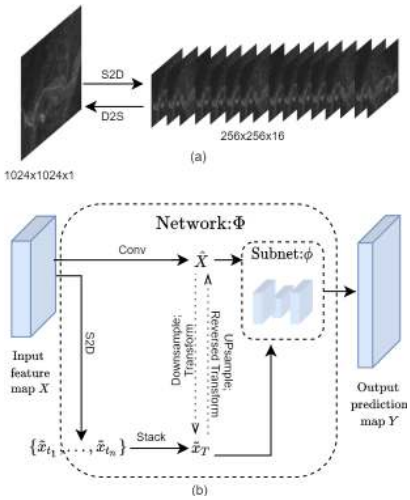

Fig. 4. (a) Visualisation of S2D and D2S. (b) The role of S2D in a network. Dashed arrows demonstrate the relationships between slices after S2D and the original branch but do not actually take effect when processing.

The use of Space-to-Depth (S2D) and Depth-to-Space (D2S) operations can alleviate these problems [47], [48]. As

illustrated in Fig. 4(a), S2D moves pixels from spatial locations to channel dimensions, while D2S is the inverse operation. With S2D operations, more local features can be preserved for the decoder process. Moreover, the S2D operation can be treated as an intra-model augmentation as shown in Fig. 4(b) and offers views of inputs with different pixel shifts. Vice versa, D2S is an alternative to transposed convolutions for up-sampling due with two advantages: (1) D2S is parameter-free while keeping all the responses from previous layers, and (2) it merges information across feature map channels to allow effective feature exchange instead of focussing on individual channels when applying transposed convolutions.

B. Loss functions

Our loss function comprises two terms, a pixel-wise loss term and a region-based loss term. For the pixel-wise term, we use the binary cross-entropy loss, calculated as

$$loss_{BCE} = -\frac{1}{|\Omega|}\sum_{i\in\Omega}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)], \quad (2)$$

where Ω denotes all pixels in the predicted map, y_i is the label of the i-th pixel (0 or 1), and $\hat{y}$ is the predicted probability of pixel i.

In contrast to the pixel-wise loss term, the region-based loss term focusses on optimising the smoothness of regions to improve the mIoU (mean Intersection over Union over all classes). Commonly used region-based loss terms include the Dice coefficient loss, Jaccard loss [49], and Lovász-Softmax loss [50]. The former two are more suitable for imbalanced data training, and we thus use only them. The Dice coefficient loss is defined as

$$loss_{Dice} = 1 - \frac{2\left|\hat{Y} \cap Y\right|}{\left|\hat{Y}\right| + |Y|}, \quad (3)$$

where $\hat{Y}$ represents the predicted map and Y the label mask, while the Jaccard loss is calculated as

$$loss_{Jaccard} = 1 - \frac{\left|\hat{Y} \cap Y\right|}{\left|\hat{Y} \cup Y\right|}. \quad (4)$$

The multi-task loss is then calculated as

$$loss = \alpha \cdot loss_{BCE} + (1-\alpha)\cdot loss_{Dice} \quad (5)$$

and

$$loss = \alpha \cdot loss_{BCE} + (1-\alpha)\cdot loss_{Jaccard}, \quad (6)$$

respectively, where $\alpha \in [0;1]$ is a weight to balance the two terms. A larger value of α will favour the binary cross-entropy loss and thus improve detection of small water bodies (e.g., paddy fields and streams) while a lower weight will enhance regional smoothness of the segmentation. The optimal setting of α is assessed in our ablation study.

The end-to-end training of the model is performed by backpropagation [51] through the loss function. After gradient calculation, the parameter set of the network is updated as

$$\theta = \theta - \epsilon\hat{g}, \quad (7)$$

where ϵ denotes the learning rate. We use Adam [52] as the optimiser.

V. Experimental results and analysis

A. Experimental setup

Our data is split into disjoint training, validation and test sets. From the original satellite imagery we first generate 441 blocks of size 1024×1024. After image splitting, which generates 9 patches from each 1024×1024 block, we thus have a total of 3969 image patches. 300 1024×1024 blocks are used to generate 2700 training patches, and we divide the remaining 141 blocks into a validation set of 33 blocks and 297 patches and a test set of 108 blocks and 972 patches. We further divide the training and validation sets into three folds (each of 300 blocks training and 33 block validation) to test the robustness of the proposed method. We report mIoU results in terms of average and standard deviation on the test set when trained using the three trained models.

We compare our proposed MC-WBDN model with commonly used RGB-based segmentation architectures and some of the latest multi-band methods. In particular, we use U-Net [30][2], D-LinkNet [53], vanilla Sharpmask [29], DeeplabV3+ [28] and the method by Kemker *at al.* [37], which merges the encoder structure of Sharpmask and the decoder structure of RefineNet[3], in our evaluation.

We train each deep learning model for a minimum of 100 and a maximum of 300 epochs with an early-stopping mechanism that terminates learning when performance on the validation set does not improve for five consecutive epochs.

[2]Our implementation of U-Net involves no pre-training.

[3]Our implementation uses the five bands mentioned, while the original method uses six spectral bands.

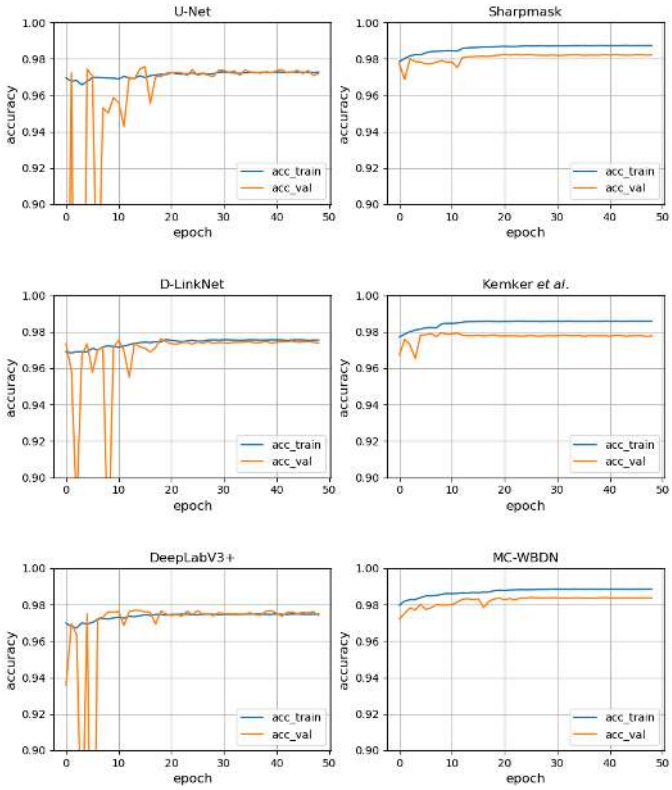

Fig. 5. Learning curves of different models.

TABLE III
MIOU RESULTS ON TEST SET FOR ALL METHODS.

	method	**bands**	**parameters/band** $[10^6]$	**ImageNet pre-trained**	**mIoU [%]**
traditional indices	NDWI [9]	Green+SWIR	n/a	n/a	1.77
	NDMI [10]	NIR+SWIR	n/a	n/a	3.28
	MNDWI [11]	Green+SWIR	n/a	n/a	10.44
RGB-based DCNNs	U-Net [30]	RGB	10.36	no	54.44 $\pm$ 6.02
	D-LinkNet [53]	RGB	10.37	yes	54.39 $\pm$ 1.60
	DeepLabV3+ [28]	RGB	8.66	yes	58.96 $\pm$ 9.58
multi-spectral DCNNs	Sharpmask [29]	RGB+NIR+SWIR	5.78	no	70.60 $\pm$ 2.30
	Kemker *et al.* [37]	RGB+NIR+SWIR	39.66	no	64.70 $\pm$ 4.11
	MC-WBDN	RGB+NIR+SWIR	6.85	yes	74.42 $\pm$ 0.65

We use an initial learning rate of $1e$-3, an end threshold of $1e$-8 and a decay rate of 0.5 in our experiments. Training takes, depending on the model, about 10-50 hours to convergence using 4 NVIDIA GTX1080ti GPUs with a batch size of 16. All the experiments were carried using the PyTorch deep learning framework [54]. We note that only Sharpmask, Kemker *et al.* and our MC-WBDN take multi-spectral images as input. In addition to deep learning models, we also evaluate some traditional index methods, namely NDWI, NDMI, and MNDWI, for which the established thresholds of 0.34, 0.3, and 0.1, respectively, are used.

B. Experimental results

We plot learning curves depicting accuracy over the first 50 training epochs in Fig. 5. From there, we can see that U-Net shows the lowest training and validation accuracy. Higher accuracies are achieved by D-LinkNet and DeepLabV3+, but it is clear that the multi-spectral models do better still, reaching both higher accuracies on both training and validation sets and more stable performance, indicating that RGB features are insufficient to successfully learn a more general water representation. Our proposed MC-WBDN approach outperforms all other models and yields the highest training accuracy of 0.988 as well as the highest validation accuracy of 0.984, while the training process takes less than 20 hours due to the simplicity and efficiency of our network architecture.

mIoU results, in terms of average and standard deviation on the test set are given in Table III for all evaluated models.

From there, it is immediately apparent that the results for the traditional water extraction indices are extremely inferior, even when compared to the worst performing deep learning method. While MNDWI improves upon NDWI and NDMI, the achieved mIoU of just over 10% is far too low to be useful. Also, since these models rely on (fixed) thresholds, they lack flexibility while the indices themselves exploit only linear relationships between the selected bands.

In contrast, deep learning-based approaches support non-linear representation ability and are able to learn useful features from large parameter spaces, leading to significantly better segmentation performance. Looking closer at the obtained results, we can see that generally approaches that use only the RGB bands are inferior to those that also incorporate NIR and SWIR bands to exploit the additional information contained there. Of the RGB models, the best results are obtained using DeepLabV3+, while U-Net and D-LinkNet yield relatively poor performance.

Similar to our proposed MC-WBDN approach, Sharpmask and Kemker *et al.* also use multi-spectral input (the identical bands in our experiments). The method by Kemker *et al.* fails to generalise well and consequently yields lower performance in comparison to Sharpmask. Our proposed MC-WBDN however clearly outperforms Sharpmask and all other evaluated methods, giving the best segmentation results with an mIoU of 74.42%, based on an equal weighting of pixel and region loss function terms and Jaccard loss for the latter. In addition, MC-WBDN also yields the lowest standard deviation and thus is the most robust of the evaluated DCNN models.

Table III also shows the number of trainable parameters per band. Due to heavy usage of transposed convolutions in the bottom-up phase, the number of parameters almost doubles for U-Net and D-LinkNet in comparison to Sharpmask which has the lowest number of parameters per band. By far the most parameters are used in the model by Kemker *et al.*, while the parameter space of our proposed MC-WBDN model is relatively small and only somewhat larger than that of Sharpmask.

Fig. 6 shows several typical test patches together with their ground truth segmentations and the outputs of the six deep learning models, while Fig. 7 gives results for further, more challenging, test patches under low lighting conditions.

As can be seen from these examples, the multi-spectral models such as Sharpmask, Kemker *et al.* and our MC-WBDN outperform RGB-only models (i.e., U-Net, D-LinkNet and DeepLabV3+). In particular, for areas that contain complex urban scenarios (such as row 6 in Fig. 6), RGB-only models tend to fail as they are unable to handle the wider range of colour shifts. Also, for patches containing both shallow and wide water bodies (e.g., rows 3 and 5 in Fig. 6), our MC-WBDN is able to deliver improved detection due to both the additional information in the NIR and SWIR bands and the multi-scale features leaned by the EASPP module.

While the performance improvement of our MC-WBDN model is relatively minor for the samples in Fig. 6, it becomes more apparent for areas under low lighting conditions such as the examples shown in Fig. 7. In particular for the patch in the second row, which requires consistent prediction of scattered water ponds, we can notice a vast improvement.

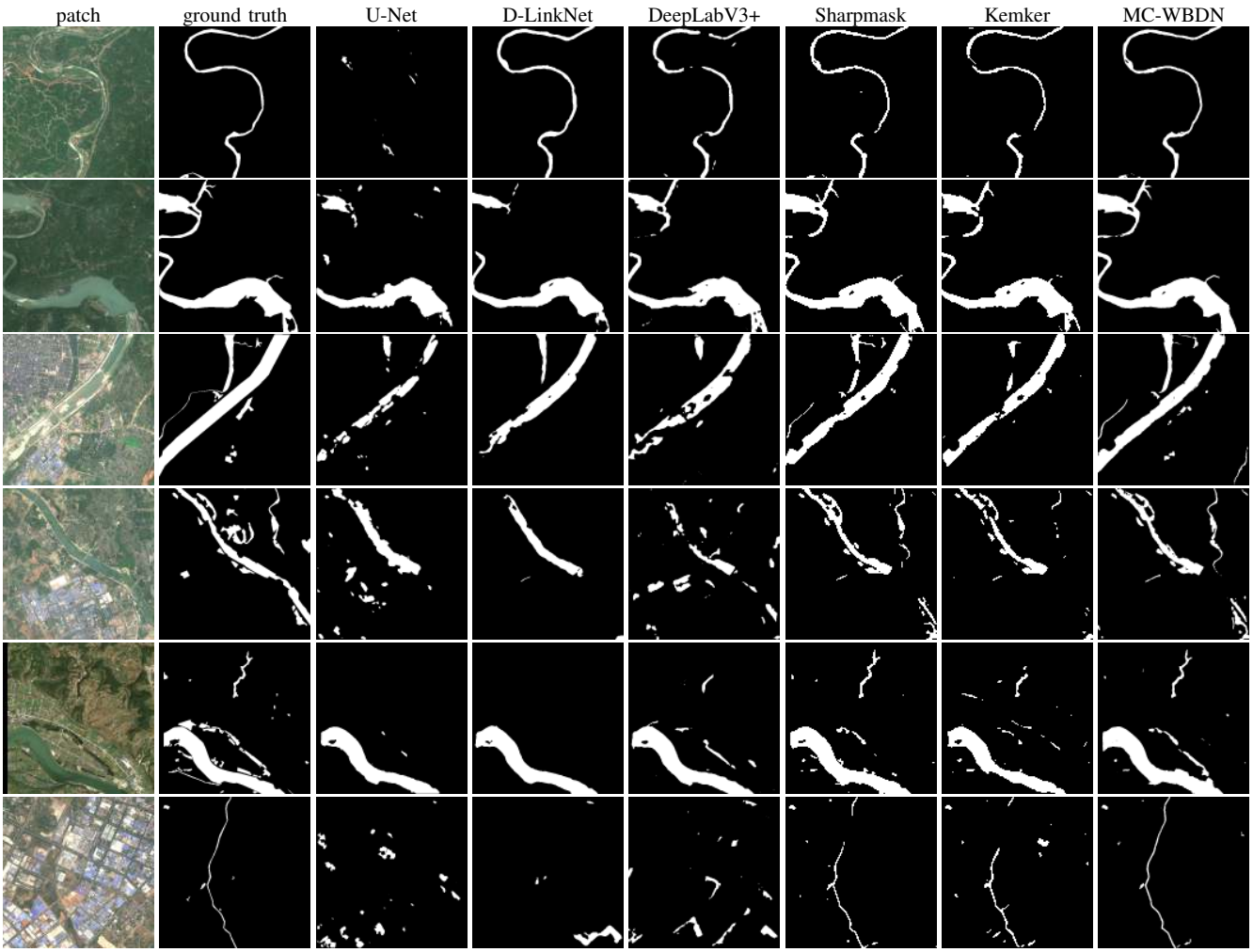

Fig. 6. Test results for different models on example patches.

C. Ablation study

We perform a thorough ablation study where we investigate the effect of each introduced component. The results are given in Table IV which shows the obtained performances (based on the models trained on the first partition of the dataset) when employing multi-channel fusion versus standard pan-sharpening, employing our proposed EASPP module versus standard ASPP, and using D2S/S2D versus standard bilinear interpolation.

TABLE IV
RESULTS OF ABLATION STUDY (BASED ON $\alpha = 0.1$ AND DICE REGION LOSS [55]). DEFAULT PROCESSING (THE NOES IN THE TABLE) INVOLVES PAN-SHARPENING, ASPP, AND BILINEAR INTERPOLATION.

MC fusion	EASPP	D2S/S2D	mIoU [%]
no	no	no	67.82
yes	no	no	70.41
no	yes	no	68.60
no	no	yes	69.49
yes	yes	no	72.59
no	yes	yes	69.91
yes	no	yes	71.41
yes	yes	yes	74.14

As we can see from Table IV, the baseline results are relatively modest with an mIoU of 67.82, while introduction of each component (MC fusion, EASPP, D2S/S2D) on its own is shown to lead to an improvement. It is however the three working together in tandem that really boosts the water detection performance, to an mIoU of 74.14, and does so by more than the sum of the individual improvements, thus confirming the effectiveness of our proposed model and its careful design.

D. Loss function evaluation

As explained in Section IV-B, our loss function comprises a pixel-based component and a region-based component while the latter is based on either Dice coefficient loss or Jaccard loss. Tuning the α parameter that balances the two components, one can thus put more emphasis on pixel- or region-based labelling. We evaluate three different settings, namely $\alpha = 0.1$, $\alpha = 0.5$, and $\alpha = 0.9$, together with the two region-based loss terms and show the obtained results (again, based on models trained on the first partition of the dataset) in Table V.

Looking at the obtained results, we notice that better performance is achieved using the Jaccard loss compared to the Dice coefficient loss, which is not surprising since the standard performance measure of mIoU corresponds to the Jaccard

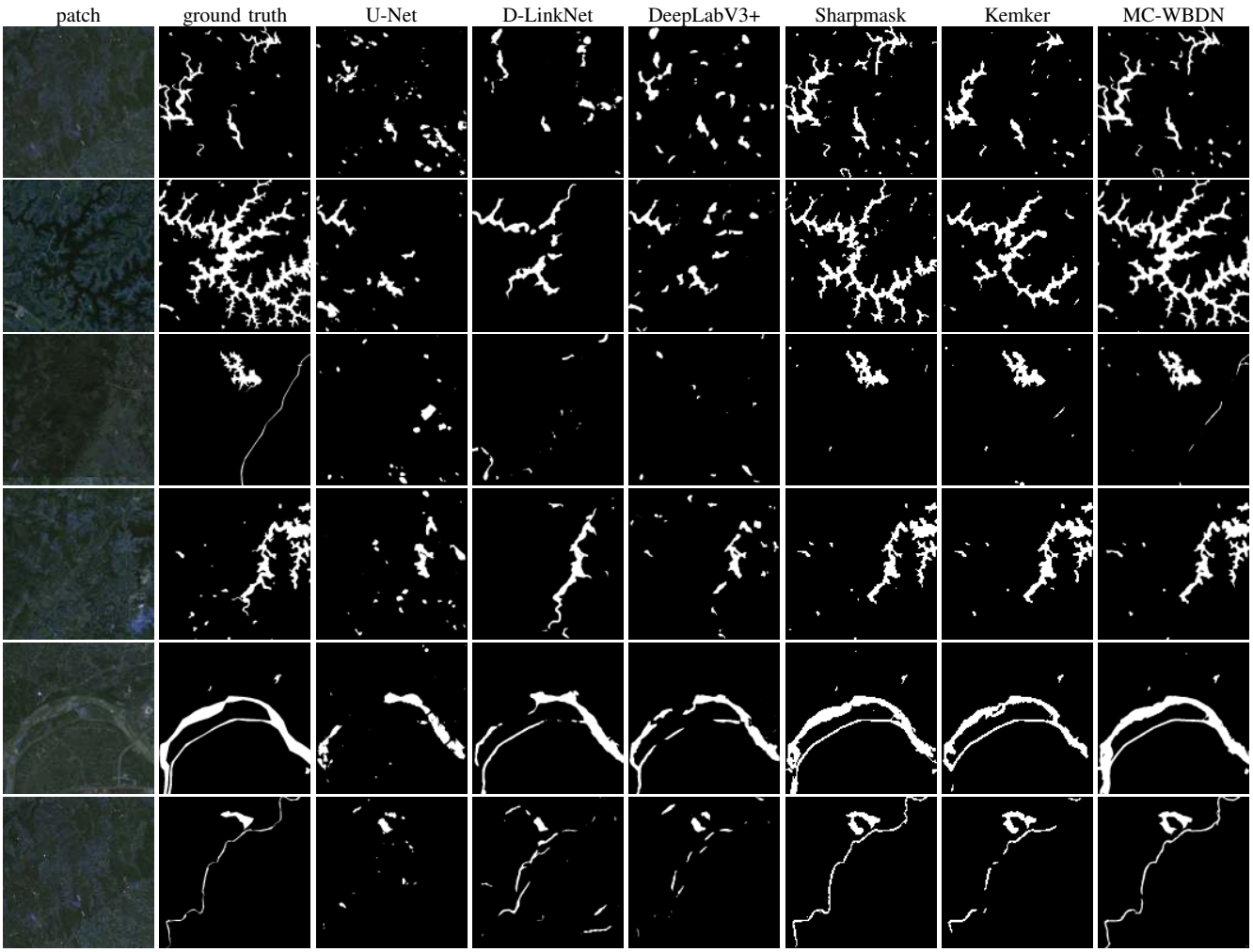

Fig. 7. Test results for different models on dark example patches.

TABLE V
RESULTS FOR DIFFERENT LOSS FUNCTION SETUPS.

region loss	α	mIoU [%]
Dice	0.1	74.14
Dice	0.5	73.96
Dice	0.9	74.80
Jaccard	0.1	75.06
Jaccard	0.5	75.13
Jaccard	0.9	74.71

index. The best weighting between pixel and region (Jaccard) loss is obtained by setting α to 0.5, which justifies the setup of our loss function. A best test mIoU of 75.13%, is obtained based on the first training-validation partition. After repeating training on the three partitions while using Jaccard region loss and $\alpha = 0.5$, an average mIoU of 74.42% on the test set is achieved, which is also the result reported in Table III.

E. Robustness against light and weather variations

One of the challenges of automated remote sensing is that the captured information of the same area can vary drastically due to environmental changes such as differing lightning and weather conditions. To evaluate robustness against varying conditions, we measure the performance of the various models on image patches of the same area as the original test samples but taken at different times (in late 2018 and in early 2019).

Fig. 8 shows a collection of sample patches for all three timestamps together with the water areas detected by our MC-WBDN model. As we can see, the variations in terms of colour shifts and cloud cover are quite apparent. In addition, we can notice some artefacts that come from the satellite built-in pre-processing and lead to rather different appearances within the same patch such as in the two middle patches for timestamp 3. Despite these difficulties, the performance of MC-WBDN is relatively consistent including for the very challenging patch at the top right.

Fig. 9 shows the results obtained by all deep learning models for the first area patch of Fig. 8. From there, we can observe that RGB-only models are greatly affected when lightning condition change and in particular fail completely for the cloudy scenario for the third timestamp. In contrast, the multi-band models exhibit improved robustness due to their ability to incorporate information from the NIR and SWIR bands also. MC-WBDN gives the best results across the three timestamps, followed by Sharpmask.

Table VI gives the results over all test patches for all three

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3098678, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

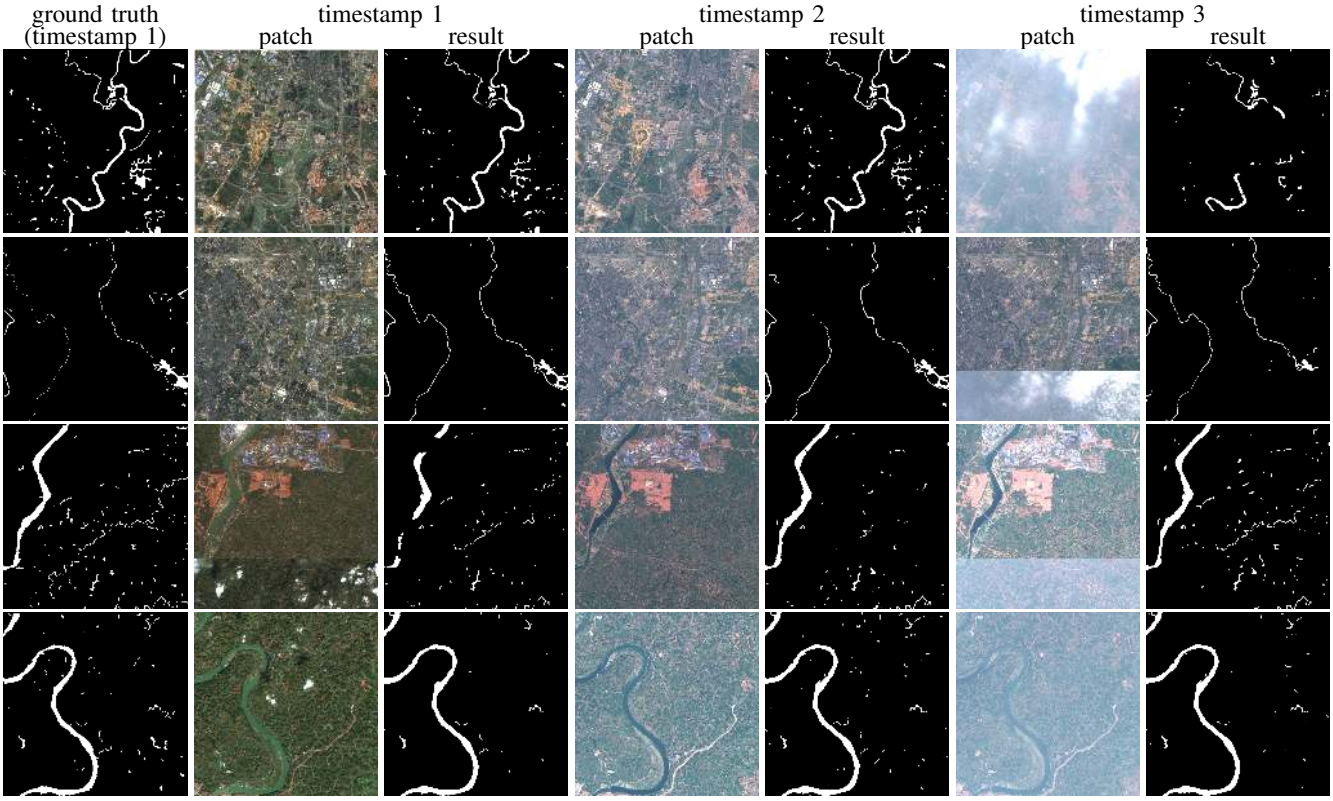

Fig. 8. MC-WBDN water body detection result examples across different timestamps

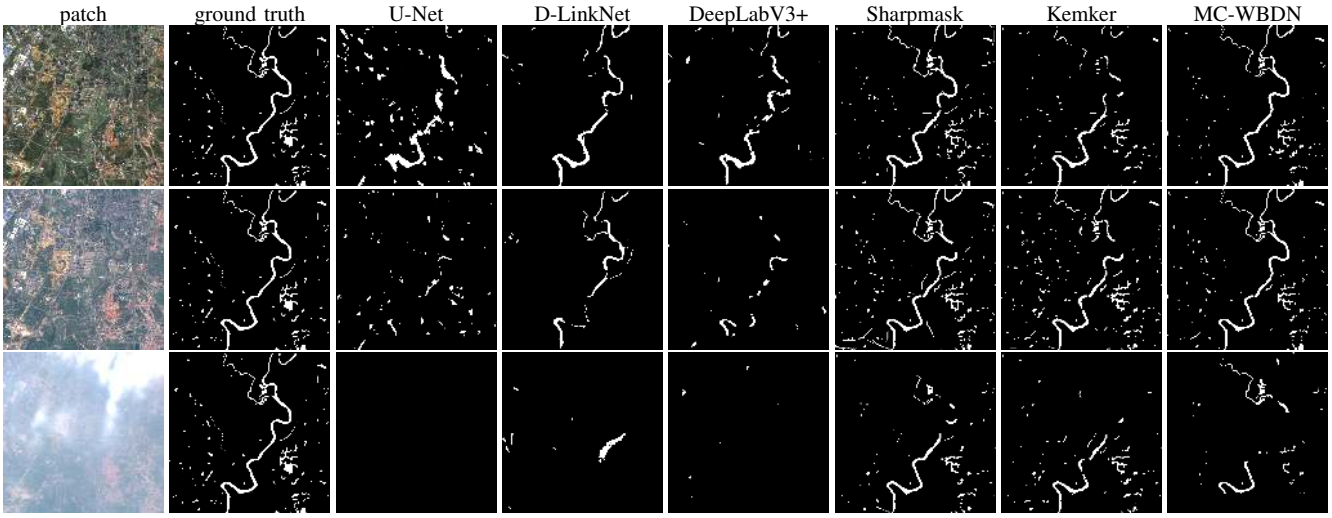

Fig. 9. Results on sample patch across three timestamps for all models.

timestamps and for all deep learning models (based on the models trained on the first partition of the dataset). From there it is clear that the RGB-only models fail to generalise well for the other two timestamps. For example, for DeepLabV3+, the mIoUs for timestamps 2 and 3 are almost 20 points lower than for timestamp 1. Significantly better results are achieved by the multi-band methods of Sharpmask and Kemker *et al.* which clearly outperform the RGB-only networks. The overall best performance across the different timestamps, with an average mIoU of 73.56, is obtained by MC-WBDN which thus confirms that our proposed model does not only outperform the other ones but also allows for water body detection that is robust with respect to light and weather conditions.

F. Discussion

In this paper, we design a novel effective DCNN model for water segmentation from satellite imagery which can be trained in an end-to-end fashion through backpropagation. Deep learning approaches allow for adaptive training on large and varied datasets, in contrast to traditional index-based water extraction methods that work only within a small range but largely fail in open areas and complex scenes due to the

TABLE VI
RESULTS, IN TERMS OF MIOU [%], ACROSS ALL TEST PATCHES FOR THE THREE DIFFERENT TIMESTAMPS.

	timestamp 1	**timestamp 2**	**timestamp 3**	**average**
U-Net	61.39	50.07	50.06	53.84
D-LinkNet	52.56	56.31	55.68	54.85
DeepLabV3+	70.20	51.76	51.59	57.85
Sharpmask	68.47	67.90	72.44	69.60
Kemker *et al.*	59.98	64.89	66.69	63.85
MC-WBDN	75.13	72.52	73.03	73.56

difficulty in selecting appropriate thresholds and inability to learn non-linear feature representations.

Instead of upsampling satellite bands captured at lower resolutions, our MC-WBDN approach directly fuses multi-band inside the network. This fusion head is then connected to an elegant state-of-the-art semantic segmentation DeepLabV3-like architecture, resulting in only a small increase of tuneable parameters but yielding noticeable performance increases compared to standard pan-sharpening as reported in Table IV. Compared with RGB-only data, inclusion of multi-spectral bands (in our approach NIR and SWIR bands) provides additional useful information for water extraction, leading to more accurate segmentation of water bodies, although further work to investigate the contribution of the additional bands over RGB-only data should prove useful in order to define a more detailed relationship between the bands and object features for water detection.

The proposed EASPP module is better capable of integrating representations from previous convolutional layers, leading to a significant performance boost as can be seen from Table IV. In addition, we also introduce Space-to-Depth/Depth-to-Space operations to enhance the reconstruction performance in the bottom-up phase of our model. Benefitting from the ability to exchange arbitrary pixel information from feature map channel dimensions to spatial dimensions and vice versa, this supports improved detection as confirmed in Table IV, while the overall MC-WBDN provides better generalisation ability and excellent water body detection performance, also in comparison to previous work, as shown in Table III and illustrated by the examples in Figs. 6 and 7. The impact of S2D/D2S can be seen as a combination of lower resolution feature maps and a splitting of the high-resolution feature map. Currently, this is only supported for image dimensions that are a power of 2. With an arbitrary sampling rate, such a reconstruction from dense feature maps using D2S could potentially replace the 1×1 convolutions that are currently dominant but require a higher memory allocation [47], [48].

Emphasis on pixelwise and region-based classification respectively is possible through adjustment of the weight parameter α in the loss function. Overall, our MC-WBDN model reaches the best performance by equally weighting the two loss terms as shown in Table V, thus paying equal attention to pixelwise and region-based classification. Further work can focus to identify if skeleton features are favourable for e.g. rivers while fine-granularity features are beneficial for larger water bodies such as lakes.

Consistent prediction on samples taken at different timestamps demonstrate the advantage of our MC-WBDN model as shown in Figs. 8 and 9 and Table VI. In contrast, other models fail to provide consistently high detection ability across timestamps, in particular when light and weather conditions vary more dramatically.

As mentioned in Section II-C, traditional baseline methods, e.g. FCN, U-Net along with refinement modules such as CRF, have been previously used in water body detection research. In addition, we also compare our proposed method with other generic methods such as Sharpmask and Kemker et al.'s work, since these methods have reported better performance in various applications compared to FCN and U-Net. These benchmarking methods also share their open-source code, thus allowing for objective and reproducible performance comparison. In future work, we plan to investigate more algorithms including further methods that been specifically developed for water body detection.

VI. CONCLUSIONS

Motivated by the success of deep learning methods and their applications to remote sensing, in this paper, we have introduced a novel approach to satellite-based water body extraction, accomplished through an effective deep convolutional neural network that incorporates several contributions. While RGB, being the basis of both the human visual system and common camera systems, has been frequently used for remote sensed analysis, we demonstrate that additional wavelength bands (NIR and SWIR) allow for improved segmentation. Given Sentinel-2 satellite data, we effectively exploit its multi-spectral information to aid our network model in successfully recognising water areas. Information from bands captured at different resolutions is appropriately fused directly in the network avoiding the need for image interpolation methods. We also incorporate Space-to-Depth/Depth-to-Space operations which are memory efficient and allow to retain better features, while we have presented an Enhanced ASPP to appropriately extract multi-receptive features from multiple scales. Experimental results have demonstrate excellent water detection capability of our MC-WBDN model, outperforming other evaluated models including traditional water detection indices and state-of-the-art deep models based on RGB and multi-spectral input, as well as showing improved robustness against light and weather variations. In future work, we aim to further use the proposed method in applicable hydrological studies.

ACKNOWLEDGEMENTS

We would like to thank the government of Chengdu City for partly funding this research, and also acknowledge the European Space Agency for providing open access to their Sentinel-2 satellite data.

REFERENCES

[1] Z. Shao, H. Fu, D. Li, O. Altan, and T. Cheng, "Remote sensing monitoring of multi-scale watersheds impermeability for urban hydrological evaluation," *Remote Sensing of Environment*, vol. 232, p. 111338, 2019.

[2] X. Wang and H. Xie, "A review on applications of remote sensing and geographic information systems (GIS) in water resources and flood risk management," *Water*, vol. 10, p. 608, 05 2018.

[3] Z. Miao, K. Fu, H. Sun, X. Sun, and M. Yan, "Automatic water-body segmentation from high-resolution satellite images via deep networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 4, pp. 602–606, 2018.

[4] J. Zhang, M. Xing, G.-C. Sun, J. Chen, M. Li, Y. Hu, and Z. Bao, "Water body detection in high-resolution SAR images with cascaded fully-convolutional network and variable focal loss," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[5] A. Ben Hamida, A. Benoit, P. Lambert, L. Klein, C. Ben Amar, N. Audebert, and S. Lefèvre, "Deep learning for semantic segmentation of remote sensing images with rich spectral content," in *IEEE International Geoscience and Remote Sensing Symposium*, 2017, pp. 2569–2572.

[6] M. Wurm, T. Stark, X. X. Zhu, M. Weigand, and H. Taubenböck, "Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 150, pp. 59–69, 2019.

[7] P. Thanh Noi and M. Kappas, "Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery," *Sensors*, vol. 18, no. 1, p. 18, 2018.

[8] G. Chen, X. Zhang, Q. Wang, F. Dai, Y. Gong, and K. Zhu, "Symmetrical dense-shortcut deep fully convolutional networks for semantic segmentation of very-high-resolution remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 5, pp. 1633–1644, 2018.

[9] S. McFeeters, "The use of normalized difference water index (NDWI) in the delineation of open water features," *International Journal of Remote Sensing*, vol. 17, pp. 1425–1432, 05 1996.

[10] B.-C. Gao, "Normalized difference water index for remote sensing of vegetation liquid water from space," in *Imaging Spectrometry*, vol. 2480. International Society for Optics and Photonics, 1995, pp. 225–237.

[11] H. Xu, "Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery," *International Journal of Remote Sensing*, vol. 27, no. 14, pp. 3025–3033, 2006.

[12] G. L. Feyisa, H. Meilby, R. Fensholt, and S. R. Proud, "Automated water extraction index: A new technique for surface water mapping using landsat imagery," *Remote Sensing of Environment*, vol. 140, pp. 23–35, 2014.

[13] Y. Zhang, X. Liu, Y. Zhang, X. Ling, and X. Huang, "Automatic and unsupervised water body extraction based on spectral-spatial features using gf-1 satellite imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 6, pp. 927–931, 2018.

[14] J. Geng, J. Fan, H. Wang, X. Ma, B. Li, and F. Chen, "High-resolution sar image classification via deep convolutional autoencoders," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 11, pp. 2351–2355, 2015.

[15] Y. Zhou, H. Wang, F. Xu, and Y.-Q. Jin, "Polarimetric SAR image classification using deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1935–1939, 2016.

[16] W. Feng, H. Sui, W. Huang, C. Xu, and K. An, "Water body extraction from very high-resolution remote sensing imagery using deep U-Net and a superpixel-based conditional random field model," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 4, pp. 618–622, 2018.

[17] L. Li, Z. Yan, Q. Shen, G. Cheng, L. Gao, and B. Zhang, "Water body extraction from very high spatial resolution remote sensing data based on fully convolutional networks," *Remote Sensing*, vol. 11, no. 10, p. 1162, 2019.

[18] G. Wang, M. Wu, X. Wei, and H. Song, "Water identification from high-resolution remote sensing images based on multidimensional densely connected convolutional neural networks," *Remote Sensing*, vol. 12, no. 5, p. 795, 2020.

[19] K. Makantasis, A. Doulamis, N. Doulamis, and A. Voulodimos, "Common mode patterns for supervised tensor subspace learning," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2927–2931.

[20] Kaggle, "Dstl satellite imagery feature detection, 2017," https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection , 2017, [Online; accessed 25-March-2020].

[21] J. Mukherjee, J. Mukherjee, and D. Chakravarty, "Automated seasonal separation of mine and non mine water bodies from landsat 8 oli/tirs using clay mineral and iron oxide ratio," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2550–2556, 2019.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[23] J. Guo, H. Zhou, and C. Zhu, "Cascaded classification of high resolution remote sensing images using multiple contexts," *Information Sciences*, vol. 221, pp. 84–97, 2013.

[24] M. Wang, Y. Wan, Z. Ye, and X. Lai, "Remote sensing image classification based on the optimal support vector machine and modified binary coded ant colony optimization algorithm," *Information Sciences*, vol. 402, pp. 50–68, 2017.

[25] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *arXiv preprint arXiv:1404.1869*, 2014.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[27] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

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

[29] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *European Conference on Computer Vision*, 2016, pp. 75–91.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[31] G. Lin, A. Milan, C. Shen, and I. D. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5168–5177, 2017.

[32] L. Chan, M. S. Hosseini, and K. N. Plataniotis, "A comprehensive analysis of weakly-supervised semantic segmentation in different image domains," *International Journal of Computer Vision*, pp. 1–24, 2020.

[33] A. Tao, K. Sapra, and B. Catanzaro, "Hierarchical multi-scale attention for semantic segmentation," *arXiv preprint arXiv:2005.10821*, 2020.

[34] C. Peng, Y. Li, L. Jiao, Y. Chen, and R. Shang, "Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 8, pp. 2612–2626, 2019.

[35] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 9, pp. 3252–3261, 2018.

[36] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[37] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 60–77, Nov. 2018. [Online]. Available: https://doi.org/10.1016/j.isprsjprs.2018.04.014

[38] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1–9.

[39] W. Sun and R. Wang, "Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 3, pp. 474–478, 2018.

[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[41] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.

[42] P. Ramachandran, B. Zoph, and Q. Le, "Searching for activation functions. arxiv 2017," *arXiv preprint arXiv:1710.05941*, 2017.

[43] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *CoRR*, vol. abs/1511.07122, 2016.

[44] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 9, pp. 3252–3261, 2018.
[45] M. Lan, Y. Zhang, L. Zhang, and B. Du, "Global context based automatic road segmentation via dilated convolutional neural network," *Information Sciences*, vol. 535, pp. 156–171, 2020.
[46] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya, "Learning and incorporating shape models for semantic segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 203–211.
[47] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," *CoRR*, vol. abs/1605.08803, 2017.
[48] T.-J. Yang, M. D. Collins, Y. Zhu, J. Hwang, T. Liu, X. Zhang, V. Sze, G. Papandreou, and L.-C. Chen, "DeeperLab: Single-shot image parser," *CoRR*, vol. abs/1902.05093, 2019.
[49] J. Cai, L. Lu, Y. Xie, F. Xing, and L. Yang, "Improving deep pancreas segmentation in CT and MRI images via recurrent neural contextual learning and direct loss function," *arXiv preprint arXiv:1707.04912*, 2017.
[50] M. Berman, A. R. Triki, and M. B. Blaschko, "The Lovasz-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. [Online]. Available: https://doi.org/10.1109/cvpr.2018.00464
[51] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognitive Modeling*, vol. 5, 1988.
[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.
[53] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 192–1924, 2018.
[54] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
[55] W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, and X. Xie, "AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy," *Medical Physics*, vol. 46, no. 2, pp. 576–589, 2019.

Kunhao Yuan received his B.Sc. and M.Sc. degree from Northeastern University, China, 2017 and Loughborough University, 2018 respectively. From later 2018 to first quarter of 2020, he was an algorithmic engineer in UnionBigData, Chengdu, China. He is now a PhD student at Loughborough University. His research interests include computer vision, self-supervised learning and representation learning.

Xu Zhuang received his B.S. degree in Computer Science from Southwest Jiaotong University (SWJTU), Chengdu, China, in 2006, and is currently pursuing the Ph.D. degree in the Software Engineering Laboratory in Southwest Jiaotong University (SWJTU), Chengdu, China. His research interests include data mining, data hiding and information security.

Gerald Schaefer gained his PhD in Computer Vision from the University of East Anglia. He worked at the Colour & Imaging Institute, University of Derby, in the School of Information Systems, University of East Anglia, in the School of Computing and Informatics at Nottingham Trent University, and in the School of Engineering and Applied Science at Aston University before joining the Department of Computer Science at Loughborough University. His research interests are mainly in the areas of computer vision, colour image analysis, medical imaging, and computational intelligence. He has published extensively in these areas with a total publication count of about 500, has been invited as keynote or tutorial speaker to numerous conferences, is the organiser of various international workshops and special sessions at conferences, and the editor of several books, conference proceedings and special journal issues.

Jianxin Feng received a Ph.D (2005) degree from Northeastern University, China. She was teacher in the Information Science and Engineering Institution of Northeastern University from 1999 to 2012. She is currently an associate professor at the Information Engineering College of Dalian University, China. She was a visiting scholar in the Computer Science department of Liverpool John Moores University from 2018 to 2019. Her current research interests include network optimization, wireless communication, intelligent control and image processing.

Lin Guan is a Senior Lecturer in the Department of Computer Science at Loughborough University. Her research interests focus on performance modelling/evaluation of heterogeneous computer networks and systems (or system of systems); QoX-QoS (Quality of Service, Quality of Resilience, Quality of Experience) analysis, provisioning and enhancements; caching, edge/fog Computing; vehicular ad-hoc networks (VANET); software defined networks (SDN); cloud computing and security, mobile computing, wireless and wireless sensor networks; multimedia systems and Model Based System Engineering (MBSE) with QoS attributes. She has published over 100 journal and conference papers and she has been serving as guest editor for several international journals, such as those published by Elsevier and Springer. She is currently on Editorial Board of Elsevier Journal of Systems and Software (ranked #2 for SE venues in Google Scholar) and Editor in Elsevier Simulation Modelling Practice and Theory. During her PhD, she was awarded the British Federation of Women Graduates Foundation Main Grant in 2004. She then held two EPSRC/industry CASE awards, one EPSRC/BAE EngD projects and two industrial subcontracts on feasibility study and consultancy. She received a prestigious award as Royal Society Industry Fellow and EPSRC KTA project. One of her current projects works on EPSRC/Rolls Royce funded 4 years project Model Based System Engineering (MBSE) with QoS Attributes.

Hui Fang received the B.S. degree from the University of Science and Technology, Beijing, China, in 2000 and the Ph.D. degree from the University of Bradford, U.K., in 2006. He is currently with the Computer Science Department at Loughborough University. Before, he has carried out research at several world-leading universities, such as University of Oxford and Swansea University. His research interests include computer vision, image/video processing, pattern recognition, machine learning, data mining, scientific visualisation, visual analytics, and artificial intelligence. Recently, he was awarded several grants as PI and co-PI, including Innovate UK funded "An agent-based modelling solution for reliable decision making in crisis and market turmoil in consumer retail", EPSRC funded "RAMP VIS: Making Visual Analytics an Integral Part of the Technological Infrastructure for Combating COVID-19", and NIHR funded "Computer vision to automatically monitor urine output". During his career, he has published more than 60 journal and conference papers.