IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE GEOSCIENCE AND REMOTE SENSING SOCIETY SECTION

Received 21 December 2023, accepted 5 January 2024, date of publication 17 January 2024, date of current version 25 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3355154

RESEARCH ARTICLE

An Attentive Hough Transform Module for Building Extraction From High Resolution Aerial Imagery

SOUAD YAHIA BERROUIGUET[1,2], EHLEM ZIGH[3], AND MOHAMMED DJEBOURI[1]

[1]Department of Electronics, Faculty of Electrical Engineering, University Djillali Liabes of Sidi Bel Abbes, Sidi Bel Abbès 22000, Algeria
[2]National Higher School of Telecommunication and I. C. T. Abdelhafid Boussouf, Oran 31000, Algeria
[3]Laboratoire de Codage et de la Sécurité de l'information, Department of Electronics, Faculty of Electrical Engineering, University of Science and Technology of Oran—Mohamed Boudiaf, Oran 31000, Algeria

Corresponding author: Souad Yahia Berrouiguet (syberrouiguet@ensttic.dz)

ABSTRACT In the era of abundant high-resolution aerial imagery, the automatic extraction of buildings is indispensable for applications like disaster response, environmental monitoring, and urban growth analysis. Deep learning approaches, particularly fully convolutional networks, have exhibited remarkable performance in this challenging task. Nevertheless, the accurate identification and delineation of building boundaries pose persistent challenges hindering further improvements in building extraction precision. To tackle these, we introduce a novel deep learning architecture explicitly designed for building extraction in high-resolution aerial images. Our method addresses the precise identification of building borders by combining both local and global contextual information. We efficiently preserve object boundaries and optimize the representation of straight lines within buildings through the integration of the Attentive Hough Transform and Inverse Hough Transform (AttHT-IHT) module into the U-Net architecture. Extensive experiments on the Potsdam dataset showcase substantial enhancements in building extraction accuracy, with a 97.73% accuracy rating and a 96.42% recall rate. Generalization capability on the WHU satellite dataset I was assessed to validate the adaptability of our proposed method.

INDEX TERMS Aerial images, AttHT-IHT, building extraction, deep learning, U-Net.

I. INTRODUCTION

The rapid advancements in remote sensing technology has rendered high-resolution aerial imagery more accessible than ever before. This democratization has bestowed the ability to discern and extract small man-made objects, particularly buildings. The significance of building extraction spans a plethora of practical domains, encompassing disaster management [1], urban development analysis [2], [3], and environmental monitoring [4]. However, the path to automated building extraction from remote sensing imagery remains strewn with challenges ranging from the diversity of building appearances and sizes to the intricacies of scene complexities and incomplete cue extraction [5]. In this context, deep learning techniques have risen as a formidable answer to the challenges within the realm of computer vision tasks [6], [7], [8], [9]. More specifically, Convolutional Neural Networks (CNNs) have achieved exceptional results [10], [11]. Unlike traditional methodologies that require manual feature extraction, CNNs excel in automated feature extraction and subsequent classification through their convolutional and fully connected layers. CNNs offer an integrated solution by merging feature extraction and classification into a unified model. Furthermore, they often showcase heightened generalization capabilities by directly imbibing knowledge pertaining to feature extraction from the dataset.

Additionally, Fully Convolutional Networks (FCNs) have ushered in a pioneering approach, augmenting the capabilities of CNNs with a specific focus on semantic segmentation

The associate editor coordinating the review of this manuscript and approving it for publication was Stefania Bonafoni.

tasks, thus exerting a profound influence on the domain of building extraction from remote sensing images. Researchers have diligently probed various architectural designs and methodologies to amplify the precision and effectiveness of this crucial task. Notably, Sherrah [12] introduced an FCN refinement that substantially enhanced building delineation accuracy. Maggiori et al. [13] devised a multiscale structure to surmount the intricate trade-off between context expansion and parameter augmentation. Meanwhile, Liu et al. [14] introduced an ingenious dense FCN architecture to bolster building recognition.

Another remarkable architecture, U-Net, which was initially conceived by Ronneberger et al. [15] for medical image segmentation, has become pivotal in semantic segmentation. U-Net excels in capturing intricate details whereas preserving spatial information, leveraging upsampling and downsampling techniques for contextual understanding and precise localization. It outperforms sliding window convolution networks in terms of performance, even with minimal training data, and efficient feature extraction, effectively addressing the limitations inherent in FCNs.

DeepLab models [16], [17] also marked significant progress, harnessing Atrous convolutions and integrating conditional random fields for refined post-processing. These models have served as the foundation for the development of several enhanced models [18], [19], [20], [21], [22]. Additionally, with the emergence and advancement of attention mechanisms such as spatial attention [23], self-attention [24], squeeze and excitation networks [25] and Convolutional Block Attention Module (CBAM) [26], significant progress has been achieved in building extraction. Various studies have leveraged these attention mechanisms to boost network capabilities, contributing unique insights into the field [19], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36].

For example, Yang et al. [27] utilized spatial attention to enrich deep feature maps.Whereas Hosseinpoor and Samadzadegan [29] recalibrated intermediate features using a spatial attention module. Ye et al. [30] focused on precise building extraction by employing re-weighting techniques and a joint attention module. Shi et al. [31] combined HRNET [32] with Class-Oriented Region Attention and Context Fusion Modules, facilitating meaningful connections between classes and regions. Innovative approaches such as the integration of boundary-aware loss, as demonstrated by Barnet [33] within a multilevel feature fusion block, have been employed to enhance edge sharpness.

In SCAttNet [34], the authors applies sequential channel and spatial attention modules to improve feature quality for semantic segmentation in remote sensing images. In [35], the authors use a lightweight RegNet network and a multiscale depthwise separable atrous spatial pyramid pooling structure for feature extraction in the encoding stage. They then employ squeeze-and-excitation attention and lightweight residual blocks, to refine and reconstruct building features in the decoding stage. In [36], the authors introduce the SSDBN model, incorporating an enhanced Res2Net encoder and a dual-branch decoder with CBAM. This design emphasizes the model's proficiency in capturing both global and local context details.

Simultaneously, other research efforts [37], [38], [39], [40] explored the concept of incorporating prior knowledge to enhance network performance and reduce reliance on extensively annotated datasets.

In our study, we draw inspiration from the advantages offered by both attention mechanisms and prior knowledge. Through a synergistic fusion of these two mechanisms, we introduce an innovative and effective approach tailored to address the specific challenges associated with building extraction. We present the AttHT-IHT module as the cornerstone of our proposed methodology, which unites these principles to achieve remarkable results in building extraction tasks.

The contributions of this study include the following:

- We introduce an innovative AttHT-IHT module, skillfully integrated into an enhanced U-Net network, aiming to achieve precise building edge detection and enhanced building extraction accuracy. To the best of our knowledge, this combination has not been explored or implemented previously.
- We validate the efficacy of the proposed module by comparing the building extraction results obtained with and without the inclusion of the AttHT-IHT module in the U-Net architecture.
- We conduct extensive experiments by comparing the new U-Net-based architecture with a range of state-of-the-art building extraction methods.
- We evaluate the overall architecture's accuracy by deploying it on two distinct datasets, namely, the ISPRS Potsdam dataset and the WHU satellite dataset I.

II. OUR APPROACH/METHOD

This paper presents an architecture specifically designed for building extraction from aerial images, which combines the U-Net framework, attentive mechanisms, and the ASPP block with different dilation rates. The overall architecture is illustrated in Fig.1. In the following sections, we provide detailed explanations of each component of our architecture and how they work together to achieve state-of-the-art performance in building extraction from remote sensing imagery.

A. U-NET NETWORK

In this study, we utilize a version of U-Net as the baseline design, which is a well-established encoder-decoder model commonly used in tasks involving medical and remote sensing image segmentation [15], [41], [42]. Fig.1 illustrates the feature maps represented by the blue rectangular blocks following the U-Net structure. In the encoder part, we employ max pooling (red arrow) to reduce the dimensions of the feature maps for efficient computation. Conversely,in the decoder section, we use up-sampling techniques (purple arrow) to restore input image dimensions. The green

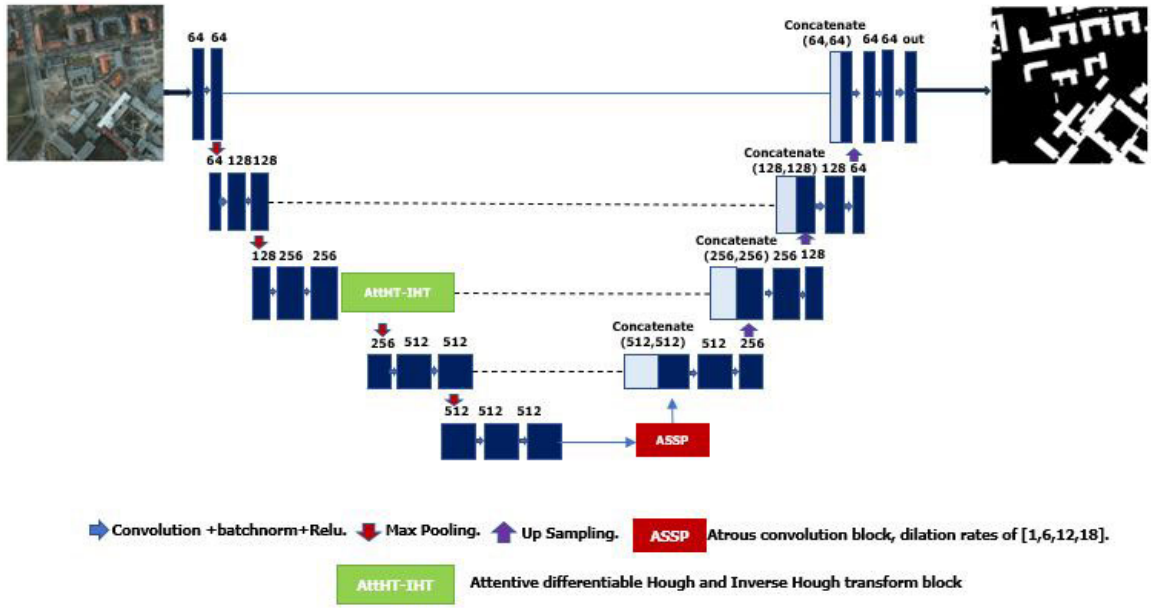

FIGURE 1. Our proposed architecture.

rectangle illustrate how we have incorporated the AttHT-IHT block to enhance the architecture's ability to identify patterns related to buildings. Serving as a connection between the encoder and decoder we have included an ASPP block that helps gather information from the context. Finally, in order to generate binary predictions we employ a 1×1 convolution, with a Sigmoid activation function as the last layer of the U-Net model.

Algorithm 1 Hough Transform

function HoughTransform($\mathcal{F}$, N_{offsets}, N_{angles})
  Initialize an empty Hough histogram H of size [N_{offsets} x N_{angles}]
  for each pixel (x_i, y_i) in image $\mathcal{F}$ **do**
    for each angle index from 0 to $N_{\text{angles}} - 1$ **do**
      Calculate ρ as $x_i \cdot \cos(\text{angle\_index} \cdot (360/N_{\text{angles}})) + y_i \cdot \sin(\text{angle\_index} \cdot (360/N_{\text{angles}}))$
      if ρ is in the range $[0, N_{\text{offsets}}]$ **then**
        Increment H at bin (ρ, angle_index) by $\mathcal{F}(x_i, y_i)$
      end if
    end for
  end for
  Remove all-zero lines in H for efficiency
  return the resulting Hough histogram H
end function

B. ATTHT-IHT BLOCK

While deep networks excel in general feature extraction, the Hough transform specifically enhances the model's performance in detecting linear structures, such as building edges. Its robustness to small gaps, noise, and partial occlusion makes it a valuable complement to deep networks, ensuring accurate and comprehensive feature representation for precise building extraction. This is particularly crucial in scenarios where deep networks face challenges related to the detection of intricate linear details. In the context of our approach, and drawing inspiration from previous research [37], we enhanced the existing module by integrating the attention mechanism using CBAM to form the AttHT-IHT block. This block operates as a trainable attention module within the encoder, inferring attention maps along channel and spatial dimensions. By harmonizing local learned image features with global predictions from Hough lines, the AttHT-IHT block sequentially refines features and optimizes the attentive Hough transform weight matrix during the training process. The structure of the AttHT-IHT block, illustrated in Fig.2, comprises three key stages: HT, CBAM Integration, and IHT. The appropriate pseudocode for each stage is also provided below.

Step 1: Hough Transform application

The initial step involves applying the Hough transform to the input feature map. This is accomplished by computing votes of pixels along lines in the image, generating a Hough histogram. The channels of the Hough histogram are subsequently filtered and refined to extract relevant line features.

Step 2: CBAM integration

The CBAM module [26] is strategically incorporated between the HT and IHT stages. It generates attention maps for both channel and spatial positions, assigning importance weights to different features based on their relevance to the task.These attention maps are then element-wise multiplied with the input feature map, facilitating adaptive feature refinement guided by learned attention patterns. CBAM

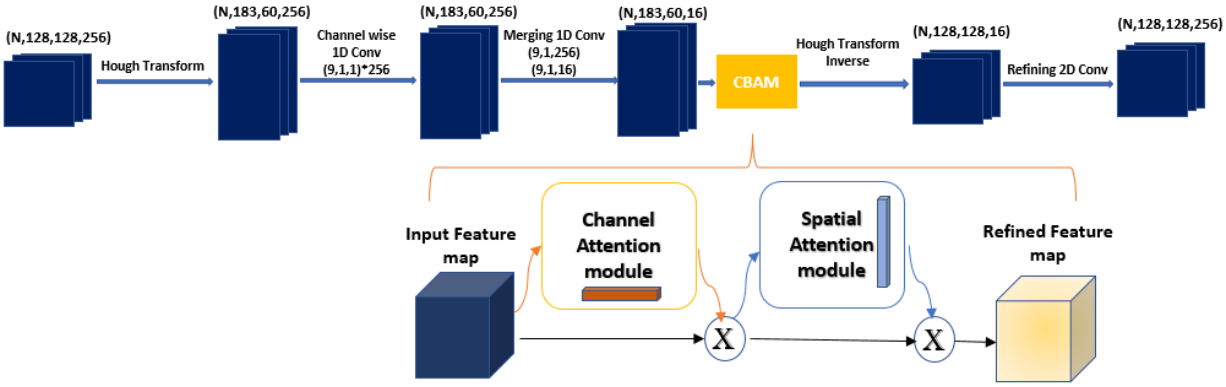

FIGURE 2. The proposed AttHT-IHT block structure.

also enhances features related to building structures by emphasizing the most pertinent lines for building delineation. In other words, CBAM functions as a selector mechanism within the block, prioritizing details deemed relevant for line extraction whereas mitigating the influence of less useful information. This enables the model to focus on crucial information related to building boundaries.

Algorithm 2 CBAM Module

function CBAM($\mathcal{F}$)
  Compute Channel Attention Map A_c using global average pooling of $\mathcal{F}$
  Compute Spatial Attention Map A_s using convolutional layers and sigmoid activation
  Apply Channel Attention to $\mathcal{F}$: $\mathcal{F}_c = \mathcal{F} \cdot A_c$
  Apply Spatial Attention to $\mathcal{F}_c$: $\mathcal{F}_{cs} = \mathcal{F}_c \cdot A_s$
  return $\mathcal{F}_{cs}$
end function

Step 3: Application of IHT

The final stage entails the application of the inverse Hough transform, which converts the Hough histogram into a feature map in the image domain. This feature map is prepared for precise building contour extraction.

In summary, our AttHT-IHT block seamlessly combines three crucial steps: HT, CBAM, and IHT, to maximize accuracy and quality in building extraction. The HT phase lays the foundation by translating semantic features into Hough space, whereas CBAM enhances feature refinement and line selection. Finally, IHT converts the enhanced Hough histogram back into a feature map for precise building contour extraction. This sequence of transformations ensures optimal accuracy and quality in obtaining building contours from very high-resolution aerial images.

C. ASPP FOR EXTRACTING MULTISCALE FEATURES

The ASPP (Atrous Spatial Pyramid Pooling) module [16] is a key component of our architecture, designed to effectively capture contextual information at various scales. It employs Atrous convolutions at different rates to combine the benefits

Algorithm 3 Inverse Hough Transform (IHT) Module

function IHT($\mathcal{F}_{cs}$, N_{offsets}, N_{angles})
  Initialize an empty output feature map $\mathcal{F}_{iht}$ of the same size as $\mathcal{F}_{cs}$
  for each pixel (x_i, y_i) in $\mathcal{F}_{cs}$ **do**
    Initialize sum sum_iht as 0
    for each angle index from 0 to $N_{\text{angles}} - 1$ **do**
      Calculate ρ as $x_i \cdot \cos(\text{angle\_index} \cdot (360/N_{\text{angles}})) + y_i \cdot \sin(\text{angle\_index} \cdot (360/N_{\text{angles}}))$
      if ρ is in the range $[0, N_{\text{offsets}}]$ **then**
        Increment sum_iht by $H(\rho, \text{angle\_index})$
      end if
    end for
    $\mathcal{F}_{iht}(x_i, y_i) = \frac{\text{sum\_iht}}{N_{\text{angles}}}$
  end for
  return $\mathcal{F}_{iht}$
end function

of this technique, enabling the network to consider a broader or narrower context without adding extra parameters. In our architecture, the ASPP module consists of a 1×1 convolution layer and three branches of Atrous convolutions with rates of 6, 12, and 18 (Fig.3). This integration allows the network to analyze aerial images at different scales, improving the detection of buildings of varying sizes and the capture of fine details, ultimately enhancing the accuracy of building segmentation in aerial images.

III. DATASETS USED AND EVALUATION METRICS

In this section, we present an overview of the datasets used. We delve into detailed discussions on data processing methods and experimental settings. Additionally, we introduce the evaluation metrics employed in this study.

A. DATASETS

1) POTSDAM DATASET

The performance of the proposed network was assessed using the well-established ISPRS 2D Potsdam semantic benchmark dataset, a highly regarded dataset in the remote

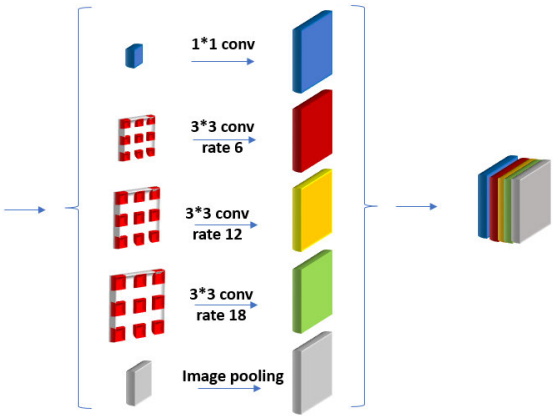

FIGURE 3. The ASPP structure in our network.

TABLE 1. Potsdam dataset characteristic.

Items	Training	Testing
Bands used	RGB	RGB
Ground samplaning distance	6000 × 6000	6000 × 6000
Sample size	5cm	5cm
Use size	512 × 512	512 × 412
Sample number	24	14

sensing field [43]. This dataset comprises high-resolution aerial images taken over Potsdam, Germany, with a total of 38 patches, each measuring 6000 × 6000 pixels. The dataset includes a true orthophoto (TOP), a digital surface model (DSM), and utilizes the near infrared (NIR), red (R), green (G), and blue (B) bands as features. The dataset is categorized into six classes, each manually labeled using distinctive colors: cluster/background (red), Impervious surfaces (white), trees (green), low vegetation (cyan), cars (yellow), and buildings (blue). Only RGB bands of TOP are utilized as features, excluding other bands.

To mitigate over-fitting concerns, the dataset is partitioned into two subsets: a training set and a testing set. The training set contains 24 patches, whereas the testing set contains 14 patches. Due to hardware limitations, the data is sliced into smaller patches measuring 800 × 800 pixels. The training datasets incorporate a 200 pixel overlap to minimize the potential impact of the slicing process. Table 1 provides an overview of the dataset's specific characteristics. Additionally, Fig.4 visually presents sample images from the dataset, accompanied by their corresponding labels.

2) WHU SATELLITE DATASET I

The WHU Satellite Dataset I (global cities) [44], is a mosaic of remote sensing data gathered from various satellites, including QuickBird, worldview series, IKONOS, and ZY-3. It encompasses 204 images with resolutions ranging from 0.3 to 2.5 meters. This dataset's distinctive feature lies in its diverse surface building textures and shapes, making it an ideal yet challenging benchmark for evaluating the robustness of building extraction algorithms. We specifically used this dataset to assess the generalization capability of our method.

B. DATA PROCESSING

To address the limited number of labeled images available in the Potsdam database and to mitigate over-fitting, data augmentation was employed as an effective approach to expand the dataset. In this research, data augmentation involved applying vertical and horizontal flipping operations. These augmentation techniques contribute to enlarging the dataset and enhancing its diversity, thereby improving the generalization capabilities of the model and reducing the risk of over-fitting [45].

C. EXPERIMENT SETTINGS

The experiments were executed with the PyTorch [46] deep learning framework, employing the Adam optimizer [47] for training. Batch size and learning rate were determined through experiments. In our approach, we used a loss function known as the Binary Cross Entropy (BCE). It's designed specifically for tasks which we have two categories, such as determining whether a pixel in the input image belongs to a building or not. BCE helps us measure how well our model's predicted probabilities match the actual classification of each pixel in the image.

D. EVALUATION METRICS

To ensure a fair comparison with existing literature, identical metrics are used to evaluate the performance of our network. We employed overall accuracy to assess the overall performance of our network, along with four quantitative evaluation metrics:

IoU, a widely employed metric for evaluating image semantic segmentation results, is utilized to compare the predicted output with the corresponding ground truth. The IoU is calculated using (1):

$$IoU = \frac{TP}{TP + FN + FP} \tag{1}$$

Precision refers to the fraction of positive samples that are correctly predicted as positive samples out of all positive samples. The Precision calculation is displayed in (2).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall, also known as true positive rate, measures the fraction of actual positive samples that are correctly identified as positive by the model. The recall computation is presented in (3).

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Precision and recall are fully considered by F1-Score. The F1-Score computation is given in (4).

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where TP represents correctly predicted positives (buildings), FP is for incorrectly predicted negatives as positives, TN stands for correctly predicted negatives, and FN denotes incorrectly predicted positives as negatives.

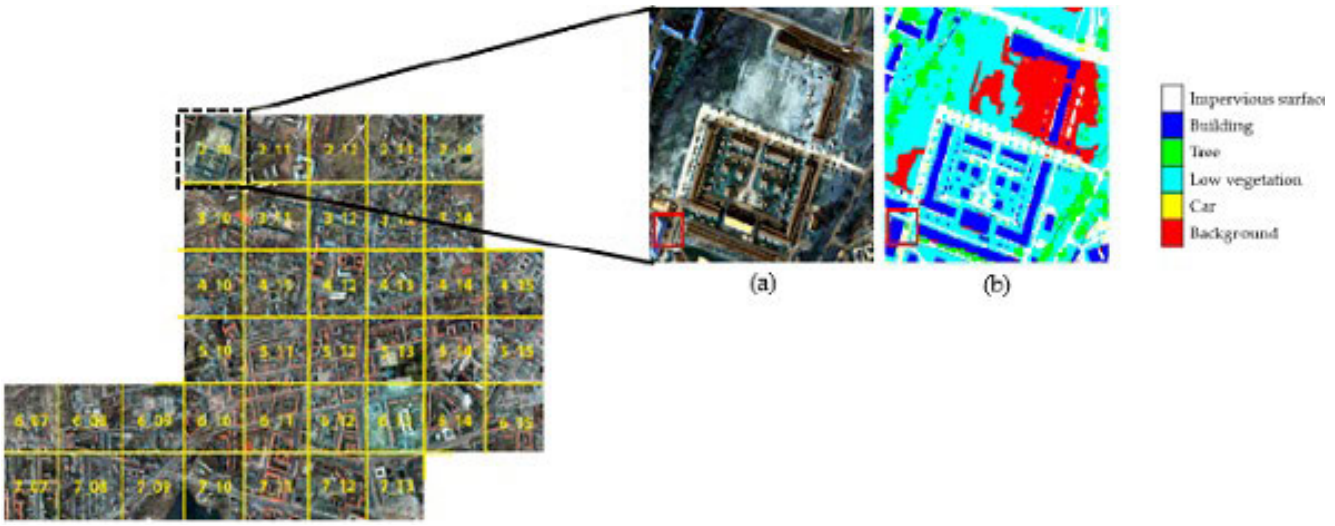

FIGURE 4. Potsdam Dataset [43].

IV. RESULTS

A. EXPERIMENTAL RESULTS ON THE POTSDAM DATASET

Over the Potsdam dataset, the architecture achieves overall accuracy of 97.73% and F1-Score of 96.28%, the deep learning frame performs exceptionally well in building extraction. As observed in Fig.5, our model (column c) demonstrates accurate and efficient building extraction capabilities. It successfully captures buildings of various shapes, including highly complex structures, closely resembling the ground truth (column b). Notably, our model excels in accurately extracting small buildings, showcasing its ability to handle diverse architectural structures. These results highlight the effectiveness and versatility of our proposed model in accurately identifying and preserving the integrity of buildings, regardless of their size or complexity.

B. COMPARISON EXPERIMENTS

In this section, we conduct a comprehensive comparison between our modeling framework and several alternative methods that employ attention mechanisms for building extraction from the Potsdam dataset. It is important to note that approach presented in [48] stands out from the others as it utilizes direct fusion of 2D and 3D data where the 3D data is obtained by converting Digital Surface Models (DSMs) data. Table 2 presents a detailed overview of the performance of these methods, focusing on quantitative metrics.

Our proposed method achieved the highest IoU score of 92.88%, outperforming the second-best method by Barnet [33] which achieved an IoU score of 92.24%. The difference between our method and the second-best is 0.64%. In terms of Precision, our method ranked second among the compared studies, with a Precision score slightly below that of Jin et al. [33], who obtained the highest Precision of 98.64%. Regarding Recall, our method demonstrated the highest score of 96.42%, surpassing the second-best score reported by Barnet [33] by 1.3%. In terms of F1-Score, our method achieved the second-highest score of 96.28%, with Jin et al. [33] obtaining the highest F1-Score of 96.84%. Furthermore, our method achieved an accuracy of 97.73%, which is the second-highest among the compared studies, although the highest accuracy score was not reported for all studies.

Overall, our proposed method consistently demonstrates strong performance across multiple quantitative metrics, positioning it as a competitive approach for building extraction from high-resolution aerial imagery.

Moreover, the qualitative comparison confirms the obtained results. Through visual examination of our method compared to method [29] which used a spatial attention module only, it is evident that our approach successfully extracts buildings with remarkable precision and consistency. The building boundaries are well-defined and accurate, and the structural details are preserved.

Additionally, our method demonstrates the ability to identify buildings of various shapes, sizes, and orientations, including those that other methods, such as Hosseinpoor and Samadzadegan [29] failed to detect, as indicated by the blue squares in the results. Furthermore, our approach excels in detecting buildings with straight lines, as depicted by the yellow circles in Fig.6. In contrast, the attention-based model proposed by Hosseinpoor and Samadzadegan [29] struggles to accurately capture the precise outlines of these straight buildings. This highlights the superiority of our approach, leveraging the AttHT-IHT attention module, in accurately extracting straight building lines. However, it's important to note that there are certain limitations in our approach compared to that of Hosseinpoor and Samadzadegan [29], as indicated by the red circles in the results.

V. DISCUSSION

In this section, we analyze our proposed architecture, examining its strengths, potential limitations, and key design

TABLE 2. Potsdam dataset quantitative results compared to those of other authors. The best value is in bold, and the second-best value is underlined.

Methods	Year & Venue	IoU	Precision	Recall	F1-score	Accuracy
Xu *et al.* [19]	2018 Remote sensing	-	94.71	82.51	88.19	94.25
Yang *et al.* [27]	2018	86.71	95.21	90.66	92.56	96.16
Fu *et al.* [28]	2019 CVPR	88.19	95.63	94.30	94.97	-
Hosseinpoor *et al.* [29]	2019 ISPRS	-	95.70	95.90	95.80	-
Ye *et al.* [30]	2019 Remote sensing	90.02	-	-	94.75	**98.84**
Barnet [33]	2021 MDPI	<u>92.24</u>	**98.64**	95.12	**96.84**	-
Xie *et al.* [48]	2023 Elsvier	90.25	-	-	94.88	97.59
Ours	2023	**92.88**	<u>96.14</u>	**96.42**	<u>96.28</u>	<u>97.73</u>

decisions. We explore how the foundational U-Net model, along with the strategic placement of the AttHT-IHT module, interacts and impacts building segmentation effectiveness. Additionally, we address the generalization ability of our model.

A. RESIDUAL CONNECTIONS IN ATTHT-IHT BLOCK: TO USE OR NOT?

In the context of employing U-Net as the foundational model for our architecture, the primary forward stream of U-Net processes image information through encoding and decoding stages. This sequential flow ensures the preservation of original features without alteration. Notably, U-Net incorporates skip connections to facilitate the recovery and integration of low-level features with the original feature maps. Our methodology builds upon this foundation by selectively preserving the most relevant features using AttHT-IHT. These preserved features are then merged with the decoded features, contributing to an enhanced overall representation.

Given the Hough transform's exclusive application to deep-level features, our approach deems it unnecessary to reintroduce features before the Hough block.

B. PLACEMENT OF ATTHT-IHT

In our approach, the integration of the Differentiable Hough Transform (DHT) into the AttHT-IHT at a specific deep feature scale is a strategic decision driven by both efficiency considerations and alignment with previously successful applications [37], [38]. Applying AttHT-IHT at this level is intended to amplify and emphasize the critical lines essential for our primary objective of building segmentation. This deliberate focus helps distinguish these significant lines from less relevant ones that may emerge due to low-level feature noise. Furthermore, to enhance the Hough Transform's performance in identifying key line features, we introduced an attention mechanism within the Hough domain. This mechanism selectively concentrates on line features most relevant to our task, thereby boosting the overall effectiveness of our approach. Additionally, the spatial proximity of the Hough block to the original features during the encoding process within the U-Net architecture ensures the effective retention of essential image information in our forward stream.

TABLE 3. Evaluation Results on the Potsdam dataset with and without AttHT-IHT block.

Methods	F1-score	OA	IoU
Without AttHT-IHT	95.84	97.40	92.10
With AttHT-IHT	**96.28**	**97.73**	**92.88**

TABLE 4. Evaluation Results on the Potsdam dataset with and without ASPP block.

Methods	F1-score	OA	IoU
Without ASPP	95.25	97.14	91.10
With ASPP	**96.28**	**97.73**	**92.88**

C. IMPACT OF ATTHT-IHT AND ASPP BLOCKS

In this section, we explore the pivotal impact of the AttHT-IHT and ASPP blocks on our segmentation model's performance, emphasizing their crucial roles in improving accuracy and overall effectiveness.

1) IMPACT OF ATTHT-IHT

To strengthen the credibility of our model's outcomes and emphasize the significance of the AttHT-IHT module, we conducted supplementary experiments. The ablation study on the Potsdam dataset, delineated in Table 3, meticulously dissects the quantitative performance with and without the AttHT-IHT module. The results showcase a noteworthy improvement, yielding an F1-score of 96.28%, an Overall Accuracy of 97.73%, and an IoU of 92.88% with the assimilation of AttHT-IHT. These outcomes emphatically underscore the substantial positive influence of integrating the AttHT-IHT module, accentuating its contribution to the enhancement of building segmentation performance.

2) IMPACT OF ASPP MODULE

Our choice to adopt the U-Net architecture enhanced with ASPP as the foundational model is based on current research trends and advanced methodologies, highlighting its consistently superior performance, as demonstrated in previous studies [22]. Furthermore, to scrutinize the influence of ASPP in our approach, we undertake a comparative analysis between our method with and without ASPP. The outcomes reveal that the incorporation of ASPP results in elevated overall accuracy, F1-score, and IoU on the Potsdam dataset, showcasing its impact on model performance, as illustrated in Table 4.

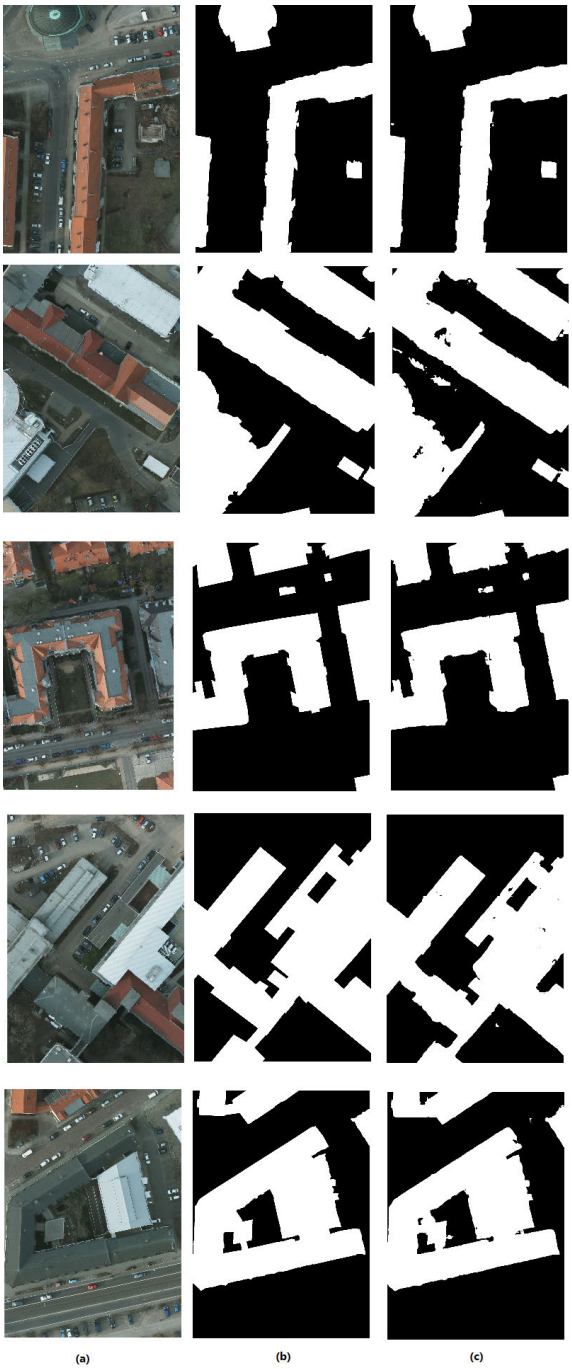

FIGURE 5. Building extraction maps obtained from the Potsdam dataset (a) Original image. (b) Ground truth.(c) Our proposed model.

D. GENERALIZATION PERFORMANCE ASSESSMENT

In the assessment of generalization performance, we utilized the WHU Satellite Dataset I. Table 5 presents the performance of our model across various metrics. Specifically, our model achieved outstanding scores on the WHU Satellite Dataset I, with an Overall Accuracy of 92.13%, IoU of 73.08%, F1-score of 84.32%, Recall of 82.80%, and Precision of 85.91%.

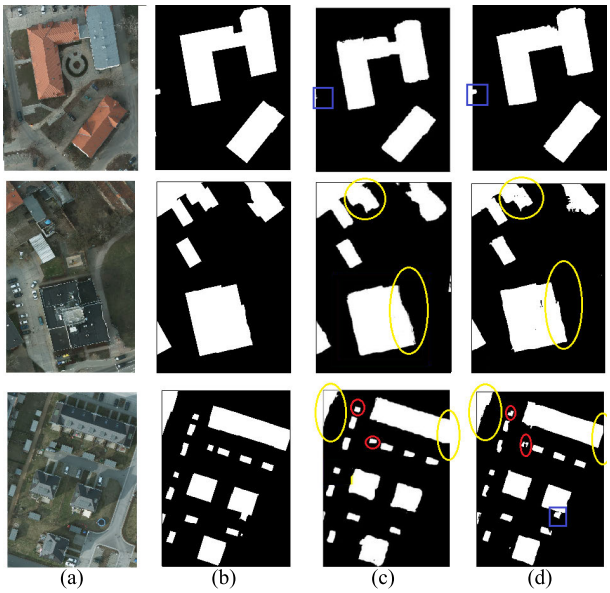

FIGURE 6. Building Extraction Maps: Comparative Analysis. (a) Original image. (b) Ground truth. (c) Extracted building map from [29] model. (d) Proposed model.

TABLE 5. Results on the WHU satelllite Dataset I.

Methods	OA	IoU	F1-score	Recall	Precision
SSDBN [36]	-	72.00	68.41	-	-
Ours	**92.13**	**73.08**	**84.32**	**82.80**	**85.91**

Furthermore, a direct comparison to SSDBN [36] in Table 5 reveals our model's superior performance, particularly in IoU and F1-score metrics. Notably, our model exhibited a significant 15.9% increase in F1-score compared to SSDBN [36], underscoring positive implications for its generalization ability.

E. LIMITATIONS

While our approach prioritizes overall accuracy, focusing on regional building characteristics, the decision to exclude specific contour-related metrics was made thoughtfully. This choice is driven by the primary goal of our application, constraints within our experimental framework aligned with state-of-the-art methodologies, and the lack of sufficient data in our chosen datasets for a detailed contour analysis. However, the current lack of such metrics in our assessment underscores a limitation to be addressed in our future perspectives.

VI. CONCLUSION

In this article, we present a CNN-based method for semantic segmentation of remote sensing images. Our architecture incorporates specialized modules to tackle challenges concerning global context and boundary intricacies. Among these modules is AttHT-IHT module, which effectively enhances the quality of extracted results by isolating pertinent straight lines that significantly contribute to defining building

shapes. By integrating the AttHT-IHT modules into our model, we achieve a fusion of global line priors with locally learned image features. This fusion substantially bolsters the architecture's ability to identify straight lines associated with buildings in aerial images. Additionally, we incorporate the ASPP module into the U-Net architecture, allowing us to capture multiscale features and refine classification accuracy. The Encoder-Decoder network demonstrates adeptness in both restoring image resolution and adeptly handling segmentation intricacies. However, the current approach for building extraction using RGB imagery does not capitalize on additional types of information, such as multispectral data and digital surface models. In future research, we will explore efficient methods to incorporate this extra information into deep learning models, aiming to enhance building extraction. Furthermore, the utilization of more advanced deep learning architectures like FCN and Mask R-CNN holds the potential for further improvements in building extraction and scene.

VII. ACKNOWLEDGMENT

The authors would like to thank the comments and constructive suggestions provided by the reviewers and also would like to thank the time and expertise devoted to evaluating their work. S. Yahia Berrouiguet would like to thank Tliba Marouane for his help and pertinent advice and Chatgpt 3.5 who helped in the improvement of English.

REFERENCES

[1] S. Cho, H. Xiu, and M. Matsuoka, "Backscattering characteristics of SAR images in damaged buildings due to the 2016 Kumamoto earthquake," *Remote Sens.*, vol. 15, no. 8, p. 2181, Apr. 2023.

[2] Y. Su, L. Gao, M. Jiang, A. Plaza, X. Sun, and B. Zhang, "NSCKL: Normalized spectral clustering with kernel-based learning for semisupervised hyperspectral image classification," *IEEE Trans. Cybern.*, vol. 53, pp. 1–14, 2022.

[3] D. Domingo, J. van Vliet, and A. M. Hersperger, "Long-term changes in 3D urban form in four Spanish cities," *Landscape Urban Planning*, vol. 230, Feb. 2023, Art. no. 104624.

[4] J. Chen, M. Xia, D. Wang, and H. Lin, "Double branch parallel network for segmentation of buildings and waters in remote sensing images," *Remote Sens.*, vol. 15, no. 6, p. 1536, Mar. 2023.

[5] M. Awrangjeb, X. Hu, B. Yang, and J. Tian, "Editorial for special issue: 'Remote sensing based building extraction,'" *Remote Sens.*, vol. 15, no. 4, p. 998, 2020.

[6] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[7] A. Sharma, X. Liu, X. Yang, and D. Shi, "A patch-based convolutional neural network for remote sensing image classification," *Neural Netw.*, vol. 95, pp. 19–28, Nov. 2017.

[8] X. Pan, L. Gao, A. Marinoni, B. Zhang, F. Yang, and P. Gamba, "Semantic labeling of high resolution aerial imagery and LiDAR data with fine segmentation network," *Remote Sens.*, vol. 10, no. 5, p. 743, May 2018.

[9] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[10] L. Zhang, D. Zhang, and F. Tian, "SVM and ELM: Who wins? Object recognition with deep convolutional features from ImageNet," 2016, *arXiv:1506.02509*.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[12] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," 2016, *arXiv:1606.02585*.

[13] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution semantic labeling with convolutional neural networks," 2016, *arXiv:1611.01962*.

[14] 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. 76–85.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.

[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[17] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[18] A. Khalel and M. El-Saban, "Automatic pixelwise object labeling for aerial imagery using stacked U-nets," 2018, *arXiv:1803.04953*.

[19] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sens.*, vol. 10, no. 1, p. 144, Jan. 2018.

[20] Y. Liu, J. Zhou, W. Qi, X. Li, L. Gross, Q. Shao, Z. Zhao, L. Ni, X. Fan, and Z. Li, "ARC-net: An efficient network for building extraction from high-resolution aerial images," *IEEE Access*, vol. 8, pp. 154997–155010, 2020.

[21] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet—A: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[22] H. Wang and F. Miao, "Building extraction from remote sensing images using deep residual U-net," *Eur. J. Remote Sens.*, vol. 55, no. 1, pp. 71–85, Dec. 2022.

[23] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1–9.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

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

[26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[27] H. Yang, P. Wu, X. Yao, Y. Wu, B. Wang, and Y. Xu, "Building extraction in very high resolution imagery by dense-attention networks," *Remote Sens.*, vol. 10, no. 11, p. 1768, Nov. 2018.

[28] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.

[29] H. R. Hosseinpoor and F. Samadzadegan, "Attention based convolutional neural network for building extraction from very high resolution remote sensing image," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 507–512, 2019.

[30] Z. Ye, Y. Fu, M. Gan, J. Deng, A. Comber, and K. Wang, "Building extraction from very high resolution aerial imagery using joint attention deep neural network," *Remote Sens.*, vol. 11, no. 24, p. 2970, Dec. 2019.

[31] W. Shi, W. Qin, Z. Yun, P. Ping, K. Wu, and Y. Qu, "Attention-based context aware network for semantic comprehension of aerial scenery," *Sensors*, vol. 21, no. 6, p. 1983, Mar. 2021.

[32] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[33] Y. Jin, W. Xu, C. Zhang, X. Luo, and H. Jia, "Boundary-aware refined network for automatic building extraction in very high-resolution urban aerial images," *Remote Sens.*, vol. 13, no. 4, p. 692, Feb. 2021.

[34] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 905–909, May 2021.
[35] J. Liu, H. Huang, H. Sun, Z. Wu, and R. Luo, "LRAD-net: An improved lightweight network for building extraction from remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 675–687, 2023.
[36] Y. Li, H. Lu, Q. Liu, Y. Zhang, and X. Liu, "SSDBN: A single-side dual-branch network with encoder–decoder for building extraction," *Remote Sens.*, vol. 14, no. 3, p. 768, Feb. 2022.
[37] Y. Lin, S. L. Pintea, and J. C. van Gemert, "Deep Hough-transform line priors," 2020, *arXiv:2007.09493*.
[38] Y. Lin, S.-L. Pintea, and J. van Gemert, "Semi-supervised lane detection with deep Hough transform," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1514–1518.
[39] K. Zhao, Q. Han, C.-B. Zhang, J. Xu, and M.-M. Cheng, "Deep Hough transform for semantic line detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4793–4806, Sep. 2022.
[40] W. Zhao, C. Persello, and A. Stein, "End-to-end roofline extraction from very-high-resolution remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. IGARSS*, Jul. 2021, pp. 2783–2786.
[41] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested U-net architecture for medical image segmentation," 2018, *arXiv:1807.10165*.
[42] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected Unet for medical image segmentation," 2020, *arXiv:2004.08790*.
[43] *ISPRS 2D Potsdam Semantic Benchmark Dataset*. Accessed: May 2021. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/datarequest-form2.html
[44] 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. 2019.
[45] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.
[46] A. Paszke et al., (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. [Online]. Available: https://pytorch.org
[47] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR'15)*, San Diego, CA, USA, 2015, p. 500.
[48] Y. Xie, J. Tian, and X. X. Zhu, "A co-learning method to utilize optical images and photogrammetric point clouds for building extraction," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 116, Feb. 2023, Art. no. 103165.

SOUAD YAHIA BERROUIGUET is currently pursuing the Ph.D. degree in signal and image processing with the Electronics Department, University Djillali Liabes of Sidi Bel Abbes, Algeria. Additionally, she is also a Teacher with the National Higher School of Telecommunication and I. C. T. Abdelhafid Boussouf, Algeria. Her research interests include remote sensing and deep learning.

EHLEM ZIGH received the Habilitation degree in electronic from University Djillali Liabes of Sidi Bel Abbes, Belabes, Algeria, in 2017, and the Ph.D. degree from the University of Sciences and Technology of Oran—Mohammed Boudiaf, Algeria, in 2014. She is currently a Professor with the University of Science and Technology of Oran—Mohamed Boudiaf. She is also a member of Laboratoire de Codage et de la Sécurité de l'information, Mohamed Boudiaf University. She has around 20 national and international communications and 11 international publications. She has published a book chapter in a *Handbook of Research on Artificial Intelligence Techniques and Algorithms* (Malaysia). She is a Reviewer of the *International Journal of Energy Optimization and Engineering*, *IGI Global*, *European Journal of Remote Sensing*, and *Interactive Learning Environment* journal.

MOHAMMED DJEBOURI has been a Professor with the Department of Electronics, University Djillali Liabes of Sidi Bel Abbes, since 1991. His research interests include digital signal processing, imagery, embedded systems, mobile satellite communications, and GNSS.