Fusing Convolution and Vision Transformer Encoders for Object Height Estimation from Monocular Satellite and Aerial Images

Supplementary Material

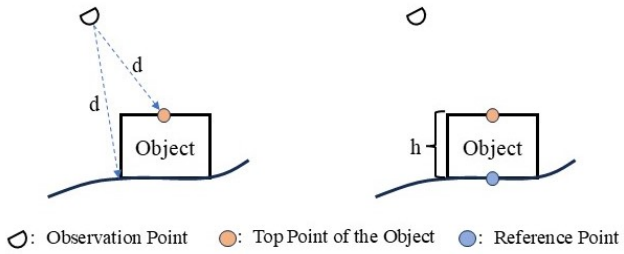

Figure 1. Visually comparison of the depth and height.

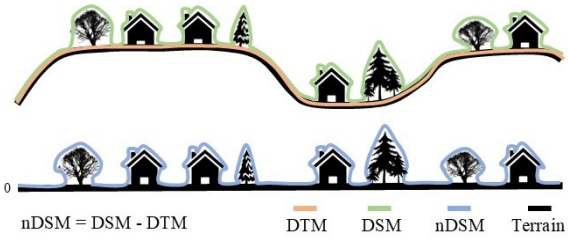

Figure 2. Representative illustration of Digital Terrain Model (DTM), Digital Surface Model (DSM), and Normalized Digital Surface Model (nDSM).

1. Data types

Depth and height information are extracted from various data sources, depending on the specific applications. In remote sensing (RS), height data is typically derived from satellite or aerial imagery, measuring the elevation of objects relative to the Earth's surface. This section introduces the primary data types utilized in this study, which focuses on height estimation from monoscopic satellite images.

Although depth and height represent different lengths, they are terms used to express the distance between two points. As illustrated in Fig. 1, while the term depth gives the distance from an observation point to the observed object, the height represents the distance of the observed object to another reference point. In the field of RS, this reference point is often the Earth. Both data types can be used in different fields. For example, depth information is used in autonomous driving cars; the car's distance to other objects represents the depth, and according to depth information, the car automatically decides what to do next. In this application, a car is an observation point, and the objects around the car are the observed objects. Height information is used in RS applications by obtaining the distance of an object from the earth's surface point and can be used in 3D earth surface models.

Datasets for height estimation in RS are composed of satellite or aerial images and their corresponding height data, which can be represented as Digital Surface Models (DSM), Digital Terrain Models (DTM), or normalized Digital Surface Models (nDSM). As illustrated in Fig. 2, DSM contains height information of both the Earth's surface and objects on it, while DTM only represents the natural surface, excluding objects like buildings or trees. nDSM is derived by subtracting DTM from DSM, showing only the height of objects on the surface. These data types contain pixel-wise elevation information obtained through traditional methods like photogrammetry, SAR, and LiDAR, and the height values differ based on the reference surface used, such as ellipsoidal or geoid heights.

DSM, DTM, and nDSM are raster data containing elevation data in each pixel. However, they should be defined correctly as each represents elevation data obtained with different approaches. Height represents the distance from the top point of an object to the reference point. For example, the distance from the top point of buildings and trees to the ground represents the object's height. This height may be different in data such as DSM and DTM. This depends on the reference surface used. Smooth imaginary ellipsoidal surfaces closest to the Earth's surface are generally used to facilitate mathematical calculations. In addition, ellipsoidal heights can be converted to geoid heights using local vertical datums. Depending on the applications in which these data types will be used, the height data they represent may vary depending on the reference surface. Reference surfaces and different definition of the heights are illustrated in Fig. 3

Object height is calculated using reference surfaces like the geoid or ellipsoid, which are mathematical models and not captured by cameras in the real world. DSM and DTM depend on these reference surfaces, varying based on the application. To determine the exact height of an object, nDSM is used, as it takes the terrain as the reference surface, which cameras can detect. As illustrated in Fig. 3, the reference surfaces help differentiate between terrain and object heights effectively. DSM and nDSM data are critical for height estimation models, as they contain pixel-wise elevation information. DSM includes topographic details as well as object heights, while Ndsm focuses solely on the heights of objects, excluding topographic variations. This makes nDSM a more suitable training dataset for object height estimation from satellite images.

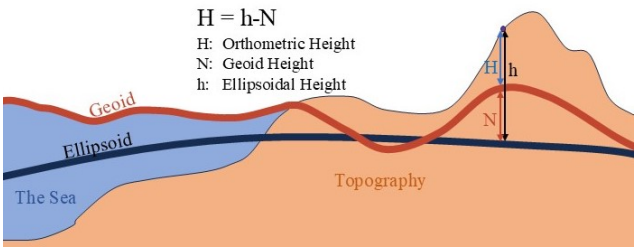

Figure 3. An illustration of definitions of heights according to different reference surfaces. Ellipsoidal Height, Geoid Undulation, and Orthometric height.

2. Further visual results

Fig. 4 illustrates the results of models trained on the DFC2019 dataset applied to test images. The fused method effectively detects building boundaries and edges while delivering accurate results in uniform areas like roads. Additionally, the proposed model shows consistent performance even in extreme conditions, such as images captured in snowy weather, as illustrated in the example in the 2nd row of Fig. 4. However, all models struggled to predict forest areas' complex and rough terrain.

Fig. 5 shows the FusedSeg-HE model's results on test images, including absolute errors and histograms. The histograms indicate that the proposed model estimates height values in each pixel close to real values. However, absolute error maps reveal lower errors within object bodies but higher errors at the edges. This highlights the challenge of accurately detecting objects and assigning pixels at boundaries, where complex structures and shadows cause information loss. In the first example, the model predicted missing object annotations in the ground truth, demonstrating its robustness in handling incomplete or inaccurate labels. Despite boundary errors, the model performs well overall in height estimation. Fig. 6 shows the ground truth and prediction comparison of three monocular satellite images in the 3D reconstructed view.

3. Discussions

Monocular satellite image height estimation models provide many innovations and advantages. One of its most significant advantages is that it fully automates the height estimation problem by eliminating the manual parameters and human supervision required by traditional methods. Another key benefit is that these models can estimate height from a single image, significantly reducing the need for images. Such capabilities of monocular height estimation from a single image enable a rapid evaluation process, which is crucial in emergency management scenarios. Although monocolur height estimation models have a robust infrastructure in terms of innovation, they have some areas for improvement and failure cases. In the monocular approach, where object height is determined using a single image, the features represented by the objects within that image become crucial. Accordingly, the first disadvantage is that a single image cannot extract features from occlusion areas. As moving away from the Nadir point in the satellite image, occlusion areas formed by relief displacement effects can make it challenging to extract information about objects. Another weakness is the requirement for representative features that give insight into height values, such as shadow and object associations, which would provide better accuracy. Estimating the height of objects close to the Nadir point from the top view image is difficult. At this point, the shadows and shadow lengths of the object are essential for the information extraction of the models.

The performance of the inference images is also an important point that should be considered in addition to the standard evaluation of the height estimation from monoscopic satellite images. Fig. 7 shows the results of the inference images, which are not included during training and have different spatial resolutions than the FusedSeg-HE training set. Model results are visualized on images taken from different angles and at various times for the same building object. In the left column, the object exhibits a representative feature closer to the orthophoto because it is relatively close to the Nadir point. However, this image does not contain features such as shadows that make it easier to extract information about height.

In the middle column, an image giving the impression of an orthophoto taken from almost the same angle as the RGB image in the left column is given. However, this image contains the shadow feature that facilitates height estimation. In the far-right column, an oblique image containing the shadow features of the building object is given. If we analyze these three images and their results, it is obvious that the shadow feature is an undeniable inference tool in the height estimation task. In addition, while oblique images can provide more detail about building height, they make inference difficult in determining building boundaries.

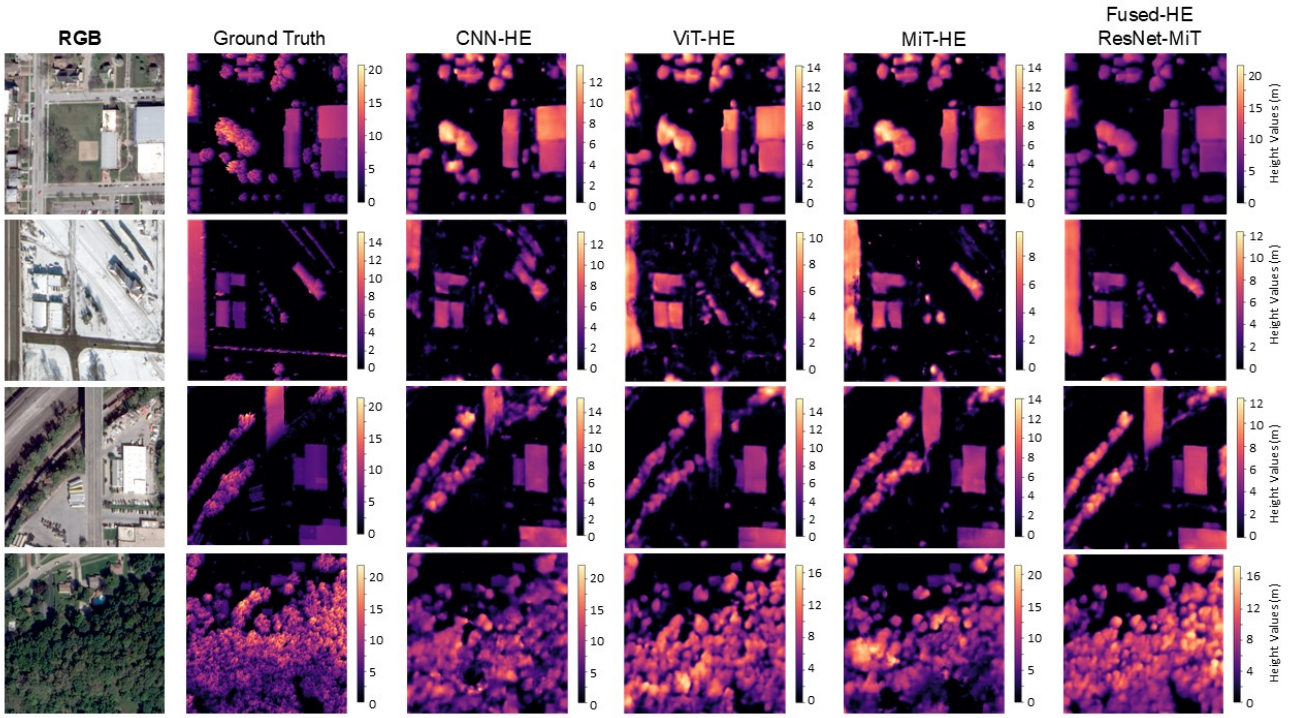

Figure 4. Trained model visual results on four different satellite images in DFC2019 test data. The scale bars represent height values in meters.

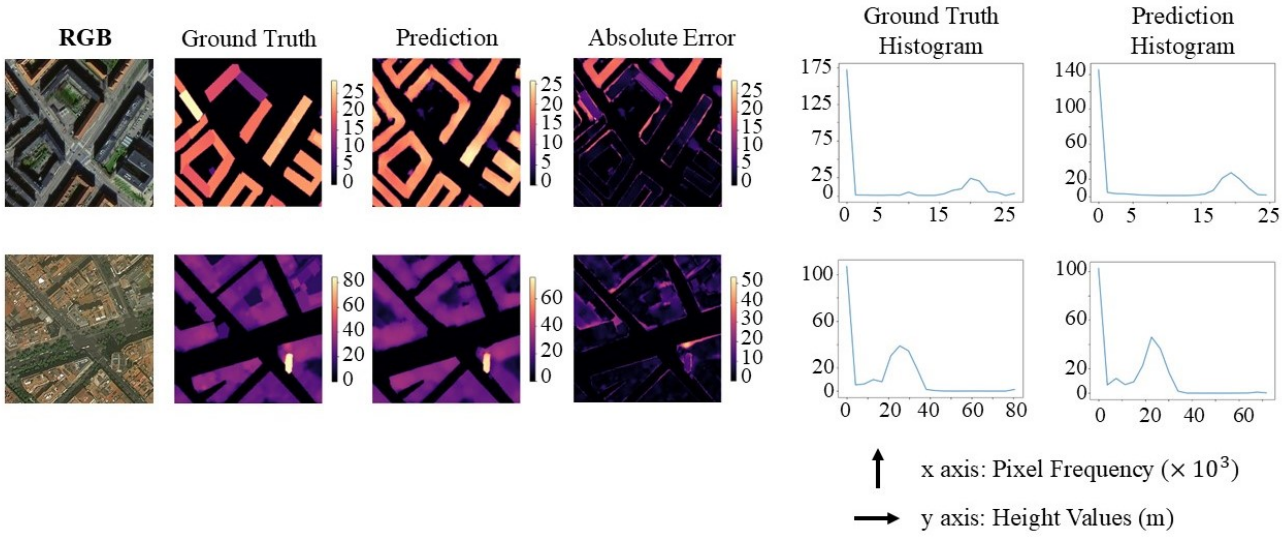

Figure 5. FusedSeg-HE model visual results, absolute error, and histogram on two different scene satellite images in DFC2023 test data. The scale bars represent height values in meters.

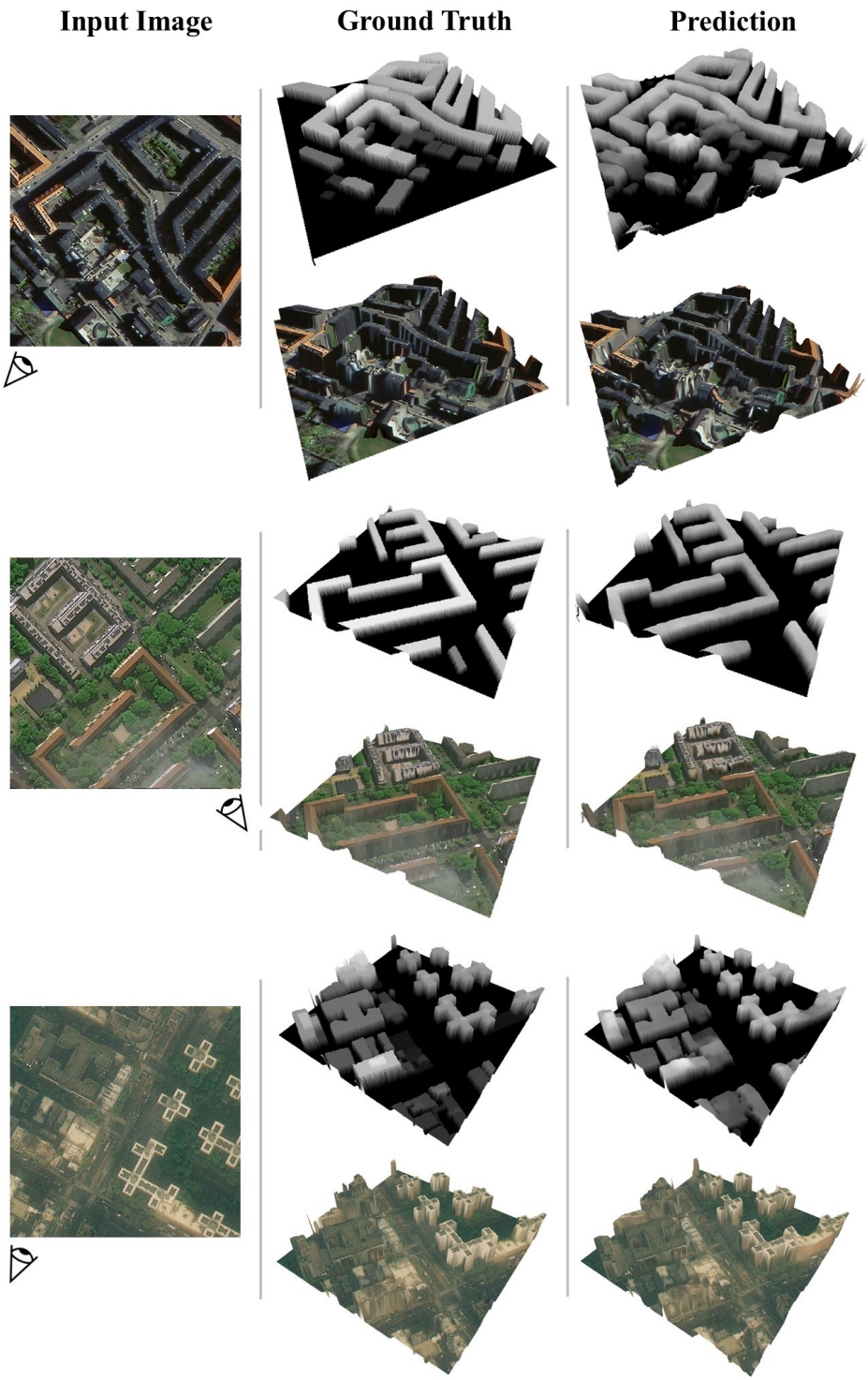

Figure 6. 3D visualization of FusedSeg-HE predictions on DFC2023 test data.

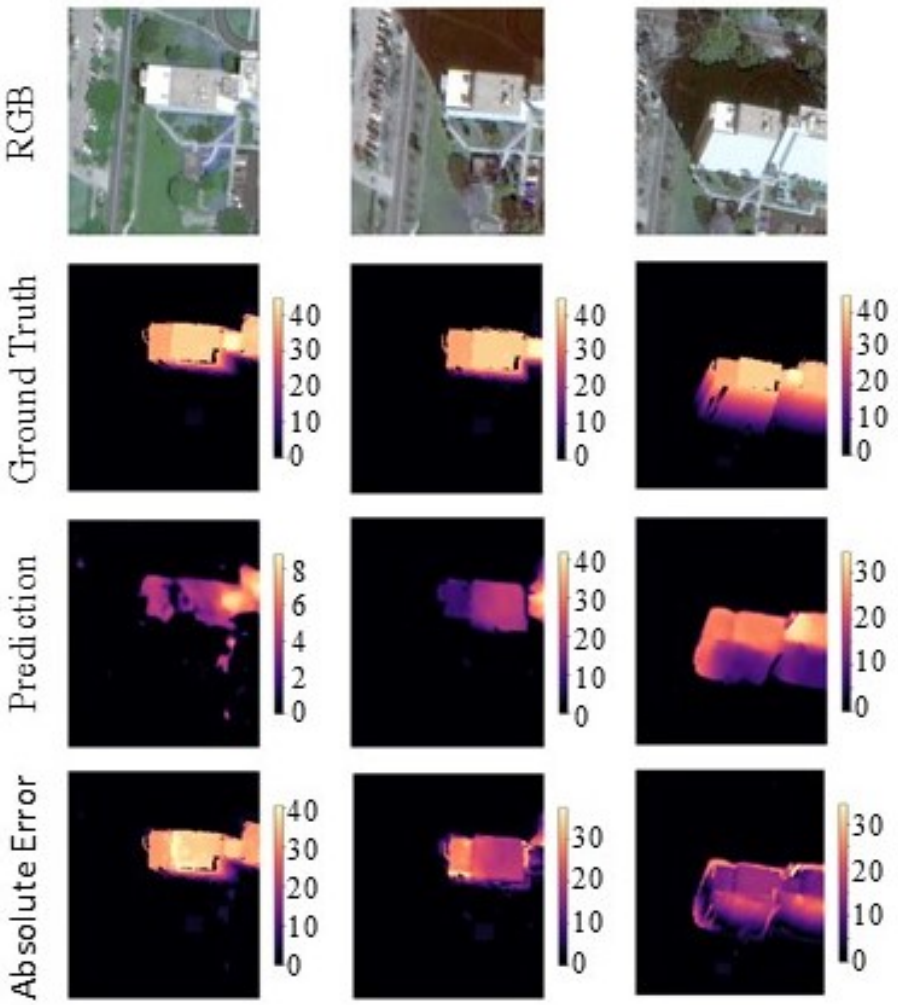

Figure 7. Inference results of FusedSeg-HE model on WV3 satellite image with 30cm GSD from DFC2019 dataset that represents the same object captured at different times at different angles and predictions of the model trained with DFC2023 dataset. Left column: The object is relatively close to the Nadir point without shadow and height estimation results. Middle column: The same object is relatively close to the Nadir point with shadow. Right column: The same object in oblique view with shadow. The scale bars represent height values in meters.