A Comparative Study on Vision Transformers in Remote Sensing Building Extraction

Georgios-Fotios Angelis, Armando Domi, Alexandros Zamichos, Maria Tsourma, Ioannis Manakos, Anastasios Drosou and Dimitrios Tzovaras

Information and Technologies Institute, Centre for Research and Technology Hellas,

Keywords: Remote Sensing, Transformers, Building, Extraction, Segmentation.

Abstract: Data visualization has received great attention in the last few years and gives valuable assets for better understanding and extracting information from data. More specifically, in Geospatial data, visualization includes information about the location, the geometric shape of elements, and the exact position of elements that can lead in enhances downstream applications such as damage detection, building energy consumption estimation, urban planning and change detection. Extracting building footprints from remote sensing (RS) imagery can help in visualizing damaged buildings and separate them form terrestrial objects. Considering this, the current manuscript provides a detailed comparison and a new benchmark for remote sensing building extraction. Experiments are conducted in three publicly available datasets aiming to evaluate accuracy and performance of the compared Transformer-based architectures. MiTNet and other five transformers architectures are introduced, namely DeepViTUNet, DeepViTUNet++, Coordformer, PoolFormer, EfficientFormer. In these choices we study design adjustments in order to obtain the best trade off between computational cost and performance. Experimental findings demonstrate that MitNet, which learns features in a hierarchical manner can be established as a new benchmark.

1 INTRODUCTION

Effective information visualiztion from RS imagery is a vital and useful step to applications, such as urban planning, damage detection and land use management. The acquisition of buildings footprints from remote sensing images was an open issue for discussion for researchers, but nowadays it can be considered as a mature research. Although, the recent advancements in artificial intelligence have given accurate solutions in many computer vision tasks, like building extraction from RS imagery, many challenges persist.

In the last few years, Transformers (Vaswani et al. (2017)) have demonstrated exceptional predictive performance in a large variety of natural language processing tasks, (Liu et al. (2020), Zhang et al. (2021)). Their performance results, lead research community to apply them in computer vision ((Liu et al., 2021)). Currently, they managed to stand as a state of the art solution image segmentation with several works proposing Transformer-based models as a solution. Moreover, this work (Xie et al. (2021)) introduced SegFormer, which utilized a hierarchical Transformer as an encoder and lightweight Multi Layer Perceptrons (MLPs) in the decoder part. Another interesting model presented in (Chen et al. (2021a)) where UNet was combined with Vision Transformer for medical image segmentation. The proposed TransUNet combined the individual advantages of the two networks and achieved superior results. Another approach is Trans4pass (Zhang et al. (2022)), that was originally proposed for panoptic segmentation. The authors introduce a Deformable Patch Embedding (DPE) which is applied both on the encoder and the decoder of the architecture.

Except their ability in understanding geometric object in scenes, Transformers have successfully applied in RS building extraction task. In this work, (Chen et al. (2021b)) the authors employed a Sparse token transformer, referred as STTNet for building extraction. Instead of using convolutional layers they utilized a spatial and channel transformer to receive a global receptive field. Additionally, they generated semantic sparse tokens in the low-resolution feature map to make their architecture computationally efficient. Another method that has achieved high seg-

Angelis, G., Domi, A., Zamichos, A., Tsourma, M., Manakos, I., Drosou, A. and Tzovaras, D.
A Comparative Study on Vision Transformers in Remote Sensing Building Extraction.
DOI: 10.5220/0011787800003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 3: IVAPP*, pages 222-229
ISBN: 978-989-758-634-7; ISSN: 2184-4321

mentation accuracy was (Wang et al. (2022a)). The authors designed a model that utilized as backbone Swin Transformer to extract context information and a novel model as a decoder DCFAM that was responsible to produce the final segmentation mask. Hatamizadeh et al. (2022) proposed UNetFormer, a model that was based on (Wang et al. (2022b)) for RS building footprint extraction. The authors in their proposed model they replaced the convolution layers in the UNet decoder with Transformer blocks. They also utilized Global-local attention to preserve and enhance the capturing of local and global information.

Considering their differences on experimental settings and dedicated vision tasks, it is necessary to categorize and evaluate the existing architectures under the same scenarios. Several comparative studies (Li et al. (2021); Han et al. (2022)) tried to address this issue aiming to provide extensive analysis and fair comparisons among domains, tasks, and performance. Focusing on building footprint extraction from RS imagery, which is tackled as semantic segmentation task, handful of publications exist that performed review, evaluated and summarized the current status of the literature (Han et al. (2022); Sariturk et al. (2022)). However, to the best of our knowledge, a study that evaluates the ability of state of the art Transformer-based architectures to extract building footprints under the same settings is missing. This manuscript aims to overcome this issue, presenting an extensive comparison of different Transformer-based models on three aerial imagery datasets (Inria Aerial Image Labeling dataset, WHU building dataset, WHU Satellite Dataset I (Global Cities)). Furthermore, the state of the art literature baselines are modified and several other Transformer-based variants are introduced (DeepViTUNet, DeepViTUNet++, Coordformer, MiTNet, PoolFormer, EfficientFormer) to explore the learning ability of Vision Transformers in RS imagery. The aforementioned modifications are based on two principles, first learning efficiently through hierarchical structure and secondly replacing the canonical self attention. Experimental findings show that hierarchical structure can learn from structured datasets efficiently, using a small number of network parameters. Considering these, the main contributions of this paper are the following:

- Extensive comparison and analysis of state-of-the-art Transformer based in buliding footprint extraction task.
- MitNet, a lightweight new benchmark approach that presents a trade-off between speed and accuracy.
- Modifications on the current Transformer models are being performed and five new architectures are being presented to handle building footprint extraction task.
- Evaluation in three publicly available datasets, that represent diverse conditions and settings.

The rest of the paper is organized as follows. First, MiTNet and the evaluated Transformer architectures are presented in section 2. The Section 3 presents the data of study and the experimental settings. The results of the building extraction from aerial imagery are presented and discussed in Section 4. The concluding remarks and future work are given in Section 5.

2 VISION TRANSFORMER MODELS

In this section before the models utilized for evaluation are introduced, a brief formulation on Vision Transformers is presented.

2.1 Preliminaries on Vision Transformers

Dosovitskiy et al. (2020), originally presented Vision Transformer (ViT), aiming to replace completely convolutions with Transformer blocks for image recognition task. The methodology of Vision Transformer is based on five steps. First of all, since the original Transformer architecture Vaswani et al. (2017) is taking $1D$ sequences as an input, the input image $x \in \mathcal{R}^{H \times W \times C}$ is converted into a sequence of flattened patches $x \in \mathcal{R}^{N \times P^2 C}$. The terms (H, W) correspond to the height and the width of the original image, C to the channels, (P, P) the width and height of the resulting image patch and finally $N = HW/P^2$ the number of patches that are created. Afterwards class tokens and positional encodings for each sequence are extracted with the methodology proposed in (Devlin et al. (2018)). Then positional encodings that are created to are added to the patch embeddings in order to hold positional information for each patch. The created embedding vector is utilized as input in the Transformer encoder. The main part in the encoder is the self-attention layer, that is responsible for computing the similarities between elements in the input, and more specifically between queries and keys. The self-attention is described from the following equation:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \qquad (1)$$

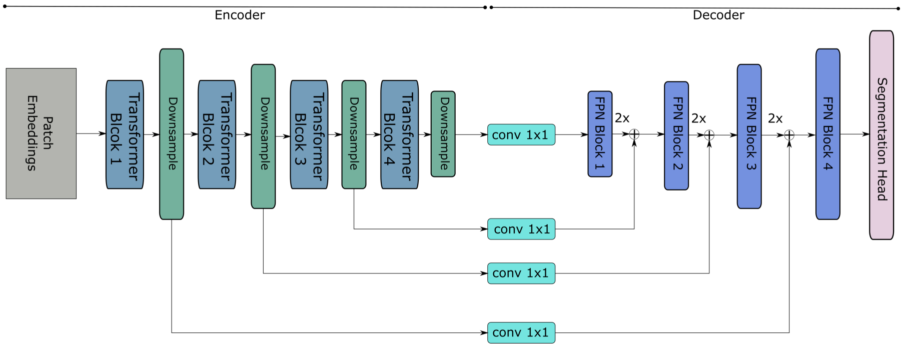

Figure 1: MiTNet complete architecture.

Where Q, K, V represent queries, keys and values that in the multi-head attention have the same dimensions $N \times C$, and N describes the length of the input sequence that is the product of the input patch width and height, $N = H \times W$. Except of the Multi-Head Self Attention (MHSA) layers, the Transformer encoder is composed of Feed forward networks (FFN) and Layer Normalization (LN) and residual connections.

2.2 MitNet: A Lightweight Approach

One of the key contributions of this manuscript is MiTNet, a network that presents a trade-off between performance and accuracy. As it can be depicted in Figure 1, MiTNet is composed of two parts the encoder and the decoder. The encoder is a the Hierarchical Transformer Encoder, originally presented in (Xie et al. (2021)). The transformer blocks are utilized, in our case. After each downsample layer the resolution of the input representation is reduced. In the downsampling instead of using regular convolutional layers, we employ linear layers with depthwise convolutions. Given an input image with dimensions $H \times W \times 3$, the input resolution after the downsampling process is $\frac{H}{2^{i+1}} \times W2^{i+1} \times C_i$ with $i \in \{1,2,3,4\}$. By this the encoder is able to produce multi-level features similar to CNNs. Furthermore another two core elements of the encoder are the efficient self-attention that replaces the regular self-attention introduced in 2.1 and the Mix-FFN. Regarding the first part, its core difference with vanilla self-attention is that utilizes the input sequence length reduction methodology that was applied in (Wang et al. (2021)). As a result the computational complexity is reduced from $O(N^2)$ to $O(\frac{N^2}{K})$ where K is manually selected. For the FFN, a depthwise convolutional layer with 3×3 kernel and zero padding and GeLU as activation function to mix it with FFN network, aiming to enhance the positional information that is captured in the Transformer blocks.

For the decoder part, an Feature Pyramid Network (FPN) decoder is utilized that was proposed for object detection task (Lin et al. (2017)). The decoder is composed of four convolutional parts, while also an 1×1 convolutional layer is applied before each downsampled transformer block is fed into the corresponding FPN part. In the current architecture layer normalization is added in this blocks These 1×1 lateral connections provide strong semantic features in each block directly from the encoder. After each FPN block the output representation is fed into an upsampling layer and then is added with the output of the lateral connection. Each FPN block is composed of 3×3 convolutional layer with zero padding and ReLU as activation function. After the fourth FPN block the output representation is passed into the segmentation head, which is a 1×1 convolutional layer that produces the final segmentation mask.

2.3 Revisiting Existing Vision Transformers

Except MiTNet five other architectures are built in order to provide a detailed evaluation. In this subsection the structure and characteristics are described.

- **Metaformer:** Metaformer is general architecture abstracted from Transformer by not specifying the token mixer (Yu et al. (2021)).The Metaformer architecture is considered Transformer/ MLP-like models depending of using attention/spatial MLP as the token mixer. The authors propose to replace the attention module in Transformers with a simple pooling operator as token mixer and introducing a new model named PoolFormer. In this manuscript, the PoolFormer-S12 model is used as feature extractor equipped with FPN decoder. This model is designed in order to compare with an architecture that utilizes also hierarchical structure, as MiTNet, but without using self-attention modules.

- **EfficientFormer:** In this work (Li et al. (2022)), the authors proposed a new dimension-consistent design paradigm for vision transformers in order to achieve low latency on mobile devices while maintaining high performance. They suggest a simple but efficient latency-driven slimming method to create a new family of models called EfficientFormers. The proposed EfficientFormer comprises patch embedding and a stack of meta transformer blocks, where each block contains different token mixer followed by a MLP block. The network has four stages, each serves as an embedding operation that maps the embedding dimensions and downsamples token length. The EfficientFormer-L1 is selected as backbone in conjunction with FPN decoder.
- **DeepViTUNet:** In this study, a variation of TransUNet is designed that uses DeepViT (Zhou et al. (2021)) in the bottleneck part of the UNet model. The proposed architecture employs deeper ViT modules, to increase the depth of the architecture. Moreover, another difference between the two TransUNet and DeepViTUNet, lies in the Transformer block, where DeepViT replaces the self-attention module with re-attention, to address attention collapse. For the DeepViTUNet also the same Base-16 heads architecture is employed.
- **DeepViTUNet++:** Based on the combination of UNet with Transformers, the latest modification of UNet series is applied, namely UNet++ (Zhou et al. (2018)). Based on this, a new architecture is proposed in this manuscript DeepViTUNet++. It follows the same encoder – decoder structure as its predecessors but its main difference is the redesigned path skip-ways that combine the feature representations in the two subnetworks. In that case, the architecture utilizes a dense convolution block whose number of layers depends on the pyramid level. Similar to the TransUNet and DeepViTUNet, the features from the DeepViT serve as a second input to the decoder.
- **Coordformer:** The core of this model was based on the UNetFormer, but the global-local attention was replaced with coordinate attention, initially proposed in (Hou et al. (2021)). Coordinate attention, was utilized to reduce computational resources, aiming at creating an attention mechanism suitable for devices with low computational power. The authors proposed to reshape channel attention by performing two separate $1D$ calculations and then aggregating the produced features into spatial dimension. The authors also proved that coordinate attention is also a proper candidate for several visual tasks. In the Coordformer model, we replace the global-local attention mechanisms inside the Transformer modules with coordinate attention, to create a more computationally efficient architecture.

3 EXPERIMENTAL SETTINGS

3.1 Dataset

In order to evaluate the aforementioned architectures, three publicly available datasets are utilized.

1. Inria Aerial Image Labeling (INRIA) dataset (Maggiori et al. (2017)) is a widely used and challenging database that contains urban settlements over five different cities. The spatial resolution is 0.3 m and the complete dataset covers 81 km^2 for each region. The final publicly available INRIA dataset includes 36 ortho-rectified images for each location, sized 5000×5000 pixels. Since there isn't a complete test set released, the training set is divided into a training and a test set with the ratio of 8:2.
2. WHU building dataset (Ji et al. (2018)) contains both aerial and satellite imagery with 0.075m spatial resolution and includes countryside, residential, culture, and industrial areas with more than 187000 building footprints. The dataset is composed of 8188 extracted tiles, 4736 utilized for training, 1036 for validation and 2416 for testing.
3. Moreover a partition of WHU building dataset is utilized, the WHU Satellite Dataset I (Global Cities). It contains 204 satellite images with multiple spatial resolutions collected from various satellite sources. The 75% of the total dataset samples is utilized for training while the rest 25% for testing purposes.

3.2 Experimental Setup

In this subsection, the preprocessing steps are described and the methodology utilized for training and testing is presented. For the INRIA dataset, since the original database includes images with high resolution, the included fine resolution images were divided into 512×512 overlapping patches to reduce complexity and the stride was set to 32. For the WHU and Global cities datasets the input image resolution remained the same at 512×512. Afterwards all the three datasets were augmented by randomly rotating, resizing, contrasting, transposing and horizontal axis flipping. Data augmentation helps in building a strong

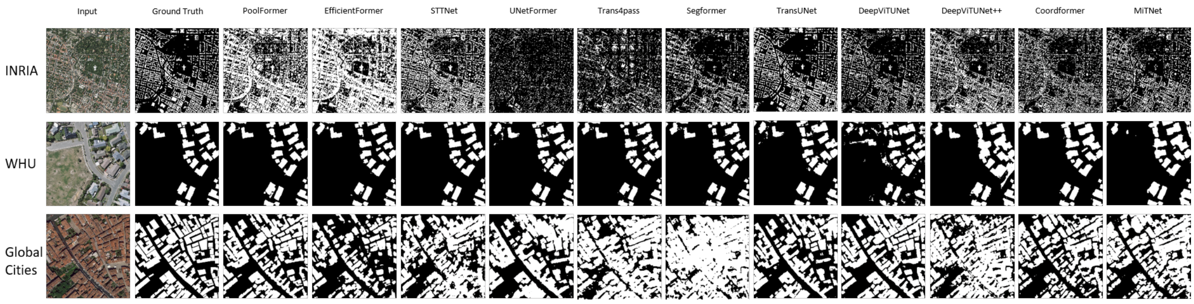

Figure 2: Qualitative comparison between Tranformer-based models in a single image in INRIA (first row) WHU (second row) and Global Cities (third row) datasets.

model, which is less dependent on input image orientation. This is very helpful for our model to generalize to different regions other than regions in training set.

All of the training procedures were implemented using PyTorch on a single NVIDIA RTX 3090 and the specific gradient descent variant used for the training of the models was the Adam optimizer with learning rate initial value of 1e−4 while the objective function that were utilized in all the cases except Coordformer, UNetFormer and STTNet, was cross entropy loss. In these three models a combination of cross entropy with dice loss (Jadon (2020)) was employed. Furthermore, in order to quantitatively evaluate the effectiveness of the proposed method, three different widely used metrics were employed, the overall accuracy (OA), mean Intersection over Union (*mIoU*), and F1-score (F1).

4 RESULTS

In this section, we present the numerical results of the experimental evaluation of the eleven aforementioned Transformer-based networks over three building extraction datasets, with the symbol $\star$ the architectures that are introduced in this manuscript are denoted. In Table 1, the accuracy results from all three datasets are illustrated.

Considering INRIA dataset, it can be distinguished that STTNet and TransUNet stand among the top performing models, in the *mIoU* metric, outperforming all the other architectures and achieving significant better value. Specifically for the STTNet the introduced sparse token sampler seems to enhance the prediction accuracy, with a relatively small number of parameters. Regarding the inference results in WHU dataset, it can be noticed that all the compared architectures are managing to extract buildings from remote sensing imagery more accurately. One obvious reason is that the WHU is a dataset that includes more training instances from the other two, that helps models to learn to segment objects efficiently. Moreover, comparing to the INRIA dataset, WHU includes smaller spatial resolution, where visual objects are more clearly described and as a consequence it affects positively the performance of the evaluated models. Regarding the prediction accuracy, it can be depicted that MiTNet presents the best results considering *OA* and *mIoU* metrics. STTNet maintains its accuracy while TransUNet which has attained decent results in the INRIA dataset, presents predictive accuracy degradation in WHU dataset. Moreover, architectures with hierarchical Transformers structure (PoolFormer, EfficientFormer, MiTNet) capture better simple low-level visual information and for this reason they manage to perform better in WHU dataset. Additionally, lower scale models like MiTNet can produce features more directly and extract more efficiently buidling masks. TransUNet manages to obtain better results in datasets with a larger spatial resolution, as it can operate better low level information. In overall, we can conclude that MiTNet outperforms all other approaches in two out of three metrics. More specifically, it reaches 93.27 on the *mIoU* metric outperforming all other methods. Additionally, the Global Cities dataset, consists of a limited number of training examples and spatial resolution similar to INRIA, which explains the lower predictive performance. Again, STTNet and TransUNet are the top performing approaches in all three metrics, and more specifically TransUNet is achieving the best results in all three metrics.

Further insights are provided from Figure 2 were inference results from a single image on each dataset are illustrated. The first two parts of each row of the image grid represent the input and the ground truth image, while the rest of them depict the predictions from the evaluated architectures. In the INRIA dataset, TransUNet achieves the best prediction results compared to all others. It captures the existence of buildings, manages to decouple them from impervious surfaces and roads. Additionally, it produces an accurate final result when it has to handle

Table 1: Results of Transformer-based models in all three datasets.

Models	Params (M)	INRIA			WHU building			Global Cities		
		OA	mIoU	F1-score	OA	mIoU	F1-score	OA	mIoU	F1-score
PoolFormer ⋆	13.2	93.03	70.75	80.93	98.28	91.89	95.66	86.49	70.30	81.84
EfficientFormer ⋆	13.2	90.98	69.03	79.80	98.37	92.30	95.90	86.54	70.61	82.10
STTNet	18.8	95.54	84.44	**91.21**	98.52	92.96	95.69	87.14	71.72	82.91
UNetFormer	11.7	94.05	74.73	84.19	95.98	92.45	**98.38**	87.95	73.62	84.31
Trans4pass	39.7	93.32	70.77	80.89	98.00	90.07	95.00	86.25	70.12	81.73
Segformer	13.6	87.61	57.65	68.59	91.57	85.09	96.41	84.87	68.46	80.58
TransUNet	21.5	**95.88**	**84.82**	91.03	91.09	71.57	82.05	**88.49**	**74.55**	
DeepViTUNet ⋆	20.1	91.45	64.92	75.63	83.29	58.99	71.63	88.09	74.14	84.70
DeepViTUNet++ ⋆	90.9	91.18	63.24	73.91	97.83	89.97	94.55	86.74	70.55	82.01
Coordformer ⋆	11.5	93.86	77.04	86.08	97.62	89.21	94.10	87.65	72.91	83.80
MiTNet ⋆	15.0	91.33	65.75	76.51	**98.60**	**93.27**	96.44	87.02	71.76	82.96

[1] Symbol ⋆ refers to the architectures that are introduced in this manuscript

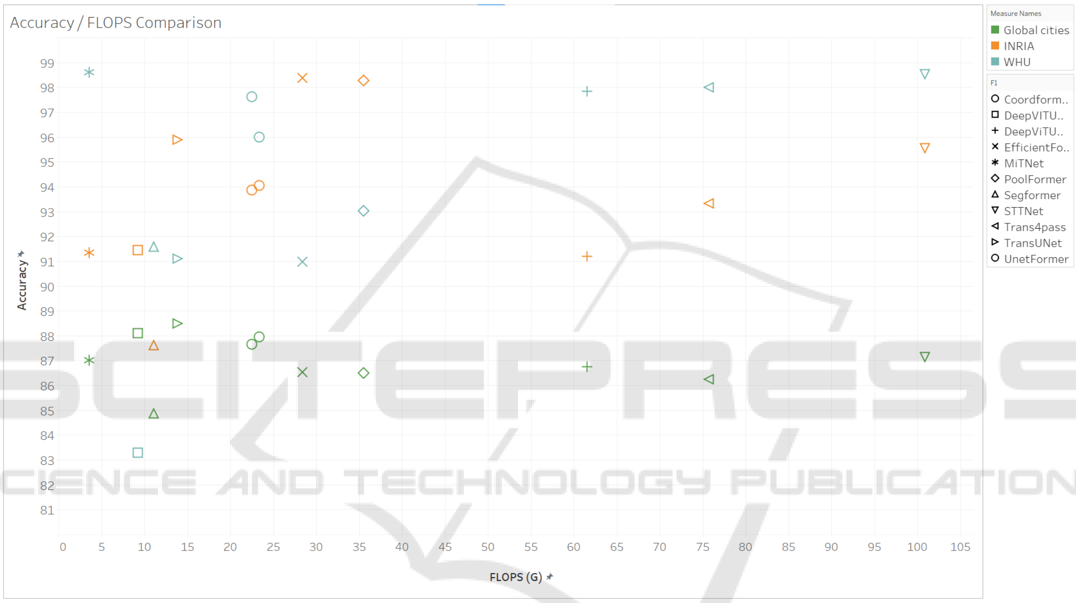

Figure 3: Accuracy - GFLOPS comparison for all the evaluated architectures on all three datasets. The legend on the left depicts the symbol that describes each evaluated architecture and the color refers to the dataset.

multiple tiles from a very high resolution image, returning a reconstructed mask without artifacts. Furthermore, the DeepViTUNet and MitNet also manage to produce better segmentation masks, especially compared to the RS literature baselines STTNet and UNetFormer, where the first one struggles to separate buildings from surfaces. For the WHU dataset, Coordformer, UNetFormer and STTNet produce segmentation maps more closely to the ground truth. They manage to predict segmentation masks with finer details while the building boundaries in a decent way. For the Global cities dataset TransUNet, DeepViTUNet and UNetFormer they are more sensitive to the context of the image and they predict segment buildings more accurately. However, all these three approaches face difficulty to generalize in the diverse conditions. In overall we can conclude that the lower spatial resolution enhances predictive performance and helps all the evaluated models to extract building segments.

For a real-time urban application to be feasible, metrics such as complexity, memory and speed are crucial. The performance of the eleven Transformer-based networks is presented in terms of the computation complexity measured in GFLOPs (G), the inference speed measured by frames per second (FPS), as well as the memory footprint measured with megabytes (MB). The inference speed is measured with input size of 512 × 512 on a single NVIDIA GTX 3090. The comparison results are presented in Table 2. The top performing models in terms of speed are Coordformer, UNetFormer, Segformer and

Table 2: Performance Comparison of Transformer-based models.

Models	Complexity (G) (↓)	Memory (Mb) (↓)	Speed (FPS) (↑)
PoolFormer ⋆	35.52	501.62	10.75
EfficientFormer ⋆	28.38	580.95	10.92
STTNet	100.9	1721.96	30.70
UNetFormer	23.38	386.19	118.76
Trans4pass	75.72	2075.61	12.50
Segformer	11.10	270.43	102.04
TransUNet	13.9	113.07	10.02
DeepViTUNet ⋆	9.24	**97.56**	11.73
DeepViTUNet++ ⋆	61.54	1131.91	22.16
Coordformer ⋆	22.52	345.39	**136.05**
MiTNet ⋆	**3.56**	211.62	101.11

[1] Symbol ⋆ refers to the architectures that are introduced in this manuscript

MiTNet. The first two models have a comparable inference speed, Coordformer achieves 136.05 FPS and UNetFormer achieves 118.76 FPS. In comparison with the STTNet, Coordformer is approximately 5 times faster and UNetFormer is approximately 4 times faster. In terms of computational complexity, MiTNet is the more efficient approach with significant difference from the second best DeepViTUNet, while it manages to surpass STTNet by 33 times.

Summarizing the experimental results from all the three datasets, and observing Figure 3, it can be concluded that MiTNet presents the best trade-off between performance and accuracy, as it presents decent levels of accuracy, with the smaller number of GFLOPS. For instance, in INRIA and Global cities datasets, it can be observed that except MiTNet, all the other top-performing models have more than 25 GFLOPS. Nevertheless, after 25 GFLOPs the improvement in accuracy of the models is around $1-2\%$, whereas the increase in computation complexity is significantly large. However, in WHU Building dataset, it outperforms all other approaches, while it is the most efficient architecture in terms of GFLOPS, with a significant difference from the second best. Apart from this, we can also observe that STTNet presents the most stable and accurate performance in all three datasets but with the bigger number of GFLOPs. Furthermore, it can be observed that TransUnet presents the second best predictive performance. Especially in comparison with all the Transformer - UNet variants, it can be claimed that Vision Transformer helps more effectively the model to extract buildings. However in all Transformer - UNet variants we can notice big deviations in accuracy results between different datasets. This observation raises concerns about the ability of the vanilla Vision Transformer to be robust solution in Remote Sensing imagery.

5 CONCLUSION

This paper investigated Vision Transformers in building footprint extraction from remote sensing imagery task, by performing analytical comparison between eleven different segmentation architectures and proposed a new benchmark model, MiTNet. All different architectures were trained and tested on three different publicly available datasets, aiming to evaluate the predictive performance in different scenarios and cities. MiTNet managed to present the best trade-off between speed and accuracy, and could be more suitable for practical applications. Additionally, is the top-performing approach in one out of three datasets. Moreover, five other Vision Transformer building footprint mask extraction models were introduced, where modifications on their structure were employed, aiming to monitor the effects predictive performance and computational efficiency. Future steps are focused on introducing an architecture that is entirely relied on Transformers aiming to exploit their properties on learning effectively low and high level features in computationally efficient manner.

ACKNOWLEDGEMENTS

This research was supported by grants from Horizon 2020, the European Union's Programmes for Research and Innovation under grant agreement No. 870373-SnapEarth, No. 101037648-Socio-Bee.

REFERENCES

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y. (2021a). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Chen, K., Zou, Z., and Shi, Z. (2021b). Building extraction from remote sensing images with sparse token transformers. *Remote Sensing*, 13(21):4441.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., and Tao, D. (2022). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hatamizadeh, A., Xu, Z., Yang, D., Li, W., Roth, H., and Xu, D. (2022). Unetformer: A unified vision transformer model and pre-training framework for 3d medical image segmentation. *arXiv preprint arXiv:2204.00631*.

Hou, Q., Zhou, D., and Feng, J. (2021). Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722.

Jadon, S. (2020). A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7. IEEE.

Ji, S., Wei, S., and Lu, M. (2018). Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586.

Li, C., Yang, J., Zhang, P., Gao, M., Xiao, B., Dai, X., Yuan, L., and Gao, J. (2021). Efficient self-supervised vision transformers for representation learning. *arXiv preprint arXiv:2106.09785*.

Li, Y., Yuan, G., Wen, Y., Hu, J., Evangelidis, G., Tulyakov, S., Wang, Y., and Ren, J. (2022). Efficientformer: Vision transformers at mobilenet speed.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.

Liu, X., Duh, K., Liu, L., and Gao, J. (2020). Very deep transformers for neural machine translation. *arXiv preprint arXiv:2008.07772*.

Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., Zhang, Y., Shi, Z., Fan, J., and He, Z. (2021). A survey of visual transformers. *arXiv preprint arXiv:2111.06091*.

Maggiori, E., Tarabalka, Y., Charpiat, G., and Alliez, P. (2017). Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3226–3229. IEEE.

Sariturk, B., Seker, D. Z., Ozturk, O., and Bayram, B. (2022). Performance evaluation of shallow and deep cnn architectures on building segmentation from high-resolution images. *Earth Science Informatics*, pages 1–23.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, L., Li, R., Duan, C., Zhang, C., Meng, X., and Fang, S. (2022a). A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5.

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., and Atkinson, P. M. (2022b). Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090.

Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., and Yan, S. (2021). Metaformer is actually what you need for vision.

Zhang, J., Chang, W.-C., Yu, H.-F., and Dhillon, I. (2021). Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 34:7267–7280.

Zhang, J., Yang, K., Ma, C., Reiß, S., Peng, K., and Stiefelhagen, R. (2022). Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16917–16927.

Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q., and Feng, J. (2021). Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer.