Abstract—Semantic segmentation necessitates approaches that learn high-level characteristics while dealing with enormous quantities of data. Convolutional neural networks (CNNs) can learn unique and adaptive features to achieve this aim. However, due to the large size and high spatial resolution of remote sensing images, these networks cannot efficiently analyze an entire scene. Recently, deep transformers have proven their capability to record global interactions between different objects in the image. In this article, we propose a new segmentation model that combines CNNs with transformers and show that this mixture of local and global feature extraction techniques provides significant advantages in remote sensing segmentation. In addition, the proposed model includes two fusion layers that are designed to efficiently represent multimodal inputs and outputs of the network. The input fusion layer extracts feature maps summarizing the relationship between image content and elevation maps [digital surface model (DSM)]. The output fusion layer uses a novel multitask segmentation strategy where class labels are identified using class-specific feature extraction layers and loss functions. Finally, a fast-marching method (FMM) is used to convert unidentifiable class labels into their closest known neighbors. Our results demonstrate that the proposed method improves segmentation accuracy compared with state-of-the-art techniques.

Index Terms—Convolutional neural networks (CNNs), EfficientNet, fusion networks, semantic segmentation, transformers.

I. INTRODUCTION

In recent years, with the continual advancement of remote sensing technology, high-resolution remote sensing satellites have been remarkably used, and the resolution of remote sensing images has considerably improved [1]. As a result, understanding detailed, high-resolution remote sensing images has become a significant challenge [2]. Semantic image segmentation, also known as pixel-level categorization, is a critical computer vision challenge and is a vital technology for remote sensing image understanding [3]. In semantic segmentation, the goal is to assign a class label to each image pixel [4]. There are two types of image semantic segmentation methods: conventional and deep-learning-based [5]. Traditionally, machine learning approaches have used handmade features, whereas deep learning approaches show higher performance by simultaneously learning feature representation and classifier parameters [3].

Deep learning methods, particularly convolutional neural networks (CNNs), have been used successfully in solving multiple remote sensing problems. The main power of the CNN-based methods comes from the efficient capturing of local context information compared with traditional machine learning methods for segmentation, such as the support vector machine (SVM) [6], random forest, and conditional random field (CRF) [7]. However, in semantic segmentation, the per-pixel classification is frequently unclear if only local information is modeled, whereas global contextual information improves the semantic content of each pixel [8], [9].

Many attempts have been investigated to address the aforementioned issue, such as the self-attention mechanism [10], [11], or transformer-based models [12], [13]. A self-attention mechanism can be used for efficient feature fusion between multisource data, as well as multilevel feature fusion between high-level abstract characteristics and low-level spatial features. Furthermore, the transformer architecture is renowned for its use of attention to model long-term relationships in data. The transformer model was designed for sequence modeling and transduction activities. Its phenomenal performance in the language domain has prompted researchers to develop models that use it in computer vision applications, where it has recently exhibited promising results on a variety of tasks, including image classification [14] and joint vision-language modeling [15]. However, attention modules can limit local feature extraction and representation. In addition, despite the transformer’s benefits, its computational complexity is significantly higher than that of the CNN-based encoders because of its square-complexity self-attention mechanism [10]. This has a significant impact on the encoder’s potential and viability for urban-related real-time applications.

It has been noted in earlier works that hybridizing transformers and CNNs may overcome the limitations of a pure transformer segmentation network [16], [17]. For instance, in medical image segmentation TransFuse [16] processed CNNs and transformers in parallel and merged the two sets of extracted features. To compensate for CNN’s shortcomings in global modeling, ST-UNet [17] used the Swin transformer to support a UNet in medical image segmentation. Inspired by these developments, we propose a novel hybrid model to address the global–local and data fusion challenges in remote sensing semantic segmentation. Here, we propose an end-to-end fusion framework that combines U-Nets and transformers. Fig. 1 describes the architecture of the proposed model. The U-Net component models the dense connections

Manuscript received 3 March 2023; accepted 7 April 2023. Date of publication 18 April 2023; date of current version 2 May 2023. This work was supported by the Information and Communication Technology Fund (ICT), Telecommunications Regulatory Authority (TRA), Abu Dhabi, United Arab Emirates. (Corresponding author: Lyes Saad Saoud.)

Hasan AlMarzouqi is with the Electrical Engineering and Computer Science Department, Khalifa University, Abu Dhabi 127788, United Arab Emirates (e-mail: hasan.almarzouqi@ku.ac.ae).

Lyes Saad Saoud is with the Mechanical Engineering Department, Khalifa University, Abu Dhabi 127788, United Arab Emirates (e-mail: lyes.saoud@ku.ac.ae).

Digital Object Identifier 10.1109/TGRS.2023.3268159

1558-0644 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
between pixels, while the transformer-based component models the context using a token-based technique. The features extracted by the EfficientUNet model accentuate local relationships between pixel intensities, while transformer features emphasize global interactions. The developed segmentation system consists of four major parts: an input fusion layer, a transformer-based network, an EfficientUNet network, and an output fusion layer. In the input fusion layer, features representing the multimodal input data, i.e., image content, and elevation maps [digital surface model (DSM)], are extracted. DSM data offer substantial information that was shown to improve segmentation accuracy [18], [19], [20], [21]. The fused features are processed in parallel using the U-Net model and a transformer model. Finally, the resulting feature maps are passed through a novel multitask segmentation strategy that identifies class labels using class-specific feature extraction layers and loss functions. The use of these class-specific features and loss functions is shown to further improve the network performance.

The contributions of this article can be summarized as follows.

1) An efficient mixture model, EfficientUNetTransformer, is developed for the semantic segmentation of high-resolution remote sensing images. In this model, we combine transformers and the U-Net network to better represent global and local contexts, leading to more consistent labeling outcomes in complex urban constructions.

2) A novel multitask segmentation approach that identifies class labels using class-specific feature extraction layers and loss functions.

3) A novel input fusion layer that multiplies the normalized features extracted from two separate convolutional units and adds a residual component to the result.

4) Extensive experiments on two publicly available datasets demonstrate the performance of the proposed model. The proposed model yields higher accuracy than the purely convolutional equivalent and outperforms several recently proposed attention-based semantic segmentation algorithms.

The remainder of this article is organized as follows. Section II reviews some related work. Section III describes the proposed model in detail. Section IV presents the experimental results and discussions. Finally, Section V summarizes the conclusions.

II. RELATED WORK

A. CNN-Based Methods

The fully convolutional network (FCN) was the first successful CNN-based method investigated to solve end-to-end semantic segmentation problems [22]. Since then, the CNN-based models have shown their power over traditional models. For instance, a CNN-based model depending on a downsampling-then-upsampling architecture was used for the semantic labeling of subdecimeter resolution images [23]. An object-based classification technique was integrated with a deep learning model in [24] to improve remote sensing image classification accuracy. Zhao et al. [25] used CNNs to explore semantic segments, and a CRF was used to model the contextual information between them. A rotation equivariance CNN architecture was used for high-resolution land cover mapping in [26]. Bergado et al. [27] presented a single-stage approach that embeds the processing stages in a recurrent multiresolution convolutional network. A self-cascaded network was used to improve labeling coherence using a sequential global-to-local context aggregation method [28]. For the encoder–decoder models, Sun et al. [29] offered ensemble approaches and a residual architecture to reduce the detrimental impacts of structural stereotypes and address the issue of inadequate learning. A differentiable decision forest was used to segment remote sensing data semantically [3]. Zhong et al. [30] adapted the conventional FCN-8s/16s/32s models to extract roads and buildings from remote sensing RGB images. Audebert et al. [31] suggested and evaluated their FCN-based semantic segmentation approaches using infrared, red, and green (IRRG) and DSM as input data. A residual dense U-Net was proposed for pixelwise sea–land segmentation in complex and high-density remote sensing images [32]. Symmetrical dense-shortcut U-Net was used to segment high-resolution remote sensing images [33]. The DeepLab semantic segmentation model and object-based image analysis were used to segment high-resolution remote sensing images in [34].

B. CNNs With Self-Attention

Convolutional segmentation models rely on learnable convolutions that extract semantically significant features. Unfortunately, the local scope of convolutional filters restricts access to the relationships between distant image pixel intensities.
Global features are particularly critical in remote sensing segmentation because image patch labeling is frequently dependent on the global context. To circumvent this limitation, Wang et al. [11] introduced nonlocal neural networks for video classification, which were motivated by the traditional nonlocal means filter (which can be viewed as a form of self-attention) in computer vision [35]. In comparison to 2-D and 3-D convolutional networks, nonlocal neural networks, which use a few nonlocal blocks, were shown to be highly effective in classifying videos. Furthermore, nonlocal neural networks are more computationally efficient than the 3-D convolutional equivalents. Remote sensing image segmentation networks can benefit from attention processes as well. An attention-fused network combining the high-level and low-level semantic information was used for remote sensing image segmentation in [36]. Lightweight spatial and channel attention modules were incorporated to optimize the semantic characteristics of high-resolution remotely sensed images [5]. A local attention block with an embedded module was developed to collect additional contextual information in [20]. Li et al. [37] proposed a linear attention mechanism to decrease the computational complexity and enhance the performance. However, the aforementioned attention modules remain biased toward local interactions because they rely extensively on convolutions [13], [38].

C. Transformer-Based Methods

Recently, transformer models have gained immense interest in solving computer vision tasks due to their enhanced performance levels [12]. The majority of available transformers for semantic segmentation adhere to the encoder–decoder structure [38]. The pure transformer structure is composed of a transformer-based encoder and a transformer-based decoder. Some notable models that belong to this category are the Segmenter [13], SegFormer [39], and SwinUNet [40]. A second class of transformers uses a hybrid structure that consists of a transformer-based encoder and a CNN-based decoder. For example, UNetFormer [38] uses a hybrid architecture that consists of a transformer-based decoder and an encoder that is based on CNNs. The TransUNet architecture used the hybrid vision transformer [14] as the encoder and obtained the state-of-the-art results in medical image segmentation [41]. For fine-resolution remote sensing image segmentation, Panboonyuen et al. [42] used the Swin transformer as the encoder and used a variety of CNN-based decoders, including UNet [43], Panoptic feature pyramid (PFP) networks [44], and pyramid scene parsing (PSP) network [45] as decoders. Wang et al. [46] proposed the pyramid vision transformer (PVT), which mimicked the properties of CNN backbones.

D. Using DSMs

Regular aerial photographs are often captured from a single perspective, leaving out valuable geographical location details [47]. Fortunately, high-resolution aerial images, such as DSM images, offer a wealth of geographic details, that are useful in describing challenging frontier areas such as roads/buildings and low vegetation/trees [31]. It has been observed that models that only use optical images frequently make errors in shadows and other dark areas found in remotely sensed images and that these errors can be corrected using DSM data [18]. Therefore, DSM images can greatly increase the performance of remote sensing segmentation systems as demonstrated in [18], [19], [20], and [21]. For example, Sun and Wang [18] used color images to compute initial label maps that are refined using ground surface information extracted from DSM data. Marmanis et al. [19] extracted boundaries between the different classes using color images and DSM data to counter the loss of detail that can be caused by the convolutional filters. Ding et al. [20] concatenated DSM data with RGB images.

III. PROPOSED MODEL

A flowchart of the proposed model is presented in Fig. 1. The developed segmentation system consists of four major parts: 1) a transformer-based network that can extract global high-level semantic characteristics from the input image; 2) an EfficientUNet network that focuses on the extraction of local features from the input image; 3) the input fusion network that merges the IRRG or the RGB image with its DSM image; and 4) the output fusion layer that splits the sum of local and global features into six separate binary subclasses by passing them through tokenizers and transformers. In what follows, we provide a detailed description of the different components of the system.

A. Input Fusion Layer

DSM contains single-channel data that describe the depth of the scene captured by the three channels in RGB or IRRG images. We propose using an input fusion network that uses learned convolutional parameters to extract the set of features that describe the complex relationship between DSM data and each one of the RGB/IRRG channels (Fig. 2). This fusion layer passes both IRRG/RGB and DSM images through a sequence of convolution, batch normalization (BN), and rectifier linear unit (ReLU) blocks to extract high-resolution features, and then a dot multiplication is applied between the two obtained feature maps followed by a BN layer. The dot multiplication can be viewed as a separator between features of objects at different elevation levels. Finally, an addition operation is applied between the resulting feature map and the original IRRG/RGB image. This addition provides a direct path for
RGB/IRRG data and acts as a residual connection that can help in avoiding the vanishing gradient problem.

B. Transformer Path

An EfficientNet B7 [48] deep neural network architecture is used to extract relevant features from image patches. First, we removed the last stage (i.e., the head stage) from the original EfficientNetB7, which originally contained nine stages. Next, the input features are tokenized and sent to a transformer encoder. The tokenizer (Fig. 1) takes the extracted features of each stage (MSA) and feedforward (FF) blocks (see Fig. 4). At each layer, the input to self-attention is a triple (query Q, key K, and value V) computed from the input T\((l-1)\). We apply the layer normalization immediately before the MSA/multilayer perceptron (MLP) units. The MSA unit can be described as follows:

\[ MSA(q, k, v) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \]  

where

\[ \text{head}_j = \text{Attention}(q W^q_j, k W^k_j, v W^v_j) \]

where \( q W^q_j, k W^k_j, \) and \( v W^v_j \) are the linear projection matrices, and \( h \) is the number of attention heads. The multi-head attention block receives three components the query \( Q \), the key \( K \), and the value \( V \) to compute the self-attention output

\[ \text{Attention}(Q, K, V) = \sigma \left( \frac{Q K^T}{\sqrt{d}} \right) V \]

where \( d \) represents the channel dimension of the three components, and \( \sigma \) is the softmax function applied to the channel dimension. Finally, the transformer output is upsampled to match the dimension of features extracted from the EfficientUNet network. The transformer decoder comprises \( L_D = 6 \) layers of multi-head cross attention (MCA) and...
FF blocks. The encoder’s patch-level encodings are mapped to patch-level class scores by the decoder. The decoder and encoder configurations are equivalent. The MCA receives the query from the extracted features $X$, the key, and the value from the tokens $T$ generated by the transformer encoder (Fig. 4). It should be noted that we followed the vision transformer architecture (ViT) [14] and use prenorm residual units, which means that layer normalization takes place before the MSA/MLP components, in contrast to the original transformer, which uses the postnorm residual unit. Prenorm has been demonstrated to be more capable and stable than the alternative [49], [50].

C. EfficientUNet

The second part of the proposed model is the U-Net segmentation model that uses an architecture from the EfficientNet family of networks as a backbone. EfficientNet image classification models apply a compound-scaling approach that consistently adjusts the network depth, width, and resolution for increased performance using a given set of scaling parameters [48]. Scaling the network incrementally increases the model performance by balancing the architecture’s breadth, depth, and image resolution compound coefficients. EfficientNet is built using mobile inverted bottleneck convolution (MBConv), as shown in Fig. 5(b). The proposed model uses the swish activation function [51] instead of the widely used ReLUs. When going from EfficientNetB0 to EfficientNetB7, the depth, width, resolution, and model size increase, while the accuracy improves [48]. The architecture used in the proposed model is the EfficientNetB7 which has 55 basic building MBConv blocks as shown in Fig. 5(a). The components used in these blocks are shown in Fig. 5(b). The proposed U-Net encoder is based on the EfficientNetB7 model. The decoder is constructed using a reversed version of the encoder model with upsampling units. The encoder outputs of layers 3, 10, 17, and 27 are concatenated with their corresponding decoder outputs, as shown in Fig. 1. Transposed convolution layers are used to build the decoder, which doubles the size of a feature map while decreasing the number of channels by half. An upsampling layer followed by the double convolution layers are applied after each concatenation operation. Double convolution layers apply the sequence of convolution, BN, and ReLU operations twice.

D. Output Fusion Layer

This module is composed of a CNN that receives as input the sum of the final feature mappings from the two deep networks: the transformer and EfficientUNet. Input feature maps were summed and fed into a shallow CNN, which is a double convolution layer with 32 input channels and six output channels. The CNN’s output was sent to six separate tokenizers plus transformer encoders, representing the six binary classes, and then passed through logarithmic softmax functions (Fig. 6).

Next, the logistic loss is used as the loss function and was computed using a logarithmic softmax layer and averaged across the entire patch [31]

$$\text{loss} = \frac{1}{N} \sum_{i=1}^{N} y^i \log \left( \frac{\exp(y_i^i)}{\sum_{j=1}^{N} \exp(y_j^i)} \right) \quad (5)$$

where $N$ is the number of pixels in the input image, $k$ equals two classes, and for each pixel $i$, $y_i$ and $\hat{y}_i$ are the true and predicted labels, respectively. The different parts of the networks were trained together to find the best model that can predict the right labels without using special pre- or postprocessing operations.

The semantic segmentation maps are obtained by combining the six binary classes using the Add&Inpaint layer (Fig. 6). This layer combines all the class labels computed in the output fusion layer in the output segmentation map and then replaces unclassified pixels by their nearest classified neighbor using a fast-marching method (FMM) [52]. Using this method, color information is propagated inward from the region boundaries, i.e., known image information is used to fill the empty areas. The propagation method is based on moving an image smoothness operator down the gradient. The image smoothness operator is calculated by taking a weighted average across a neighborhood of the pixel to inpaint [52].

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Datasets

The proposed model was evaluated on two challenging public datasets for semantic labeling.
TABLE I

| Input fusion | Tokenizer | Output fusion | F1 score | IoU | F1 score | IoU | F1 score | IoU | F1 score | IoU | MiIoU | Kappa | Total accuracy |
|--------------|-----------|---------------|----------|-----|----------|-----|----------|-----|----------|-----|--------|--------|---------------|
| No           | No        | No            | 90.27    | 82.28| 94.84    | 90.19| 77.04    | 82.65| 90.19    | 82.13| 80.70  | 67.65  | 85.72         |
| No           | No        | Yes           | 91.69    | 84.65| 96.35    | 92.95| 79.55    | 66.04| 90.24    | 82.32| 86.03  | 75.49  | 88.99         |
| No           | Yes       | No            | 90.21    | 82.16| 94.79    | 90.10| 77.47    | 63.22| 90.01    | 81.84| 85.29  | 74.35  | 85.70         |
| No           | Yes       | Yes           | 87.60    | 77.94| 89.70    | 81.33| 74.99    | 63.25| 60.79    | 79.72| 75.92  | 61.18  | 80.57         |
| Yes          | No        | No            | 83.08    | 71.03| 91.51    | 84.35| 69.96    | 53.80| 84.52    | 73.19| 84.90  | 73.76  | 78.32         |
| Yes          | No        | Yes           | 88.71    | 79.71| 95.38    | 91.16| 79.00    | 65.29| 90.74    | 83.11| 83.71  | 71.94  | 85.95         |
| Yes          | Yes       | No            | 91.90    | 85.01| 95.97    | 92.25| 76.79    | 62.32| 90.54    | 82.71| 82.05  | 69.57  | 85.15         |
| Yes          | Yes       | Yes           | 91.47    | 84.28| 96.38    | 93.02| 79.42    | 65.86| 90.90    | 83.31| 84.12  | 78.76  | 87.70         |

TABLE II

| Input fusion | Tokenizer | Output fusion | F1 score | IoU | F1 score | IoU | F1 score | IoU | F1 score | IoU | MiIoU | Kappa | Total accuracy |
|--------------|-----------|---------------|----------|-----|----------|-----|----------|-----|----------|-----|--------|--------|---------------|
| No           | No        | No            | 90.38    | 82.79| 94.83    | 90.16| 85.19    | 75.16| 87.22    | 77.33| 93.25  | 87.35  | 75.86         |
| No           | No        | Yes           | 91.60    | 84.50| 86.86    | 95.76| 97.83    | 76.92| 88.07    | 78.68| 92.42  | 85.91  | 78.15         |
| No           | Yes       | No            | 92.40    | 85.87| 95.66    | 91.32| 84.50    | 72.87| 85.52    | 76.70| 93.72  | 88.18  | 86.33         |
| No           | Yes       | Yes           | 92.31    | 85.72| 97.80    | 95.69| 87.06    | 77.09| 88.46    | 79.31| 93.35  | 87.52  | 77.93         |
| Yes          | No        | No            | 90.55    | 82.74| 92.25    | 83.62| 85.41    | 74.35| 87.08    | 77.12| 93.20  | 87.26  | 72.28         |
| Yes          | No        | Yes           | 92.86    | 86.67| 97.41    | 94.95| 85.30    | 75.91| 88.17    | 78.85| 95.35  | 91.12  | 78.81         |
| Yes          | Yes       | No            | 91.50    | 84.24| 93.41    | 87.64| 86.16    | 75.68| 88.98    | 78.69| 94.10  | 88.86  | 75.03         |
| Yes          | Yes       | Yes           | 92.86    | 84.67| 97.77    | 95.64| 87.30    | 77.46| 88.72    | 79.72| 93.43  | 87.70  | 86.55         |

1) ISPRS Vaihingen Challenge Dataset: This dataset served as the baseline for the ISPRS 2-D Semantic Labeling Challenge in Vaihingen [53]. It consists of three bands of IRRG image data and DSM, and normalized DSM (NDSM) data [54]. The Vaihingen dataset consists of 33 images of different sizes (the average size of the training dataset is 2494 × 2064). We split the dataset into three parts: 12 images were used for training, four images ("5," "21," "15," and "30") for validation (hyperparameter tuning), and 17 images for testing. It should be noted that all the models presented in Table V use the same testing dataset.

2) ISPRS Potsdam Challenge Dataset: The second dataset used in this study belongs to the Potsdam ISPRS 2-D Semantic Labeling Challenge [53]. It is made up of four-band IRRG and blue (IRRGB) image data and matching DSM and NDSM data. The Potsdam dataset includes 38 orthorectified aerial IRRGB images that have a resolution of 5 cm and a size of 6000 × 6000 pixels. We applied the same strategy used in the Vaihingen dataset here. We use 17 images for training and seven images ("3_11," "3_12," "4_11," "5_10," "6_9," "6_12," "7_11") for validation and 14 images for testing.

C. Input Fusion, Tokenizer, and Output Fusion Effects

We tested the proposed model using all possible combinations of the input fusion, tokenizer, and output fusion units to test the performance of the proposed system in different settings.

D. Implementation Details and Evaluation Metrics

The proposed system was implemented in PyTorch using an Nvidia RTX A6000 GPU. All our models were trained using stochastic gradient descent (SGD) with a base learning rate of 0.01, momentum of 0.9, weight decay of 0.0005, and batch size of 10. The encoder–decoder weights were randomly initialized. We divided the learning rate by 10 after 25 and 45 epochs (out of a total of 100 epochs used for training). We use the overall pixelwise accuracy (OA), average F1 score, intersection-over-union (IoU), and Cohen’s kappa coefficient κ as performance metrics. The F1 score, IoU, and kappa coefficient κ for class i are defined as follows:

\[
F_1 = \frac{2TP_i}{2TP_i + FP_i + FN_i}
\]  

\[
IoU = \frac{TP_i + FP_i + FN_i}{TP_i}
\]  

\[
\kappa = \frac{p_o - p_e}{1 - p_e}
\]

where \(TP_i\) is the number of true positives for class i, \(FP_i\) is the number of false positives for class i, \(FN_i\) is the number of false negatives for class i, \(p_o\) is the relative observed agreement among raters, and \(p_e\) is the hypothetical probability of chance agreement. In compliance with the competition organizers’ assessment guidelines, these metrics were derived after eroding the boundaries with a three-pixel radius circle and deleting those pixels [31]. Next, we present several experimental results to test the performance of the proposed system in different settings.
output fusion units significantly increases the accuracy of the extracted bounding boxes.

D. Selecting an EfficientNet Model

The original EfficientNet model has eight variants that differ in the number of parameters named B0 to B7. We demonstrate the relationship between model size and the performance of the proposed model in Tables III and IV. We can see that EfficientNet B7 produces the best results for all the metrics for most classes in the Vaihingen and Potsdam datasets. Therefore, we used the B7 model in the remaining experiments in the article.

E. Performance Comparison

We compared the performance of the proposed model against several recently proposed models using validation and testing datasets.

1) On Validation Data: For both the datasets, the proposed model was compared against the following seven deep learning models: fully connected networks (FCN-8s) [22], U-Net [55], SegNet [56], PSPNet [45], Segmenter [13], SegFormer [39], and deep transformers [50]. In addition, the proposed model was compared with the EfficientUNet model, a model that is similar to the proposed model but has only the EfficientUNet branch, and the Double EfficientUNet model, a model that uses two EfficientUNet branches to illustrate the importance of including the transformer architecture.

The results obtained for the Vaihingen challenge dataset are presented in Fig. 8 and Table V, and those for the Potsdam challenge dataset are given in Fig. 9 and Table VI. One can see in Tables V and VI that the UNet, FCN-8s, and PSPNet models achieve low-quality results compared with SegNet, Transformers, and the proposed model. For example, all the convolutional models (i.e., UNet, FCN-8s, PSPNet, and SegNet) show low performance in car segmentation; however, the transformer-based models (i.e., transformer alone or our fusion models) classify that category accurately (improvement by more than 3%). This result demonstrates the superior ability of transformers in representing dynamically changing object classes.
The results of the Potsdam dataset show that the transformer model alone improved the results by more than 2% in kappa and by around 1% in total accuracy compared with the SegNet model, which achieved the closest results. The fusion adds 1% to the overall accuracy. The transformer alone and fused with the EfficientUNet each improve the tree’s classification quality by approximately 4% compared with the conventional models, such as SegNet or UNet. To demonstrate the ability of the proposed model when dealing with unclear objects, we took a sample that contained such objects (Fig. 10). We can see that the proposed model predicts these objects well, compared with other models.

In addition, the results in Tables V and VI highlight the effects of combining EfficientUNets and transformers. For example, the MIoU value achieved using Transformers is 58.78% for the Vaihingen dataset and 62.88% for the Potsdam dataset. While the proposed hybrid model that combines EfficientUNets and Transformers achieves an MIoU value of 73.19% and 80.55% on the Vaihingen and Potsdam datasets, respectively. Similar improvements in performance can be observed between the proposed model and the EfficientUNet model. Sample results that illustrate how the combination of EfficientUNet and transformers provide an added advantage are presented in Fig. 11. We can see that the transformer can

---

**Fig. 8.** Performance comparison with other deep learning models using the Vaihingen validation dataset. The labels represent the impervious surface in white, buildings in blue, low vegetation in cyan, trees in green, cars in yellow, and clutter in red.

**TABLE V**

| Model         | FL score | IoU | FL score | IoU | FL score | IoU | FL score | IoU | FL score | IoU | FL score | IoU | MIoU | Kappa | Total accuracy | Number of Parameters | Training time (h) |
|---------------|----------|-----|----------|-----|----------|-----|----------|-----|----------|-----|----------|-----|------|-------|-----------------|---------------------|-------------------|
| FCN8s         | 82.14    | 69.69 | 86.94    | 76.90 | 64.29    | 47.39 | 85.31    | 74.38 | 57.95    | 40.80 | 56.22    | 75.10 | 81.56 | ~137 | 4.22            |
| SegNet        | 83.34    | 74.43 | 91.12    | 83.69 | 70.63    | 54.60 | 85.69    | 74.97 | 23.32    | 13.20 | 50.15    | 79.11 | 84.53 | ~48  | 43.81          |
| SegFormer     | 88.57    | 79.48 | 91.38    | 84.13 | 75.30    | 60.65 | 89.08    | 80.31 | 66.45    | 49.76 | 59.05    | 82.85 | 87.27 | ~18  | 14.78          |
| UNet          | 89.71    | 81.34 | 93.59    | 87.96 | 76.64    | 61.87 | 89.41    | 80.84 | 64.50    | 47.60 | 59.93    | 84.64 | 88.83 | ~31  | 3.45           |
| Transformer   | 89.01    | 80.20 | 94.22    | 89.07 | 76.60    | 62.07 | 89.73    | 81.37 | 57.08    | 39.94 | 58.78    | 84.72 | 88.70 | ~66  | 16.35          |
| EfficientUNet | 89.37    | 80.78 | 93.89    | 88.48 | 77.41    | 63.15 | 90.26    | 82.25 | 87.59    | 77.92 | 65.43    | 85.18 | 88.99 | ~78  | 13.97          |
| PSPNet        | 91.50    | 84.33 | 95.34    | 91.10 | 78.92    | 65.18 | 89.91    | 81.66 | 79.26    | 65.65 | 64.65    | 86.69 | 90.11 | ~66  | 7.58           |
| SegNet        | 91.22    | 83.87 | 95.56    | 91.49 | 79.16    | 65.50 | 90.45    | 82.57 | 83.12    | 71.13 | 66.50    | 86.96 | 90.31 | ~29  | 4.80           |
| Double EfficientUNet | 90.98 | 83.46 | 96.34    | 93.02 | 79.04    | 65.34 | 90.59    | 82.80 | **88.37** | **79.16** | 67.30    | 87.25 | 90.53 | ~282 | 19.56          |
| Proposed      | 91.47    | 84.32 | 96.38    | 93.02 | 79.42    | 65.86 | 90.90    | 83.31 | 84.12    | 78.76 | **73.19** | **87.70** | **90.87** | ~143 | 24.96          |

---

**Fig. 9.** Performance comparison with other deep learning models using the Potsdam validation dataset. The labels represent the impervious surface in white, buildings in blue, low vegetation in cyan, trees in green, cars in yellow, and clutter/background in red.
extract the global shape of objects in the figure, but is not accurately describing local information (the second car is not detected by the transformer model).

2) ISPRS Benchmark Dataset: In this experiment, the proposed model was tested using the testing datasets available on the ISPRS challenge website [53]. The model was trained using both the training and validation datasets. Its performance was compared with the top-performing scores presented on the competition website. Selected methods achieved the highest accuracy and had a detailed description of their implementation published in a journal or conference proceedings. Some authors tested different versions of their methods that essentially use the same strategy. In these cases, we selected the best-performing method. The proposed model was also compared with recently published methods like [21] and [66]. The selected methods for comparison are as follows.

1) SegNet + DSM + NDSM (ONE_7) [31]: The authors used a late fusion of two trained SegNets, using the IRRG image and a composite image that contained normalized digital vegetation index (NDVI), DSM, and NDSM data.

2) Self-cascaded+ResNet (CASIA2) [28]: A single self-cascaded network with an encoder that is based on a variant of the ResNet-101 network [45].

3) CNN + HCF + CRF (ADL_3) [59]: This model used a combination of CNN and hand-crafted features to extract representative image features and produce per-pixel category probabilities; then, a CRF was applied as a postprocessing step to find the predicted labels.

4) FCN + SegNet + NDSM (DLR_9) [19]: The authors used an ensemble model that contained an FCN and SegNet based on IRRG and NDSM images to find the best segmentation labels.

5) HUSTW3 [29]: The authors developed a residual architecture for the encoder–decoder models to address the issues of inadequate learning and receptive field imbalance faced by the encoder–decoder models.

6) FCN + DSM + RF + CRF (DST_2) [60]: IRRG and DSM data were used as inputs to the model based on an FCN trained with no downsampling and used random forest classifier to find the output features; next, a CRF was used as a postprocessing step.

7) Gated Segmentation Network (GSN3) [65]: A gated segmentation network was proposed. ResNet-101 was used as the feature extractor in the encoder portion, and an entropy control module was used for feature fusion in the decoder. A residual convolution module (RCM) was used as the basic processing unit.

8) CNN + NDSM + Deconvolution (UZ_1) [23]: The model comprises a CNN that has been trained to learn a series of downsampling operations (i.e., a regular CNN) and a sequence of nonlinear upsampling blocks using deconvolutions back to the original input size.

9) Dilated Convnet (UFMG_4) [62]: The authors proposed a series of dilated convolutions [62]. The main idea is to train a dilated network with different patch sizes to collect multicontext features from diverse contexts.

10) SegNet + FCN (RIT_7) [63]: SegNet was fused with an FCN for pixelwise semantic classification.

11) LANet [20]: The authors proposed a local attention network to improve the segmentation results by enhancing the scene-related representation in both the encoding and decoding phases.
Fig. 11. Segmentation results before and after combining EfficientUNet and Transformer. The labels represent the impervious surface in white, buildings in blue, and cars in yellow.

| Method                      | Imp. Surface | Building | Low Veg. | Tree | Car | Overall Accuracy |
|-----------------------------|--------------|----------|----------|------|-----|-----------------|
| Segmenter [13]              | 86.6         | 89.1     | 75.8     | 84.7 | 32.2 | 84.2            |
| SVL [54] *                  | 86.6         | 91.0     | 77.0     | 85.0 | 55.6 | 84.8            |
| SegFormer [39]              | 87.9         | 88.8     | 80.9     | 88.4 | 61.4 | 86.3            |
| MF-DFNet [57] *             | 88.8         | 93.1     | 78.4     | 84.0 | 77.5 | 86.2            |
| UZL [23]                    | 89.2         | 92.5     | 81.6     | 86.9 | 57.3 | 87.3            |
| RIT_L [58] *                | 89.6         | 92.2     | 81.6     | 88.6 | 76.0 | 87.8            |
| ADL_3 [59]                  | 89.5         | 93.2     | 82.3     | 88.2 | 63.3 | 88.0            |
| DST_2 [60] *                | 90.5         | 93.7     | 83.4     | 89.2 | 72.6 | 89.1            |
| CBGFNet [61] *              | -            | -        | -        | -    | -    | 89.3            |
| UFMG_4 [62] *               | 91.1         | 94.5     | 82.9     | 88.8 | 81.3 | 89.4            |
| ONB_7 [31] *                | 91.0         | 94.5     | 84.4     | 89.9 | 77.8 | 89.8            |
| RIT_7 [63] *                | 91.7         | 95.2     | 83.5     | 89.2 | 82.8 | 89.9            |
| LANet [20] *                | 92.4         | 94.9     | 82.9     | 88.9 | 81.3 | 89.9            |
| G2GNet [64] *               | 92.1         | 94.8     | 83.8     | 89.6 | 85.4 | 90.2            |
| GSN3 [65] *                 | 92.2         | 95.1     | 83.7     | 89.9 | 82.4 | 90.3            |
| DLR [19] *                  | 92.4         | 95.2     | 83.9     | 89.9 | 81.2 | 90.3            |
| Swin-B-CNN+BD [66] *        | 92.2         | 95.3     | 83.6     | 89.6 | 86.9 | 90.4            |
| SBA Net [21] *              | 94.4         | 92.9     | 83.4     | 89.6 | 91.4 | 90.5            |
| HiSTW [29] *                | 92.1         | 95.3     | 85.6     | 90.5 | 78.3 | 90.7            |
| DGCR [67]                   | 92.9         | 95.8     | 84.7     | 90.1 | 86.5 | 91.1            |
| CASIA2 [28]                 | 93.2         | 96.0     | 84.7     | 89.9 | 86.7 | 91.1            |
| AREANs-ResNeSt [68] *       | 93.2         | 96.1     | 84.8     | 90.2 | 90.5 | 91.3            |
| Proposed model without DSM data | 93.5     | 95.7     | 84.9     | 89.9 | 87.8 | 91.5            |
| Proposed model without the decoder unit * | 93.5     | 96.2     | 84.8     | 90.4 | 87.1 | 91.5            |
| AFNet [69]                  | 93.4         | 95.9     | 86.0     | 90.7 | 87.2 | 91.6            |
| Swin-S [70]                 | 93.6         | 96.2     | 85.8     | 90.4 | 87.6 | 91.6            |
| Proposed model *            | 94.7         | 98.0     | 89.4     | 90.4 | 96.0 | 91.8            |

* Models using DSM or nDSM data

12) **Swin-B-CNN + BD [66]:** The Swin transformer and a CNN were used as encoders and decoders. The CNN was used to recover the size of feature maps and generate semantic segmentation results.

13) **AREANs-ResNeSt [68]:** Attention-Residual block-Embedded Adversarial Network was used to learn local-to-global contextual information through semantic and position information.

14) **MF-DFNet [57]:** A multiscale feature network and a discriminative feature network were proposed to solve the issues of intra-class inconsistency and the difficulty in locating and identifying targets.

15) **DGCR [67]:** A dynamic graph contextual reasoning module over global reasoning networks was presented for capturing long-range dependencies in feature representations.

16) **Swin-S [70]:** The Swin transformer and a densely connected feature aggregation module were used as the encoder and decoder units, respectively.

17) **CEGFNet [61]:** An end-to-end common extraction and gate fusion network was proposed to solve the problem of misclassification of small objects.

18) **AFNet [69]:** In this article, multiscale and multilevel CNN based maps were combined.

19) **SBA Net [21]:** A semantic boundary awareness network was developed to extract full and crisp borders from remote sensing images.

20) **SVL-DSM-CRF (SVL) [54]:** The baseline approach used by the challenge’s organizer [54]. In addition to the
TABLE VIII

| Method                        | F1 Score | Overall Accuracy |
|-------------------------------|----------|------------------|
|                              | Imp. Surface | Building | Low Veg. | Tree | Car |
| SegFormer [39]               | 79.7      | 83.7           | 70.4     | 57.0 | 62.1 | 73.3 |
| SVL [54] *                   | 83.5      | 91.7           | 72.2     | 63.2 | 62.2 | 77.8 |
| Segmenter [13]               | 85.5      | 87.7           | 79.2     | 73.6 | 85.8 | 81.2 |
| CEGFNet [61] *               | -         | -              | -        | -    | -    | 85.2 |
| UZ [23]                      | 89.3      | 95.4           | 81.8     | 80.5 | 86.5 | 85.8 |
| RIT-L [58] *                 | 91.2      | 94.6           | 85.1     | 85.1 | 92.8 | 88.4 |
| DST [60] *                   | 92.5      | 96.4           | 86.7     | 88.0 | 94.7 | 90.3 |
| RIT [63] *                   | 92.6      | 97.0           | 86.9     | 87.4 | 95.2 | 90.3 |
| G2GNet [64] *                | 93.0      | 96.5           | 86.3     | 88.2 | 95.8 | 90.3 |
| LANet [20] *                 | 93.1      | 97.2           | 87.3     | 88.0 | 94.2 | 90.8 |
| Swin-B-CNN+BD [66] *         | 93.6      | 96.7           | 88.0     | 88.0 | 96.3 | 91.0 |
| CASIA2 [28]                  | 93.3      | 97.0           | 87.7     | 88.4 | 96.2 | 91.1 |
| HUSTW3 [29] *                | 93.6      | 97.6           | 88.5     | 88.8 | 94.6 | 91.6 |
| DGCN [67]                    | 94.1      | 97.3           | 88.3     | 88.9 | 93.0 | 91.8 |
| Proposed model without DSM data | 93.5      | 97.2           | 86.9     | 88.5 | 96.2 | 91.8 |
| AREANs-ResNetSt [68]         | 94.3      | 97.4           | 88.5     | 89.6 | 97.0 | 91.9 |
| Swin-S [70]                  | 94.2      | 97.6           | 88.6     | 89.6 | 96.3 | 92.0 |
| AFNet [69]                   | 94.2      | 97.2           | 89.2     | 89.4 | 95.1 | 92.2 |
| Proposed model without the decoder unit * | 94.0      | **98.0**       | 89.2     | 90.0 | 95.8 | 92.5 |
| SBA-Net [21] *               | 93.8      | **98.0**       | 89.0     | 89.5 | 94.7 | 92.8 |
| Proposed model *             | 94.8      | **98.0**       | **89.5** | **90.5** | 94.6 | **92.9** |

* Models using DSM or nDSM data

V. CONCLUSION

In this article, we proposed a novel fusion deep learning model for investigating the semantic labeling of multimodal high-resolution urban remote sensing data. We showed that the fusion of deep transformers and conventional neural networks (i.e., the U-Net model) is an effective method for recognizing the relationships between objects and scenes, leading to consistent labeling outcomes for complex urban objects.

We evaluated the proposed method on two challenging publicly available datasets, and the results demonstrate its efficacy and efficiency. The proposed model outperforms the existing frameworks, producing more accurate labeling outcomes with higher consistency.

ACKNOWLEDGMENT

The authors would like to thank Dr. Markus Gerke, Technische Universität Braunschweig, Braunschweig, Germany, for providing the normalized digital surface model (NDSM) data.

REFERENCES

[1] Z. Wang, J. Wang, K. Yang, L. Wang, F. Su, and X. Chen, “Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with DeepLabv3+,” Comput. Geosci., vol. 158, Jan. 2022, Art. no. 104969.

[2] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, “Deep learning in remote sensing applications: A meta-analysis and review,” ISPRS J. Photogramm. Remote Sens., vol. 152, pp. 166–177, Jun. 2019.
M. Volpi and D. Tuia, “Dense semantic labeling of subdecimeter resolution images with convolutional neural networks,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 2, pp. 881–893, Feb. 2016.

D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, “Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models,” ISPRS J. Photogramm. Remote Sens., vol. 145, pp. 96–107, Nov. 2018.

J. R. Bergado, C. Persello, and A. Stein, “Recurrent multiresolution convolutional networks for VHR image classification,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 11, pp. 6361–6374, Jun. 2018.

Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, “Semantic labeling in very high resolution images via a self-cascaded convolutional neural network,” ISPRS J. Photogramm. Remote Sens., vol. 145, pp. 78–99, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271617303854

Y. Sun, Y. Tian, and X. Yu, “Problems of encoder–decoder framework for high-resolution remote sensing image segmentation: Structural stereotype and insufficient learning,” Neurocomputing, vol. 330, pp. 297–304, Feb. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231218313821

Z. Zhong, J. Li, W. Cui, and H. Jiang, “Fully convolutional networks for building and road extraction: Preliminary results,” in Proc. Int. Conf. Int. Geosci. Remote Sens. Symp. (IGARSS), Jul. 2016, pp. 1591–1594.

N. Audebert, B. Le Saux, and S. Lefèvre, “Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks,” ISPRS J. Photogramm. Remote Sens., vol. 140, pp. 20–32, Jun. 2018.

P. Shamsolmoali, M. Zareapoor, R. Wang, H. Zhou, and J. Yang, “A novel deep structure U-Net for sea-land segmentation in remote sensing images,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 12, no. 9, pp. 3210–3223, Sep. 2019.

W. Feng, H. Sui, L. Hua, C. Xu, G. Ma, and W. Huang, “Building extraction from VHR resolution imagery by combining an improved deep convolutional encoder–decoder architecture and historical land use vector map,” Int. J. Remote Sens., vol. 41, no. 17, pp. 6595–6617, Sep. 2020, doi: 10.1080/01431161.2020.1742944.

Z. Zhang, J. Li, W. Cui, and H. Jiang, “Fully convolutional networks for building and road extraction: Preliminary results,” in Proc. Int. Conf. Int. Geosci. Remote Sens. Symp. (IGARSS), Jul. 2016, pp. 1591–1594.

S. Du, S. Bu, B. Liu, and X. Zhang, “Incorporating DeepLabv3+ and object-based image analysis for segmentation of very high resolution remote sensing images,” Int. J. Digit. Earth, vol. 14, no. 3, pp. 537–578, Mar. 2021, doi: 10.1080/17538947.2020.1851307.

A. Buades, B. Coll, and J.-M. Morel, “A non-local algorithm for image denoising,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2, Jun. 2005, pp. 60–65.

W. Zhao, S. Du, and W. J. Emery, “Object-based convolutional neural network for high-resolution imagery classification,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 10, no. 7, pp. 3386–3396, Jul. 2017.

W. Zhao, S. Du, Q. Wang, and W. J. Emery, “Contextually guided very-high-resolution imagery classification with semantic segments,” ISPRS J. Photogramm. Remote Sens., vol. 132, pp. 48–60, Oct. 2017.

J. Chen et al., “TransUNet: Transformers make strong encoders for medical image segmentation,” 2021, arXiv:2102.04306.

T. Panboonyuen, “Transformer-based decoder designs for semantic segmentation on remotely sensed images,” Remote Sens., vol. 13, no. 24, p. 5100, 2021, doi: 10.3390/rs13245100.

O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedial image segmentation,” in Medical Image Computing and Computer-Assisted Intervention—MICCAI, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241.
[44] A. Kirillov, R. Girshick, K. He, and P. Dollar, “Panoptic feature pyramid networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6399–6408.
[45] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6230–6239.
[46] W. Wang et al., “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,” 2021, arXiv:2102.12122.
[47] Y.-C. Li, H.-C. Li, W.-S. Hu, and H.-L. Yu, “DSPCANet: Dual-channel scale-aware segmentation network with position and channel attentions for high-resolution aerial images,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 8552–8565, 2021.

T. Q. Nguyen and J. Salazar, “Transformers without tears: Improving the normalization of self-attention,” in Proc. 16th Int. Conf. Spoken Lang. Transl. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 1–11. [Online]. Available: https://aclanthology.org/2019.iswitl-1.17

[48] H. Chen, Z. Qi, and Z. Shi, “Remote sensing image change detection with transformers,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5607514.
[49] M. Sandier, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 4510–4520.
[50] A. Telea, “An image inpainting technique based on the fast marching method,” J. Graph. Tools, vol. 9, no. 1, pp. 23–34, 2004, doi: 10.1080/10867651.2004.10487596.

[51] ISPRS. (2016). International Society for Photogrammetry and Remote Sensing. 2D Semantic Labeling Challenge. Accessed: Sep. 30, 2021. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html
[52] M. Gerke, “Use of the stair vision library within the ISPRS 2D semantic labeling benchmark (Vaihingen),” Dept. Earth Observ. Sci., Univ. Twente, Enschede, The Netherlands, Tech. Rep., Jan. 2015, doi: 10.13140/2.15015.9683.

[53] Z. Zhang, Q. Liu, and Y. Wang, “Road extraction by deep residual U-Net,” IEEE Geosci. Remote Sens. Lett., vol. 15, no. 5, pp. 749–753, May 2018.
[54] 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, Jan. 2017.
[55] S. Zhang, C. Wang, J. Li, and Y. Sui, “MF-DiNet: A deep learning method for pixel-wise classification of very high-resolution remote sensing images,” Int. J. Remote Sens., vol. 43, no. 1, pp. 330–348, Jan. 2022, doi: 10.1080/01431161.2021.2018147.

[56] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, “Dense semantic labeling of very-high-resolution aerial imagery and LiDAR with fully-convolutional neural networks and higher-order CRFs,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 1561–1570.
[57] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. V. D. Hengel, “Effective semantic pixel labelling with convolutional networks and conditional random fields,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2015, pp. 36–43.

[58] J. Sherrah, “Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery,” 2016, arXiv:1606.02585.
[59] W. Zhou, J. Jin, J. Lei, and J.-N. Hwang, “CEGFNet: Common extraction and gate fusion network for scene parsing of remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5405110.
[60] K. Nogueira, M. D. Mura, J. Chanussot, W. R. Schwartz, and J. A. D. Santos, “Dynamic multicontext segmentation of remote sensing images based on convolutional networks,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 10, pp. 7503–7520, Oct. 2019, doi: 10.1109/TGRS.2019.2913861.

[61] S. Piramanayagam, E. Saber, W. Schwartzkopf, and F. Koehler, “Supervised classification of multisensor remotely sensed images using a deep learning framework,” Remote Sens., vol. 10, no. 9, p. 1429, Sep. 2018. [Online]. Available: https://www.mdpi.com/2072-4292/10/9/1429
[62] X. Zheng, X. Wu, L. Huan, W. He, and H. Zang, “A gather-to-guided framework for remote sensing semantic segmentation and auxiliary image,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5404915.

[63] H. Wang, Y. Wang, Q. Zhang, S. Xiang, and C. Pan, “Gated convolutional neural network for semantic segmentation in high-resolution images,” Remote Sens., vol. 9, no. 5, p. 446, May 2017. [Online]. Available: https://www.mdpi.com/2072-4292/9/5/446
[64] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, “Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 4408820.

[65] Y. Su, J. Cheng, W. Wang, H. Bai, and H. Liu, “Semantic segmentation for high-resolution remote-sensing images via dynamic graph context reasoning,” IEEE Geosci. Remote Sens. Lett., vol. 19, 2022, Art. no. 6807505.

[66] K. Yan, H. Wang, S. Bu, L. Yang, and J. Li, “Scene parsing for very high resolution remote sensing images using on attention-residual block-embedded adversarial networks,” Remote Sens. Lett., vol. 12, no. 7, pp. 625–635, Jul. 2021, doi: 10.1016/j.rsl.2021.1910362.

[67] R. Liu, L. Mi, and Z. Chen, “AFNet: Adaptive fusion network for remote sensing image semantic segmentation,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 9, pp. 7871–7886, Sep. 2021.

[68] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, “A novel transformer based segmentation scheme for fine-resolution remote sensing images,” IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.

[69] S. Gould, O. Russakovsky, I. Goodfellow, P. Baumstarck, A. Y. Ng, and D. Koller, “The STAIR vision library (v2.4),” Stanford Univ., Stanford, CA, USA, 2010. [Online]. Available: http://ai.stanford.edu/~sgould/svla

[70] Z. Yang, Z. Yan, X. Sun, W. Diao, Y. Yang, and X. Li, “Category correlation and adaptive knowledge distillation for compact cloud detection in remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5623315.

[71] Z. Yang et al., “Adaptive knowledge distillation for lightweight remote sensing object detectors optimizing,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5623715.

[72] Y. Zhang, Z. Yan, X. Sun, W. Diao, K. Fu, and L. Wang, “Learning efficient and accurate detectors with dynamic knowledge distillation in remote sensing imagery,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5613819.

Hasan AlMarzouqi (Senior Member, IEEE) received the bachelor’s (Hons.) and M.Sc. degrees in electrical and computer engineering from Vanderbilt University, Nashville, TN, USA, in 2004 and 2006, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2014. His research interests include deep learning, artificial intelligence, digital rock physics, and bioinformatics.

Dr. Almarzouqi is a member of the IEEE Signal Processing Society.

Lyes Saad Saoud received the D.Sc. and M.C.A. degrees (empowerment to direct research) in physics, specialty renewables energies from the University of Boumerdes, Boumerdes, Algeria, in January 2015 and February 2017, respectively. He was an Associate Professor with the University of Boumerdes. From February 2019 and October 2019, he became the Head and a Scientific Member of the Department of Electrical Systems Engineering and a Scientific Member of the Faculty of Technology, University of Boumerdes. He is currently an engineer and a magister in electronics. His recent research interests include advanced neural networks, fuzzy logic, hybrid systems, renewable energy, sustainability, and environment, energy engineering, and cybersecurity.

Dr. Saad Saoud has served as a member of the Technical Committee for several international specialized conferences. He is a Reviewer of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE ACCESS, Transactions of the Institute of Measurement and Control (SAGE Publications), Journal of Renewable and Sustainable Energy (the American Institute of Physics), Neurocomputing (Elsevier), and Neural Processing Letters (Springer).

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.