SARFormer – An Acquisition Parameter Aware Vision Transformer for Synthetic Aperture Radar Data

Supplementary Material

Contents

8. Additional Information: Data

Imaging Modes The TerraSAR-X satellite, like other SAR platforms, offers multiple imaging modes, each providing different spatial resolutions and aerial coverage. In this work, we utilize four imaging modes: *StripMap* (SM), *SpotLight* (SL), *High Resolution SpotLight* (HS), combining both 150 MHz and 300 MHz bandwidths, and *Staring SpotLight* (ST). The key differences among these modes are summarized in Tab. 4. It is important to note that a SAR image with lower spatial resolution is not equivalent to a downsampled high-resolution image. This distinction arises primarily because the speckle effect – a granular noise inherent to coherent imaging systems like SAR, resulting from the constructive and destructive interference of backscattered signals from multiple scatterers within a resolution cell – has a significantly greater impact on images with larger resolution cells than on fine-resolution SpotLight images. Fig. 4 compares the same area captured in the four different modes at similar incidence angles (note: different acquisition dates).

Radar Geometry In SAR imaging, two common geometries are used to represent spatial dimensions: *slant range* and *ground range*. The slant range is the direct line-of-sight distance between the radar antenna and a target on the Earth's surface, measured along the radar beam's path. In a typical slant range SAR image, each column represents the slant range distance from the radar to the target, effectively mapping the across-track (range) dimension. Each row corresponds to a different position along the radar platform's flight path, representing the along-track (azimuth) dimension. This means the image grid is organized such that columns increase with distance from the radar, and rows progress with the movement of the radar platform. Compare the first to columns in Fig. 5: One can observe that although the scenes were captured from opposite directions,

Table 4. Properties of the different imaging modes of TerraSAR-X as used in this work

Imaging Mode	Scene Size [km]	Slant Range Resolution [m]	Azimuth Resolution [m]	Looking Angle
SM	30 x 50	1.2	3.3	20° – 45°
SL	10 x 10	1.2	1.7	20° – 55°
HS	10 x 5	0.6 – 1.2	1.1	20° – 55°
ST	4 x 3.7	0.6	0.24	20° – 45°

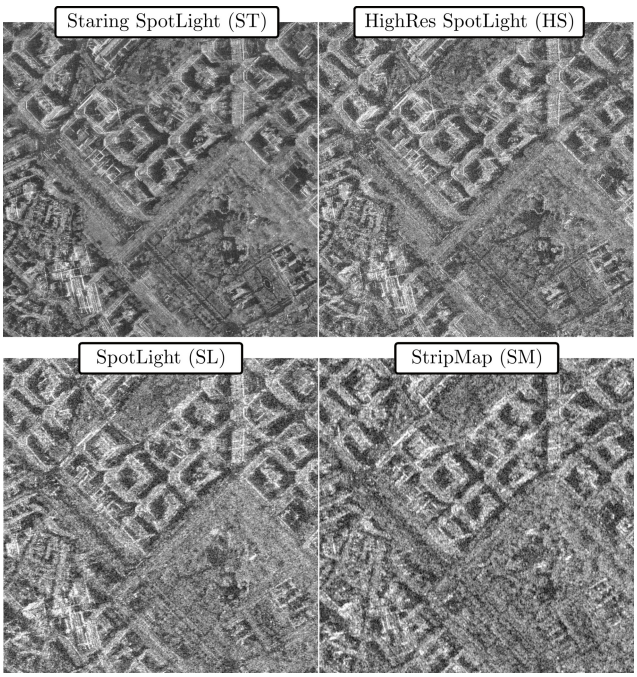

Figure 4. Comparison between the four different imaging modes being used in this work. The images depict the same scene captured with similar viewing angles but in different imaging modes.

the top of the Eiffel Tower points to the left side of the image because it is closer to the sensor than the base of the tower. Further, the north direction on the Earth's surface does not correspond to the top in the slant SAR image. Consequently, two SAR images in slant range cannot be accurately coregistered unless they were acquired from the same orbit with identical acquisition parameters.

On the other hand, in ground range geometry, these slant range measurements are projected onto a horizontal plane (or, as in this case, a terrain model), as if viewed from di-

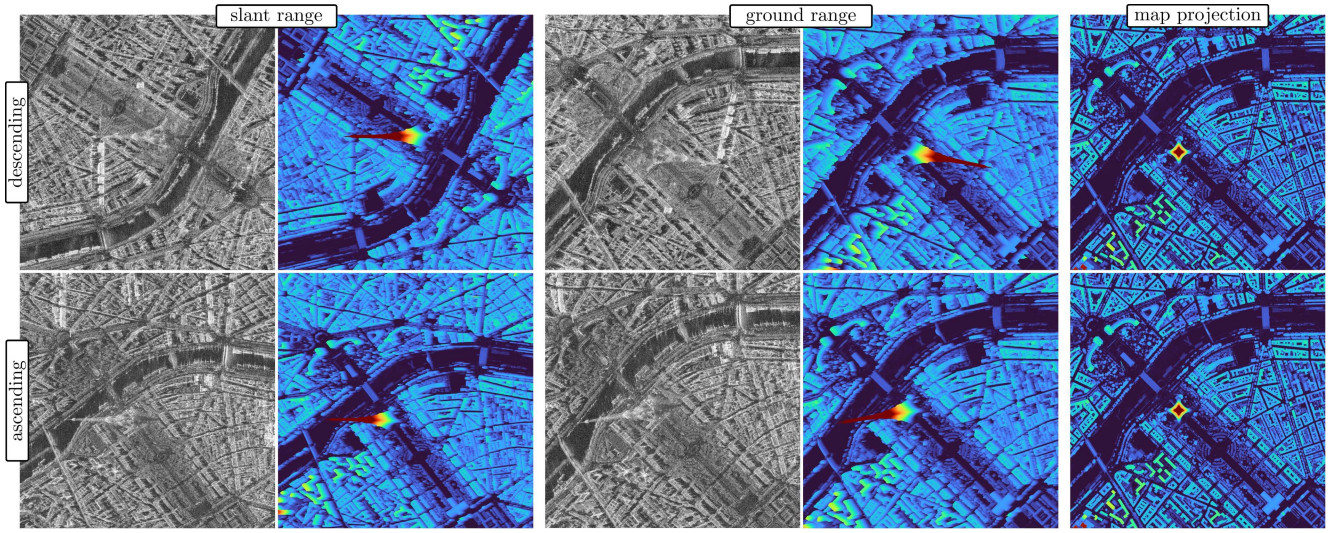

Figure 5. Comparison of the different SAR image geometries. Each row shows a SAR image of the same area taken from different directions, along with the corresponding height above ground values. Columns 1 and 2 display the images in their native slant-range geometry, where the columns represent the distance to the sensor from left to right. In Columns 2 and 3, the images are projected onto a terrain model, making each pixel correspond to one meter on the Earth's surface. The far-right column shows the height values in a map projection, independent of the image geometries, and thus identical for both acquisitions.

rectly above. By removing elevation-induced distortions, ground range images simplify spatial interpretation and, importantly, enable fusion/coregistration with different SAR images or other types of geospatial data. However, objects not included in the terrain model used for projection – i.e. buildings and vegetation – still appear distorted. In the third and fourth columns of Fig. 5, the two SAR images alongside their respective image-specific heights in ground range projection are shown. The Eiffel Towers are now pointing in different directions (towards the respective sensor), but ground pixels appear at the same position in both of the images, enabling pixel-by-pixel superimposition. We refer to the heights projected from slant range geometry onto the DTM as heights in image geometry or still as *slant* heights since these height values originate from slant range and retain the characteristics of the original geometry being image-specific. The last column in Fig. 5 displays the corresponding heights in a map projection, which is independent of the image geometry and thus identical for both SAR images shown.

Dataset Limitations and Challenges The dataset utilized in this study presents several limitations that may introduce uncertainties into the model's performance. First, there are inconsistencies in the acquisition dates of different images capturing the same geographic scene, meaning that physical or structural changes could have occurred in the intervening period. For instance, a newly built building visible in only one of the provided views could confuse the model. Additionally, discrepancies between the acquisition dates of images and corresponding ground truth data – whether LiDAR or building footprint annotations – introduce errors both in training and validation, as both imagery and ground truth may not accurately reflect the same spatial conditions. Furthermore, the digital terrain model used for terrain correction bears inherent inaccuracies, with deviations that can reach several meters, particularly in densely constructed urban areas. These inaccuracies propagate into geolocation errors, potentially distorting spatial alignment between SAR images captured from varying perspectives and with respect to the ground truth data.

Parameter Preprocessing The acquisition parameters, which are incorporated into the transformer via the APE module, undergo preprocessing to enhance their interpretability for the neural network, as described in Eq. (2). To preserve its cyclic nature and avoid artificial discontinuities, the azimuth angle Az, which spans from $0°$ to $360°$, is represented using its sine and cosine components. The looking angle θ is transformed using the cotangent function, as this effectively approximates the ratio between building height and the corresponding layover extent on a tangential plane. The imaging mode m is mapped to a single-digit identifier. While the imaging mode could alternatively be substituted by the sensor's resolution – given their direct correlation – it is included here as an example of how semantic or non-numerical metadata, such as sensor type, input modality, or polarization, can be incorporated into the model.

9. Additional Information: Metrics

Segmentation Metrics To evaluate the performance of the binary building footprint segmentation, we used the overall accuracy (OA)

$$\text{OA} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

with TP and FP the true and false positives, and TN and FN the true and false negatives, and the mean Intersection over Union (mIoU):

$$\text{mIoU} = \frac{1}{2}\left(\frac{TP}{TP + FP + FN} + \frac{TN}{TN + FP + FN}\right). \tag{5}$$

Regression Metrics To assess the performance of the height reconstruction (both in map and image geometries), we utilized the Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{6}$$

with n as the number of data points, y_i is the actual target value for pixel i, and $\hat{y}_i$ as the predicted value for pixel i, the Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{7}$$

and the Structural Similarity Index Measure (SSIM) [30] as some sort of relative metric, which is intended to reflect people's perception:

$$\text{SSIM} = \frac{(2\mu_y\mu_{\hat{y}} + C_1)(2\sigma_{y\hat{y}} + C_2)}{(\mu_y^2 + \mu_{\hat{y}}^2 + C_1)(\sigma_y^2 + \sigma_{y\hat{y}}^2 + C_2)}, \tag{8}$$

where y and $\hat{y}$ are the predicted and target images, μ_y and $\mu_{\hat{y}}$ their average pixel intensities, σ_y^2 and $\sigma_{y\hat{y}}^2$ the corresponding variances, $\sigma_{y\hat{y}}$ representing the covariance, and C_1 and C_2 as constants to numerically stabilize the division.

10. Additional Information: Training

Loss Function Since we are dealing with a long-tailed distribution, characterized by a substantial number of ground pixels at a height of zero, the models usually tend to underestimate heights. This observation underscores the rationale for employing an asymmetric loss function:

$$l_{\text{asym}} = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} w_{\text{underestimated}} \cdot |\hat{y}_i - y_i|, & \text{if } \hat{y}_i < y_i \\ w_{\text{overestimated}} \cdot |\hat{y}_i - y_i|, & \text{if } \hat{y}_i \geq y_i \end{cases} \tag{9}$$

with $\hat{y}$ the predictions and y the target, n the number of pixels, $w_{\text{underestimated}} = 1.5$ and $w_{\text{overestimated}} = 1$. To penalize errors along edges and to further improve small details, the gradient loss

$$l_{\text{grad}} = \sum_{d\in\{x,y\}} \|\nabla_d(\hat{y}) - \nabla_d(y)\|_1 \tag{10}$$

where ∇_d is the spatial derivative in the dimension d (determined through Sobel operator), and normal loss is added (details can be found in [12]):

$$l_{\text{normal}} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{\langle \vec{n}_i^{\hat{y}}, \vec{n}_i^{y}\rangle}{\sqrt{\langle \vec{n}_i^{\hat{y}}, \vec{n}_i^{\hat{y}}\rangle}\sqrt{\langle \vec{n}_i^{y}, \vec{n}_i^{y}\rangle}}\right) \tag{11}$$

with $\vec{n}_i^{\mathrm{x}} = [-\nabla_x(\mathrm{x}_i),\ -\nabla_y(\mathrm{x}_i),\ 1]^\top,\ \mathrm{x}\in\{y,\hat{y}\}$ and $\langle\cdot,\cdot\rangle$ the inner product of vectors. The combined loss function for the regression task is the weighted sum of the losses above:

$$\mathcal{L}_{\text{regression}} = \alpha \cdot l_{\text{asym}} + \beta \cdot l_{\text{norm}} + \gamma \cdot l_{\text{grad}} \tag{12}$$

with $\alpha = \beta = 1$ and $\gamma = 0.1$. For the segmentation task, i.e. the building footprints, the binary cross entropy loss is used, which can be expressed as (using logits $\hat{y}$):

$$\mathcal{L}_{\text{segmentation}} = l_{\text{BCE}} = \\ -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log(\sigma(\hat{y}_i)) + (1 - y_i)\log(1 - \sigma(\hat{y}_i))\right) \tag{13}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. The total loss function is formulated as a weighted sum of the individual task-specific losses:

$$\mathcal{L}_{\text{MTL}} = \mathcal{L}_{\text{height}} + \mathcal{L}_{\text{height slant}} + 0.1 \cdot \mathcal{L}_{\text{buildings}}. \tag{14}$$

Multitask DPT Setup We adopt the dense prediction strategy using ViT-based models as proposed by [19] (DPT). The projection of features derived from multiple views and their associated metadata tokens is described in detail in the main manuscript. For the 12-layer *ViT-Base* configuration, features are extracted after layers 3, 6, 9, and 12. For the *ViT-Large* configuration, features are extracted after layers 5, 12, 18, and 24. Similarly, for the *ViT-Huge* configuration, features are extracted after layers 8, 16, 24, and 32.

To infer annotations for the three tasks in the multitask framework, we extend the final convolutional block of the DPT with three task-specific blocks. Each block consists of five convolutional layers, each followed by a *LeakyReLU* activation function.

Masking strategies for MAE pre-training In the context of remote sensing images, traditional masking strategies employed in masked autoencoders must be rethought to address the fundamental differences from natural object-centered images. Unlike photographs of objects (such as dogs for instance), where masking significant portions still

Table 5. Numerical results from fully-supervised experiments (no pre-training) using the **ViT-Base backbone** in the 2-view scenario. Performance gains achieved with the APE module are comparable to those reported for the ViT-Large configuration (refer to the main paper).

		Classification Footprints		**Regression Height**			**Regression Height (Slant)**		
# Views	Model	mIoU	OA	MAE	RMSE	SSIM	MAE	RMSE	SSIM
2	DPT-Base [19]	**0.73**	**0.92**	4.86	7.23	**0.85**	5.06	7.38	**0.88**
	SARFormer (ours)	**0.73**	**0.92**	**4.44**	**6.68**	**0.85**	**4.69**	**6.94**	**0.88**

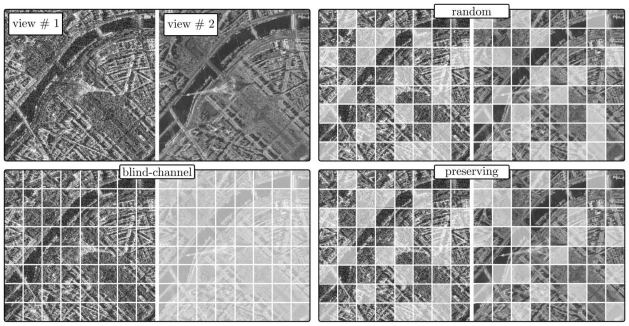

Figure 6. Different masking strategies for multi-view scenarios. The *preserving* masking strategy (lower right) ensures that at least one view remains active for all locations across views #1 and #2. In contrast, the *random* masking strategy (upper right) does not guarantee this consistency. The *blind-channel* masking strategy (lower left) is a special case of *preserving*, where one view is entirely masked while the other remains fully active.

allows recognition due to the structured and object-centric nature of the image, remote sensing images often lack such intrinsic coherence. For example, if a single building within a forested area is masked, it is basically impossible for the model to reconstruct it due to the absence of sufficient contextual information. To address this issue, we introduce novel masking strategies tailored to remote sensing data. Our approach leverages multiple views of the same scene captured under varying acquisition conditions, such as differing resolutions, angles, and directions. By ensuring that masking does not occlude the same patch across all views, the model retains at least one perspective for reference, thereby enhancing reconstruction potential while keeping the intrinsic complexity arising from acquisition variability. We refer to this strategy as *preserving*. A notable extreme of this strategy, termed *blind-channel* masking, involves completely masking one view, challenging the MAE to reconstruct it solely from the unmasked complementary view. The *blind-channel* scenario necessitates encoding the acquisition parameters (as being done by the APE module) since these cannot be inferred from the data in the masked view. These strategies exploit the rich heterogeneity of remote sensing data, fostering more robust and semantically meaningful representations. Compare Fig. 6 for a visual example of the different strategies.

Table 6. Comparison between different model size configurations (on a subset of the metrics). The setting was chosen to the best-performing: two views, active APE module, pre-trained using the *preserving* strategy.

Model Size	**mIoU**	**MAE** (map)	**MAE** (slant)
ViT-Base	0.73	4.26	4.39
ViT-Large	0.74	4.12	**3.96**
ViT-Huge	**0.76**	**4.04**	**3.96**

11. Additional Results

Effect of Backbones As discussed in the main paper, we trained our best-performing configuration (2 views, active APE, and *preserving* masking during pre-training) using three different backbone architectures: ViT-Base, ViT-Large, and ViT-Huge. Tab. 6 presents a subset of evaluation metrics on the test set for these configurations, demonstrating a consistent trend where larger model sizes lead to improved performance. Furthermore, Tab. 5 illustrates the effect of incorporating the APE module into the ViT-Base backbone, evaluated in the 2-view setup. Notably, the inclusion of the *metatoken* demonstrates significant benefits, particularly for the height estimation task.

Fine-Tuning on limited Labels To further highlight the effectiveness of the proposed pre-training paradigm, we minimized the labeled dataset for fine-tuning to just two SM images from a single location, Paris. This setup introduces a significant domain shift in multiple regards during testing, as it includes data from different locations, acquisition modes, and looking angles. Fig. 7 presents a visual comparison of outputs – specifically, height maps and building footprints – generated by a UNet (trained from scratch), the non-pre-trained *SARFormer*, and the pre-trained *SARFormer*. Two HS scenes from Berlin served as model input. Although performance remains below that achieved with the full dataset, the benefits of pre-training are evident, underscoring its value, particularly for few-label or out-of-

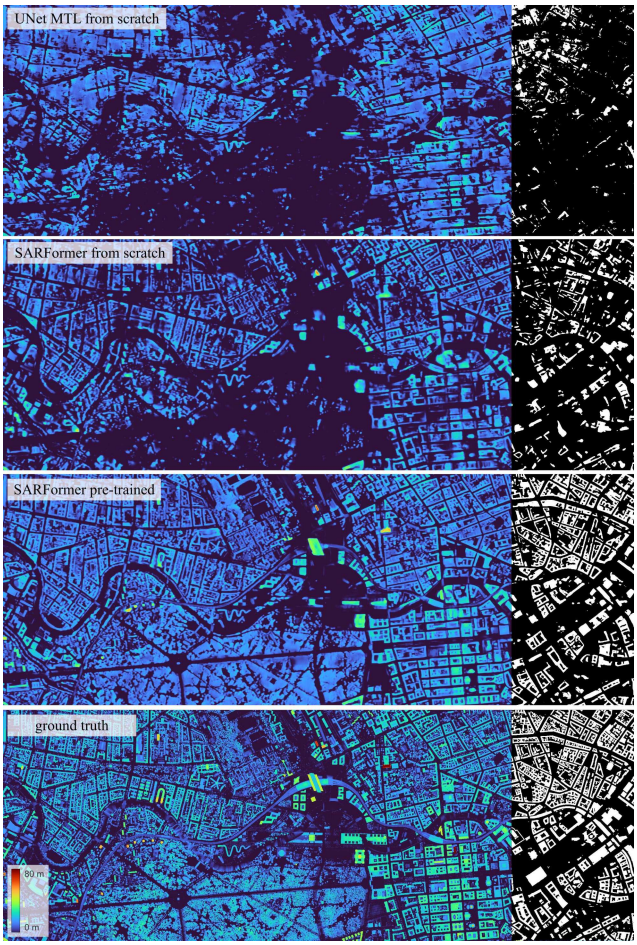

Figure 7. Next to the ground truth (bottom), we present model outputs trained on an extremely limited dataset consisting of only two SM images of Paris. Inference was conducted on two HS images of Berlin captured from different viewing angles than those used in training. Notably, the pre-trained *SARFormer* (third row) demonstrates the highest resilience to this multifactorial domain shift, encompassing location, resolution, and geometric differences. For comparison, we also display results from UNet and *SARFormer* trained from scratch (first and second rows, respectively).

domain scenarios. Tab. 3 shows the error metrics for the entire test set (the same as all other experiments were evaluated on) in the limited-label scenario.

12. Additional Visual Results

Demonstration of best-performing configuration Fig. 9 presents the outputs of all three tasks on three representative SAR scenes. These results were generated using the pre-trained *SARFormer* with the ViT-Huge backbone, activated APE module, and the *preserving* masking strategy during pre-training. For each example, the top row shows the model outputs, while the bottom row displays the corresponding ground truth data. The depicted scenes are from Vancouver and Berlin, both of which were entirely excluded from the training set. In the first example, two different imaging modes were utilized to reconstruct a complex scene containing multiple high-rise buildings. The results, in both slant and map geometries, are closely aligned with the ground truth. The second example illustrates a challenging case involving SM input data, characterized by low spatial resolution and high noise levels. While the performance is inferior compared to Spotlight images, the model still achieves acceptable results. An intriguing detail in the final example is the absence of the Berlin TV Tower in the model's prediction, which is distinctly visible as the tallest structure in the ground truth. This omission is very likely due to the weak radar response of the tower. Only the sphere at the top of the structure is faintly discernible in the SAR images, a feature detectable only by trained observers. Here, the methodology reaches its physical limitations.

Extension to other missions Although a detailed description of integrating various satellite missions into the *SARFormer* framework is beyond the scope of this manuscript, it is important to note that such integration is straightforward. Figure 8 presents an exemplary output from a *SARFormer* variant that was pre-trained and fine-tuned on an extended version of the dataset described here. This extended dataset includes imagery from *ICEYE*, *Umbra*, and *Capella Space* in addition to the previously mentioned *TerraSAR-X* data. The only modification relative to the manuscript was to replace the encoding of a discretized imaging mode m with the encoding of the azimuth and range resolutions of the corresponding product since the nomenclature of the imaging modes differs between providers.

Comparison to Baseline Fig. 10 compares the outputs of four different models with the corresponding ground truth. The baseline is a UNet in the multi-task configuration, evaluated for both single-view and two-view scenarios. In contrast, we include results from the pre-trained *SARFormer* (ViT-Large), also evaluated for single-view and two-view cases. The performance comparison across the four illustrated scenes highlights several key observations. First, the addition of a second view significantly enhances the reconstruction capability of the models, both in terms of height accuracy and building shapes. Furthermore, SARFormer demonstrates superior performance compared to the baseline, in both single-view and two-view scenarios. For the third scene, no DSM as ground truth was available, so the comparison is limited to building footprints. In this case, the *SARFormer* models again produced results that align more closely with the labels compared to the baseline models. Overall, it was observed that the proposed *SARFormer* architecture particularly excels in complex scenarios, such as those involving low-resolution data, small structures, or heavily mixed layover signals.

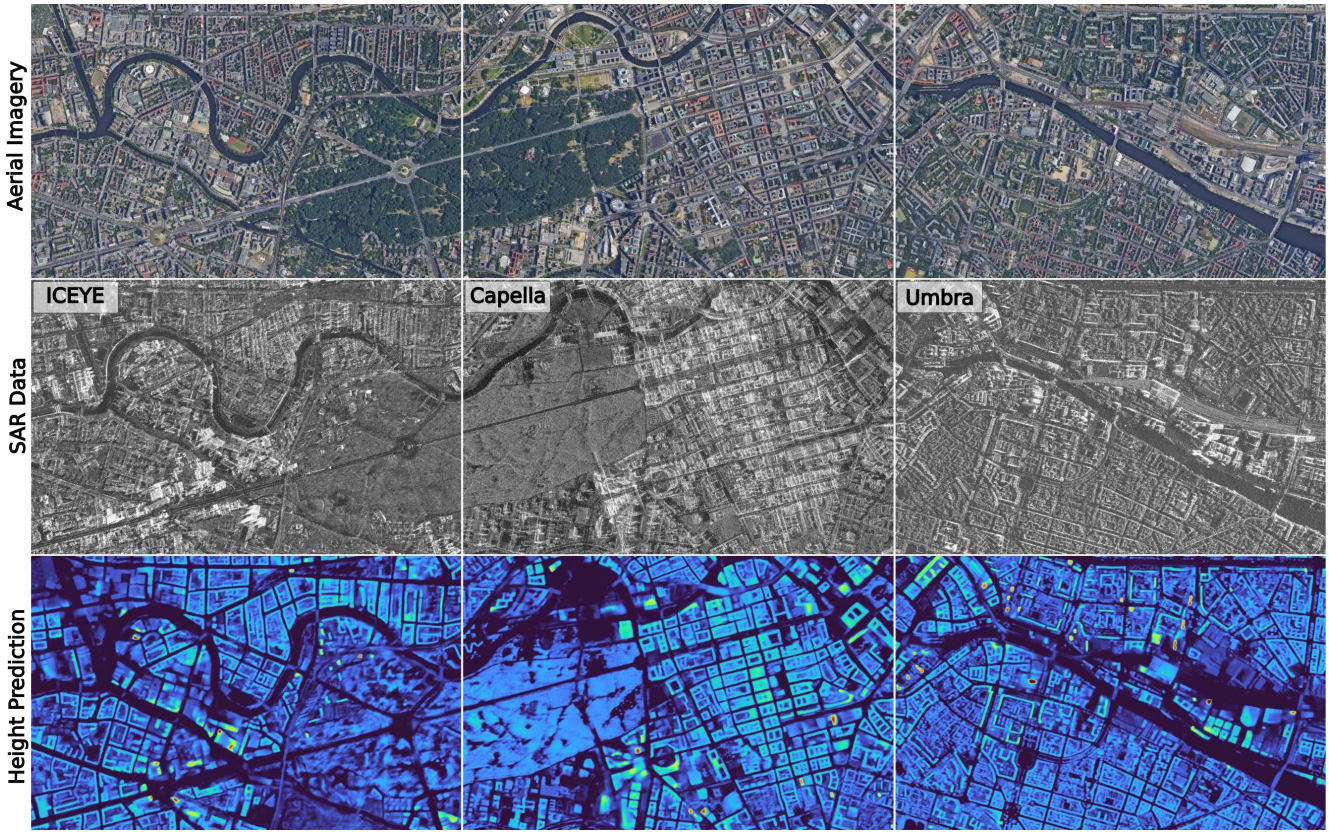

Figure 8. Three SAR spotlight acquisitions over the city of Berlin, provided by *ICEYE*, *Capella Space*, and *Umbra*, were used to infer the *SARFormer* model. Despite differences in acquisition characteristics, the combined nDSMs (bottom row) result in a homogeneous and coherent reconstruction. Aerial imagery is taken from Google, ©2025.

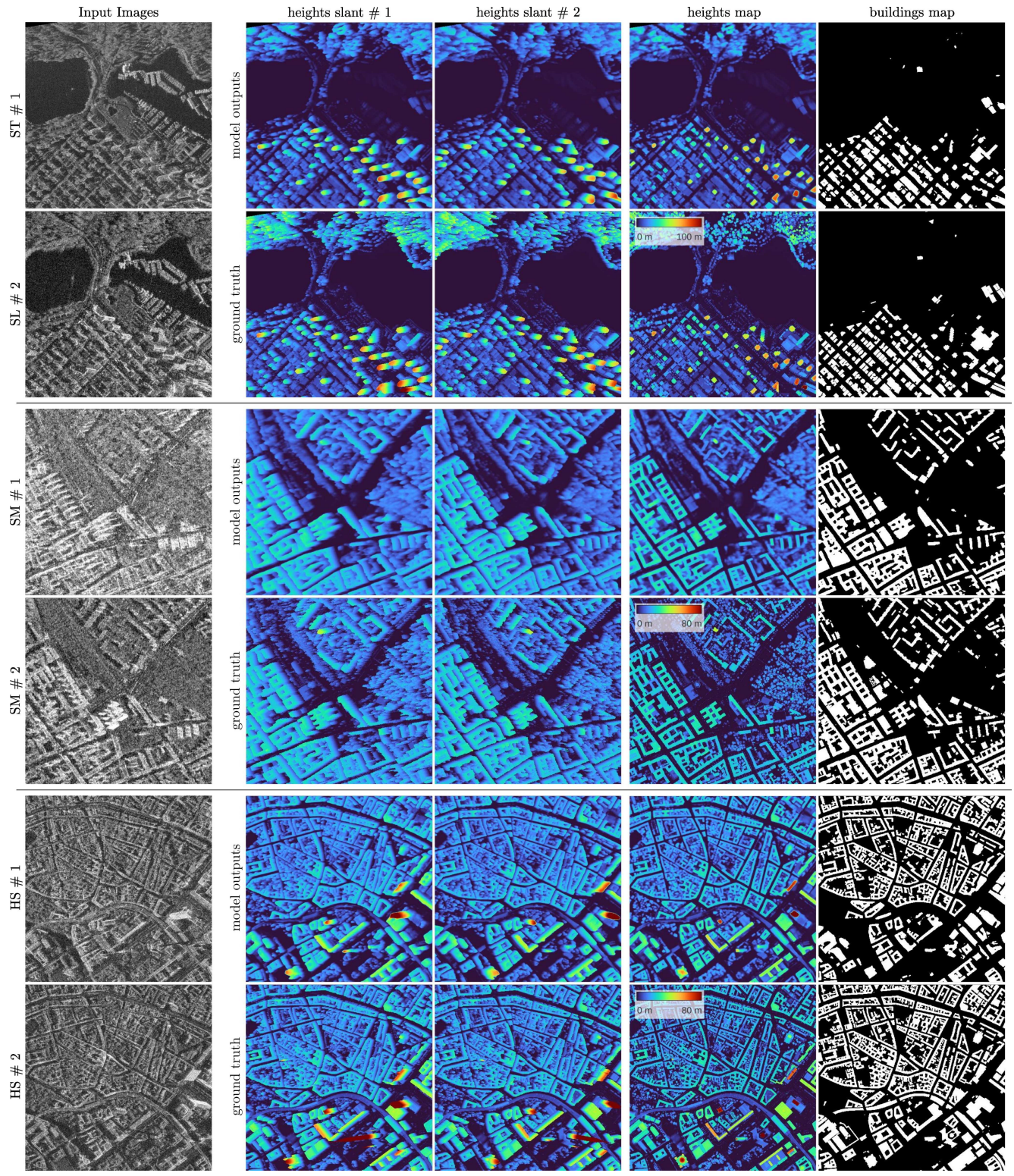

Figure 9. Model outputs for three scenes generated by the pre-trained *SARFormer* (ViT-Huge) using 2 input views. The upper rows present the model's predictions for the three downstream tasks, while the lower rows display the corresponding ground truth data.

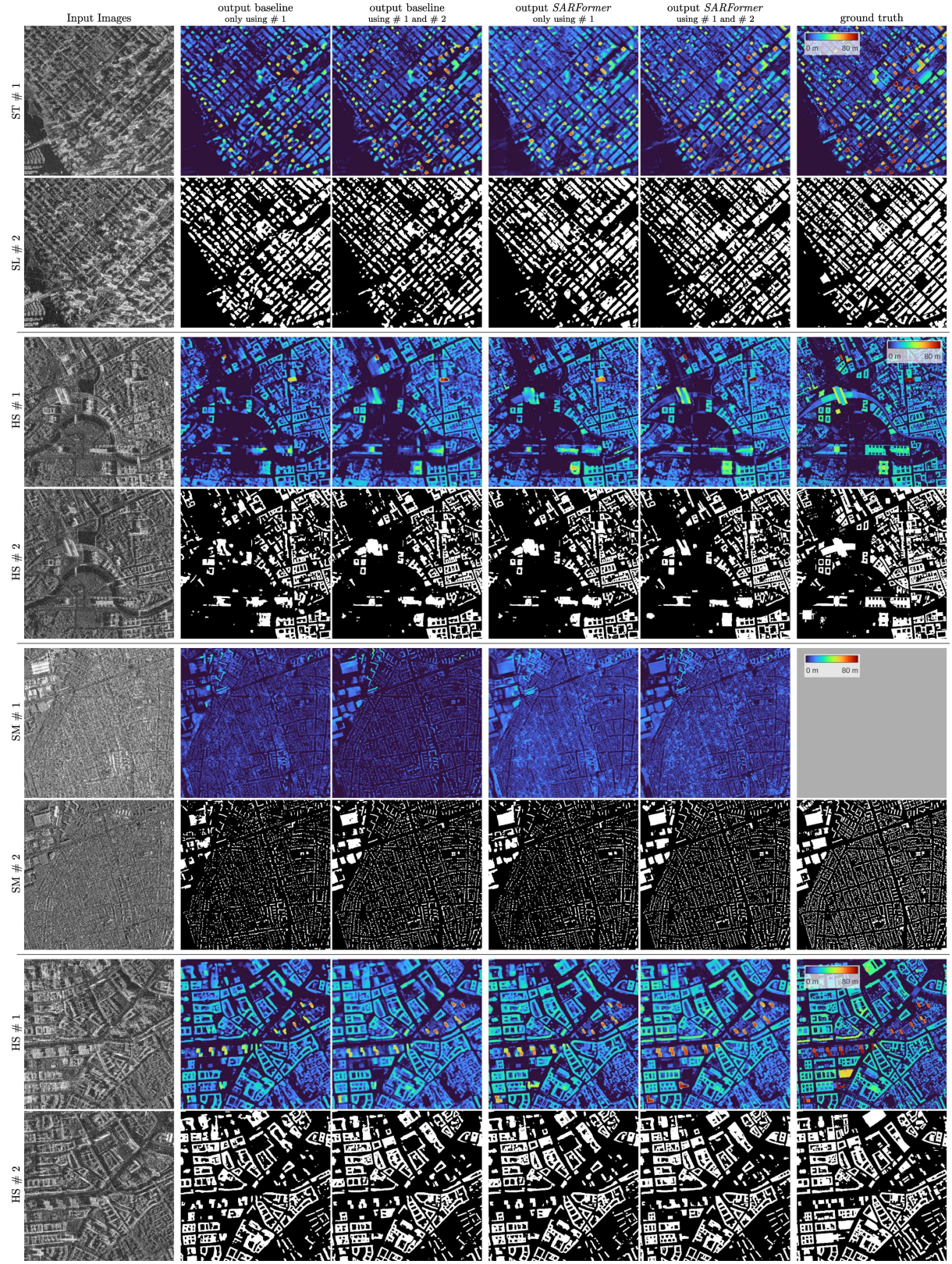

Figure 10. Comparison of SARFormer and baseline models in both single-view and two-view scenarios. The final column displays the ground truth data. Note that height labels were unavailable for the third scene.