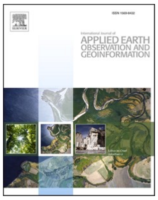
International Journal of Applied Earth Observations and Geoinformation 111 (2022) 102826

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Super-resolving and composing building dataset using a momentum spatial-channel attention residual feature aggregation network

Hongjie He [a], Kyle Gao [a], Weikai Tan [a], Lanying Wang [a], Nan Chen [b], Lingfei Ma [c,*], Jonathan Li [a,d,*]

[a] *Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada*
[b] *College of Geological Engineering and Geomatics, Chang'an University, Xi'an SX710054, China*
[c] *School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 102206, China*
[d] *Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada*

ARTICLE INFO

Keywords:
Building footprint extraction
Generalization error
Building dataset composition
Image super-resolution

ABSTRACT

Model generalizability is crucial in the deployment of deep learning (DL) techniques. When trained on specific datasets, generalizability problems arise across many applications of DL including building extractions. Apart from regularizing the training process, collecting data with distinctive characteristics or distributions can be a promising solution. Over the past decade, several open building datasets have been released. However, in practice, a single dataset cannot overcome the generalization error. By unifying the spatial resolution and spectral bands of different datasets, those datasets could be integrated to relieve the generalization error in building footprint extraction. In this work, we focused on the difference in the spatial resolution between different building datasets. We first examined state-of-the-art super-resolution methods and proposed our own method based on Residual Feature Aggregation Network (RFANet), which we named Momentum and Spatial-Channel Attention RFANet (MSCA-RFANet). We then benchmarked our MSCA-RFANet in a comparative study; our new method achieved higher performance on spatial resolution enhancement. Specifically, in the four times spatial resolution enhancement on the SWOOP 2010 Dataset, our MSCA-RFANet result's peak signal-to-noise ratio (PSNR) of 30.72 dB exceeded that of RFANet (30.66 dB). Likewise, we achieved a lower mean squared error (MSE) of 36.64 compared to RFANet's 36.94. With detailed benchmarks against Second-order Attention Network (SAN) and Residual Channel Attention Network (RCAN), we confirmed the superior performance of our method in enhancing the spatial resolution of high-spatial-resolution images. Then, we explored the impact of super-resolution resolution and data composition on building footprint extraction. Our building footprint extraction experiments demonstrated the positive impact of super-resolution and data composition. These promising results showed that our method is suitable to integrating existing public building dataset to overcome generalization error in DL-based building footprint extraction.

1. Introduction

Deep learning (DL) methods are widely used in different fields after its revival in the last decade. Despite recent advances, generalization of models to out of training set data is still a problem. Applying certain methods such as regularization in model training can alleviate the problem to some extent. Another way is to construct a dataset with a wide variety of different characteristics and distributions (Lambert et al., 2020). For building footprint extraction, given the wide application of high spatial resolution images in building footprint (Cai et al., 2021) and abundant public building datasets released in the past decade (Rottensteiner et al., 2013; Mnih, 2013; Maggiori et al., 2017; Ji et al., 2018; Van Etten et al., 2018; Roscher et al., 2020; He et al., 2021), composing these datasets to overcome generalization errors can be a more promising solution. As these datasets have different spatial resolutions ranging from 0.05 m (the ISPRS Vaihingen and Potsdam Datasets (Rottensteiner et al., 2013)) to 1 m (the Massachusetts Building Dataset (Mnih, 2013)) and different spectral bands, spatial resolution

* Corresponding authors at: Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada (J. Li). School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, 102206, China (L. Ma).

E-mail addresses: l53ma@cufe.edu.cn (L. Ma), junli@uwaterloo.ca (J. Li).

https://doi.org/10.1016/j.jag.2022.102826
Received 22 February 2022; Received in revised form 13 May 2022; Accepted 15 May 2022
Available online 25 May 2022

enhancement and bands selection are required to process these datasets. For bands selection, we can preserve the most discriminative and commonly used red, green, and blue bands (Chen et al., 2018b). For spatial resolution enhancement, DL-based super-resolution methods achieve the highest performance in remote sensing. In DL-based building footprint extraction, super-resolution can not only help data composition but also super-resolve low-spatial resolution images.

Super-resolution methods are commonly categorized into two groups: joint image super-resolution (Marivani et al., 2020) and single-image super-resolution (SISR) (Yang et al., 2014). The former is usually applied to hyperspectral images; it utilizes spectral information from low resolution hyperspectral images and spatial information from multi-spectral images (Zhang et al., 2020a). SISR methods directly process a lower spatial resolution image and output a higher spatial super-resolved one, which are flexible and easy to-use. Pre-trained SISR models can be easily applied to new images. Moreover, building extraction datasets are not usually created with joint hyperspectral and multi-spectral images; therefore, we used SISR techniques to super-resolve building datasets.

Deep convolution neural network (DCNN) based SISR methods were developed along with deep learning techniques and grew rapidly in sophistication and performance. From Super-Resolution CNN (SRCNN) (Dong et al., 2015) to Residual Feature Aggregation Network (RFANet) (Liu et al., 2020), the performance of SISR methods increased significantly over the years. Residual Channel Attention Network (RCAN) (Zhang et al., 2018), Second-order Attention Network (SAN) (Dai et al., 2019) and RFANet represent the state-of-the-art techniques in this field. Each of them reached the highest performance on different datasets. By re-examining them, as well as applying newly developed deep learning techniques, these state-of-the-art methods can be improved further.

In this paper, we have two main objectives. The first one is to examine super-resolution and dataset composition for the improvement of building footprint extraction. Specifically, we compare the performance of the same DL-based building footprint extraction model on the original datasets, the super-resolved datasets, and the composited dataset. The second one is to benchmark our newly developed Momentum Spatial-Channel Attention RFANet (MSCA-RFANet), which adopts the advantages of RFANet, residual channel attention mechanism and share-source skip connection. A comparative study is conducted with four other DL-based methods as well as bicubic interpolation. The contribution of this paper includes:

(1) Exploring and examining the effects of super-resolved and composited dataset in building footprint extraction, and
(2) Presenting a new SISR method: MSCA-RFANet and benchmarking it against the state-of-art SISR methods.

This paper is structured as follows. Section 2 provides a brief literature review related to this work, which includes the development of SISR, recently released building datasets and efforts paid to overcome generalization errors with respect to data composition. Section 3 describes the datasets used for SISR network training and building footprint extraction, as well as the architecture of our MSCA-RFANet. Section 4 presents and analyzes the experimental results. Section 5 discusses the methods we tried which did not improve MSCA-RFANet. Section 6 concludes the paper by summarizing our findings.

2. Related work

2.1. A review of DCNN-based SISR

Once DCNN based SISR network surpassed conventional SISR methods (Dong et al., 2015), they drew much attention from different research communities which further accelerated the development of new DCNN based SISR methods in the last decade. To improve the accuracy of SISR, networks became increasingly deeper. In this direction, VDSR (Kim et al., 2016), Deeply Recursive Convolutional Network (DRCN) (Kim et al., 2016), Residual Encoder-Decoder Networks (RED-Net) (Mao et al., 2016), and Deep Recursive Residual Network (DRRN) (Tai et al., 2017) were proposed. With the development of new techniques in deep learning, such as transposed convolution and dense block, Laplacian Pyramid Super-Resolution Network (LapSRN) (Lai et al., 2017), SRDenseNet (Tong et al., 2017), super-resolution generative adversarial network (SRGAN) (Ledig et al., 2017), Enhanced deep super-resolution network (EDSR) and multi-scale deep super-resolution system (MDSR) (Lim et al., 2017) were proposed. In recent years, the attention mechanism became widely used in DCNN. Recent DCNN based SISR methods also made use of this innovation. The state-of-the-art methods in SISR proposed in recent years applied attention mechanism to boost the performance of image super-resolution. Some examples of which are RCAN (Zhang et al., 2018), SAN (Dai et al., 2019), and RFANet (Liu et al., 2020). RCAN and SAN apply channel attention, whereas RFANet uses spatial attention. Efficient Sub-Pixel Convolutional Neural Network (ESPCN) (Shi et al., 2016), one of the classic SISR networks, was used in those three methods as an up-sampling module. To develop our new SISR network, we chose RCAN, SAN and RFANet as representative baseline SISR networks.

2.2. Publicly available building datasets

Over the past decades, several building datasets were released and widely used for building footprint extraction, including the ISPRS Vaihingen and Potsdam Datasets (Rottensteiner et al., 2013), the Massachusetts Building Dataset (Mnih, 2013), the Inria Dataset (Maggiori et al., 2017), the Wuhan University (WHU) Building Dataset (Ji et al., 2018), the SpaceNet Building Dataset (Van Etten et al., 2018), the Aerial Imagery for Roof Segmentation (AIRS) Dataset (Chen et al., 2018a), and the Semcity Toulouse Dataset (Roscher et al., 2020). Other than the aforementioned datasets, there are also the Waterloo Building Dataset (He et al., 2021), datasets from the Open Cities AI Challenge (GFDRR Labs, 2020), and datasets from the Crowd-AI Mapping Challenge (Mohanty et al., 2020). The Waterloo Building Dataset is a city-scale building dataset, which covers the Kitchener-Waterloo area in Ontario, Canada. The dataset from the Open Cities AI Challenge is a building footprint dataset across 10 cities in Africa and is known for its inconsistent annotation accuracy. All the aforementioned datasets have a spatial resolution ranging from 0.05 m (the ISPRS Vaihingen and Potsdam Datasets) to 1 m (the Massachusetts Building Dataset). According to Nyquist-Shannon Sampling theorem (Farrow et al., 2011), for better representation, sampling frequency must be equal or higher than twice the highest spatial frequency of the signal (Duveiller and Defourny, 2010). For building footprint extraction, it means the spatial resolution of images is determined by the level of detail required in certain tasks (Farrow et al., 2011). In this work, we define our ideal spatial resolution to be 0.3 m, which would allow us to resolve sub meter-level detail without being too computationally expensive to process. Many publicly available datasets are also at this resolution.

2.3. Data composition

In computer vision, there are single-domain data mixing and cross-domain data mixing (Lambert et al., 2020). In single-domain data mixing, datasets for same specific purpose are mixed, such as combining various driving datasets. On the other hand, Lambert et al. (2020) merged multi-domain datasets for semantic segmentation. They presented MSeg dataset, which included COCO, ADE20K, Mapillary, IDD, BDD, Cityscapes, and SUN RGB-D datasets. Their experiments showed that the model trained on MSeg is more robust compared to models trained on single dataset or single domain mixed datasets. For building footprint extraction, the Inria Dataset (Maggiori et al., 2017) was released to address the generalization error. A total of 10 cities of the U. S. and Austria were split into training and testing dataset. As the dataset was released for exploring generalization ability, the split was made such that no adjacent images exist in training or testing dataset. However, the dataset is limited to two countries. In this work, we aim at mixing all publicly available datasets which cover both the Northern and Southern hemispheres and most continents.

2.4. Deep learning-based building footprint extraction

In recent years, deep learning (LeCun et al., 2015) methods have been broadly utilized in various remote sensing image–based applications (Zhu et al., 2017; Liu and Abd-Elrahman, 2018). For building footprint extraction, the applications can be categorized into three stages. In first stage, deep learning models are used as feature extractors to generate features. Mnih (2013) and Shu (2014) applied deep learning in building footprint extraction in this way. The last layer is a fully connected feature classifier to generate the label of each pixel.

With the proposal of fully convolutional networks (Long et al., 2015), the applications of deep learning in building footprint extraction can be seen as the second stage. In this stage, deep learning models proposed in computer vision were widely used and compared in building footprint extraction tasks by the remote sensing community. Comparative studies conducted by Nogueir et al. (2017), Kemker et al. (2018), Liu et al. (2018), Yi et al. (2019), ERDEM and AVDAN (2020), and Cai et al. (2021) showed the high performance of deep learning models in building footprint extraction. In this stage, there were also many deep learning models proposed by the remote sensing community specifically for building footprint extraction. These methods include but not limited to multiple-feature reuse network (Li et al., 2018), dual-resolution U-Net (Lu et al., 2018), Efficient Separable Factorized Network (Lin et al., 2019), Deep Encoding Network (Liu et al., 2019), Efficient Non-Local Residual U-shape Network (Wang et al., 2020) and Capsule Feature Pyramid Network (Yu et al., 2020). Accuracy and speed of deep learning model in building footprint extraction were improved by the introduction of advanced deep learning techniques in this stage.

Vector maps of building footprint are the ultimate data used in practice for analysis and statistics. In previous methods, vector maps were not considered or were generated via a post-processing step based on extraction results. In recent work, extracting vector maps of building footprint from images in an end-to-end manner has drawn much attention. The first method proposed in this stage was Deep Structured Active Contours (DSAC), which embedded active contour model in a CNN model. Similarly, Li et al. (2019) and Zhao et al. (2021) connected a Recurrent Neural Network to a CNN model, which showed high performance in the end-to-end manner of building footprint extraction. However, with the latest development in deep learning and the high-quality training data, the performance of building footprint extraction is expected to be improved further.

3. Datasets and methods

3.1. Datasets

3.1.1. Dataset used for SISR

In this work, the selected aerial images from the Southwestern Ontario Orthophotography Project 2010 (SWOOP 2010)[1] were super-resolved using SISR. Specifically, 1,127, 4,910, 2,582, 1,478 and 750 aerial images from Brant, Bruce, Chatham-Kent, Dufferin and Elgin, in Ontario, Canada (as shown in Fig. 1 and Table 1), respectively, were selected from SWOOP 2010 Dataset. Those images have a spatial resolution of 0.2 m with red, green, and blue bands. Each image has 5,000 × 5,000 pixels and cover 1 km^2 area. In addition, to add more training data, we collected 274 aerial images[2] covering the Kitchener-Waterloo area, with a spatial resolution of 0.12 m of size 8,350 × 8,350 pixels, from the Regional Municipality of Waterloo. Images from SWOOP 2010 and Waterloo were resized and cropped into small patches with a size of 256 × 256 pixels as High Resolution (HR) (0.25 m) images and processed further to a size of 64 × 64 pixels as Low Resolution (LR) (1 m) images. In the rest of the paper, we refer to this dataset as the SWOOP 2010 Dataset. Consequently, 1,708,032, 284,672 and 854,272 pairs of patches were prepared for SISR network training, validation and testing, respectively.

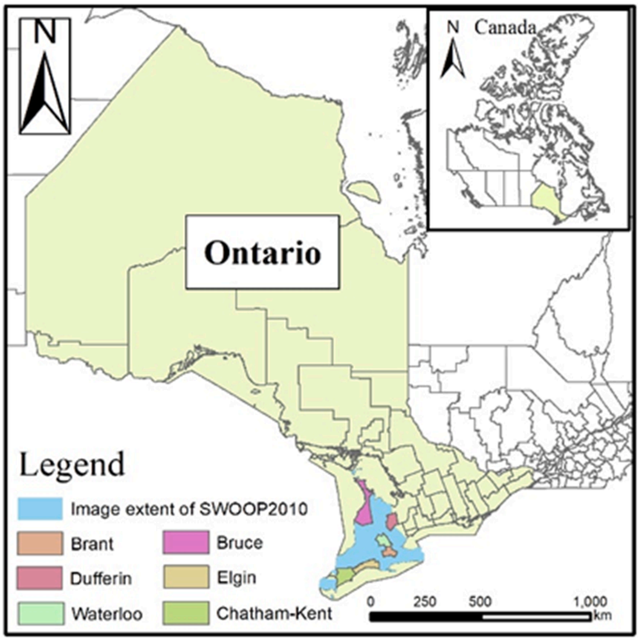


Fig. 1. The extension of datasets used for SISR(Administrative areas shapefiles are downloaded from http://www.diva-gis.org/gdata. SWOOP extent area shapefile is acquired from https://www.arcgis.com/home/item.html?id=2d424be6b6054bd091023df227ea73da).

Table 1
The details of images for SISR.

Area	#Images	Image size	Pixel size
Brant	1127	5000 × 5000	0.20 m
Bruce	4910	5000 × 5000	0.20 m
Chatham-Kent	2582	5000 × 5000	0.20 m
Dufferin	1478	5000 × 5000	0.20 m
Eligin	750	5000 × 5000	0.20 m
Waterloo	274	8350 × 8350	0.12 m

3.1.2. Datasets for building footprint extraction

In this work, we selected the Massachusetts Building Dataset, WHU Building Dataset and our Waterloo Building Dataset for building footprint extraction and data composition. As we mentioned in Section 2.2, most public building datasets have a spatial resolution of 0.3 m/pixel, we unified the spatial resolution of selected datasets to 0.3 m/pixel. The Massachusetts Building Dataset was selected to explore the effect of SISR on building footprint extraction. The other two datasets were selected to explore the impact of data fusion on building footprint extraction. In addition, these three datasets can also be used to examine the generalizability of trained models.

The Massachusetts Building Dataset has a spatial resolution of 1 m

[1] Produced by the Ontario Ministry of Natural Resources under License with the Ontario Ministry of Natural Resources © Queen's Printer for Ontario, 2010–2011.

[2] Those aerial images are used for constructing Waterloo building dataset.

with three spectral bands (red, green and blue). In this dataset, a total of 151 aerial images covering 340 km^2 in the Boston area, USA, as well as paired ground truth images are split into 137 pairs, 4 pairs and 10 pairs of images for training, validation and test, respectively. The WHU Building Dataset has a spatial resolution of 0.3 m. It is composed of total of 8,189 aerial RGB images covering 450 km^2 (about half the area of San Antonio, Texas) in Christchurch, New Zealand. They are divided into 4,736, 1,036 and 2,416 items with matched ground truth images, split into training, validation and test sets, respectively. The Waterloo Building Dataset consists of 242 RGB aerial images at a spatial resolution of 0.12 m, covering the Kitchener-Waterloo area, Ontario, Canada. A total of 242 aerial images are digitalized and converted to ground truth images. After splitting into small patches with a size of 512 × 512, as well as removing geometrically distorted patches, 42,147, 6,887 and 20,768 pairs of patches were obtained for the training, validation, and test sets, respectively. The details of three building datasets were listed in Table 2.

3.2. Methods

In this section, we describe our proposed SISR method MSCA-RFANet, building footprint extraction method and evaluation metrics.

3.2.1. MSCA-RFANet

The state-of-the-art SISR deep learning methods typically have three parts (Liu et al., 2020): the head part, the trunk part (base modules) and the reconstruction part (as shown in Fig. 2). They are responsible for shallow feature extraction, deep feature extraction and image reconstruction, respectively. As RFANet is the most recent and powerful method in SISR field, we developed our method based on its core architecture using its RFA and Enhanced Spatial Attention (ESA) modules. As for the head and the reconstruction parts, it is standard to use a standard convolution layer and ESPCN for shallow feature extraction and image reconstruction. We made our modifications on the trunk part of RFANet.

In the trunk part, inspired by recent work (Chen et al., 2017; Woo et al., 2018; Zhao et al., 2020; Zhang et al., 2020b), we added the Channel Attention (CA) block after the ESA block resulting in a spatial-channel attention block (SCA block). In this way, the network could focus on both informative regions and features. We named the modified RFA module as RFA +. Each RFA + module was skip connected to previous one with a momentum term. Share-source skip connection would relieve the deep model training and benefit the information flow, which was also used in the shared source residual group of SAN (Dai et al., 2019). The difference between skip connection with and without momentum term is described below:

$$\textit{Normal skip connection (ResNet)}: x_{n+1} = x_n + f(x_n, \theta_n) \tag{1}$$

Skip connection with momentum term(Momentum ResNet):

$$\begin{cases} v_{n+1} = \gamma v_n + (1-\gamma) f(x_n, \theta_n) \\ x_{n+1} = x_n + v_{n+1} \end{cases} \tag{2}$$

where x_n represents the convolutional layer generated feature. $f(x_n, \theta_n)$ stands for the convolution block in ResNet, in which θ_n are learnable parameters in each block. γ is a constant between 0 and 1. v_n is the momentum at layer n. The initial momentum can be 0 or pre-defined function. As described by Sander et al. (2021), momentum ResNet could achieve same accuracy as ResNet on image classification with smaller memory footprint and benefits transfer learning (Sander et al., 2021). We added a batch normalization layer after each skip connection.

Following RFANet, we set the number of RFA + modules as 30. The head part and the reconstruction part were detailed in Table 3 and 6. We also detailed the ESA and CA modules in Table 4 and 5. ESA and CA blocks were connected to construct the RFA + module as we presented in bottom left of Fig. 2. A total of 30 RFA + were connected in the share-source skip connection manner (Dai et al., 2019) with a momentum term which was initiated to 0. The initial learning rate was set as 5e-5 and decreased by half every 2e5 iterations. It is worth noting that when using our own learning rate schedule as opposed to the original authors chosen learning rate, the final evaluation metrics of these networks may differ from their original values. In addition, we used Adam as the optimizer and MAE as the loss function. We trained both SISR models 20 epochs with batch size of 16. Since we applied these models to remote sensing, all images were directly input to network and evaluated in RGB space rather than YCBCr space which is commonly used in the SISR field. In this work, all experiments were implemented on a single GeForce RTX 3090 GPU and CUDA 11.2.

Table 2
The details of datasets for building footprint extraction.

Dataset		#Images	Image size	Pixel size
Waterloo Building Dataset	Train	42,147	512 × 512	0.12 m
	Validation	6,887		
	Test	20,768		
WHU Building Dataset	Train	4,736	512 × 512	0.30 m
	Validation	1,036		
	Test	2,416		
Massachusetts Building Dataset	Train	137	1500 × 1500	1 m
	Validation	4		
	Test	10		

3.2.2. Methods for building footprint extraction

In building footprint extraction, both semantic and instance segmentation methods have been widely used (Cai et al., 2021; Roscher et al., 2020). In this work, we did not focus on comparing sophisticated building footprint extraction methods. Therefore, we selected high-resolution network v2 (HRNet v2) (Sun et al., 2019), a powerful recently proposed network, for building footprint extraction. This network aimed at maintaining high-resolution representations. To do so, four levels of features with different spatial resolution were preserved and sequentially concatenated in four stages. With the exception of the first, these stages consisted of repeated modularized multi-resolution blocks. Each block had a multi-resolution group convolution and a multi-resolution convolution. The detailed information about the architecture can be found in Sun et al. (2019).

For the training of building extraction models, we used Adam optimizer due to its high performance, instead of Stochastic gradient descent (SGD) which was used in the original paper (Sun et al., 2019). The learning rate was set as a constant 1e-4. The Jaccard loss (Berman et al., 2018) was used as loss function to address binary class imbalance. It is worth noting that for fair comparison, all HRNet v2 models discussed below were trained with the same number of iterations. Specifically, for the training of each HRNet v2, we set batch size as 8 and iterated the optimizer 5,400 times per epoch for 100 epochs.

3.3. Metrics used for evaluating SISR

To evaluate super-resolved images, MSE, RMSE, PSNR and SSIM are calculated respectively by.

$$MSE = \frac{1}{N}\sum\nolimits_{n=1}^{N} \left(\hat{g}_i - g_i\right)^2 \tag{3}$$

$$RMSE = \sqrt{MSE} \tag{4}$$

$$PSNR = 20\log_{10}\frac{L}{RMSE} \tag{5}$$

$$SSIM = \frac{\left(2\mu_{\hat{g}}\mu_g + C_1\right)\left(2\sigma_{\hat{g}}\sigma_g + C_2\right)}{\left(\mu_{\hat{g}}^2 + \mu_g^2 + C_1\right)\left(\sigma_{\hat{g}}^2 + \sigma_g^2 + C_2\right)} \tag{6}$$

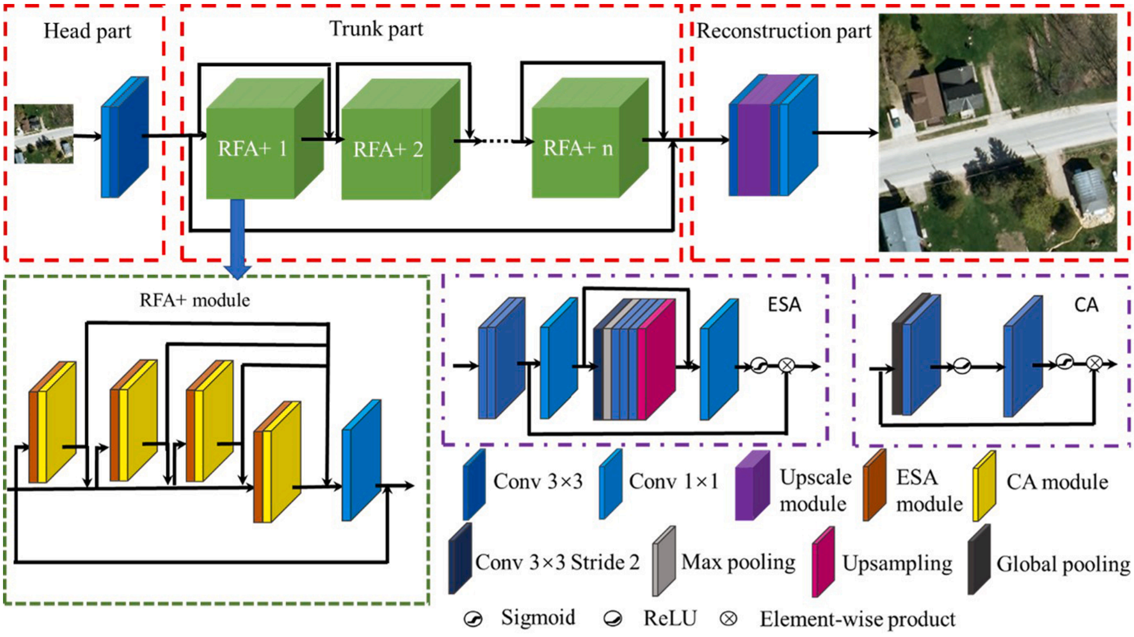


Fig. 2. Architecture of our proposed MSCA-RFANet(Element-wise addition in skip connection unless specified otherwise.).

Table 3
The head part of our mSCA-RFANet.

Layer types	Filters	Size	Strides	Output size
Input				h × w × 3
Convolutional layer	3	1 × 1	1	h × w × 3
Convolutional layer	64	3 × 3	1	h × w × 64

Table 4
The ESA modules in the trunk part of our mSCA-RFANet.

Layer types	Filters	Size	Strides	Output size
Convolutional layer	64	3 × 3	1	h × w × 64
ReLU				h × w × 64
Convolutional layer	64	3 × 3	1	h × w × 64
Residual_ESA1				h × w × 64
Convolutional layer	16	1 × 1	1	h × w × 16
Residual_ESA2				h × w × 16
Convolutional layer	16	3 × 3	2	h/2 × w/2 × 16
Maxpooling		8 × 8	2	h/4 × w/4 × 16
Convolutional layer	16	3 × 3	1	h/4 × w/4 × 16
Convolutional layer	16	3 × 3	1	h/4 × w/4 × 16
Convolutional layer	16	3 × 3	1	h/4 × w/4 × 16
Upsampling		4 × 4		h × w × 16
Add: +=Residual_ESA2				h × w × 16
Convolutional layer	64	1 × 1	1	h × w × 64
Sigmoid				h × w × 64
Multiply: ×=Residual_ESA1				h × w × 64

Table 5
The CA modules in the trunk part of our mSCA-RFANet.

Layer types	Filters	Size	Strides	Output size
Residual_CA				h × w × 64
Global pooling				1 × 1 × 64
Convolutional layer	4	3 × 3	1	1 × 1 × 4
ReLU				1 × 1 × 4
Convolutional layer	64	3 × 3	1	1 × 1 × 64
Sigmoid				1 × 1 × 64
Multiply: ×=Residual_CA				h × w × 64

Table 6
The reconstruction part of our mSCA-RFANet.

Layer types	Filters	Size	Strides	Output size
Convolutional Layer	64	3 × 3	1	h × w × 64
Convolutional Layer	256	3 × 3	1	h × w × 256
Depth_to_space				2 h × 2w × 64
Covnolutional layer	256	3 × 3	1	2 h × 2w × 256
Depth_to_space				4 h × 4w × 64
Convolutional Layer	64	3 × 3	1	4 h × 4w × 64
Convolutional Layer	3	3 × 3	1	4 h × 4w × 3

where $\widehat{g}$ and g refer to the super-resolved images and ground truth of high spatial resolution images. N is the number of pixels in images, i indexes individual pixels which ranges from i = 1 to i = N. L in the calculation of PSNR denotes the max value of pixels in images given certain bit depth. For example, if images are normalized into 0 ~ 1, L will be 1. For images with unsigned int 8 bits depth, L is 255. $\mu_{\widehat{g}}$ and μ_g are mean value of all pixels in super-resolved images and original high spatial resolution images, respectively. Similarly, $\sigma_{\widehat{g}}$ and σ_g are their unbiased standard deviation. In the evaluation process of our experiments, we calculated MSE, RMSE, PSNR and SSIM for each image and reported the averaged MSE, RMSE, PSNR and SSIM to show the performance of different super-resolution methods. In addition, we also recorded the total number of trainable parameters and calculated the floating-point operations (FLOPs) (Molchanov et al., 2019) to show the model size and the computational complexity of each super-resolution model.

To evaluate the accuracy of segmentation results, we used Overall Accuracy (OA), Intersection of Union (IoU), mean IoU (mIoU), precision, recall and F_1 score. OA indicates how many pixels were correctly classified; mIoU represents the average IoU of negative class and positive class; F_1 score stands for the harmonic mean of recall and precision, which is more representative than other three metrics. As those metrics are widely used in all segmentation tasks, we direct the readers to Cai et al. (2021) for detailed information.

4. Results and analysis

4.1. Performance of super-resolution

4.1.1. Qualitative evaluation

To visually compare the performance of our MSCA-RFANet t with other super-resolution methods used in our experiments, we selected three images from our SWOOP 2010 Dataset. As shown in Fig. 3, the images on the first row to the last row were: the low-resolution images with a pixel size of 1 m, the high-resolution images with a pixel size of 0.25 m, the bicubic interpolated images, the RCAN super-resolved images, the SAN super-resolved images, the RFANet super-resolved images and our method's super-resolved images, which were denoted as “LR images”, “HR images”, “BI images”, “RCAN images”, “SAN images”, “RFANet images”, “MSCA-RFANet images” in the first row of Fig. 3. As can be seen from the figure, bicubic interpolation can generate high-resolution images but features in the images were blurred. The CNN-based super-resolution methods can generate high-resolution buildings and roads but also blur trees in first column and last column. From the figure, it was hard to tell the difference between CNN-based super-resolution methods. Therefore, we did quantitative evaluation in next section.

4.1.2. Quantitative evaluation

In this section, we explored the performance of our method in super-resolving the SWOOP 2010 Dataset and.

the down sampled WHU Building Dataset which was generated by bicubically interpolating all images in the original WHU Building Dataset to 1.2 m/pixel spatial resolution. We also trained RCAN (Zhang et al., 2018), SAN (Dai et al., 2019) and RFANet (Liu et al., 2020) on SWOOP 2010 Dataset as baseline and tested their performance on the SWOOP 2010 (Table 7) and the WHU Building Datasets (Table 8). In these tables, “Bicubic” refers to bicubic interpolation. RCAN, SAN, RFANet and MSCA-RFANet represents three state-of-the-art methods, as well as our own method. SCA-RFANet denotes the method which only applies to SCA block on top of RFANet. The performance of SCA-RFANet is provided here to explore the contribution of SCA block and the share-source skip connection between RFA + modules by comparing it to RFANet and MSCA-RFANet.

As shown in Tables 7 and 8, all DL-based SISR methods significantly outperform bicubic interpolation in terms of all metrics. As shown in Table 7, on the SWOOP 2010 Dataset, our MSCA-RFANet outperforms other state-of-the-art methods. Specifically, our MSCA-RFANet has a PSNR value of 30.72 dB, which exceeds that of RCAN, SAN, RFANet by 0.31 dB, 0.21 dB and 0.06 dB, respectively. On the WHU Building Dataset (Table 8), our MSCA-RFANet has a PSNR value of 20.01 dB, which is higher than RCAN, RFANet by 0.02 dB and 0.03 dB, respectively. We omit the evaluation scores of the SAN model because using our computational resource, the SAN could not process the down sampled WHU Building Dataset while other methods could. By examining the performance of SCA-RFANet in Tables 7 and 8, we noticed the positive contribution of using both spatial attention (ESA) and CA or SCA block in SISR. For instance, the PSNR value of SCA-RFANet in the WHU Building Dataset is increased from 20.35 dB to 20.42 dB. The contribution of the share-source skip connection between RFA + modules need further investigation because the share-source skip connection shows positive effect in spatial resolution enhancement of the SWOOP 2010 Dataset but negative effect in that of the WHU Building Dataset. Overall, given these results, we can conclude that our MSCA-RFANet achieved superior performance.

4.2. Impact of SISR on building footprint extraction

4.2.1. Semantic models trained on single building dataset

For ease of comparison, we arranged the evaluation metrics of extraction results from two experiments according to training dataset in Tables 9 and 10. It is worth noting that for images in the Massachusetts Building Dataset, we used SISR methods to super-resolve images. However, we used bicubic interpolation, which is sufficient to preserve simple geometric shapes, to super-resolve ground truth images. The interpolated ground truth images represented buildings' locations and shapes well, as shown in Fig. 4. For Waterloo Building dataset, we also applied bicubic interpolation to down sample images and ground truth images as shown in Fig. 5.

As shown in Table 9, according to OA, mIoU and F_1 score, there were two trends. Firstly, the more similar the test set was to the training set with respect to data distribution and spatial resolution, the higher the model scored on the evaluation metrics. Secondly, after super-resolution, the quality of extraction results improved in general. This was true unless there was too large of a discrepancy between the test set and training set in terms of spatial resolution and data statistics, as can be seen for the models trained on the 1 m Massachusetts dataset. For example, the model trained on the original Waterloo Building Dataset obtained its highest mIoU of 87.12% on the original Waterloo Building Dataset but achieved its lowest mIoU of 43.31% on the original Massachusetts Building Dataset. This model achieved its second highest mIoU of 69.31% on the bicubically interpolated Waterloo Building Dataset. In addition, the model achieved higher mIoU on the super-resolved Massachusetts Building Dataset than on the bicubically interpolated and original Massachusetts Building Dataset. Specifically, the mIoU value increased from 43.31% to 45.46% and 48.00% by interpolating and super-resolving the Massachusetts Building Dataset.

Same trends could also be found in the extraction results on the test sets using models trained on the interpolated Waterloo Building Dataset (Table 9), the WHU Building Dataset (Table 9), the bicubically interpolated Massachusetts Building Dataset (Table 10), the super-resolved Massachusetts Building Dataset (Table 10) and the original Massachusetts Building Dataset (Table 10). One interesting thing we found is that HRNet v2 model trained on the interpolated Waterloo Building Dataset achieved very poor results on the WHU Building Dataset. We believe that is caused by different building types and minor interpolation errors.

We initially believed that MSCA-RFANet would outperform RFANet in all scenarios. We noticed from the results in Table 9 and Table 10 that in general, super-resolving the test set with RFANet achieved slightly better results. However, super-resolving the training set using MSCA-RFANet produced significantly better results when testing on test-sets different from the training set in terms of resolution (the 1 m Massachusetts Building Dataset) or building distribution (the WHU Building Dataset). We believed that training on MSCA-RFANet made the model more generalizable to strong distribution shift between training and test datasets. However, using RFANet on the test set better minimized small distribution shifts between test set and training set. This effect was also noticeable in training set D and training set E of Table 11 in the next section.

4.2.2. Semantic models trained on composed building dataset

We investigated the effect of SISR on dataset composition for building footprint extraction. We trained the HRNet v2 on the combination of the Waterloo Building Dataset, the WHU Building Dataset and Massachusetts Dataset. In training set A, we used all three datasets as is. In training set B, we bicubically interpolated the Massachusetts Building Dataset to 0.3 m/pixel. In training set C, we bicubically interpolated both the Waterloo Building Dataset and Massachusetts Building Dataset to 0.3 m/pixel. In training set D, we interpolated the Waterloo Building Dataset and use RFANet to super-resolve the Massachusetts Building Dataset to 0.3 m/pixel. In training set E, we interpolated the Waterloo Building Dataset and use our MSCA-RFANet to super-resolve the Massachusetts Building Dataset to 0.3 m/pixel.

In addition to the trends, we found above, the performance improvement caused by composition is noticeable. For example, OA on the original Waterloo Building Dataset test set increased from 88.63 % of the model trained on the WHU Building Dataset (Table 9) to 94.36% of

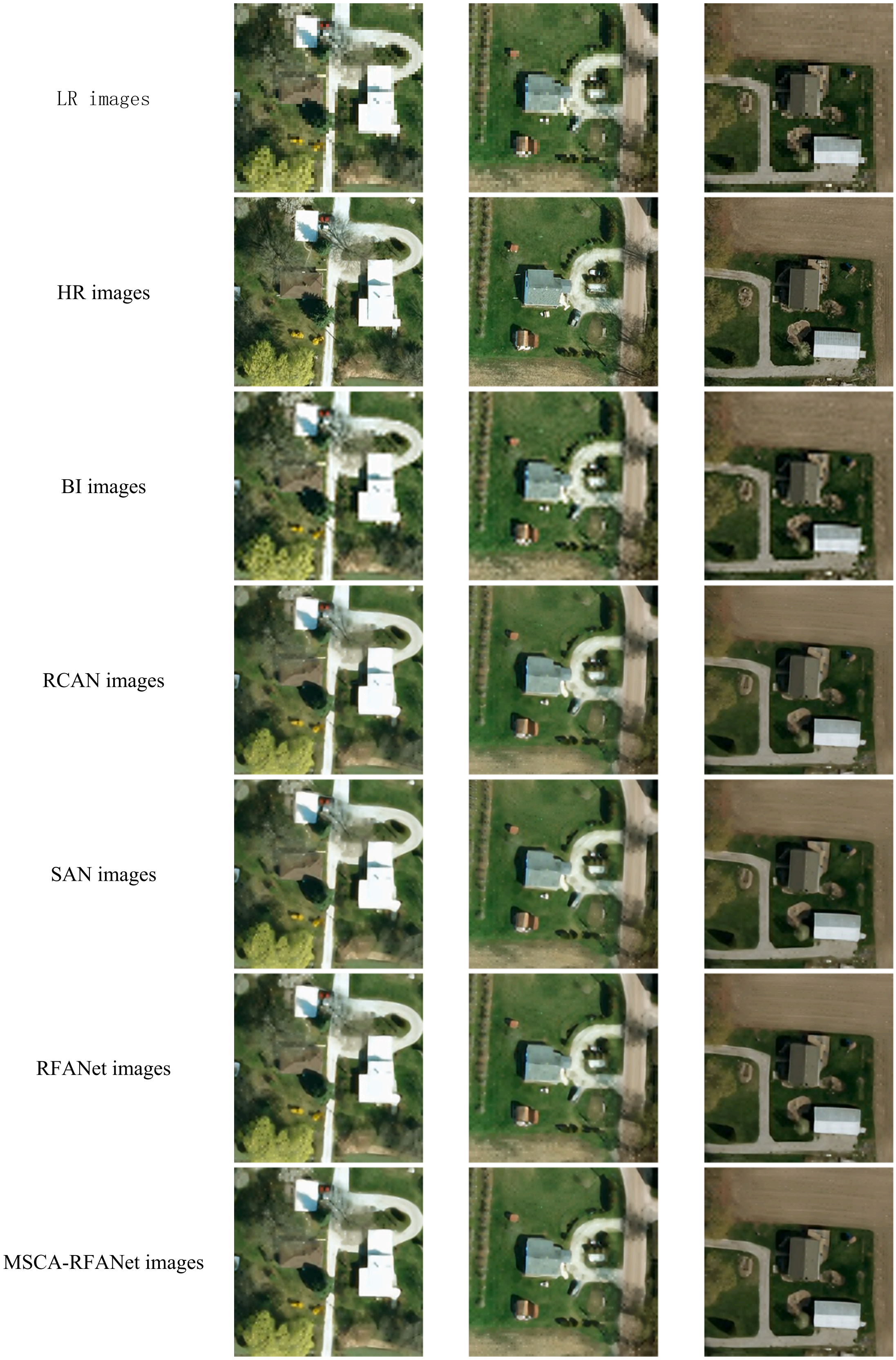


Fig. 3. Examples of super-resolution.

Table 7
Performance of SISR models (tested on the SWOOP 2010 Dataset).

Models	MSE	RMSE	PSNR (dB)	SSIM[1]	#Parameters	FLOPs
Bicubic	43.05	6.33	29.13	0.69	0	0
RCAN	38.04	5.91	30.41	0.73	16,406,409	135.14G
SAN	37.47	5.87	30.51	0.74	15,936,553	179.36G
RFANet	36.94	5.81	30.66	0.75	10,692,489	87.76G
SCA-RFANet	36.89	5.81	30.68	0.75	11,245,449	87.83G
mSCA-RFANet	**36.64**	**5.79**	**30.72**	**0.75**	**11,245,449**	**87.84G**

[1] Because RFANet, SCA-RFANet and mSCA-RFANet have similar performance, the SSIM score was the same up to two decimal places which we used for the table.

Table 8
Performance of SISR models (tested on the WHU Building Dataset).[7]

Models	MSE	RMSE	PSNR (dB)	SSIM	#Parameters	FLOPs
Bicubic	76.54	8.73	19.39	0.44	0	0
RCAN	69.10	8.29	20.36	0.50	16,406,409	540.54G
SAN	–	–	–	–	–	–
RFANet	69.42	8.31	20.35	0.50	10,692,489	351.06G
SCA-RFANet	68.88	8.27	20.42	0.50	11,245,449	351.31G
mSCA-RFANet	**68.97**	**8.28**	**20.38**	**0.50**	**11,245,449**	**351.37G**

[7] As we described in Section 3.3, MSE and RMSE are averaged values of all images. Therefore, RMSE values in our results do not equal the squared MSE values.

the model trained on the training set A (Table 11), although it is lower than the 97.78% achieved by the model trained on the original Waterloo Building Dataset where both training and test sets were split from the same dataset (Table 9). In other words, by simply composing datasets, the generalizability of the trained model improved significantly. By super-resolving the Massachusetts Building Dataset in composed dataset, this improvement becomes more obvious. Specifically, for all composed training sets, all evaluation scores increased across the different test sets except for the original Massachusetts Building Dataset, likely due to the large spatial resolution difference, which was overcome with SISR super-resolution (or bicubic super-resolution to a lesser degree). For instance, when training on any composed datasets and applying super resolution as a preprocessing step, the model achieved a high degree of generalizability. We also noticed the same effect here. Super-resolving the test-set using RFANet achieved the best results. However, super-resolving the training set using our MSCA-RFANet made the model more generalizable and achieved better results.

4.2.3. *Impact visualization*

In this section, we first visually showed the generalization errors in Fig. 6, and the impact of super-resolution and combining super-resolution and data composition on building footprint extraction in Fig. 7 and Fig. 8, respectively.

As shown in Fig. 6, from the first to last rows, samples of super-resolved Massachusetts Building Dataset and ground truth, extraction results generated by models trained on the Waterloo Building Dataset with spatial resolution of 0.12 m/pixel and 0.3 m/pixel, the WHU Building Dataset and the Massachusetts Building Dataset. Among these four models, the model trained on bicubic interpolated Waterloo Building Dataset showed higher performance than that trained on original Waterloo Building Dataset; the model trained on the WHU Building Dataset showed the poorest performance; and the model trained on the Massachusetts Building Dataset showed the highest performance. We can conclude from Fig. 6 with our previously mentioned findings: the more similar the test set was to the training set with respect to data distribution and spatial resolution, the higher the model scored on the evaluation metrics.

As shown in Fig. 7, from the first to last rows, samples of super-resolved Massachusetts Building Dataset and ground truth, extraction results generated by models trained on the original, the bicubic interpolated, the RFANet super-resolved, and the MSCA-RFANet super-resolved Massachusetts Building Dataset were listed respectively. Among these models, models trained on super-resolved dataset showed better performance than those trained on the original and bicubic interpolated dataset. The results confirmed our second finding that after super-resolution the quality of extraction results improved in general.

Similarly, in Fig. 8, we showed the samples of super-resolved Massachusetts Building Dataset and ground truth, extraction results generated by models trained on the original Massachusetts Building Dataset, training set A, B, C, D and E. Visualization results also followed the findings we mentioned above and was under our expectation. The

Table 9
Performance of building footprint extraction results using models trained on Waterloo Building Dataset and WHU Building Dataset (in %).[8]

Training data	Test data	OA	IoU	mIoU	Precision	Recall	F_1 score
Waterloo (0.12)	**Waterloo (0.12)**	**97.78**	**76.63**	**87.12**	**92.48**	**81.72**	**86.77**
	Waterloo (BI: 0.3)	94.67	44.18	69.31	79.07	50.03	61.29
	WHU (0.3)	89.54	18.62	53.95	58.04	21.52	31.39
	Massachusetts (BI: 0.3)	73.08	19.75	45.46	32.45	33.55	32.99
	Massachusetts (RFA: 0.3)	74.31	23.95	48.00	36.58	40.97	38.65
	Massachusetts (ours:0.3)	74.18	23.63	47.78	36.24	40.44	38.22
	Massachusetts (1)	80.71	6.14	43.31	40.66	6.75	11.58
Waterloo (BI: 0.3)	Waterloo (0.12)	79.21	25.49	51.55	27.25	79.79	40.62
	Waterloo (BI: 0.3)	**83.99**	**30.55**	**56.66**	**32.50**	**83.57**	**46.80**
	WHU (0.3)	50.26	15.76	30.46	16.26	83.61	27.23
	Massachusetts (BI: 0.3)	73.49	21.01	46.25	33.80	35.70	34.73
	Massachusetts (RFA: 0.3)	75.48	29.37	51.04	40.53	51.62	45.41
	Massachusetts (ours:0.3)	75.09	28.65	50.48	39.75	50.65	44.54
	Massachusetts (1)	76.97	11.44	43.85	28.95	15.90	20.52
WHU (0.3)	Waterloo (0.12)	88.63	15.30	51.85	31.29	23.05	26.55
	Waterloo (BI: 0.3)	87.37	18.93	52.96	29.21	34.99	31.84
	WHU (0.3)	**95.22**	**67.74**	**81.21**	**73.09**	**90.26**	**80.77**
	Massachusetts (BI: 0.3)	75.68	6.56	40.91	21.40	8.65	12.32
	Massachusetts (RFA: 0.3)	75.75	21.66	47.83	37.44	33.94	35.61
	Massachusetts (ours:0.3)	75.81	21.83	47.94	37.63	34.21	35.83
	Massachusetts (1)	80.15	17.98	48.61	44.20	23.26	30.48

[8] In Tables 9, 10 and 11, we note the Waterloo Building Dataset, the WHU Building Dataset and the Massachusetts Building Dataset as “Waterloo”,” WHU” and “Massachusetts”. We denote bicubic interpolation (BI), super-resolution using RFANet(RFA) and MSCA-RFANet(ours).

Table 10
Performance of building footprint extraction results using models trained on the Massachusetts Building Dataset (in %).

Training data	Test data	OA	IoU	mIoU	Precision	Recall	F_1 score
Massachusetts (BI: 0.3)	Waterloo (0.12)	65.13	10.74	37.17	12.21	47.04	19.39
	Waterloo (BI: 0.3)	75.39	12.96	43.71	15.58	43.45	22.94
	WHU (0.3)	77.37	23.52	49.60	27.38	62.53	38.08
	Massachusetts (BI: 0.3)	**81.48**	**44.88**	**61.54**	**52.14**	**76.32**	**61.95**
	Massachusetts (RFA: 0.3)	80.98	45.20	61.32	51.19	79.45	62.26
	Massachusetts (ours:0.3)	79.89	43.95	60.04	49.43	79.86	61.06
	Massachusetts (1)	52.03	24.76	33.90	25.95	84.38	39.69
Massachusetts (RFANet: 0.3)	Waterloo (0.12)	68.82	9.61	38.68	11.48	37.21	17.54
	Waterloo (BI: 0.3)	76.83	11.21	43.67	14.21	34.70	20.16
	WHU (0.3)	77.61	19.45	47.89	24.49	48.59	32.57
	Massachusetts (BI: 0.3)	81.75	42.83	60.85	52.91	69.21	59.97
	Massachusetts (RFA:0.3)	**84.57**	**49.49**	**65.66**	**58.34**	**76.54**	**66.21**
	Massachusetts (ours:0.3)	83.81	48.34	64.63	56.67	76.67	65.17
	Massachusetts (1)	53.51	23.29	34.58	25.20	75.47	37.79
Massachusetts (ours: 0.3)	Waterloo (0.12)	64.06	10.17	36.35	11.57	45.66	18.46
	Waterloo (BI: 0.3)	74.23	10.25	41.85	12.67	34.90	18.59
	WHU (0.3)	82.48	15.18	48.55	24.76	28.18	26.36
	Massachusetts (BI: 0.3)	78.08	38.52	56.55	46.34	69.56	55.62
	Massachusetts (RFA:0.3)	80.69	44.40	60.79	50.73	78.07	61.50
	Massachusetts (ours:0.3)	79.60	43.11	59.49	48.97	78.29	60.25
	Massachusetts (1)	64.92	20.77	41.07	26.45	49.16	34.40
Massachusetts (1)	Waterloo (0.12)	80.29	4.81	42.45	7.79	11.18	9.19
	Waterloo (BI: 0.3)	82.22	5.09	43.57	8.48	11.32	9.69
	WHU (0.3)	87.68	21.09	54.18	42.34	29.59	34.84
	Massachusetts (BI: 0.3)	79.85	8.94	44.19	45.45	10.01	16.41
	Massachusetts (RFA: 0.3)	81.78	21.28	51.06	59.23	24.93	35.10
	Massachusetts (ours: 0.3)	**81.55**	**20.86**	**50.73**	**57.69**	**24.62**	**34.51**
	Massachusetts (1)	**87.46**	**47.68**	**66.76**	**68.49**	**61.08**	**64.57**

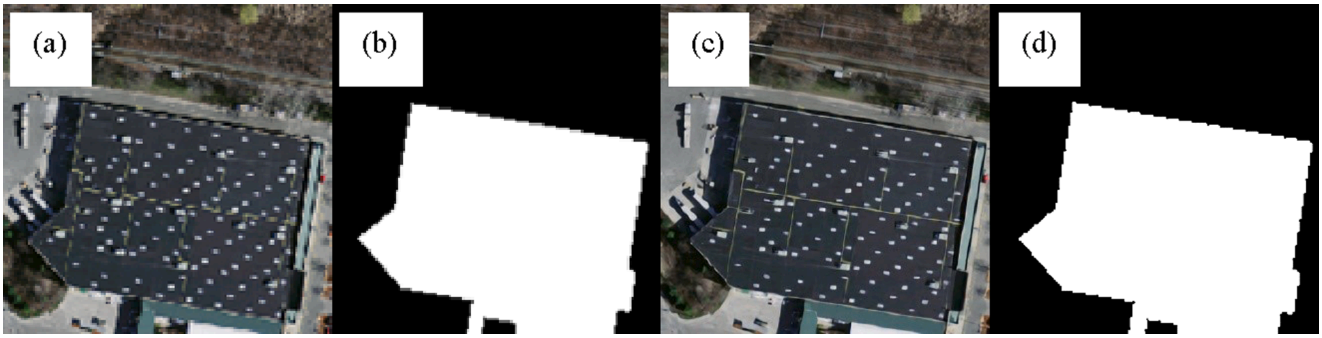


Fig. 4. Example of super-resolved Massachusetts Building Dataset. (a-b) An original image and the matched original mask (1 m/pixel); (c-d) The matched super-resolved image and the interpolated mask (0.3 m/pixel).

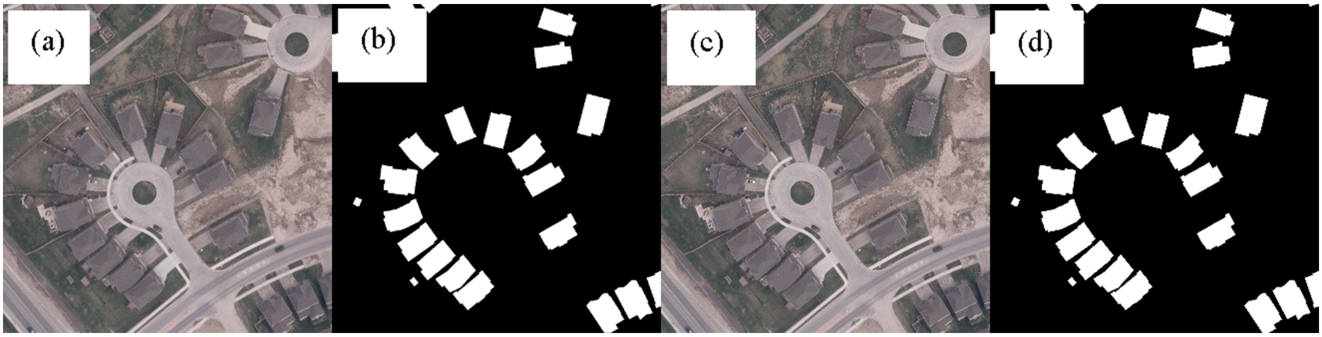


Fig. 5. Example of processed Waterloo building dataset. (a-b) An original image and the matched original mask (0.12 m/pixel); (c-d) The matched interpolated image and the interpolated mask (0.3 m/pixel).

extraction results in last two rows confirmed the positive impact of combining super-resolution and data composition on building footprint extraction.

4.2.4. *Test on "unknown" data*

To further test the impact of super-resolution and data composition on building footprints, as well as the findings we mentioned above, we evaluated building footprint extraction models generated in our experiments on the Inria Building Dataset and compared them with the model

Table 11
Effect of SISR on data fusion (in %).

Training data	Test data	OA	IoU	mIoU	Precision	Recall	F_1 score
A	Waterloo (0.12)	94.36	58.21	76.04	63.14	88.16	73.58
	Waterloo (BI: 0.3)	92.72	43.11	67.70	55.84	65.40	60.25
	WHU (0.3)	94.76	65.06	79.63	71.62	87.67	78.84
	Massachusetts (BI: 0.3)	81.25	12.58	46.66	61.44	13.66	22.35
	Massachusetts (RFA: 0.3)	84.92	34.96	59.28	70.25	41.04	51.81
	Massachusetts (ours:0.3)	84.76	33.98	58.72	70.18	39.72	50.73
	Massachusetts (1)	88.17	51.93	69.19	68.42	68.29	68.36
B	Waterloo (0.12)	79.97	21.48	50.14	24.82	61.46	35.36
	Waterloo (BI: 0.3)	87.11	27.55	57.00	34.37	58.13	43.20
	WHU (0.3)	93.46	59.62	76.20	65.61	86.73	74.71
	Massachusetts (BI: 0.3)	78.71	41.12	58.06	47.54	75.25	58.27
	Massachusetts (RFA:0.3)	79.27	41.68	58.67	48.39	75.03	58.84
	Massachusetts (ours:0.3)	78.30	40.70	57.60	46.93	75.41	57.85
	Massachusetts (1)	77.89	23.14	49.72	39.82	35.58	37.58
C	Waterloo (0.12)	81.65	24.65	52.56	27.99	67.35	39.55
	Waterloo (BI: 0.3)	89.19	37.56	63.00	42.26	77.15	54.61
	WHU (0.3)	93.93	61.47	77.37	67.67	87.03	76.14
	Massachusetts (BI: 0.3)	82.57	43.77	61.80	54.68	68.68	60.89
	Massachusetts (RFA: 0.3)	82.97	45.48	62.81	55.29	71.95	62.53
	Massachusetts (ours:0.3)	82.71	45.09	62.47	54.74	71.90	62.16
	Massachusetts (1)	80.87	15.93	48.04	47.21	19.38	27.48
D	Waterloo (0.12)	92.02	40.79	66.17	54.61	61.71	57.95
	Waterloo (BI: 0.3)	96.04	62.12	78.94	76.22	77.05	76.63
	WHU (0.3)	95.45	68.14	81.55	75.54	87.43	81.05
	Massachusetts (BI: 0.3)	84.67	42.04	62.39	62.39	56.31	59.20
	Massachusetts (RFA: 0.3)	87.67	52.92	69.30	68.29	70.15	69.21
	Massachusetts (ours:0.3)	87.32	52.09	68.69	67.27	69.78	68.50
	Massachusetts (1)	81.54	14.51	47.73	52.06	26.75	25.34
E	Waterloo (0.12)	92.78	41.41	66.90	59.96	57.25	58.57
	Waterloo (BI: 0.3)	94.83	56.27	75.36	66.25	78.88	72.01
	WHU (0.3)	94.16	61.80	77.68	69.43	84.91	76.39
	Massachusetts (BI: 0.3)	86.35	40.98	62.95	73.77	47.97	58.14
	Massachusetts (RFA: 0.3)	88.32	54.21	70.33	70.63	69.99	70.31
	Massachusetts (ours:0.3)	88.25	54.00	70.18	70.41	69.85	70.13
	Massachusetts (1)	82.79	16.29	49.24	64.32	17.91	28.02

trained on the Inria Building Datset. We selected Inria Building Dataset here because this dataset was created to benchmark out-of-distribution generalization errors. As the test set of Inria Dataset was not released, we splitted its training set into training and test set with a ratio of 7:3. As shown in Table 12, although the model trained on 0.12 m resolution Waterloo Building Dataset gave a high OA, the 0.3 m resolution Waterloo Building Dataset and WHU Building Dataset gave high scores in other metrics. In addition, same results can be seen from four different versions of the Massachusetts Building Datasets. The results confirmed that the more similar the test set was to the training set with respect to data distribution and spatial resolution, the higher the model scored on the evaluation metrics. The performance of models trained on C, D, E confirmed our second findings: after super-resolution, the quality of extraction results improved in general. The higher performance of models trained on A and B can be explained as larger data volume used in model training. Consequently, with the best performance of model trained on training set E, the experiment on the "unknown" dataset demonstrated the good performance of combining super-resolution and data composition in construction training dataset for building footprint extraction.

5. Discussion

5.1. Accuracy improvement

Key modules in RCAN and SAN played important roles in performance improvement. In RCAN, long skip connection (LSC), short skip connection and CA are key strategies, which were explored in times two spatial resolution enhancement of Set 5 dataset (Zhang et al., 2018). LSC, which fuses features from the head part and feature from the trunk part via pixel-wise addition, contributed to a 0.32 dB increase in PSNR. Short skip connection, which fuses features from input to that from output of each module, contributed to a 0.36 dB increase in PSNR. CA block contributed to a 0.07 dB increase in PSNR. Short skip connection was inherited in SAN and RFANet; LSC was inherited in RFANet and upgraded to share-source residual group (SSRG) in SAN; CA was embedded in the SCA blocks in our MSCA-RFANet. Therefore, both key modules were inherited or investigated in our MSCA-RFANet.

In SAN, region-level non-local module (RL-NL), SSRG, first-order channel attention (FOCA) and second-order attention (SOCA) were major modules, which were explored in spatial resolution enhancement of Set 5 dataset (Dai et al., 2019). By considering feature interdependencies, SOCA outperformed FOCA and was adopted in SAN, while the implementation of SOCA needed the matrix calculation of large size covariance matrix limiting the size of input images and then the performance of SISR (Dai et al., 2019). Therefore, we did not adopt SOCA in our MSCA-RFANet, although it gave 0.16 dB increase in PSNR in Dai et al. (2019). Share-source skip connection, which is the skip connection between each basic module (RFA + module in our MSCA-RFANet), brought a 0.07 dB increase in PSNR in Dai et al. (2019), which was adopted in our MSCA-RANet and discussed in Section 4.2. RL-NL modules in SAN evenly split input features into top left, top right, bottom left and bottom right and apply non-local modules on each part, which computed long-range dependencies in images. By adding a RL-NL module before and after the trunk part of SAN, the PSNR value increased by 0.04 dB and 0.06 dB, respectively. RL-NL modules could in theory improve our MSCA-RFANet further. However, given the better performance of global context (GC) module in recent work (Cao et al., 2019) compared to non-local module, we explored the former rather than the latter in this work.

Fig. 9 shows the difference in architecture between the NL module and the GC module. The detailed information about the GC module can be found in Cao et al. (2019). The effect of GC module on the performance of MSCA-RFANet is provided in Table 13. We denote the model

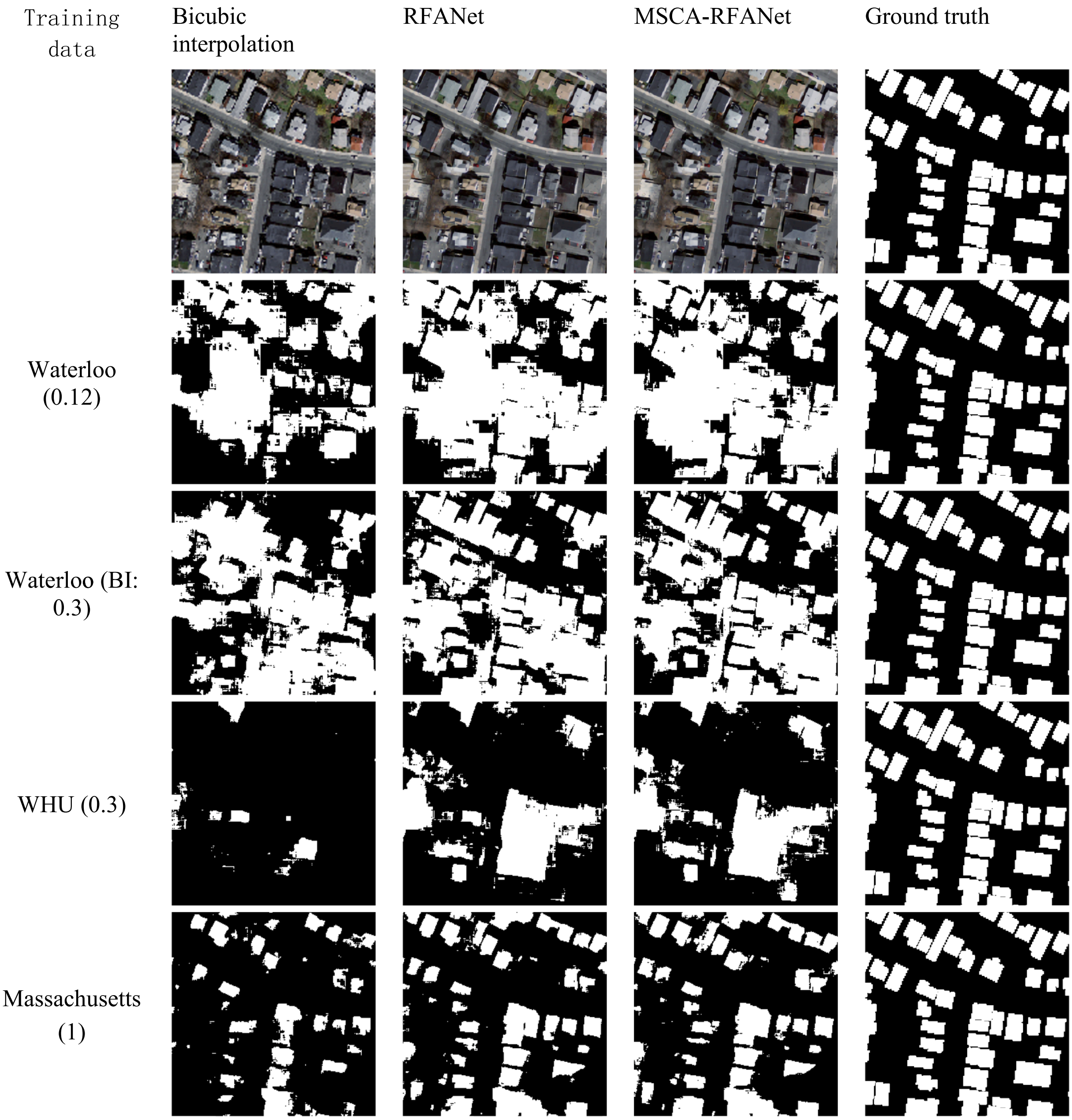


Fig. 6. Visualization of generalization errors and extraction results using models trained on the Waterloo Building Dataset with the pixel size of 0.12 m and 0.3 m (the second and third row), the WHU Building Dataset (the fourth row) and the original Massachusetts Building Dataset (the last row).

with GC block before and after the trunk part of MSCA-RFANet as "+ GC block". As shown in Table 13, the super-resolution performance decreased after adding GC blocks for both datasets. For example, the PSNR value of super-resolution performance on the WHU Building Dataset significantly decreased from 20.38 dB to 20.01 dB after adding GC blocks to MSCA-RFANet. The experiment's result showed the detrimental effect of GC blocks on the performance of our SISR method. In the end we confirmed that the combination of key modules from modules from RCAN, SAN and RFANet used in MSCA-RFANetis is optimal. To further improve SISR performance, powerful networks, such as capsule network (Sabour et al., 2017) and transformer networks (Dosovitskiy et al., 2020) should be considered.

5.2. *Speed improvement*

In this section, we take RFANet as an example and explored the performance of low-precision training and separable convolution methods on accelerating SISR methods. Low-precision training employs the fact that current GPUs (such as Nvidia V100) perform low precision floating point operations much faster than full precision floating point operations (He et al., 2019). Separable convolution defines a convolution group which has fewer parameters compared to standard convolution in calculation (Chollet, 2017). For a detailed introduction, we direct the authors to the original works. After exploration, we applied the most promising acceleration method on top of RFANet and tested its performance of super-resolution on SWOOP 2010 Dataset.

As shown in Table 14, we identify low-precision training (Mixed precision) as a viable acceleration method. The low speed of model training with separable convolution is unexpected. Theoretically, reducing the number of trainable parameters would boost the speed of model training. We believe the low training speed using separable convolution was caused by a non-optimized network implementation in

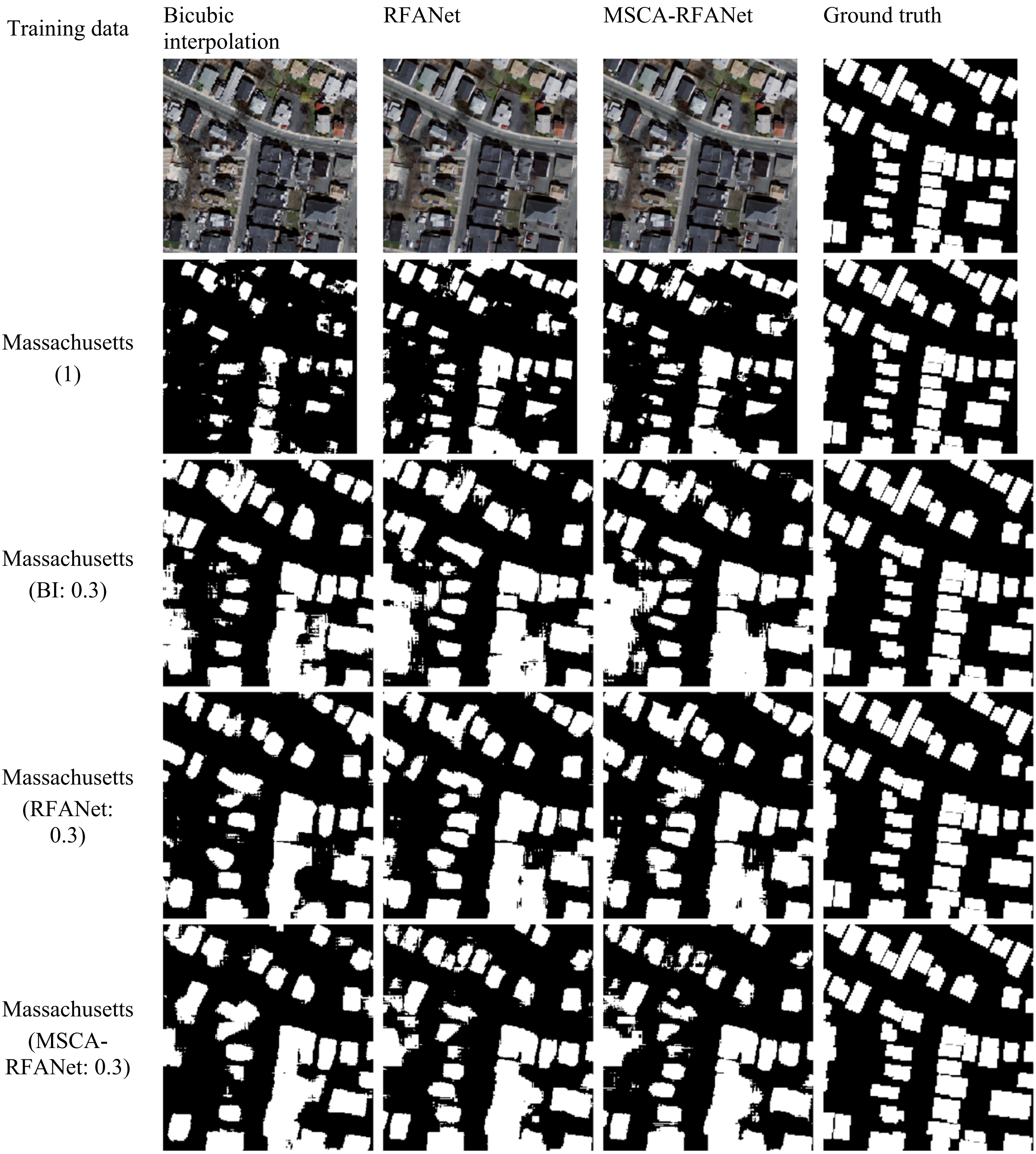


Fig. 7. Visualization of the impact of super-resolution on building footprint extraction.

the deep learning framework (Qin et al., 2018), which could not make full use of GPU capacity.

Following the initial exploration, we applied low-precision training to RFANet training to explore its impact on the super-resolution performance. In Table 15, we denote RFANet with and without low-precision training as "+low-precision training" and "RFANet". As shown in Table 15, by applying low-precision training, the PSNR value of RFANet is significantly dropped from 30.66 dB to 30.03 dB. Although the PSNR value was still higher than that of bicubically interpolated images, it was unacceptable given its low accuracy compared to DL-based SISR methods in this work and low speed compared to bicubic interpolation method. In other words, for our purposes, the speed gain brought by low-precision training could not make up for the accuracy loss.

6. Conclusion

In this paper, we proposed to combine super-resolution and data composition to overcome the generalization errors and improve the accuracy in building footprint extraction. We first proposed a new super-

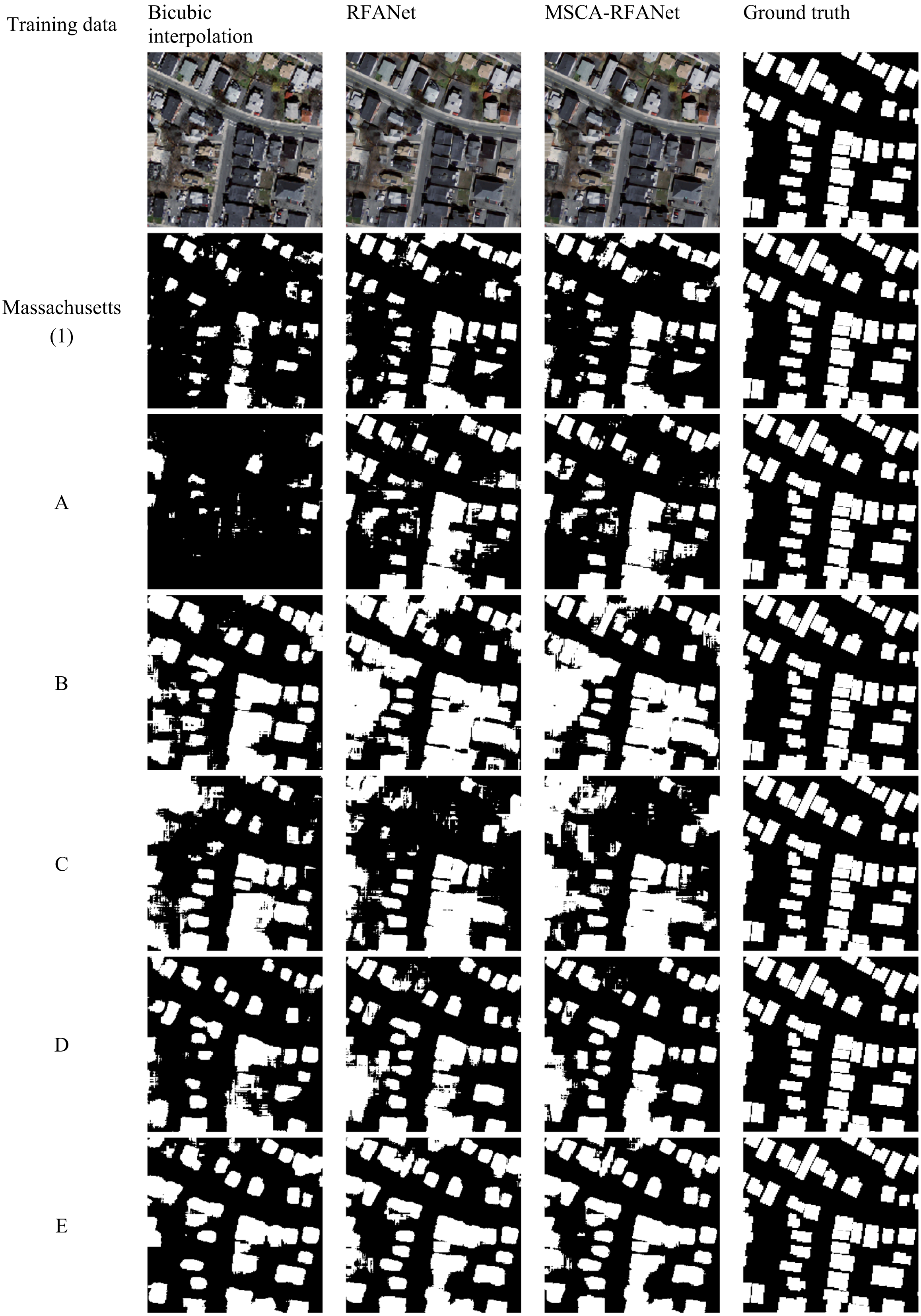


Fig. 8. Visualization of the impact of data composition and super-resolution on building footprint extraction.

Table 12
Test on “unknown” Inria dataset (in %).

Extraction models trained on	OA	IoU	mIoU	Precision	Recall	F_1 score
The Inria Building Dataset	92.35	59.49	75.44	74.01	75.20	74.60
Waterloo (0.12)	86.18	15.90	50.85	63.53	17.49	27.43
Waterloo (0.3)	83.24	34.39	58.01	45.29	58.83	51.18
WHU (0.3)	82.06	26.70	53.76	40.65	43.75	42.15
Massachusetts (BI: 0.3)	83.92	30.27	56.49	46.19	46.76	46.47
Massachusetts (RFA: 0.3)	83.91	25.18	54.09	45.19	36.26	40.23
Massachusetts (ours: 0.3)	83.14	25.04	53.59	42.70	37.70	40.04
Massachusetts (1)	86.19	12.35	49.13	70.31	13.03	21.99
A	88.11	36.24	61.74	64.52	45.26	53.20
B	86.13	33.38	59.24	54.15	46.53	50.05
C	86.41	32.99	59.21	55.57	44.81	49.61
D	87.60	33.14	59.96	62.93	41.19	49.79
E	88.78	39.93	63.90	66.53	49.97	57.07

resolution method based on state-of-the-art methods, named MSCA-RFANet, and then examined the impact of SISR and data composition on building footprint extraction. In the comparison study of different SISR methods, our MSCA-RFANet showed higher performance on both the SWOOP 2010 Dataset and the WHU Building Dataset compared to bicubic interpolation, RCAN, SAN and RFANet. In the super-resolution impact examination, our experimental results showed that using super-resolution to match spatial resolution across datasets resulted in higher performance of building footprint extraction. In addition, data composition achieved a positive impact on building footprint extraction resulting in higher generalizability of trained models. We noticed that by unifying the spatial resolution of different datasets, and training on the resulting composed dataset, the building extraction performance is greatly improved. For building footprint extraction, not only can MSCA-RFANet be used to compose the training set by unifying candidate training datasets to a single spatial resolution, but also as a pre-processing step during testing or deployment to up-sample input images to the spatial resolution used during training. Doing so would, according to our results, greatly alleviate the generalization error in the practical application of building footprint extraction models. We noticed that our MSCA-RFANet achieved very similar results to RFANet, which were superior to those of other methods. We discovered that when super-resolving the test set, despite being the better SISR network as demonstrated by the super-resolution metrics, using our MSCA-RFANet yielded slightly worse building extraction results than the RFANet it was based on, with around 0.1% to 1.1% OA difference. However, when super-resolving the training set (e.g., the Massachusetts dataset from 1 m to 0.3 m), using our MSCA-RFANet produced better building extraction results than when using RFANet on test-sets significantly different from the training set in terms of resolution (11.4% OA improvement on the 1 m Massachusetts Building Dataset) or building distribution (4.87% OA improvement on the WHU Building Dataset). In general, both methods outperformed the other SISR models we tested whether when super-resolving the training or the test set. We believed is caused by how the two SISR models affected the distribution shift across training and test

Table 13
Effect of GC blocks on the performance of our MSCA-RFANet.

Datasets	Models	MSE	RMSE	PSNR (dB)	SSIM
SWOOP	MSCA-RFANet	**36.64**	**5.79**	**30.72**	**0.75**
	+GC block	36.70	5.79	30.70	0.75
WHU	MSCA-RFANet	**68.97**	**8.28**	**20.38**	**0.50**
	+GC block	71.01	8.40	20.01	0.47

Table 14
Time consumed for model training in first epoch.

Models	Time consumed (min)
Original model	**806**
Mixed precision	525
Separable convolution	3645

Table 15
Performance of super-resolution with low-precision training.

Models	MSE	RMSE	PSNR (dB)	SSIM
RFANet	**36.94**	**5.81**	**30.66**	**0.75**
+ low-precision training	39.47	6.06	30.03	0.72

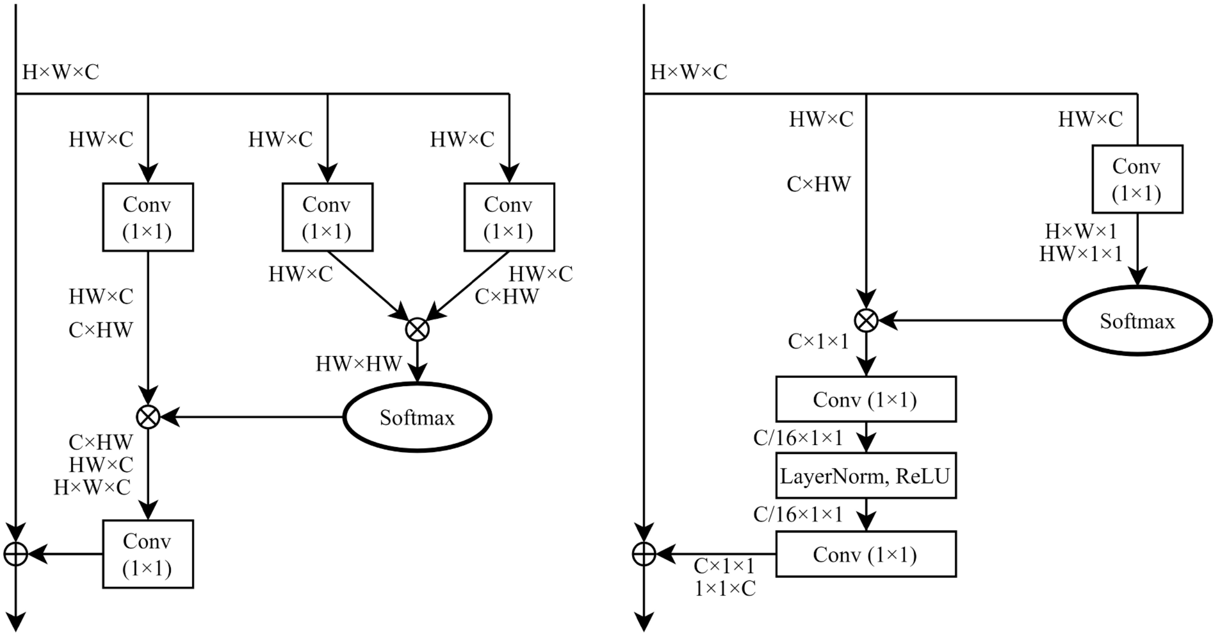


Fig. 9. Left: Non-local (NL) module, right: GC module. H: height, W: width, C: channel of features.

sets, which we plan to investigate further.

Among currently released building dataset, to overcome the generalization error, building datasets with variety building types should be considered. In addition, the availability of the data should also be taken into consideration. Our results showed that in general, when training on a composite dataset by mixing different training sets, the model was more robust to out-of-distribution testing on an unknown (Inria) dataset, achieving up to 88.78% OA compared to low to mid 80′s for single dataset training. In this paper, we recommend using the WHU Building Dataset and the SpaceNet Building Dataset as a base composite dataset, as they include images acquired from Oceania, North America, Europe, Africa and South America and have similar spatial resolution. The Inria Building Dataset, the Waterloo Building Dataset, the Semicity Toulouse Dataset, the SpaceNet Building Dataset, and the ISPRS Vaihingen and Potsdam Datasets should also be considered to enrich the aforementioned datasets which only cover six cities. Further consideration should be made to the sampling of building types to ensure many types of architectures are evenly represented in the composite dataset or that uneven distributions are properly accounted for. Future datasets can offer even more opportunities to enrich future model generalizability, as well as help the remote sensing community better understand cross-dataset differences.

CRediT authorship contribution statement

Hongjie He: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Kyle Gao:** Investigation, Visualization, Writing – review & editing. **Weikai Tan:** Investigation, Writing – review & editing. **Lanying Wang:** Investigation, Writing – review & editing. **Nan Chen:** Investigation. **Lingfei Ma:** Writing – review & editing, Supervision, Funding acquisition. **Jonathan Li:** Resources, Writing – review & editing, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was supported by the Emerging Interdisciplinary Project of Central University of Finance and Economics from China. The first author acknowledges the China Scholarship Council for their support via a doctoral scholarship [No. 201906180088]. The Geospatial Centre of the University of Waterloo is acknowledged for providing the SWOOP 2010 Dataset.

References

Berman, M., Triki, A.R., Blaschko, M.B., 2018. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. Proc. CVPR 4413–4421. https://doi.org/10.1109/CVPR.2018.00464.

Cai, Y., He, H., Yang, K., Fatholahi, S.N., Ma, L., Xu, L., Li, J., 2021. A comparative study of deep learning approaches to rooftop detection in aerial images. Can. J. Remote Sens. 47 (3), 413–431. https://doi.org/10.1080/07038992.2021.1915756.

Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H., 2019. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. Proc. ICCVW 1971–1980. https://doi.org/10.1109/ICCVW.2019.00246.

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S., 2017. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. Proc. CVPR 5659–5667. https://doi.org/10.1109/CVPR.2017.667.

Chen, Q., Wang, L., Wu, Y., Wu, G., Guo, Z., and Waslander, S.L. 2018a. Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings. arXiv 2018. arXiv preprint arXiv:1807.09532. https://doi.org/10.1016/j.isprsjprs.2018.11.011.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. Proc. CVPR 1251–1258. https://doi.org/10.1109/CVPR.2017.195.

Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L., 2019. Second-order attention network for single image super-resolution. In Proc. CVPR 11065–11074. https://doi.org/10.1109/CVPR.2019.01132.

Dong, C., Loy, C.C., He, K., Tang, X., 2015. Image super-resolution using deep convolutional networks. IEEE Trans. Patt. Anal. Mach. Intell. 38 (2), 295–307. https://doi.org/10.1109/TPAMI.2015.2439281.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Duveiller, G., Defourny, P., 2010. A conceptual framework to define the spatial resolution requirements for agricultural monitoring using remote sensing. Remote Sens. Environ. 114 (11), 2637–2650. https://doi.org/10.1016/j.rse.2010.06.001.

Farrow, C.L., Shaw, M., Kim, H., Juhás, P., Billinge, S.J., 2011. Nyquist-Shannon sampling theorem applied to refinements of the atomic pair distribution function. Phys. Rev. B 84 (13), 134105. https://doi.org/10.1103/PhysRevB.84.134105.

GFDRR Labs, 2020. Open Cities AI Challenge Dataset, Version 1.0, Radiant MLHub. https://doi.org/10.34911/rdnt.f94cxb.

He, H., Jiang, Z., Gao, K., Fatholahi, S.N., Cai, Y., Tan, W., Hu, B., Qing, L., Xu, H., Li, J., 2021. Waterloo Building Dataset, V1. Harvard Dataverse. https://doi.org/10.7910/DVN/EXRA2V.

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M., 2019. Bag of tricks for image classification with convolutional neural networks. In Proc. CVPR 558–567. https://doi.org/10.1109/CVPR.2019.00065.

Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans. Geosci. Remote Sens. 57 (1), 574–586. https://doi.org/10.1109/TGRS.2018.2858817.

Kim, J., Kwon Lee, J., Mu Lee, K., 2016. Deeply-recursive convolutional network for image super-resolution. In Proc. CVPR 1637–1645. https://doi.org/10.1109/CVPR.2016.181.

Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H., 2017. Deep Laplacian pyramid networks for fast and accurate super-resolution. In Proc. CVPR 624–632. https://doi.org/10.1109/CVPR.2017.618.

Lambert, J., Liu, Z., Sener, O., Hays, J., Koltun, V., 2020. MSeg: A composite dataset for multi-domain semantic segmentation. In Proc. CVPR 2879–2888. https://doi.org/10.1109/CVPR42600.2020.00295.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W., 2017. Photo-realistic single image super-resolution using a generative adversarial network. In Proc. CVPR 4681–4690. https://doi.org/10.1109/CVPR.2017.19.

Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K., 2017. Enhanced deep residual networks for single image super-resolution. In Proc. CVPRW 136–144. https://doi.org/10.1109/CVPRW.2017.151.

Liu, J., Zhang, W., Tang, Y., Tang, J., Wu, G., 2020. Residual feature aggregation network for image super-resolution. In Proc. CVPR 2359–2368. https://doi.org/10.1109/CVPR42600.2020.00243.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In Proc. IGARSS 3226–3229. https://doi.org/10.1109/IGARSS.2017.8127684.

Mao, X.J., Shen, C., Yang, Y.B., 2016. Image restoration using convolutional auto-encoders with symmetric skip connections. In Proc. NeurIPS 29, 2802–2810.

Marivani, I., Tsiligianni, E., Cornelis, B., Deligiannis, N., 2020. Joint image super-resolution via recurrent convolutional neural networks with coupled sparse priors. In Proc. ICIP 868–872. https://doi.org/10.1109/ICIP40778.2020.9190644.

Mnih, V., 2013. Machine learning for aerial image labeling. University of Toronto. PhD Thesis.

Mohanty, S.P., Czakon, J., Kaczmarek, K.A., Pyskir, A., Tarasiewicz, P., Kunwar, S., Rohrbach, J., Luo, D., Prasad, M., Fleer, S., Göpfert, J.P., Tandon, A., Mollard, G., Rayaprolu, N., Salathe, M., Schilling, M., 2020. Deep learning for understanding satellite imagery: An experimental survey. Front. Artif. Intell 3. https://doi.org/10.3389/frai.2020.534696.

Molchanov, P., Tyree, S., Karras, T., Aila, T., Kautz, J., 2019. Pruning convolutional neural networks for resource efficient inference. In 5th International Conference on Learning Representations, ICLR 2017-Conference Track Proceedings.

Qin, Z., Zhang, Z., Li, D., Zhang, Y., Peng, Y., 2018. Diagonalwise refactorization: An efficient training method for depthwise convolutions. In Proc. IJCNN 1–8. https://doi.org/10.1109/IJCNN.2018.8489312.

Roscher, R., Volpi, M., Mallet, C., Drees, L., Wegner, J.D., 2020. SemCity Toulouse: A benchmark for building instance segmentation in satellite images. ISPRS Annals 5, 109–116. https://doi.org/10.5194/isprs-annals-V-5-2020-109-2020.

Sabour, S., Frosst, N., Hinton, G.E., 2017. December. Dynamic routing between capsules. In Proc. NeurIPS, 3859–3869.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proc. CVPR 1874–1883. https://doi.org/10.1109/CVPR.2016.207.

Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W. and Wang, J., 2019. High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514.

Tai, Y., Yang, J., Liu, X., 2017. Image super-resolution via deep recursive residual network. In Proc. CVPR 3147–3155. https://doi.org/10.1109/CVPR.2017.298.

Tong, T., Li, G., Liu, X., Gao, Q., 2017. Image super-resolution using dense skip connections. In Proc. CVPR 4799–4807. https://doi.org/10.1109/ICCV.2017.514.

Van Etten, A., Lindenbaum, D., and Bacastow, T.M. 2018. SpaceNet: A remote sensing dataset and challenge series. arXiv preprint arXiv:1807.01232.

Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In Proc. ECCV 3–19. https://doi.org/10.1007/978-3-030-01234-2_1.

Yang, C.Y., Ma, C., Yang, M.H., 2014. Single-image super-resolution: A benchmark. In Proc. ECCV 372–386. https://doi.org/10.1007/978-3-319-10593-2_25.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y., 2018. Image super-resolution using very deep residual channel attention networks. In Proc. ECCV 286–301. https://doi.org/10.1007/978-3-030-01234-2_18.

Zhao, P., Zhang, J., Fang, W., Deng, S., 2020. SCAU-Net: Spatial-channel attention U-Net for gland segmentation. Front. Bioeng. Biotech. 8, 670. https://doi.org/10.3389/fbioe.2020.00670.