ISPRS Open Journal of Photogrammetry and Remote Sensing 21 (2026) 100139

Contents lists available at ScienceDirect

ISPRS Open Journal of Photogrammetry and Remote Sensing

journal homepage: www.journals.elsevier.com/isprs-open-journal-of-photogrammetry-and-remote-sensing

Disentangling geometry and texture: An investigation into large-scale RGB-DSM pre-training for 3D urban understanding

Guneet Mutreja *, Ksenia Bittner

German Aerospace Center (DLR), Münchener Straße 20, Weßling, 82234, Bavaria, Germany

ARTICLE INFO

Dataset link: https://github.com/guneetmutreja/HiResFused-MIM

Keywords:
Masked image modeling
Multi-modal learning
Digital surface models
Self-supervised learning
Foundation models
Urban 3D reconstruction

ABSTRACT

Recent Earth observation foundation models have achieved strong transfer performance using large-scale self-supervised pre-training, but most remain primarily spectral or RGB-based and do not explicitly exploit geometric information from digital surface models (DSMs). This limits their effectiveness in dense urban scenes, where height, building boundaries, and vertical structure are essential for tasks such as building extraction and height estimation.

In this work, we present HiRes-FusedMIM, an empirical study of large-scale multi-modal pre-training for high-resolution urban remote sensing. We curate a dataset of 1.2 million paired RGB and DSM image tiles from native source products primarily between 0.1 to 0.5 m spatial resolution, processed to training resolutions of 0.2 to 0.5 m where appropriate, making it, to our knowledge, one of the largest publicly available high-resolution paired RGB–DSM datasets. Specifically, we compare shared and modality-specialized dual encoders, masked image modeling and contrastive alignment, and small- versus large-scale pre-training.

Our results show that architecture is critical for geometry-sensitive transfer. A shared encoder suffers from modality interference and drops to 18.0 AP50 on UBCv2 instance segmentation, whereas the dual encoder reaches 31.5 AP50. Similarly, the dual encoder improves LoveDA semantic segmentation from 41.58 to 52.28 mIoU and reduces height-estimation error from 9.26 m to 7.89 m RMSE. We further find that global contrastive alignment benefits some scene classification tasks but degrades dense geometric prediction. Finally, scaling unique training samples from 368k to 1.2M improves UBCv2 AP50 by 3.0 points, indicating that data diversity is more effective than repeated training on a smaller corpus. These findings provide practical guidance for designing geometry-aware foundation models for urban remote sensing.

1. Introduction

Earth observation (EO) is undergoing a notable shift driven by the growing availability of large-scale remote sensing archives and the adoption of foundation models. In computer vision, self-supervised learning (SSL) has enabled models with strong generalization capabilities by leveraging massive unlabeled datasets. Recent efforts in the EO domain, such as SatMAE (Cong et al., 2023) and RingMo (Sun et al., 2023), have successfully adapted these paradigms to multi-spectral and temporal satellite imagery. However, urban environments present unique challenges that spectral data alone cannot fully resolve. The high variance of urban morphology, characterized by severe occlusions and cast shadows, often leads to spectral confusion where impervious surfaces (e.g., concrete roofs versus roads) exhibit identical reflectance signatures (Rottensteiner et al., 2014; Soergel, 2010). Furthermore, the integration of heterogeneous data sources remains non-trivial due to the statistical domain gap between optical intensity and geometric elevation fields (Ghamisi et al., 2018).

Despite the critical importance of geometry, current geospatial foundation models remain predominantly reliant on spectral, optical, or SAR imagery, exhibiting a limitation we term "height blindness" (Cong et al., 2023; Jakubik et al., 2023; Hong et al., 2024; Guo et al., 2024; Diao et al., 2025). By failing to exploit the vertical structural information encoded in DSMs, these models can be limited in dense urban scenarios where 3D geometry is essential for tasks such as building instance separation and height estimation (Hattula et al., 2024; Diao et al., 2025). RGB imagery provides rich texture, material, and contextual cues, but it can be ambiguous when roofs, roads, courtyards, and other impervious surfaces exhibit similar reflectance, or when cast shadows obscure object boundaries (Peng et al., 2019; Fu et al., 2025). In such cases, DSMs provide complementary height-derived information, including relative elevation, height discontinuities, roof morphology, and vertical separation between adjacent structures (Hattula et al., 2024; Fu et al., 2025; Hazaymeh et al., 2023). These cues

* Corresponding author.
E-mail address: guneet.mutreja@dlr.de (G. Mutreja).

https://doi.org/10.1016/j.ophoto.2026.100139
Received 23 January 2026; Received in revised form 22 July 2026; Accepted 27 July 2026
Available online 17 August 2026

are particularly important for distinguishing buildings from spectrally similar surfaces, separating neighboring buildings, delineating building boundaries, and estimating urban morphology, where the target output depends not only on semantic appearance but also on the spatial arrangement and vertical structure of objects (Hattula et al., 2024; Fu et al., 2025).

The integration of RGB and DSM data, however, is non-trivial due to the modality gap: RGB imagery captures high-frequency texture and appearance cues, whereas DSMs represent continuous surface elevation with fundamentally different gradient statistics and noise characteristics (Li et al., 2022). While supervised deep learning has explored fusion architectures (Peng et al., 2020; Albanwan and Qin, 2021), these insights have yet to be systematically evaluated in the context of large-scale self-supervised pre-training. This study therefore asks how RGB–DSM foundation models should fuse texture and geometry: whether through shared representations or modality-specialized streams, and whether generative reconstruction or contrastive alignment better preserves geometry-sensitive urban information.

A primary bottleneck in advancing geometry-aware foundation models is the scarcity of high-resolution aligned training data. Unlike the massive repositories available for Sentinel-2 or Landsat (Stewart et al., 2023), high-resolution (0.2–0.5 m) paired RGB-DSM data are fragmented and costly to acquire. Prior multi-modal pre-training efforts, such as DeCUR (Wang et al., 2024) or msGFM (Han et al., 2024), have typically relied on datasets containing fewer than 50,000 paired samples. At such modest scales, models often struggle to learn robust and transferable geometric representations and risk overfitting to local urban morphologies. As a result, it remains unclear how architectural choices (e.g., shared versus dual encoders) and learning objectives (e.g., contrastive versus generative) interact when training foundation models at scale.

In this work, we present HiRes-FusedMIM, a comprehensive empirical investigation into large-scale multi-modal pre-training for high-resolution urban understanding. We address the data bottleneck by curating and publicly releasing a dataset of 1.2 million paired RGB and DSM image tiles, representing a >30× increase over previously available paired RGB–DSM datasets used in multi-modal pre-training (Han et al., 2024; Wang et al., 2024). Leveraging this unprecedented scale, we conduct a rigorous $2 \times 2 \times 2$ factorial analysis examining the interplay of architectural design, learning objectives, and data scale. Our experiments reveal that standard Shared Encoder architectures which are common in diverse multi-modal settings, suffer from modality interference that degrades dense geometric prediction. Conversely, we show that a modality-specialized Dual Encoder trained with a pure masked image modeling (MIM) objective better preserves high-frequency spatial detail and achieves strong transfer performance on building instance segmentation and height regression.

The main contributions of this work are summarized as follows:

- **Curating the Largest High-Resolution Paired RGB-DSM Dataset:** We introduce and publicly release a dataset of 1.2 million aligned RGB-DSM image tiles derived from native source products primarily between 0.1–0.5 m spatial resolution and processed to training resolutions of 0.2–0.5 m where appropriate. This dataset represents a >30× increase over existing paired datasets, addressing the data scarcity that has previously constrained geometry-aware pre-training.
- **Identifying Modality Interference in Shared Architectures:** Through a rigorous factorial analysis, we demonstrate that standard Shared Encoder architectures suffer from severe modality interference when fusing texture and elevation. We show that modality-specialized Dual Encoders are required to prevent a performance collapse on dense geometric tasks (e.g., recovering a 13.5 AP loss on instance segmentation).
- **Demonstrating the Role of Data Scale:** We provide empirical evidence that data diversity is more effective than repetition under our fixed-compute comparison. Our 1.2M model outperforms a baseline trained for 4× as many epochs on a smaller dataset, suggesting that scaling unique samples is an efficient path toward more robust urban representations.
- **Uncovering the Invariance–Equivariance Trade-off:** We show that while global contrastive alignment can aid some scene-level classification tasks, it degrades fine-grained geometric reasoning in our dense prediction experiments. Within the evaluated configurations, generative Masked Image Modeling provides stronger transfer for geometry-sensitive urban tasks such as building extraction and height estimation.

2. Related work

2.1. Foundation models in earth observation

The use of self-supervised learning (SSL) has gained increasing attention in Earth observation following its success in natural image analysis. As discussed in recent surveys (Y. Wang et al., 2022; Gupta et al., 2022), current research is moving toward general-purpose feature representations that can support a range of downstream tasks. Early adaptations of Transformer-based architectures, such as SatMAE (Cong et al., 2023), demonstrated that masked image modeling can be effectively applied to multi-spectral data through channel-aware masking and temporal encoding strategies. These models showed strong transfer performance on land cover classification and change detection benchmarks, often outperforming supervised baselines. Similarly, RingMo (Sun et al., 2023) demonstrated that generative pre-training can serve as an effective alternative to ImageNet initialization for high-resolution optical imagery, particularly for small-object recognition.

Recent EO foundation models have further expanded the scale, modality coverage, and transfer capabilities of remote sensing pre-training. SkySense (Guo et al., 2024) explores large-scale multi-modal EO pre-training, while RingMo-Aerial (Diao et al., 2025) and RingMoE (Bi et al., 2025) extend aerial and remote-sensing foundation modeling to broader downstream evaluations, including geometry-related tasks such as height estimation. These works highlight the rapid progress of EO foundation models and the growing interest in urban 3D understanding. However, they do not specifically investigate large-scale paired RGB–DSM pre-training or systematically analyze how RGB–DSM fusion architecture and learning objective affect dense geometric transfer.

Despite this progress, the current generation of EO foundation models remains predominantly 2D-centric. While recent advancements like Scale-MAE (Reed et al., 2023) have incorporated ground sample distance (GSD) as a conditioning variable to handle multi-scale inputs, and models like Prithvi (Jakubik et al., 2023) and SpectralGPT (Hong et al., 2024) have extended SSL to temporal and hyperspectral domains, they largely overlook the vertical dimension. Even vision-language models such as RemoteCLIP (Liu et al., 2024), which leverage text supervision for zero-shot recognition, operate primarily at a global semantic level and lack the dense geometric understanding required for urban reconstruction. By not explicitly modeling DSMs, these foundation models show limited sensitivity to vertical structure, which constrains their performance on geometry-dependent tasks such as building height estimation and dense urban monitoring (Rottensteiner et al., 2014).

2.2. Multi-modal fusion: From supervised to self-supervised

The integration of heterogeneous modalities, specifically optical texture (RGB) and elevation geometry (DSM) has been extensively studied in the supervised learning literature. Early fusion approaches, which

stack modalities channel-wise, often struggle due to the disparate statistical distributions of pixel intensities versus height fields (Schindler, 2012). Conversely, Late Fusion strategies employing separate encoders, such as FuseNet (Hazirbas et al., 2016) or multi-stream CNNs (Audebert et al., 2018), have historically yielded superior results by allowing modality-specific feature extraction before semantic integration. These supervised studies consistently show that while RGB captures high-frequency appearance, DSMs provide complementary structural information that benefits from distinct inductive biases.

In the context of self-supervised foundation models, this architectural trade-off remains an active area of investigation. Recent multimodal SSL frameworks such as CROMA (Fuller et al., 2023) and DeCUR (Wang et al., 2024) have proposed hybrid architectures to align radar (SAR) and optical imagery. DeCUR, for instance, explicitly decouples "common" vs. "unique" features to prevent modality collapse. However, these methods have largely been evaluated on relatively small paired datasets (<50k samples) or focus on SAR-optical fusion, where the physical scattering mechanisms differ fundamentally from the RGB-DSM relationship. The question of whether shared or specialized encoders are optimal for high-resolution urban geometry, at the scale of millions of samples, has not yet been systematically examined at this scale.

Furthermore, the evolution of fusion mechanisms offers critical context. While early works relied on simple concatenation, more sophisticated mechanisms such as Cross-Attention and Gated Fusion have recently gained traction in supervised settings (Huang et al., 2024). For instance, CMX (Zhang et al., 2023) proposed a cross-modal feature rectification module for RGB-X semantic segmentation, effectively calibrating the interaction between modalities. However, these specialized modules introduce significant computational overhead. Our work investigates whether a pure Dual Encoder design can achieve similar disentanglement efficiency without the complexity of dense cross-attention layers.

2.3. Cross-domain insights: RGB-Depth (RGB-D) learning

While RGB-DSM pre-training is still limited in remote sensing, the computer vision (CV) community has extensively studied RGB-Depth (RGB-D) representation learning. Bachmann et al. (2022) introduced MultiMAE, demonstrating that masked autoencoding can effectively learn cross-modal correlations by treating depth as a distinct tokenized modality rather than a pixel channel. Similarly, Girdhar et al. (2023) proposed ImageBind, which learns a joint embedding space across six modalities; however, its emphasis on global, embedding-level alignment does not explicitly preserve the dense geometric detail required for photogrammetric applications. In this work, we adapt these masked reconstruction principles, specifically the masked-out token strategy to the continuous elevation surfaces found in DSMs.

Successful CV approaches often employ modality-specific tokenization or separate backbones to preserve the geometric integrity of the depth channel. Our work adapts these insights to the geospatial domain. Unlike indoor depth maps, which often contain holes and sensor noise, geospatial DSMs represent continuous elevation surfaces with specific artifacts (e.g., occlusion shadows, interpolation smoothing). By bridging the gap between CV-based RGB-D learning and EO-specific data characteristics, we investigate architectural designs that are robust to the specific morphology of urban landscapes.

2.4. Learning objectives: Invariance vs. Equivariance

A critical design choice in multi-modal pre-training is the selection of the learning objective. Two primary paradigms dominate the landscape: Contrastive Learning (e.g., InfoNCE (van den Oord et al., 2019), CLIP (Radford et al., 2021)) and Masked Image Modeling (e.g., MAE (He et al., 2021), SimMIM (Xie et al., 2022)). Contrastive objectives encourage semantic invariance by mapping augmented views or paired modalities to similar representations. While this is highly effective for global scene classification (e.g., identifying a residential area), it can be detrimental to dense prediction tasks. By enforcing invariance, the model may learn to discard spatial details such as precise object orientation or fine-grained boundary information (Tschannen et al., 2020; Wang and Isola, 2020).

In contrast, masked image modeling (MIM) is an equivariant objective: it requires the model to reconstruct the exact input signal from partial observations, thereby forcing the preservation of high-frequency spatial details. For geometry-centric tasks such as building delineation or height regression, preserving this equivariant behavior is likely to be important. Our work empirically tests this theoretical distinction, investigating whether the global semantic alignment offered by contrastive learning complements or conflicts with the pixel-level precision required for 3D urban understanding.

In summary, existing literature has advanced along orthogonal axes: large-scale spectral pre-training (Cong et al., 2023), supervised multimodal fusion (Hazirbas et al., 2016), or semantic alignment via contrastive learning (Liu et al., 2024). However, a critical gap remains at the intersection of these fields: no prior work has systematically investigated large-scale generative pre-training for high-resolution geometric modalities. Our work addresses this gap by curating the largest paired RGB–DSM dataset to date and performing a controlled factorial analysis to isolate the effects of encoder specialization and objective design. In this context, masked image modeling is especially relevant because reconstructing masked RGB and DSM tokens requires the encoder to preserve local spatial context, texture boundaries, and elevation discontinuities, directly linking the pre-training objective to the dense geometric information needed for urban 3D understanding.

3. Methodology

3.1. Problem formulation

We formulate the task of geometry-aware urban pre-training as a self-supervised learning problem over a joint distribution of spectral and elevation data. Let $\mathcal{D} = \{(x_i^{rgb}, x_i^{dsm})\}_{i=1}^{N}$ denote an unlabeled dataset of N aligned image pairs, where each sample consists of an RGB orthophoto tile $x^{rgb} \in \mathbb{R}^{3\times H\times W}$ and a corresponding DSM tile $x^{dsm} \in \mathbb{R}^{1\times H\times W}$.

Our objective is to learn a parameterized encoder function f_θ that maps this multi-modal input to a unified latent representation z:

$$z = f_\theta(x^{rgb}, x^{dsm}) \tag{1}$$

Depending on the fusion architecture, f_θ is instantiated either as a shared encoder or as a composition of modality-specific encoders.

The quality of this representation is evaluated based on its transferability to downstream tasks, specifically its ability to preserve both semantic class separability (for classification) and high-frequency geometric details (for dense prediction). Unlike standard RGB-only pre-training, the core challenge lies in effectively fusing the two modalities without suppressing the unique statistical properties of the elevation data. An overview of the three evaluated configurations and their associated fusion and learning objectives is provided in Fig. 1, while the individual components are described in Section 3.7.

3.2. RGB–DSM pre-training dataset

The HiRes-FusedMIM pre-training corpus consists of 1.2 million paired RGB–DSM image chips curated from 56 urban regions. The data were collected primarily from open geospatial data portals provided by national, regional, and local government agencies. The corpus is dominated by high-resolution European coverage, particularly Germany, France, and Switzerland, and is complemented with additional global anchor regions to increase morphological diversity. The geographic and morphological distribution of the training chips is summarized in

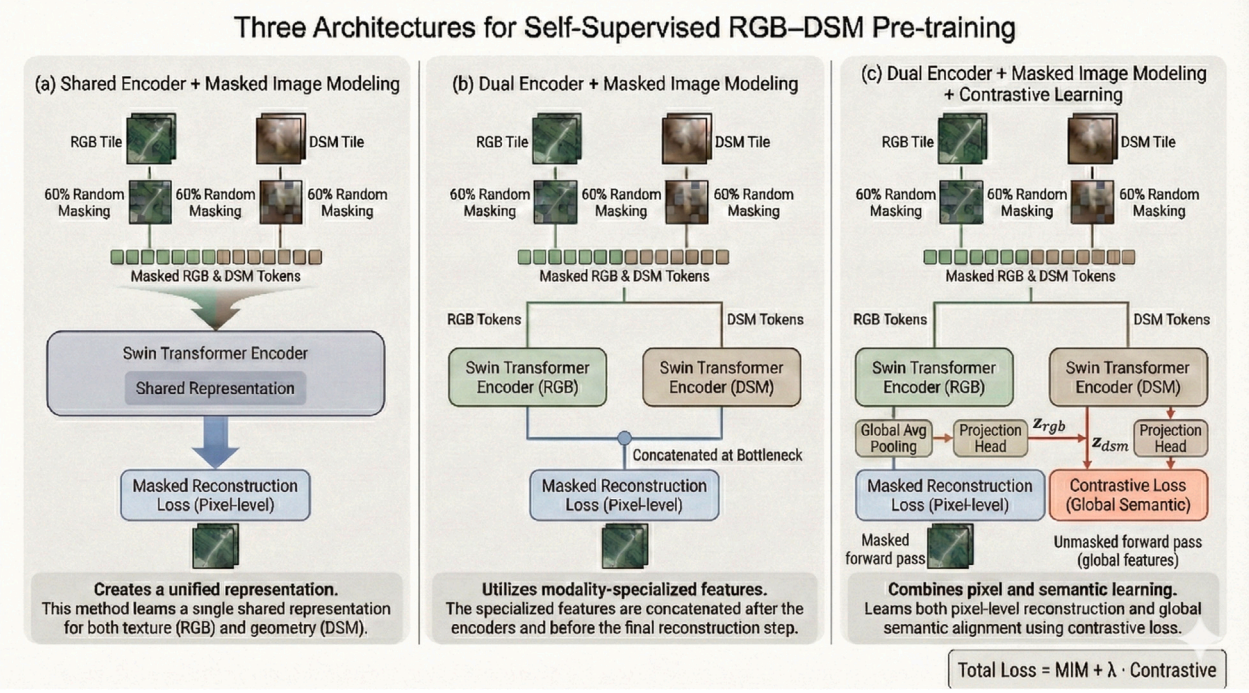

Fig. 1. Overview of the evaluated RGB–DSM fusion and pre-training strategies. We compare a Shared Encoder, which fuses RGB and DSM tokens early into a single spatial token grid, with a Dual Encoder, which processes each modality separately and fuses features later. For the dual-encoder setup, we evaluate masked image modeling alone (MIM) and masked image modeling with a global contrastive objective (MIM + Contrastive), resulting in three configurations. For this image, Generative AI was used only to assist with the initial visual layout of this schematic. The model configurations, data flow, labels, and scientific content were specified, manually revised, and verified by the authors; no experimental data or results were AI-generated.

Supplementary Table S1 and visualized in Supplementary Figure S1. The RGB data consist primarily of digital orthophotos derived from aerial surveys, while the DSM layers are official gridded surface-model products provided by the corresponding mapping agencies. Where true orthophotos were available from the source agency, we used them to reduce relief-displacement effects and improve spatial consistency with the DSM, especially near building edges. Since acquisition and production pipelines differ across regions, we provide a source-level provenance table in Supplementary Table S2, reporting the RGB product, DSM product, nominal spatial resolution, documented orthophoto type, DSM production source where specified, source category, and number of sampled chips. When a provider does not explicitly specify the orthophoto type or DSM-generation workflow, the corresponding metadata field is marked as not specified rather than inferred.

The source products have native ground sampling distances primarily between 0.1 m and 0.5 m, with a small number of regions available at coarser resolution. To balance computational efficiency with the preservation of structural features, we enforce a targeted spatial normalization strategy rather than global scaling: datasets with extremely fine native resolutions (such as the 0.1 m Swiss coverage) are downsampled to a 0.5 m grid, whereas datasets with native resolutions of 0.2 m to 0.5 m are kept completely intact. For city-level processing, the original RGB and DSM tiles are mosaicked into region-level rasters using GDAL virtual rasters (`gdalbuildvrt`) and exported as tiled GeoTIFFs at these designated training resolutions. When a common grid is required within a city or region, cubic convolution interpolation is used. During pre-training, the dataloader samples paired RGB and DSM chips of size 224×224 pixels from these aligned mosaics and applies identical spatial crops and geometric augmentations to both modalities.

This construction preserves the spatial characteristics of the original open geodata products while providing a unified paired RGB–DSM sampling interface for self-supervised learning. The resulting corpus covers diverse urban morphologies, including dense historical European centers, modern urban areas, alpine terrain, high-density megacity regions, and orthogonal grid-like layouts.

3.3. Use of pre-training data and leakage considerations

Consistent with standard practices for large-scale self-supervised foundation model pre-training, such as MAE (He et al., 2021) and SatMAE (Cong et al., 2023), we use the full 1.2M paired RGB–DSM tile corpus for unsupervised representation learning in order to maximize geographic and morphological diversity. Since the corpus contains no semantic labels, instance annotations, class labels, or height-regression targets, we do not impose an internal train/validation split during the pre-training phase.

To mitigate spatial data leakage, we examined the relationship between the pre-training corpus and the downstream evaluation datasets used in this study, including scene classification, semantic segmentation, instance segmentation, and height-regression tasks. Some downstream benchmarks may cover broad geographic regions that are similar to, or partially overlap with, regions represented in the unlabeled pre-training corpus. However, the pre-training data differ from the downstream benchmarks in one or more of acquisition time, sensor characteristics, source product, spatial resolution, tiling scheme, and annotation protocol. More importantly, no downstream labels, masks, instance annotations, height targets, validation samples, or test annotations are used during pre-training.

We therefore interpret the setting as domain-adaptive self-supervised pre-training over high-resolution urban RGB–DSM data, rather than transductive learning on labeled downstream test sets. All downstream results are obtained only after task-specific fine-tuning and evaluation using the official or established dataset splits where available, or using a fixed held-out split for the internally generated RGB-to-DSM height-regression diagnostic.

3.4. Data preprocessing and normalization

Proper normalization is required when fusing modalities with disparate physical units (radiance vs. meters).

DSM preprocessing. Raw DSM values represent absolute elevation above sea level, which varies significantly across geographic regions. To make the representation translation-invariant, we employ a local mean-centering strategy. For each DSM patch, we compute the local mean μ_{patch} over valid pixels and subtract it:

$$x_c^{dsm} = x^{dsm} - \mu_{patch} \tag{2}$$

This operation removes the absolute terrain bias while preserving relative building heights, which are the features of interest for urban understanding. To handle the high dynamic range of elevation data, we normalize the centered values by a global standard deviation σ_{dsm}, estimated robustly from the training corpus by discarding the top and bottom 5% of outliers:

$$\overline{x}^{dsm} = \frac{x_c^{dsm}}{\sigma_{dsm}} \tag{3}$$

RGB preprocessing. The spectral channels are normalized using standard z-score normalization based on dataset-level statistics calculated over the entire 1.2M tile corpus:

$$\tilde{x}_c^{rgb} = \frac{x_c^{rgb} - \mu_c}{\sigma_c}, \quad c \in \{R, G, B\} \tag{4}$$

where x_c^{rgb} denotes the cth RGB channel.

Finally, to ensure spatial correspondence during training, identical geometric augmentations (random resized crops and horizontal flips) are applied synchronously to both the RGB and DSM inputs.

3.5. Backbone and tokenization

We employ a Hierarchical Vision Transformer (Swin Transformer) (Liu et al., 2021) as the backbone encoder. This architecture has become widely adopted in dense remote sensing tasks due to its shifted window mechanism, which models local interactions efficiently while maintaining a global receptive field, which is well suited for analyzing elongated urban structures such as bridges and roads (He et al., 2022; L. Wang et al., 2022).

The input images are first serialized into a sequence of discrete tokens. Given an input image x (either RGB or DSM), a patch embedding layer partitions it into non-overlapping patches of size $P \times P$ and projects them into a C-dimensional embedding space:

$$t = \text{PatchEmbed}(x) \in \mathbb{R}^{L \times C}, \quad \text{where } L = \frac{H}{P} \times \frac{W}{P} \tag{5}$$

This tokenization step is the entry point for fusion. As detailed in Section 3.7, the strategy for processing these tokens constitutes the primary architectural variable in our investigation.

3.6. Masked image modeling objective

We adopt a MIM objective to encourage the model to learn structure-preserving representations.

Masking strategy. We employ a random masking strategy with a high masking ratio of $r = 60\%$. A binary mask $m_{mask} \in \{0, 1\}^L$ is generated, where 1 indicates a masked patch. The visible tokens are kept, while masked tokens are replaced by a learnable mask token t_{mask}:

$$\tilde{t} = (1 - m_{mask}) \odot t + m_{mask} \odot t_{mask} \tag{6}$$

Reconstruction loss. A lightweight decoder reconstructs the original pixel values from the latent features of the visible patches. The loss is computed only on the masked regions Ω using the $\mathcal{L}_1$ distance, which promotes sharper boundaries than $\mathcal{L}_2$:

$$\mathcal{L}_{MIM} = \underbrace{\frac{1}{|\Omega|} \sum_{i \in \Omega} \|\hat{x}_i^{rgb} - x_i^{rgb}\|_1}_{\text{Texture Reconstruction}} + \underbrace{\frac{1}{|\Omega|} \sum_{i \in \Omega} \|\hat{x}_i^{dsm} - x_i^{dsm}\|_1}_{\text{Geometry Reconstruction}} \tag{7}$$

This dual reconstruction objective prevents the model from minimizing the loss by relying on a single modality.

For RGB–DSM pre-training, this reconstruction formulation is particularly important because the two modalities encode complementary spatial signals. RGB reconstruction encourages preservation of texture, material transitions, and local appearance cues, whereas DSM reconstruction encourages preservation of elevation discontinuities, roof boundaries, and relative height structure. Since the loss is evaluated on masked regions, the encoder must infer missing local structure from the surrounding spatial context rather than only learning global semantic invariances. This makes MIM well aligned with dense urban tasks such as building delineation, instance separation, and height regression, where fine-grained geometric detail is required during transfer.

3.7. RGB–DSM fusion architectures

To isolate the effects of architectural integration, we evaluate two distinct fusion paradigms: a parameter-efficient Shared Encoder (Early Fusion) and a modality-specialized Dual Encoder (Late Fusion). Although masked image modeling and contrastive learning are established self-supervised objectives, their direct application to paired RGB–DSM pre-training requires architectural choices that account for the heterogeneous nature of texture and elevation. The structural focus of our design is therefore not a new self-supervised loss, but the adaptation and controlled evaluation of fusion mechanisms for high-resolution RGB–DSM data.

Specifically, we separate the modality interface from the learning objective by comparing two fusion strategies under the same backbone family, masking policy, reconstruction loss, and training schedule. The Shared Encoder variant tests whether RGB and DSM tokens can be processed in a common Transformer space after modality-specific patch embedding and explicit modality encoding. In contrast, the Dual Encoder variant assigns independent Swin Transformer streams to RGB and DSM inputs, allowing texture- and height-specific features to be extracted before late bottleneck fusion. Both variants use synchronized spatial masking and reconstruct both RGB and DSM targets, forcing the representation to preserve information from both appearance and geometry. This controlled design enables us to isolate whether performance differences arise from the fusion architecture itself rather than from differences in objective, data, or optimization protocol.

3.7.1. Variant A: Shared encoder (spatial token concatenation)

This variant explores whether a single Transformer backbone can jointly model spectral (RGB) and geometric (DSM) information. To accommodate heterogeneous input channels while preserving spatial structure, we adopt a spatial token concatenation strategy. RGB and DSM inputs are first patch-embedded independently:

$$t^{rgb} = \text{PatchEmbed}_{rgb}(x^{rgb}), \quad t^{dsm} = \text{PatchEmbed}_{dsm}(x^{dsm}), \tag{8}$$

where $t^{rgb}, t^{dsm} \in \mathbb{R}^{H_p W_p \times C}$. To retain modality identity and geographic context, learnable modality embeddings (e_m) and sinusoidal geolocation embeddings (e_{geo}) are added:

$$z^m = t^m + e_m + e_{geo}, \quad m \in \{rgb, dsm\}. \tag{9}$$

The resulting token grids are reshaped to (H_p, W_p) and concatenated along the spatial width dimension, forming a fused grid of size $(H_p, 2W_p)$, which is then flattened in row-major order and processed by a single Transformer encoder:

$$Z_{\text{input}} = \text{Flatten}\big(\text{Concat}([z^{rgb}, z^{dsm}], \text{dim} = W)\big). \tag{10}$$

This early fusion forces self-attention to model cross-modal interactions from the first layer, which can lead to modality interference as attention heads must jointly capture appearance-driven semantics and height-driven geometric continuity.

3.7.2. Variant B: Dual encoder (modality-specialized streams)

This architecture decouples feature extraction using two parallel Swin encoders, E_{rgb} and E_{dsm}, which share no weights:

$$r^{rgb} = E_{rgb}(x^{rgb}, m_{mask}), \quad r^{dsm} = E_{dsm}(x^{dsm}, m_{mask}) \tag{11}$$

Fusion occurs only at the bottleneck stage, where features are concatenated before decoding. This design allows each stream to learn modality-specific inductive biases while reducing gradient interference between RGB texture and DSM elevation features.

3.8. Global contrastive alignment

For the MIM + Contrastive variant, we introduce a global cross-modal alignment objective. To extract stable semantic representations, we perform a separate forward pass on the unmasked RGB and DSM inputs. The encoder outputs are globally aggregated using Global Average Pooling (GAP) and passed through modality-specific projection heads $h(\cdot)$ to obtain normalized global embeddings:

$$z_i^{rgb} = \frac{h_{rgb}(v_i^{rgb})}{\|h_{rgb}(v_i^{rgb})\|_2}, \quad z_i^{dsm} = \frac{h_{dsm}(v_i^{dsm})}{\|h_{dsm}(v_i^{dsm})\|_2}, \tag{12}$$

where v_i^{rgb} and v_i^{dsm} denote the pooled encoder features for the ith RGB and DSM sample, respectively.

We minimize a symmetric InfoNCE loss with temperature $\tau = 0.07$:

$$\mathcal{L}_{con} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp\left(z_i^{rgb}\cdot z_i^{dsm}/\tau\right)}{\sum_{j=1}^{B}\exp\left(z_i^{rgb}\cdot z_j^{dsm}/\tau\right)} + \log\frac{\exp\left(z_i^{dsm}\cdot z_i^{rgb}/\tau\right)}{\sum_{j=1}^{B}\exp\left(z_i^{dsm}\cdot z_j^{rgb}/\tau\right)}\right], \tag{13}$$

where B is the batch size.

The total objective is $\mathcal{L}_{total} = \mathcal{L}_{MIM} + \lambda\mathcal{L}_{con}$, with $\lambda = 0.05$.

The weight $\lambda = 0.05$ was empirically selected to ensure that the backpropagated gradient magnitudes of the contrastive loss $\mathcal{L}_{con}$ and the reconstruction loss $\mathcal{L}_{MIM}$ remained comparable during training. This prevents either objective from dominating the optimization dynamics and is consistent with gradient balancing principles commonly adopted in multi-task learning (Chen et al., 2018; Kendall et al., 2018).

3.9. Training protocol

All models are implemented in PyTorch and trained on 4 NVIDIA A100 GPUs.

- **Optimization:** AdamW optimizer, base learning rate 2×10^{-4} scaled linearly with batch size ($lr = \text{base\_lr} \times \frac{\text{batch\_size}}{512}$), cosine decay with 5-epoch warm-up.
- **Compute Budget:** Fixed budget of 100 epochs, batch size 512.
- **Augmentation:** Synchronized random cropping, flipping, and rotation. Color jitter is applied only to RGB.

For the data scale ablation, the 368k baseline was trained for 400 epochs to ensure a comparable total compute budget to the 1.2M model (trained for 100 epochs)

4. Results

4.1. Evaluation protocol

We evaluate the transferability of the pre-trained representations across a diverse suite of downstream tasks. To ensure rigorous benchmarking, we compare our HiRes-FusedMIM variants (trained on 1.2M pairs) against a standard supervised baseline initialized with ImageNet-1k weights. This allows us to isolate the specific contribution of domain-specific multi-modal pre-training versus generic visual feature learning.

The evaluated variants are:

- **SharedEnc (1.2M):** The Shared Encoder fusion baseline (Variant A).
- **Dual-MIM (1.2M):** The modality-specialized Dual Encoder with pure MIM (Variant B).
- **Dual-MIM+Con (1.2M):** The Dual Encoder augmented with global contrastive alignment.

4.2. Downstream evaluation benchmarks

To evaluate the transferability of the learned representations, we use downstream tasks covering scene-level classification, semantic segmentation, instance segmentation, and height regression. These tasks probe complementary aspects of the representation, including global semantic separability, dense boundary localization, instance-level building separation, and continuous geometric prediction. Table 1 summarizes the datasets, input modalities used during fine-tuning, evaluation metrics, and the role of each benchmark in the evaluation. For height regression, we use Track 2 of the 2023 IEEE GRSS Data Fusion Contest, which defines joint building extraction and height estimation for semantic urban reconstruction. The dataset covers 17 cities across six continents and provides satellite imagery with normalized DSM (nDSM) height references. The optical/SAR imagery is collected from SuperView-1/GaoJing-1 and Gaofen-2/Gaofen-3 satellites at approximately 0.5–1 m spatial resolution, while the nDSM references are generated from Gaofen-7 and WorldView stereo imagery at approximately 2 m GSD. We adapt this benchmark to RGB-to-height regression by using only RGB imagery as input and nDSM as output, fine-tuning with a modified MMSegmentation image-to-image regression pipeline. We train on 1474 files and report RMSE in meters on 300 validation images.

4.3. Qualitative analysis: Reconstruction fidelity

The reconstruction quality (Fig. 2) serves as our first diagnostic tool for modality interference.

Visual evidence of interference. Comparing the DSM reconstructions (columns 3 vs. 4), the Shared Encoder consistently produces smoothing artifacts around building boundaries. For example, in the dense residential blocks shown in Row 2, individual building footprints are blurred into a continuous elevation mound. This observation aligns with established findings in urban photogrammetry, where surface smoothing is a known artifact when trying to model sharp discontinuities (e.g., building eaves) using continuous basis functions or band-limited features (Qin et al., 2016; Nex and Remondino, 2014). Furthermore, the statistical domain gap between the high-frequency texture of RGB and the piecewise-smooth nature of DSMs often leads to feature contention in shared operational spaces, a phenomenon also observed in recent multi-modal medical imaging studies (Huang et al., 2023). In contrast, the Dual Encoder preserves the sharp, high-frequency discontinuities required to distinguish adjacent structures. This visual degradation in the Shared Encoder directly predicts the downstream failures in instance segmentation.

4.4. The cost of modality interference: Instance segmentation

Instance segmentation on UBCv2 is the most geometry-sensitive task in our suite. The results in Table 2 reveal the most critical finding of our study.

Table 1
Downstream evaluation benchmarks. The evaluation covers scene classification, semantic segmentation, instance segmentation, and height regression. For Vaihingen and GeoNRW, we report both RGB-only transfer, where only the RGB encoder is used, and RGB–DSM transfer, where both modality-specific encoders are used. The height-regression dataset is 2023 IEEE Data Fusion Contest Track-2 dataset and is used only to assess geometry-aware transfer.

Dataset	Task	Fine-tuning input	Metric	Purpose in evaluation
UCM	Scene classification	RGB	Accuracy	Scene-level semantic transfer
PatternNet	Scene classification	RGB	Accuracy	Large-scale aerial scene recognition
NWPU-RESISC45	Scene classification	RGB	Accuracy	Cross-scene generalization under high class diversity
AID	Scene classification	RGB	Accuracy	Aerial scene classification with varied spatial layouts
WHU Aerial	Semantic segmentation	RGB	mIoU	Building/urban segmentation in high-resolution aerial imagery
LoveDA	Semantic segmentation	RGB	mIoU	Urban–rural domain generalization
SpaceNetV1	Semantic segmentation	RGB	mIoU	Building extraction on a distinct urban benchmark
Vaihingen (RGB)	Semantic segmentation	RGB encoder only	mIoU	RGB-only dense urban segmentation transfer
Vaihingen (RGB–DSM)	Semantic segmentation	RGB and DSM encoders	mIoU	Effect of using both modalities in dense urban segmentation
GeoNRW (RGB)	Semantic segmentation	RGB encoder only	mIoU	RGB-only transfer on high-resolution urban scenes
GeoNRW (RGB–DSM)	Semantic segmentation	RGB and DSM encoders	mIoU	Geometry-sensitive segmentation using paired RGB–DSM inputs
UBCv2	Instance segmentation	RGB	AP50	Building instance separation and boundary-sensitive transfer
IEEE DFC 2023	Height regression	RGB	RMSE (m)	Continuous geometric prediction from learned image features

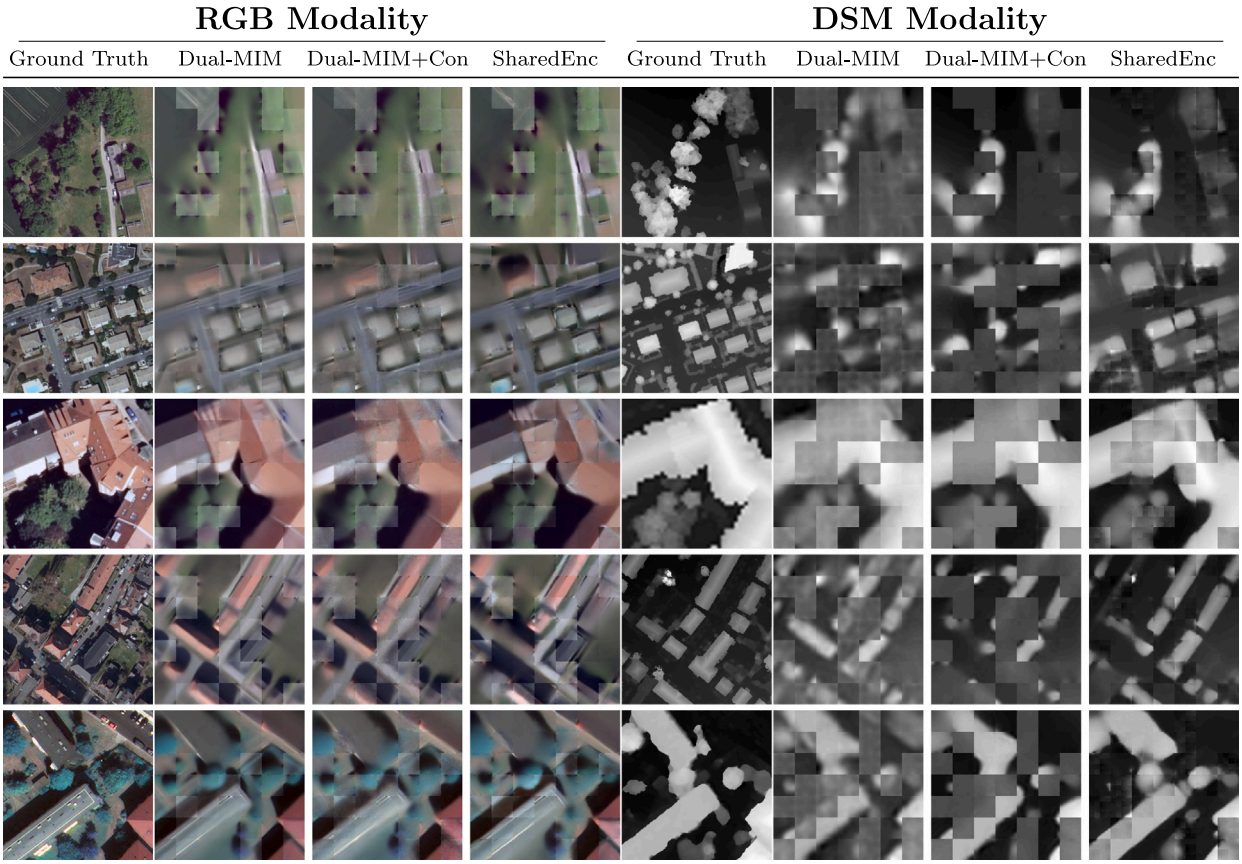

Fig. 2. Qualitative comparison of reconstruction fidelity for RGB and DSM modalities. The SharedEnc produces smoothed DSM surfaces that blur building boundaries, while the Dual-MIM variants preserve sharp geometric discontinuities. These reconstruction differences foreshadow downstream failures in geometry-sensitive tasks such as instance segmentation.

Failure of early fusion. Table 2 shows that SharedEnc (1.2M) performs substantially worse than all other initialization strategies, reaching only 18.0 AP50 on UBCv2. In contrast, Dual-MIM (1.2M), which uses modality-specialized RGB and DSM encoders with the same MIM objective, reaches 31.5 AP50. This corresponds to a 13.5 AP50 improvement over SharedEnc (1.2M) and a small gain over the ImageNet-1k baseline (31.2 AP50). The result indicates that the performance loss is not caused by RGB–DSM pre-training itself, but by forcing texture and elevation information into a shared encoder too early. The qualitative examples in Fig. 3 support this quantitative result: SharedEnc frequently merges adjacent buildings into single instances, whereas Dual-MIM better preserves individual building boundaries and instance separation.

Effect of contrastive alignment. Dual-MIM+Con (1.2M) reaches 29.0 AP50, which is 2.5 AP50 below Dual-MIM (1.2M). This suggests that the global contrastive objective can reduce boundary-level precision required for instance separation, even though it may help some scene-level classification tasks. Therefore, for UBCv2 instance segmentation, the pure MIM objective paired with modality-specialized fusion provides the strongest transfer performance.

Fig. 3. Instance segmentation results on the dense UBCv2 benchmark. This task requires precise separation of adjacent structures. The Shared Encoder (rightmost column) suffers from "blobbing", merging neighboring buildings into single instances due to smoothed geometric features. The Dual Encoder (Column 3) maintains the necessary high-frequency precision to delineate individual instances, explaining the massive +13.5 AP performance gap observed quantitatively.

Table 2
Instance segmentation on UBCv2. We report AP50 for building instance segmentation. The comparison includes ImageNet-1k initialization, the earlier 368k MIM baseline, and the three 1.2M HiRes-FusedMIM variants. The SharedEnc (1.2M) variant drops to 18.0 AP50, whereas Dual-MIM (1.2M) reaches 31.5 AP50, recovering 13.5 AP50 over the shared-encoder design and slightly exceeding the ImageNet-1k baseline. Higher is better.

Model	Pre-training setting	AP50
ImageNet-1k	Supervised initialization	31.2
MIM (368k)	RGB–DSM, Dual encoder	28.5
SharedEnc (1.2M)	RGB–DSM, Shared encoder + MIM	18.0
Dual-MIM (1.2M)	RGB–DSM, Dual encoder + MIM	**31.5**
Dual-MIM+Con (1.2M)	RGB–DSM, Dual encoder + MIM + Contrastive	29.0

4.5. Semantic segmentation and generalization

We evaluate semantic segmentation across high-resolution remote-sensing benchmarks representing different urban morphologies and annotation settings. Table 3 consolidates representative external remote-sensing pre-training baselines with our evaluated RGB–DSM pre-training variants. The external methods are included to contextualize performance against established EO representation-learning approaches. However, because these methods differ in pre-training corpus, input modality, backbone, decoder, and fine-tuning protocol, the external rows should not be interpreted as a strictly controlled leaderboard. The controlled comparison in this paper is given by the HiRes-FusedMIM variants evaluated under the same downstream setup.

Comparison with external EO pre-training baselines. Table 3 shows that the proposed Dual-MIM representation is competitive with established remote-sensing pre-training methods across the evaluated high-

Table 3
Semantic segmentation comparison on high-resolution remote-sensing benchmarks We report mIoU for representative remote-sensing pre-training methods and our evaluated RGB–DSM pre-training variants. External methods are included for contextual comparison because they differ in pre-training corpus, input modality, backbone, decoder, and fine-tuning protocol. The controlled comparison in this paper is given by the HiRes-FusedMIM variants evaluated under the same downstream setup. For readability, the best, second-best, and third-best reported values in each column are shown in **bold**, underline, and *italics*, respectively. These ranks are computed only among reported values and should be interpreted contextually rather than as a strict leaderboard. The names SharedEnc, Dual-MIM, and Dual-MIM+Con denote the same 1.2M RGB–DSM variants used throughout the ablation study. Higher is better.

Method	Backbone	WHU	LoveDA	Vaihingen	GeoNRW	SpaceNetV1
External/published baselines						
ImageNet-1k	Swin-B	88.50	**54.37**	*73.60*	60.30	76.42
SeCo (Mañas et al., 2021)	ResNet-50	86.70	43.63	68.90	–	73.89
GASSL (Ayush et al., 2022)	ResNet-50	–	48.76	–	–	**78.51**
SatMAE (Cong et al., 2023)	ViT-L	82.50	–	70.60	–	78.07
GFM (Mendieta et al., 2023)	Swin-B	*90.70*	–	**75.30**	–	72.81
BillionFM (Cha et al., 2026)	ViT-L	–	52.38	–	–	–
Ours: controlled RGB–DSM pre-training variants						
SharedEnc (1.2M)	Swin-B	88.98	41.58	69.15	59.10	77.62
Dual-MIM (1.2M)	Swin-B	**91.40**	*52.28*	74.58	**61.68**	78.37
Dual-MIM+Con (1.2M)	Swin-B	91.08	52.04	72.54	*60.02*	*78.08*

resolution segmentation benchmarks. Dual-MIM achieves the best reported result on WHU and GeoNRW, the second-best reported result on Vaihingen and SpaceNetV1, and remains among the top reported methods on LoveDA. We emphasize that these rankings are contextual because the compared methods use different pre-training data, input modalities, backbones, decoders, and fine-tuning protocols. Therefore, the table should be read as evidence of competitiveness rather than as a strict state-of-the-art ranking. Direct controlled comparison with recent large-scale models such as RingMo-Aerial, RingMoE, and SkySense for height estimation is left for future work because their public checkpoints, supported input modalities, pre-training corpora, and downstream fine-tuning protocols are not directly aligned with the paired RGB–DSM setting evaluated here.

Controlled comparison among RGB–DSM variants. The main controlled conclusion comes from the three HiRes-FusedMIM variants, which are evaluated under the same downstream setup. Across the semantic segmentation benchmarks, the Dual Encoder variants consistently outperform the Shared Encoder. The difference is particularly pronounced on LoveDA, where the Shared Encoder drops to 41.58 mIoU compared with 52.28 mIoU for Dual-MIM, corresponding to a 10.7 point gap. Similar degradation is observed on Vaihingen and GeoNRW. These results support the modality-interference hypothesis: forcing RGB texture and DSM-derived geometric information into a shared representation can reduce transfer performance, whereas modality-specific encoders better preserve geometry-sensitive features before fusion. The qualitative examples in Fig. 4 are consistent with these results: SharedEnc exhibits boundary erosion and missed building-edge pixels, whereas the Dual Encoder variants preserve sharper building perimeters across the evaluated datasets.

RGB-only versus RGB–DSM transfer. For datasets where paired elevation information is available, we additionally compare RGB-only fine-tuning with two-encoder RGB–DSM fine-tuning. In the RGB-only setting, only the RGB encoder is used, whereas in the RGB–DSM setting both modality-specific encoders are used during downstream training. The results show that the benefit of DSM information is dataset-dependent. On GeoNRW, using both RGB and DSM improves the Dual-MIM result from 59.32 to 61.68 mIoU, indicating that elevation cues are useful when the downstream labels and scene structure are strongly aligned with geometry. On Vaihingen, RGB-only transfer remains slightly stronger, suggesting that DSM benefits depend on factors such as DSM quality, co-registration, annotation protocol, and the relevance of height cues for the target classes. We therefore interpret DSMs as complementary geometric information rather than as a universally superior replacement for RGB-only transfer.

Table 4
Scene classification accuracy. We report top-1 accuracy (%) on four aerial scene classification datasets. The contrastive objective improves UCM but does not consistently improve all scene-level benchmarks; Dual-MIM remains stronger on PatternNet, RESISC45, and AID. Higher is better.

Dataset	UCM	PatternNet	RESISC45	AID
ImageNet-1k	99.0	97.97	89.2	87.8
SharedEnc (1.2M)	96.4	98.31	93.2	88.8
Dual-MIM (1.2M)	97.6	**99.54**	**95.1**	**96.7**
Dual-MIM+Con	**98.4**	99.44	94.6	96.3

Pre-training versus ImageNet initialization. The ImageNet-1k baseline remains competitive on some RGB-only semantic segmentation benchmarks, especially LoveDA, indicating that generic visual features can be strong for appearance-driven segmentation. However, Dual-MIM remains competitive across all reported datasets and performs particularly well on benchmarks where geometric structure is more relevant, such as GeoNRW and SpaceNetV1. Together with the height-regression results in Section 4.7, this suggests that the benefit of RGB–DSM pre-training is most evident when the downstream task requires geometry-aware representations rather than only semantic appearance cues.

4.6. Scene classification

In scene classification, the relative behavior of the objectives differs from the dense prediction tasks. As shown in Table 4, the Dual-MIM+Con variant improves over Dual-MIM on UCM by 0.8 percentage points, while Dual-MIM remains stronger on PatternNet, RESISC45, and AID. This indicates that global contrastive alignment can benefit simpler scene-level semantic categorization, but the gain is not uniform across all classification benchmarks.

The **Dual-MIM+Con** variant achieves the highest performance among our RGB–DSM variants on UCM, improving over Dual-MIM from 97.6% to 98.4%. However, this benefit does not generalize uniformly across all scene-classification datasets: Dual-MIM performs slightly better on PatternNet (99.54% vs. 99.44%), RESISC45 (95.1% vs. 94.6%), and AID (96.7% vs. 96.3%). Compared with ImageNet-1k initialization, the best RGB–DSM pre-trained model improves substantially on RESISC45 (+5.9 percentage points) and AID (+8.9 percentage points), while ImageNet remains slightly stronger on UCM. These results suggest that domain-specific RGB–DSM pre-training can improve transfer to more diverse aerial scene datasets, whereas the additional contrastive objective provides task-dependent rather than universal gains.

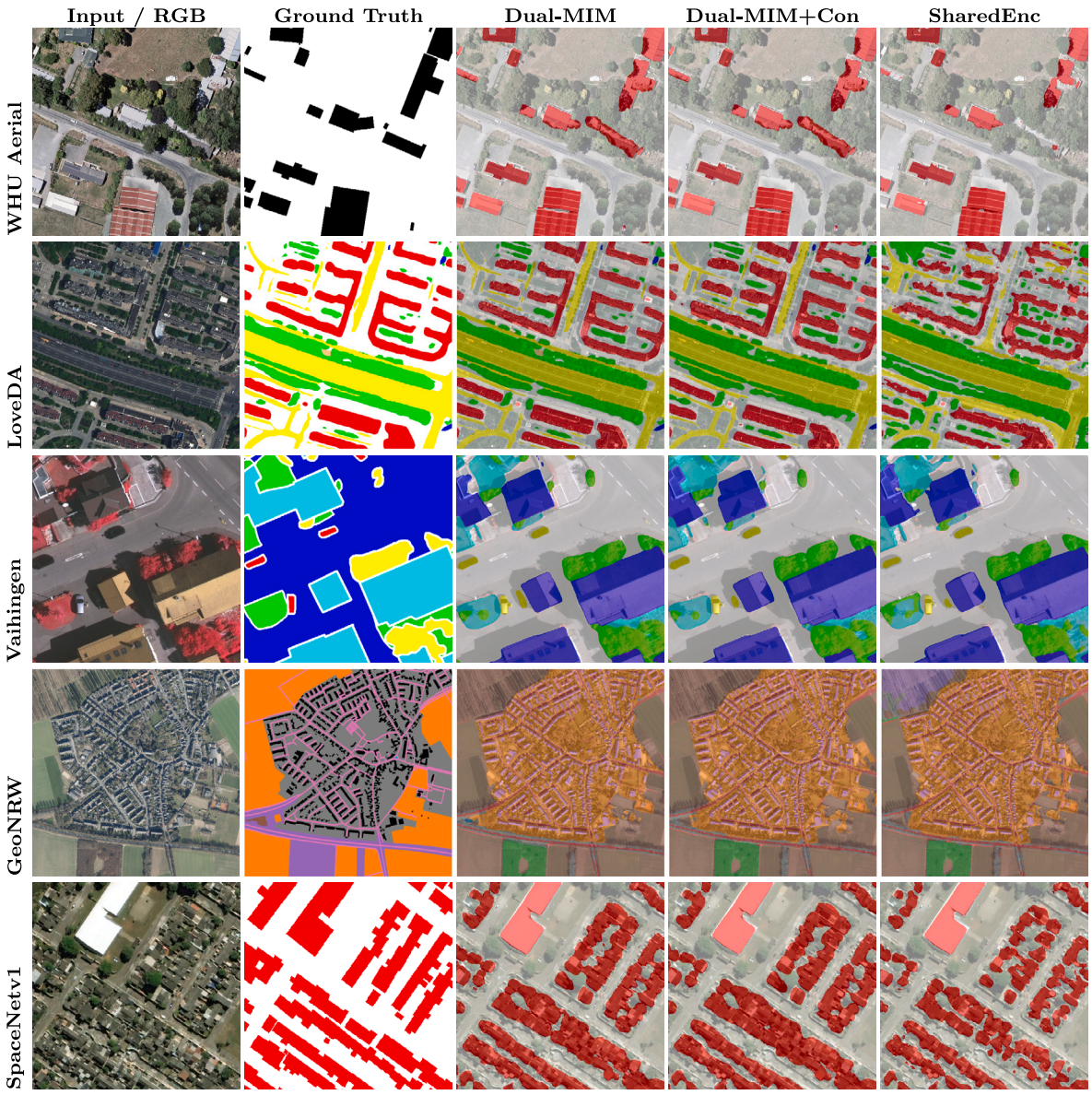

Fig. 4. Qualitative comparison of semantic segmentation performance. The Shared Encoder (Right Column) exhibits "boundary erosion", frequently misclassifying building edges as background due to geometric smoothing. This aligns with the "Modality Interference" hypothesis, where high-frequency structural details are lost in the shared latent space. In contrast, the Dual Encoder (Middle Column) recovers crisp building perimeters, significantly reducing false negatives at boundaries and driving the superior quantitative performance (e.g., +10.7 mIoU on LoveDA).

Table 5
Height estimation performance. Quantitative comparison of RGB-to-height regression for the evaluated RGB–DSM pre-training variants. Lower RMSE indicates better performance.

Pre-training variant	Objective	RMSE (m) ↓
Shared encoder	MIM	9.26
Dual encoder	MIM	7.89
Dual encoder	MIM + Contrastive	**7.82**
Shared – Dual-MIM	–	+1.37

4.7. Geometry-centric regression

We further evaluate geometry-aware transfer using single-image height estimation, formulated as RGB-to-height regression. This task is complementary to segmentation because it requires the representation to preserve continuous geometric structure rather than only semantic class separability. All variants are fine-tuned under the same downstream protocol and evaluated using root mean squared error (RMSE), where lower values indicate better height prediction accuracy.

As shown in Table 5, both Dual Encoder variants outperform the Shared Encoder on height regression. The Shared Encoder increases the error by 1.37 m RMSE compared with the Dual Encoder trained using the same MIM objective. This supports the modality-interference observation from the segmentation experiments: forcing RGB texture and DSM-derived geometric information into a shared representation can reduce the fidelity of height-sensitive features. The Dual Encoder variants better preserve geometric structure, with the MIM+Contrastive model achieving the lowest RMSE. Unlike instance segmentation, where the contrastive objective reduces AP50, height regression appears to benefit slightly from the additional global alignment, possibly because global scale consistency is useful for continuous height prediction. This result indicates that the effect of contrastive alignment

Table 6
Impact of fusion architecture. Comparison between the Shared encoder and Dual encoder under the same MIM objective and 1.2M pre-training setting. The Shared encoder shows lower transfer performance across geometry-sensitive downstream metrics. Lower is better for RMSE, while higher is better for AP50 and mIoU.

Architecture	Instance Seg. (UBCv2 AP50)	Semantic Seg. (LoveDA mIoU)	Height Est. (RMSE, m)
Shared encoder	18.0	41.58	9.26
Dual encoder	**31.5**	**52.28**	**7.89**
Shared – Dual	−13.5 AP50	−10.7 mIoU	+1.37 m

is task-dependent: it can reduce boundary-level precision for instance separation while still providing useful global consistency for continuous height regression.

Visual analysis of error distribution. Fig. 5 shows that the Shared Encoder often produces smoother height predictions but introduces larger localized errors around sharp height discontinuities, including courtyard boundaries, building edges, and elongated structures. In addition, some small or low-height buildings are suppressed by the Shared Encoder and, in certain cases, also by the Dual-MIM+Con variant, whereas Dual-MIM preserves these weak geometric signals more faithfully. This qualitative behavior is consistent with the quantitative RMSE results in Table 5, where the Dual Encoder variants reduce height-estimation error relative to the Shared Encoder. It also supports the interpretation that modality-specialized generative pre-training is particularly useful for preserving fine-grained urban geometry.

4.8. Summary of empirical findings

Synthesizing these results, we derive three design principles:

1. **Avoid Shared Encoders for Geometry:** The consistent failure of the Shared Encoder (18.0 AP on UBCv2, 41.58 mIoU on LoveDA) indicates that early fusion can induce severe modality interference. Architectural specialization appears important for geometry-centric tasks within the evaluated setting.
2. **Task-Specific Objectives:** There is no "one size fits all" objective. Contrastive learning aids global classification (UCM +0.8%) but harms dense segmentation (UBCv2 −2.5 AP). These results suggest that the choice of pre-training objective should be guided by the target task granularity (semantic vs. geometric).
3. **Avoiding Degradation from Improper Fusion:** The fact that improper fusion (SharedEnc) performs worse than ImageNet highlights a critical risk in multi-modal learning. Our Dual-MIM architecture provides a robust baseline that consistently matches or exceeds ImageNet initialization, while avoiding the failure modes observed in shared architectures.

5. Ablation studies

In this section, we systematically analyze the impact of the three core design pillars: fusion architecture, learning objective, and pre-training data scale. All ablation experiments are conducted under controlled conditions, varying only a single factor while keeping the remaining components fixed to ensure statistical validity.

5.1. Effect of fusion architecture: The modality gap

We first quantify the impact of architectural specialization by comparing the Shared Encoder (Variant A) against the Dual Encoder (Variant B), both trained with the MIM objective on the full 1.2M dataset.

As shown in Table 6, the choice of fusion architecture has a large observed effect on dense geometry-sensitive transfer within the evaluated setting. The Shared Encoder shows substantial degradation compared to the Dual Encoder on instance segmentation (−13.5 AP50), LoveDA semantic segmentation (−10.7 mIoU), and height estimation (+1.37 m RMSE). These results suggest that RGB texture and DSM elevation do not benefit from being forced into a fully shared weight space at the earliest stages of representation learning. We interpret this degradation as evidence of modality interference in the tested Shared Encoder design, where the network may suppress or average out high-frequency geometric signals that are important for instance separation and height-sensitive prediction. The Dual Encoder mitigates this effect by allowing modality-specific feature extraction before late fusion.

Table 7
Impact of learning objective. Adding global contrastive alignment (MIM+Con) yields mixed results: it improves simple semantic categorization (UCM) but degrades performance on fine-grained recognition (RESISC45) and dense geometry (UBCv2).

Objective	Classification (%)		Dense tasks	
	UCM	RESISC45	UBCv2 (AP)	Height (m)
MIM Only	97.6	**95.1**	**31.5**	7.89
MIM + Contrastive	**98.4**	94.6	29.0	**7.82**

5.2. Effect of learning objective: Invariance vs. Equivariance

Next, we investigate whether adding a global contrastive alignment loss ($\mathcal{L}_{con}$) complements the dense Masked Image Modeling objective ($\mathcal{L}_{MIM}$).

The results in Table 7 challenge the common assumption that contrastive alignment is universally beneficial.

- **Semantic Trade-off:** While contrastive learning boosts performance on UCM (+0.8%), it actually hurts performance on the more challenging RESISC45 dataset (−0.5%) and UBCv2 instance segmentation (−2.5 AP).
- **Mechanism:** Contrastive objectives optimize for semantic invariance, effectively suppressing nuisance details. However, in urban analysis, those details, such as exact orientation, object boundaries, and small height discontinuities, are often the target signal. The qualitative height-regression examples further suggest that global contrastive alignment may suppress weak local geometric structures, such as small or low-height buildings, even when the overall height field remains plausible. This suggests that pure generative reconstruction is particularly effective for boundary-sensitive dense tasks such as instance segmentation, while the effect of global contrastive alignment remains task-dependent.

5.3. Effect of pre-training data scale and efficiency

Finally, we validate the necessity of our large-scale data curation by comparing the Dual Encoder model trained on the initial pilot dataset (368k pairs) versus the full release (1.2M pairs).

Training protocol difference. Crucially, the training budget reveals a fundamental efficiency insight. The pilot 368k model was trained for 400 epochs (≈147 million total samples seen), whereas the final 1.2M model was trained for only 100 epochs (≈120 million total samples seen).

As shown in Table 8, increasing the unique data volume yields a substantial **+3.0 AP** gain on the challenging UBCv2 task. While the 1.2M model (100 epochs) outperforms the 368k model (400 epochs) despite lower total compute, we acknowledge that scale and diversity are conflated. However, given that the 368k subset was already comparable in scale to datasets commonly used for training standard CNN-based models, the performance jump suggests that the morphological variance introduced by the additional 800k+tiles (including global

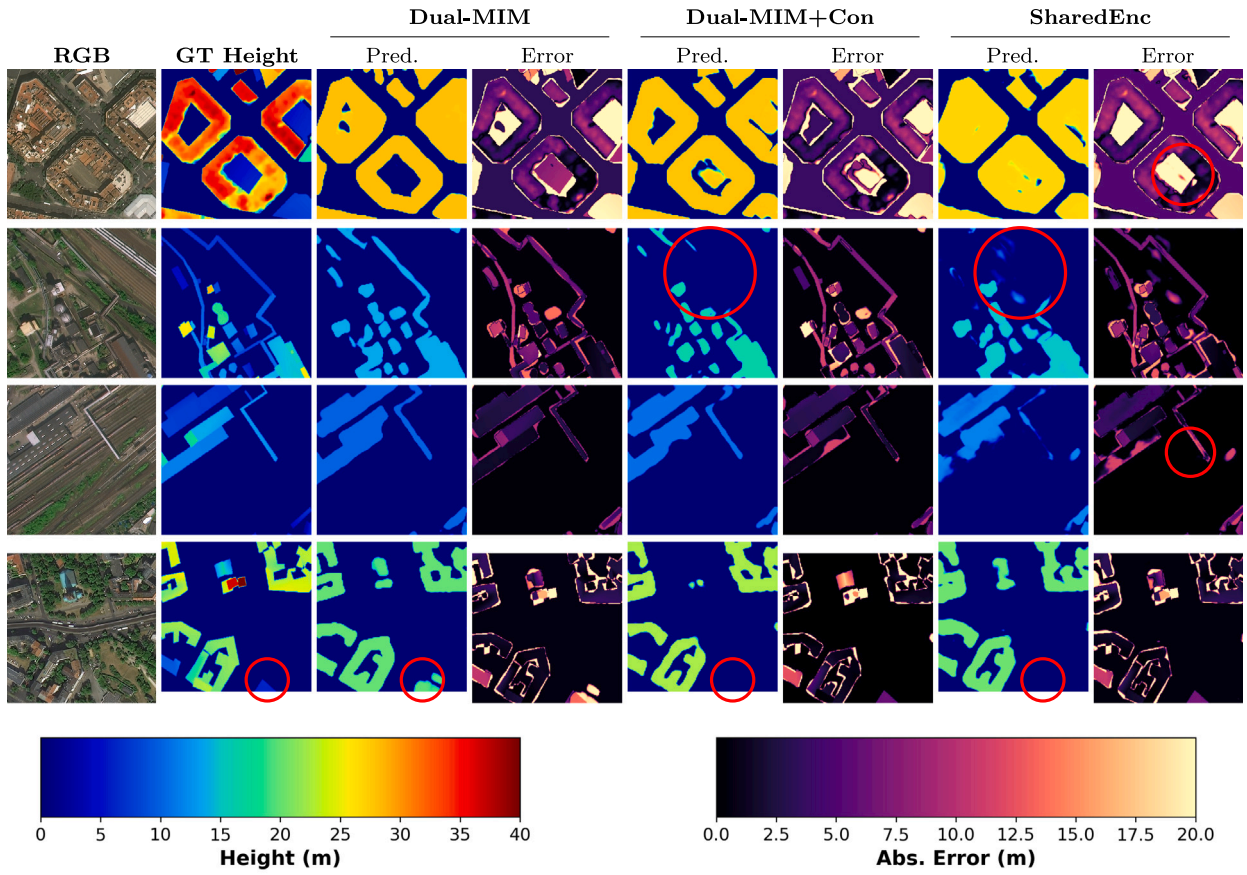

Fig. 5. Height regression results. Columns show the input RGB image, ground-truth height, and prediction/error pairs for the three model variants. Error maps show absolute height error. The highlighted regions illustrate two qualitative failure modes. First, the Shared Encoder produces smoother predictions and larger localized errors around sharp height discontinuities, including courtyards, elongated structures, and building edges. Second, small or low-height structures can be suppressed by SharedEnc and, in some cases, by Dual-MIM+Con, whereas Dual-MIM preserves these weak geometric signals more clearly.

Table 8
Impact of data scale. Despite training for fewer total iterations (100 epochs vs. 400 epochs), the 1.2M model outperforms the 368k baseline, highlighting that data diversity is more valuable than repetition.

Data scale	Epochs	Total samples	UBCv2 AP50	WHU mIoU
368k (Pilot)	400	≈147*M*	28.5	91.28
1.2M (Full)	**100**	**≈ 120M**	**31.5**	**91.40**
Improvement			**+3.0 AP**	+0.12 mIoU

anchors) provides a stronger learning signal than repeated exposure to the narrower pilot set. This aligns with recent findings in foundation models that favor unique tokens seen over total tokens processed.

For a visual breakdown of the geographic distribution and morphological diversity of the 1.2M training tiles, please refer to Figure S1 in the Supplementary material.

5.4. Summary of design choices

Based on these ablations, we identify the **Dual Encoder + Pure MIM** trained on **large-scale data** as the most effective configuration observed in our study for geometry-sensitive urban transfer. The three design factors influence performance in different ways. The largest observed changes on dense prediction tasks arise from the fusion architecture: replacing the Shared Encoder with the Dual Encoder improves UBCv2 instance segmentation by 13.5 AP50, LoveDA semantic segmentation by 10.7 mIoU, and height estimation by 1.37 m RMSE. In comparison, the global contrastive objective shows task-dependent behavior: it improves UCM scene classification and slightly improves height RMSE, but reduces UBCv2 instance-segmentation AP50, suggesting that global alignment can conflict with boundary-level spatial precision. Increasing the number of unique pre-training samples from 368k to 1.2M further improves transfer performance, supporting the importance of data diversity. These results do not imply a universal ranking of all possible RGB–DSM design choices; rather, they indicate that, within the evaluated configurations, modality-specialized fusion is particularly important for preserving geometric information, while objective design and data scale provide additional task-dependent effects.

6. Discussion

This work set out to systematically investigate the design space of foundation models for high-resolution urban remote sensing. By disentangling the effects of architecture, objective, and scale, several critical insights emerge that challenge current conventions in Earth Observation pre-training.

6.1. The modality interference hypothesis

One of the key findings of this study is the limited suitability of shared encoders for modeling heterogeneous geometric data. While shared encoders have become a standard in vision-language models, our results (e.g., the **13.5 AP drop** on instance segmentation) suggest that they are poorly suited for fusing RGB texture with DSM elevation. We attribute this to modality interference: the statistical distribution of elevation fields (continuous, gradient-based) is orthogonal to that of optical texture (high-frequency, reflectance-based). Forcing a single set

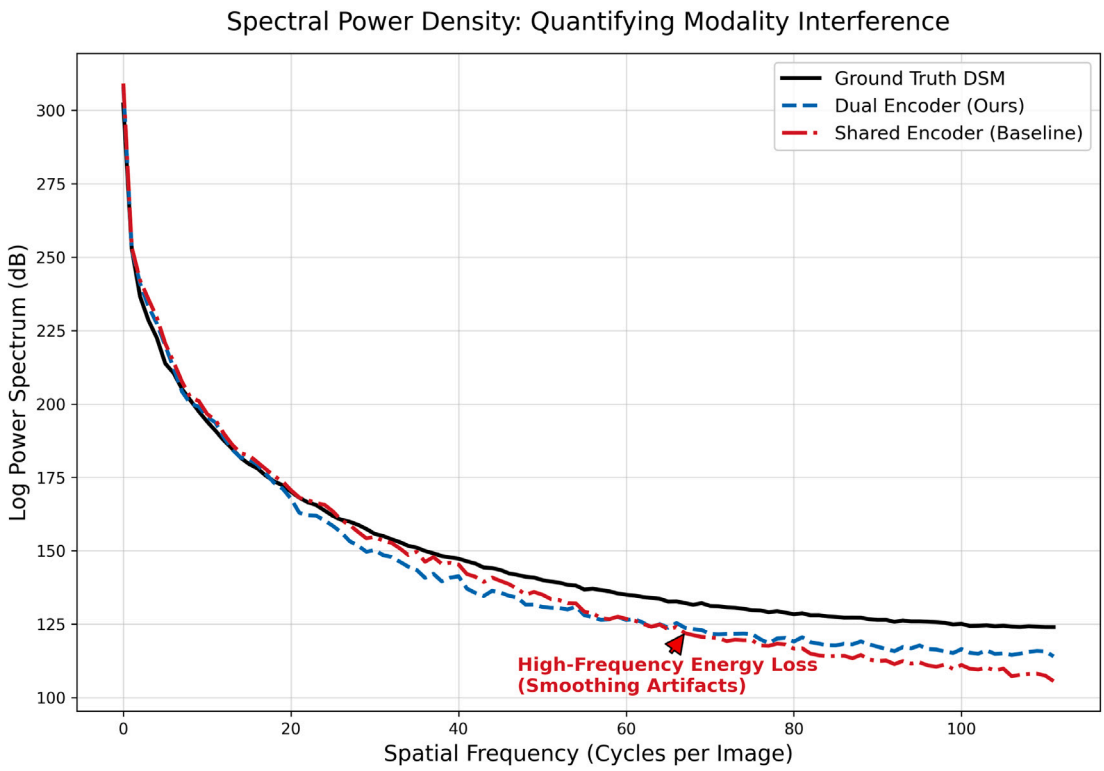

Fig. 6. Quantifying modality interference via spectral analysis. We plot the azimuthally averaged power spectrum of the DSM reconstructions. While both architectures capture global topography (low frequencies), the **Shared Encoder (Red)** suffers from significant energy loss at high frequencies compared to the **Dual Encoder (Blue)**. This spectral gap quantifies modality interference as a low-pass filtering effect, where the shared encoder fails to preserve the high-frequency discontinuities (edges) required for urban geometry.

of attention weights to process both distributions creates a bottleneck in which geometric precision is progressively attenuated. To quantify this interference mechanism, we perform a frequency-domain analysis of the reconstructed DSM surfaces, as shown in Fig. 6.

Spectral analysis protocol. To formalize the interference, we conducted a frequency domain analysis on $N = 100$ randomly sampled validation tiles. For each sample, we computed the 2D Fast Fourier Transform (FFT) of the mean-centered DSM predictions. The magnitude spectrum was azimuthally averaged to produce a 1D radial power profile, representing the energy retention across spatial frequencies.

The resulting analysis reveals that the Shared Encoder acts as a low-pass filter, retaining significantly less energy at high spatial frequencies (>40 cycles/image) compared to the Dual Encoder. This spectral gap confirms that the observed performance degradation is not merely semantic but stems from a fundamental inability to preserve high-frequency structural edges when weights are shared. This suggests that future geospatial foundation models should carefully balance parameter efficiency with modality specialization when geometry-sensitive transfer is the target.

6.2. Diversity supersedes repetition

Our data scaling ablation highlights the role of data diversity in representation learning. The 1.2M model, trained for 100 epochs (≈120M total samples), outperforms the 368k model trained for 400 epochs (≈147M total samples). This result indicates that increasing the number of unique training samples can be more effective than extending training duration on a fixed dataset.

Exposure to a broader range of urban morphologies in the expanded 1.2M corpus appears to support the learning of more transferable geometric representations than repeated exposure to a smaller and more homogeneous dataset. These findings suggest that, under a fixed compute budget, scaling data diversity is a practical and effective strategy for improving robustness in urban foundation models.

6.3. The invariance–equivariance trade-off

Our results reveal a consistent trade-off between learning objectives. We do not suggest that contrastive learning is inherently unsuitable for 3D tasks. Instead, our experiments show that global contrastive alignment (InfoNCE) favors semantic invariance, improving scene classification performance (e.g., +0.8% on UCM), while reducing performance on geometry-sensitive tasks such as instance segmentation (−2.5 AP on UBCv2).

By encouraging different views to be mapped to a shared latent representation, global objectives tend to down-weight spatial variations related to scale, orientation, and fine structural boundaries. While this behavior is beneficial for holistic scene understanding, it can limit performance in tasks that require precise geometric reasoning. Future work may explore local or dense contrastive formulations, such as pixel-level alignment, to better balance semantic consistency with spatial fidelity.

6.4. Scope of architectural analysis

Our analysis focuses on two contrasting architectural choices: fully shared encoders (early fusion) and fully disjoint modality-specific streams (late fusion). Within this scope, we observe that full parameter sharing without modality-specific conditioning performs poorly when modeling the heterogeneous statistical properties of RGB imagery and DSM elevation data.

We do not argue that shared encoders are universally unsuitable for multi-modal Earth observation tasks. Intermediate designs, including approaches based on low-rank adaptation (LoRA), gated cross-attention, modality-specific normalization, or hierarchical token-level

fusion, represent promising alternatives that warrant further investigation. Nonetheless, our results indicate that the modality gap between texture and elevation is sufficiently large that naive shared Transformer architectures struggle to model both modalities effectively without explicit architectural mechanisms.

Several additional limitations should be considered when interpreting the results. First, although the pre-training corpus is substantially larger than prior paired RGB–DSM datasets, its geographic coverage is influenced by the availability of open high-resolution geospatial products, with stronger representation from European regions. We partially mitigate this through additional global anchor regions, but further expansion to more geographic, climatic, and architectural contexts remains important. Second, the external comparisons reported in this paper are contextual rather than strictly controlled, because existing EO foundation models differ in pre-training data, input modality, backbone, decoder, and fine-tuning protocol. Therefore, our strongest conclusions are drawn from the controlled comparisons among the HiRes-FusedMIM variants. Finally, public redistribution of some RGB–DSM tiles may depend on the licensing conditions of the original geospatial data providers; for restricted sources, we provide metadata, provider links, and preprocessing scripts to support reproducibility.

7. Conclusion

In this work, we presented HiRes-FusedMIM, an empirical study of self-supervised multi-modal pre-training for high-resolution urban remote sensing. By curating and releasing a dataset of 1.2 million paired RGB and digital surface model (DSM) tiles, we enabled a systematic analysis of how foundation models can better incorporate geometric information that is absent in spectral-only pre-training.

Our experiments lead to three main observations:

1. **Architectural Specialization Matters:** Modality-specialized Dual Encoders consistently outperform parameter-efficient Shared Encoders on geometry-sensitive tasks, with substantial gains observed on instance segmentation benchmarks (e.g., +13.5 AP on UBCv2). This highlights the importance of architectural specialization when modeling heterogeneous spectral and elevation data.
2. **Objective Choice is Task-Dependent:** Pure Masked Image Modeling provides stronger transfer for boundary-sensitive dense prediction, particularly instance segmentation, whereas the additional contrastive objective shows mixed behavior and slightly improves height-regression RMSE. This suggests that preserving high-frequency spatial information through generative reconstruction is important for geometry-aware representation learning, while global alignment may be useful when continuous height prediction benefits from broader scale consistency.
3. **Data Diversity over Training Duration:** Increasing the diversity of unique training samples appears more effective than extending training duration on smaller datasets under the evaluated fixed-compute budget. Models trained on the full 1.2M corpus outperform those trained for substantially more epochs on reduced data, emphasizing the role of large-scale and diverse urban coverage.

We believe that the released dataset, together with the empirical insights reported in this study, will support future research on geometry-aware foundation models for urban Earth Observation and related 3D understanding tasks.

CRediT authorship contribution statement

Guneet Mutreja: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ksenia Bittner:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Funding

This work was supported by the DLR/DAAD Research Fellowship Programme – Doctoral Studies in Germany (grant no. 57575487).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ophoto.2026.100139.

Data and code availability

The data and code associated with this study are publicly available through the HiRes-FusedMIM repository: https://github.com/guneetmutreja/HiResFused-MIM. The repository provides access to the aligned RGB orthophoto and DSM GeoTIFF data used for pre-training, source-level metadata, trained model weights, and preprocessing and sampling scripts. During pre-training, approximately 1.2 million paired RGB–DSM image patches were dynamically sampled from the source rasters. For source regions where redistribution is restricted,source metadata, provider links, and scripts to reproduce the preprocessing and sampling pipeline are provided.

References

Albanwan, H., Qin, R., 2021. Adaptive and non-adaptive fusion algorithms analysis for digital surface model generated using census and convolutional neural networks. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. XLIII-B2-2021, 283–288. http://dx.doi.org/10.5194/isprs-archives-XLIII-B2-2021-283-2021, URL: https://isprs-archives.copernicus.org/articles/XLIII-B2-2021/283/2021/.

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

Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., Ermon, S., 2022. Geography-aware self-supervised learning. URL: https://arxiv.org/abs/2011.09980. arXiv:2011.09980.

Bachmann, R., Mizrahi, D., Atanasov, N., Pfrommer, B., Zamir, A.R., 2022. MultiMAE: Multi-modal multi-task masked autoencoders. In: European Conference on Computer Vision. ECCV, Springer, pp. 348–367.

Bi, H., et al., 2025. RingMoE: Mixture-of-modality-experts multi-modal foundation models for universal remote sensing image interpretation. arXiv preprint arXiv:2504.03166.

Cha, K., Seo, J., Lee, T., 2026. A billion-scale foundation model for remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 19, 16363–16378. http://dx.doi.org/10.1109/jstars.2024.3401772.

Chen, Z., Badrinarayanan, V., Lee, C.-Y., Rabinovich, A., 2018. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. URL: https://arxiv.org/abs/1711.02257. arXiv:1711.02257.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D.B., Ermon, S., 2023. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. URL: https://arxiv.org/abs/2207.08051. arXiv:2207.08051.

Diao, W., et al., 2025. RingMo-aerial: An aerial remote sensing foundation model with affine transformation contrastive learning. IEEE Trans. Geosci. Remote Sens..

Fu, Y., et al., 2025. DDDMNet: A DSM difference normalization module network for urban building change detection. ISPRS Int. J. Geo-Inform..

Fuller, A., Millard, K., Green, J.R., 2023. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. URL: https://arxiv.org/abs/2311.00566. arXiv:2311.00566.

Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., Bovolo, F., Chi, M., Anders, K., Gloaguen, R., Atkinson, P.M., Benediktsson, J.A., 2018. Multisource and multitemporal data fusion in remote sensing. URL: https://arxiv.org/abs/1812.08287. arXiv:1812.08287.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I., 2023. ImageBind: One embedding space to bind them all. URL: https://arxiv.org/abs/2305.05665. arXiv:2305.05665.

Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., He, H., Wang, J., Chen, J., Yang, M., Zhang, Y., Li, Y., 2024. SkySense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 27662–27673.

Gupta, T., Kamath, A., Kembhavi, A., Hoiem, D., 2022. Towards general purpose vision systems. URL: https://arxiv.org/abs/2104.00743. arXiv:2104.00743.

Han, B., Zhang, S., Shi, X., Reichstein, M., 2024. Bridging remote sensors with multisensor geospatial foundation models. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 27852–27862.

Hattula, E., Zhu, L., Raninen, J., 2024. Building extraction in urban and rural areas with aerial and LiDAR DSM. ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci. X-4/W4-2024, 73–79. http://dx.doi.org/10.5194/isprs-annals-X-4-W4-2024-73-2024.

Hazaymeh, K., Almagbile, A., Alsayed, A., 2023. A cascaded data fusion approach for extracting the rooftops of buildings in heterogeneous urban fabric using high spatial resolution satellite imagery and elevation data. Egypt. J. Remote. Sens. Space Sci..

Hazirbas, C., Ma, L., Domokos, C., Cremers, D., 2016. FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In: Asian Conference on Computer Vision. ACCV, Springer, pp. 213–228.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2021. Masked autoencoders are scalable vision learners. URL: https://arxiv.org/abs/2111.06377. arXiv:2111.06377.

He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., Xue, Y., 2022. Swin transformer embedding UNet for remote sensing image semantic segmentation. IEEE Trans. Geosci. Remote Sens. 60, 1–15. http://dx.doi.org/10.1109/TGRS.2022.3144165.

Hong, D., Zhang, B., Li, X., Li, Y., Li, C., Yao, J., Yokoya, N., Li, H., Ghamisi, P., Jia, X., Plaza, A., Gamba, P., Benediktsson, J.A., Chanussot, J., 2024. SpectralGPT: Spectral remote sensing foundation model. IEEE Trans. Pattern Anal. Mach. Intell. 46 (8), 5227–5244. http://dx.doi.org/10.1109/tpami.2024.3362475.

Huang, K., Shi, B., Li, X., Li, X., Huang, S., Li, Y., 2024. Multi-modal sensor fusion for auto driving perception: A survey. URL: https://arxiv.org/abs/2202.02703. arXiv:2202.02703.

Huang, S.-C., et al., 2023. Modality-specific learning rates for multi-modal medical image analysis. In: Medical Image Computing and Computer Assisted Intervention. MICCAI, Springer, pp. 1–11.

Jakubik, J., Roy, S., Phillips, C.E., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., Kimura, D., Simumba, N., Chu, L., Mukkavilli, S.K., Lambhate, D., Das, K., Bangalore, R., Oliveira, D., Muszynski, M., Ankur, K., Ramasubramanian, M., Gurung, I., Khallaghi, S., Hanxi, Li, Cecil, M., Ahmadi, M., Kordi, F., Alemohammad, H., Maskey, M., Ganti, R., Weldemariam, K., Ramachandran, R., 2023. Foundation models for generalist geospatial artificial intelligence. URL: https://arxiv.org/abs/2310.18660. arXiv:2310.18660.

Kendall, A., Gal, Y., Cipolla, R., 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. URL: https://arxiv.org/abs/1705.07115. arXiv:1705.07115.

Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., Chanussot, J., 2022. Deep learning in multimodal remote sensing data fusion: A comprehensive review. Int. J. Appl. Earth Obs. Geoinf. 112, 102926. http://dx.doi.org/10.1016/j.jag.2022.102926, URL: https://www.sciencedirect.com/science/article/pii/S1569843222001248.

Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J., 2024. Remoteclip: A vision language foundation model for remote sensing. URL: https://arxiv.org/abs/2306.11029. arXiv:2306.11029.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 9992–10002.

Mañas, O., Lacoste, A., Giró-i Nieto, X., Vazquez, D., Rodríguez, P., 2021. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 9394–9403.

Mendieta, M., Han, B., Shi, X., Zhu, Y., Chen, C., 2023. Towards geospatial foundation models via continual pretraining. In: 2023 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 16760–16770.

Nex, F., Remondino, F., 2014. UAV for 3D mapping applications: A review. Appl. Geomat. 6 (1), 1–15. http://dx.doi.org/10.1007/s12518-013-0120-x.

van den Oord, A., Li, Y., Vinyals, O., 2019. Representation learning with contrastive predictive coding.

Peng, Y., Sun, S., Wang, Z., Pan, Y., Li, R., 2019. Robust semantic segmentation by dense fusion network on blurred vhr remote sensing images. arXiv preprint arXiv:1903.02702.

Peng, Y., Sun, S., Wang, Z., Pan, Y., Li, R., 2020. Robust semantic segmentation by dense fusion network on blurred vhr remote sensing images. In: 2020 6th International Conference on Big Data and Information Analytics (BigDIA). pp. 142–145. http://dx.doi.org/10.1109/BigDIA51454.2020.00031.

Qin, R., Tian, J., Reinartz, P., 2016. 3D change detection – approaches and applications. ISPRS J. Photogramm. Remote Sens. 122, 41–56. http://dx.doi.org/10.1016/j.isprsjprs.2016.09.013.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. URL: https://arxiv.org/abs/2103.00020. arXiv:2103.00020.

Reed, C.J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T., 2023. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In: 2023 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 4065–4076.

Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J.D., Breitkopf, U., Jung, J., 2014. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction. ISPRS J. Photogramm. Remote Sens. 93, 256–271.

Schindler, K., 2012. An overview and comparison of smooth labeling methods for land-cover classification. IEEE Trans. Geosci. Remote Sens. 50 (11), 4534–4545.

Soergel, U. (Ed.), 2010. Radar Remote Sensing of Urban Areas. Springer, Dordrecht, The Netherlands.

Stewart, A.J., Lehmann, N., Corley, I.A., Wang, Y., Chang, Y.-C., Braham, N.A.A., Sehgal, S., Robinson, C., Banerjee, A., 2023. SSL4EO-L: Datasets and foundation models for landsat imagery. URL: https://arxiv.org/abs/2306.09424. arXiv:2306.09424.

Sun, X., Wang, P., Lu, W., Zhu, Z., Lu, X., He, Q., Li, J., Rong, X., Yang, Z., Chang, H., He, Q., Yang, G., Wang, R., Lu, J., Fu, K., 2023. RingMo: A remote sensing foundation model with masked image modeling. IEEE Trans. Geosci. Remote Sens. 61, 1–22.

Tschannen, M., Djolonga, J., Rubenstein, P.K., et al., 2020. On mutual information maximization for representation learning. arXiv preprint arXiv:1907.13625.

Wang, Y., Albrecht, C.M., Braham, N.A.A., Liu, C., Xiong, Z., Zhu, X.X., 2024. Decoupling common and unique representations for multimodal self-supervised learning. URL: https://arxiv.org/abs/2309.05300. arXiv:2309.05300.

Wang, Y., Albrecht, C.M., Braham, N.A.A., Mou, L., Zhu, X.X., 2022. Self-supervised learning in remote sensing: A review. URL: https://arxiv.org/abs/2206.13188. arXiv:2206.13188.

Wang, T., Isola, P., 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. Int. Conf. Mach. Learn. (ICML).

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M., 2022. UNetFormer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J. Photogramm. Remote Sens. 190, 196–214.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H., 2022. SimMIM: A simple framework for masked image modeling. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 9643–9653.

Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., Stiefelhagen, R., 2023. CMX: Cross-modal fusion for RGB-x semantic segmentation with transformers. URL: https://arxiv.org/abs/2203.04838. arXiv:2203.04838.