International Journal of Remote Sensing

ISSN: 0143-1161 (Print) 1366-5901 (Online) Journal homepage: https://www.tandfonline.com/loi/tres20

A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery

Shunping Ji, Shiqing Wei & Meng Lu

To cite this article: Shunping Ji, Shiqing Wei & Meng Lu (2019) A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery, International Journal of Remote Sensing, 40:9, 3308-3322, DOI: 10.1080/01431161.2018.1528024

To link to this article: https://doi.org/10.1080/01431161.2018.1528024

Published online: 16 Oct 2018.

Submit your article to this journal

Article views: 1189

View related articles

View Crossmark data

Citing articles: 40 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tres20

INTERNATIONAL JOURNAL OF REMOTE SENSING
2019, VOL. 40, NO. 9, 3308–3322
https://doi.org/10.1080/01431161.2018.1528024

Taylor & Francis
Taylor & Francis Group

Check for updates

A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery

Shunping Ji[a], Shiqing Wei[a] and Meng Lu[b]

[a]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China; [b]Department of physical geography, Utrecht University, Utrecht, The Netherlands

ABSTRACT

Identifying buildings from remote sensing imagery has been a challenge due to uncertainties from remote sensing imagery and variations in building structure and texture. In this study, we develop a scale robust CNN structure to improve the segmentation accuracy of building data from high-resolution aerial and satellite images. Based on a fully convolutional network, we introduce two Atrous convolutions on the first two lowest-scale layers, respectively, in the decoding step, aiming at enlarging the sight-of-view and integrate semantic information of large buildings. Then, a multi-scale aggregation strategy is applied. The last feature maps of each scale are used to predict the corresponding building labels, and further up-sampled to the original scale and concatenated for the final prediction. In addition, we introduce a combined data augmentation and relative radiometric calibration method for multi-source building extraction. The method enlarges sample spaces and hence the generalization ability of the deep learning models. We validate our developed methods with an aerial dataset of more than 180, 000 buildings with various architectural types, and a satellite image dataset consists of more than 29,000 buildings. The results are compared with several most recent studies. The comparison result shows our neural network outperformed other studies, especially in segmenting scenes of large buildings. The test on transfer learning from aerial dataset to satellite dataset showed our augmentation strategy significantly improved the prediction accuracy; however, further studies are needed to improve the generalization ability of the CNN model.

ARTICLE HISTORY
Received 28 April 2018
Accepted 9 September 2018

1. Introduction

Building extraction has important implications in urban planning, population estimation, and topographic map creating and updating. Automatic building detection from aerial and satellite images has been extensively studied during the last thirty years. In 1989, Liow and Pavlidis (Liow and Pavlidis 1989) proposed methods to integrate region growing and edge detection to extract buildings in aerial images. The commonly used feature metrics include colour (Sirmacek and Unsalan 2008), spectrum (Zhong, Huang, and Xie 2008; Zhang 1999), length, edge (Li and Huayi 2008; Ferraioli 2010), shape

CONTACT Shunping Ji jishunping@whu.edu.cn School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan, China

© 2018 Informa UK Limited, trading as Taylor & Francis Group

(Dunaeva, Valer'evna, and Kornilov 2017), texture (Awrangjeb, Zhang, and Fraser 2011), shadow (Liow and Pavlidis 1989; Sirmacek and Unsalan 2008; Chen, Shang, and Chengdong 2014), height, semantic (Zhong et al. 2015), etc. In (Ma and Meyer 2005), a method is proposed to extract ground points to generate DEM and to detect points that belong to buildings. Similarly, a building detection technique by only using the height information of DSM has been proposed (Zhou et al. 2013). (Sirmacek and Unsalan 2009) proposed the use of scale invariant feature transform (SIFT) and graph theoretical tools to detect buildings in satellite images. (Rottensteiner et al. 2005) presented a building detection method by the fusion of the first and last pulse laser scanner data and multi-spectral images. However, due to the diverse building appearance and environmental characters such as light and background, as well as the variety in sensors and scale, these manually designed metrics and features could vary largely. Therefore, it is difficult to generalize the characteristics of buildings. Building detection using these methods are subject to practical applications and a general and reproducible method was basically unavailable.

In recent years, deep learning has been applied to remote sensing measurements including building segmentation, in replacement with the empirical feature design process by automatically learning multi-level representations (Maggiori et al. 2016; Yuan 2017). Since 2012, CNN has been extensively applied to image-wise classification, and novel CNN structures such as AlexNet (Krizhevsky, Sutskever, and Hinton 2012), VGGNet (Simonyan and Zisserman 2014), GoogLeNet (Szegedy et al. 2015), and ResNet (He et al. 2016) have shown to be successful.

Since 2015, special CNN structures are developed to pixel-wise semantic segmentation and object detection. A group of special CNN, known as region-based model, detects objects by predicting a bounding box of each object. These region-based models include, for example, R-CNN (Girshick et al. 2015), Fast R-CNN (Girshick 2015), Faster R-CNN (Ren et al. 2015) and Mask R-CNN (He et al. 2017). Another mainstream special CNN that is semantic segmentation oriented is the fully convolutional networks (FCN) (Shelhamer, Long, and Darrell 2014), which replaces fully connected layers in traditional CNN with convolutional and transposed convolutional layers. A variety of FCNs have been proposed, such as SegNet (Badrinarayanan, Kendall, and Cipolla 2015), DeconvNet (Noh, Hong, and Han 2015), U-net (Ronneberger, Fischer, and Brox 2015). Up to now, a continuously updated network named DeepLab (Chen et al. 2015, 2018a, 2017, 2018b) has been a new benchmark of semantic segmentation. The most recent studies in building extraction exclusively utilized FCN-based methods. (Maggiori et al. 2016) designed a two-scale neuron module in an FCN to reduce the trade-off between recognition and precise localization. Studies in (Yuan 2017; Maggiori et al. 2017) integrated multiple layers of activation into pixel level prediction based on FCN. Wu et al. (Wu et al. 2018) designed a multi-constraint FCN that utilizes multi-layer outputs.

Although CNN has shown promising ability to detect buildings from remote sensing images, scale invariance and generalization ability are two remaining challenges. Scale invariance considers extracting buildings of heterogeneous sizes from high-resolution images. Since in most cases cropping remote sensing images is currently unavoidable due to limited GPU memory, large buildings occupying most part of an image tile input becomes a common case. The model proposed in (Maggiori et al. 2016) attempts to solve this multi-scale problem by revising the first convolutional layer of an FCN to a

two-scale neuron module; however, this may not be optimal as the multi-scale information can hardly be completely propagated to the very late layers with only a top-down route of a feature pyramid. In (Wu et al. 2018), multi-scale outputs are weighted for training a U-Net based model. Two problems may exist with this strategy. Firstly, the outputs are not physically combined and do not fully leverage the information of feature pyramids in decoding. The model is trained with multi-scale labels but predicts only for the original inputs. Second, for a large building, the FCN and U-Net backbone with conventional convolution can hardly grasp the whole semantic information due to lacking of enough field-of-view. For example, for a typical 4-scale encoding structure, the largest scope of 3×3 kernel is 48×48 pixels, less than one-tenth of a 512×512 input image in width.

Other possible strategies to address multi-scale problem include Atrous convolution (Chen et al. 2018a) and multi-scale aggregation, which have not been applied to building extraction. The former can tremendously expand the field-of-view and will benefit for large object segmentation. A specially designed form of the latter could physically combine all of the multi-scale outputs instead of weighting them to obtain better scale invariance. We develop our building segmentation method with the two new features based on an FCN structure.

The generalization ability of a CNN on multi-source building extraction is another challenge. Extracting Buildings from combined aerial and satellite data is a common case but has not been addressed in recent studies since CNN has been applied to extract buildings. The similarity between training and test dataset affects significantly the performance of a deep learning based method. A pre-trained model may fail to direct application on different remote sensing sources. The radiometric variation of source and target dataset is the main obstacle.

A commonly used strategy is data augmentation, which resamples original input under various situations including geometric and/or radiometric transformations to expand sample space and hence the generalization ability of the trained model. Ronneberger et al. (Ronneberger, Fischer, and Brox 2015) used random elastic deformation to expand the manifold space of training samples and obtained improved training results. In our study, we introduce a data augmentation strategy that combines a relative radiometric calibration to source and target datasets and a radiometric resampling to improve the generalization ability of multi-source deep learning.

Our main contribution lies in proposing a scale robust CNN structure combining Atrous convolution and multi-scale aggregation for extracting buildings from high-resolution remote sensing images, which is described in section 2. In addition, we introduce a combined data augmentation strategy that combines a relative radiometric calibration and a radiometric calibration for transfer learning from aerial dataset to satellite dataset and evaluate the generalization ability of CNN on the multi-source building extraction. In section 3, we compare our method with both recent building extraction methods and benchmark algorithms of general semantic segmentation on an open aerial and satellite dataset. Discussion that addresses the CNN model transfer learning from aerial to satellite data sources and conclusions are given in section 4 and 5 respectively.

2. Network structure

A typical CNN structure for pixel-wise segmentation consists of an encoder and a decoder with fully convolutional layers. In an encoder, the original inputs are convoluted and down-sampled layer by layer, to obtain higher semantic features with lower spatial resolution until the feature maps with the coarsest spatial resolution are obtained. The encoder could be any popular CNN backbone, such as the VGG (Simonyan and Zisserman 2014) used in the FCN (Shelhamer, Long, and Darrell 2014) and the U-Net (Ronneberger, Fischer, and Brox 2015), the ResNet in (He et al. 2016), the Xception in DeepLabv3+ (Chen et al. 2018b). In this work, we select VGG-16 as our encoder which has been experimentally showed to have high performance and efficiency.

In a decoder, the lower layer features are up-convoluted layer by layer to the original scale for training with given labels or for prediction. The right part of each photograph in Figure 1 shows some typical decoder structures for semantic segmentation. The FCN (Shelhamer, Long, and Darrell 2014) and Deeplabv3+ (Chen et al. 2018b) expand the coarsest feature map directly to the original size whereas the U-net (Ronneberger, Fischer, and Brox 2015) and the feature pyramid network (FPN) (Lin et al. 2017) use a ladder structure as that is in the encoding. The FPN predicts in all scales, different from the other encoders that predict only in the original scale. We developed our scale-invariant decoder based on the previous semantic segmentation studies.

Our network is shown in Figure 2 and we call it a scale-robust FCN (SR-FCN). Firstly, we concatenate the multi-layer features extracted in VGG-16 encoder to the features of corresponding scales in decoding (green arrows shown in Figure 2). This is similar to U-Net and FPN but different from FCN and DeepLab where the features in encoding are not fully integrated into features in decoding.

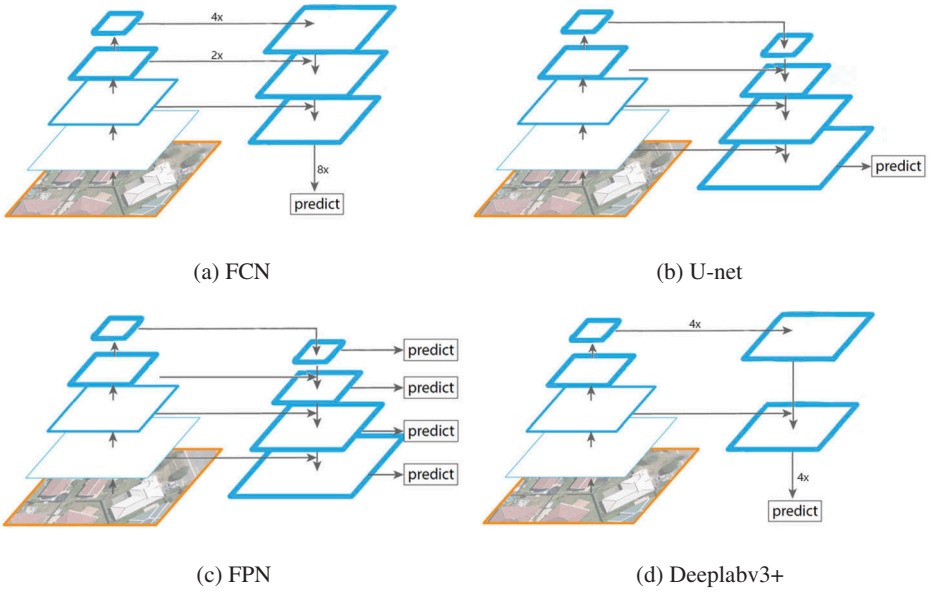

Figure 1. Decoders used in classic CNNs for semantic segmentation. (a) FCN. (b) U-net. (c) FPN. (d) Deeplabv3 +.

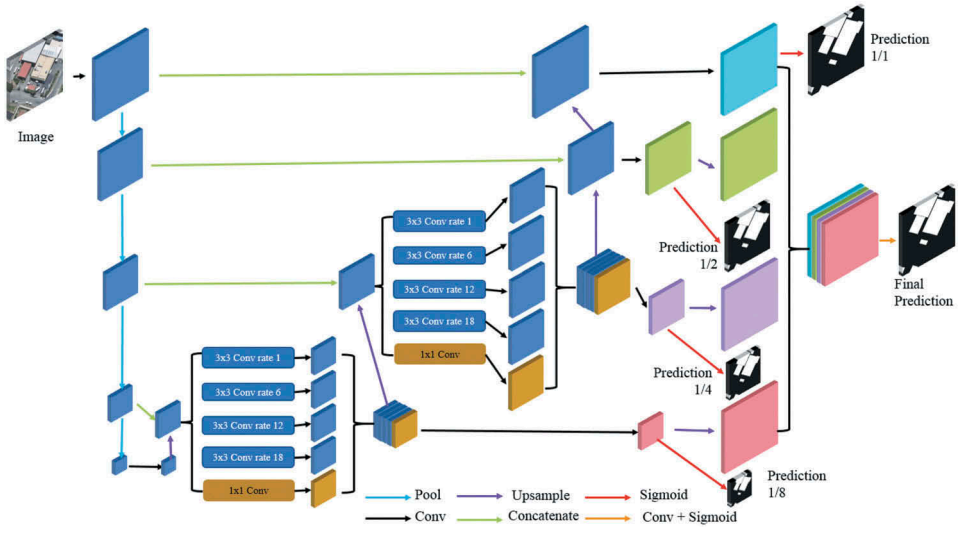

Figure 2. The structure of SR-FCN.

Second, we introduce two Atrous convolutions in the 1/4 and 1/8 scales to extract features with much larger field-of-view. Each Atrous layer consists of four 3 × 3 Atrous convolution with the rate r = 1, 6, 12, and 18, respectively, and one 1 × 1 convolution. The rate is related to stride, and kernel size can be retrieved according to k+(k-1)×(r-1), where k is the original kernel size (here 3). For a 512 × 512 input, an Atrous kernel with the rate 18 has a kernel size of 37 × 37 in 1/8 scale, equalling a field-of-view of 296 × 296 pixels in original scale. The expansion of the field-of-view positively functions on our high-resolution aerial building dataset where many large buildings occupying the most part of input images. Our Atrous convolution is different from the work of DeepLab v3+ (Chen et al. 2018b) where only one multi-scale Atrous convolution is used in 1/8 scale in encoder. As we have additionally concatenated features in encoding to the corresponding features in decoding, it is reasonable to apply the Atrous convolution in the concatenated features.

Third, all features from different scales are up-sampled to the original scale and concatenated, followed by the final prediction. Except this aggregated output, we found predicting results on separating four scales (the four red arrows) can help training more accurate model and get better results. Thus, the final loss of the SR-FCN is:

$$\text{Loss} = \sum_{i=1}^{n+1} \lambda_i L_i \tag{1}$$

Where n = 4, λ = 1 in our study and each Loss L is computed utilizing a cross-entropy:

$$L = -\frac{1}{n}\sum_{i=1}^{n} g^i \ln p^i + (1 - g^i)\ln(1 - p^i) \tag{2}$$

This multi-scale aggregation strategy utilizes all the information from different scales both in separate and integrated manner. Whereas in U-Net and FCN, only the last layer with full-resolution is used for prediction; in C-Unet (Wu et al. 2018) the separate multi-scale outputs are weighted but not physically combined; in DeepLabv3+ (Chen et al. 2018b) the feature map in 1/8 scale are simply up-sampled for the final single output.

3. Datasets experimental setting and results

3.1. *Datasets and settings*

A large datasets collection of aerial and satellite datasets called the WHU Building dataset (Ji, Wei, and Meng 2018) (http://study.rsgis.whu.edu.cn/pages/download/) is utilized for testing our method. The WHU aerial dataset (Figure 3) covers 18,7000 buildings of various architectures and colours. The whole image and the corresponding vector shape file were seamlessly cropped into 8189 512 × 512 tiles with 0.3 m ground resolution. The area in the blue box contains 130,000 buildings and is used for training, the area in the yellow box containing 14,500 buildings is used for validation and the rest in red box containing 42,000 buildings is used for testing.

The WHU satellite dataset consists of 6 neighbouring satellite images covering 550 km^2 on East Asia with 2.7 m ground resolution (Figure 4). The images with different colours and from different sensors and seasons form a challenging case for automatic building extraction. The vector building map is available containing 29,085 buildings. The whole image is also seamlessly cropped into 17,388 512 × 512 tiles for convenient training and testing with the same processing as in the aerial dataset. Among them, 21,556 buildings (13,662 tiles) are separated for training and the rest 7529 buildings (3726 tiles) are used for testing.

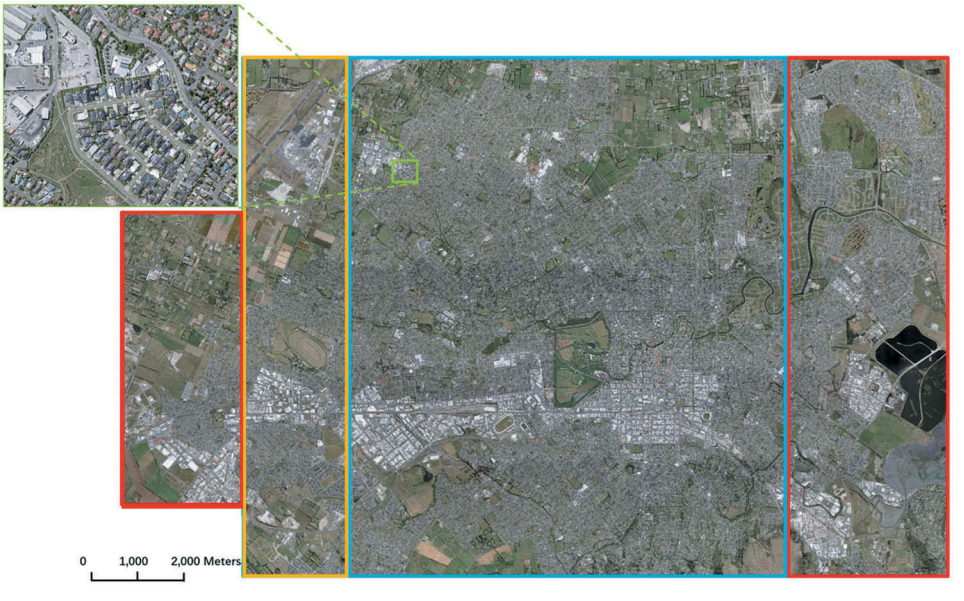

Figure 3. The aerial dataset in Christchurch, New Zealand, covering 450 km^2. The areas in the blue, yellow and red bounding boxes are used for training, validation and test, respectively.

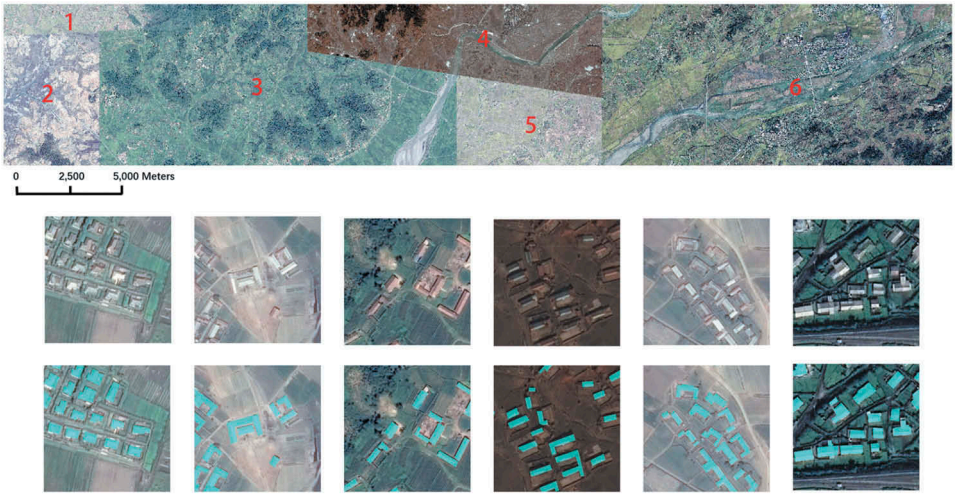

Figure 4. Six neighbouring satellite images covering 550 km^2 on East Asia with 2.7 m ground resolution.

The implementation is based on the Keras platform using a TensorFlow backend. An Adam (adaptive moment estimation) algorithm is used as a random gradient descent optimization with 6 image tiles as a mini-batch. The learning rate is set to 0.0001. The weights of all filters are initialized according to a normal distribution initialization method (He et al. 2015), and all biases are initialized to zeros. On a Nvidia Titan Xp GPU, 3 h are required to train the aerial data and 1 h to train the satellite data.

All the output values from a sigmoid classifier are translated to binary values through given a threshold 0.5. Three indicators, IoU, precision, and recall, are used to evaluate the performance of all the methods used in experiments. IoU is the ratio between the intersection of the detected building pixels and the true positive pixels and their union. Precision is the percentage of the true positive pixels among all detected building pixels. Recall is the percentage of the true positive pixels among building pixels in ground truth.

3.2. *SR-FCN on the aerial dataset*

Table 1 shows the results of our method, Deeplabv3+ (Chen et al. 2018b), C-Unet (Wu et al. 2018), U-net (Ronneberger, Fischer, and Brox 2015), FCN-8s (Shelhamer, Long, and Darrell 2014) and 2-scale FCN (Maggiori et al. 2016) on the aerial dataset. Our result showed 1.8% improvement compared to the latest DeepLabv3+ and C-Unet, and 2.1%,

Table 1. Comparison of the most recent deep learning based methods on the aerial dataset.

Method	IoU	Precision	Recall
Ours	0.889	0.944	0.939
DeepLabv3+	0.871	0.916	0.946
C-Unet	0.871	0.917	0.946
U-Net	0.868	0.945	0.914
FCN-8s	0.854	0.892	0.953
2-scale FCN	0.701	0.758	0.903

3.5% and 18.8% improvement compared to the U-net, FCN-8s and 2-Scale FCN, respectively. As a high level of segmentation accuracy has been achieved (87% IoU), an increase of 1.8 IoU points could be significant. This increase comes from introducing the network with both the aggregation of multi-scale outputs and the Atrous convolutions that help cover a much larger scene. Whereas Deeplabv3+ only utilizes the encoding output by 8× up-sampling as the final output; C-Unet only weighted the first and last scale outputs without the Atrous convolution. For high-resolution images, the Atrous convolution and separate outputs especially in lower spatial resolution grab the high-level semantic information. Additionally, the aggregation of multi-scale outputs contributes to improve the precision score (2.8% improvement compared to DeepLabv3 + and C-Unet) as a more rigorous constraint.

These findings are further shown in Figure 5. The prediction results of the first four rows show that our method is better than other methods in classifying large buildings.

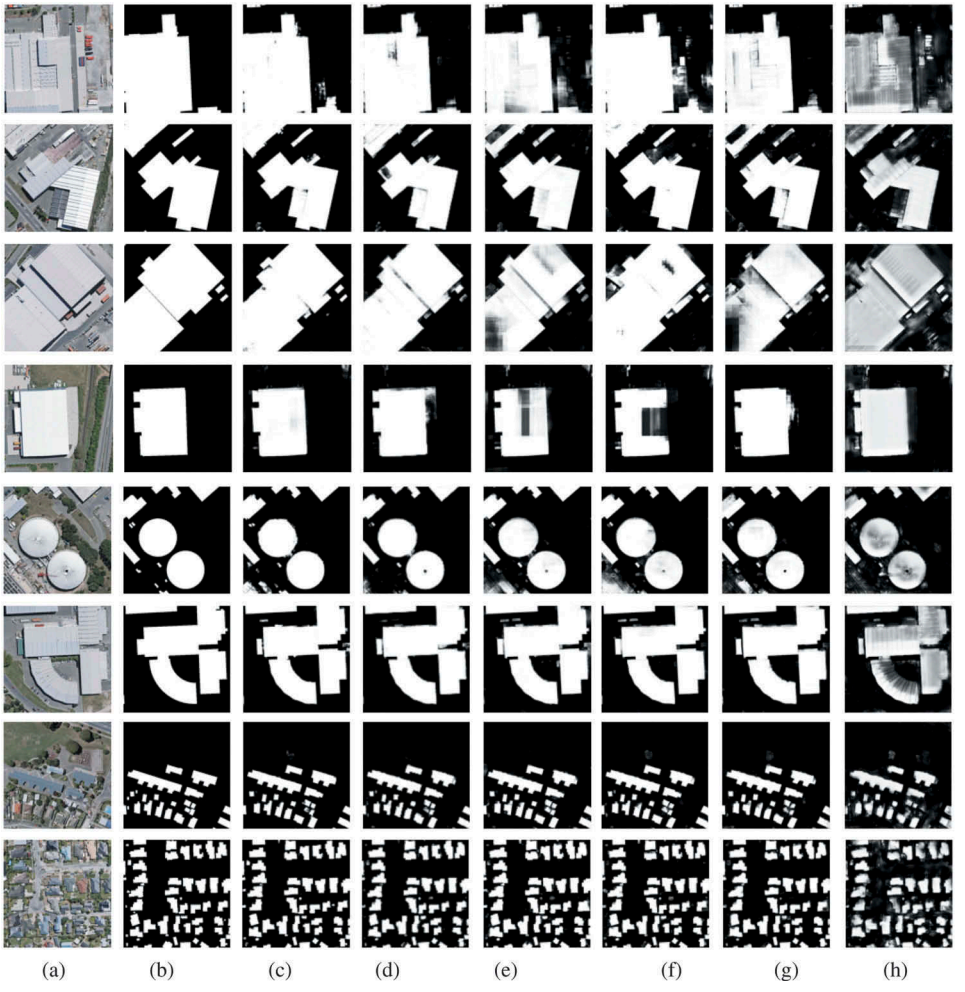

Figure 5. Comparison between our method (c), DeepLab (d), C-Unet (e), Unet (f), FCN-8s (g), 2-scale FCN (h) on the WHU aerial dataset. (a) and (b) are images and labels.

Table 2. Comparison of recent deep learning based methods on the satellite dataset.

Method	IoU	Precision	Recall
Ours	0.640	0.790	0.770
DeepLabv3+	0.635	0.787	0.766
C-Unet	0.640	0.717	0.856
U-Net	0.627	0.812	0.734

Due to the limited sight-of-view and lack of multi-scale aggregation, the other methods occasionally missed a small or big part of a whole building although the pixels of the building exhibit the same colour and texture and are seemingly easy to be segmented completely. For the case of classifying small buildings as in the last two rows in Fig. 9, the top four methods perform almost the same. These results indicate that our method segments large objects more accurately, which is beneficial for high-resolution image classification or segmentation.

3.3. *SR-FCN on the satellite dataset*

We apply the four best-performenced methods on the aerial dataset to the satellite dataset and list the prediction results in Table 2. From Table 2, our method performed similarly to the C-Unet and outperform the DeepLabv3+ and the U-Net marginally. As the satellite dataset possesses a 2.7 m ground resolution (compared to 0.3 m in the aerial dataset) and few large buildings exist in this area, it is reasonable that the Atrous convolution has not much effects and both our methods and C-Unet won other methods with the multi-scale aggregation.

In the first row of Figure 6, our method seems clearer and performs better. The predictions in the second row show a challenge case with low contrast between buildings and background. Our methods perform slightly better but all methods failed to extract buildings clearly. In the third-row, algorithms are required to predict a long building under low contrast. The shortcoming of the U-Net is visually demonstrated (the third row of Figure 6(c)): it predicts only on the original scale by single-channel layer-to-layer information transfer in decoding step, which results information loss. Our method and the C-Unet overcome this with multi-scale aggregation whereas the DeepLabv3 + addresses this by direct upsampling from low spatial, high semantic layer. The last row of Figure 6 showed an uncommon large building in this area, which also has roofs with unusual different colours. All methods failed to extract the whole building but DeepLabv3+ perform best again for its mechanism of a simple decoder that only depends on the lowest resolution features.

4. Discussion

4.1. *Data augmentation and relative radiometric calibration for multi-source datasets*

In this section, we focus on transfer learning of a CNN model that is trained from an aerial image dataset and applied to a satellite dataset. CNN network trained using an aerial dataset may not perform well on a satellite dataset as radiometric distortions exist.

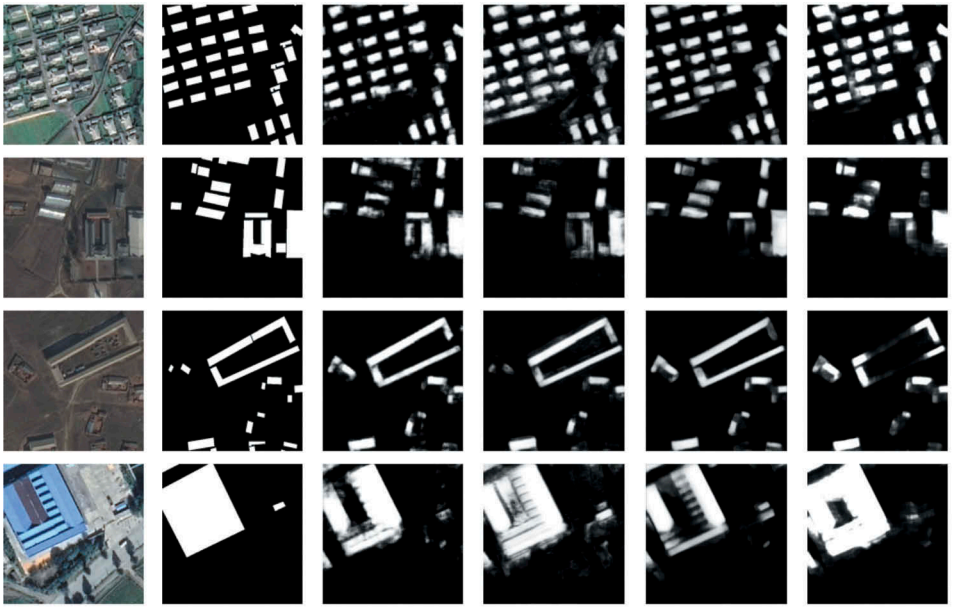

Figure 6. Comparison between our method (c), DeepLab (d), C-Unet (e), U-Net (f) on the WHU satellite dataset. (a) and (b) are images and labels.

We introduce a data augmentation and relative radiometric calibration strategy and evaluate their performances. The former expands the sample space and hence improve the generalization ability of a deep learning model, and the latter transforms the sample space of the target dataset similar to the sample space of the source dataset.

The relative radiometric correction is implemented with a Wallis filtering (Zhang, Zhang, and Zhang 1999), which transfers the mean and covariance of pixels of the reference image to the target images. The Wallis filtering formula is:

$$g(x,y) = (f(x,y) - m_s)\frac{cv_d}{cv_s + (1-c)v_d} + bm_d + (1-b)m_s \tag{3}$$

Where $g(x,y)$ and $f(x,y)$ are pixel values of the target image and source image in position (x, y). m and v are the mean and variance of the image pixels. The subscript s and d represent the source image and the target image, respectively, c and b are the specified coefficients between 0 ~ 1.

In our test, the source image is obtained from a large down-sampled part of the whole aerial training area that could reflect the general radiometric feature.

Data augmentation strategy is applied to all input images. Under the same bird's eye view, we only concern the radiometric distortion caused by atmospheric radiation transmission. A counterpart generator first randomly draws sample values from the given intervals and distributions of parameters and then resample the original image to a new sample. The parameter set consists of linear stretching, histogram equalization, blur, and salt noise. Histogram equalization changes the grey histogram of the original image to a uniform distribution in the entire grey scale range.

Table 3. Performance evaluation on the satellite dataset by direct prediction using the model pretrained on the aerial dataset. 'Data augmented' indicates the model is pretrained on the data-augmented aerial dataset. 'Calibrated' means the target satellite images have been radiometrically calibrated according to the aerial image.

Method	IoU	Recall	Precision
Direct prediction	0.040	0.109	0.069
Data augmented	0.172	0.359	0.249
Augmented & calibrated	0.181	0.356	0.269

In Gaussian blur, we set the convolution kernel size to 3 or 5, and sigma to 1.5, 2.2, or 3. The ratio of randomly generated salt and pepper noise is set between 0 and 0.02. Plenty of subsamples will be generated during data augmentation, however, only one image is taken at a time for training. The number of iterations is then correspondingly extended for converge.

4.2. *Direct transfer learning from aerial to satellite dataset via data augmentation*

The SR-FCN model trained from aerial images is directly applied to the satellite dataset with fully automatic manner and without training samples. Table 3 shows the results applying SR-FCN on the satellite dataset without augmentation, data augmentation, and both augmentation and relative radiometric correction, respectively. It could be observed that, without data preprocessing, the model has almost no generalization ability (only 4% IoU). The reason could be a combination of radiometric distortion and the dissimilarity between buildings (including buildings styles and backgrounds) in the two datasets. After introducing data augmentation that specially designed for radiometric augmentation, the IoU is increased to 17.2% (330% relative improvement). When relative radiometric calibration is added, the IoU is slightly increased to 18.1%. It indicates that relative radiometric calibration has only slight help when the training samples have been radiometrically augmented.

Figure 7 shows four prediction examples. We can observe that without data augmentation the model predicts almost nothing. When data augmentation and relative radiometric correction are applied, the impact of radiometric distortion between source and target datasets on model transfer is largely reduced, and the effects of all predictions except the third row are significantly improved. However, in general, 18.1% IoU indicates the directly transferred model is basically lacking the generalization ability for multi-source building extraction. Further studies in radiometric calibration, preparing larger training data including various building styles and backgrounds are required to achieve a satisfactory transfer learning between multi-source building extraction.

5. Conclusion

We proposed a scale robust FCN to extract buildings from a large open aerial and satellite dataset, which outperformed the recent studies including the developed methods for building extraction and current benchmarks for general semantic segmentation.

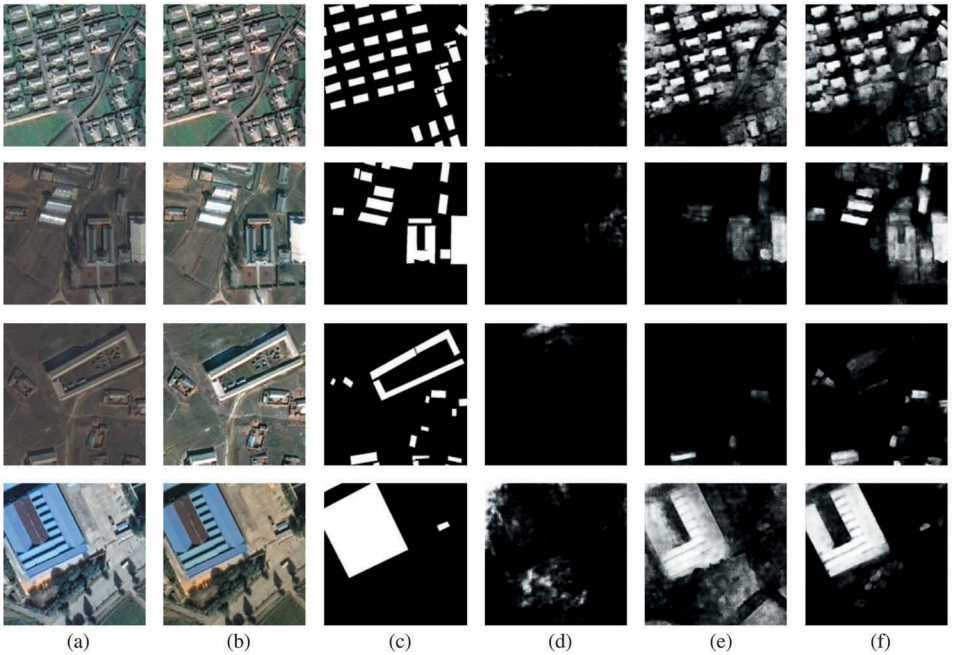

Figure 7. Direct prediction on the satellite dataset using the model pretrained on the aerial dataset. (a) and (c) are images and labels; (b) is the image after relative radiometric calibration; (d), (e) and (f) are the results of original SR-FCN, SR-FCN with radiometric augmentation and SR-FCN with both radiometric augmentation and relative radiometric calibration, respectively.

The designed network structure with the Atrous convolutions and multi-scale aggregation could accurately segment buildings, especially with large covers. We also attempt to extract buildings from different data sources and introduce a radiometric augmentation strategy combining spectral resampling and relative radiometric correction. With this strategy, the model pretrained on the aerial dataset achieved significant improvement when directly applied to the satellite dataset. However, the generalization ability of deep learning based methods on multi-source building extraction should be further improved. The current work on pixel-wise segmentation could be extended to segment single building instances, and ultimately draw a building vector map directly for surveying, mapping and updating. Larger building datasets covering various building styles and complex backgrounds, and better data augmentation strategies may be necessary to further increase the generalization ability of deep learning models.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the the National Natural Science Foundation of China [41471288].

References

Awrangjeb, M., C. Zhang, and C. S. Fraser. 2011. "Improved Building Detection Using Texture Information." *ISPRS - International Archives of the Photogrammetry* XXXVIII-3/W22 (XXXVIII–3/W22): 143–148. doi:10.5194/isprsarchives-XXXVIII-3-W22-143-2011.

Badrinarayanan, V., A. Kendall, and R. Cipolla. 2015. "Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation." *arXiv preprint arXiv:1511.00561*. doi:10.1109/TPAMI.2016.2644615.

Chen, D., S. Shang, and W. Chengdong. 2014. "Shadow-Based Building Detection and Segmentation in High-Resolution Remote Sensing Image." *Journal of Multimedia,9,1(2014-01-01)* 9 (1). doi:10.4304/jmm.9.1.181-188.

Chen, L. C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2015. "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs." *Computer Science* 4: 357–361.

Chen, L. C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2018a. "Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (4): 834–848. doi:10.1109/TPAMI.2017.2699184.

Chen, L. C., G. Papandreou, F. Schroff, and H. Adam. 2017. "Rethinking Atrous Convolution for Semantic Image Segmentation." *arXiv preprint*. *arXiv:1706.05587*.

Chen, L. C., Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. 2018b. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation." *arXiv preprint arXiv:1802.02611*.

Dunaeva, A., Valer'evna, and F. A. Kornilov. 2017. "Specific Shape Building Detection from Aerial Imagery in Infrared Range." *Vestnik Yuzhno-Ural'skogo Gosudarstvennogo Universiteta. Seriya" Vychislitelnaya Matematika i Informatika"* 6 (3): 84–100. doi:10.14529/cmse170306.

Ferraioli, G. 2010. "Multichannel InSAR Building Edge Detection." *IEEE Transactions on Geoscience & Remote Sensing* 48 (3): 1224–1231. doi:10.1109/TGRS.2009.2029338.

Girshick, R. 2015. "Fast R-CNN." Paper Presented at the IEEE International Conference on Computer Vision, Santiago, Chile, December 7–13. doi:10.1109/ICCV.2015.169.

Girshick, R., J. Donahue, T. Darrell, and J. Malik. 2015. "Region-Based Convolutional Networks for Accurate Object Detection and Segmentation." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 38 (1): 142–158. doi:10.1109/TPAMI.2015.2437384.

He, K., G. Gkioxari, P. Dollar, and R. Girshick. 2017. "Mask R-CNN." *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 99: 1. doi:10.1109/ICCV.2017.322.

He, K., X. Zhang, S. Ren, and J. Sun. 2015. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification." Paper Presented at the Proceedings of the IEEE International Conference on Computer Vision, CentroParque Convention Center in Santiago, Chile, December 7–13. doi:10.1109/ICCV.2015.123.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. "Deep Residual Learning for Image Recognition." Paper Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, June 26 –July 1. doi:10.1109/CVPR.2016.90.

Ji, S., S. Wei, and L. Meng. 2018. "Fully Convolutional Networks for Multi-Source Building Extraction from an Open Aerial and Satellite Imagery Dataset." *IEEE Transactions on Geoscience and Remote Sensing*. doi:10.1109/TGRS.2018.2858817.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." Paper Presented at the International Conference on Neural Information Processing Systems, Doha, Qatar, November 12–15. doi:10.1145/3065386.

Li, Y., and W. Huayi. 2008. "Adaptive Building Edge Detection by Combining LiDAR Data and Aerial Images." Paper Presented at the Proceedings of the International Society for Photogrammetry and Remote Sensing, Beijing, China, July 3–11.

Lin, T. Y., P. Dollar, R. Girshick, H. Kaiming, B. Hariharan, and S. Belongie. 2017. Feature Pyramid Networks for Object Detection. Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii July 21–26. doi:10.1109/CVPR.2017.106

Liow, Y. T., and T. Pavlidis. 1989. "Use of Shadows for Extracting Buildings in Aerial Images." *Computer Vision Graphics & Image Processing* 48 (2): 242–277. doi:10.1142/9789814368292_0010.

Ma, R., and W. Meyer. 2005. "DEM Generation and Building Detection from Lidar Data." *Photogrammetric Engineering & Remote Sensing* 71 (7): 847–854. doi:10.14358/PERS.71.7.847.

Maggiori, E., Y. Tarabalka, G. Charpiat, and P. Alliez. 2016. "Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification." *IEEE Transactions on Geoscience & Remote Sensing* 55 (2): 645–657. doi:10.1109/TGRS.2016.2612821.

Maggiori, E., Y. Tarabalka, G. Charpiat, and P. Alliez. 2017. Can Semantic Labeling Methods Generalize to Any City? the Inria Aerial Image Labeling Benchmark. Paper presented at the IGARSS 2017-2017 IEEE International Geoscience and Remote Sensing Symposium, Fort Worth, Texas, USA, July 23–28. doi:10.1109/IGARSS.2017.8127684

Noh, H., S. Hong, and B. Han. 2015. Learning Deconvolution Network for Semantic Segmentation. Paper presented at the Proceedings of the IEEE international conference on computer vision, CentroParque Convention Center in Santiago, Chile, December 7–13. doi:10.1109/ICCV.2015.178

Ren, S., H. Kaiming, R. Girshick, and J. Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Paper presented at the International Conference on Neural Information Processing Systems, Istanbul, Turkey, November 9–12. doi:10.1109/TPAMI.2016.2577031

Ronneberger, O., P. Fischer, and T. Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. Paper presented at the International Conference on Medical image computing and computer-assisted intervention, Munich, Germany, October 5–9. doi:10.1007/978-3-319-24574-4_28

Rottensteiner, F., J. Trinder, S. Clode, and K. Kubik. 2005. "Using the Dempster–Shafer Method for the Fusion of LIDAR Data and Multi-Spectral Images for Building Detection." *Information Fusion* 6 (4): 283–300. doi:10.1016/j.inffus.2004.06.004.

Shelhamer, E., J. Long, and T. Darrell. 2014. "Fully Convolutional Networks for Semantic Segmentation." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 39 (4): 640–651. doi:10.1109/TPAMI.2016.2572683.

Simonyan, K., and A. Zisserman. 2014. ""Very Deep Convolutional Networks for Large-Scale Image Recognition." *arXiv Preprint*." arXiv:1409.1556.

Sirmacek, B., and C. Unsalan. 2008. Building Detection from Aerial Images Using Invariant Color Features and Shadow Information. Paper presented at the International Symposium on Computer and Information Sciences, Istanbul, Turkey, October 27–29. doi:10.1109/ISCIS.2008.4717854

Sirmacek, B., and C. Unsalan. 2009. "Urban-Area and Building Detection Using SIFT Keypoints and Graph Theory." *IEEE Transactions on Geoscience and Remote Sensing* 47 (4): 1156–1167. doi:10.1109/TGRS.2008.2008440.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going Deeper with Convolutions. Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, June 8–10. doi:10.1109/CVPR.2015.7298594

Wu, G., X. Shao, Z. Guo, Q. Chen, W. Yuan, X. Shi, X. Yongwei, and R. Shibasaki. 2018. "Automatic Building Segmentation of Aerial Imagery Using Multi-Constraint Fully Convolutional Networks." *Remote Sensing* 10 (3): 407. doi:10.3390/rs10030407.

Yuan, J. 2017. "Learning Building Extraction in Aerial Scenes with Convolutional Networks." *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 99: 1. doi:10.1109/TPAMI.2017.2750680.

Zhang, L., Z. Zhang, and J. Zhang. 1999. "The Image Matching Based on Wallis Filtering." *Journal of Wuhan Technical University of Surveying & Mapping* 24 (1): 24–35.

Zhang, Y. 1999. "Optimisation of Building Detection in Satellite Images by Combining Multispectral Classification and Texture Filtering." *Isprs Journal of Photogrammetry & Remote Sensing* 54 (1): 50–60. doi:10.1016/S0924-2716(98)00027-6.

Zhong, C., X. Qizhi, F. Yang, and H. Lei. 2015. Building Change Detection for High-Resolution Remotely Sensed Images Based on a Semantic Dependency. Paper presented at the IGARSS

2015-2015 IEEE International Geoscience and Remote Sensing Symposium, Milan, Italy, July 26–31. doi:10.1109/IGARSS.2015.7326535

Zhong, S. H., J. J. Huang, and W. X. Xie. 2008. A New Method of Building Detection from A Single Aerial Photograph. Paper presented at the International Conference on Signal Processing, Leipzig, Germany, May 10–11. doi:10.1109/ICOSP.2008.4697350

Zhou, S., M. Liang, H. Chen, and Y. Geng. 2013. Building Detection in Digital Surface Model. Paper presented at the IEEE International Conference on Imaging Systems and Techniques, Beijing, China, October 22–23. doi:10.1109/IST.2013.6729690