Large-Scale Semantic 3D Reconstruction: Outcome of the 2019 IEEE GRSS Data Fusion Contest - Part A

Saket Kunwar, *Member, IEEE*, Hongyu Chen, Manhui Lin, Hongyan Zhang, *Senior Member, IEEE*, Pablo D'Angelo, Daniele Cerra, Seyed Majid Azimi, Myron Brown, *Senior Member, IEEE*, Gregory Hager, *Senior Member, IEEE*, Naoto Yokoya, *Member, IEEE*, Ronny Hänsch, *Senior Member, IEEE*, and Bertrand Le Saux, *Member, IEEE*

***Abstract*—In this paper, we present the scientific outcomes of the 2019 Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society. The 2019 Contest addressed the problem of 3D reconstruction and 3D semantic understanding on a large scale. Several competitions were organized to assess specific issues, such as elevation estimation and semantic mapping from a single view, two views, or multiple views. In this Part A, we report the results of the best-performing approaches for semantic 3D reconstruction according to these various set-ups, while 3D point cloud semantic mapping is discussed in Part B [1].**

***Index Terms*—Image analysis and data fusion, data fusion contest, stereo, multi-view, 3D reconstruction, height estimation, elevation model, point-cloud, semantic labeling, semantic mapping, classification, LiDAR, deep learning, convolutional neural networks.**

I. Introduction

One of the challenges inherent in Earth observation is to add a new dimension to the representation of the world. Multiple 2D imagery, with various sensors and resolutions, are currently available with which the surface of the Earth can be observed from above. However, for critical applications such as flight management, urban planning, and the environmental monitoring of forests, floods, and landslides, 3D models of the ground are significant sources of information.

Capturing this 3D information on a large scale is extremely challenging. Two approaches are currently used: active and passive. Active methods include Light Detection and Ranging (LiDAR) acquisition, which is primarily carried out using airborne sensors in large aerial laser scanning campaigns [2]. Satellite LiDARs, such as the Geoscience Laser Altimeter System (GLAS) instrument on IceSAT (launched in 2003), are also in operation but with lower-resolution products. Passive approaches include *structure from motion* and *multi-view stereo* and leverage multiple optical images that correspond to the same ground site in order to estimate common 3D points. These approaches are cheaper, yield high-resolution and accurate elevation models, benefit from developments that span over four decades [3], and include various estimation models that can be used by all the different satellite generations [4]–[6]. The appeal of using passive methods has increased significantly with the unprecedented number of imaging satellites currently in orbit. This led to the development of a benchmark on multi-view stereo 3D mapping [7] by the Johns Hopkins University (JHU) Applied Physics Laboratory (APL) in 2016, using data consisting of 47 WorldView-3 images of the same area over San Fernandino (Argentina), with airborne LiDAR used to define the reference data. The winning approach of this challenge used multiple two-view stereo methods to generate 3D models, which were then fused [8].

A new benchmark was co-organized by JHU/APL with the Image Analysis and Data Fusion Technical Committee (IADF TC) of the IEEE Geoscience and Remote Sensing Society (GRSS) in 2019, on the topic of semantic 3D with two sites and more images (69 overall), while addressing additional scientific issues. The IADF TC is an international network of scientists working on Earth observation, geospatial data fusion, and algorithms for image analysis. It aims at connecting people and resources, educating students and professionals, and promoting theoretical advances and best practices in image analysis and data fusion. The IADF TC has coordinated a challenge in order to foster ideas and progress in remote sensing every year since 2006, distributing novel data and benchmarking analysis methods, known as the Data Fusion Contest (DFC) [9]–[21]. The 2019 DFC (DFC19) aimed at large-scale semantic 3D reconstruction, encompassing 3D

Manuscript received xxx 2020;

S. Kunwar is with nestAI, Nepal (e-mail: saketkunwar2005@gmail.com).

H. Chen, M. Lin and H. Zhang are with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, 430079 Wuhan, China (e-mails: {hongyuchen, mhlin425, zhanghongyan}@whu.edu.cn).

P. D'Angelo, D. Cerra, S. M. Azimi are with the German Aerospace Center (DLR), 82234 Weßling, Germany (e-mails: pablo.angelo@dlr.de; daniele.cerra@dlr.de; seyedmajid.azimi@dlr.de).

M. Brown is with the Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland, USA (e-mail: myron.brown@jhuapl.edu).

G. Hager is with the Johns Hopkins University, Baltimore, Maryland, USA (e-mail: hager@cs.jhu.edu).

N. Yokoya is with the Graduate School of Frontier Sciences, the University of Tokyo, 277-8561 Chiba, Japan, and also with the RIKEN Center for Advanced Intelligence Project, 103-0027 Tokyo, Japan (e-mail: yokoya@k.u-tokyo.ac.jp).

R. Hänsch is with the German Aerospace Center (DLR), 82234 Weßling, Germany (e-mail: ronny.haensch@dlr.de).

B. Le Saux is with ESA/ESRIN, Φ-Lab, I-00044 Frascati (RM), Italy. (e-mail: bls@ieee.org)

M. Brown acknowledges this work was supported by IARPA contract no. 2017-17032700004.

modeling of the Earth's surface from satellite imagery with the automated cartography of its physical aspects.

DFC19 used the Urban Semantic 3D (US3D) data [22] to deliver an unprecedented number of images and 3D references, producing more than 320 GB of data that span roughly 20 km^2 over the urban areas of Jacksonville (Florida) and Omaha (Nebraska) in the United States. The data comprised WorldView-3 satellite images (courtesy of Maxar), both panchromatic and eight-band visible and near-infrared, with ground sampling distances (GSDs) of 35 cm and 1.3 m, respectively. Second, 3D data were provided as point clouds or digital surface models (DSMs), produced using airborne LiDAR at a resolution of 80 cm. Finally, semantic labels were produced for urban classes including buildings, elevated roads and bridges, high vegetation, ground, and water.

DFC19 consisted of four parallel challenges that were organized as four independent tracks. Tracks 1, 2, and 3 were dedicated to semantic 3D reconstruction with various levels of input data. Participants were able to submit semantic maps and Digital Elevation Models (DEMs) resulting from single-view semantic 3D methods (Track 1), two-view stereo semantic 3D methods (Track 2), and multi-view stereo semantic 3D algorithms (Track 3). Track 4 addressed a related but different problem: large-scale 3D point cloud semantic labeling.

The present paper is the first of a two-part manuscript that aims to present and critically discuss the scientific outcomes of the 2019 Contest. This first Part A focuses on semantic 3D reconstruction and covers Tracks 1, 2, and 3. Complementary, Part B [1] is dedicated to large-scale point cloud classification and reports on the results of Track 4.

In detail, we describe the relevant datasets in Section II, and discuss the overall results of the 3D reconstruction challenges of the contest in Section III. We focus on the approaches proposed by the winning teams in each of the 3D reconstruction challenges, reporting on single-view 3D estimation in Section IV, pairwise semantic stereo reconstruction in Section V, and multi-view stereo reconstruction in Section VI. Finally, we present our concluding remarks in Section VII.

II. The Data of the 3D Reconstruction Challenges of the Data Fusion Contest 2019

Data from US3D were provided for all DFC19 challenge tracks. US3D is a large-scale public dataset including multi-date, multi-view, and multi-band satellite images and ground truth geometric and semantic labels covering approximately 100 square kilometers over Jacksonville, Florida and Omaha, Nebraska, in the United States [22]. The diversity of image viewpoints, resolutions, and months over which data were collected is shown in Figure 1. Training and test datasets were provided for each challenge track in the contest, which included approximately 20% of the total US3D data. Details of the data provided for the 3D point cloud semantic labeling Track 4 are presented in Part B [1]. The following were provided for Tracks 1, 2, and 3:

- Multi-date WorldView-3 panchromatic and 8-band visible and near-infrared (VNIR) satellite images were provided courtesy of Maxar. The source data consist of 26 images collected between 2014 and 2016 over Jacksonville, Florida and 43 images collected between 2014 and 2015 over Omaha, Nebraska, in the United States. GSDs of approximately 35 cm and 1.3 m were used for the panchromatic and VNIR images, respectively. The VNIR images were all pan-sharpened. The satellite images were provided in geographically non-overlapping tiles, whereas the airborne LiDAR data and semantic labels were projected onto the same plane. Unrectified images (for Tracks 1 and 3) and epipolar rectified image pairs (for Track 2) were provided as TIFF files.
- Airborne LiDAR data from the Homeland Security Infrastructure Program (HSIP) were used to provide a reference geometry with an aggregate nominal pulse spacing of approximately 80 cm. The training data derived from LiDAR included reference of the Above Ground Level (AGL) height images for Track 1, pairwise disparity images for Track 2, and Digital Surface Models (DSMs) for Track 3, which were all provided as TIFF files. The images were not collected concurrently with the LiDAR, indicating that the solutions had to address noise in the training labels caused by changes in the scenery.
- Semantic labels were provided as TIFF files for each geographic tile in Tracks 1-3. Semantic classes in the contest included ground, trees or high vegetation, buildings, water, and elevated roads or bridges.

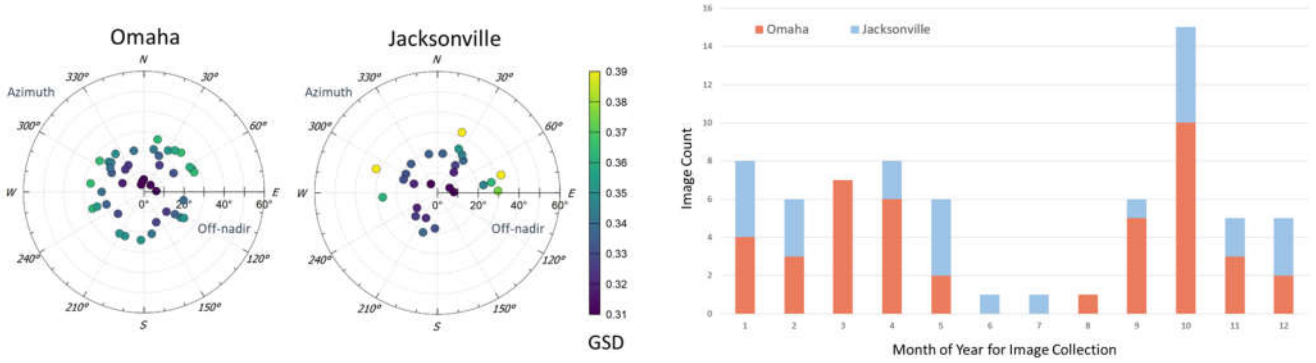
The above datasets were only provided for the training regions. The reference data for the validation and test regions remained undisclosed and were used for evaluation of the results. The training and test sets used in the contest included dozens of images for each geographic $500m \times 500m$ tile: 111 tiles for the training set, 10 tiles for the validation set, and 10 tiles for the test set. The training and test datasets were selected to ensure similar semantic and geometric distributions, as shown in Figure 2. After the contest was completed, we also released an extended training dataset, including RGB and VNIR images, AGL heights and the semantic labels for 756 geographic tiles. Additional reference layers were included in the extended training set, including building footprints, facades, and shadow masks, as described in [23]. The contest data and extended training data are available on the IEEE DataPort [24], [25].

III. Organization, Submissions and Results

Three parallel and independent tracks were dedicated to semantic 3D reconstruction in DFC19. Each track addressed a specific task and followed a different set-up, as described in Sections III-A, III-B, and III-C. Because each of these three tracks was concerned with semantic annotation as well as one form of 3D reconstruction, performance was assessed using pixel-wise mean Intersection over Union (mIoU), for which true positives must have both the correct semantic label as well as a 3D estimate within a certain error range. We call this metric mIoU-3.

Participation is analyzed in Section III-E and the winning approaches in Section III-F, while Section III-D discusses the provided baseline solutions.

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.2020.3032221, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing



(a) (b)

Fig. 1. Satellite image statistics. (a) Viewpoints and pixel ground sample distance for each image. (b) Seasonal distribution of image collection dates.

A. Track 1: Single-view semantic 3D

An unrectified single-view image was provided for each geographic tile. The objective was to predict semantic labels and normalized DSM (nDSM) above ground heights, as shown in Figure 3. Participants in Track 1 were tasked with submitting 2D semantic maps and AGL maps in raster format (similar to the TIFF files in the training set). Performance was assessed using the mIoU-3 metric with a height error of less than a threshold of one meter.

B. Track 2: Pairwise semantic stereo

In this case, a pair of epipolar rectified images was provided for each geographic tile. The objective was to predict semantic labels and stereo disparities, as shown in Figure 4. Participants of Track 2 were tasked with submitting 2D semantic maps and disparity maps in raster format (similar to the TIFF files in the training set). Performance was assessed using mIoU-3 with a threshold of 3 pixels for disparity values.

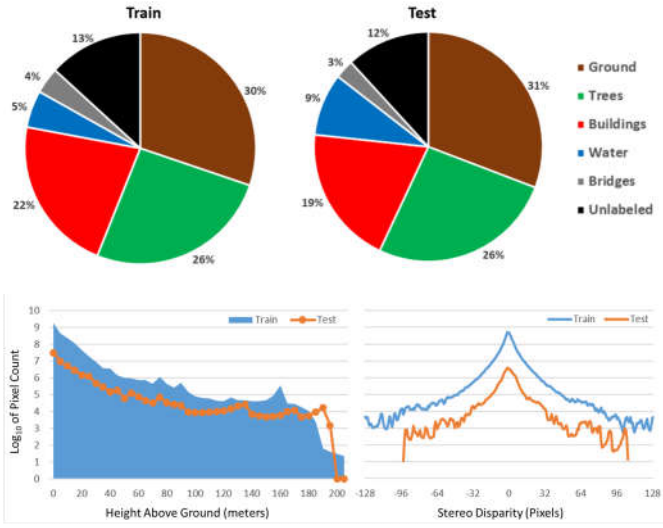


Fig. 2. Distribution of semantic labels for Track 1, height above ground values in Track 1, and stereo disparity values in Track 2.

(a) (b) (c)

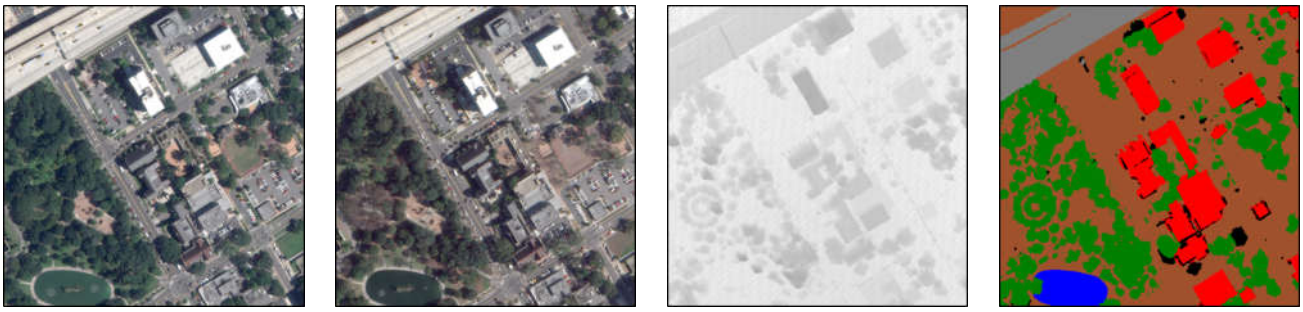
Fig. 3. Track 1: From a single image (a), predict height above ground (b) and semantic label (c) for each pixel.

C. Track 3: Multi-view semantic stereo

With multi-view images provided for each geographic tile, the objective in this task was to predict semantic labels and a DSM. Unrectified images were provided with Rational Polynomial Coefficients (RPC) metadata that had already been adjusted with LiDAR such that registration was not required for the evaluation, indicating that the solutions could focus on the methods used to select images, the correspondence, semantic labeling, and the multi-view fusion. Because this track relies on RPC metadata, which everyone may not be familiar with, the baseline algorithm provided also included simple Python code with which the RPC could be manipulated for epipolar rectification and triangulation. Participants of Track 3 were tasked with submitting 2D semantic maps and DSMs in raster format (similar to the TIFF files in the training set). Performance was assessed using mIoU-3 with a threshold of 1 m for the DSM height values.

D. Baseline Solutions

Lightweight baseline solutions were developed for all the challenge tracks and used to validate the data and characterize the expectations in terms of minimum performance. These solutions were made available on GitHub for the contest participants [26]. The Track 1 single-view semantic 3D baseline solution combines semantic segmentation and height regression deep network models, both based on a ResNet34 [27] encoder and U-Net [28] decoder implemented in Keras and TensorFlow [29]. The Track 2 pairwise semantic stereo solution combines an ICNet [30] semantic segmentation model with a DenseMapNet [31] stereo disparity regression model.



(a) (b) (c) (d)

Fig. 4. Track 2: From an epipolar rectified pair of multi-date images (a and b), predict stereo disparity (c) and semantic label (d) for each pixel.

The Track 3 multi-view semantic stereo baseline solution combines semi-global matching [32] for disparity estimation with the same ICNet model that was used in Track 2 for semantic segmentation. These baseline solutions do not exploit the complementary nature of semantic segmentation and 3D reconstruction tasks. An initial experimental evaluation using the baselines was reported in [22].

E. Participation

A total of 710 unique registrations were received from 45 countries for downloading the DFC19 data. A total of 41, 37, and 23 teams finally entered Tracks 1-3, which received 337, 264, and 142 submissions, respectively, on the Codalab competition websites during the test phase. These numbers indicate that the single-view semantic 3D and pairwise semantic stereo challenge were the most popular, with less demand for the multi-view semantic stereo challenge. Tasks that required more specialized knowledge were subjected to higher entry levels, and the single-view semantic 3D that was composed of estimating semantic labels and nDSM, which has been actively studied with machine learning, was easier in terms of participation.

F. Best-performing approaches and discussion

The first and second ranked teams in all tracks were awarded winning places. The winners presented their solutions during the 2019 IEEE International Geoscience and Remote Sensing Symposium in Yokohama, Japan. The six winning teams for Tracks 1–3 were:

- **1st place in Track 1**: The *nest* team; Saket Kunwar from NestAI, Nepal; with an ensemble of a few varied backbones employed in a U-Net architecture [33].
- **2nd place in Track 1**: The *RSIDEA-WHU* team; Zhuo Zheng, Yanfei Zhong, and Junjue Wang from Wuhan University, China; with a pyramid on pyramid network based on an encoder-dual decoder framework [34].
- **1st place in Track 2**: The *BurningAllthing* team; Hongyu Chen, Manhui Lin, Hongyan Zhang, Guangyi Yang, Guisong Xia, Xianwei Zheng, and Liangpei Zhang from Wuhan University, China; with a modified version of the pyramid stereo matching network (PSMNet) and disparity fusion segmentation net (DFSN) [35].
- **2nd place in Track 2**: The *qin.324* team; Rongjun Qin, Xu Huang, Wei Liu, and Changlin Xiao from Ohio Ohio State University, US; with an U-Net and pyramid stereo matching network (PSMNet) [36].
- **1st place in Track 3**: The *Panoptes* team; Pablo d'Angelo, Daniele Cerra, Seyed Majid Azimi, Nina Merkle, Jiaojiao Tian, Stefan Auer, Miguel Pato, Raquel de los Reyes, Xiangyu Zhuo, Ksenia Bittner, Thomas Krauss, and Peter Reinartz from German Aerospace Center, Germany; with semi-global matching and an ensemble of CNN classifiers with ad hoc detectors [37].
- **2nd place in Track 3**: The *qin.324* team; Rongjun Qin, Xu Huang, Wei Liu, and Changlin Xiao from the Ohio State University, US; with semi-global matching and U-Net [38].

Table I is a summary of the teams ranked top three for Tracks 1–3 and their approaches. The overall trend was that the top-ranked teams in each task extended well-established techniques with practical tricks. All winners adopted fully convolutional neural networks (FCNs) and their variations (e.g., U-Net and LinkNet) for semantic segmentation. The methods used to estimate height differed significantly depending on the tracks. Both winners of Track 1 developed U-Net-based regression models for the nDSM estimation that leveraged global statistics in each class. The winners of Tracks 2 and 3 tackled height estimation using PSMNet and semi-global matching (SGM), respectively. Many teams achieved further improvements in performance by ensembling multiple prediction models and post-processing.

In Sections IV, V, and VI, we present the solutions proposed by the first ranked teams of Tracks 1, 2, and 3, respectively. The winning classification methodologies are summarized and an in-depth analysis of the pros and cons of each solution is provided.

IV. First Place in the Single-view Semantic 3D Challenge: NestAI team

A. Method: U-Net ensemble for semantic and height estimation using coarse-map initialization

We used a U-Net [28], which is an architecture that is widely used for its strong performance in semantic segmentation tasks, and evaluated it using mIoU. It is a variation of the classic auto-encoder architecture and consists of using an

TABLE I
TOP RANKED TEAMS AND APPROACHES.

Track	Rank	Team	mIoU-3	Affiliation	Approach				
					FCN	PSMNet	SGM	Ensemble	Postprocess.
1	1	nest	0.5571	NestAI	✓			✓	
	2	RSIDEA-WHU	0.5340	Wuhan University	✓				
	3	nick_ncc	0.5209	Wuhan University	✓				✓
	–	Baseline	0.456	–	✓				
2	1	BurningAllthing	0.7775	Wuhan University	✓	✓		✓	✓
	2	qin.324	0.7724	The Ohio State University	✓	✓		✓	✓
	3	Kampai	0.7606	Xidian University	✓	✓			✓
	–	Baseline	0.608	–					
3	1	Panoptes	0.7461	German Aerospace Center	✓		✓	✓	✓
	2	qin.324	0.7300	The Ohio State University	✓		✓		✓
	3	Midkey_zhong	0.7282	Xidian University	✓		✓		✓
	–	Baseline	0.550	–					

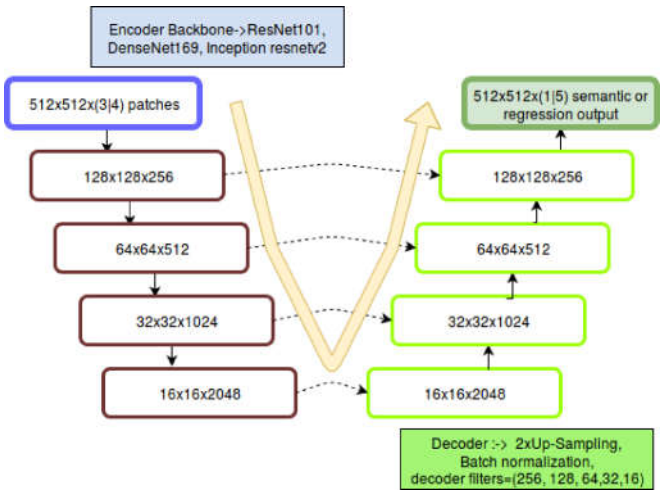


Fig. 5. U-Net Encoder-Decoder Architecture

encoder followed by a decoder preserving the input-output dimensionality. The novelty that was introduced alongside the U-Net consists of a feature passing mechanism that used lateral connections from each stage of the encoder to the corresponding decoder stage, as shown in Figure 5, enabling precise localization. A variety of backbones pretrained on large datasets such as ImageNet can be used for semantic segmentation, provided the decoder part is suitably constructed to reflect the up-sampling or deconvolution necessary to output the segmentation map.

We evaluated three backbones: ResNet101 [27], Inception ResNetV2 [39], and DenseNet169 [40]. ResNet101, which has 101 layers, uses deep residual learning with skip connections that allow for deeper networks and avoid performance saturation. As a deep network, it has a hierarchical structure, with each stage providing a specific feature representation, such as the haar-like features from the first stage. Through the use of skip connections at each stage, the residual blocks enable better refinements of the feature representation, leading to better segmentation maps. In the Inception ResNetV2 architecture, the layers are wider using filters of different size, in addition to the residual connection for increasing depth. Batch normalization is used to reduce the covariate shift between the layers. The change in the distribution of pixel intensity for the same conditional output distribution or covariate shift is a problem that is particularly prevalent in remote sensing imagery. The relatively better performance of the inception network suggests that batch normalization adds to its effectiveness. It should be noted though that batch normalization addresses the internal covariate shift and therefore deals with the shift in the given dataset. When train and test images are from different locations, an external covariate shift can occur. The batch statistics collected during training may therefore be unsuitable and may lead to poor results. For example, if the dataset used for training and testing had been obtained from Africa or Asia as opposed to nearby cities, as with this dataset, other approaches would have been required. DenseNet [40] expands upon the deeper and wider network architecture using short-cuts between the layers, and was also evaluated during the competition.

To provide context to the higher resolution stages, we added a feature pyramid network [41] to the encoder that started at the low-resolution stage. The top-down pathways were constructed via successive up-sampling. We evaluated the feature pyramid network for the semantic prediction task as well, but found its performance slightly worse in this case. The backbones, which were primarily developed for classification tasks with high-resolution images, downsample the input by a factor of four in the first stage. For remote sensing imagery, this can be too severe as the input resolution is low compared to that of natural images, leading to difficulties in fine-grained object segmentation or detection. A possible fix is to upsample the input image by a factor of two or to reduce the stride in the convolution layer of the first stage of the backbone to two. However, this was not performed in this challenge because mIoU was used as an evaluation metric, which rewards semantic accuracy rather than detection.

1) Semantic Prediction: From the eight multispectral bands of Worldview-3 images, bands 7 and 8, and the infrared bands including band 6, the red-edge band, provide spectral signatures that are most suitable for delineating urban areas. We created patches of 512×512 3-band combinations from these bands as well as RGB bands from the given 1024×1024 dimensional images. The model performance on a local validation set for the three backbones (ResNet101 [27], Inception ResNetv2 [39], and DenseNet169 [40]) with four different band combinations is shown in Table II. We determined that Inception ResNetV2 with the band combination of 7,

6, 4 provided the best overall single model performance. Nadam [42], which is an optimizer with momentum, was used beginning at a learning rate of 0.002 for 30 epochs and then fine-tuned for 5 more epochs at a learning rate of 0.0002.

We used the Hybrid Jaccard loss (HJL), a weighted sum of binary cross-entropy loss (BCE) and the Jaccard loss (JL), shown in Equation 1. The dataset contains classes that have imbalanced pixels as a result of the used images. Ground and water are generally balanced, while roads and bridges tend to face a foreground-background imbalance problem. Introduced in RetinaNet [43], focal loss can help with this imbalance, and in combination with cross-entropy loss, can enhance the performance of the imbalanced classes, as noted for JL above. Conceptually, focal loss down-weights the contribution of easy predictions and instead focuses on more difficult samples, while JL works by penalizing incorrect predictions.

The final prediction during the competition phase was a simple mean ensemble of all the semantic models, as seen in Table II. Vertical flips, horizontal flips, and rotation were used to augment the dataset during training. The test time augmentation consisted of one additional vertical flip.

TABLE II
MODELS PERFORMANCE FOR SEMANTIC SEGMENTATION

	Inception ResNetV2	Inception ResNetV2	DenseNet 169	ResNet 101
Bands	4,6,7	1,2,4	4,5,6	5,6,7
Ground	**0.891**	0.885	0.868	0.861
Trees	**0.701**	0.696	0.648	0.640
Roof	**0.862**	0.847	0.825	0.812
Water	0.940	**0.941**	0.932	0.915
Bridge	**0.819**	0.802	0.771	0.750
mIoU	**0.843**	0.834	0.809	0.795

$$HJL = 0.75JL + 0.25BCE \quad (1)$$
$$= 0.75\frac{y_{\text{pred}} \cap y_{\text{true}}}{y_{\text{pred}} \cup y_{\text{true}}} - 0.25\sum_{i==1}^{n} y_{\text{true}_i} \cdot log(y_{\text{pred}_i})$$

After the competition, we evaluated the combination of focal loss and cross-entropy loss against the combination of JL and cross-entropy loss, while maintaining the same training and model parameters, except for the loss. As shown in Table III, the focal loss performed better when evaluated with two different models and two different band combinations. This was only carried out for the imbalanced bridge class.

TABLE III
PERFORMANCE COMPARISON FOR FOCAL AND JACCARD LOSS

	BCE + Jaccard	BCE + Focal
Class = Bridge Model = ResNet101 Bands = 5, 6, 7	0.750	0.762
Class = Bridge Model = DenseNet 169 Bands = 4, 5, 6	0.771	0.793

2) Height Prediction: Predicting height from a single optical image without any additional supporting data or metadata is a challenging task. It is easy to overfit when a small validation dataset is used, especially for buildings and bridges. This is demonstrated using an experiment following a simple method to fill in the height values with one globally computed mean or median. In Table IV 65 instances are extracted from different tiles (i.e., JAX_204, JAX_224, and OMA_292), and the AGL is filled-in after the semantic segmentation. Each label was replaced with a corresponding single global height value. A change of 1 m in the global value for buildings and 5 m for bridges can generate mIoU-3 variation of as much as 8%. The selected global value may work for this particular set but may fail to generalize for other test sets such as those with predominantly high-rise buildings or urban areas in the city outskirts. Exceptions to this rule can be observed in the ground and water classes, as there is a 1 m tolerance for mIoU-3 calculations with these classes and the deviation from the mean value for these classes lies within this tolerance limit.

TABLE IV
VARIATION OF MIOU-3 SCORES OF HEIGHT PREDICTIONS BASED ON COARSE DEFAULT VALUES AND THE ESTIMATE OF A REGRESSION MODEL.

Building	value =5.3	value =6.3	value =7.3	Regression model (with coarse-map)
mIoU3	0.64	0.59	0.47	**0.72**
Bridges	value =7.3	value =10.3	value =15.3	Regression model (with coarse-map)
mIoU3	0.268	0.372	0.453	**0.46**

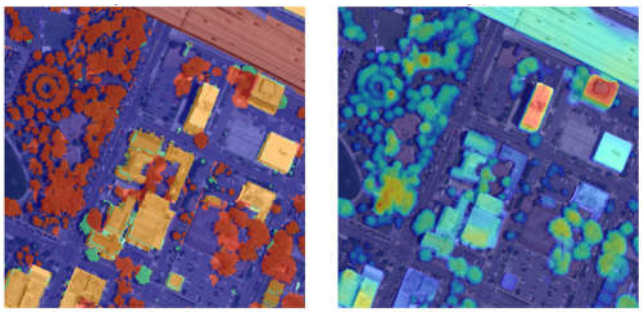


Fig. 6. Based on the initial coarse AGL (left) the final AGL (right) is predicted.

Instead of using a single model to predict the semantic class and height values, we opted to use two models to predict the height once the semantic class is mapped. To introduce cues from which a deep network can bootstrap, we added a coarse channel to the 3-band optical imagery (see Figure 6) initialized from pre-computed global statistics. Addition of a coarse-map was also performed by [44] for depth prediction using RGB images, as the map provides essential cues for depth estimation, such as starting or terminating points. We trained a U-Net regression model and used rmsProp [45] for the height prediction.

In addition to the model with a coarse channel, a model without a coarse channel constituted part of the class-specific ensembles (see Table V). For the Tree class, the model with the coarse channel achieved a worse mIoU-3 score than the model

without it, indicating that its contribution was down-weighted. Similar to semantic segmentation, the height prediction test time augmentation included one vertically flipped image. Ensembling and test time augmentation leads to significantly better results in terms of height prediction.

TABLE V
HEIGHT MODELS WEIGHTED ACCORDING TO SEMANTIC CLASS

	Band 4 6 7 Inception ResNetV2 (With coarse)	Band 4 6 7 Inception ResNetV2 (Without coarse)	Band 4 5 6 DenseNet (With coarse)	Band 4 5 6 DenseNet (Without coarse)	Ensemble
Buildings	0.33	0.16	0.33	0.16	Contributing Fraction
mIoU3 Public Validation	0.4080	0.3803	0.3920	0.3750	0.4383
mIoU3 Public Test	0.4147	0.4060			0.4230
Trees	0.16	0.33	0.16	0.33	Contributing Fraction
mIoU3 Public Validation	0.2745	0.2830	0.2703	0.2780	0.2974
mIoU3 Public Test	0.2394	0.2598			0.2714

B. Results and Discussion

In this report, we show that an ensemble consisting of a few varied backbones with different band combinations in a U-Net architecture, with hybrid JL and a Nadam optimizer, provides the best semantic segmentation result for the grss_dfc_2019 dataset. For the challenging task of height prediction, we show that the addition of a coarse-map initialized using global statistics provides a model that can generalize well and converges rapidly. Our band selection strategy proved successful; however, the key to the challenge-specific performance was the use of coarse maps, while utilizing essential cues for the height estimation. Methods that utilize the predicted height and perform further iterative refinement based on local and global contexts with the addition of pyramid features should be investigated. Another avenue for improvement is to replace dice loss with focal loss [43], as the road and highway classes are generally imbalanced.

V. FIRST PLACE IN THE PAIRWISE SEMANTIC STEREO CHALLENGE: WUHAN UNIVERSITY TEAM

In this section, we describe the winning algorithm proposed for the pairwise semantic stereo challenge. This method is based on a fusion-based framework in which the semantic segmentation and disparity estimation tasks are performed simultaneously. A deep neural network is utilized to extract multi-scale features of the input epipolar rectified stereo image pair. The construction and minimization of the cost volume finally allows to apply regression to infer the disparity map. For the semantic segmentation task, the multi-receptive fields semantic features of the stereo image pair are exploited, before being fused with the disparity features. The segmentation accuracy is improved using information about the semantic context of an object as well as the disparity information that contains cues regarding object elevation. Further, high vegetation, elevated roads, and buildings are refined with the help of class elevation priors, and noisy pixels are eliminated through morphological operators during post-processing.

A. Proposed Framework

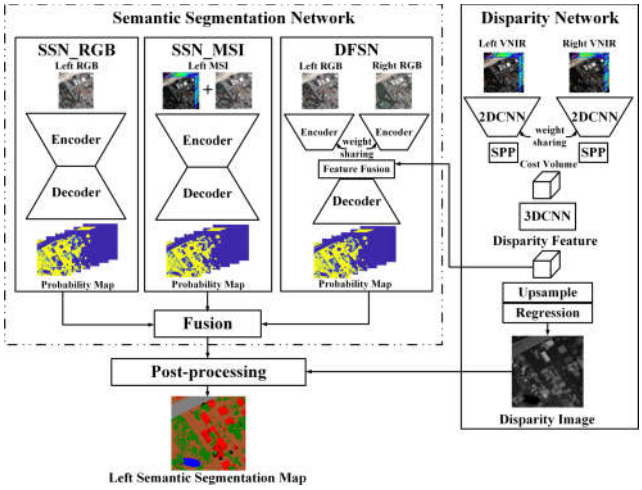


Fig. 7. The flow chart of the proposed method for pairwise semantic stereo.

The flow chart of our approach is shown in Figure 7 [35]. The framework can be summarized as follows:

- For the disparity estimation branch, a disparity network is developed that considers the VNIR stereo pair as input and generates the disparity map in an end-to-end manner.
- For the semantic segmentation branch, the left image is fed into our proposed single segmentation network (SSN) to predict the left segmentation map. With respect to the different data types, the SSN-RGB is built for the RGB inputs and the SSN-MSI for the 11-band multispectral images (MSIs), i.e., the concatenation of RGB and VNIR.
- For further mining of 3D stereo segmentation information, we propose the disparity fusion segmentation network (DFSN), which aggregates the semantic features from the stereo pair and fuses them with the disparity features from the disparity network.
- To fully combine the complementary information of the three segmentation models, we apply a pixel-wise median filter on the three probability maps of SSN-RGB, SSN-MSI, and DFSN, and determine the ensemble prediction. After post-processing, with the assistance of the disparity results, the final segmentation map is obtained.

B. Disparity Estimation

Owing to the different acquisition times of the stereo image pairs, there are obvious differences in the irradiation and environmental changes in the image pairs, which will significantly interfere with the accuracy of the traditional disparity estimation algorithms based on low-level image features (geometry, texture, color, etc.). Therefore, we need to utilize a deep learning method that generates high-level semantic

features, use these features to construct the cost volume, and finally regress the final disparity map.

As shown in Figure 7, we adopt the PSMNet [46] to enable end-to-end disparity estimation. The atrous convolution with a downsampling rate of 8 is not only used to increase the receptive field for semantic features, but also to reduce the computational costs of the network. Meanwhile, the Spatial Pyramid Pooling (SPP) module helps to fuse the contextual information at different scales. The SPP features from stereo pairs are then utilized to form the 4D ($height \times width \times disparity \times channel$) cost volume by concatenating the stereo feature maps across each disparity level. Moreover, stacked hourglass 3D convolutional layers are employed to aggregate the feature information along the disparity and spatial dimensions. Finally, the stereo features are regressed to the disparity map. For the contest, the disparity range was set to $[-96, 96]$, according to the statistics of the training set.

C. Semantic Segmentation

In this section, our fusion-based segmentation method for deep learning-based semantic segmentation is elaborated in three parts: 1) SSN, 2) DFSN, and 3) probability fusion. The architectures of the proposed SSN and DFSN are shown in Figure 8 [35], where the dash-dot arrowed line is used to denote the SSN and the dashed line describes the DFSN.

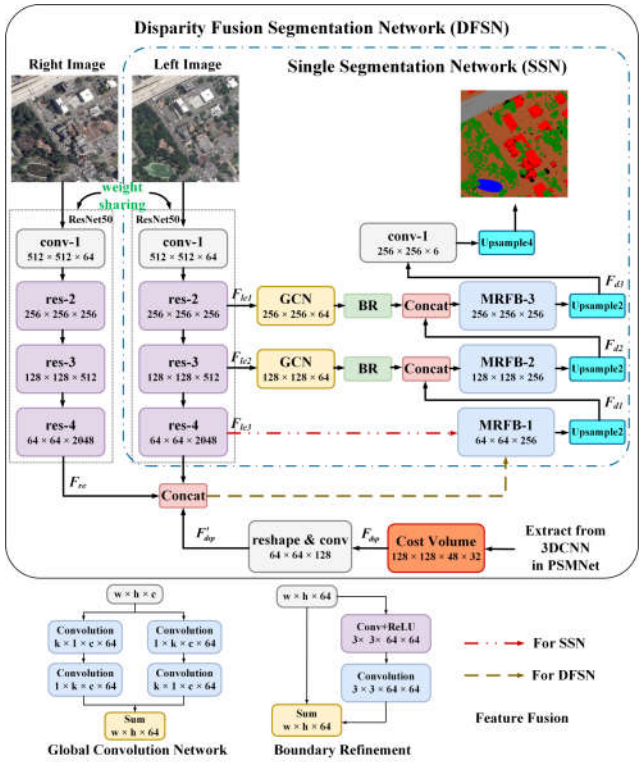


Fig. 8. The architecture of the proposed SSN and DFSN for semantic segmentation.

1) Single Segmentation Network: The SSN is designed in an encoder-decoder architecture. A pretrained ResNet-50 [27] is adopted as the backbone for the encoder, which is known to extract multi-level features effectively. The remainder of this section discusses the special block that was introduced into the decoder.

(a) (b)

Fig. 9. (a) Objects of the same class in different scales. (b) Similar color and texture between different classes.

The complex urban scene also leads to different characteristics within objects of the same class; the intra-class variance tends to be large. For example, it is apparent from Figure 9(a) that the size and shape of buildings varies considerably. As a result, extracting semantic features with a constant receptive field size weakens the inference ability of the network to identify the same object in different scenes. Therefore, the network should be designed to automatically extract semantic information from different receptive fields. In addition, the network should be able to adaptively weigh the features of different receptive fields through learning from mass data. Given these requirements, we propose the Multi-Receptive Fields Block (MRFB) as the basic block in our decoder to segment objects on multiple scales.

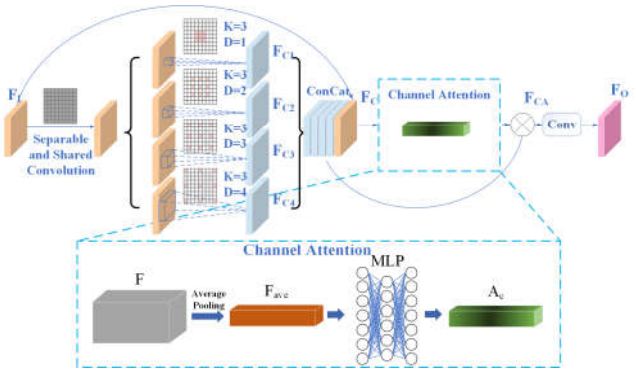


Fig. 10. The architecture of the proposed MRFB.

The architecture of the proposed MRFB is shown in Figure 10 [35]. The MRFB benefits from atrous SPP (ASPP) in deeplabv3+ [47], where four atrous convolutions with different dilation rates are applied in parallel to extract multi-context features. The original input feature F_i is then concatenated with these features for more stable training. Meanwhile, the discriminative contextual feature $F_c = [F_i, F_{c1}, F_{c2}, F_{c3}, F_{c4}]$ is obtained for five different receptive fields. The atrous convo-

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.2020.3032221, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

lution is well-known to impose the expansion of the receptive fields without the loss of resolution or coverage. However, it can sometimes result in gridding artifacts. Therefore, separable and shared convolution [48] are employed in the MRFB to establish dependencies among the input channels.

Simply concatenating the features of different receptive fields is inefficient. Hence, in order to understand the relative significance of different contexts, channel attention [49] is introduced to weigh the features in respect of their importance. Specifically, average pooling is adopted to distill a feature vector from the input feature maps, and the multi-layer perceptron (MLP) is used to obtain the channel attention vector. The MLP plays an essential part as it automatically learns the nonlinear mapping of a channel's significance via back-propagation in the training process. The formulation of the channel attention vector is:

$$A_c(F_i) = \theta(MLP(F^c_{i(ave)})) \in R^{c\times 1\times 1} \tag{2}$$

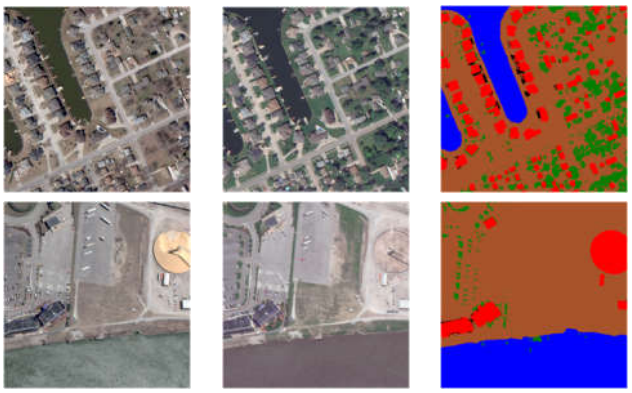
where θ denotes the sigmoid function. We use the channel attention vector A_c to weight the contextual features of the different receptive fields, as described below:

$$F_{CA} = F_i \otimes A_c(F_i) \tag{3}$$

where $\otimes$ denotes element-wise multiplication. The 3×3 convolutional layer is utilized to fuse the weighted feature F_{CA} and reduce the channel number to 256 for concentration.

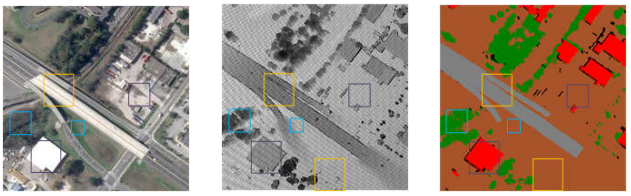
The US3D dataset covers extensive scenes. As a result, objects of different classes in the US3D can appear to be similar. For instance, the characteristics of the buildings and ground in terms of color and texture are very similar in Figure 9(b). Thus, the network learns more robust features and makes more accurate inferences only if the long-range relationship between the input pixels is fully utilized. In order to establish long-range dependence among these features, a global convolutional layer is required, which has a large number of parameters and high computational costs. The Global Convolutional Network (GCN, [50]) module builds dense connections within the global feature maps with a size of $k\times k$ through a combination of $1\times k+k\times 1$ and $k\times 1+1\times k$ convolutions. Because deeper features can lose large amounts of structural information compared with shallower ones, the shallow features are transmitted to the deeper layers by GCN and boundary refinement (BR) in order to better recover the spatial information.

2) Disparity Fusion Segmentation Network: Since the stereo images in US3D were not acquired in the same time phase, the two acquisition times could be in different seasons. As shown in Figure 11, objects in the same scene may appear different in different seasons: the flourishing and withering of trees and changes in the color or the physical state of water. This nonconformity led to great challenges for the network. To better resolve this issue, we include a stereo image pair in the network inputs. By combining semantic information from different time phases, the representative power of the network is enhanced, which increases the robustness to seasonal change. Therefore, on the basis of the SSN, the semantic features of the right image are extracted through the weight-sharing encoder and concatenated with the features of the left image.



(a) (b) (c)

Fig. 11. Different seasons between the epipolar rectified pair (a and b), and their semantic label (c).



(a) (b) (c)

Fig. 12. The disparity cues contain important information for semantic segmentation. (a) The left-view image, (b) the disparity map and (c) the semantic label

It is apparent from Figure 12 that the elevated and non-elevated roads share similar geometric and textural characteristics. Similar observations can be found in the pairs including high vegetation and grasslands, or buildings and parking lots. It is not easy to discriminate these highly correlated objects using only the semantic features. However, the disparity information inherently contributes to the discrimination between easily confused objects at different elevations. Hence, a straightforward approach is to incorporate the features of the disparity network into the semantic segmentation network. Our solution is to reshape the 4D stereo features from the PSMNet, transform them in channels via a 1×1 convolution, and feed them into the segmentation net in concatenation with the semantic features of the left and right images. Additionally, the disparity information also plays an important role in the alignment of the left and right images. At this point, the entire DFSN model is built, as shown in Figure 8.

3) Probability Fusion: As shown in Figure 7, three probability maps are predicted by SSN-RGB, SSN-MSI, and DFSN. Then, a median filter is used to perform post-inference fusion. Subsequently, the fused map is normalized, and the ensemble segmentation map is obtained by considering the class with the maximum probability at each pixel.

D. Implementation Details

Since the RGB and VNIR epipolar rectified images were valued in different ranges, we first pre-processed all the samples via the well-known z-score normalization, with the channel-wise mean and variance calculated on the training set.

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.2020.3032221, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

In the PSMNet the loss function is the mean absolute error (MAE).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i^p| \tag{4}$$

where y_i^p and y_i denote the estimated disparity and ground truth, respectively. Cross-entropy loss was used in the SSN and DFSN.

We selected the Adam optimizer [51] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ for all networks. Regarding the learning rate γ, we employ an auto-attenuating learning rate that halved every 30 epochs, which initializes at 0.01 in the PSMNet and 0.001 in the SSN and DFSN.

Four different 3×3 atrous convolutions are used in the MRFB, with dilation rates set to $[1, 4, 8, 16]$, corresponding to a kernel size of 31 for the separable and shared convolution. For the segmentation task, random flipping and random rotation are adopted to augment the training data. With respect to the RGB inputs, a ResNet-50 pretrained on ImageNet is used, significantly accelerating the convergence. As for the MSI inputs, the input convolutional layer in ResNet-50 is modified to match the channels. The parameters of the pretrained PSMNet are frozen during training for training stability of the DFSN.

E. Post-Processing

Because all the classes in the US3D dataset have natural characteristics in terms of elevation, we propose a post-processing method based on class elevation priors in order to correct confused labels and remove noise caused by misclassification.

- Correction for high vegetation: Although the proposed DFSN has to some extent addressed the difficult classification problem under seasonal change, high vegetation is sometimes confused with the ground class, or the miscellaneous collection of low ground objects. This is because of the variable appearance of high vegetation and its similar spectral characteristics with those of grasslands. To separate the high vegetation from the ground class, corrections are made on the basis of prior knowledge of elevation. We first remove the average disparity of the ground class from the original disparity map and consider the absolute values, obtaining an estimation of the pseudo DSM. Then, the ground pixels with a pseudo DSM value higher than 3.0 and the probability of high vegetation higher than 0.3 are set as high vegetation.
- Removal of noise in the segmentation map: The building and elevated road classes are known to have distinct geometric and elevation characteristics. To fill in the potential holes and gaps caused by the inside of buildings and elevated roads, we create a binary segmentation map of the buildings and elevated roads, and use a morphological method to extract all 8-connected components. The connected domain filtering algorithm is then applied to fill the holes, with thresholds of 600 for the pixel number and 5.0 for the average pseudo DSM in the component. We also remove the noise caused by water and elevated roads via a simple global filtering strategy. Compared with other classes, the spatial distribution of water and elevated roads is usually clustered within a few isolated points. Based on this prior knowledge, if the total number of pixels denoting water or elevated roads in a segmentation map is less than 500, the predicted class for these pixels is set to ground.

F. Results and Discussion

In this subsection, we evaluate the performance of our proposed framework using a test set of 50 images. Our method achieves an average endpoint error (EPE) of 1.3966, an erroneous pixel score (D1) of 0.0906, and a final mIoU-3 of 0.7775. An ablation study is conducted to verify the influence of different components in our framework, including segmentation network architectures and post-processing. A brief introduction to the methods used for comparison is given as follows.

- **SSN-RGB**: The single segmentation network with RGB input.
- **SSN-MSI**: The single segmentation network with MSI input.
- **DFSN**: The disparity fusion segmentation network.
- **Fusion**: The fusion of the proposed SSN-RGB, SSN-MSI, and DFSN via a median filter.
- **Post-fusion**: Fusion with post-processing.

TABLE VI
EXPERIMENTAL RESULTS WITH DIFFERENT STRATEGIES

class	SSN-RGB	SSN-MSI	DFSN	Fusion	Post-fusion
Ground	0.7925	0.7928	0.8026	**0.8114**	0.8093
High vegetation	0.5178	0.5333	0.5417	0.5495	**0.5692**
Building	0.7744	0.7703	0.7767	0.7964	**0.7964**
Water	0.9381	0.9421	0.9435	0.9488	**0.9511**
Elevated road	0.7459	0.7690	0.7785	0.8013	**0.8264**
All	0.7537	0.7615	0.7686	0.7815	**0.7905**

The results of all the compared methods are shown in Table VI and Figure 13 [35]. First, it is not difficult to ascertain that SSN-MSI achieves better results for high vegetation (51.78% versus 53.33%) and water (93.78% versus 94.21%) than SSN-RGB. This is because MSI contains more spectral information, and thus includes rich prior knowledge for the identification of vegetation and water. Additionally, the significant accuracy improvement of the elevated road class is due to the fact that the rich spectral information diminishes confusion with other classes. Second, from the experimental results of SSNs and DFSNs, there are obvious improvements in the classes with elevation characteristics, including high vegetation, buildings, and elevated road classes. Therefore, the elevation information carried by disparity cues is shown to play a significant role in semantic segmentation. Moreover, the increase in the ground, vegetation, and water classes, which are sensitive to seasonal changes, affirms the effectiveness of the fusion of the multi-date stereo pair. Finally, it is clear from the results that by combining the elevation information and the semantic confidence maps in our post-processing method, the accuracy concerning high vegetation and elevated roads is further improved.

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.2020.3032221, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

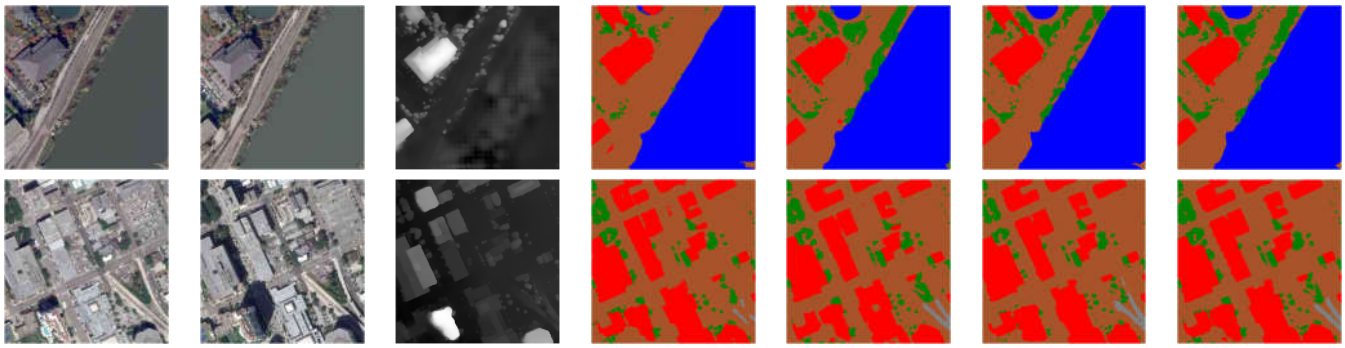


Fig. 13. Disparity estimation (c) between the epipolar rectified pair (a and b), and semantic segmentation results of (d) SSN-RGB, (e) SSN-MSI, (f) DFSN, and (g) Fusion-post

In post-processing, we introduce the pseudo DSM to perform class correction and denoising, and several thresholds are manually set. For future research, it is important to consider how to incorporate these steps into network training and build an end-to-end framework. In addition, since the disparity estimation accuracy benefits from the semantic information in the same manner, a joint multi-task framework can be established, where the semantic segmentation task receives guidance from the disparities, and vice versa. This will be an interesting topic to discuss in the future.

VI. First Place in the Multi-view Semantic Stereo Challenge: DLR team

In this section, we describe the winning algorithm proposed for the multi-view semantic stereo challenge. We used the procedure shown in Fig. 14 to obtain the best score on the multi-view semantic stereo challenge. After refining image orientation using bundle block adjustment, Semi-Global Matching (SGM) was used to produce height maps, Digital Surface Models (DSM), and normalized DSM (nDSM). Convolutional Neural Network (CNN) based semantic segmentation was performed on the RGB, multi-spectral images (MSI) and height maps and projected into UTM coordinates. Pixel-wise detectors were applied to the orthorectified MSI images, deriving binary maps for the classes high vegetation and water. An ensemble of three CNN classifiers was merged with the ad hoc detectors to obtain the final semantic segmentation maps, after an additional step of morphological filtering.

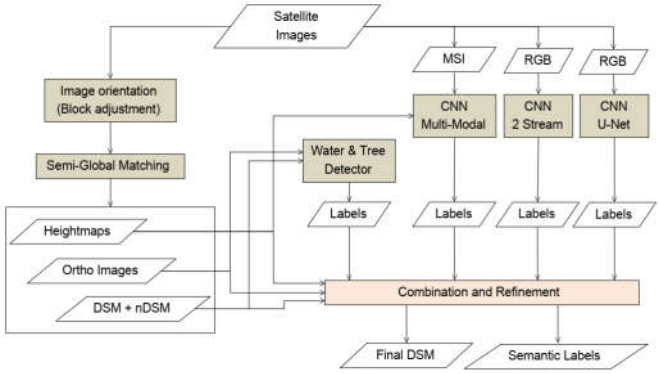


Fig. 14. Processing steps for multi-view semantic segmentation.

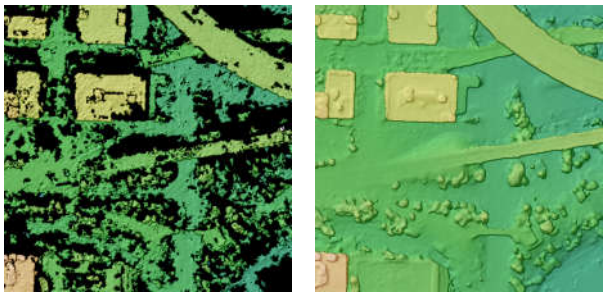


Fig. 15. DSM after matching (a) single stereo pair and (b) merging of 50 stereo pairs.

A. Image Orientation and Multi-View 3D Reconstruction

Before performing dense matching, a good relative image orientation is required. The contest dataset is only coarsely aligned to the reference data, leading to height offsets of more than 1 m for some stereo pairs. Additionally, relative orientation is required to avoid systematic height offsets between individual stereo pairs. For image orientation and dense matching, a synthetic panchromatic image is generated by averaging the red, green, and blue channels of the MSI images. Multi-ray tie points are matched using SIFT, refined, and transferred to unmatched images using local least-squares matching. Bias-corrected RPCs are then obtained using bundle block adjustment [52].

Following [53], dense stereo matching is performed using pairwise SGM using CENSUS as the matching cost. Owing to the difference in image acquisition time, dense matching of single stereo pairs yields incomplete results, particularly in areas that have undergone changes or include vegetation (cf. Fig. 15). All possible stereo pairs with a convergence angle above a predefined threshold are ranked based on the number of tie points found in the image orientation step, matching the 50 pairs with the highest number of tie points. Each pair is matched in both directions, resulting in 100 height maps.

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.2020.3032221, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

We compute height clusters for every pixel in the final DSM and select the mean height of the cluster with the highest number of points. In addition to the DSM heights, we produce quality layers containing the number of matches and standard deviations of all height values. The remaining holes are filled using interpolation. Finally, all images are orthorectified using the DSM. Additionally, normalized DSM [54] and dense height maps are generated for each input image, allowing the use of height information during semantic segmentation of the original images in sensor geometry.

B. Semantic Segmentation

Semantic classification is performed by utilizing three different neural network architectures, and two ad hoc approaches for the classes high vegetation and water. The RGB and MSI images, along with the dense height maps generated during our stereo matching, are used as input for the classification. Note that the third network of choice is the provided baseline U-Net network[1] and is therefore not described below.

The most interesting aspect for the ad hoc detection of the classes water and elevated roads is the use of information derived from multi-view processing, detailed below.

In order to initialize the water mask, pixels with low consistency in the DEM are selected, where the consistency is estimated as the number of DEMs in which an image element has been matched, by selecting pixels matched in less than 5 stereo pairs. Such pixels correlate well with the presence of water, which makes it difficult to obtain reliable matchings and height values. The Normalized Difference Water Index (NDWI) [55] is computed for these image elements: the water mask is initialized by maintaining higher values whenever a bimodal distribution of the NDWI is detected by an adaptive Otsu threshold, indicating the presence of bodies of water. The results are then refined with morphological filtering.

The elevated road labels are refined by adding to the class elevated pixels with a similar gradient to the objects detected, assuming that such a gradient is homogeneous for a small scene. This improves the accuracy of this class by 1.5%.

C. Results

TABLE VII
HEIGHT STATISTICS FOR DIFFERENT DSM FUSION AND POST-PROCESSING ALGORITHMS.

Method	Postproc.	Height accuracy	Completeness
Median		0.411	0.654
Median	VegHeight	0.408	0.658
Cluster		0.356	0.671
Cluster	VegHeight	0.355	0.675

Table VII reports the DSM statistics for the different processing options. On the one hand, the cluster-based merging leads to significant improvement in height accuracy, increasing IoU-3 for ground and building classes by 0.5% and 0.4%, respectively. On the other hand, the systematic vegetation height difference correction only has a small impact.

[1]https://github.com/pubgeo/dfc2019

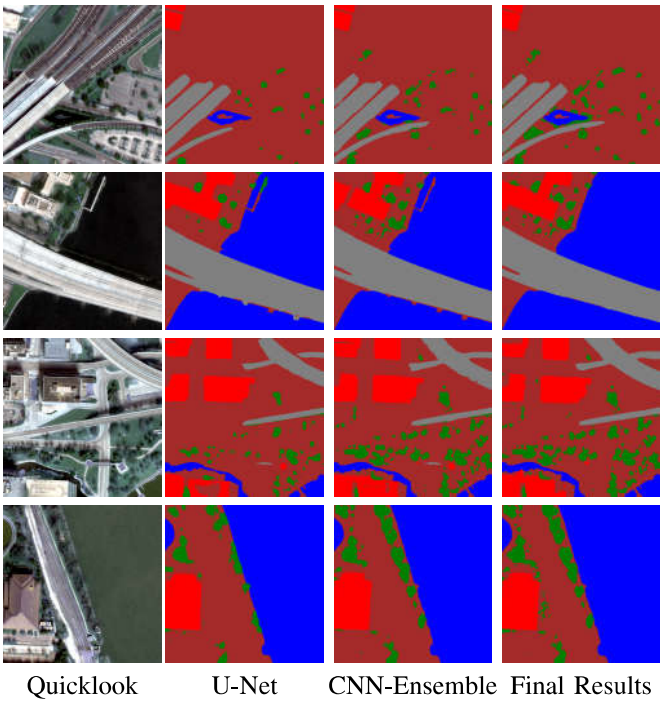


Fig. 16. Examples for semantic classification results. From left to right: true color combination of the median multispectral image, baseline U-Net classification, CNN-Ensemble classification, and final results after post-processing. A water mask is computed separately and overlaid on all results. Ref. Fig. 2 for a legend of the classes.

The impact of different classification and post-processing steps is shown in Table VIII. The first row uses a basic median fusion for the DSM generation and the baseline U-Net and NDWI-based water detection without morphological refinement. The second row reports the results of the CNN ensemble, without further refinements in the classification. The classification benefits from the fusion of the network outputs, particularly the class high vegetation, for which an accuracy improvement of approximately 5% is observed. Finally, the last row shows the final results of the complete process. Further post-processing includes morphological filtering, improving the accuracy for each class slightly by up to 0.5% for the class buildings.

Examples of the different classification strategies are reported in Fig.16. Improvements in the classes buildings, elevated roads, and high vegetation are evident when switching from the U-Net to the ensemble CNN classifier, and after post-processing.

The results show that the basic CNN ensemble with overlaid water masks, without extensive post-processing, would have been sufficient to win the contest. Nevertheless, the cluster-based height merging, water mask, high vegetation, and elevated roads post-processing led to a further overall improvement of 1%.

The semantic classification yielded the best performance across all tracks using multispectral images (Tracks 1, 2, and 3). This may be due to the inclusion of parameters derived from the multi-view processing, which allows for improvements to the water and elevated road classes, which

TABLE VIII
EVALUATION SCORES OF DIFFERENT CLASSIFIERS AND DSM COMBINATIONS.

DSM Fusion	Semantic Segmentation	mIOU-3	mIOU	Ground	High Vegetation	Building	Water	Bridges
Median-Fusion	UNet + Water Mask	0.718	0.782	0.819	0.509	0.809	0.949	0.823
Median-Fusion	CNN-Ensemble + Water Mask	0.736	0.798	0.827	0.564	0.803	0.953	0.843
Cluster-Fusion	All	0.746	0.806	0.831	0.571	0.814	0.958	0.855

is not possible for Tracks 1 and 2.

D. Discussion

The final contest results show that, while CNNs are indispensable for high-quality semantic segmentation, they can still be improved using traditional methods for specific tasks such as water detection. For DSM generation from multi-view data, classical non-deep learning approaches based on SGM were used by the top three entries, indicating that more work needs to be done on CNN-based stereo algorithms to reach the accuracy of traditional methods, especially when many stereo pairs are available. While our work includes some integration between semantic segmentation and DSM generation in the form of using height information in the multimodal fusion network, and semantic segmentation results during DSM merging, future work could further benefit from a tighter integration of both semantic segmentation and stereo matching. Our entry won the competition by a margin of 1.46%, of which 1% was the result of post-processing aimed at improving the semantic segmentation of large buildings, bridges, and high vegetation. In spite of the final semantic segmentation score of the top two teams being comparable, the better DSM owing to bundle adjustment and mature implementation of SGM led to a final difference of 1.46% for mIoU-3.

VII. CONCLUSIONS

Geometric and semantic analyses of images have long been treated independently. However, the increasing abundance of available images and reference data as well as the use of mature methods for image-based 3D reconstruction and the semantic analysis of 2D and 3D information, have allowed the research directions to be combined in recent years. Corresponding approaches solve both tasks simultaneously and are able to provide semantically annotated 3D models, which are of great importance in a wide range of applications ranging from urban planning to monitoring of natural environments.

The 2019 Data Fusion Contest of the Image Analysis and Data Fusion (IADF) Technical Committee of the IEEE Geoscience and Remote Sensing Society addressed the challenges within the context of semantic 3D reconstruction for various aspects by providing high-resolution image data, LiDAR based 3D reference data, and semantic annotations for two different sites, resulting in 69 image tiles with more than 320 GB of data. This allows for the efficient benchmarking of methods that aim to solve large-scale semantic 3D reconstruction tasks.

The contest was arranged in four different tracks corresponding to different input modalities. Tracks 1–3 addressed image-based semantic 3D reconstruction based on a single image, a stereo image pair, or multiple images, respectively, while Track 4 addressed the semantic annotation of point clouds. This first Part A of a two-part paper describes the data modalities, challenges, performance metrics, and winning approaches of Tracks 1–3, while Part B provides an in-depth discussion on Track 4.

Despite the challenges of this contest, for example, the large amount of data, participation numbers continued to increase compared to previous years [21]. With a total of 45 participating countries and winning approaches from China, Germany, Nepal, and the USA, DFC19 was a truly global event. All winners used FCNs for the semantic analysis of the data, while 3D estimation was performed via modern deep learning methods (e.g., PSMNet [46], as in Section V) or more traditional approaches (such as SGM [53], as in Section VI). Finally, various methods for post-processing as well as using ensembles of multiple predictors, allowed for significant performance gains.

After the contest, the data remained accessible for further research on the globally accessible data platform IEEE DataPort[2] and the evaluation servers were re-opened and made accessible on the contest website[3]. While addressing semantic 3D at such a scale was unprecedented, many promising improvements can already be foreseen. In addition to scaling up (with more scenes and more semantic classes, including objects and underrepresented land-use classes), new problems can also be addressed, such as estimation of physical variables (e.g., albedo, atmosphere, and aerosol measurements) or the evolution along the time dimension (temporal changes in semantics and in 3D). Moreover, it will be crucial to investigate how currently emerging machine-learning techniques such as weak supervision and self-supervised learning can be harnessed to build well-performing models.

ACKNOWLEDGMENTS

The authors would like to thank K. Foster, G. Christie, M. Bosch, S. Almes, and S. Wang for their contributions in preparing the dataset and baseline code.

Preparation and public release of the challenge data and code was supported by the Intelligence Advanced Research Projects Activity (IARPA). Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA or the U.S. Government.

[2]https://ieee-dataport.org/open-access/data-fusion-contest-2019-dfc2019
[3]http://www.grss-ieee.org/community/technical-committees/data-fusion/2019-ieee-grss-data-fusion-contest/

REFERENCES

[1] Y. Lian, T. Feng, J. Zhou, M. Jia, A. Li, Z. Wu, L. Jiao, M. Brown, G. Hager, N. Yokoya, R. Hänsch, and B. Le Saux, "Large-Scale Semantic 3D Reconstruction: Outcome of the 2019 IEEE GRSS Data Fusion Contest - Part B," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, submitted.

[2] A. Wehr and U. Lohr, "Airborne Laser Scanning - An introduction and overview," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 54, no. 2, pp. 68 – 82, 1999.

[3] D. Poli and T. Toutin, "Review of developments in geometric modelling for high resolution satellite pushbroom sensors," *The Photogrammetric Record*, vol. 27, pp. 58–73, 2012.

[4] T. Westin, "Precision rectification of spot imagery," *Photogrammetric Engineering & Remote Sensing*, vol. 56, pp. 247–253, 1990.

[5] Z. Li and A. Gruen, "Automatic DSM generation from linear array imagery data," *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 35, 01 2004.

[6] K. Jacobsen, "Dem generation by spot hig resolution stereo," *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 35, 01 2004.

[7] M. Bosch, Z. Kurtz, S. Hagstrom, and M. Brown, "A multiple view stereo benchmark for satellite imagery," in *IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, Oct 2016, pp. 1–9.

[8] G. Facciolo, C. de Franchis, and E. Meinhardt-Llopis, "Automatic 3D Reconstruction From Multi-Date Satellite Images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[9] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce, "Comparison of pansharpening algorithms: Outcome of the 2006 GRS-S Data Fusion Contest," *IEEE Trans. Geosci. Remote Sensing*, vol. 45, no. 10, pp. 3012–3021, 2007.

[10] F. Pacifici, F. Del Frate, W. J. Emery, P. Gamba, and J. Chanussot, "Urban mapping using coarse SAR and optical data: Outcome of the 2007 GRSS Data Fusion Contest," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 3, pp. 331–335, July 2008.

[11] G. Licciardi, F. Pacifici, D. Tuia, S. Prasad, T. West, F. Giacco, J. Inglada, E. Christophe, J. Chanussot, and P. Gamba, "Decision fusion for the classification of hyperspectral data: Outcome of the 2008 GRS-S data fusion contest," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3857–3865, 2009.

[12] N. Longbotham, F. Pacifici, T. Glenn, A. Zare, M. Volpi, D. Tuia, E. Christophe, J. Michel, J. Inglada, J. Chanussot, and Q. Du, "Multi-modal change detection, application to the detection of flooded areas: outcome of the 2009-2010 Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 331–342, 2012.

[13] F. Pacifici and Q. Du, "Foreword to the special issue on optical multiangular data exploitation and outcome of the 2011 GRSS Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 3–7, 2012.

[14] C. Berger, M. Voltersen, R. Eckardt, J. Eberle, T. Heyer, N. Salepci, S. Hese, C. Schmullius, J. Tao, S. Auer, R. Bamler, K. Ewald, M. Gartley, J. Jacobson, A. Buswell, Q. Du, and F. Pacifici, "Multi-modal and multi-temporal data fusion: Outcome of the 2012 GRSS Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 3, pp. 1324–1340, 2013.

[15] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens, A. Pizurica, S. Gautama, W. Philips, S. Prasad, Q. Du, and F. Pacifici, "Hyperspectral and LiDAR Data Fusion: Outcome of the 2013 GRSS Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, 2014.

[16] W. Liao, X. Huang, F. V. Coillie, S. Gautama, A. Pizurica, W. Philips, H. Liu, T. Zhu, M. Shimoni, G. Moser, and D. Tuia, "Processing of multiresolution thermal hyperspectral and digital color data: Outcome of the 2014 ieee grss data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 6, pp. 2984–2996, June 2015.

[17] M. Campos-Taberner, A. Romero-Soriano, C. Gatta, G. Camps-Valls, A. Lagrange, B. Le Saux, A. Beaupère, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, M. Ferecatu, M. Shimoni, G. Moser, and D. Tuia, "Processing of extremely high resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS Data Fusion Contest. Part A: 2D contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 12, pp. 5547–5559, 2016.

[18] A.-V. Vo, L. Truong-Hong, D. Laefer, D. Tiede, S. d'Oleire Oltmanns, A. Baraldi, M. Shimoni, G. Moser, and D. Tuia, "Processing of extremely high resolution LiDAR and RGB data: Outcome of the 2015 IEEE GRSS Data Fusion Contest. Part B: 3D contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 12, pp. 5560–5575, 2016.

[19] L. Mou, X. Zhu, M. Vakalopoulou, K. Karantzalos, N. Paragios, B. Le Saux, G. Moser, and D. Tuia, "Multi-temporal very high resolution from space: Outcome of the 2016 IEEE GRSS Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3435–3447, 2017.

[20] N. Yokoya, P. Ghamisi, J. Xia, S. Sukhanov, R. Heremans, I. Tankoyeu, B. Bechtel, B. Le Saux, G. Moser, and D. Tuia, "Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1363–1377, May 2018.

[21] Y. Xu, B. Du, L. Zhang, D. Cerra, M. Pato, E. Carmona, S. Prasad, N. Yokoya, R. Hänsch, and B. Le Saux, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS Data Fusion Contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, June 2019.

[22] M. Bosch, K. Foster, G. Christie, S. Wang, G. D. Hager, and M. Brown, "Semantic stereo for incidental satellite images," in *Winter Conf. on Applications of Computer Vision (WACV)*, 2019, pp. 1524–1532.

[23] G. Christie, R. Munoz, K. Foster, S. Hagstrom, G. Hager, and M. Brown, "Learning geocentric object pose in oblique monocular images," in *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[24] B. Le Saux, N. Yokoya, R. Hänsch, and M. Brown, "IEEE Dataport: Data Fusion Contest 2019," 2019. [Online]. Available: http://dx.doi.org/10.21227/c6tm-vw12

[25] K. Foster, G. Christie, and M. Brown, "IEEE Dataport: Urban semantic 3d dataset," 2020. [Online]. Available: http://dx.doi.org/10.21227/9frn-7208

[26] "GitHub: Data Fusion Contest 2019." [Online]. Available: https://github.com/pubgeo/dfc2019

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

[28] 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.

[29] P. Yakubovskiy, "Segmentation models," https://github.com/qubvel/segmentation_models, 2019.

[30] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for Real-Time Semantic Segmentation on High-Resolution Images," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[31] R. Atienza, "Fast disparity estimation using dense networks," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3207–3212.

[32] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008. [Online]. Available: https://doi.org/10.1109/TPAMI.2007.1166

[33] S. Kunwar, "U-Net ensemble for semantic and height estimation using coarse-map initialization," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Yokohama, Japan, 2019.

[34] Z. Zheng, Y. Zhong, and J. Wang, "Pop-Net: Encoder-dual decoder for semantic segmentation and single-view height estimation," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Yokohama, Japan, 2019.

[35] H. Chen, M. Lin, H. Zhang, G. Yang, G. Xia, X. Zheng, and L. Zhang, "Multi-level fusion of the multi-receptive fields contextual networks and disparity network for pairwise semantic stereo," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Yokohama, Japan, 2019.

[36] R. Qin, X. Huang, W. Liu, and C. Xiao, "Pairwise stereo image disparity and semantics estimation with the combination of U-Net and pyramid stereo matching network," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Yokohama, Japan, 2019.

[37] P. d'Angelo, D. Cerra, S. M. Azimi, N. Merkle, J. Tian, S. Auer, M. Pato, R. de los Reyes, X. Zhuo, K. Bittner, T. Krauss, and P. Reinartz, "3D Semantic Segmentation from Multi-View Optical Satellite Images," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Yokohama, Japan, 2019.

[38] R. Qin, X. Huang, W. Liu, and C. Xiao, "Semantic 3D reconstruction using multi-view high-resolution satellite images based on U-Net and image-guided depth fusion," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Yokohama, Japan, 2019.

[39] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[41] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[42] I. Sutskever, J. Martens, G. E. Dahl, and G. E. Hinton, "On the importance of initialization and momentum in deep learning." *ICML (3)*, vol. 28, no. 1139-1147, p. 5, 2013.

[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[44] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366–2374.

[45] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent," *Cited on*, vol. 14, 2012.

[46] J. Chang and Y. Chen, "Pyramid stereo matching network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 5410–5418.

[47] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 833–851.

[48] Z. Wang and S. Ji, "Smoothed dilated convolutions for improved dense prediction," in *24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 07 2018, pp. 2486–2495.

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

[50] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters - improve semantic segmentation by global convolutional network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 1743–1751.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[52] P. d'Angelo, "Automatic orientation of large multitemporal satellite image blocks," in *International Symposium on Satellite Mapping Technology and Application*, November 2013, pp. 1–6. [Online]. Available: https://elib.dlr.de/88017/

[53] P. d'Angelo and G. Kuschk, "Dense multi-view stereo from satellite imagery," in *IGARSS 2012*, no. DOI: 10.1109/IGARSS.2012.6352565. IEEE Press, November 2012, pp. 6944–6947.

[54] R. Perko, H. Raggam, K. Gutjahr, and M. Schardt, "Advanced DTM generation from very high resolution satellite stereo images," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, no. Volume II-3/W4, 2015.

[55] S. K. McFeeters, "The use of the normalized difference water index (NDWI) in the delineation of open water features," *International Journal of Remote Sensing*, vol. 17, no. 7, pp. 1425–1432, 1996. [Online]. Available: https://doi.org/10.1080/01431169608948714