remote sensing

MDPI

Article

Urban Sprawl Monitoring by VHR Images Using Active Contour Loss and Improved U-Net with Mix Transformer Encoders

Miguel Chicchon [1], Francesca Colosi [2], Eva Savina Malinverni [3] and Francisco James León Trujillo [4,*]

[1] Facultad de Ingeniería, Universidad Tecnológica del Perú, Lima 150101, Peru; c11201@utp.edu.pe
[2] Istituto di Scienze del Patrimonio Culturale, Consiglio Nazionale delle Ricerche, Area della Ricerca Roma 1, Via Salaria km 29,300, 00016 Rome, Italy; francesca.colosi@cnr.it
[3] Dipartimento di Ingegneria Civile, Edile e Architettura, Università Politecnica delle Marche, Via Brecce Bianche, 12, 60131 Ancona, Italy; e.s.malinverni@staff.univpm.it
[4] Generales Ciencias, Universidad Continental, Cusco 08000, Peru
* Correspondence: fleont@continental.edu.pe or francisco.leon2903@gmail.com

Abstract: Monitoring the variation of urban expansion is crucial for sustainable urban planning and cultural heritage management. This paper proposes an approach for the semantic segmentation of very-high-resolution (VHR) satellite imagery to detect the changes in urban sprawl in the surroundings of Chan Chan, a UNESCO World Heritage Site in Peru. This study explores the effectiveness of combining Mix Transformer encoders with U-Net architectures to improve feature extraction and spatial context understanding in VHR satellite imagery. The integration of active contour loss functions further enhances the model's ability to delineate complex urban boundaries, addressing the challenges posed by the heterogeneous landscape surrounding the archaeological complex of Chan Chan. The results demonstrate that the proposed approach achieves accurate semantic segmentation on images of the study area from different years. Quantitative results showed that the U-Net-scse model with an MiTB5 encoder achieved the best performance with respect to SegFormer and FT-UNet-Former, with IoU scores of 0.8288 on OpenEarthMap and 0.6743 on Chan Chan images. Qualitative analysis revealed the model's effectiveness in segmenting buildings across diverse urban and rural environments in Peru. Utilizing this approach for monitoring urban expansion over time can enable managers to make informed decisions aimed at preserving cultural heritage and promoting sustainable urban development.

Keywords: building footprints; neural network; semantic segmentation; OpenEarthMap dataset; transformers; Chan Chan

Academic Editor: Tania Stathaki

Received: 8 February 2025
Revised: 17 March 2025
Accepted: 26 March 2025
Published: 30 April 2025

Citation: Chicchon, M.; Colosi, F.; Malinverni, E.S.; León Trujillo, F.J. Urban Sprawl Monitoring by VHR Images Using Active Contour Loss and Improved U-Net with Mix Transformer Encoders. *Remote Sens.* **2025**, *17*, 1593. https://doi.org/10.3390/rs17091593

1. Introduction

In recent years, rapid urban growth has been observed, with urbanization projected to increase from 55% of the world's population in 2018 to 68% in 2050, according to United Nations estimates [1]. This contributes significantly to environmental degradation and poses serious problems, such as soil and riverbank erosion, caused by excessive land consumption, habitat loss, as well as climate change [2–4]. Another problem observed is an increase in the amount of damage to cultural heritage (CH) sites due to rapid urban growth around them. Historical sites and monuments cannot be viewed as isolated elements within a territory. Their significance, uniqueness, and integrity are intrinsically connected to the historical landscape in which they are found. Urbanization and the resulting destruction of the original landscapes lead to the rapid loss of the historical and symbolic meaning of these cultural assets along with their inherent beauty. This process leaves them disconnected

from their context, rendering them isolated within a historical and cultural environment to which they do not belong [5–7].

It is therefore desirable to develop a new methodology that is capable of providing the greatest possible amount of information for monitoring the environment and for the construction of a model for the prevention and prediction of environmental risk. In recent decades, the application of remote sensing (RS) and the geographic information system (GIS) in the field of CH has demonstrated that both are powerful tools for the collection, analysis, management, and digitization of spatial data, both in terms of quantity and of scope. The purposes of these applications are specific to or included in the monitoring of CH, particularly the detection of changes in land use within or in the vicinity of archaeological sites, including urban centers. In this sense, through different RS methods, it is possible to detect significant changes related to the phenomenon of urban sprawl and construction projects, agricultural or temporary activities, among others, all related to the alteration of the cultural landscape [5,8–13].

The use of artificial intelligence (AI) is growing across scientific disciplines, helping to integrate massive datasets, refine measurements, guide experimentation, explore data-compatible theories, and provide processable and reliable models integrated with scientific workflows for autonomous discovery. The discoveries and research efforts on land cover change detection have benefited from the significant integration of AI and satellite RS. Since 2016, automatic methods, i.e., machine learning and classification techniques, have been using very-high-resolution (VHR) multispectral satellite imagery, as well as VHR synthetic aperture radar (SAR), for the purpose of archaeological cultural landscape prospection, analysis, and monitoring processes [14–18]. The emergence of deep learning (DL) in the early 2010s greatly expanded the scope and ambition of land use and land cover (LULC) mapping processes, the first step in monitoring urban sprawl [19–22].

Among deep learning network architectures, convolutional neural networks (CNNs) have achieved good performance in many computer vision (CV) tasks for Earth observations [23]. For the purpose of building extractions, the use of CNNs has been positive because it generates end-to-end solutions for many computer vision tasks, such as object detection [23–25], semantic segmentation [26–31], and instance segmentation [32–35]. In this case, semantic segmentation uses DL architectures based on encoder–decoder architectures to classify each pixel of an input image into distinct classes. To extract building footprints from VHR satellite images, semantic segmentation architectures such as SegNet [36], U-Net [37], UNet++ [38], PSPNet [39], FPN [40], and DeepLab V3+ [41] have been used with several backbones. In particular, the U-Net architecture was modified, for example, in Tiramisu [42], a combination of the U-Net structure and DenseNet [43] building blocks, as well as in UNet++, where nested dense building blocks perform convolution on the skip connections. In [44], three U-Nets consisting of normal convolution blocks, residual units, and inception modules, named Normal-U-Net, Residual-U-Net, and Inception-U-Net, respectively, were designed, trained, and validated on the Massachusetts building dataset, and it was shown that the Inception-U-Net was the most competitive architecture comprehensively. In addition, the three single U-Nets were combined into a stronger model with remarkable performance called EU-Net. In [29], the DeepLab V3+ model, using the Xception backbone, showed more efficiency with respect to the U-Net, FPN, and PSPNet models. In [45], U-Net, UNet++, DeepLab V3+, FPN, and PSPNet architectures, using an SE-ResNeXt101 encoder, were pre-trained with ImageNet [46], and the best solutions were the U-Net and UNet++ architectures. Then, both architectures were trained on the Istanbul dataset, using the same encoder, and UNet++ outperformed U-Net in all accuracy metrics except recall. An ensemble DL model named Seg-Unet, a combination of U-Net and SegNet techniques, was presented in [47] to extract building footprints from a public

dataset (Massachusetts building dataset), obtaining an overall average accuracy of 92.73% and demonstrating improved performance compared to fully convolutional neural network (FCN), SegNet, and U-Net models.

Extracting complete information from a global scene from RS data with complex backgrounds is still a challenge. Recently, Vision Transformers (ViTs) are overcoming the limitations of CNN-based models associated with capturing global contextual information, and they present a higher level of data representation capabilities in, for example, tasks such as image recognition [48,49], semantic segmentation [50,51], and object detection [52,53]. Nevertheless, ViTs have some drawbacks compared to CNNs, such as large memory requirements and the need for large datasets [47]. Self-supervised approaches can alleviate some of these drawbacks and further improve ViTs [54,55]. In [56], 10 CNN and Transformer models were generated, and comparisons were made. The proposed Residual-Inception U-Net (RIU-Net), as well as U-Net, Residual U-Net, Attention Residual U-Net, and four CNN architectures (Inception, Inception-ResNet, Xception, and MobileNet) were implemented as encoders to U-Net-based models, and two Transformer-based approaches (Trans U-Net and Swin U-Net) were also used, all trained and evaluated on the Massachusetts buildings dataset and Inria aerial image labeling dataset, with results showing significant success of RIU-Net. In [57], a Transformer-based decoder and a UNet-type Transformer (UNetFormer) were proposed for real-time urban scene segmentation. The UNetFormer, with the lightweight ResNet18 as an encoder, developed an efficient global–local attention mechanism to model both global and local information at the decoder, showing that it not only performed faster, but also produced higher accuracy compared to state-of-the-art lightweight models. In a more detailed analysis, the proposed Transformer-based decoder combined with a Swin Transformer encoder also achieved the best results (91.3% F1 and 84.1% mIoU) on the Vaihingen dataset.

In general, deep neural networks may encounter challenges in predicting precise semantic labels around object boundaries. Traditional methods such as active contours enable accurate delineation of object boundaries; however, they are sensitive to initialization [58]. In recent years, research has been presented wherein integration allows for leveraging the strengths of both methods [59,60]. For example, level set methods are formulated in a self-supervised way by minimizing energy functions such as the Mumford–Shah functional, being still useful in generating label-free segmentation masks. The features of buildings are unique, having different shapes. When affected by light in VHR images, artificial surface features are difficult to extract by employing traditional image segmentation methods. Manual methods are laborious, and deep learning-based approaches fail to delineate all construction features and to do so with good accuracy. In [61], trainable deep active contours (TDAC), an automatic image segmentation framework that intimately links convolutional neural networks (CNNs) and active contour models (ACMs), was proposed. TDAC provides fast, accurate, and fully automatic simultaneous delineation of an arbitrary number of buildings in the image. In [62], an active contour model (ACM) based on a convolutional neural network (CNN) was proposed, integrating prior knowledge and constraints of the ACM, such as boundary continuity, smooth edges, and geometric features of buildings into the learning process of the CNN to achieve tight unity with the ACM, and it proved to have good performance.

This study aims to segment buildings in very-high-resolution (VHR) satellite imagery utilizing an improved U-Net architecture and active contour loss to detect changes in urban sprawl in the vicinity of Chan Chan, Peru. The structure of this article is as follows: The first section provides a concise introduction to the problem and relevant literature. The second section delineates the methods employed, encompassing a brief overview of the study area, dataset presentation, description of the network architecture for semantic

segmentation, and experimental design. The third section presents the results, focusing on segmentations in images from various Peruvian cities. The fourth section comprises the discussion, concentrating on the results of building segmentation and an analysis of urban expansion in three areas around Chan Chan. The final section offers conclusions and proposes future research directions.

2. Methods

2.1. Study Area

The archaeological complex of Chan Chan, the capital of the pre-Inca Chimú kingdom, is located along the northern coast of Peru, near the city of Trujillo, a center with strong economic and urban growth (Figure 1). Chan Chan is the largest adobe settlement in Latin America (approximately 20 square kilometers) and has been listed on the UNESCO World Heritage List since 1986 [63–66]. Unfortunately, the fragility of the construction material and the threats compromising its integrity have led to the site's inclusion on the List of World Heritage in Danger [67]. This prompted the Ministry of Culture of Peru to draft the Plan de Manejo Complejo Arqueológico Chan Chan (Plan Maestro), approved by the Peruvian government in 2000 and updated for the decade 2021–2031 [68]. The Plan Maestro includes a series of actions aimed at protecting the monuments from environmental threats that affect their walls, such as sea salts, the corrosive action of the wind, and torrential rains caused by the ENSO phenomenon. However, it is less effective in addressing the primary danger to Chan Chan, which is the rapid and uncontrolled growth of the nearby urban center of Trujillo.

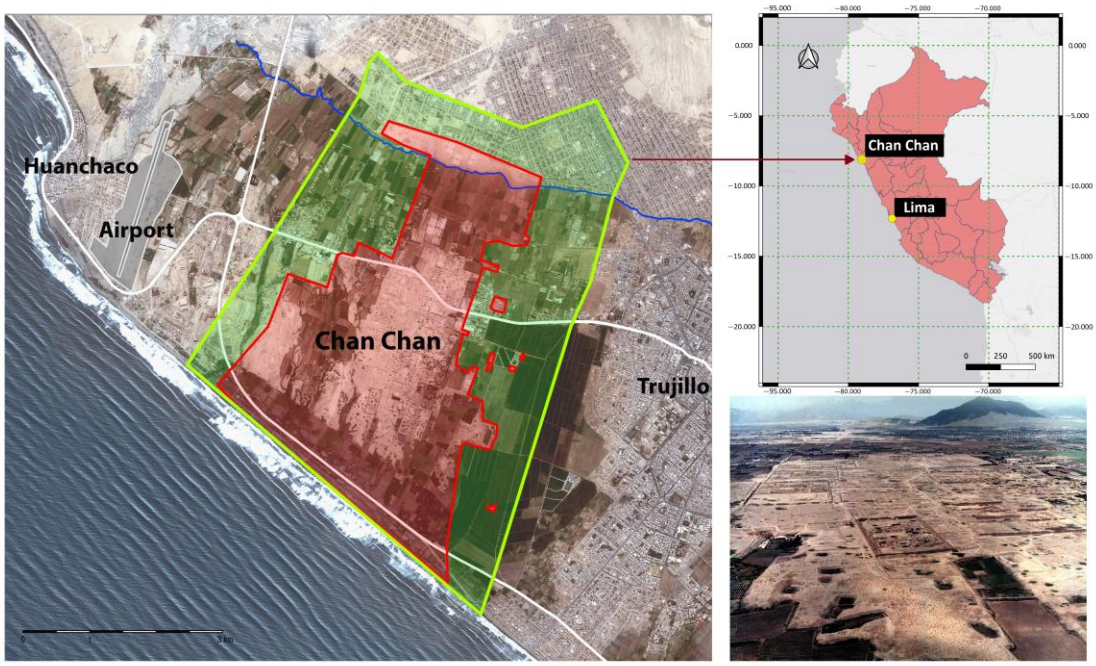

Figure 1. In the lower right-hand corner, the study area covering the surrounding archaeological area of Chan Chan, Trujillo, Peru. On the satellite image, the yellow polygon is Chan Chan's buffer zone, while the red polygon is its intangible zone.

Since 2002, the Italian Mission in Peru (MIPE), led by the National Research Council—Institute of Heritage Science (CNR-ISPC), has been operating in Chan Chan in partnership with Università Politecnica delle Marche. With the scientific and operational support of the Ministry of Culture of Peru, the mission collaborates closely on the realization of the Plan Maestro, working in direct coordination with Proyecto Especial Complejo

Arqueológico Chan Chan (PECACH), the Ministry's specialized body responsible for the site's preservation and conservation. MIPE focuses on documenting and studying adobe architecture, conserving the urban center and its landscape, and promoting the archaeological site within the international community [69–73]. As part of the mission's operational activities, an area of respect was defined, with its boundaries approved by the Ministry of Culture in 2010 [74]. Unfortunately, this buffer zone lacks specific legislative regulation and is generally included in Trujillo's urban development plan.

In the frame of the CNR-CONCYTEC bilateral agreement 2021–2022, a joint research project was conducted, focused on diagnosing the buffer zone using RS techniques and direct field research. With support from the National Center for Disaster Risk Estimation, Prevention, and Reduction (CENEPRED), the research team investigated the buffer zone, documenting the extent and effects of urbanization on the UNESCO heritage archaeological complex. The resulting picture is devastating, marked by the constant loss of Chimú monuments and infrastructures and the consequent rapid disappearance of the cultural landscape [5,75].

There is a clear urgency for a methodological procedure to monitor urban growth in the buffer zone periodically. This would provide local operators with the necessary data to plan sustainable urban development that respects the exceptional universal value of Chan Chan, thereby preventing the complex from being removed from the World Heritage List.

2.2. Dataset

The experimental phase of this study employed two distinct datasets. The primary dataset, designed for change analysis and semantic segmentation model testing within the Chan Chan region, consisted of 50 georeferenced images. These images were extracted from Google Earth Pro, spanning the period from 2017 to 2023, and each measured 512 × 512 pixels. This dataset's temporal range allowed for the observation of land cover changes over a six-year period. The semantic relationship being investigated was the presence and change of building footprints within the Chan Chan area.

For training the semantic segmentation models, the OpenEarthMap dataset [76] was selected due to its comprehensive global land cover mapping capabilities and high spatial resolution. This dataset provided 5000 aerial and satellite images, characterized by a ground sampling distance ranging from 0.25 to 0.5 m. The images covered 97 diverse regions across 44 countries on six continents, including ten cities within Peru. The original eight land cover classes within the OpenEarthMap dataset were reclassified into a simplified binary schema: 'building footprints' and 'background'. This reclassification facilitated the model's focus on accurately delineating built structures, representing a specific semantic category, from the surrounding environment. The dataset was partitioned into training and validation subsets, with 80% of the images allocated for model training and 20% reserved for validation, ensuring robust model performance evaluation.

2.3. Network Architecture

In semantic segmentation, each pixel within an image is categorized and assigned a specific class label, effectively performing detailed classification of the entire visual scene. In this study, U-Net was employed as the base architecture to generate segmentation models. U-Net [37] is an encoder–decoder architecture based on convolutional neural networks (CNN). UNet, with its symmetrical U-shaped structure, extracts feature maps from an input image through an encoding stage, see Figure 2. In the decoding stage, the dimensions of the original image were restored. The encoder output was upsampled and merged with the feature maps propagated via skip connections from the encoder layers. In each decoder layer, scSE (simultaneous spatial and channel squeeze and excitation) blocks [77]

are integrated to restore the spatial features lost during the encoding process. The cSE blocks recalibrate the channels by integrating global spatial information. In parallel, the sSE blocks generate a spatial attention map, indicating the specific regions where the network should direct its focus to improve segmentation. Figure 3 shows the structure of a decoder layer. The scSE attentional blocks (green) are incorporated after each convolutional block (blue) in layers B-2 to B-5 of U-Net.

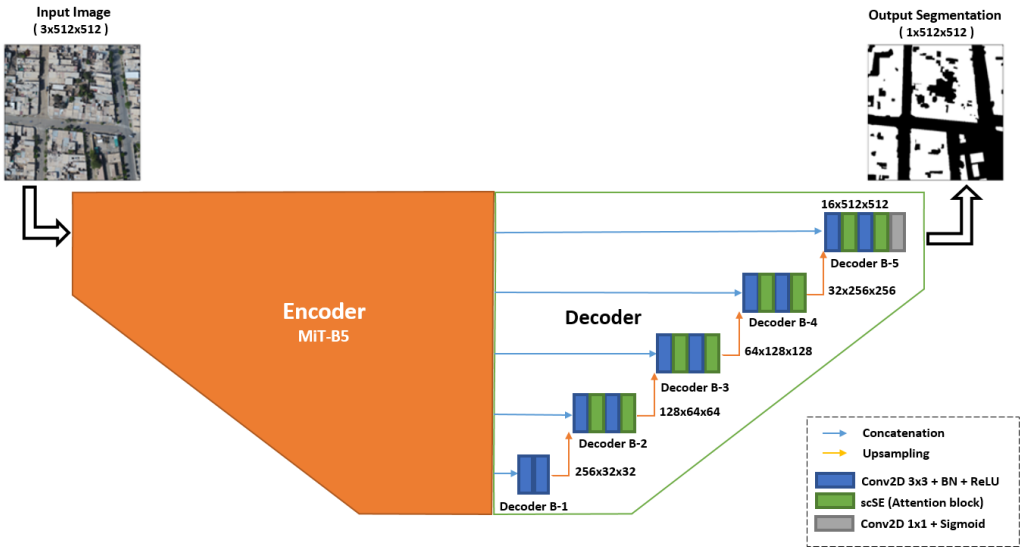

Figure 2. Improved U-Net architecture.

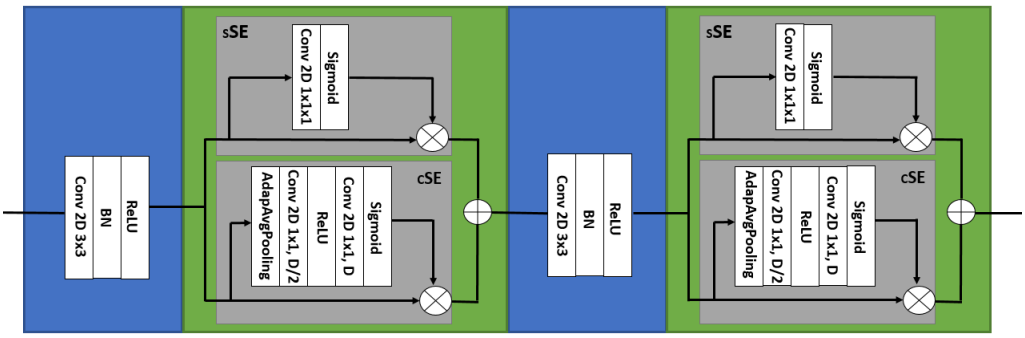

Figure 3. Decoder layer architecture.

In the U-Net encoder, EfficientNet [78] and a Mix Transformer (MiT) were utilized as the backbone for the base architecture. EfficientNet is a family of models designed to search for an optimal and computationally efficient neural architecture for classification tasks. The primary component of these models is the mobile inverted bottleneck (MBConv), which is responsible for the model compression and optimization of neuron excitation. The initial version, comprising eight models (B0-B7), was subsequently updated with four models (S, M, L, and XL) in a second version that incorporates Fused-MBConv blocks [79]. Conversely, the Mix Transformer serves as the backbone of the SegFormer segmentation architecture [80]. The Mix Transformer family of models (B0–B5) consists of hierarchically structured transformers designed to generate multiscale features.

The FT-U-NetFormer [57] architecture demonstrated favorable outcomes when evaluated on the OpenEarthMap dataset [63], obtaining 0.8029 of IoU for the building footprint class. This encoder–decoder architecture utilizes a Swin Transformer [50] as the encoder, leveraging its capacity to reduce the computational complexity of high-resolution images. FT-U-NetFormer introduces a global–local transformer block (GLTB) in the decoder to

model global and local information. DCSWIN [81], another architecture employed in land cover segmentation, similarly utilizes a Swin Transformer within its encoder. This architecture incorporates a densely connected feature aggregation module (DCFAM) in the decoder stage with the objective of capturing enhanced semantic features of multi-scale relationships to restore resolution and generate the final segmentation map.

2.4. Experimental Protocol and Setting

The objective of this study was to identify building footprints in very-high-resolution (VHR) satellite imagery utilizing deep semantic segmentation techniques. The methodology employs a public dataset to train and evaluate the performance of the model in identifying the building footprints. By applying this approach to the temporal satellite imagery of Chan Chan and its surroundings, researchers can analyze urban expansion patterns over time, providing valuable insights for CH preservation and sustainable urban planning efforts. Figure 4 illustrates the workflow. The initial step involved creating an image collection based on Google Earth Pro images spanning the period 2017 to 2023. Subsequently, the OpenEarthMap dataset was used for the network training. The improved U-Net was trained using a combined loss function. Finally, the test set was employed to analyze of urban sprawl in Chan Chan.

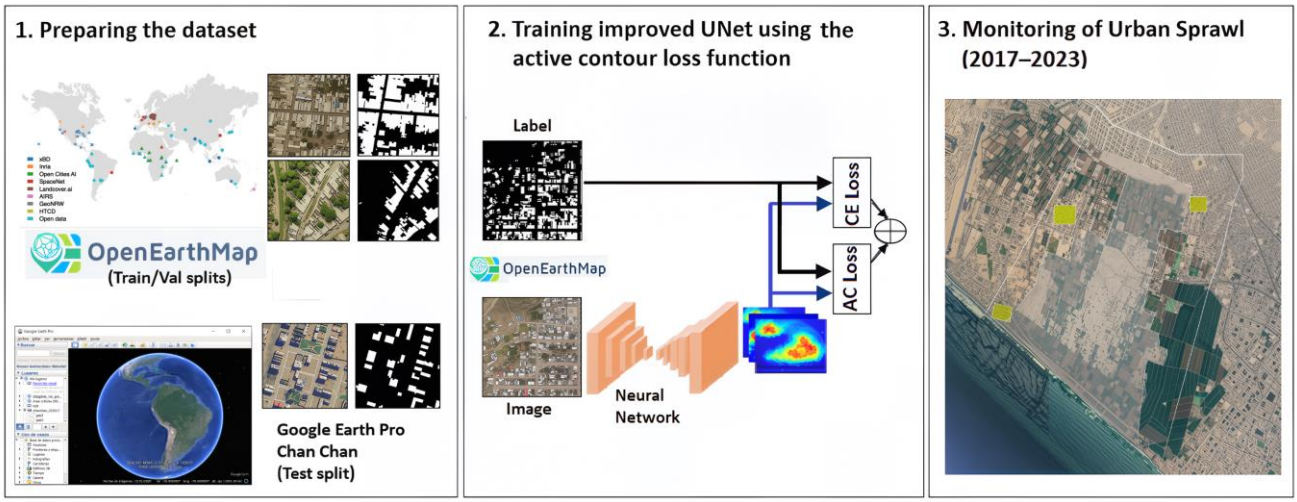

Figure 4. Workflow scheme.

The models implementations were based on the PyTorch 2.5.1 deep learning framework. All models were trained for 200 epochs on an NVIDIA RTX 4090 GPU, AMD EPYC 7282 16-Core Processor, and 128 GB of RAM. The Adam optimizer was employed, with the initial learning rate set to 0.001 for the decoder and 0.0001 for the backbone. Additionally, the ReduceLROnPlateau scheduler was used with a patience of 10 epochs and a reduction factor of 0.1. To mitigate overfitting, an early stopping strategy was implemented, with a patience of 20 epochs.

2.5. Data Augmentation

Several data augmentation techniques were employed in the training partition to enhance the training dataset.

- The input was subjected to various transformations with 50% probability for each operation. These transformations included a horizontal flip around the y-axis (HorizontalFlip), flip around the x-axis (VerticalFlip), transposition, and a 90-degree rotation;
- Four complementary strategies were implemented to obtain 512 × 512 pixel images: (1) 512 × 512 pixel resizing, (2) 512 × 512-pixel random crop, (3) cropping between

256 and 768 pixels of a random portion of the input, followed by 512 × 512 pixel resizing, and (4) 15° rotation, followed by 512 × 512 pixel center crop. For images with dimensions smaller than 768 pixels, padding was applied to the edges when the dimensions were below 512 pixels. In case (3), the random section was cropped with dimensions ranging from 256 to 512 pixels. Each of these strategies had an equal probability of being applied to the input data;

- Three complementary pixel intensity variation strategies with 25% probability of application were employed: (1) random brightness and contrast variations, (2) alterations in RGB color representations, and (3) modifications to hue and saturation values.

2.6. Loss Function

Active contour models entail the fitting of a curve around an object within an image subject to specific constraints. These models dynamically adjust the shape of the curve and are designed to terminate at the edges of an object. We utilize the loss function presented in [82] and defined by

$$L_{AC} = \mu \times \text{Length} + \lambda_1 \times \text{Area\_in} + \lambda_2 \times \text{Area\_out}, \quad (1)$$

where the initial term corresponds to the length of the evolving curve, and the subsequent terms represent the weights of the pixels situated within and outside the curve of the class under consideration. The weighting parameters were set to $\mu = 0.01$, $\lambda_1 = 1$, and $\lambda_2 = 1$.

To enhance the network's ability to learn more discriminative features of objects, the active contour loss function (L_{AC}) was integrated with the cross-entropy loss function (L_{CE}) as follows:

$$\text{Loss} = a \times L_{CE} + b \times L_{AC}, \quad (2)$$

where a and b are weighting parameters.

2.7. Metrics

In the context of image segmentation, intersection over union (IoU) and the Dice coefficient are crucial metrics for evaluating the performance of models. The IoU quantifies the degree of similarity between the predicted result and the ground truth. The Dice coefficient, also known as the F1-score, measures the overlap between the predicted segmentation and ground truth but is generally considered less stringent than IoU [83]. These metrics are defined as follows:

$$\text{IoU} = TP/(TP + FP + FN), \quad (3)$$

$$\text{Dice} = 2 \times TP/(2 \times TP + FP + FN), \quad (4)$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

3. Results

In this section, the quantitative results of the proposed method based on the selected metrics are presented. Subsequently, a qualitative analysis of the urban footprint segmentation in Peruvian cities was conducted.

3.1. Quantitative Results

The performance of all the models was evaluated using the OpenEarthMap and Chan Chan datasets. For the improved U-Net (U-Net-scse) encoder, EfficientNet V1 (B0, B2, B4, B6, B7), EfficientNet (s, m, l, xl), and Mix Transformers (B0, B2, B5) were examined. The segmentation capabilities of these models were compared with U-Net as a baseline and the

state-of-the-art models SegFormer, FT-UNet-Former, and DCSWIN, which achieved the best results in building segmentation on the OpenEarthMap dataset [76]. The quantitative results calculated based on IoU and Dice metrics are presented in Table 1. For both metrics, the mean values and standard deviations (in parentheses) obtained from the evaluations of the five trained models for each network architecture are shown. It can be observed that U-Net obtained the lowest result, with an IoU of 0.7746 in OpenEarthMap and 0.4280 in Chan Chan, demonstrating generalization issues. Regarding U-Net-scse with the EfficientNet V1 encoder, the best results in OpenEarthMap were obtained for B6 and B7 with approximately 0.82 IoU, while in Chan Chan, B6 achieved the best result with approximately 0.58 IoU. For U-Net-scse with the EfficientNet V2 encoder, the best result in OpenEarthMap was obtained for S with an IoU of 0.8173 and in Chan Chan with an IoU of 0.5878. In the case of U-Net-scse with a Mix Transformer encoder, the best result in OpenEarthMap was obtained for B5 with an IoU of 0.8284 and in Chan Chan with an IoU of 0.66465. For SegFormer, the best result in OpenEarthMap was obtained for B5 with an IoU of 0.8190 and in Chan Chan, with an IoU of 0.6386. Regarding the models based on the Swin Transformer, FT-UNet-Former obtained better results in both OpenEarthMap and Chan Chan, with IoUs of 0.8217 and 0.6100, respectively.

Table 1. Quantitative comparison of various networks on the OpenEarthMap and Chan Chan datasets. The best results are indicated with bold.

Network	**OpenEarthMap**		**Chan Chan**	
Architecture	**IoU**	**Dice**	**IoU**	**Dice**
U-Net	$0.7746_{(0.0127)}$	$0.8730_{(0.0081)}$	$0.4280_{(0.0070)}$	$0.5994_{(0.0068)}$
U-Net-scse (EffV1B0)	$0.8066_{(0.0015)}$	$0.8929_{(0.0009)}$	$0.5748_{(0.0232)}$	$0.7318_{(0.0189)}$
U-Net-scse (EffV1B2)	$0.8105_{(0.0046)}$	$0.8954_{(0.0028)}$	$0.5587_{(0.0137)}$	$0.7168_{(0.0112)}$
U-Net-scse (EffV1B4)	$0.8153_{(0.0037)}$	$0.8982_{(0.0022)}$	$0.5709_{(0.0146)}$	$0.7267_{(0.0118)}$
U-Net-scse (EffV1B6)	$0.8202_{(0.0006)}$	$0.9012_{(0.0004)}$	$0.5797_{(0.0349)}$	$0.7334_{(0.0282)}$
U-Net-scse (EffV1B7)	$0.8207_{(0.0013)}$	$0.9015_{(0.0008)}$	$0.5679_{(0.0295)}$	$0.7241_{(0.0295)}$
U-Net-scse (EffV2s)	$0.8173_{(0.0022)}$	$0.8995_{(0.0014)}$	$0.5878_{(0.0076)}$	$0.7404_{(0.0060)}$
U-Net-scse (EffV2m)	$0.8141_{(0.0030)}$	$0.8975_{(0.0018)}$	$0.5758_{(0.0238)}$	$0.7306_{(0.0194)}$
U-Net-scse (EffV2l)	$0.8138_{(0.0036)}$	$0.8973_{(0.0022)}$	$0.5776_{(0.0175)}$	$0.7328_{(0.0134)}$
U-Net-scse (EffV2xl)	$0.8077_{(0.0074)}$	$0.8937_{(0.0046)}$	$0.5731_{(0.0240)}$	$0.7273_{(0.0197)}$
U-Net-scse (MiTB0)	$0.8064_{(0.0022)}$	$0.8929_{(0.0013)}$	$0.6081_{(0.0063)}$	$0.7563_{(0.0049)}$
U-Net-scse (MiTB2)	$0.8229_{(0.0023)}$	$0.9029_{(0.0014)}$	$0.6162_{(0.0086)}$	$0.7630_{(0.0071)}$
U-Net-scse (MiTB5)	$\mathbf{0.8284_{(0.0016)}}$	$\mathbf{0.9062_{(0.0010)}}$	$\mathbf{0.6465_{(0.0023)}}$	$\mathbf{0.7853_{(0.0017)}}$
SegFormer (B0)	$0.7934_{(0.0013)}$	$0.8848_{(0.0008)}$	$0.5837_{(0.0090)}$	$0.7371_{(0.0072)}$
SegFormer (B2)	$0.8172_{(0.0012)}$	$0.8994_{(0.0007)}$	$0.6270_{(0.0106)}$	$0.7707_{(0.0081)}$
SegFormer (B5)	$0.8190_{(0.0012)}$	$0.9005_{(0.0008)}$	$0.6386_{(0.0080)}$	$0.7794_{(0.0060)}$
FT-UNet-Former	$0.8217_{(0.0024)}$	$0.9021_{(0.0014)}$	$0.6100_{(0.0097)}$	$0.7577_{(0.0075)}$
DCSWIN	$0.8154_{(0.0042)}$	$0.8983_{(0.0026)}$	$0.5958_{(0.0265)}$	$0.7465_{(0.0210)}$

Table 2 presents the batch size utilized for each model, constrained by the memory capacity of the GPU employed in the experiment and the size of the model itself. This table also displays the total number of parameters and inference time for each model. Comparing the models with superior performance in OpenEarthMap and Chan Chan, U-Net-scse with Mix Transformer encoders yielded the most favorable results. Although the MiTB5 encoder

achieved the highest IoU, this encoder, in comparison to MiTB2, comprises three times the number of parameters and requires twice the inference time.

Table 2. Total number of parameters and inference time for models studied.

Network Architecture	BS	Total Parameters (M)	Inference Time (ms)
U-Net	8	31.0	11.38
U-Net-scse (EffV1B0)	16	6.3	4.29
UNet-scse (EffV1B2)	16	10.1	5.28
U-Net-scse (EffV1B4)	16	20.3	6.65
U-Net-scse (EffV1B6)	8	43.9	8.81
U-Net-scse (EffV1B7)	6	67.2	11.27
U-Net-scse (EffV2s)	16	22.1	5.71
U-Net-scse (EffV2m)	16	55.2	7.31
U-Net-scse (EffV2l)	8	119.9	9.43
U-Net-scse (EffV2xl)	5	209.6	11.45
U-Net-scse (MiTB0)	16	5.6	3.21
U-Net-scse (MiTB2)	16	27.6	5.42
U-Net-scse (MiTB5)	8	84.8	10.71
SegFormer (B0)	16	3.7	2.17
SegFormer (B2)	16	24.7	5.15
SegFormer (B5)	8	90.0	12.83
FT-UNet-Former	8	96.0	11.69
DCSWIN	8	118.9	10.83

Table 3 provides a detailed description of the investigation results. Network architecture models that performed the best on the OpenEarthMap and Chan Chan datasets are presented. These models were trained using a combined loss function based on the weighting of the cross-entropy loss function (L_{CE}) and active contour loss function (L_{AC}). The weighting parameters that best performed the combined loss function were a = 1 and b = 10. A UNet-scse model with an MiTB5 encoder achieved the optimal performance on OpenEarthMap, with IoU and Dice scores of 0.8288 and 0.9064, respectively, while UNet-scse with an MiTB5 encoder obtains superior results on Chan Chan, with IoU and Dice scores of 0.6743 and 0.8055. The UNet-scse model with an MiTB5 encoder demonstrates notable superiority over other methods in building extraction within the Chan Chan area, improving the IoU by 5.73% and 5.52% compared to the FT-UNet-Former and DCSWIN models, respectively.

Table 3. Quantitative comparison of the best models of each network architecture on the OpenEarthMap and Chan Chan datasets. The best results are indicated with bold.

Model	OpenEarthMap		Chan Chan	
	IoU	Dice	IoU	Dice
U-Net	0.7867	0.8806	0.4453	0.6162
U-Net-scse (EffV1B6)	0.8208	0.9016	06100	0.7578
U-Net-scse (EffV1B7)	0.8222	0.9024	0.5968	0.7475
U-Net-scse (EffV2s)	0.8194	0.9008	0.5944	0.7456
U-Net-scse (EffV2xl)	0.8160	0.8987	0.6038	0.7530
U-Net-scse (MiTB2)	0.8260	0.9047	0.6263	0.7702
U-Net-scse (MiTB5)	**0.8288**	**0.9064**	**0.6503**	**0.7881**
SegFormer (MiTB2)	0.8187	0.9003	0.6350	0.7768
SegFormer (MiTB5)	0.8203	0.9013	0.6468	0.7855
FT-UNet-Former	0.8244	0.9037	0.6150	0.7616
DCSWIN	0.8185	0.9002	0.6163	0.7626

Figure 5 presents box plots as a descriptive statistical representation of the models detailed in Table 3, evaluated using the OpenEarthMap and Chan Chan datasets. The median is depicted as an orange horizontal line within the box, while the first and third quartiles are represented as the lower and upper bounds of the box, respectively. The 1.5 interquartile range is indicated by black vertical lines, and outliers are denoted by black circles. The loss values of the models in Table 3 for the training and validation data are presented in Figure 6. The IoU values for the building footprint pixels compared to the training epoch are shown in Figure 7. Figure 6 shows that the U-Net-scse models with EffV1B6 and EffV1B7 encoders can achieve lower loss values, whereas the U-Net-scse (MiTB2), U-Net-scse (MiTB5), and FT-UNet-Former models obtain the highest IoU values in validation.

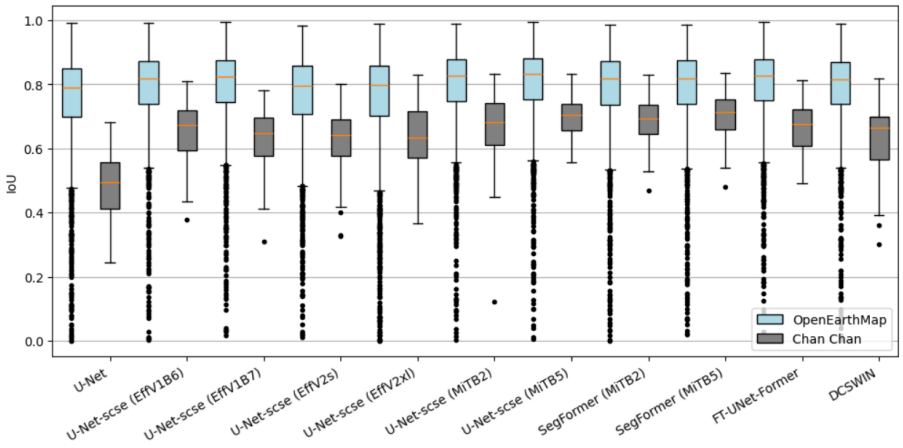

Figure 5. Dot and box plots of the IoU values of the models shown in Table 3, color-coded according to the dataset evaluated. OpenEarthMap (light blue) and Chan Chan (gray).

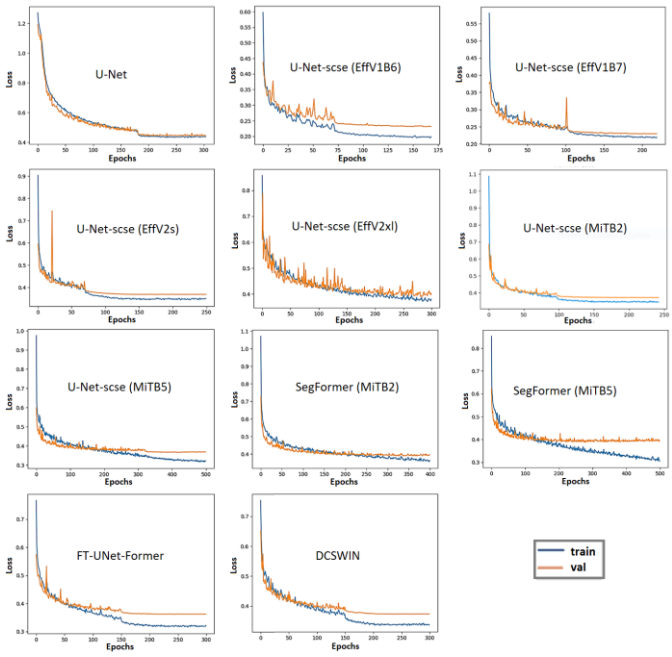

Figure 6. Loss curves of the models shown in Table 3, color-coded according to the training (blue) and validation (orange) process.

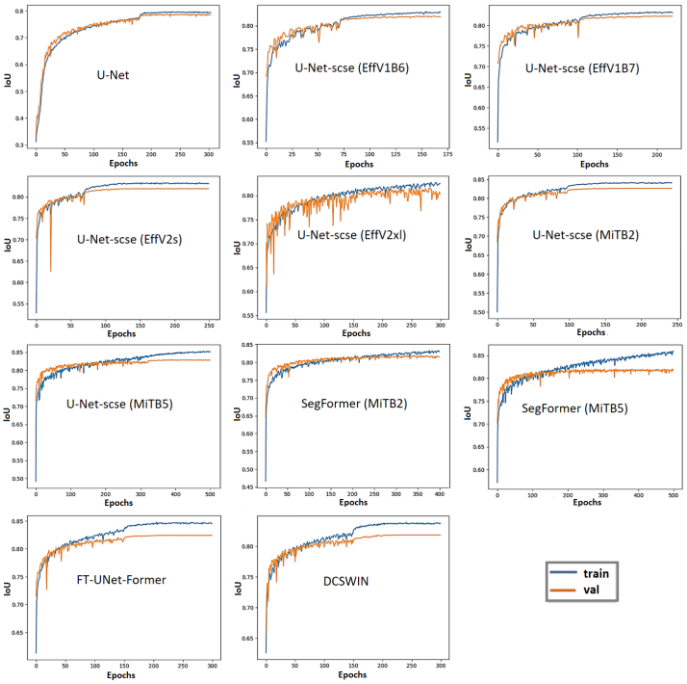

Figure 7. IoU curves of the models shown in Table 3, color-coded according to the training (blue) and validation (orange) process.

3.2. Qualitative Results in Peruvian Cities

The qualitative analysis of the segmentation of the models we examined was carried out in several cities in northwestern Peru, where urban expansion in the surrounding rural areas is comparable to that observed in the buffer zone of Chan Chan.

Figure 8 presents the results of building footprint segmentation in rural areas using images of cities in northwestern Peru from the OpenEarthMap dataset. Figure 8a displays the original RGB image, Figure 8b shows the manual semantic segmentation or ground truth, Figure 8c illustrates the semantic segmentation results of the U-Net-scse (MiTB2) model, Figure 8d depicts the semantic segmentation results of the U-Net-scse (MiTB5)

model, and Figure 8e demonstrates the semantic segmentation results of the FT-UNet-Former model. The first row corresponds to an image of Lambayeque city, where FT-UNet-Former achieves the highest IoU of 0.9258, while U-Net-scse (MiTB5) yields the lowest IoU of 0.6599. The second row pertains to an image of Viru city, where FT-UNet-Former attains the highest IoU of 0.8442, while U-Net-scse (MiTB2) produces the lowest IoU of 0.8130. The third row represents an image of Sechura city, where FT-UNet-Former achieves the highest IoU of 0.8065, while U-Net-scse (MiTB2) yields the lowest IoU of 0.3923. The fourth row corresponds to an image of Piura city, where U-Net-scse (MiTB5) attains the highest IoU of 0.7811, while FT-UNet-Former produces the lowest IoU of 0.6935.

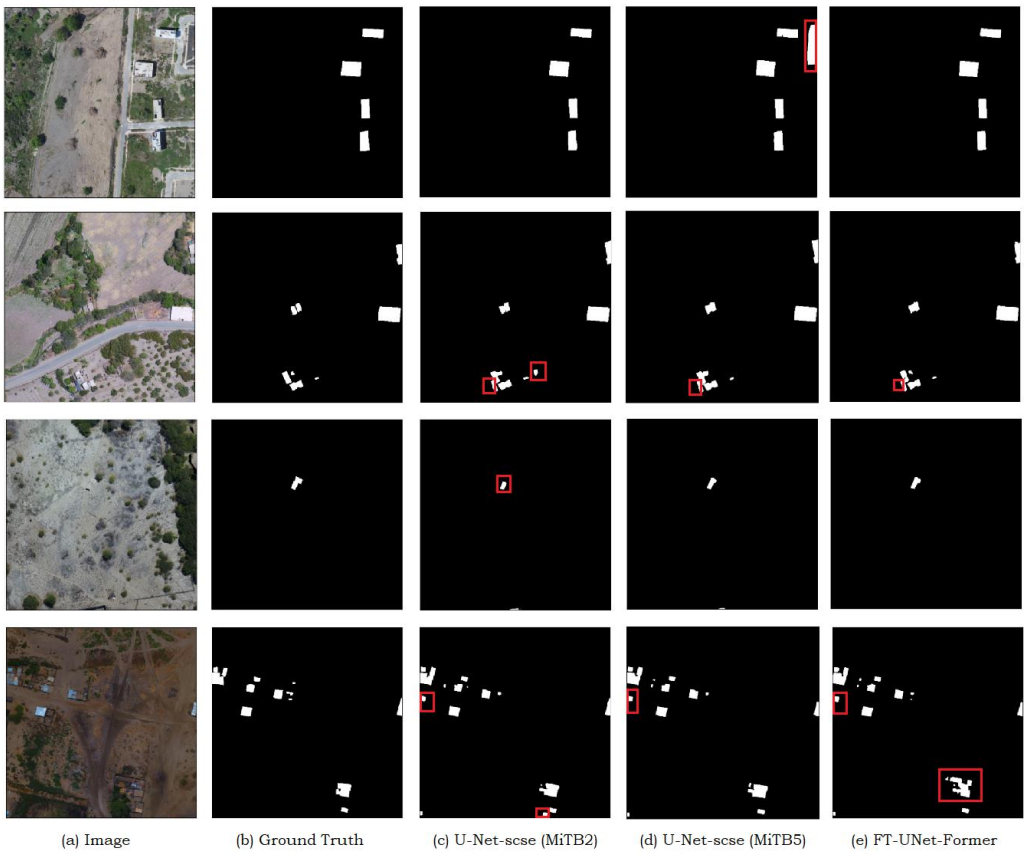

Figure 8. Experimental results on the OpenEarthMap dataset in rural areas of cities in northwestern Peru. The red boxes show the inconsistencies of the respective model on the labels in the study area.

Figure 9 presents the results of building footprint segmentation in high-density urban areas derived from images of Peruvian cities in the OpenEarthMap dataset. Figure 9a represents the original RGB image, Figure 9b depicts the manual semantic segmentation or ground truth, Figure 9c illustrates the semantic segmentation results of the U-Net-scse (MiTB2) model, Figure 9d shows the semantic segmentation results of the U-Net-scse (MiTB5) model, and Figure 9e displays the semantic segmentation results of the FT-UNet-Former model. The first row corresponds to an image of Chiclayo city, where U-Net-scse (MiTB5) achieved the highest IoU of 0.9072, while FT-UNet-Former yielded the lowest IoU of 0.8868. The second row pertains to an image of Lambayeque city, with U-Net-scse (MiTB2) achieving the highest IoU of 0.8957, and FT-UNet-Former producing the lowest IoU of 0.8803. The third row relates to an image of Piura city, where U-Net-scse (MiTB5) obtains the highest IoU of 0.8801, while FT-UNet-Former generates the lowest IoU of 0.8691. The fourth row corresponds to an image of Viru city, with U-Net-scse (MiTB5) achieving the highest IoU of 0.7528, and FT-UNet-Former yielding the lowest IoU of 0.7454.

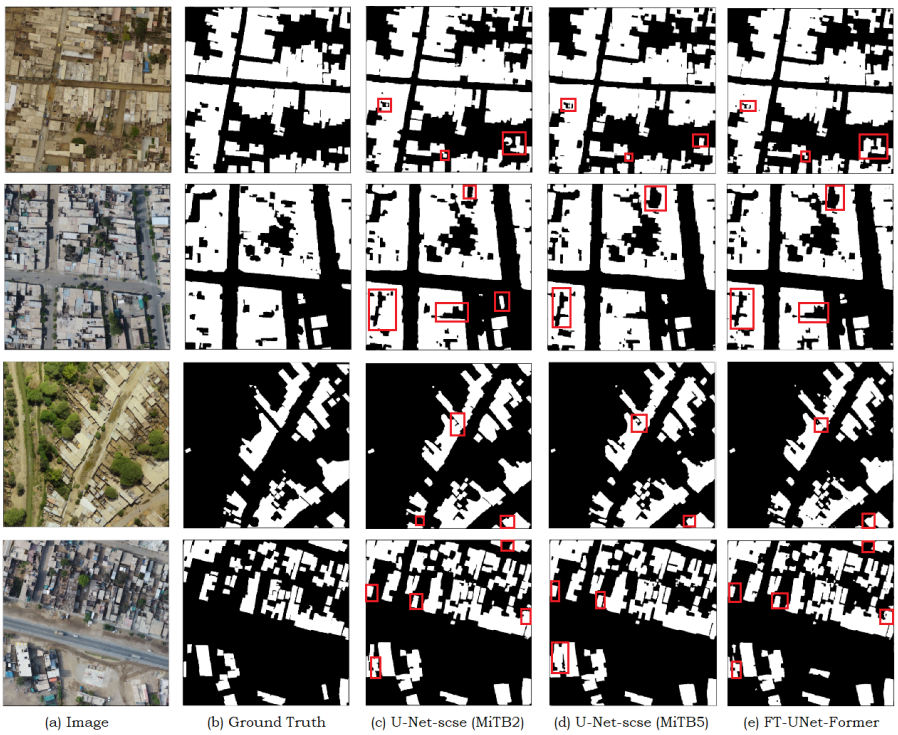

Figure 9. Experimental results on the OpenEarthMap dataset in urban areas of cities in northwestern Peru. The red boxes show the inconsistencies of the respective model on the labels in the study area.

Figure 10 presents the results of building footprint segmentation in the vicinity of the archaeological complex of Chan Chan, where Figure 10a displays the original RGB image and Figure 10b shows the manual semantic segmentation or ground truth. Figure 10c–e correspond to the semantic segmentation results of the three best-performing models from Table 3, specifically SegFormer (B2), U-Net-scse (MiTB5), and SegFormer (B5). In the first row, U-Net-scse (MiTB5) achieves the highest IoU of 0.7103, whereas in the second row, SegFormer (B5) attains the highest IoU of 0.6492. Finally, in the third row, U-Net-scse (MiTB5) achieves the highest IoU of 0.6539.

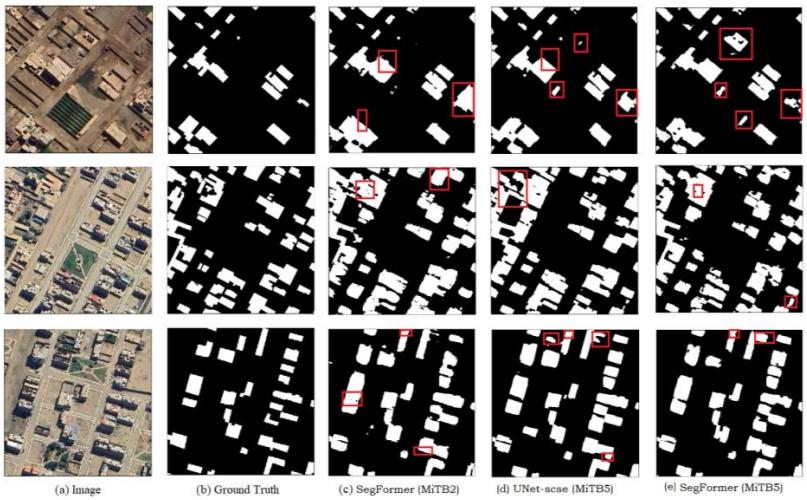

Figure 10. Experimental results on the surroundings of the archaeological complex of Chan Chan. The red boxes show the inconsistencies of the respective model on the labels in the study area.

4. Discussion

4.1. Building Semantic Segmentation in Peruvian Cities

From the qualitative results, it was possible to identify areas where the models exhibited limitations and challenges in the segmentation of building footprints. Specifically, in areas highlighted in red boxes, it can be observed that the models may encounter difficulties in accurately delineating buildings, resulting in a lower intersection over union (IoU) score. Upon further analysis, it is evident that building boundaries represent a significant challenge, influenced by disorganized layouts and variable sizes, as noted in previous studies [31,76]. Specifically, erroneous inferences in building boundaries were observed in areas with height differences between buildings and gaps in between (Figure 9), as well as in areas with buildings with varying roof materials and complex structures (Figure 8).

Additionally, Figure 10 illustrates the influence of the satellite incidence angle on the model perception, particularly in distinguishing between tall building facades and rooftops. These observations align with previous evidence, indicating that the network may encounter difficulties in differentiating between architectural components in situations where facades are visible in the final images, potentially affecting segmentation accuracy [45]. Consequently, detecting and rectifying limitations in building footprint segmentation in rural environments is critically important to enhance model performance in accurately identifying building boundaries, considering the structural complexity and variable satellite data acquisition conditions.

4.2. Urban Sprawl Analysis in Chan Chan

The experimental survey was conducted in three test areas within the buffer zone, each characterized by unique environmental features and at different stages of urban development (Figure 11).

Figure 11. Test areas on the surroundings of the archaeological complex of Chan Chan.

Zone 1 (Z1) is located along the coastal strip, which, during the Chimú empire, was occupied by totorales—low-lying fields irrigated by groundwater to cultivate the typical totora reed. Totorales, a natural environment characteristic of Peru's coastline, are protected wetlands that represent an example of biodiversity deserving preservation. However, Z1 is currently undergoing significant urbanization, with the presence of warehouses and

poultry farms as well as the subdivision of land into building plots near Huanchaquito beach, which are gradually being developed into residential areas (Figure 12).

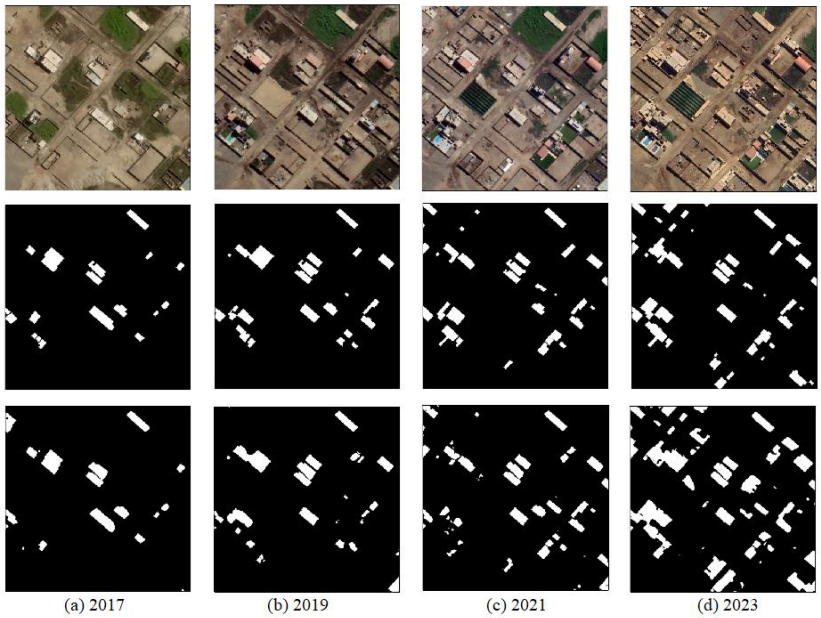

Figure 12. Segmentation results using U-Net-scse (MiTB5) on images of Zone 1 in Chan Chan for the years 2017–2023. Row 1: image, row 2: ground truth, row 3: U-Net-scse-MiTB5.

Zone 2 (Z2), Urbanización Ramón Castilla, is characterized by a more advanced stage of urbanization, with blocks predominantly occupied by residential buildings and few green spaces. Some areas, such as the one analyzed in the images (Figure 13), present lots still under construction interspersed with small gardens. Finally, Zone 3 (Z3), Urbanización Rosas del Valle, is an area undergoing urban expansion located near the limit of the core zone of the Chan Chan archaeological site. It is the only zone still relatively free of residential development within a broader context of intense and consolidated urbanization (Figure 14).

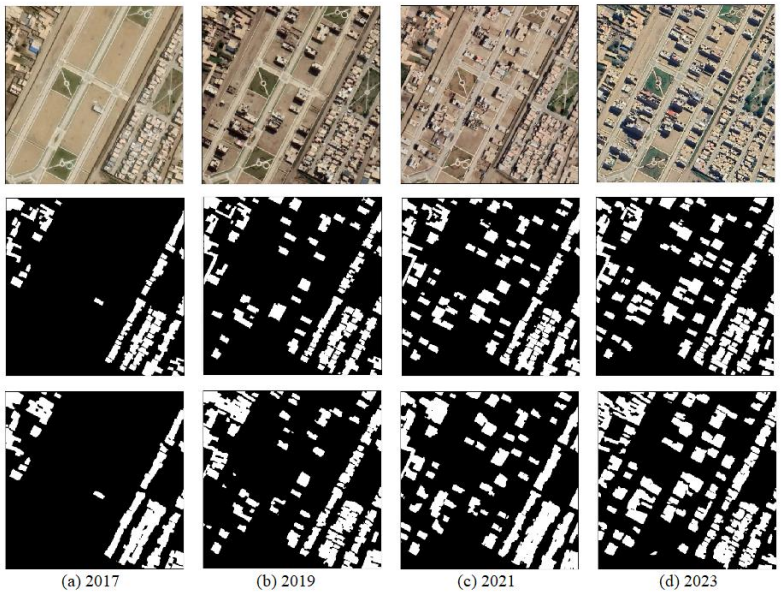

Figure 13. Segmentation results using U-Net-scse (MiTB5) on images of Zone 2 in Chan Chan for the years 2017–2023. Row 1: image, row 2: ground truth, row 3: U-Net-scse-MiTB5.

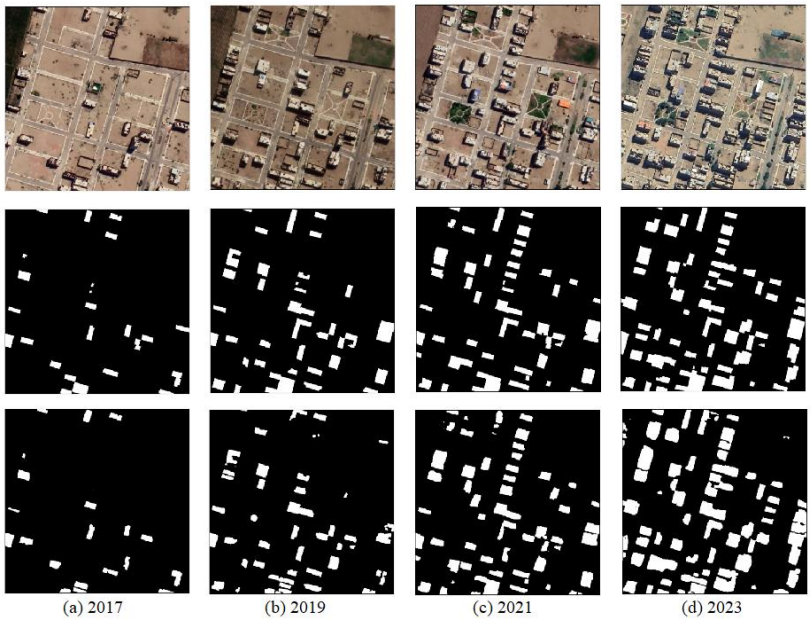

Figure 14. Segmentation results using U-Net-scse (MiTB5) on images of Zone 3 in Chan Chan for the years 2017–2023. Row 1: image, row 2: ground truth, row 3: U-Net-scse-MiTB5.

The image analysis accurately reflects the situations described. In Z1, a marked increase in built-up areas was observed in 2023, although large open spaces remain. In Z2, a previously undeveloped area in 2017 began to see construction in 2019, with a noticeable acceleration in development from 2021, immediately following the pandemic. A similar trend was observed in Z3: vacant or recently subdivided plots in 2017 were sparsely built upon by 2019, with significant growth in built-up areas recorded from 2021 onward.

It can be concluded that semantic segmentation highlights the progressive occupation of the Chan Chan buffer zone, with the consequent loss of open or green spaces. Urban development starts in consolidated residential sectors, often originating from informal settlements, and extends into areas increasingly closer to the archaeological complex. This ongoing phenomenon experienced significant acceleration between 2019 and 2021, possibly facilitated by the lack of territorial control during the pandemic years.

The analysis of the surface area of urbanized zones has shown that, during the period under study, approximately 11% of the buffer zone has been subdivided into plots and occupied by dwellings, while approximately 33% has undergone a change in land use. Statistical data published by the INEI (Instituto Nacional de Estadística e Informática del Perú) show a 120% increase in the number of apartments within residential complexes over the last 10 years, highlighting the exponential growth of multi-story buildings. This rapid urban expansion and the increased population density have created significant land management challenges, particularly in terms of traffic congestion and solid waste production [5].

5. Conclusions

In recent years, the cultural landscape of many UNESCO World Heritage Sites has been threatened by natural and anthropogenic factors. Among the latter, urban expansion, due to population growth, is generating a notorious threat to the conservation and management of the territory associated with heritage.

An improved U-Net architecture incorporating Mix Transformer encoders and active contour loss was developed in this research for the semantic segmentation of building footprints in high-resolution satellite images. The integration of an active contour loss function

further enhances the model's ability to delineate complex urban boundaries, addressing the challenges posed by the heterogeneous landscape surrounding the archaeological complex of Chan Chan.

Quantitative results showed that the U-Net-scse model with an MiTB5 encoder achieved the best performance with respect to SegFormer and FT-UNet-Former, with an IoU score of 0.8288 on OpenEarthMap and 0.6743 on Chan Chan images. This represents a significant improvement over previous methods, particularly for the challenging Chan Chan dataset. Qualitative analysis revealed the model's effectiveness in segmenting buildings across diverse urban and rural environments in Peru.

Application of the proposed method to multi-temporal imagery of the archaeological complex of Chan Chan enabled detailed monitoring of urban sprawl in the site's buffer zone between 2017 and 2023. The results clearly demonstrated progressive urbanization encroaching on open spaces and approaching the archaeological complex, with acceleration of development observed after 2019.

The rapid and uncontrolled occupation of Chan Chan's buffer zone is causing irreversible damage to the archaeological complex and its landscape, leading to the loss of its integrity and consequently its exceptional universal value as recognized by UNESCO. Not only has construction destroyed significant archaeological evidence in various parts of the buffer zone, but ancient Chan Chan is now completely surrounded by modern buildings that are visible from every part of the monumental complex, diminishing the charm of a cultural landscape that is unique in the world.

In addition, the lack of proper land use planning generates a number of challenging problems, including environmental pollution, traffic management, and solid waste generation. Although these factors cannot be directly measured through remote sensing analysis, they still result in significant land use transformations, such as road changes, debris accumulation, and the formation of open dumps. In further research, we aim to test useful algorithms to analyze and quantify these changes.

If legislative measures to regulate the buffer zone are not implemented as soon as possible, Chan Chan risks being removed from the World Heritage List. This approach for monitoring urban expansion over time can enable managers to take the necessary protective measures and to make informed decisions aimed at preserving cultural heritage sites and promote sustainable urban development around them. The methodology could also be extended to map other land cover types relevant to cultural landscape preservation.

Author Contributions: Conceptualization: M.C., F.J.L.T., F.C. and E.S.M.; methodology: F.J.L.T., M.C., F.C. and E.S.M.; software: M.C.; validation: M.C., F.C. and F.J.L.T.; formal analysis: F.J.L.T., M.C., F.C. and E.S.M.; investigation: F.J.L.T. and M.C.; data curation: M.C.; writing—original draft preparation: F.J.L.T. and M.C.; writing—review and editing: F.J.L.T., M.C. and F.C.; project administration: F.J.L.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by Universidad Continental.

Data Availability Statement: The data are available on the public dataset OpenEarthMap. Access to this dataset is through the following url: https://zenodo.org/records/7223446 (accessed on 16 October 2024).

Acknowledgments: We would like to thank the Dirección de Investigación de la Universidad Continental, who have always supported our research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. United Nations, Department of Economic and Social Affairs, Population Division. *World Urbanization Prospects: The 2018 Revision (ST/ESA/SER.A/420)*; United Nations, Department of Economic and Social Affairs, Population Division: New York, NY, USA, 2019. Available online: https://population.un.org/wup/assets/WUP2018-Report.pdf (accessed on 14 May 2024).
2. Gantulga, N.; Iimaa, T.; Batmunkh, M.; Surenjav, U.; Tserennadmin, E.; Turmunkh, T.; Denchingungaa, D.; Dorjsuren, B. Impacts of natural and anthropogenic factors on soil erosion. *Proc. Mong. Acad. Sci.* **2023**, *63*, 3–18. [CrossRef]
3. Chidi, C.L. Urbanization and Soil Erosion in Kathmandu Valley, Nepal. In *Nature, Society, and Marginality*; Perspectives on Geographical Marginality Series; Pradhan, P.K., Leimgruber, W., Eds.; Springer International Publishing: Cham, Switzerland, 2022; Volume 8, pp. 67–83. [CrossRef]
4. Santos, F.; Calle, N.; Bonilla, S.; Sarmiento, F.; Herrnegger, M. Impacts of soil erosion and climate change on the built heritage of the Pambamarca Fortress Complex in northern Ecuador. *PLoS ONE* **2023**, *18*, e0281869. [CrossRef]
5. Colosi, F.; Bacigalupo, C.; De Meo, A.; Kobata Alva, S.A.; Orazi, R.; Rojas Vasquez, G.E.; León Trujillo, F.J. Chan Chan y la pérdida de su paisaje cultural. *ACE Archit. City Environ.* **2025**, in press.
6. Gainullin, I.I.; Khomyakov, P.V.; Sitdikov, A.G.; Usmanov, B.M. Study of anthropogenic and natural impacts on archaeological sites of the Volga Bulgaria period (Republic of Tatarstan) using remote sensing data. In Proceedings of the Fourth International Conference on Remote Sensing and Geoinformation of the Environment, Paphos, Cyprus, 12 August 2016; SPIE: Bellingham, WA, USA, 2016. [CrossRef]
7. Agapiou, A.; Lysandrou, V.; Alexakis, D.D.; Themistocleous, K.; Cuca, B.; Argyriou, A.V.; Sarris, A.; Hadjimitsis, D.G. Cultural heritage management and monitoring using remote sensing data and GIS: The case study of Paphos area, Cyprus. *Comput. Environ. Urban Syst.* **2015**, *54*, 230–239. [CrossRef]
8. Agapiou, A. Multi-Temporal Change Detection Analysis of Vertical Sprawl over Limassol City Centre and Amathus Archaeological Site in Cyprus during 2015–2020 Using the Sentinel-1 Sensor and the Google Earth Engine Platform. *Sensors* **2021**, *21*, 1884. [CrossRef]
9. Xiao, D.; Lu, L.; Wang, X.; Nitivattananon, V.; Guo, H.; Hui, W. An urbanization monitoring dataset for world cultural heritage in the Belt and Road region. *Big Earth Data* **2021**, *6*, 127–140. [CrossRef]
10. Moise, C.; Dana Negula, I.; Mihalache, C.E.; Lazar, A.M.; Dedulescu, A.L.; Rustoiu, G.T.; Inel, I.C.; Badea, A. Remote Sensing for Cultural Heritage Assessment and Monitoring: The Case Study of Alba Iulia. *Sustainability* **2021**, *13*, 1406. [CrossRef]
11. Tang, Y.; Chen, F.; Yang, W.; Ding, Y.; Wan, H.; Sun, Z.; Jing, L. Elaborate Monitoring of Land-Cover Changes in Cultural Landscapes at Heritage Sites Using Very High-Resolution Remote-Sensing Images. *Sustainability* **2022**, *14*, 1319. [CrossRef]
12. Yao, Y.; Wang, X.; Luo, L.; Wan, H.; Ren, H. An Overview of GIS-RS Applications for Archaeological and Cultural Heritage under the DBAR-Heritage Mission. *Remote Sens.* **2023**, *15*, 5766. [CrossRef]
13. Cuca, B.; Agapiou, A. The Potentials of Large-Scale Open Access Remotely Sensed Ready Products: Use and Recommendations when Monitoring Urban Sprawl Near Cultural Heritage Sites. In Proceedings of the 2024 IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), Oran, Algeria, 15–17 April 2024; pp. 300–305. [CrossRef]
14. Megarry, W.P.; Cooney, G.; Comer, D.C.; Priebe, C.E. Posterior Probability Modeling and Image Classification for Archaeological Site Prospection: Building a Survey Efficacy Model for Identifying Neolithic Felsite Workshops in the Shetland Islands. *Remote Sens.* **2016**, *8*, 529. [CrossRef]
15. Balz, T.; Caspari, G.; Fu, B.; Liao, M. Discernibility of Burial Mounds in High-Resolution X-Band SAR Images for Archaeological Prospections in the Altai Mountains. *Remote Sens.* **2016**, *8*, 817. [CrossRef]
16. Stewart, C.; Oren, E.D.; Cohen-Sasson, E. Satellite Remote Sensing Analysis of the Qasrawet Archaeological Site in North Sinai. *Remote Sens.* **2018**, *10*, 1090. [CrossRef]
17. Cuypers, S.; Nascetti, A.; Vergauwen, M. Land Use and Land Cover Mapping with VHR and Multi-Temporal Sentinel-2 Imagery. *Remote Sens.* **2023**, *15*, 2501. [CrossRef]
18. Bachagha, N.; Elnashar, A.; Tababi, M.; Souei, F.; Xu, W. The Use of Machine Learning and Satellite Imagery to Detect Roman Fortified Sites: The Case Study of Blad Talh (Tunisia Section). *Appl. Sci.* **2023**, *13*, 2613. [CrossRef]
19. Zarro, C.; Cerra, D.; Auer, S.; Ullo, S.L.; Reinartz, P. Urban Sprawl and COVID-19 Impact Analysis by Integrating Deep Learning with Google Earth Engine. *Remote Sens.* **2022**, *14*, 2038. [CrossRef]
20. Gu, Z.; Zeng, M. The Use of Artificial Intelligence and Satellite Remote Sensing in Land Cover Change Detection: Review and Perspectives. *Sustainability* **2024**, *16*, 274. [CrossRef]
21. Southworth, J.; Smith, A.C.; Safaei, M.; Rahaman, M.; Alruzuq, A.; Tefera, B.B.; Muir, C.S.; Herrero, H.V. Machine learning versus deep learning in land system science: A decision-making framework for effective land classification. *Front. Remote Sens.* **2024**, *5*, 1374862. [CrossRef]
22. Chicchon, M.; Malinverni, E.S.; Sanità, M.; Pierdicca, R.; Colosi, F.; Trujillo, F.J.L. Building Semantic Segmentation Using UNet Convolutional Network on SpaceNet Public Data Sets for Monitoring Surrounding Area of Chan Chan (Peru). *Geomat. Environ. Eng.* **2024**, *18*, 25–43. [CrossRef]

23. Hoeser, T.; Kuenzer, C. Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends. *Remote Sens.* **2020**, *12*, 1667. [CrossRef]
24. Xie, Y.; Cai, J.; Bhojwani, R.; Shekhar, S.; Knight, J. A locally-constrained YOLO framework for detecting small and densely-distributed building footprints. *Int. J. Geogr. Inf. Sci.* **2019**, *34*, 777–801. [CrossRef]
25. Nurkarim, W.; Wijayanto, A.W. Building footprint extraction and counting on very high-resolution satellite imagery using object detection deep learning framework. *Earth Sci. Inform.* **2023**, *16*, 515–532. [CrossRef]
26. Pan, Z.; Xu, J.; Guo, Y.; Hu, Y.; Wang, G. Deep Learning Segmentation and Classification for Urban Village Using a Worldview Satellite Image Based on U-Net. *Remote Sens.* **2020**, *12*, 1574. [CrossRef]
27. Rastogi, K.; Bodani, P.; Sharma, S.A. Automatic building footprint extraction from very high-resolution imagery using deep learning techniques. *Geocarto Int.* **2020**, *37*, 1501–1513. [CrossRef]
28. Mou, L.; Hua, Y.; Zhu, X.X. Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7557–7569. [CrossRef]
29. Adamiak, M.; Jażdżewska, I.; Nalej, M. Analysis of Built-Up Areas of Small Polish Cities with the Use of Deep Learning and Geographically Weighted Regression. *Geosciences* **2021**, *11*, 223. [CrossRef]
30. Huang, Y.; Jin, Y. Aerial Imagery-Based Building Footprint Detection with an Integrated Deep Learning Framework: Applications for Fine Scale Wildland–Urban Interface Mapping. *Remote Sens.* **2022**, *14*, 3622. [CrossRef]
31. Amirgan, B.; Erener, A. Semantic segmentation of satellite images with different building types using deep learning methods. *Remote Sens. Appl. Soc. Environ.* **2024**, *34*, 101176. [CrossRef]
32. Wen, Q.; Jiang, K.; Wang, W.; Liu, Q.; Guo, Q.; Li, L.; Wang, P. Automatic Building Extraction from Google Earth Images under Complex Backgrounds Based on Deep Instance Segmentation Network. *Sensors* **2019**, *19*, 333. [CrossRef]
33. Tian, D.; Han, Y.; Wang, B.; Guan, B.; Gu, H.; Wei, W. Review of object instance segmentation based on deep learning. *J. Electron. Imaging* **2021**, *31*, 041205. [CrossRef]
34. Amo-Boateng, M.; Nkwa Sey, N.E.; Ampah Amproche, A.; Kyereh Domfeh, M. Instance segmentation scheme for roofs in rural areas based on mask R-CNN. *Egypt. J. Remote Sens. Space Sci.* **2022**, *25*, 569–577. [CrossRef]
35. Chen, S.; Ogawa, Y.; Zhao, C.; Sekimoto, Y. Large-Scale Building Footprint Extraction from Open-Sourced Satellite Imagery via Instance Segmentation Approach. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; IEEE: New York, NY, USA, 2022; pp. 6284–6287. [CrossRef]
36. Badrinarayanan, V.; Handa, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv* **2015**, arXiv:1505.07293.
37. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Lecture Notes in Computer Science; Navab, N., Hornegger, J., Wells, W., Frangi, A., Eds.; Springer: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]
38. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Proceedings of the 4th International Workshop, DLMIA 2018 and 8th International Workshop, ML-CDS 2018 Held in Conjunction with MICCAI 2018; Granada, Spain, 20 September 2018*; Lecture Notes in Computer Science; Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R.S., Bradley, A., Papa, J.P., Belagiannis, V., et al., Eds.; Springer: Cham, Switzerland, 2018; pp. 3–11. [CrossRef]
39. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: New York, NY, USA, 2017; pp. 6230–6239. [CrossRef]
40. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; IEEE: New York, NY, USA, 2017; pp. 2980–2988.
41. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Computer Vision—ECCV 2018*; Lecture Notes in Computer Science; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Cham, Switzerland, 2018; pp. 833–851. [CrossRef]
42. Jégou, S.; Drozdzal, M.; Vázquez, D.; Romero, A.; Bengio, Y. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: New York, NY, USA, 2017; pp. 1175–1183. [CrossRef]
43. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: New York, NY, USA, 2017; pp. 2261–2269. [CrossRef]
44. Zhang, W.; Tang, P.; Zhao, L.; Huang, Q. A comparative study of U-nets with various convolution components for building extraction. In Proceedings of the 2019 Joint Urban Remote Sensing Event (JURSE), Vannes, France, 22–24 May 2019; IEEE: New York, NY, USA, 2019; pp. 1–4. [CrossRef]

45. Bakirman, T.; Komurcu, I.; Sertel, E. Comparative analysis of deep learning based building extraction methods with the new VHR Istanbul dataset. *Expert Syst. Appl.* **2022**, *202*, 117346. [CrossRef]
46. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*; Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2012; Volume 25, pp. 1097–1105.
47. Abdollahi, A.; Pradhan, B.; Alamri, A.M. An ensemble architecture of deep convolutional Segnet and Unet networks for building semantic segmentation from high-resolution aerial images. *Geocarto Int.* **2022**, *37*, 3355–3370. [CrossRef]
48. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929. [CrossRef]
49. Chen, C.-F.R.; Fan, Q.; Panda, R. Crossvit: Cross-attention multi-scale vision transformer for image classification. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: New York, NY, USA, 2021; pp. 347–356. [CrossRef]
50. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: New York, NY, USA, 2021; pp. 9992–10002. [CrossRef]
51. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: New York, NY, USA, 2021; pp. 6877–6886. [CrossRef]
52. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2020**, arXiv:2010.04159.
53. Dai, X.; Chen, Y.; Yang, J.; Zhang, P.; Yuan, L.; Zhang, L. Dynamic DETR: End-to-End Object Detection with Dynamic Attention. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: New York, NY, USA, 2021; pp. 2968–2977. [CrossRef]
54. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: New York, NY, USA, 2021; pp. 9630–9640. [CrossRef]
55. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: New York, NY, USA, 2021; pp. 7242–7252. [CrossRef]
56. Sariturk, B.; Seker, D.Z. A Residual-Inception U-Net (RIU-Net) Approach and Comparisons with U-Shaped CNN and Transformer Models for Building Segmentation from High-Resolution Satellite Images. *Sensors* **2022**, *22*, 7624. [CrossRef]
57. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
58. Chan, T.F.; Vese, L.A. Active Contours without Edges. *IEEE Trans. Image Process.* **2001**, *10*, 266–277. [CrossRef]
59. Kim, B.; Ye, J.C. Mumford–Shah loss functional for image segmentation with deep learning. *IEEE Trans. Image Process.* **2019**, *29*, 1856–1866. [CrossRef]
60. Ma, J.; He, J.; Yang, X. Learning Geodesic Active Contours for Embedding Object Global Information in Segmentation CNNs. *IEEE Trans. Med. Imaging* **2020**, *40*, 93–104. [CrossRef]
61. Hatamizadeh, A.; Sengupta, D.; Terzopoulos, D. End-to-End Trainable Deep Active Contour Models for Automated Image Segmentation: Delineating Buildings in Aerial Imagery. In *Computer Vision—ECCV 2020*; Lecture Notes in Computer Science; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer: Cham, Switzerland, 2020; pp. 730–746. [CrossRef]
62. Mengjia, L.; Peng, L.; Bingze, S.; Yuwei, Z.; Luo, Z. Active Contour Building Segmentation Model based on Convolution Neural Network. *IOP Conf. Ser. Earth Environ. Sci.* **2022**, *1004*, 012015. [CrossRef]
63. Moseley, M.E.; Day, K.C. (Eds.) *Chan Chan: Andean Desert City*, 1st ed.; The University of New Mexico Press: Albuquerque, NM, USA, 1982.
64. Campana, C. *Chan Chan del Chimo: Estudio de la Ciudad de Adobe Más Grande de América Latina*; Editorial ORUS: Lima, Peru, 2006.
65. Vergara Montero, E.; Valle Álvarez, L. *Chan Chan: Ayer y Hoy*; Ediciones SIAN: Trujillo, Peru, 2012.
66. Rengifo, C. (Ed.) *Chan Chan: Esplendor y Legado. Redescubriendo la Antigua Capital del Chimor*; Ministerio de Cultura del Perú: Lima, Peru, 2020. Available online: https://ddclalibertad.gob.pe/transparencia/pdf/LibroChanChanEsplendorlegado.pdf (accessed on 15 January 2025).
67. Sitios del Patrimonio Mundial del Perú. Lista del Patrimonio Mundial en Peligro. Available online: https://patrimoniomundial.cultura.pe/listadelpatrimoniomundial/listadelpatrimoniomundialenpeligro (accessed on 1 October 2024).

68. Ministerio de Cultura. Resolución Ministerial N.° 130-2021-DM-MC, Plan Maestro para la Conservación y Manejo del Complejo Arqueológico de Chan Chan 2021–2031. 10 de Mayo de 2021. Available online: https://www.gob.pe/institucion/cultura/normas-legales/1915213-130-2021-dm-mc (accessed on 15 August 2024).
69. Colosi, F.; Fangi, G.; Gabrielli, R.; Orazi, R.; Angelini, A.; Bozzi, C.A. Planning the Archaeological Park of Chan Chan (Peru) by means of satellite images, GIS and photogrammetry. *J. Cult. Herit.* **2009**, *10* (Suppl. S1), e27–e34. [CrossRef]
70. Colosi, F.; Gabrielli, R.; Malinverni, E.S.; Orazi, R. Discovering Chan Chan: Modern technologies for urban and architectural analysis. *Archeol. Calc.* **2013**, *24*, 187–207. Available online: http://www.archcalc.cnr.it/indice/PDF24/09_Colosi_et_al.pdf (accessed on 16 January 2025).
71. Colosi, F.; Orazi, R. Integridad Material e Inmaterial de Chan Chan. In Proceedings of the En Actas VIII Congreso Nacional de Arqueología (Virtual), Lima, Peru, 16–21 August 2021; Ministerio de Cultura: Lima, Peru, 2022; pp. 131–144.
72. Colosi, F.; Orazi, R. Chan Chan Archaeological Park: Looming threats and suggested remedies. In *World Heritage and Ecological Transition, Proceedings of the Le Vie dei Mercanti XX International Forum, Napoli-Capri, Italy, 8–10 September 2022*; Ciambrone, A., Ed.; Architecture Heritage and Design; Gangemi Editore: Rome, Italy, 2022; Volume 10, pp. 452–460.
73. Colosi, F.; Malinverni, E.S.; Leon Trujillo, F.J.; Pierdicca, R.; Orazi, R.; Di Stefano, F. Exploiting HBIM for historical mud architecture: The huaca Arco Iris in Chan Chan (Peru). *Heritage* **2022**, *5*, 2062–2082. [CrossRef]
74. Instituto Nacional de Cultura. Resolución Directoral Nacional N° 1383/INC. 23 de Junio de 2010. Diario Oficial El Peruano, Normas Legales, 421488–421489. Available online: https://elperuano.pe/NormasElperuano/2010/06/30/512071-2.html (accessed on 4 February 2025).
75. Colosi, F.; Leon Trujillo, F.J.; Malinverni, E.S.; Kobata Alva, S.; Orazi, R. Multidisciplinary analisys and HBIM metodology for the risk management of harmfull events: The large earth complex of Chan Chan (Trujillo, Peru). Restauro Archeologico, Special Issue. In Proceedings of the Convegno Internazionale 1972–2022—II Patrimonio Mondiale alla Prova del Tempo: A Proposito di Gestione, Salvaguardia e Sostenibilità, Florence, Italy, 18–19 November 2022; pp. 24–31.
76. Xia, J.; Yokoya, N.; Adriano, B.; Broni-Bediako, C. OpenEarthMap: A Benchmark Dataset for Global High-Resolution Land Cover Mapping. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2–7 January 2023; IEEE: New York, NY, USA, 2023; pp. 6243–6253. [CrossRef]
77. Roy, A.G.; Navab, N.; Wachinger, C. Concurrent Spatial and Channel 'Squeeze & Excitation' in Fully Convolutional Networks. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*; Lecture Notes in Computer Science; Frangi, A., Schnabel, J., Davatzikos, C., Alberola-López, C., Fichtinger, G., Eds.; Springer: Cham, Switzerland, 2018; pp. 421–429. [CrossRef]
78. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 6105–6114. Available online: http://proceedings.mlr.press/v97/tan19a.html (accessed on 11 July 2024).
79. Tan, M.; Le, Q.V. EfficientNetV2: Smaller Models and Faster Training. In Proceedings of the 38th International Conference on Machine Learning, PMLR 139, Virtual Event, 18–24 July 2021; pp. 10096–10106. Available online: https://proceedings.mlr.press/v139/tan21a/tan21a.pdf (accessed on 12 July 2024).
80. Xie, E.; Wang, W.; Álvarez, J.; Anandkumar, A.; Yu, Z.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021), Proceeding of the NeurIPS 2021, Thirty-Fifth Annual Conference on Neural Information Processing Systems, Virtual Conference, 6–14 December 2021*; Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Wortman Vaughan, J., Eds.; NeurIPS: San Diego, CA, USA, 2021. Available online: https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf (accessed on 12 July 2024).
81. Wang, L.; Duan, C.; Fang, S.; Li, R.; Meng, X.; Zhang, C. A Novel Transformer Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6506105. [CrossRef]
82. Chicchon, M.; Bedon, H.; Del-Blanco, C.R.; Sipiran, I. Semantic Segmentation of Fish and Underwater Environments Using Deep Convolutional Neural Networks and Learned Active Contours. *IEEE Access* **2023**, *11*, 33652–33665. [CrossRef]
83. Widyaningrum, R.; Aulianisa, R.; Aji, N.R.A.S.; Candradewi, I. Comparison of Multi-Label U-Net and Mask R-CNN for panoramic radiograph segmentation to detect periodontitis. *Imaging Sci. Dent.* **2022**, *52*, 383–391. [CrossRef]