Transferring Transformer-Based Models for Cross-Area Building Extraction From Remote Sensing Images

Chunping Qiu, He Li, Wenyue Guo, Xin Chen, Anzhu Yu, Xiaochong Tong, and Michael Schmitt, *Senior Member, IEEE*

***Abstract*—Extracting buildings from remote sensing (RS) images is an important task with a variety of applications. Considerable attention has focused on achieving new state-of-the-art (SOTA) accuracy with more and more advanced deep learning (DL) models. However, the developed models still hardly generalize across geographical areas, hindering the practical use of SOTA approaches. To attack this problem, we established a baseline for model cross-area generalization ability using available datasets for building extraction (BE). In addition to two popular fully convolutional neural network (FCN) based models, we first adapted two novel transformer-based models, shifted windows (Swin) transformer and SegFormer, which are all able to output SOTA accuracy with no big difference when tested within one area. However, experimental results show that all models fail to generalize to a different area. We then propose to fine-tune pretrained models from one area on a small subset of an unseen area, the effectiveness of which depends on the model choice and the data size for tuning. By jointly taking advantage of the transfer learning idea and the multiscale feature learning ability of SegFormer, a distinct improvement has been achieved compared to results from Swin transformer and FCN-based models trained on the same amount of data. Commonly used metric, Intersection over Union, can be increased from 38.97% to 70.86%, and from 48.36% to 74.51%, when using 10% and 30% subset of the targeting area, respectively. The influence of model choice and data size for tuning has also been investigated. Our work contributes to complementing the algorithm development and within-area model evaluation in the hot field of BE from RS images.**

***Index Terms*—Benchmark dataset, building extraction (BE), convolutional neural networks (CNNs), generalization ability, remote sensing (RS), transformer.**

Manuscript received February 21, 2022; revised April 24, 2022; accepted May 6, 2022. Date of publication May 16, 2022; date of current version June 1, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42101458, Grant 41801388, and Grant 42130112, in part by the Natural Science Foundation of Henan Province of China under Grant 212300410096 and Grant 222300420592, and in part by the National Key Research and Development Program of China under Grant 2018YFB0505304. *(Correspondong authors: He Li; Wenyue Guo.)*

He Li, Wenyue Guo, Xin Chen, Anzhu Yu, and Xiaochong Tong are with the PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China (e-mail: lihe_5115@zxiat.org; guowyer@163.com; xinchen_cosmos@126.com; anzhu_yu@126.com; txchr@aliyun.com).

Chunping Qiu is with the PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China, and also with the Research Center for Artificial Intelligence, National Innovation Institute of Defense Technology, Academy of Military Science, Beijing 100850, China (e-mail: chunping.qiu@outlook.com).

Michael Schmitt is with the Bundeswehr University Munich, 85579 Neubiberg, Germany (e-mail: michael.schmitt@unibw.de).

Digital Object Identifier 10.1109/JSTARS.2022.3175200

I. Introduction

BUILDING extraction (BE) from remote sensing (RS) images is a hot topic because buildings are fundamental to a wide variety of high-level tasks such as land-use statistics, urban planning, and population evaluation [1], [2]. It remains a challenging semantic segmentation task due to a large within-class and small between-class variance, diverse shapes of buildings, complex backgrounds, and high requirements on the boundaries [3].

In recent years, deep convolutional neural networks (CNNs) and fully convolutional neural networks (FCNs) have demonstrated state-of-the-art (SOTA) performance, the same as in other visual tasks in the field of earth observation [4]. Because of the powerful feature learning ability, FCN and its variants have been the mainstream for BE. This kind of model features an encoder–decoder structure, i.e., a contracting path gradually down-sampling the resolution of feature maps and an expanding path progressively restoring the resolution. This way, hierarchical features and contextual information can be modeled for accurate prediction of buildings from backgrounds. To achieve accurate results and sharp boundaries, an enormous number of variants have been proposed in the past years. The work [5] used a distance map as a building mask to pose more constraints for feature learning. The work [6] designed a dense spatial pyramid pooling to extract dense and multiscale features simultaneously and used a focal loss to suppress the impact of incorrect labels. The work [7] proposed a boundary-aware perceptual loss, consisting of a loss network and transfer loss functions, which achieves much improvement over the cross-entropy loss when tested with PSPNet and UNet. Similarly, [8] also used a structural feature constraint module for boundary refinement. For postprocessing, regularization algorithms were developed to refine mixed pixels and building boundaries into regular shapes. In [9] and [10], boundaries were generalized with detected edges and feature points. In [11], a dense conditional random field was used to smoothen extraction results. Based on this, [12] developed a graph-based conditional algorithm to further solve the boundary problem by combining some strategies to extract and fuse multilevel and multiresolution features. To summarize, main strategies include global information mining, boundary contour refinement, dilated convolution, multiscale prediction [13], and feature fusion [3].

While being the standard paradigm for semantic segmentation tasks, FCN-based models are not good at modeling global contextual information, limited by their receptive field. To tackle this problem, there have been various solutions. One is based on attention mechanisms, e.g., integrating attention modules into FCN models [14]. Another one is resort to transformer architectures that translate 2-D image-based tasks into 1-D sequence-based tasks and feature powerful sequence-to-sequence modeling [15], [16]. Recently, the computation issue of transformers motivated novel approaches such as shifted windows (Swin) transformer and SegFormer [17], [18].

There has been an effort to adapt or improve Swin transformer for multiclass semantic segmentation tasks using RS images. The work [19] proposed a memory-augmented transformer using a memory-based global relationship guidance module and a transformer-based local feature extraction module. The work [20] proposed context transformer to combine the extracted feature from a global and a local CNN branch. The work [21] proposed a bilateral awareness network, consisting of a two-path network, one based on transformer (ResT) and the other based on CNNs, for feature extraction and an attention-based feature aggregation module. The authors in [22] proposed an efficient hybrid transformer by combining a CNN-based encoder and a Swin transformer-based decoder, achieving efficiency-accuracy balance for urban scene segmentation. The authors in [23] adaptively fused multilevel features from CNNs and Swin transformer with a self-attentive mechanism. The work [24] used Swin transformer as a backbone for semantic segmentation from fine-resolution RS images, coupled with a proposed decoder for feature aggregation. The authors in [25] proposed an efficient transformer with multilayer perception (MLP) head and modules for edge enhancement, after investigating Swin transformer. Besides semantic segmentation, other RS applications based on transformer include object detection and instance segmentation [26].

There are also attention and transformer-based studies with a focus on BE from RS images. The authors in [27] used a HRNet-like architecture based on multibranch feature encoding and attention mechanism. The authors in [28] proposed a U-Net that combines self-attention and reconstruction-bias modules. The authors in [29] combined a U-shaped encoder–decoder structure and an asymmetric pyramid nonlocal block. Also using nonlocal block, [13] proposed a global multiscale encoder–decoder network. The work [30] proposed to adaptively fuse the multiscale learned features by Swin transformer.

Most of these newly developed models are benchmarked on publicly available datasets and demonstrate outperformance, while some application-oriented studies also created their own datasets [14], [31]. For example, by exploiting low- and high-level features and designing a boundary refinement module, the Intersection over Union (IoU) metric has been improved from 90.86% to 91.4% when tested on the WHU dataset [32]. These works can provide interesting methodological insights, even though there is usually only one percentage point rise in accuracy. On the other hand, the generalization ability of the proposed models, particularly across geographical areas are rarely tested. This hinders application-oriented studies, as it is usually uncertain how a deep learning (DL) model will perform when tested on data subject to domain shift [33]. This, in general, is an interesting and important topic when using machine learning in RS and quick answers are transfer learning and domain adaptation [34]–[37]. While there are BE challenges targeting on cross-area generalization, such as the DeepGlobe 2018 challenge [38], the datasets are underused in literature due to the unavailability of testing labels. Consequently, there is still little reference when it comes to generalizing DL models for BE from RS images.

In this study, we proposed a practical approach to improving model performance in geographically distinct areas by adapting advanced transformer-based models coupled with a transfer learning idea. We reviewed related work to identify suitable datasets and models and experimentally demonstrated the need for advanced models and strategies to enable cross-area BE applications. Our investigation provides a new perspective to the accuracy convergence in the hot BE topic. The particular contributions of this work are as follows.

1) We adapted two transformer-based models, Swin and SegFormer, for BE from RS images, and we proposed to test model performance in a realistic cross-area setting by adjusting public datasets.
2) We showed that Swin and SegFormer can provide SOTA results in a common setup but can be worse compared to SOTA convolution-based models, depending on aspects such as data size and test setting.
3) We found that SegFormer can provide higher accuracy than both Swin-based and convolution-based models when transferred from a different area and fine-tuned in the target area. The effectiveness has been tested with several experiments on open datasets.

The rest of this article organized as follows. Section II elaborates the choice of models, and the adapted Swin and SegFormer structure for our study. Section III details descriptions about the datasets and the experimental setup for testing model cross-area generalization ability. Section IV evaluates BE accuracy of different models when tested within one dataset and in a cross-area setting. This section also visualizes and compares BE results from different settings. Section V discusses the model generalization issue described above, based on the interpretation and analysis of the achieved results, and points out the remaining challenges and the possible solutions for the future work. Finally, Section VI concludes this article.

II. Adapted Swin and SegFormer for Building Extraction

Many classic networks have been tested on the aerial WHU building dataset [39], e.g., U-Net [40], U-NetPlus [41], SegNet [42], DeconvNet [43], HRNetv2 [44], Refinenet [45], Se-U-Net [46], Ma-Fcn [10], PSPNet [47], and DeepLab V3 [48]. The best results are 91.4% IoU and about 95.51% F1-score by a coarse-to-fine boundary refinement network based on VGG-16 [32], and 90.77% IoU and 95.4% F1-score by MAP-Net [29]. These methods achieve similarly high accuracy by effectively exploiting multiscale features. Taking this point into account, we

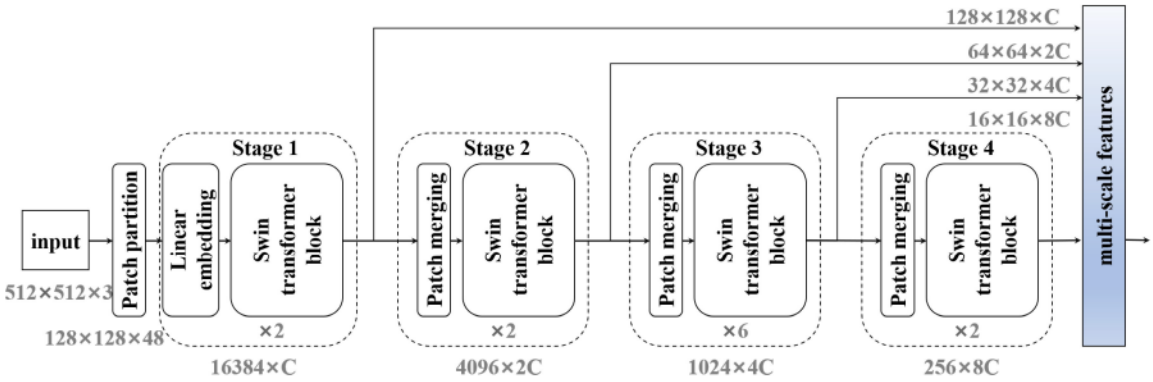


Fig. 1. Architecture of a tiny Swin transformer as a backbone for BE. The output multiscale features from all four stages are fed into a decoder. Changing C and layer numbers of each stage will result in model variants of different sizes.

took two representative FCN-based models, MAP-Net [27] and Foreground-Aware Relation Network (FarSeg) [49], as baselines for our investigation.

Transformer-based models have also been tested on the aerial WHU building dataset, and best results are 88% IoU and 88.2% F1-score by an adapted Swin transformer [30]. Considering the recent great success of transformers, we adapted two models, Swin and SegFormer. The former takes both the advantages of CNNs and transformers by using a hybrid structure combining convolutional blocks and transformer blocks, and the latter sets new SOTA for publicly available semantic segmentation datasets in the field of computer vision by designing a hierarchical transformer encoder and a lightweight All-MLP decoder. These two models can generate both high-resolution fine features and low-resolution coarse features, thus facilitating the BE purpose.

The structure of the adapted Swin transformer as a backbone is illustrated in Fig. 1. For the head part, we utilized a simple structure similar to MAP-Net head and SegFormer for feature fusion, in addition to the original UPerNet-based head [17], and the networks are referred to as Swin-M and Swin-U, respectively.

As shown in Fig. 1, Swin transformer as a backbone consists of the following core modules.

1) **Patch partition**. In this process, the input image patch is split into nonoverlapping smaller patches, ending up with some feature vectors corresponding to these small patches, which are further processed by the subsequent operations. For instance, an input image of the size of 512×512 is partitioned into $128 \times 128 = 16\ 384$ small patches with the size of $4 \times 4 \times 3$ pixels.
2) **Linear embedding** is the first operation applied to the processed feature vectors after patch partition, which outputs 16 384 new feature maps with a higher dimension of C. For tiny, small, and base model, C is 96, 96, and 128, respectively.
3) **Swin transformer block**. The embedded feature vectors then go through the first stage, where the main operation is multihead self-attention within two sequential transformer and Swin transformer blocks. During this process, window-scale and local features are efficiently learned by applying self-attention within each window, e.g., with a size of 7×7 and 12×12 patch when the input image is with a size of 224×224 and 384×384, respectively [17], which is mainly to address the computation problem. Cross-window information can also be learned by the shift of window, i.e., Swin. When the original input is with a size of 512×512, as the case for BE in this study, the window size is set as 12×12, considering the ground sampling distance of the used high-resolution images and the computation efficiency. In addition, padding is used when necessary to fit the window into the whole patch.
   The following stages consist of similar operations, except that surrounding small patches are first merged. Layer number of stages 1, 2, 3, and 4, are $\{2, 2, 6, 2\}$, $\{2, 2, 18, 2\}$, and $\{2, 2, 18, 2\}$ for tiny, small, and base model, respectively. The size and dimension of feature maps, as well as the equivalent receptive field are kept the same within each stage.
4) **Patch merging**. This is similar to a down-sampling process in CNNs. The size of the partitioned patches are 4×4, 8×8, 16×16, and 32×32 for stages 1, 2, 3, and 4, respectively. This leads to a CNN-like hierarchical structure, so that features of different scales can be progressively learned.

We used output features from all four stages to benefit from multiscale features in the head network. Features from stages 2, 3, and 4 are first up-sampled into 128×128, and then concatenated with features of stage 1. Afterwards, the combined features are squeezed in the channel dimension and progressively up-sampled into the same size as the input, i.e., 512×512, using interpolation layers followed by convolution layers to decrease the feature channels.

Fig. 2 illustrates the SegFormer architecture used in this study. Similar to Swin, its encoder also outputs hierarchical feature representation from the four-stage encoder, with the sizes of 128×128, 64×64, 32×32, and 16×16, respectively. One difference comes in the module of overlapped patch merging. Unlike the nonoverlapping patch merging between stages in Swin, it aims at preserving the local continuity around patches. Padding is needed to output feature maps with the same size as the nonoverlapping patch merging case. Another difference is the efficient self-attention module. Instead of performing with-in window self-attention, SegFormer relied on a sequence

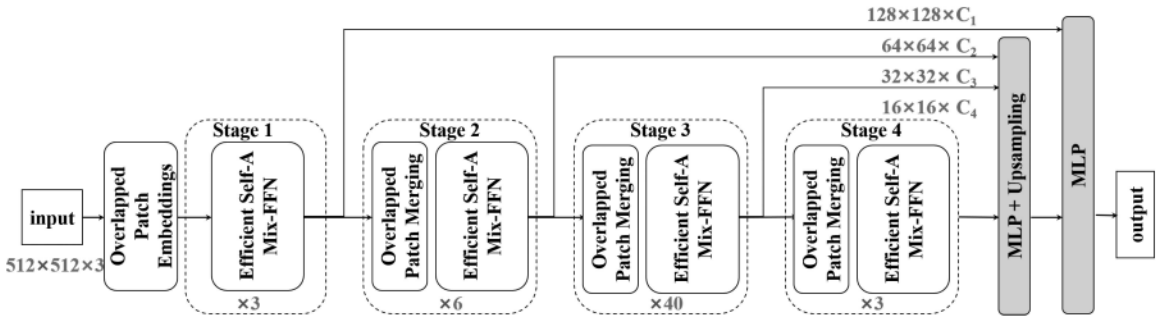


Fig. 2. Architecture of SegFormer, consisting of a four-stage encoder and an MLP-based decoder.

reduction process to address the computation problem. Besides, the mixed feed-forward network (Mix-FFN) module within each stage is to provide positional information by mixing a 3×3 convolution and an MLP. Mix-FFN takes output features from the efficient self-attention module as inputs, and these features are processed with a convolution layer with a kernel size of 1, a convolution layer with a kernel size of 3, an activation layer, and a convolution layer with a kernel size of 1. The processed results are then added with the inputs for the final output of the Mix-FFN module.

One of the novelties of SegFormer is the lightweight decoder, the goal of which is the same as MAP-Net and Swin-M, i.e., aggregating multiscale features for final dense prediction. There are four steps in the SegFormer decoder. First, one MLP unifies the multilevel features from each of the four stages in the encoder to output features of the same channel dimension C. Then, features are up-sampled to the size of 128×128 and concatenated together, which are fused subsequently by a following MLP layer. Finally, one MLP layer takes the fused features to predict a segmentation mask, which is then up-sampled to the input size of 128×128.

III. Experimental Setup

A. Datasets for Cross-Area Model Generalization Testing

Commonly used airborne datasets for BE include the WHU building dataset [50], the Massachusetts building dataset [51], the International Society for Photogrammetry and Remote Sensing (ISPRS) Vaihingen and Potsdam datasets [52], the Inria dataset [53], and the Aerial Imagery for Roof Segmentation (AIRS) dataset [54], which are all not ideal for testing cross-area model generalization ability for the following reasons. Because of the officially fixed and standard split of training, validation, and testing datasets, the WHU building dataset is often used to compare different models. The whole dataset is spatially split instead of randomly split, but it is only from one single city, i.e., Christchurch, New Zealand, with a size of 450 km^2. Therefore, it is not suitable for testing cross-area model generalization. The Massachusetts building dataset is also made up of images from only one single city, i.e., Boston, Massachusetts, and the split is not consistent in different studies, making it difficult to interpret different model performances [13], [55]. The ISPRS Vaihingen and Potsdam datasets only cover a small area with the size of 2 and 11 km^2, respectively. The Inria dataset includes images from different cities, but the test data used in the literature is a subset of the training data from the same areas, i.e., the validation set described in [53], due to the unavailability of the official ground truth (GT) labels [13], [55]. Therefore, cross-city tests of the Inria dataset have rarely been reported in the literature, and the comparison is not easy due to differences in data processing, splitting strategies, and augmentation methods [12]. Focusing on building roofs instead of building footprints, the AIRS dataset also covers only one single city, i.e., Christchurch, New Zealand, and the training, validation, and test datasets are randomly, instead of spatially, split.

Commonly used space-borne datasets include the WHU building dataset, the Urban 3-D Challenge dataset [56], and the Deep Globe Building Extraction Challenge (DG-BEC) dataset [38]. The WHU building satellite dataset is not suitable for testing cross-area model generalization, as the six satellite images are from the same geographical area. The Urban 3-D Challenge dataset is not suitable for cross-area test either, as the training, validation, and test sets are all from the same two AOIs, i.e., the test area is not entirely distinct from the training area. The DG-BEC dataset includes WorldView-3 satellite images and building labels from four areas: Las Vegas, Paris, Shanghai, and Khartoum. The data of different cities is released in separate folders instead of combined, making it possible for cross-area tests. In literature, usually only a part of the dataset, e.g., only available data form one or two cities, is used, in which case, the training and test split is usually in a random format [27], [57].

Aiming at cross-area test, the DeepGlobe-Building dataset was chosen and prepared in a novel generalization-testing manner in our study, ending up with five subdatasets. Specifically, we used pan-sharpened RGB images from three cities located on different continents, with a size of 650×650. No other bands were used without loss of generality. In addition, the WHU aerial dataset is also used to benchmark our methods since it is very often used in literature, and an official split is followed. There are three bands, R, G, and B, for each sample, with a size of 512×512. The number of image patches for training, validation, and test are listed in Table I, and the exact splits of training, validation, and test sets are fixed for all experiments. Taking the DG-BEC-Shanghai-30 and DG-BEC-Shanghai-10 dataset for example, validation and test sets are exactly the same and the 458 training patches of the smaller set is part of the larger set. This setting helps interpreting comparative results from different settings.

TABLE I
SAMPLE NUMBER OF TRAINING, VALIDATION, AND TEST SET FOR DIFFERENT EXPERIMENTS IN THIS STUDY

dataset	train	val	test
DG-BEC-Vegas	2310	386	1155
DG-BEC-Shanghai-30	1374	459	2749
DG-BEC-Shanghai-10	458	459	2749
DG-BEC-Paris-30	344	116	688
DG-BEC-Paris-10	114	116	688
WHU	4736	1036	2416

B. Implementation Details

Pretrained weights are important for transformer-based models [25]. Therefore, all transformer-based models are initialized using the pretrained weights on ImageNet-1K [17], [18]. All models were implemented using the PyTorch framework. We used binary cross-entropy as the loss function and synchronized trained models over four GPUs with a total of eight images per mini-batch (two images per GPU). The synchronized batch normalization was used for cross-GPU communication of statistics in the batch normalization layer. During training on the DG-BEC dataset, we applied common data augmentation skills including random resize with ratios 1.25 and 1.5 followed by cropping to 512 $\times$ 512, random horizontal and vertical flipping, and random rotation in four directions. We trained the models using AdamW optimizer for 500 epochs, and the learning rate was set to an initial value of 0.001. During training on the WHU dataset for 300 epochs, only random flipping and rotation was used for augmentation, for fair comparison with previous work [27]. During inference, no data augmentation is used for simplicity and fair comparison. As random cropping can result in different testing accuracy, we cropped and saved the test samples as a preprocessing setup for the DG-BEC dataset.

C. Experimental Setup to Assess Cross-Area Model Generalization Ability

To benchmark SOTA models first, we trained models using DG-BEC-Vegas training set and tested on DG-BEC-Vegas test set. This benchmarking experiment is also done on the WHU dataset because it is often used in BE studies. This way, we can compare model performance in a common experimental setting, i.e., within one area and one dataset.

In order to test cross-area model generalization ability, models are first trained on the DG-BEC-Vegas training set. Subsequently, there are different options. One is directly testing the trained models on the DG-BEC-Shanghai/Paris test set. The other is tuning the trained models with the DG-BEC-Shanghai/Paris training set, in which case, there are two settings, depending on the size of the training set. To demonstrate the effectiveness of the pretrained weights, we compared the tuning results to those without tuning, and the only difference is whether to use pretrained weights from the DG-BEC-Vegas training set when training on the DG-BEC-Shanghai/Paris training set.

As in literature, we report model performance with respect to the target class, i.e., the building class, including precision, recall, F1-score, and IoU.

TABLE II
BE ACCURACY OF DIFFERENT MODELS TESTED ON DG-BEC-VEGAS DATASET

Model	IoU	Precision	Recall	F1
FarSeg	0.8542	**0.9283**	0.9145	0.9214
MAP-Net	0.8338	0.9125	0.9063	0.9094
Swin-M-T	0.8430	0.9107	0.9189	0.9148
Swin-M-S	0.8452	0.9073	0.9251	0.9161
Swin-M-B	0.8285	0.8829	**0.9308**	0.9062
Swin-U-S	0.8400	0.9146	0.9115	0.9130
SegFormer	**0.8571**	0.9235	0.9227	**0.9231**

The best results are indicated in bold for each metric.

TABLE III
BE ACCURACY OF DIFFERENT MODELS TESTED ON WHU DATASET

Model	IoU	Precision	Recall	F1
farSeg	0.8842	0.9372	0.9399	0.9385
Mapnet	0.8827	0.9444	0.9311	0.9377
Swin-M-S	0.8963	0.9480	0.9426	0.9453
Swin-U-S	**0.9068**	**0.9504**	**0.9518**	**0.9511**
SegFormer	0.9038	0.9493	0.9496	0.9495

The best results are indicated in bold for each metric.

IV. EXPERIMENTAL RESULTS

BE results are illustrated and compared in this section. For both quantitative and qualitative comparisons, we listed results under the setting of testing within one dataset, in addition to the cross-area testing setting. This is to demonstrate the difference of these two scenarios and enable a wide interpretation of the generalization ability of transformer-based models.

A. Quantitative Assessment of Building Extraction Results

1) Comparison of Model Performance Within One Area: Table II lists BE accuracy of different models tested on the DG-BEC-Vegas dataset. From Table II, it can be seen that these models are all able to produce similar high accuracy, despite different architectures and different model sizes. The relative worse results are from MAP-Net and Swin-M-B, while the relative better results are from SegFormer. Compared to Swin-M-S, there is no benefits from the larger model, Swin-M-B. Therefore, Swin-M-B was not tested further in our study. Swin-M-T was not further tested either as it provides worse results than Swin-M-S.

BE accuracy tested on WHU dataset is listed in Table III. It can be found that there are no big differences among the results produced by different models, which is consistent as in Table II. The worse results are from MAP-Net and the better results are from Swin-U-S. The achieved SOTA accuracy on the benchmark WHU dataset shows the correctness of our implementations in the experiments.

2) Comparison of Model Cross-Area Performance: Table IV lists BE accuracy of directly applying the models trained with DG-BEC-Vegas training set on DG-BEC-Shanghai test set. Results from all the five investigated models are bad, and the better one is by SegFormer. Besides, the recall metric is much lower for all the models, meaning a high false negative in the BE results.

TABLE IV
BE ACCURACY OF DIFFERENT MODELS TRAINED ON DG-BEC-VEGAS TRAINING SET AND TESTED ON DG-BEC-SHANGHAI TEST SET

Model	IoU	Precision	Recall	F1
FarSeg	0.1902	0.5686	0.2222	0.3195
MAP-Net	0.1764	0.4095	0.2366	0.2999
Swin-M-S	0.1334	0.4903	0.1549	0.2354
Swin-U-S	0.1231	0.7506	0.1283	0.2192
SegFormer	**0.2240**	**0.8154**	**0.2360**	**0.3661**

The best results are indicated in bold for each metric.

Tables V and VI list BE accuracy after tuning models (trained with the DG-BEC-Vegas training set) on a small subset of the DG-BEC-Shanghai dataset. For fair comparison, results are also listed when not using weights from the DG-BEC-Vegas training set, i.e., models are only initialized with weights from ImageNet-1 K and directly trained on the small subset of the DG-BEC-Shanghai dataset. Namely, the same training data and weights from ImageNet-1 K is used, and pretrained weights from the DG-BEC-Vegas training set are the only difference between the left and right part in Tables V and VI.

By comparing the tuning and no-tuning settings in Tables V and VI, it can be seen that pretrained weights from the DG-BEC-Vegas training set are able to improve BE accuracy for all four models, as there is an increase for all four metrics. It can also be seen that there is a clear difference among different models. For FarSeg, the benefit from pretrained weights of the DG-BEC-Vegas training set depends on the available data size for tuning. Specifically, when there are more data available, tuning is not helpful anymore, as there is no performance improvements when tuning on 30% of DG-BEC-Shanghai dataset for FarSeg in Table VI. By contrast, the increase of MAP-Net is much larger when tuning on 30% of DG-BEC-Shanghai dataset in Table VI, compared to that in Table V. SegFormer achieves slightly and consistently better accuracy by relying on pretrained weights of the DG-BEC-Vegas training set. Lastly, the improvements for Swin-M-S are remarkable in both settings. The IoU metric increases from 38.97% to 65.57% and from 50.40% to 69.05% when tuning on 10% and 30% of DG-BEC-Shanghai dataset, respectively.

From Tables V and VI, it can also be seen that SegFormer is able to provide the best BE results in almost all cases, and the second best model is FarSeg, followed by Swin-M-S and MAP-Net.

When more data from the target area is used, 30% in Table VI in contrast to 10% Table V, an increase is obtained for all models in both settings, i.e., in eight cases. It can also be seen that MAP-Net improves more as the increase of the training data in the tuning setting, while in the no-tuning case, FarSeg and Swin-M-S improve more from Tables V and VI.

3) Improved Model Cross-Area Performance: We applied the successful solutions on another test area, the DG-BEC-Paris dataset, which achieved BE results in Table VII. As described in Section III, SegFormer is always initialized with pretrained weights from ImageNet-1 K. Our finding is consistent with precious experiments in Section IV-A2. Specifically, transferred SegFormer provides the highest cross-area accuracy. Directly applying pretrained models from DG-BEC-Vegas cannot achieve reasonable results, while pretrained weights are able to help in most cases when using a small set of the DG-BEC-Paris dataset is available, and the performance is influenced by the available training data size, and the choice of models. For instance, SegFormer tuned with 10% of the data is able to provide almost the same results with FarSeg tuned with 30% of the data.

B. Qualitative Assessment of Building Extraction Results

Qualitative comparisons were also carried out to complement the quantitative results and provide more insights into the characteristics of different models and different settings. Specifically, BE results of five different models were compared using representative samples of varying shapes and sizes from the DG-BEC-Vegas and WHU testing sets, as listed in Figs. 3 and 4, respectively.

Overall, these qualitative comparisons are consistent with what have been found in Tables II and III. It can be seen that there are no big differences among the predictions from the two FCN-based and three transformer-based models, and BE results from satellite images are worse than those from aerial images. There are minor mistakes in the GT of both datasets, as can be seen from the second columns in Figs. 3 and 4. In such case, most models can actually predict correct results, as shown by the false negative and false positive in the second columns of Figs. 3 and 4, respectively. Label noise causes difficulties in model training, as well as result interpretation, and indicates that there is little need to pursue higher and higher accuracy on one certain dataset.

Additionally, Fig. 5 compares BE results of two different settings for two transformer-based models, corresponding to Table VI. It can be seen that the cross-area test setting leads to visually worse BE predictions, with more false positives and false negatives, compared to those in Fig. 3. The same as shown in Table VI, pretrained weights can help to improve BE results, and SegFormer performs better than Swin-M-S. Still, it is a difficult dataset to accurately predict all the buildings of varying shapes with complex backgrounds. The same as in Figs. 3 and 4, there are minor mistakes in the GT. There is an additional building annotation in the last column in Fig. 5, which is correctly left out in three of the four predicted results. False positives tend to occur in impervious areas surrounding buildings. False negatives tend to occur in the boundaries and could be possibly improved by postprocessing methods.

V. DISCUSSION

In this section, we provide some empirical evidence of the experimental setup and model design, as well as addressing the problems of model generalization, based on insights gained from the experimental results presented in Section IV.

A. On the Advantage of Transformer-Based Models

Both Tables II and III show that there is a slight advantage in transformer-based models under the within-one-dataset setting. Jointly considering that the models are all among SOTA and are of different sizes (for Swin-M-T, Swin-M-S, and Swin-M-B in Table II) and that our achieved accuracy is similarly high as in

TABLE V
BE ACCURACY OF DIFFERENT MODELS TRAINED ON 10% OF DG-BEC-SHANGHAI DATASET AND TESTED ON DG-BEC-SHANGHAI TEST SET UNDER TWO SETTINGS

Model	tuning				no tuning			
	IoU	Precision	Recall	F1	IoU	Precision	Recall	F1
FarSeg	0.6333	0.8109	0.7430	0.7755	0.6122	0.7522	0.7670	0.7595
MAP-Net	0.4383	0.7918	0.4954	0.6095	0.4262	0.5569	0.6448	0.5976
Swin-M-S	0.6557	0.8004	0.7839	0.7920	0.3897	0.6540	0.4910	0.5609
SegFormer	**0.7086**	**0.8374**	**0.8218**	**0.8295**	**0.6953**	**0.8389**	**0.8024**	**0.8203**

When tuning, models are pretrained on DG-BEC-Vegas training set.
The best results are indicated in bold for each metric.

TABLE VI
BE ACCURACY OF DIFFERENT MODELS TRAINED ON 30% OF DG-BEC-SHANGHAI DATASET AND TESTED ON DG-BEC-SHANGHAI TEST SET UNDER TWO SETTINGS

Model	tuning				no tuning			
	IoU	Precision	Recall	F1	IoU	Precision	Recall	F1
FarSeg	0.7234	0.8621	0.8181	0.8395	**0.7254**	**0.8669**	0.8163	**0.8408**
MAP-Net	0.6474	0.8285	0.7475	0.7859	0.4836	0.5558	0.7882	0.6519
Swin-M-S	0.6905	0.7790	**0.8586**	0.8169	0.5040	0.6030	0.7544	0.6702
SegFormer	**0.7451**	**0.8624**	0.8457	**0.8539**	0.7210	0.8564	**0.8201**	0.8379

When tuning, models are pretrained on DG-BEC-Vegas training set.
The best results are indicated in bold for each metric.

TABLE VII
BE ACCURACY OF TWO DIFFERENT MODELS TRAINED ON 0%, 10%, AND 30% OF DG-BEC-PARIS DATASET AND TESTED ON DG-BEC-PARIS TEST SET UNDER TWO SETTINGS

Model	Pre-trained weights	Data	IoU	Precision	Recall	F1
FarSeg	DG-BEC-Vegas	0	0.3750	0.6763	0.4570	0.5454
	ImageNet-1K	10%	0.6227	0.7803	0.7551	0.7675
	DG-BEC-Vegas		0.6425	0.7719	0.7931	0.7824
	ImageNet-1K	30%	0.6792	0.8135	0.8045	0.8090
	DG-BEC-Vegas		0.6984	0.8185	0.8264	0.8224
SegFormer	DG-BEC-Vegas	0	0.3630	0.7457	0.4142	0.5326
	ImageNet-1K	10%	0.6725	0.8195	0.7894	0.8042
	DG-BEC-Vegas		0.7015	0.8343	0.8151	0.8246
	ImageNet-1K	30%	0.7122	0.8385	0.8254	0.8319
	DG-BEC-Vegas		0.7113	0.8539	0.8099	0.8313

literature [27], we can assume that the highest possible accuracy has been achieved, or we hit a bottleneck in such a testing scenario. There is probably little improvement space left after the considerable effort researchers have devoted to varying directions, such as designing loss functions and novel architectures.

The advantage of transformer-based models are still not very clear when training on smaller datasets without using weights from the DG-BEC-Vegas training set, as can be seen from the right parts (no tuning) of Tables V and VI. When only training on 10% of DG-BEC-Shanghai dataset, SegFormer provides the best accuracy, followed by FarSeg. Swin-M-S is not better than MAP-Net, providing the worst results among all four models. Swin-M-S does not provide high BE accuracy probably because it requires more data to train, compared to FarSeg and SegFormer. When increasing the size of training data, from 10% in Table V to 30% in Table VI, all four models are able to provide better results. Swin-M-S is better than MAP-Net but worse than SegFormer which is slightly worse than FarSeg. Furthermore, FarSeg and Swin-M-S improve much when using more training data. These findings show that requirements on training data size are different for different models. Transformer-based models can be worse than CNN models when not training on enough data, as can be seen from results in Tables II and III.

The advantage of transformer-based models gets clear when testing models in a cross-area setting, as can be seen from the left parts (tuning) of Tables V and VI. In both tables, Swin-M-S is clearly better than MAP-Net, and SegFormer proves to be the best model, which can achieve high accuracy even using a small training dataset. Still, FarSeg provides higher accuracy compared to Swin-M-S and MAP-Net. In summary, the outperformance of transformer-based models depends on many factors, including the test settings, the available dataset size for training and tuning, as well as the specific architecture.

B. *Cross-Area Model Generalization Ability*

In the setting of cross-area test, high accuracy can be achieved by tuning pretrained models on a small set of the testing area.

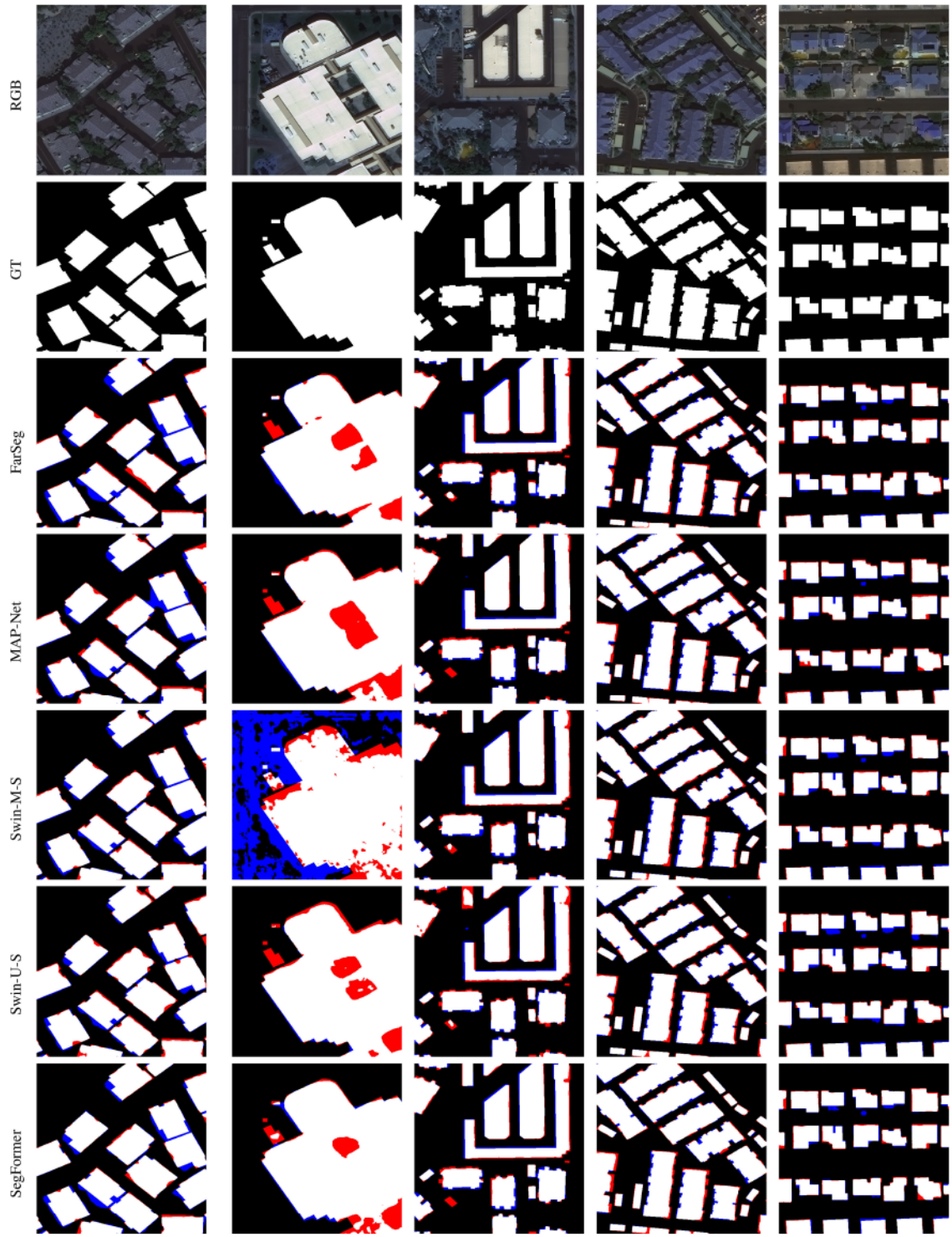


Fig. 3. Visual comparison of different models tested on DG-BEC-Vegas testing set. False negative is marked in red, and false positive is marked in blue.

This is true for all the four models. Even tuning on 10% of the DG-BEC-Shanghai dataset brings a distinct improvement, as can be seen by comparing Tables IV and V. No models can achieve reasonable BE results in Table IV, and the results are much worse, compared to directly initialized with ImageNet-1 K weights and using 10% of the whole dataset for training, indicating a big shift between data from the source area and the testing area. While this fine-tuning idea is a simple solution in a transfer learning setting, our study experimentally finds that the obtainable results depend both on the choice of the models and the size of the available dataset for tuning. For instance, when 30% of the whole dataset is available in Table VI, the fine-tuning result by FarSeg do not improve much and is similar to that by SegFormer without tuning.

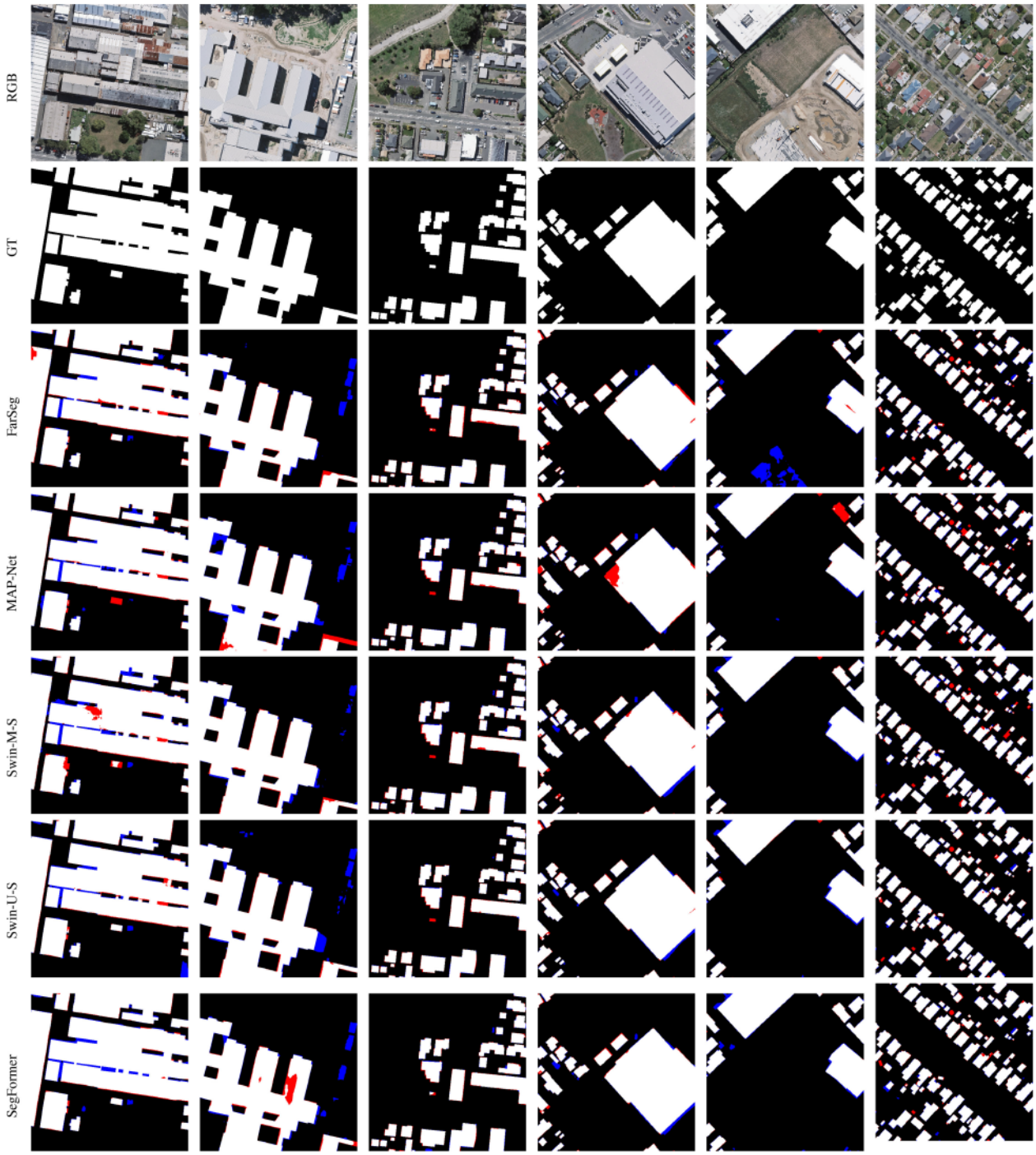


Fig. 4. Visual comparison of different models on the WHU testing set. False negative is marked in red, and false positive is marked in blue.

A general explanation of the improvements gained from pretrained weights is that a model learns more representative features and generalizes better by seeing more samples from both the pretraining and the testing areas, as it is usually true that the available samples are not enough to exploit the model's full capacity. When there are enough samples for a model, seeing samples from a different area (pretraining area) might even lead to worse results, as the case for FarSeg in Table VI.

Cross-area test, in general, belongs to a transfer learning problem, and there are many research directions and reference outside the scope of BE [36]. For instance, research on RS image scene classification suggested that optimization methods also play a role in the transfer learning results [58], and designing additional modules can help to align the source and target domains in the feature subspace [59]. Also, it has been shown that active learning is able to efficiently adapt a classifier trained on a source Sentinel-2 image to spatially distinct Sentinel-2 images for mapping poplar plantations [35]. One step further is domain adaptation without target labels [60]. All these approaches can help improve model generalization ability and facilitate automatic and applicable BE from RS images. Future work will consider further improvements by relying on specific transfer

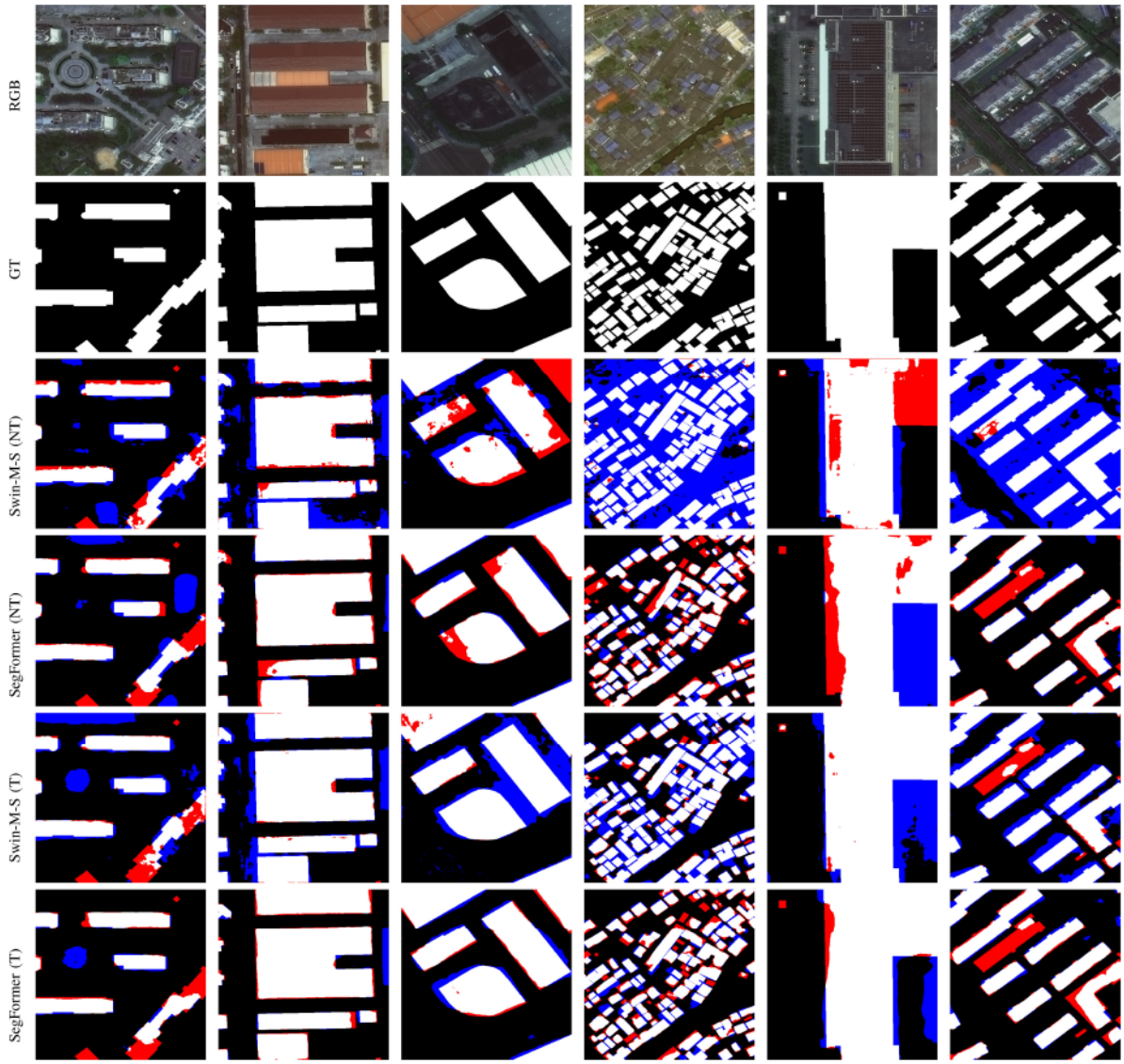


Fig. 5. Visual comparison of transformer-based models trained on 30% of DG-BEC-Shanghai dataset and tested on DG-BEC-Shanghai test set, in the case of not using (no-tuning, NT) and using (tuning, T) pretrained weights from the DG-BEC-Vegas training set. False negative is marked in red, and false positive is marked in blue.

strategies such as learning meaningful representations from a large amount of unlabeled data.

C. Cross-Data Model Generalization Ability

A close setting to cross-areas test is cross-data test, which has been studied in literature, e.g., analyzing transfer learning capabilities of FCNs for slum mapping among QuickBird, Sentinel-2, and TerraSAR-X data [61]. It has been found that trained models with the WHU aerial dataset fail to generalize to the WHU satellite data, even after data augmentation and relative radiometric correction [62]. Following the proposed tuning idea in this work, we carried out an experiment to test the cross-data generalization ability of the advantageous SegFormer. We applied SegFormer pretrained on DG-BEC-Vegas training set on the WHU aerial test set with and without finetuning on 5% of the dataset. Interestingly, the IoU metric can be increased from 81% to 83% by relying on the pretrained weights from the DG-BEC-Vegas training set, even though the testing and pretraining datasets are from different platforms.

VI. Conclusion

Extracting buildings from RS images is an important yet challenging task, attracting an increased attention these days. A wide variety of CNN-based FCN-like semantic segmentation models have been proposed and tested on benchmark datasets, and the current popular transformer-based methods have also been tested, demonstrating impressive yet convergent performance on commonly used open datasets. However, there are rare studies focusing on the cross-area model generalization ability, which is our goal in this study.

As demonstrated in this study, the best cross-area BE results can be achieved by the proposed approach, i.e., fine-tuning transformer-based model, SegFormer, with a small dataset from

the targeting area. Our main conclusions and findings can be summarized as follows.

1) We adapted two transformer-based models, Swin-M and SegFormer, for BE from RS images, which are able to achieve SOTA BE results on two open datasets when following the same setup as in literature. However, the trained models can hardly generalize to a different area, the same as previous approach.
2) Transferring pretrained SegFormer is able to achieve the best cross-area BE results, demonstrating a higher model capacity compared to both Swin-based transformers and FCN-based models. While it is a straightforward strategy to transfer pretrained weights and fine-tune on a small subset of the testing area, it does not always help a DL model generalize better. Results depend both on the choice of the models and the size of the available subset for training. An unsuitable model fails to generalize well even using this transferring learning idea.
3) The advantage of transformer-based models is influenced by the test setting and the training data size. When tested within one area, and there are many training samples, transformer-based and FCN-based models achieve similarly high BE results. And when training samples are fewer, a Swin-based model can perform much worse than a traditional FCN-based model. But still, a sophisticated transformer, SegFormer in our study, can perform consistently well even when fewer training samples are available.

Our investigation shows the potential of transformer-based models in an application-oriented and cross-area testing scenario, and future work includes preparing suitable datasets for cross-testing and developing supervised and unsupervised pretraining approaches to further improve the BE model generalization ability from RS images.

REFERENCES

[1] N. Khanal, K. Uddin, M. A. Matin, and K. Tenneson, "Automatic detection of spatiotemporal urban expansion patterns by fusing OSM and landsat data in Kathmandu," *Remote Sens.*, vol. 11, no. 19, 2019, Art. no. 2296.

[2] J. Kang *et al.*, "Disoptnet: Distilling semantic knowledge from optical images for weather-independent building segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4706315.

[3] L. Luo, P. Li, and X. Yan, "Deep learning-based building extraction from remote sensing images: A. comprehensive review," *Energies*, vol. 14, no. 23, 2021, Art. no. 7982.

[4] T. Hoeser and C. Kuenzer, "Object detection and image segmentation with deep learning on earth observation data: A review—Part I: Evolution and recent trends," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1667.

[5] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, Nov. 2018.

[6] W. Kang, Y. Xiang, F. Wang, and H. You, "EU-Net: An efficient fully convolutional network for building extraction from optical remote sensing images," *Remote Sens.*, vol. 11, no. 23, 2019, Art. no. 2813.

[7] Y. Zhang, W. Li, W. Gong, Z. Wang, and J. Sun, "An improved boundary-aware perceptual loss for building extraction from VHR images," *Remote Sens.*, vol. 12, no. 7, 2020, Art. no. 1195.

[8] C. Liao *et al.*, "Joint learning of contour and structure for boundary-preserved building extraction," *Remote Sens.*, vol. 13, 2021, Art. no. 1049.

[9] K. Zhao, J. Kang, J. Jung, and G. Sohn, "Building extraction from satellite images using mask R-CNN with building boundary regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, IEEE Computer Society, Los Alamitos, CA, USA, 2018, pp. 247–251.

[10] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020.

[11] J. Lin, W. Jing, H. Song, and G. Chen, "ESFNet: Efficient network for building extraction from high-resolution aerial images," *IEEE Access*, vol. 7, pp. 54285–54294, 2019.

[12] Q. Li, Y. Shi, X. Huang, and X. X. Zhu, "Building footprint generation by integrating convolution neural network with feature pairwise conditional random field," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7502–7519, Nov. 2020.

[13] J. Ma, L. Wu, X. Tang, F. Liu, X. Zhang, and L. Jiao, "Building extraction of aerial images by a global and multi-scale encoder-decoder network," *Remote Sens.*, vol. 12, no. 15, 2020, Art. no. 2350.

[14] W. Deng, Q. Shi, and J. Li, "Attention-gate-based encoder–decoder network for automatical building extraction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2611–2620, Feb. 2021.

[15] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[16] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.

[17] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.

[18] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Álvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," 2021, *arXiv:abs/2105.15203*.

[19] X. Zhao, J. Guo, Y. Zhang, and Y. Wu, "Memory-augmented transformer for remote sensing image semantic segmentation," *Remote Sens.*, vol. 13, no. 22, pp. 1–20, 2021.

[20] L. Ding *et al.*, "Looking outside the window: Wider-context transformer for the semantic segmentation of high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4410313.

[21] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, "Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images," *Remote Sens.*, vol. 13, no. 16, pp. 1–20, 2021.

[22] L. Wang, S. Fang, C. Zhang, R. Li, and C. Duan, "Efficient hybrid transformer: Learning global-local context for urban scene segmentation," 2021, *arXiv:abs/2109.08937*.

[23] L. Gao *et al.*, "STransFuse: Fusing Swin transformer and convolutional neural network for remote sensing image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, no. 8, pp. 10990–11003, Jul. 2021. [Online]. Available: https://www.techrxiv.org/articles/preprint/STransFuse_Fusing_Swin_Transformer_and_Convolutional_Neural_Network_for_Remote_Sensing_Image_Semantic_Segmentation/14866185/1

[24] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jan. 2022, Art. no. 6506105.

[25] Z. Xu, W. Zhang, T. Zhang, Z. Yang, and J. Li, "Efficient transformer for remote sensing image segmentation," *Remote Sens.*, vol. 13, no. 18, pp. 1–24, 2021.

[26] X. Xu *et al.*, "An improved Swin transformer-based model for remote sensing object detection and instance segmentation," *Remote Sens.*, vol. 13, no. 23, pp. 1–19, 2021.

[27] Q. Zhu, L. Cheng, H. Hu, M. Xiaoming, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2020.

[28] Z. Chen, D. Li, W. Fan, H. Guan, C. Wang, and J. Li, "Self-attention in reconstruction bias U-net for semantic segmentation of building rooftops in optical remote sensing images," *Remote Sens.*, vol. 13, no. 13, pp. 1–27, 2021.

[29] S. Wang, X. Hou, and X. Zhao, "Automatic building extraction from high-resolution aerial imagery via fully convolutional encoder-decoder network with non-local block," *IEEE Access*, vol. 8, pp. 7313–7322, 2020.

[30] W. Yuan and W. Xu, "MSST-Net: A. multi-scale adaptive network for building extraction from remote sensing images based on Swin transformer," *Remote Sens.*, vol. 13, no. 23, pp. 1–14, 2021.

[31] W. Liu *et al.*, "Building footprint extraction from unmanned aerial vehicle images via PRU-net: Application to change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, Jan. 2021, Art. no. 20322603.

[32] H. Guo, B. Du, L. Zhang, and X. Su, "A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 240–252, 2022.

[33] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 41–57, Jun. 2016.

[34] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *Adv. Data Sci. Inf. Eng.*, pp. 877–894, 2021.

[35] Y. Hamrouni, E. Paillassa, V. Chéret, C. Monteil, and D. Sheeren, "From local to global: A transfer learning-based approach for mapping poplar plantations at national scale using Sentinel-2," *ISPRS J. Photogramm. Remote Sens.*, vol. 171, pp. 76–100, 2021.

[36] M. Russwurm, S. Wang, M. Korner, and D. Lobell, "Meta-learning for few-shot land cover classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 788–796.

[37] B. Yuan, D. Zhao, S. Shao, Z. Yuan, and C. Wang, "Birds of a feather flock together: Category-divergence guidance for domain adaptive segmentation," *IEEE Trans. Image Process.*, vol. 31, Mar. 2022, Art. no. 35358045.

[38] I. Demir *et al.*, "DeepGlobe 2018: A challenge to parse the earth through satellite images," in *Proc. 2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 172–181.

[39] D. Zhou *et al.*, "Robust building extraction for high spatial resolution remote sensing images with self-attention network," *Sensors*, vol. 20, no. 24, pp. 1–19, 2020.

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.— Assist. Interv.*, 2015, pp. 234–241.

[41] S. M. Kamrul Hasan and C. A. Linte, "U-NetPlus: A modified encoder-decoder U-net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images," in *Proc. 2019 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2019, pp. 7205–7211.

[42] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[43] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1520–1528.

[44] K. Sun *et al.*, "High-resolution representations for labeling pixels and regions," 2019, *arXiv:1904.04514*.

[45] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5168–5177.

[46] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[47] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[48] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[49] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. 2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4095–4104.

[50] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2018.

[51] V. Mnih, "Machine learning for aerial image labeling," Ph.D. thesis, Univ. Toronto, Toronto, ON, Canada, 2013.

[52] F. Rottensteiner *et al.*, "The ISPRS benchmark on urban object classification and 3D building reconstruction," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. I3, pp. 293–298, 2012.

[53] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 3226–3229.

[54] Q. Chen, L. Wang, Y. Wu, G. Wu, Z. Guo, and S. L. Waslander, "Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 42–55, 2019.

[55] Y. Xie *et al.*, "Refined extraction of building outlines from high-resolution remote sensing imagery based on a multifeature convolutional neural network and morphological filtering," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, Apr. 2020, Art. no. 9082895.

[56] H. Goldberg, M. Brown, and S. Wang, "A benchmark for building footprint classification using orthorectified RGB imagery and digital surface models from commercial satellites," in *Proc. IEEE Appl. Imagery Pattern Recognit. Workshop*, 2017, pp. 1–7.

[57] W. Li, W. Zhao, H. Zhong, C. He, and D. Lin, "Joint semantic-geometric learning for polygonal building segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1–8.

[58] R. Pires de Lima and K. Marfurt, "Convolutional neural network for remote-sensing scene classification: Transfer learning analysis," *Remote Sens.*, vol. 12, no. 1, pp. 1–20, 2020.

[59] S. Song, H. Yu, Z. Miao, Q. Zhang, Y. Lin, and S. Wang, "Domain adaptation for convolutional neural networks-based remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1324–1328, Aug. 2019.

[60] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 766–785, Mar. 2021.

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

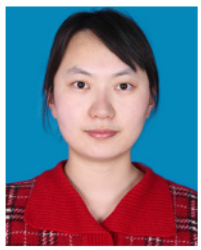
[62] S. Ji, S. Wei, and M. Lu, "A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery," *Int. J. Remote Sens.*, vol. 40, no. 9, pp. 3308–3322, 2019.

Chunping Qiu received the bachelor's and the master's degrees in photogrammetry and remote sensing from the Zhengzhou Institute of Surveying and Mapping, Zhengzhou, China, in 2013 and 2016, respectively, and the Dr.-Ing. degree in signal processing in earth observation from the Technische Universität München (TUM), Munich, Germany, in 2020.

Since 2021, she has been a Researcher with the PLA Strategic Support Force Information Engineering University, Zhengzhou, China. She is currently a Postdoctoral Researcher with the Artificial Intelligence Research Center, National Innovation Institute of Defense Technology, Academy of Military Science, Beijing, China. In 2019, she was a Guest Researcher with the Telecommunications and Remote Sensing Lab, University of Pavia, Pavia, Italy. Her main research interests are deep learning and self-supervised learning for remote sensing applications, and Big Earth Data management.

He Li received the bachelor's, master's, and Dr.-Ing. degrees in photogrammetry and remote sensing from the Zhengzhou Institute of Surveying and Mapping, Zhengzhou, China, in 2006, 2009, and 2012, respectively.

He is currently a Lecturer with the Information Engineering University, Zhengzhou, China. His research interests include geostationary Earth observation, polarimetric SAR image processing, and geographic spatiotemporal data management and analysis.



Wenyue Guo received the bachelor's and master's degrees in cartography and geographic information engineering and the Ph.D. degree in surveying and mapping from the PLA Strategic Support Force Information Engineering University, Zhengzhou, China, in 2012, 2015, and 2018, respectively.

She is currently with the PLA Strategic Support Force Information Engineering University as an Associate Professor. Her research interests include geographic information science and graph representation.

Xin Chen received the bachelor's degree in cartography and geographic information engineering in 2020 from the PLA Strategic Support Force Information Engineering University, Zhengzhou, China, where she is currently working toward the master's degree in cartography and geographic information engineering.

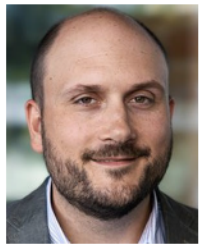
Her main research interests include spatial data processing, fusing, and updating.

Anzhu Yu received the bachelor's degree in remote sensing science and technology and the master's degree in photogrammetry and remote sensing from the PLA Strategic Support Force Information Engineering University, Zhengzhou, China, in 2011 and 2014, respectively, and the Ph.D. degree in photogrammetry and remote sensing from the Institute of Surveying and Mapping, PLA Strategic Support Force Information Engineering University, in 2017.

He is currently with the PLA Strategic Support Force Information Engineering University as an Associate Professor. His research interest includes signal processing in Earth observation.

Xiaochong Tong received the bachelor's, M.Sc., and Dr.-Ing. degrees in photogrammetry and remote sensing from the Zhengzhou Institute of Surveying and Mapping, Zhengzhou, China, in 2003, 2006, and 2010, respectively.

He took up a postdoctoral position at Beijing Normal University, Beijing, China, from 2011 to 2016. He is currently a Full Professor with Information Engineering University, Zhengzhou, China. His research interests include discrete global grid systems, remote sensing, geographic information systems, and photogrammetry.



Michael Schmitt (Senior Member, IEEE) received the Dipl.-Ing. (Univ.) degree in geodesy and geoinformation, the Dr.-Ing. degree in remote sensing, and the habilitation in data fusion from the Technical University of Munich (TUM), Munich, Germany, in 2009, 2014, and 2018, respectively.

Since 2021, he has been a Full Professor for Earth observation with the Department of Aerospace Engineering, University of the Bundeswehr Munich, Neubiberg, Germany. Before that, he was a Professor for applied geodesy and remote sensing with the Department of Geoinformatics, Munich University of Applied Sciences, Munich, Germany. From 2015 to 2020, he was a Senior Researcher and Deputy Head at the Professorship for Signal Processing in Earth Observation with TUM. In 2019, he was additionally appointed as Adjunct Teaching Professor with the Department of Aerospace and Geodesy, TUM. In 2016, he was a guest Scientist with the University of Massachusetts, Amherst, MA, USA. His research focuses on technical aspects of Earth observation, in particular, image analysis and machine learning applied to the extraction of information from multimodal remote sensing observations.

Dr. Schmitt is a Co-Chair of the Working Group "SAR and Microwave Sensing" of the International Society for Photogrammetry and Remote Sensing, and also of the Working Group "Benchmarking" of the IEEE-GRSS Image Analysis and Data Fusion Technical Committee. He frequently serves as a reviewer for a number of renowned international journals and conferences and is the recipient of several best reviewer awards. He is an Associate Editor for *IEEE Geoscience and Remote Sensing Letters*.