GeoRGMAE: Geospatially Guided Masked Autoencoders for Building Segmentation

Tuğba Eraslanoğlu[1], Guneet Mutreja[2], Martin Kada[1], Ksenia Bittner[2]

[1] Institute of Geodesy and Geoinformation Science, Technische Universität Berlin,
Straße des 17. Juni 135, 10623 Berlin, Germany - t.eraslanoglu@campus.tu-berlin.de, martin.kada@tu-berlin.de

[2] German Aerospace Center (DLR),
Münchener Straße 20, Wessling, 82234, Bavaria, Germany - (guneet.mutreja, ksenia.bittner)@dlr.de

Keywords: Semantic Segmentation, Building Segmentation, Masked Autoencoders, Masked Image Modelling, Remote Sensing.

Abstract

Accurate building segmentation from high-resolution aerial imagery is essential for various urban applications such as digital twins, geographic information system (GIS), and flood risk modelling. However, conventional supervised deep learning approaches require large amounts of pixel-level annotations, which are costly and time-consuming to obtain for large remote sensing datasets. To address this limitation, self-supervised learning (SSL) has recently emerged as an effective paradigm for learning visual representations from unlabeled data. In particular, masked autoencoders (MAE) have demonstrated strong performance by reconstructing masked image patches during pretraining. Nevertheless, conventional MAE frameworks rely on random masking strategies that ignore the spatial structure and semantic importance of regions in high-resolution remote sensing imagery. In this study, we propose GeoRGMAE, a geospatially guided masked autoencoder pretraining strategy for building segmentation. Unlike standard MAE, which rely on random masking, our approach leverages building footprint annotations available during pretraining to guide the masking process while preserving the original reconstruction objective. We introduce three masking strategies -core, balanced, and density-aware masking- that prioritize semantically relevant building regions under the varying urban densities. The core strategy focuses on building interiors, the balanced strategy distributes masking between buildings and background, and the density-aware adapts masking based on scene-level building density. Experiments on the Roof3D and WHU Building datasets demonstrate consistent, though modest, improvements over standard MAE pretraining, with the most effective masking strategy depending on dataset characteristics. These findings suggest that incorporating geospatial priors into masked image modelling (MIM) can improve representation learning for downstream building segmentation tasks.

1. Introduction

Accurate building segmentation is critical for urban applications such as urban planning, digital twins, disaster response, GIS, cadastral mapping, and 3D city models (Mutreja et al., 2025). As cities grow in complexity, the demand for automated building extraction systems capable of operating at regional or global scales has become increasingly important. Deep learning techniques have gained significant attention in the remote sensing and computer vision communities due to the availability of large-scale data, advances in algorithms, and improvements in computational hardware (Ozturk et al., 2020).

Deep learning approaches have made substantial progress and gradually replaced traditional techniques in image and building segmentation tasks (Wang et al., 2022). In particular, convolutional neural networks (CNNs) have been widely adopted due to their ability to learn hierarchical feature representations. These methods generally outperform traditional approaches in terms of accuracy and efficiency. However, most deep learning-based building segmentation methods rely heavily on large annotated datasets, which are expensive and time-consuming to produce.

To address this challenge, SSL has recently emerged as a promising paradigm for learning feature representations without labelled data (Rani et al., 2023, He et al., 2021, Dosovitskiy et al., 2021). SSL leverages the inherent structure of data to learn meaningful representations, making it particularly suitable for domains where labelled data is scarce, such as remote sensing.

MAE have proven to be an effective MIM approach within the SSL framework (He et al., 2021, Bao et al., 2022, Chen et al., 2020a). Unlike contrastive learning methods such as MoCo (Chen et al., 2020b) and SimCLR (Chen et al., 2020a), which learn representations by comparing positive and negative image pairs, MAE is a generative SSL method that reconstructs masked image patches. This allows the model to capture both global context and fine-grained spatial details, making it well-suited for segmentation tasks (He et al., 2021).

In MAE, a large random subset of image patches is masked, and the encoder processes only the visible patches. Masked tokens are introduced into the decoder, which reconstructs the original image at the pixel level. After pretraining, the decoder is discarded, and the encoder is transferred to downstream tasks (He et al., 2021).

Despite its effectiveness, the standard MAE framework employs random masking strategies that do not consider the spatial structure of the scene. While MAE performs well on conventional RGB images, where the visual context is relatively consistent and masking is effective, it faces challenges in remote sensing imagery. Due to the large spatial extent, different regions within a single image may represent entirely different geographies, making it harder for the model to learn meaningful representations. As a result, semantically important regions, such as buildings, may not be adequately emphasized during representation learning.

Building upon the principles of MAE (He et al., 2021, Li et al., 2022) and region-guided masked image modelling (RGMIM) (Li et al., 2024), this study proposes GeoRGMAE, a geospatially guided masked autoencoders framework for building segmentation. The proposed approach introduces three masking

strategies that guide the masking process toward semantically meaningful regions, particularly building areas.

The contributions of this work are threefold:

- We propose GeoRGMAE, a geospatially guided masked autoencoder pretraining approach that uses building footprint annotations to guide the masking policy while maintaining the original reconstruction objective.
- We introduce three masking strategies—core, balanced, and density-aware—that bias the masking process toward semantically relevant regions and adapt to varying urban densities.
- We provide an empirical analysis on Roof3D and WHU Building datasets, showing that region-guided masking yields consistent improvements over standard MAE pretraining and that the effectiveness of each strategy depends on scene characteristics.

2. Related Work

Building segmentation has been extensively studied in the field of remote sensing due to its importance for urban analysis applications such as GIS, digital twins, and disaster management. Traditional approaches relied on handcrafted features and rule-based methods, which often struggled to generalize across different urban environments. With the advancement of deep learning, CNNs have become the dominant approach for building segmentation tasks. These models are capable of learning hierarchical feature representations and have significantly improved segmentation accuracy compared to traditional methods (Wang et al., 2022). However, CNN-based approaches typically require large amounts of annotated data, which limits their applicability in large-scale or multi-city scenarios. More recently, transformer-based architectures have been introduced for remote sensing tasks, offering improved capability to capture long-range spatial dependencies. Despite these advances, most existing approaches remain dependent on supervised learning and large labelled datasets.

SSL has emerged as a promising paradigm for reducing the dependency on labelled data by learning representations directly from unlabeled inputs (Rani et al., 2023). SSL methods leverage intrinsic data properties to define pretext tasks, enabling models to learn meaningful feature representations without manual annotations. In remote sensing, SSL has gained increasing attention due to the abundance of unlabeled satellite and aerial imagery. Various SSL approaches have been proposed, including contrastive learning methods such as MoCo (Chen et al., 2020b), which learn representations by maximizing agreement between augmented views of the same image. While contrastive methods have shown strong performance, they often require large batch sizes and complex training strategies. As an alternative, MIM approaches have been proposed, which reconstruct masked portions of input images and provide a simpler and more scalable framework for self-supervised learning.

MAE (He et al., 2021) represent a significant advancement in SSL by introducing a reconstruction-based pretraining strategy using Vision Transformers (ViT). In MAE, a large portion of image patches is randomly masked, and the model learns to reconstruct the missing content from the visible patches. This approach has demonstrated consistent but modest improvements over standard MAE pretraining across various computer vision tasks and has been successfully applied to remote sensing imagery. The ability of MAE to capture both global and local structures makes it particularly suitable for segmentation tasks. However, a key limitation of MAE is its reliance on random masking strategies. Random masking does not consider the semantic importance of different regions within an image, which can be problematic for tasks such as building segmentation, where relevant information is concentrated in specific spatial regions.

To address the limitations of random masking, recent studies have explored region-guided masking strategies that incorporate additional information into the masking process. For example, RGMIM (Li et al., 2024) introduces semantic guidance to improve representation learning by focusing on informative regions. Similarly, other works in MIM (Hondru et al., 2025) have investigated alternative masking strategies to improve downstream task performance. These approaches suggest that guiding the masking process can enhance the quality of learned representations. However, existing region-guided methods are generally designed for natural images and do not explicitly consider the unique characteristics of remote sensing data, such as spatial structure, object sparsity, and variability across urban environments.

Despite the progress in SSL and masked image modelling, there is still a lack of approaches that explicitly incorporate geospatial information into the masking process for building segmentation tasks in remote sensing imagery. In high-resolution aerial images, buildings represent semantically important yet spatially sparse regions. Random masking strategies may fail to adequately emphasize these regions during pretraining, leading to suboptimal representation learning. To address this gap, this study proposes GeoRGMAE, a geospatially guided masked autoencoder framework that integrates building footprint information into the masking process. Unlike existing approaches, the proposed method leverages geospatial priors to guide the masking process.

3. Methodology

3.1 GeoRGMAE Overview

Instead of purely random masking, GeoRGMAE is guided by building footprint masks available during pretraining, used exclusively for masking guidance. At each training step, the model masks a fixed portion of image patches (e.g., 75%) with a bias toward building regions. Crucially, the encoder–decoder architecture and reconstruction objective remain unchanged from vanilla MAE: the encoder processes the visible patches and the decoder predicts pixel values for masked patches. Building footprint annotations are used *only* during pretraining to guide the masking process. This building-aware masking mechanism directs the model's attention to semantically important regions (buildings) while still learning from raw pixels, enabling the encoder to capture both fine structural details and surrounding context in urban scenes.

In contrast to standard MAE, which apply random masking over image patches, our approach uses building footprint annotations available in the pretraining dataset to guide the selection of masked regions. Importantly, these annotations are not used as prediction targets; the model is still trained to reconstruct the original RGB image. Instead, the annotations influence only the

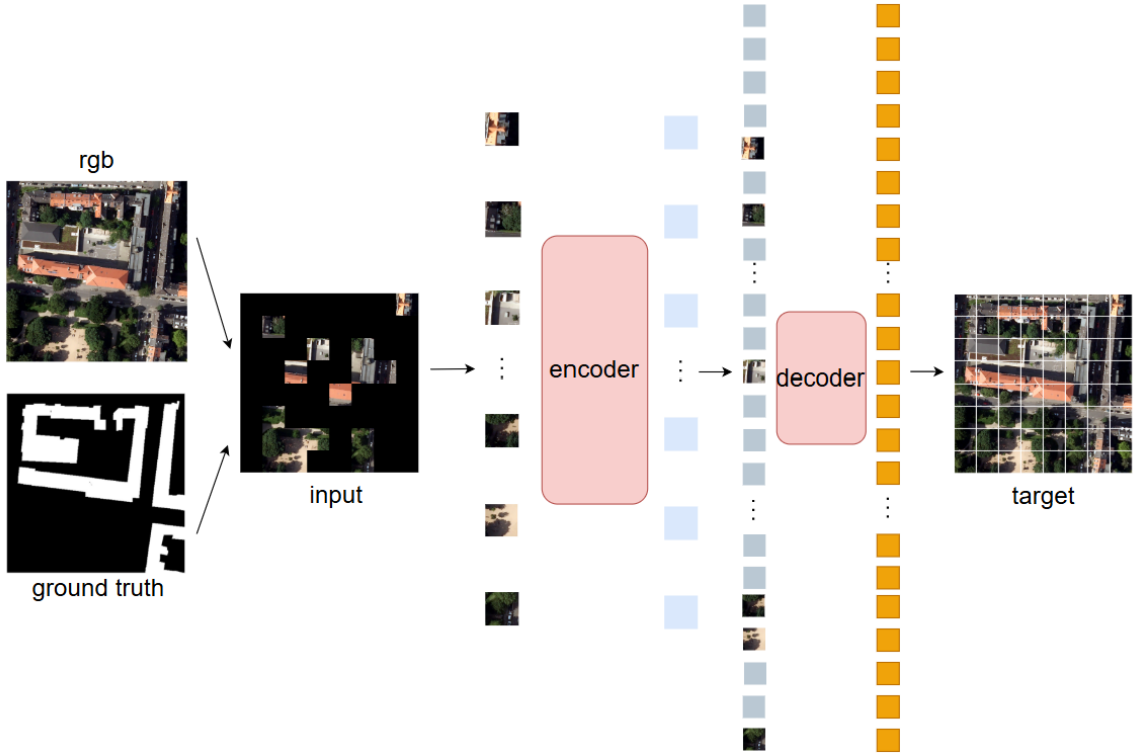

Figure 1. The overview of proposed GeoRGMAE pretraining model.

masking policy, biasing the model toward learning representations from semantically meaningful regions. As such, the proposed approach can be interpreted as a region-guided masked pretraining strategy that incorporates geospatial priors while preserving the self-supervised reconstruction objective.

3.2 Problem Formulation

An input image is partitioned into N non-overlapping patches. Each patch is assigned to either the building or background class using a binary mask.

The binary mask is initially defined at the pixel level and subsequently resized to the patch resolution. A patch is considered a building patch if the majority of its pixels belong to a building region; otherwise, it is classified as background.

Let $\{1, \ldots, N\}$ denote the set of patch indices. We define:

- $B \subseteq \{1, \ldots, N\}$ as the set of building patches,
- $G \subseteq \{1, \ldots, N\}$ as the set of background patches,

such that $B \cup G = \{1, \ldots, N\}$ and $B \cap G = \emptyset$.

Let $S \subseteq \{1, \ldots, N\}$ denote the set of masked patches and $m = |S|$ the number of masked patches determined by the global masking ratio.

In the proposed strategies, proportions such as 75% and 25% refer to the composition of the masked set S, rather than the total number of patches in the image. For instance, in the balanced masking strategy, 75% of the masked patches are sampled from building regions and 25% from background regions, subject to availability.
The masked set S is constructed by sampling from B and G according to the selected masking strategy. Sampling within each set is performed uniformly at random to avoid deterministic masking patterns and to improve generalization.

Let m_b and m_g denote the number of masked building and background patches, respectively. These quantities are defined as:

$$m_b = br \cdot m, \quad m_g = m - m_b, \tag{1}$$

where $br \in [0, 1]$ denotes the building masking ratio determined by the selected masking strategy.

These definitions control the distribution of masked patches across building and background regions.

3.3 Encoder and Decoder Architecture

Following the standard MAE framework, only the unmasked patches are processed by the encoder. Let X_U denote the set of visible patches after masking.

Let $X_U = \{1, \ldots, N\} \setminus S$ denote the set of visible patches and $u = |X_U|$.

Let $X_U = \{1, \ldots, N\} \setminus S$ denote the set of visible patch indices, and let $u = |X_U|$. Let x_i denote the i-th image patch.
Each visible patch is linearly projected into a latent embedding space:

$$\mathbf{z}_0 = [x_i \mathbf{E}]_{i \in X_U} + \mathbf{E}_{pos} \tag{2}$$

where $\mathbf{E}$ denotes the patch embedding matrix and $\mathbf{E}_{pos}$ represents positional embeddings.
The sequence is passed through L Transformer encoder blocks:

$$\mathbf{z}'_l = \mathbf{z}_{l-1} + \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{l-1})) \quad (3)$$

$$\mathbf{z}_l = \mathbf{z}'_l + \mathrm{MLP}(\mathrm{LN}(\mathbf{z}'_l)) \quad (4)$$

The final latent representation is:

$$\mathbf{y} = \mathrm{LN}(\mathbf{z}_L) \quad (5)$$

The decoder receives the encoded tokens along with mask tokens corresponding to masked patches. These tokens are reordered to their original positions and processed by a lightweight Transformer decoder to reconstruct the masked patches.

3.4 Reconstruction Objective

The model is trained to reconstruct the masked patches at the pixel level. The reconstruction loss is defined as the mean squared error (MSE) computed only on masked patches:

$$\mathcal{L} = \frac{1}{|S|} \sum_{i \in S} \|\hat{x}_i - x_i\|_2^2 \quad (6)$$

where $\hat{x}_i$ denotes the reconstructed patch and x_i is the original patch.

3.5 Masking Strategies

We introduce three novel masking strategies tailored to the distribution of buildings in each scene: *core masking*, *balanced masking*, and *density-aware masking*.

In the *core masking* strategy, the model masks out patches that cover buildings up to the target mask ratio. If there are fewer building patches than required, the model masks out the background patches until the overall masking ratio is fulfilled. This guarantees that the model focuses mostly on learning buildings. If $|B| \geq m$, then $S \subseteq B$ with $|S| = m$, sampled uniformly at random from B. Otherwise, all building patches are selected and the remaining patches are sampled from G.

In the *balanced masking* strategy, among the masked set S, 75% of masked patches are sampled from building regions and 25% from background, subject to availability. A fixed proportion of masked patches is allocated to building and background regions, such that:

$$m_b = 0.75m, \quad m_g = 0.25m \quad (7)$$

In cases where the number of available patches in a group is insufficient, the allocation is adjusted as:

$$m_b = \min(0.75m, |B|), \quad m_g = \max(m - m_b, 0) \quad (8)$$

This formulation defines the intended masking distribution. If neither group has enough patches, the allocation is adjusted to ensure that the global masking ratio is preserved. For instance, more building patches are masked out when the background patches are not enough. This method preserves the fixed-mask-ratio principle of MAE, prioritizes masking out building patches, and enables the model to learn both building and background attributes.

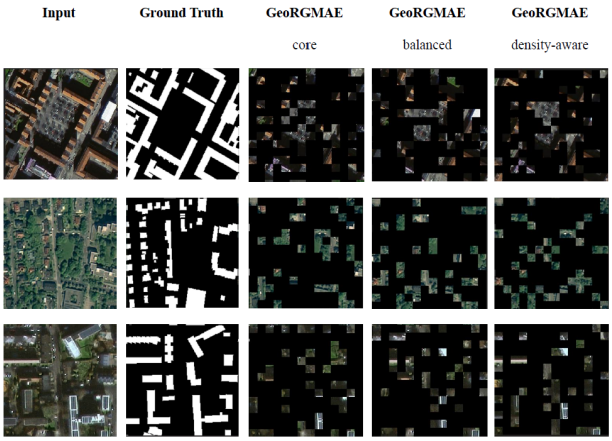

Figure 2. GeoRGMAE’s masking strategies representation. All inputs were masked out by using the same seed number.

Lastly, the *density-aware masking* strategy dynamically adjusts the fraction of masked building patches based on the scene’s building density. First, the building density is computed for each image, and different masking scenarios are applied accordingly. If the building density exceeds 30%, 85% of masked patches are sampled from building regions and 15% from background. If the density is between 10% and 30%, 70% of masked patches are sampled from building regions, with the remainder from background. For sparse scenes (below 10%), 50% of masked patches are sampled from building regions and the rest from background.
In density-aware masking, the building masking ratio br is defined as:

$$br = \begin{cases} 0.85, & \text{if density} > 0.3 \\ 0.70, & \text{if } 0.1 \leq \text{density} \leq 0.3 \\ 0.50, & \text{if density} < 0.1 \end{cases} \quad (9)$$

The thresholds used to define sparse, moderate, and dense scenes are chosen empirically to reflect common variations in urban structure, and are intended as heuristic design choices rather than theoretically derived values.

In density-aware masking, the value of br is dynamically adjusted based on the building density of the image.
If no valid building regions are present in the mask (i.e., $B = \emptyset$), the masking process falls back to uniform random sampling, which reduces GeoRGMAE to the standard MAE formulation. This ensures robustness when region annotations are missing or incomplete.
Additionally, a small stochastic variation (±5%) is applied to the masking ratio during training to improve robustness and prevent overfitting to fixed masking patterns.

$$\tilde{\sigma} = \sigma + \epsilon, \quad \epsilon \sim \mathcal{U}(-0.05, 0.05) \quad (10)$$

where $\sigma \in (0, 1)$ denotes the base masking ratio (e.g., 0.75)

As defined in Eq. 10, the masking ratio is perturbed at each iteration by a small uniformly distributed noise term, ensuring variability in the masking process and enhancing generalization. The difference among masking strategies are illustrated in Fig. 2.

4. Experiments and Results

4.1 Datasets

Three widely recognized benchmark datasets, the UBCv2 (Huang et al., 2023), Roof3D (Schuegraf et al., 2023) and the WHU Building (Ji et al., 2018), were selected in order to comprehensively assess the robustness and generalization capacity of the proposed model. While the UBCv2 dataset was used for the pretext task, the pretrained models were finetuned and tested with the Roof3D and WHU Building datasets.

The UBCv2 dataset mainly consists of a collection of labelled VHR satellite imagery from GaoFen-2 and SuperView (0.5-0.8m spatial resolution, RGB). The dataset comprises urban areas from 20 cities across five continents, selected for architectural diversity, varied spatial arrangements, and building density patterns. Cities include Beijing, Munich, Barcelona, Jacksonville, Tokyo, Berlin, São Luís, and others, with 499,339 building instances (RGB). While preparing the UBCv2 dataset for training, the annotations for 6,556 training images, split into 512×512 patches, were converted into binary mask images to guide the model.

The Roof3D is an aerial remote sensing dataset designed for individual building roof-plane and building segmentation. The dataset includes 92.5 percent real data, 5.4 percent synthetic data, and 2.1 percent manually annotated data with 0.1–0.3 m GSD. All RGB images resampled to 0.3 m by the dataset owner for consistency. The dataset covers a 22.4 km² area in German cities, Cologne and Berlin and consists of 2347 tiles, including synthetic and real images, 513 for validation, and 513 for testing, all with 512x512 pixels. Annotations were converted from the COCO format to binary mask images for use in the supervised downstream task.

The WHU Building dataset consists of more than 220,000 building instances extracted from aerial images with 0.075 m resolution and was down sampled by dataset owner to 0.3 m ground resolution. The dataset covers 450 km2 wide range of urban settlement types from Christchurch, New Zealand, and contains 4736 tiles for training, 1036 tiles for validation and 1036 tiles for testing with 512x512 pixels.

4.2 Network Training

The proposed GeoRGMAE framework is trained in two stages: self-supervised pretraining and supervised finetuning. During pretraining, five different configurations are considered, including MAE trained from scratch, ImageNet-pretrained MAE (He et al., 2021), and the proposed GeoRGMAE with core, balanced, and density-aware masking strategies. All models are pretrained on the UBCv2 dataset for 250 epochs using the MAE-ViT-Base architecture, which consists of a 12-layer encoder with 768 dimensions and an 8-layer lightweight decoder with 512 dimensions. The models are trained using the AdamW optimizer with a base learning rate of 1.5×10^{-4}, weight decay of 0.05, and $\beta_1 = 0.9$, $\beta_2 = 0.999$. The effective learning rate is scaled according to the batch size. A cosine learning rate schedule with a 20-epoch warm-up is employed. The masking ratio is set to 75%, and the reconstruction objective is defined using the L2 loss computed only on masked patches. The effective batch size is 256, achieved via gradient accumulation. During pretraining, building masks are not used as supervision labels but only to guide the masking distribution in GeoRGMAE. This preserves the self-supervised learning paradigm while incorporating geospatial priors into the training process.

In the finetuning stage, the pretrained encoder is integrated with a UPerNet decoder within the MMSegmentation framework (Contributors, 2020). The models are finetuned on Roof3D and WHU Building datasets, using a 70/15/15 train-validation-test split. The training is performed for 36,000 iterations using the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 0.05. A linear warm-up is applied for the first 1,500 iterations, followed by a polynomial learning rate decay. The training objective consists of a main decode-head cross-entropy loss and an auxiliary loss weighted by 0.4. Mixed-precision (FP16) training is employed to improve computational efficiency.

During pretraining, standard data augmentation techniques are applied to improve generalization. The input images are randomly resized and cropped to 224×224, followed by random horizontal flipping, and are normalized using the dataset mean and standard deviation. These augmentations are selected to preserve the spatial structure of remote sensing imagery. During finetuning, a broader set of data augmentation techniques is employed, including random resizing, cropping, horizontal and vertical flipping, rotation, photometric distortion, and cutout. All models are trained using a consistent set of hyperparameters to ensure a fair comparison across different masking strategies.

Although the base masking ratio is set to 75%, the effective masking ratio varies slightly at each iteration due to a stochastic perturbation (±5%), improving robustness and preventing overfitting to fixed masking patterns.

The experiments were conducted using the PyTorch framework in two separate environments for pretraining and fine-tuning. The pretraining stage was performed using Python 3.10 and PyTorch 2.0 with CUDA 11.8, while the fine-tuning stage was carried out using Python 3.8 and PyTorch 2.1 with CUDA 12.1, along with the MMSegmentation framework (Contributors, 2020) and its dependencies. All experiments were executed on an NVIDIA Quadro RTX 5000 GPU with 16 GB memory, enabling efficient training and evaluation of large-scale semantic segmentation models.

4.3 Evaluation

We validate GeoRGMAE on building segmentation benchmarks using a ViT (He et al., 2021) encoder–decoder backbone. The model is first pre-trained on a large amount of unlabeled aerial images with building footprint overlays, which is UBC-v2, then fine-tuned for semantic segmentation on labeled building datasets, which are Roof3D and WHU Building. An overview of the pre-training process is shown in Fig. 1. We compare our GeoRGMAE against two baselines: a vanilla MAE trained from scratch with random masking, and an MAE initialized with ImageNet pre-training (representative of generic features).

The performance of the proposed GeoRGMAE framework is evaluated using standard semantic segmentation metrics. In this study, Intersection over Union (IoU) and overall pixel accuracy (Acc) are reported to assess building segmentation performance. IoU measures the overlap between predicted building regions and ground-truth masks, while accuracy reflects the proportion of correctly classified pixels over the entire image.

The evaluation is conducted on two datasets with different characteristics, namely Roof3D and WHU Building. Roof3D contains heterogeneous urban scenes from multiple cities, whereas

WHU Building consists of more homogeneous building structures from a single region. For both datasets, a 70/15/15 split is used for training, validation, and testing, respectively. The validation set is used to monitor convergence during training and to select the best-performing model based on IoU.

On the Roof3D dataset, GeoRGMAE achieves the best performance with the density-aware masking strategy, reaching 72.04 % IoU, compared to 71.50 % and 71.52 % for scratch and ImageNet - pretrained MAE, respectively. On the WHU Building dataset, the core masking strategy achieves the best performance with 79.05 % IoU, slightly improving over the baselines (78.86 %). These results indicate that the effectiveness of masking strategies depends on dataset characteristics, with different strategies performing better under varying scene complexities.

Qualitatively, the GeoRGMAE models produce sharper building boundaries compared to the random-masking baseline. The quantitative results are provided in Tables 1 and 2.

model	masking type	IoU (%)	Acc (%)
MAE (scratch)	random	71.50 ± 0.48	80.80 ± 0.91
MAE (ImageNet weights)	random	71.52 ± 0.67	82.49 ± 1.44
GeoRGMAE (ours)	core	$\mathbf{71.77 \pm 0.40}$	82.12 ± 0.64
GeoRGMAE (ours)	balanced	$\mathbf{71.76 \pm 0.38}$	81.86 ± 0.38
GeoRGMAE (ours)	density-aware	$\mathbf{72.04 \pm 0.48}$	82.61 ± 0.31

Table 1. Comparison of segmentation results for different pretraining strategies on Roof3D test set.

model	masking type	IoU (%)	Acc (%)
MAE (scratch)	random	78.86 ± 0.29	85.75 ± 0.32
MAE (ImageNet weights)	random	78.86 ± 0.22	85.72 ± 0.61
GeoRGMAE (ours)	core	$\mathbf{79.05 \pm 0.34}$	85.92 ± 0.62
GeoRGMAE (ours)	balanced	78.81 ± 0.47	85.77 ± 0.09
GeoRGMAE (ours)	density-aware	$\mathbf{78.91 \pm 0.20}$	85.53 ± 0.22

Table 2. Comparison of segmentation results for different pretraining strategies on WHU Building test set.

Table 1 and Table 2 report the quantitative results for all evaluated models. To ensure robustness, all experiments were repeated with four different random seeds, and the mean and standard deviation are reported. The reported scores can be considered reliable, given that the proposed framework relies on MAE pretraining and uses limited labelled data during finetuning.

GeoRGMAE yields consistent but modest improvements over both scratch MAE and ImageNet-pretrained MAE across the evaluated datasets. On Roof3D, the density-aware mask-ing strategy achieves the best performance, suggesting that adapting the masking policy to scene-level building density is beneficial in heterogeneous urban environments. In contrast, on the WHU Building dataset, the core masking strategy performs best, indicating that focusing on building interiors may be sufficient in more homogeneous settings. These observations highlight a trade-off between emphasizing semantic regions and preserving contextual information. Strategies that strongly prioritize building regions can improve structural understanding but may reduce exposure to background context, whereas adaptive strategies such as density-aware masking can better balance these factors across diverse scenes.

For qualitative evaluation, segmentation outputs are visually compared with the corresponding ground-truth masks. Example results are presented in figures 3 and 4. The visual inspection focuses on building boundary accuracy, completeness of detected structures, and robustness in challenging regions such as shadowed areas and vegetation.

The evaluation results demonstrate that the proposed GeoRGMAE framework generalizes well across datasets with varying levels of complexity. In particular, the masking strategies enable the model to capture semantically meaningful structures during pretraining, which leads to improved segmentation performance in downstream tasks.

5. Limitations

Despite its effectiveness, the proposed approach has several limitations. First, it relies on the availability of building footprint annotations during pretraining to guide the masking policy, and is therefore not fully label-free. Second, the observed performance improvements are modest and evaluated on a limited number of datasets. Future work should explore fully annotation-free region discovery methods, and evaluate the approach on a broader range of geospatial datasets.

6. Conclusion

This paper introduces GeoRGMAE, a geospatially guided masked autoencoder pretraining approach for building segmentation in remote sensing imagery. By incorporating building footprint information into the masking process, the proposed method biases representation learning toward semantically meaningful regions while preserving the standard reconstruction objective of masked autoencoders. Experimental results on the Roof3D and WHU Building datasets demonstrate that geospatially guided masking strategies can provide consistent improvements over standard MAE pretraining, with the most effective strategy depending on dataset characteristics and urban density.

These results suggest that integrating geospatial priors into MIM is a promising approach for improving downstream performance in remote sensing tasks. Future work will explore annotation-free alternatives for region discovery and extend the approach to more diverse datasets and modalities.

7. Acknowledgements

One of the authors, Tuğba Eraslanoğlu, was supported Study Abroad Graduate Scholarship Program by The Ministry of National Education, Republic of Türkiye that is greatly appreciated.

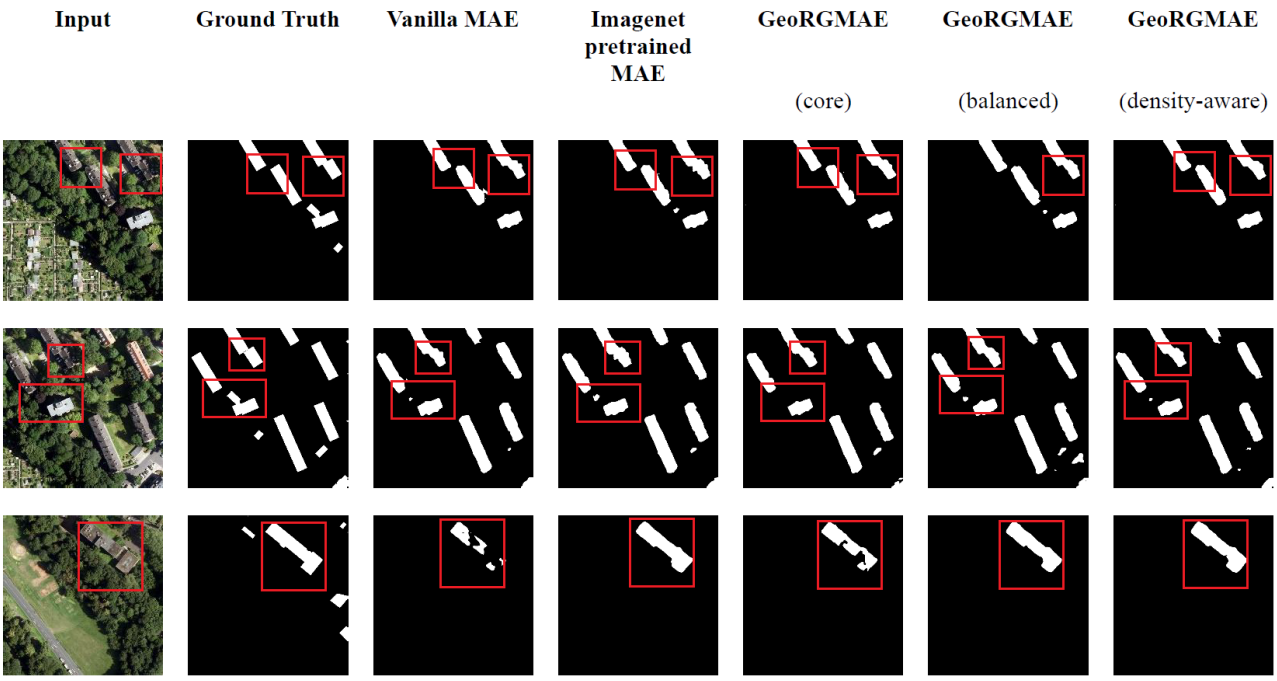

Figure 3. MAE and GeoRGMAE semantic segmentation results on the Roof3D test dataset. Note that the red boxes highlight regions with noticeable differences in building delineation quality among the models, especially in shadowed regions.

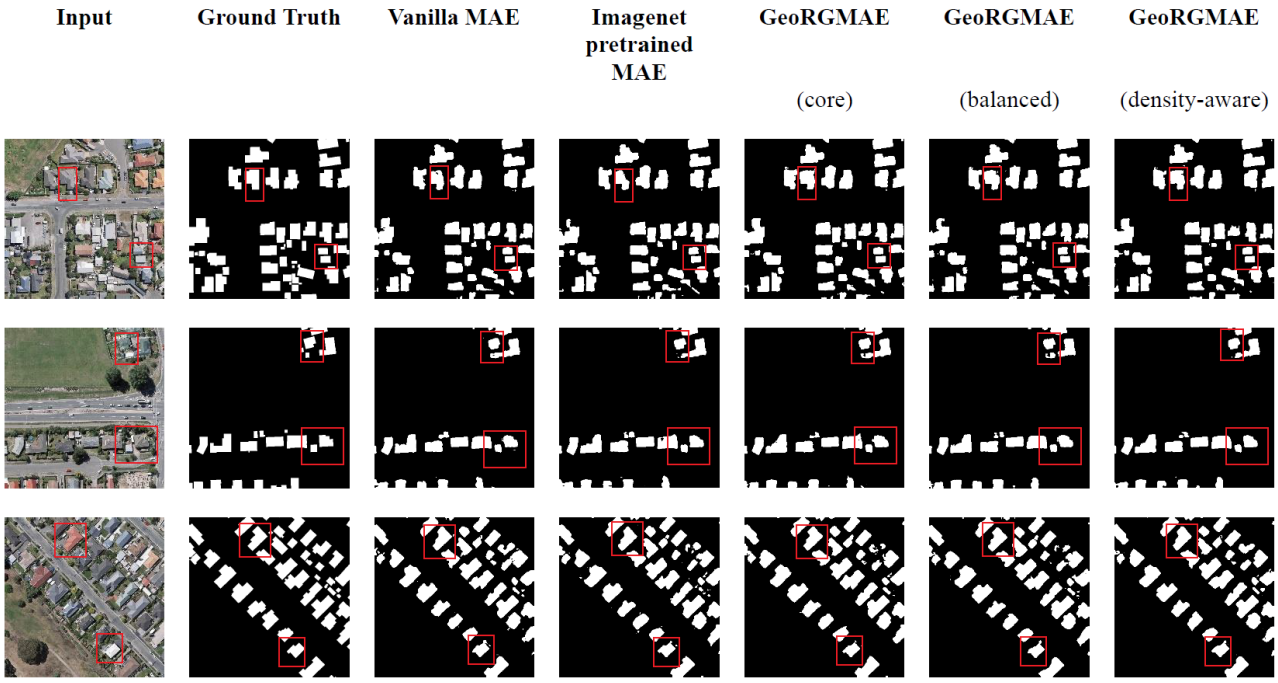

Figure 4. MAE and GeoRGMAE semantic segmentation results on the WHU Building test dataset. Note that the red boxes highlight regions with noticeable differences in building delineation quality among the models.

References

Bao, H., Dong, L., Piao, S., Wei, F., 2022. BEiT: Bert pre-training of image transformers.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations.

Chen, X., Fan, H., Girshick, R., He, K., 2020b. Improved baselines with momentum contrastive learning.

Contributors, M., 2020. MMSegmentation: Open-mmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is

worth 16x16 words: Transformers for image recognition at scale.

He, K., Chen, X., Xie, S., Li, Y., Dollar, P., Girshick, R., 2021. Masked Autoencoders Are Scalable Vision Learners. https://doi.org/10.48550/arXiv.2111.06377.

Hondru, V., Croitoru, F. A., Minaee, S., Ionescu, R. T., Sebe, N., 2025. Masked image modeling: A survey.

Huang, X., Chen, K., Tang, D., Liu, C., Ren, L., Sun, Z., Hänsch, R., Schmitt, M., Sun, X., Huang, H., Mayer, H., 2023. Urban building classification (UBC) v2—a benchmark for global building detection and fine-grained classification from satellite imagery.

Ji, S., Wei, S., Lu, M., 2018. Fully Convolutional Networks for Multi-Source Building Extraction from an Open Aerial and Satellite Imagery Dataset. *IEEE Transactions on Geoscience and Remote Sensing*.

Li, G., Togo, R., Ogawa, T., Haseyama, M., 2024. RGMIM: Region-guided masked image modeling for learning meaningful representations from x-ray images.

Li, G., Zheng, H., Liu, D., Wang, C., Su, B., Zheng, C., 2022. SemMAE: Semantic-guided masking for learning masked autoencoders.

Mutreja, G., Schuegraf, P., Bittner, K., 2025. HiRes-FusedMIM: A high-resolution rgb-dsm pre-trained model for building-level remote sensing applications.

Ozturk, O., Sariturk, B., Seker, D. Z., 2020. Comparison of fully convolutional networks (fcn) and u-net for road segmentation from high resolution imageries.

Rani, V., Nabi, S. T., Kumar, M., Mittal, A., Kumar, K., 2023. Self-supervised Learning: A Succinct Review. *Archives of Computational Methods in Engineering*, 30(4), 2761–2775.

Schuegraf, P., Fuentes Reyes, M., Xu, Y., Bittner, K., 2023. ROOF3D: A real and synthetic data collection for individual building roof plane and building sections detection.

Wang, L., Fang, S., Meng, X., Li, R., 2022. Building Extraction With Vision Transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–11. http://dx.doi.org/10.1109/TGRS.2022.3186634.