Leveraging LiDAR datasets to improve SAR tomography: a diffusion model approach

Cristiano Ulondu Mendes[1[0009−0000−9036−4761]], Loïc Denis[1,2[0000−0001−9216−8318]], Christophe Kervazo[1[0009−0008−6990−5322]], and Florence Tupin[1[0000−0002−3110−8183]]

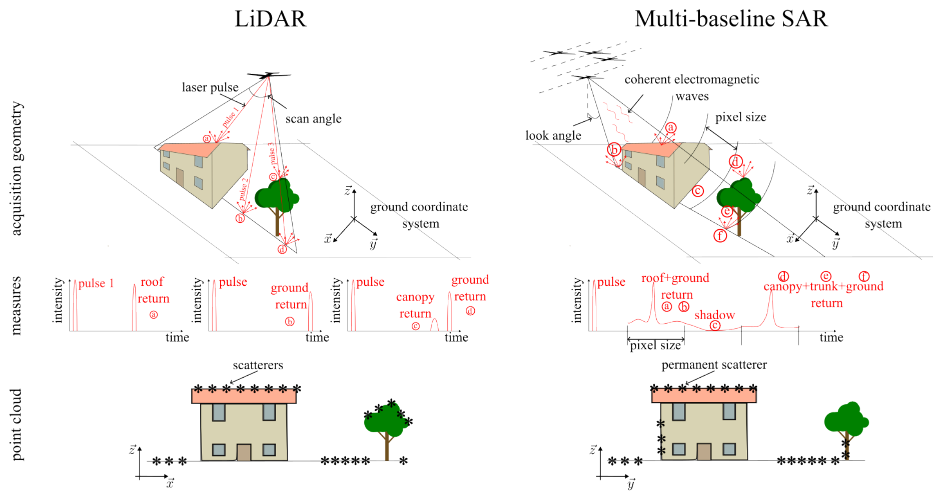
[1] Institut Polytechnique de Paris, LTCI, Télécom Paris, Palaiseau, France
{cristiano.ulondumendes,florence.tupin,christophe.kervazo}@telecom-paris.fr
[2] Université Jean Monnet, 10 rue Tréfilerie, 42023 Saint-Etienne, France
loic.denis@univ-st-etienne.fr

Abstract. Synthetic Aperture Radar (SAR) tomography is a 3D imaging technique based on the combination of multiple images acquired from slightly different angles. By analyzing the phase shift measured across the different images, it is possible to separate scatterers located at different heights. This requires solving an inverse problem and is typically performed independently for each pixel, producing a point cloud with large localization uncertainties in the elevation direction. Performing a single-step tomographic reconstruction with improved spatial regularity is difficult to achieve using a supervised approach, as simulating realistic SAR tomography data corresponding to a given 3D urban scene is a highly complex task. In this paper, we suggest a two-step approach: a first step using simple pixel-based inversion, and a second step restoring the reconstructed 3D cloud thanks to a diffusion model. We train our diffusion model on a large dataset of freely available LiDAR point clouds. A well-chosen geometrical projection is applied to represent the 3D points of the cloud visible from the radar as a 2D image. The degradation modeled by the diffusion model corresponds both to point omission (non-detections) and to mislocalizations in the elevation direction. Our diffusion model, based on the Residual Shifting method, requires as few as 20 diffusion steps to produce restored reconstructions. The paper introduces a flexible approach to leverage digital surface models from LiDAR datasets and improve 3D tomographic SAR reconstructions.

Keywords: SAR tomography · spatial regulization · learned urban structural prior · LiDAR · diffusion models · uncertainty quantification.

1 Introduction

Urban areas are densely built territories that continually evolve and expand with the influx of population. 3D urban modeling offers the ability to efficiently visualize and analyze spatial information, which is necessary for sustainable growth of cities. Remote sensing offers an alternative to the labor-intensive 3D acquisition from the ground. Airborne lateral-scanning light detection and ranging



key differences (relevant to urban area monitoring)

feature	LiDAR	SAR
electromagnetic range	infrared	microwave
view angle	near vertical	oblique
weather dependence	clear weather	all-weather
data type	elevation and intensity	amplitude and phase
coverage area	small	large

Fig. 1: Acquisition process of LiDAR and multi-baseline SAR data.

(LiDAR) is a commonly used technique in this context [13]. It involves scanning the scene linearly with discrete laser pulses and measuring the time delay of each return, as illustrated in Figure 1. Combining these measurements with the GPS position and the sensor's orientation at the time of acquisition provides accurate localization of the scatterers in the 3D ground coordinate system. This process enables the collection of a 3D point cloud of the scene. The main limitations of this technique are its sensitivity to weather conditions and its high cost. Indeed, due to its small coverage area, multiple flights are required to cover large areas. In contrast, a single flyover of a synthetic aperture radar (SAR) sensor can acquire data over a wide area, independently of weather conditions and daylight. Through the coherent processing of radar echoes collected along the sensor trajectory (called azimuth direction), SAR systems can generate a high-resolution complex-valued image. This image represents a 2D projection of the 3D imaged scene onto the azimuth-range plane, where the range of an object corresponds to its distance to the SAR sensor. As a consequence of the range-based scatterer localization of SAR imaging, scatterers located at different heights but identical distances from the sensor get projected into the same azimuth-range pixel (see the image on the right side of Figure 1). This mixing phenomenon, known as the layover effect, can be inverted by tomographic SAR (TomoSAR). Multiple acquisitions, captured with slightly different view angles, can be analyzed to separate scatterers whose echoes superimpose inside each pixel. The inversion is typically conducted separately for each pixel, leading to the localization of scatterers along the third dimension, the elevation direction, which is orthogonal to the azimuth-range plane. Since the first successful demonstration of TomoSAR on real data [20], based on the computationally efficient Fourier transform, but limited in spatial resolution by the acquisition configuration, the following works have attempted to improve the resolution [14, 15] while limiting the required number of acquisitions [35, 1]. This effort benefited from the rise of deep learning, with the introduction of algorithm unrolling [8], which led to a profusion of methods [5, 17, 26].

Because of the high spatial resolution of SAR imagery and the geometric regularity of urban structures (shoebox-like structures), adjacent azimuth-range pixels often share similar scattering responses. However, this information is not exploited in pixel-wise TomoSAR. To improve TomoSAR reconstructions, several methods have been proposed to enforce spatial regularity during the reconstruction process [19, 29, 30]. These approaches extend the pixel-wise TomoSAR model to the entire scene and incorporate priors consistent with the characteristics of urban structures. In particular, [19] formulated the reconstruction task as an optimization problem in 3D space, using handcrafted priors that promote sparsity and local smoothness of structures observed in this domain. While this allows for a significant reduction of noise in the final reconstruction, processing volume data represents a non-negligible computational burden. Moreover, handcrafted priors generally fail to capture complex features. The methods in [29] and its improved version [30] reduce the computational cost by operating on 2D slices of the 3D volume, where learned priors are used to enforce both intra-slice and inter-slice structural regularity. Priors are trained on simulated scenes made up from a collection of building models of different shapes, sizes, and orientations, covering a wide range of possible settings. [29] adopts a fully supervised training strategy requiring paired measurements and ground truth, which can lead to generalization issues when applied to real data. In contrast, [30] addresses this issue by adopting a self-supervised training framework for learning the priors during the reconstruction process. While this latter approach shows improved adaptation capability to real-world data, it substantially increases the processing time.

We aim to efficiently exploit the spatial information while keeping the processing time and dataset-specific adaptation requirements at an acceptable level for real-world applications. For this purpose, we explore a learning-based reconstruction framework that decouples TomoSAR reconstruction and spatial regularization. This enables the processing time of each component to be optimized independently and facilitates their separate analysis. Moreover, generalization to a different set of baselines (i.e., a different acquisition geometry or a change in the number of images) requires only a modification of the TomoSAR reconstruction stage. In this work, our focus will be on the spatial regularization component, which is formulated as a 3D point cloud denoising problem.

Point cloud denoising is a long-standing problem in the computer vision community. It aims to produce a dense point cloud from a sparse and noisy version while preserving the underlying structures. It typically involves outlier removal and missing data completion [21]. Point cloud denoising methods can be categorized into three types: filter-based, optimization-based, and deep learning-based methods [33]. Filter-based denoising methods assume that the noise dominates the high-frequency component of the signal and design filters that suppress noise accordingly. Optimization-based methods formulate the denoising task as an optimization problem that minimizes deviation from the input while enforcing priors through additional regularization terms. Finally, as in most computer vision subdomains, deep learning has made its way into point cloud denoising. This type of method usually offers improved performance while reducing processing time. There are several approaches to process point clouds [6]: some approaches, such as PointNet [16] and its successors such as Point Cloud Transformer [7] or Point Cloud Mamba [32], directly process the 3D points, while others use representations on rectangular grids to efficiently access neighbors, as proposed in PointProNets [21] and recent works such as [4] and [12]. Due to shadows, SAR systems typically collect a single echo in a given illumination direction: surface elements that backscatter the radar wave cast a shadow on structures located downstream. Point clouds captured by TomoSAR can then be naturally represented in 2D by projection along each line-of-sight. PointProNets [21] generate a heightmap from the point cloud. The point cloud denoising task is then recast into a 2D noisy heightmap restoration. This strategy allows to exploit architectures from the mature domain of image denoising and reduce the memory and computation required for processing 3D data. We follow the same line of approach to restore our 3D point clouds reconstructed by TomoSAR inversion. Our contributions can be summarized as follows:

1. We propose a learned spatial regularization approach to improve pixel-wise TomoSAR reconstruction
2. We formulate the process as a point cloud denoising task in a projected domain, enabling efficient restoration and an easy access to a training set through LiDAR data
3. We show that our method successfully enhances the output of a pixel-wise TomoSAR method.

2 Proposed framework

2.1 Per pixel TomoSAR 3D point reconstruction

TomoSAR imaging model TomoSAR exploits a stack of D coregistered 2D SAR images of the same scene, acquired from slightly displaced antenna positions. Given that targets are sparsely distributed along the elevation direction in urban areas, the set of pixels at the same location in the stack, $\underline{\boldsymbol{v}} \in \mathbb{C}^D$, can be modeled as [3]:

$$\underline{\boldsymbol{v}} = \sum_{m=1}^{M} \underline{u}_m \underline{\boldsymbol{a}}(h_m) + \underline{\boldsymbol{\epsilon}}, \tag{1}$$

where M is the unknown (small) number of targets, h_m is the unknown elevation of the m-th target with reflectivity $\underline{u}_m$, the vector-valued function $\underline{\boldsymbol{a}}(\cdot)$ relates the elevation h_m of the m-th target to the phase shifts seen by each antenna ($\underline{a}_d(h) = \exp(j\xi_d h)$, where ξ_d depends only on the acquisition parameters, including the wavelength and sensor trajectories), and $\underline{\boldsymbol{\epsilon}} \in \mathbb{C}^D$ is a circular complex Gaussian random vector that accounts for the thermal noise

Considering the model (1), the tomographic 3D inversion generally aims at solving the nonlinear least squares (NLS) problem:

$$\{\hat{\boldsymbol{h}}, \hat{\underline{\boldsymbol{u}}}\} = \underset{\{\boldsymbol{h},\underline{\boldsymbol{u}}\}\in[h_{\min},h_{\max}]^M\times\mathbb{C}^M}{\arg\min} \|\underline{\boldsymbol{v}} - \sum_{m=1}^{M} \underline{u}_m \underline{\boldsymbol{a}}(h_m)\|^2, \tag{2}$$

where $\boldsymbol{h} = (h_m)_{1\leq m\leq M}$ and $\underline{\boldsymbol{u}} = (\underline{u}_m)_{1\leq m\leq M}$. For a given number of targets M, the NLS estimator is statistically efficient [24].

Prior work: fast and interpretable inversion method The computational difficulty of the problem (2) has led to the development of approximate methods in the literature, including model-based approximations (e.g., compressive sensing (CS)-based methods [35, 1, 17]) and algorithmic-based approximations (e.g., greedy algorithms [11]). These methods usually have a high computational cost due to their iterative nature and/or post-processing requirements to achieve satisfactory reconstructions. Moreover, they provide target locations without associated uncertainty measures, making the reliability of the reconstructed structures difficult to assess in the absence of ground truth data. A recent work [25] addresses these limitations through a time-efficient inversion framework with accurate target detection and localization capabilities while additionally providing uncertainty estimates. This method follows a multi-stage learning framework that recovers target information from a TomoSAR measurement vector $\underline{\boldsymbol{v}}$ in a greedy fashion. At each stage, a single target is identified and characterized by its probability of existence, elevation, and associated localization uncertainty. The contribution of this target is then subtracted from the measurements and provided as input to the following stage. To fully exploit the measurements, the target identification process is carried out at each stage by a complex-valued neural network that jointly processes the amplitude and phase information. The different stages are trained end-to-end to approximate the target distribution, modeled as a collection of Dirac impulses, by a Gaussian mixture whose means, variances, and weights correspond respectively to the estimated elevations, uncertainties, and existence probabilities. The joint estimation of these parameters improves the interpretability of the reconstruction results. In particular, the probabilities enable straightforward and informed target selection while the uncertainties provide additional confidence measures regarding target localization. Hence, this method is particularly adapted as a first step for our 3D reconstruction method.

[3] complex-valued variables are underlined to distinguish them from real-valued ones, and lowercase and uppercase bold letters are used for vectors and lists of vectors, respectively.

2.2 Restoration of TomoSAR point clouds with a diffusion model

Motivation and Formulation Enforcing spatial regularity with adapted priors during the reconstruction process is a theoretically rigorous approach to remove outliers while limiting information loss [19]. However, this strategy comes at the cost of a higher complexity and faces some challenges. As hand-crafted priors fail to preserve complex structural features, learned priors were proposed with the promise of an enhanced structural feature regularization. The development of this new strategy is, however, hindered by the absence of realistic paired TomoSAR measurements and 3D ground truth. Decoupling TomoSAR reconstruction and spatial regularization reduces the complexity by performing simpler and well-documented tasks sequentially. In this setting, deep-learning can be leveraged to learn powerful structural priors. These learned priors would be independent of TomoSAR acquisition parameters and therefore more generalizable.

In this work, the regularization task is formulated as a point cloud denoising task. This allows to exploit 3D datasets over urban areas that are readily available, in particular, LiDAR-derived datasets such as the LiDAR-HD project [10] produced by the French National Institute of Geographic and Forest Information (IGN). This dataset covers France, including multiple cities with a variety of architectural styles. Such a large dataset can be used to learn rich priors of digital surface models in urban areas.

The processing of 3D data by a neural network poses a certain number of challenges, both in terms of architecture and storage. Naively extending 2D architectures can lead to sub-optimal performances along with increased computational cost [21]. To reduce the costs associated with such a strategy, we propose working with data projected into a 2D space. This also allows the use of many network architectures designed for 2D images.

Heightmap representation We aim to denoise 3D point clouds recovered from a TomoSAR image stack, where each image, obtained from a side-looking view, projects the 3D scene onto the azimuth-range $(\vec{x}, \vec{r})$ plane, in which targets with the same range overlap. We adopt, as a projection domain, the azimuth-angle $(\vec{x}, \vec{\theta})$ plane while maintaining the same side-looking configuration. For a fixed azimuth position x, the position of a target along the angle axis is defined by the angle formed by the vertical direction and the segment connecting the sensor to the target. As illustrated in Figure 2, in this geometry, layover and shadows phenomena are avoided: at each location (x, θ), a single surface element is visible to the radar. An height map in the azimuth-angle geometry indicates the height of scatterers located on that surface element. In practice, a 3D point cloud $\mathbf{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \dots, \boldsymbol{p}_N\}$ is transformed into a 2D image, referred to as a heightmap, where the pixel at position (i, j) contains the height z of the point $\boldsymbol{p}^\star \in \mathbf{P}$ that projects into this azimuth/angle cell and is the closest to the radar,

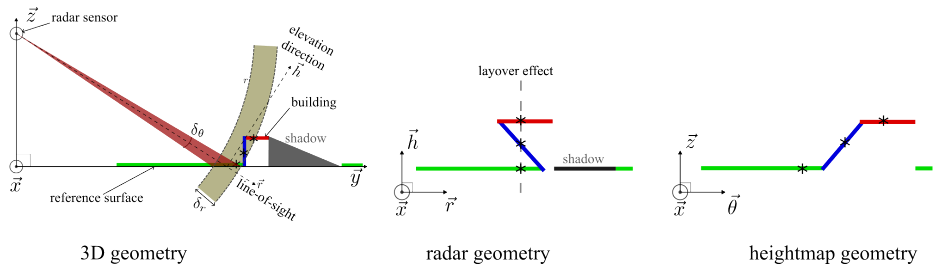


Fig. 2: Building appearance in different geometries.

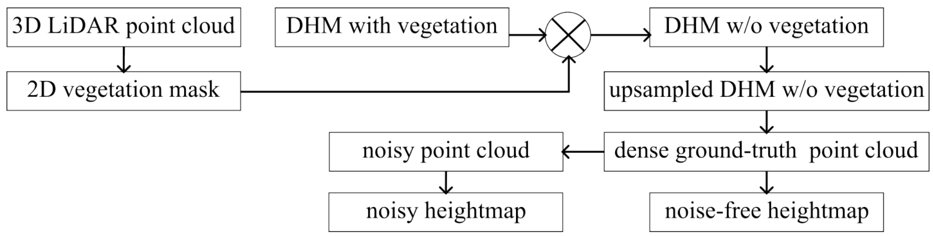


Fig. 3: Flowchart of paired noisy/noise-free heightmaps creation from LiDAR products.

i.e., such that:

$$d(\boldsymbol{p}^{\star}, \text{radar}) = \min\{d(\boldsymbol{p}, \text{radar}) : \boldsymbol{p} \in \mathbf{P},\ x_p \in [x_{\min} + (i-1)\delta_x, x_{\min} + i\delta_x[,\ \theta_p \in [\theta_{\min} + (j-1)\delta_\theta, \theta_{\min} + j\delta_\theta[\}, \tag{3}$$

where $d(\boldsymbol{p}, \text{radar})$ denotes the distance separating point $\boldsymbol{p}$ from the radar, δ_x is the azimuth resolution of the SAR sensor, and δ_θ is the angular resolution corresponding to the minimal angular difference between two consecutive ground points ($z = 0$) at the extremities of a radar pixel, as illustrated in the image on the left side of Figure 2.

Dense 3D point cloud generation The angular resolution δ_θ used in our heightmap projection depends on the range resolution δ_r of the SAR sensor. As this value decreases, the number of empty pixels in the ground truth heightmap may increase if the corresponding point cloud is not sufficiently dense. We propose a pipeline, illustrated in Figure 3, that adapts the density of the point cloud according to the desired angular resolution to avoid empty pixels in the ground truth heightmap. The process exploits both LiDAR point clouds and digital height models (DHM, nadir/down-looking view of the scene). While the DHM is the main component for creating a dense point cloud, the LiDAR point cloud (target coordinates and associated labels) is used to remove the vegetation from the DHM. This vegetation removal operation is motivated by the fact that TomoSAR 3D reconstructions of urban areas usually don't contain vegetation

(a) Digital height model (b) vegetation mask (c) Digital height model without vegetation

Fig. 4: Result of the vegetation removal process from the denoiser training dataset: the vegetation map is obtained using the labels of the LiDAR point cloud.

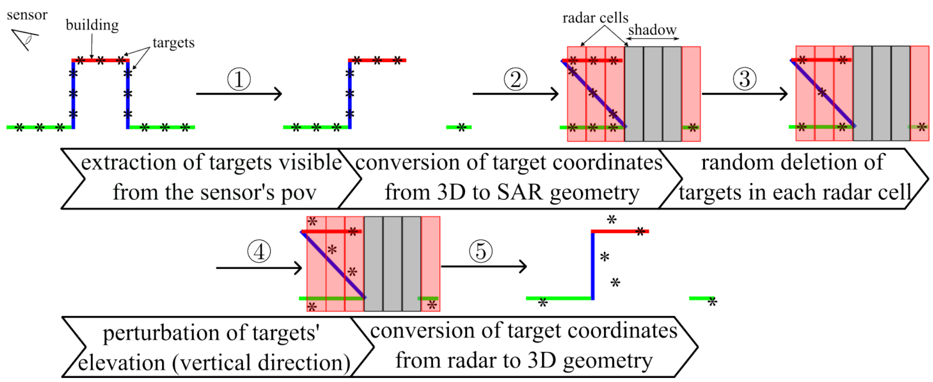
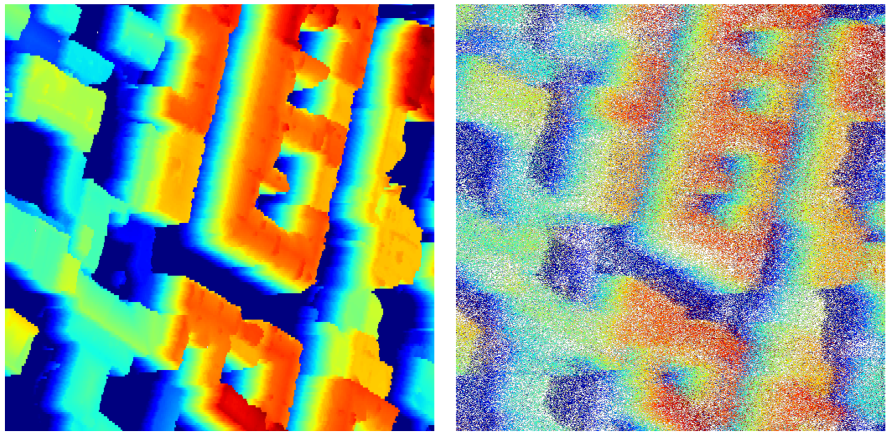


Fig. 5: Generation of noisy point clouds featuring missing targets and localization errors.

since these structures are not temporally stable or have a negligible response compared to their surroundings. This operation is useful to avoid hallucinations from the denoiser to be trained on this dataset, as the proportion of vegetation is usually non-negligible (see Figure 4). After this step, the DHM is upsampled from a 50×50 centimeters spatial resolution to a sufficient resolution, such that the angle coordinates of targets recovered from successive pixels differ by less than δ_θ if they are on the same horizontal surface. Using this property, it becomes possible to detect the position of vertical structures missing from the DHM. A dense point cloud is then obtained by adding points vertically at those positions.

Noisy heightmap generation Noisy data are generated from ground truth point clouds $\mathbf{P}$ to reproduce the characteristics of noisy TomoSAR 3D recon-



(a) noise-free heighmap (b) noisy heightmap

Fig. 6: Paired noise-free/noisy heightmap: since the noisy heightmap contains some outliers (large height values), both images are displayed in the range of the ground truth for a better visualization; the color white represents empty pixels.

structions. The different stages of this process are illustrated in Figure 5. 1 We begin by excluding targets that are invisible from the perspective of a radar sensor. These correspond to the set of targets $p \in \mathbf{P}$ satisfying the following conditions:

$$\exists \boldsymbol{p}' \in \mathbf{P} \text{ such that } \theta_p = \theta_{p'} \text{ and } d(\boldsymbol{p}, \text{radar}) > d(\boldsymbol{p}', \text{radar}) \tag{4}$$

where θ_p is the angle coordinate of $p \in \mathbf{P}$. 2 Next, the coordinates of the visible targets are converted into radar coordinates. In this domain, noisy TomoSAR reconstructions are characterized, in each radar cell, by potential missing targets and target localization errors in the elevation direction. 3 In urban areas, missing targets are not specifically localized; this results from a combination of factors, including strong target interference, large target amplitude difference, and high thermal noise. This phenomenon is thus simulated by randomly removing targets in each radar cell independently of their localization. 4 The elevation $h \in \mathbb{R}$ of a target can be modeled as a Dirac distribution $\delta(h)$ which is equal to 0 everywhere except at h where it is $+\infty$. This distribution is equivalent to the Gaussian distribution $\mathcal{N}(h, \sigma^2)$ where σ tends toward 0^+ [2]. Target localization error can therefore be simulated by allowing a larger variance around the true position. Here, the variance is drawn according to the distribution of uncertainty estimates of the pixel-wise TomoSAR inversion method proposed in [25], the first step of our approach, on real data (a log-normal distribution). This is motivated by the interesting properties of the uncertainties on simulated data. Indeed, [25] demonstrated empirically on simulated data that uncertainty estimates are well-calibrated. 5 Finally, the coordinates are reprojected into 3D geometry. Figure 6 shows a paired noise-free/noisy heightmap generated using our noising process.

Diffusion-based restoration Image restoration aims at recovering an unknown image given a corrupted version of it. This is typically ill-posed, meaning that multiple solutions are possible. Due to the demanding nature of remote sensing applications, deep learning models are expected to provide not only accurate predictions but also uncertainty estimates as a measure of prediction reliability. Conditional diffusion model-based uncertainty quantification has been shown to be more effective than other existing methods [34] and has been considered in this work.

Diffusion models were introduced with the goal of learning the distribution of a dataset in order to generate similar data [23]. They were popularized by the DDPM paper [9], in which the authors demonstrate their ability to generate very high-quality images. Training this type of model involves two main phases: the forward process and the reverse process. The forward process consists of progressively degrading an image $\mathbf{X}_0$, following the distribution $p(\mathbf{X_0})$, by injecting increasingly more Gaussian noise until pure noise (i.e., i.i.d. pixel values following a standard normal Gaussian distribution). The reverse process then aims to learn the inverse transformation. In this step, a neural network is trained to predict the noise added at each stage of the forward process. Following the demonstration presented in the paper [9], diffusion models have been applied to various image restoration tasks.

Using the notation commonly adopted in diffusion-model literature, our problem can be summarized as follows: we are given a dataset $\mathcal{D} = \{\mathbf{X}^i, \mathbf{Y}^i\}_{i=1}^N$ consisting of pairs of clean heightmaps $\mathbf{X}^i$ and noisy heightmaps $\mathbf{Y}^i$, and we aim to learn the conditional distribution $p(\mathbf{X}|\mathbf{Y})$. To address this problem, the paper [22] proposes adapting DDPM for conditional image generation. The authors introduce an additional conditioning on the noisy image $\mathbf{Y}$ during the reverse process, while the forward process remains unchanged. In practice, conditioning on $\mathbf{Y}$ simply amounts to adding $\mathbf{Y}$ as an additional input to the neural network.

Despite achieving good performance on several image restoration tasks, this method requires several hundred calls to a neural network during inference. This increases the processing time for a single image and creates a practical issue in our context. Approaches that drastically shorten the diffusion process for image restoration have been proposed. In particular, ResShift (Residual Shifting) [27] reformulates the diffusion process such that the final state of the forward process corresponds to a Gaussian distribution centered on the corrupted image $\mathbf{Y}$ rather than on 0. The image $\mathbf{X}_t$ at time t is defined as follows:

$$\mathbf{X}_t = (1 - \eta_t)\mathbf{X}_0 + \eta_t \mathbf{Y} + \sqrt{\eta_t}\kappa_t \mathbf{Z}_t \tag{5}$$

where $\eta_t \in [0, 1]$, $\mathbf{Z}_t \sim \mathcal{N}(0, \mathbf{I})$, and κ_t controls the noise level between the different stages of the process. This formulation forms a convex combination of the clean image $\mathbf{X}_0$ and the image to be denoised, $\mathbf{Y}$, to which Gaussian noise $\mathbf{Z}_t$ is added. In addition to the implicit conditioning induced by Equation

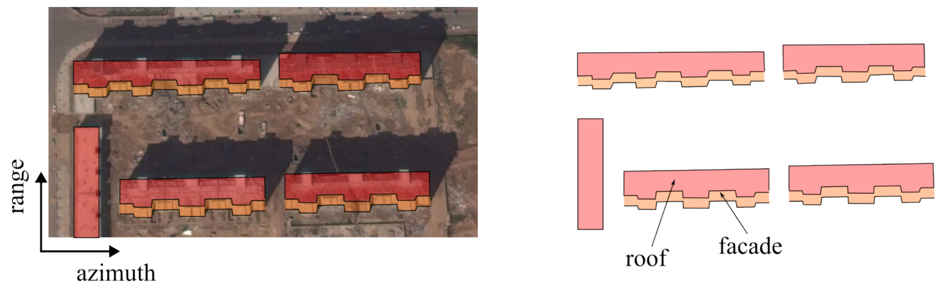


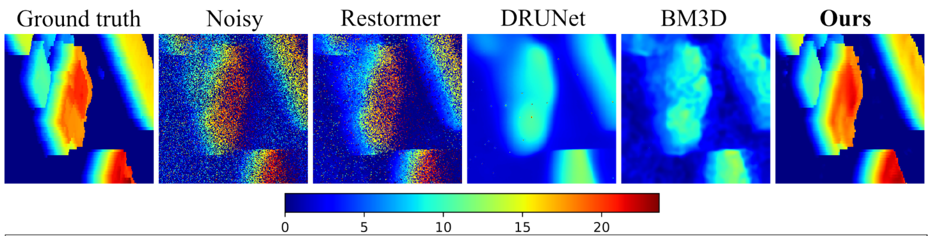
Fig. 7: Optical image (left) of the area of interest and the approximate structure composition (right).

5, ResShift also employs explicit conditioning by taking the corrupted image **Y** as an input to the network. This approach is generic. It has been shown to perform well on numerous image restoration tasks even when using fewer than 20 diffusion steps. It is well-adapted to our context.

3 Experiments

Dataset and setup The performance of the proposed framework is studied on real data using the SARMV3D Yuncheng dataset [18]. This dataset consists of a stack of 8 coregistered SAR images of the same scene acquired in a single flyover. The spatial resolution of these images is approximately 10 centimeters in azimuth and 13 centimeters in range. An optical image of the area of interest with an approximate structure composition is shown in Figure 7. The image comprises five main structures, four of which are residential buildings of approximately 54 meters high and sharing a similar shape. The spatial characteristics of this dataset (geometric baselines, spatial resolution . . .) are also used for all training data simulations. This applies to the training for both the TomoSAR inversion method and the heightmap denoiser. We followed the same training protocol for the TomoSAR inversion method as described in [25]. Regarding the heightmap denoiser architecture, a classical UNet, representing a total of 30 million parameters, is chosen. ResShift [27] is trained with $T = 15$ timesteps using the noising schedule $(\eta_t, \kappa_t)_{0\leq t\leq T}$ proposed by the authors. The training dataset is built from 96 non-overlapping tiles covering the city of Paris, each representing an area of 1km^2: 94 tiles are exploited in the learning phase, and the remaining tiles are used for model validation. By subdividing the height maps into 256×256 patches with an overlap of 128 pixels, we created two disjoint sets consisting of 252 118 images for training and 5 476 images for model validation. Inputs are scaled tile-wise, leveraging the ground truth maximum value, and empty pixels are filled with zeros.

Quantitative analysis In this paragraph, the importance of using an adapted method for the heightmap denoising task is studied on simulated data. For this

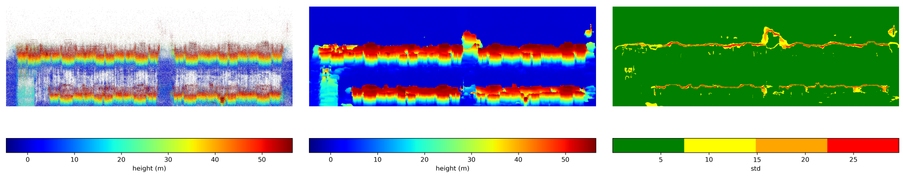


Average SNR:	-0.12 dB	1.84 dB	5.06 dB	5.23 dB	17.32 dB

Fig. 8: Filtering results on an example of the test dataset and average SNR over the entire dataset.

purpose, our model, specifically trained for this degradation, is compared to several baseline methods (standard denoisers designed for additive white Gaussian noise removal): BM3D [3], DRUNet [31], and Restormer [28], corresponding to different filtering strategies. The first is a non-blind patch-based filtering method, and the other two are non-blind and blind deep-learning-based approaches, respectively. The noise level required by the non-blind denoisers is calibrated on images from our validation dataset. All methods are evaluated on a collection of 100 images of size 256×256 (these images were neither used for training nor validation of our diffusion model). Figure 8 reports visual results on one particular image and the average SNR computed on the entire testing dataset. As expected, the Gaussian denoisers fail to correctly restore the heightmaps since the degradation significantly departs from additive Gaussian noise. In contrast, our model produces a heightmap visually close to the ground truth. These observations are in accordance with the SNR values: our method shows a significantly larger improvement, on average, of the SNR compared to the other methods (+12dB compared to the best baseline methods).

Qualitative analysis In this paragraph, we qualitatively study the results of our framework applied to the region of interest on the real Yuncheng TomoSAR dataset. The TomoSAR point cloud reconstruction step is first applied. Each target of the reconstructed point cloud is associated with an existence probability and an empirically well-calibrated uncertainty measure of its elevation. The input to our heightmap diffusion-based denoiser is generated with targets that verify a criterion defined on the probabilities, selecting the most likely targets [25]. By applying the denoiser 100 times on this data, we produced different possible heightmaps to quantify the uncertainty. Figure 9 presents one possible reconstruction and the uncertainty map computed as a pixel-wise variance of the reconstructions. Considering the optical image, the filtered heightmap 9b is fairly satisfactory. The uncertainty map 9c is also highly informative. Indeed, high uncertainty values are mostly concentrated around roof-ground transitions, but more importantly, around small structures that don't appear in the optical image. This suggests that the uncertainty measure could be used to detect hallucinations and identify challenging areas. In Figure 10, the 3D point cloud



(a) noisy heightmap (b) filtered heightmap (c) uncertainty map

Fig. 9: Filtering result of the heightmap derived from a TomoSAR reconstruction of the real scene.

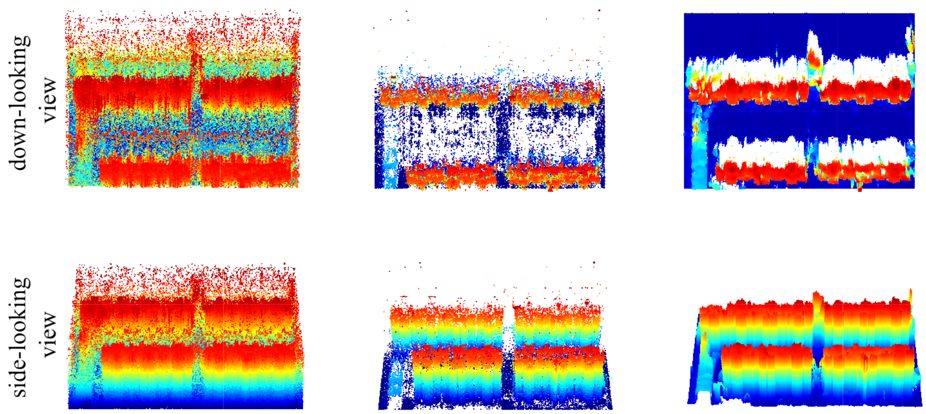


Fig. 10: Comparison of the regularized scene with low uncertainty structures from the TomoSAR inversion method: from left to right, we have the noisy point cloud, the result using the filtering method from [25], and our result.

generated from one output of the denoiser is compared to the point cloud obtained by filtering the noisy point cloud using an empirical criterion defined on the uncertainty values predicted by the TomoSAR inversion method. The different views show a strong overlap between the structures of the filtered point clouds, providing further proof that the trained denoiser recovers faithful information from the noisy heightmap. Moreover, the denoiser retrieves a denser point cloud with fewer outliers and a clear delineation of building shadows, corresponding to areas not observed by the SAR sensor (see the down-looking view in the first row of Figure 10). Overall, the trained model demonstrates good outlier removal and missing data completion capabilities. It significantly reduces noise while preserving the underlying structures. Further studies are necessary to confirm its robustness to architectural style shifts, but initial results are promising as the model was trained on data from Paris and applied to data from China.

4 Conclusion

In this paper, we propose a framework that combines per-pixel TomoSAR inversion and spatial regularization for 3D reconstructions. The regularization is framed as a point cloud denoising task and performed in a 2D projection domain, reducing the computational burden. A learning-based method is proposed to capture complex features of urban structures. And as the learning process is independent of TomoSAR data acquisition parameters, the resulting model can be applied to reconstructions from different measurement configurations without the need for retraining. A diffusion model is adopted, providing high-quality reconstructions and associated uncertainty measures for risk-aware decision making. The time-consuming aspect of the classical diffusion model is addressed by resorting to a fast-diffusion approach. Future work will aim to incorporate the empirically well-calibrated uncertainty estimates from the TomoSAR inversion model in the diffusion-based denoising process to adapt the regularization strength. The experimental evaluation of our approach will also be enriched with a comparison to baseline point cloud denoising baselines and to other TomoSAR methods with spatial regularization.

Acknowledgements This project has been funded by the Agence de l'Innovation de Défense - AID - via the Centre Interdisciplinaire d'Études pour la Défense et la Sécurité - CIEDS (project 2023 ALIA).

References

1. Budillon, A., al.: Three-Dimensional SAR Focusing From Multipass Signals Using Compressive Sampling. IEEE TGRS **49**(1), 488–499 (2011)
2. Cohen, S.B., al.: Approximating the dirac distribution for fourier analysis. Journal of Computational Physics **93**(2), 312–324 (1991)
3. Dabov, K., al.: Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. IEEE TIP **16**(8), 2080–2095 (2007)
4. Fei, B., al.: Towards unified representation of multi-modal pre-training for 3d processing. IEEE TVCG (2025)
5. Gao, J., al.: Fast Super-resolution 3D SAR Imaging Using an Unfolded Deep Network. In: IEEE ICSIDP. pp. 1–5 (2019)
6. Gao, W., al.: Deep learning for 3D point clouds. Springer (2025)
7. Guo, M., al.: Pct: Point cloud transformer. CVM **7**(2), 187–199 (2021)
8. Hershey, J.R., al.: Deep Unfolding: Model-Based Inspiration of Novel Deep Architectures. ArXiv **abs/1409.2574** (2014)
9. Ho, J., al.: Denoising Diffusion Probabilistic Models (2020)
10. IGN: Programme lidar hd : vers une nouvelle cartographie 3d du territoire (2022), https://www.ign.fr/institut/programme-lidar-hd-vers-une-nouvelle-cartographie-3d-du-territoire
11. Li, J., al.: Efficient mixed-spectrum estimation with applications to target feature extraction. IEEE TSP **44**(2), 281–295 (1996)
12. Li, J., al.: Joint point cloud upsampling and cleaning with octree-based cnns. CVM (2026)

13. Lohani, B., al.: Airborne LiDAR Technology: A Review of Data Collection and Processing Systems. Proceedings of the National Academy of Sciences, India Section A: Physical Sciences **87**, 567–579 (2017)
14. Lombardini, F., al.: Adaptive spectral estimation for multibaseline SAR tomography with airborne L-band data. In: IEEE IGARSS. Proceedings (IEEE Cat. No.03CH37477). vol. 3, pp. 2014–2016 (2003)
15. Lombardini, F., al.: Towards a Complete Processing Chain of Multibaseline Airborne InSAR Data for Layover Scatterers Separation. In: 2007 Urban Remote Sensing Joint Event. pp. 1–6 (2007)
16. Qi, C.R., al.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE CVPR. pp. 652–660 (2017)
17. Qian, K., al.: γ-Net: Superresolving SAR Tomographic Inversion via Deep Learning. IEEE TGRS **60**, 1–16 (2022)
18. QIU, X., al.: SARMV3D-1.0: Synthetic Aperture Radar Microwave Vision 3D Imaging Dataset(in English). J. Radars **10**(R21112), 485 (2021)
19. Rambour, C., al.: Introducing Spatial Regularization in SAR Tomography Reconstruction. IEEE TGRS **57**, 8600–8617 (2019)
20. Reigber, A., al.: First demonstration of airborne SAR tomography using multibaseline L-band data. IEEE TGRS **38**(5), 2142–2152 (2000)
21. Roveri, R., al.: Pointpronets: Consolidation of point clouds with convolutional neural networks. Computer Graphics Forum **37**, 87–99 (05 2018)
22. Saharia, C., al.: Image Super-Resolution via Iterative Refinement. IEEE TPAMI **45**(4), 4713–4726 (2023)
23. Sohl-Dickstein, J., al.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics (2015)
24. Stoica, P., al.: Spectral Analysis of Signals. Pearson Prentice Hall (2005)
25. Ulondu Mendes, C., al.: Fast tomographic SAR inversion by unrolling a greedy algorithm for sparse reconstruction (2026)
26. Wang, Y., al.: MAda-Net: Model-Adaptive Deep Learning Imaging for SAR Tomography. IEEE TGRS **61**, 1–13 (2023)
27. Yue, Z., al.: ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting (2023)
28. Zamir, S., al.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF CVPR. pp. 5728–5739 (2022)
29. Zeng, T., al.: Exploring Spatial Feature Regularization in Deep-Learning-Based TomoSAR Reconstruction: A Preliminary Study and Performance Analysis. IEEE TGRS **63**, 1–27 (2025)
30. Zhan, X., al.: Unified Learning and Reconstruction for Robust Tomographic SAR Reconstruction: A Model-Driven Framework. IEEE TGRS **63**, 1–27 (2025)
31. Zhang, K., al.: Plug-and-Play Image Restoration With Deep Denoiser Prior. IEEE TPAMI **44**(10), 6360–6376 (2022)
32. Zhang, T., al.: Point cloud mamba: Point cloud learning via state space model. In: Proceedings of the AAAI conference on artificial intelligence. vol. 39, pp. 10121–10130 (2025)
33. Zhou, L., al.: Point cloud denoising review: from classical to deep learning-based approaches. Graphical Models **121**, 101140 (2022)
34. Zhu, Q., al.: Capturing Uncertainty in Regression via Conditional Diffusion Models (2026)
35. Zhu, X.X., al.: Tomographic SAR Inversion by L_1-Norm Regularization-The Compressive Sensing Approach. IEEE TGRS **48**(10), 3839–3846 (2010)