Real-GDSR: Real-World Guided DSM Super-Resolution via Edge-Enhancing Residual Network

Daniel Panangian [*], Ksenia Bittner

Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling, Germany
(daniel.panangian, ksenia.bittner)@dlr.de

KEY WORDS: Super-Resolution, Digital Surface Model (DSM), Residual Network, Diffusion, Satellite Imagery

ABSTRACT:

A low-resolution digital surface model (DSM) features distinctive attributes impacted by noise, sensor limitations and data acquisition conditions, which failed to be replicated using simple interpolation methods like bicubic. This causes super-resolution models trained on synthetic data does not perform effectively on real ones. Training a model on real low and high resolution DSMs pairs is also a challenge because of the lack of information. On the other hand, the existence of other imaging modalities of the same scene can be used to enrich the information needed for large-scale super-resolution. In this work, we introduce a novel methodology to address the intricacies of real-world DSM super-resolution, named REAL-GDSR, breaking down this ill-posed problem into two steps. The first step involves the utilization of a residual local refinement network. This strategic approach departs from conventional methods that trained to directly predict height values instead of the differences (residuals) and utilize large receptive fields in their networks. The second step introduces a diffusion-based technique that enhances the results on a global scale, with a primary focus on smoothing and edge preservation. Our experiments underscore the effectiveness of the proposed method. We conduct a comprehensive evaluation, comparing it to recent state-of-the-art techniques in the domain of real-world DSM super-resolution (SR). Our approach consistently outperforms these existing methods, as evidenced through qualitative and quantitative assessments.

1. INTRODUCTION

Elevation data are essential for a wide range of applications across multiple sectors. They also contribute significantly to our improved understanding and management of the Earth's resources and environment. These data are collected through various techniques, typically aerial data capture and light detection and ranging (LiDAR) sensors. A digital elevation model (DEM) is a digital representation of the Earth's topography, specifically focusing on the bare ground surface. DEMs play a crucial role in applications like terrain analysis, hydrological modeling, geological studies, precision agriculture, and infrastructure planning. A digital surface model (DSM) in the other hand is a representation which includes all objects on it, such as vegetation and buildings. DSMs are widely applied in topographic mapping, environmental simulations (Aktaruzzaman and Schmitt, 2009), 3D city modeling and planning. Recent remote sensing technology provides several ways to measure the 3D urban morphology. Conventional ground surveying, stereo airborne or satellite photogrammetry, interferometric synthetic aperture radar (InSAR), and LiDAR are the main data sources used to obtain high-resolution DSM. However, each of these scenarios has its own set of advantages and disadvantages.

Data derived from terrestrial and airborne systems offer high spatial resolution but have limited coverage and can encounter precision-related issues. Spaceborne missions, for example Cartosat 1, provide global coverage but may lack the finest level of resolution achieved by terrestrial and airborne methods due to challenges in capturing high-resolution data from space. The resolution of such data have a substantial impact in different fields of operation. Improving measurement equipment precision is the most straightforward way to acquire high-resolution elevation data, but it is a difficult, costly, and time-consuming procedure. Therefore, generating high-resolution data without extra cost becomes a key concern of researchers from various fields. A practical way to solve this problem is to enhance the resolution of easily obtained
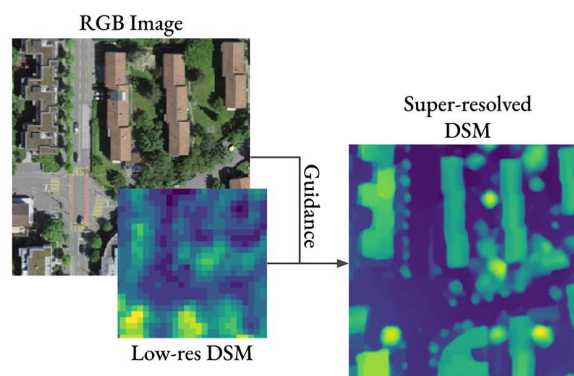

Figure 1: Guided DSM super-resolution: given a low-resolution dsm and a high-resolution guide image, our method predicts a high-resolution DSM. The figure shows an example output of the proposed method on low-resolution DSM with a factor of 10.

low-resolution data (Xu et al., 2015a). In computer vision, various algorithms have been developed for single image super-resolution task. Some of the algorithms have been explored for DEM super-resolution. Both depth estimation (Godard et al., 2017, 2019; Bhat et al., 2021) and depth super-resolution task (Voynov et al., 2018; He et al., 2021) are two intriguing applications that can be applied to DEMs and DSMs. Recently a method termed guided depth map super-resolution (GDSR) is gaining popularity. The idea is that a different imaging modality of the same object can be used as a guide for super-resolving the low-resolution image by injecting the missing high-frequency content. Research into GDSR has a long history (Patterson et al., 1992; Izraelevitz, 1994). The proposed solutions range from classical, entirely hand-crafted schemes (Ham et al., 2017) to fully learning-based methods (Hui et al., 2016a), while some recent works have combined the two with promising results (Lutio et al., 2022; Metzger et al., 2023).

---

*Corresponding author

(a) Bicubic Downsampled
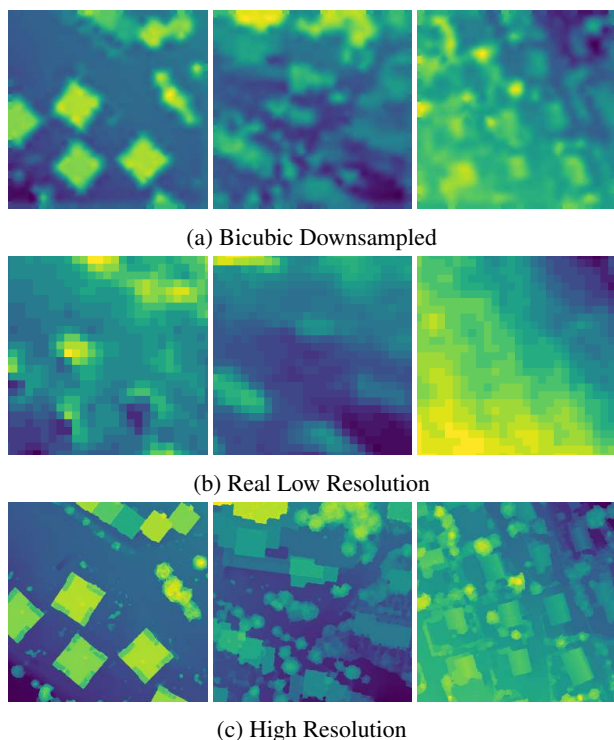

(b) Real Low Resolution

(c) High Resolution

Figure 2: Examples of real low-resolution DSMs compared to the bicubic-downsampled and high-resolution DSMs. Note that real low-resolution DSMs preserve less information in comparison to their bicubic-dowsampled counterparts.

However, generally both single image super-resolution (SISR) and GDSR methods restrict themselves to super-resolving images downsampled by a simple and uniform degradations (i.e, bicubic downsampling). Real low-resolution DSMs, in the other hand, do not preserve as much information (see Fig. 2) and applying these methods on real-world DSMs is becoming a challenge (Cai et al., 2019; Wang et al., 2021a). In this paper, we focus on super-resolving DSMs guided by their corresponding optical images, which is depicted in Fig. 1. We focus our research on urban DSMs because they contain richer information which hard to be restored. This focus also aligns with the evolving interest in leveraging DSMs for building reconstruction (Bittner et al., 2018; Partovi et al., 2019; Bittner et al., 2020; Wang et al., 2021b; Gui and Qin, 2021; Stucker and Schindler, 2022; Stucker et al., 2022).

In summary, our contributions are as follows:

1. We propose a novel super-resolution framework guided by optical images on real-world low-resolution DSMs. To the best of our knowledge, we are the first to develop a guided super-resolution network trained on real-world DSMs.

2. We achieve this by utilizing both local and global context. We improve the conventional methods in two ways: local refinement and edge-enhancing diffusion. A local refinement network with small receptive field is utilized to improve the low-resolution DSM. This shallow network can repair some missing regions and structures, according to the surrounding local regions after a coarse bicubic-interpolation stage. An edge-enhancing diffusion network is used to further smoothen the super-resolved result. This network can further improve the visual quality using the global edge information, especially for removing outliers and preserving height discontinuities.

3. We demonstrate the effectiveness of our approach by comparing it to other state-of-the-art networks, trained in the same setting. Furthermore, we show that our approach is feasible of achieving a 10x super-resolution for low-resolution DSMs from satellite data. We also found that our proposed approach outperformed the other works in terms of both qualitative and quantitative results on unseen data.

2. RELATED WORK

2.1 Single Image Super-Resolution

Single image super-resolution refers to generation of a high-resolution image from a low-resolution image. Recently SISR methods has shifted towards example-based approaches. However in those early years interpolation-based techniques like linear, bicubic or Lanczos are widely used. The idea is new pixels are estimated by interpolating given pixels. But these methods suffers from blurry results on high-frequency regions. Learning or example-based methods aim to gather insight information from paired low and high-resolution images to understand missing details in low-resolution images. SRCNN (Dong et al., 2015) is one of the first approaches to demonstrate the use of neural networks to learn the nonlinear mapping of the images in the image space. The method utilize bicubic interpolation of low-resolution image followed by high-dimensional vector representation and ended by reconstruction of the vectors to the pixel space. VSDR (Kim et al., 2016a) was designed to increase the efficiency of SRCNN, by predicting the residuals rather than the actual pixel values and, to boost the overall performance by adding more layers. Wang et al. (2015) introduced sparse coding to the training which enable the model to enlarge the images to the desired scale factor progressively. Deep recursive layers are introduced in DRCN (Kim et al., 2016b) to reduce the number of parameters.

Furthermore, the introduction of generative adversarial networks (GANs) inspired the implementation of GAN for super-resolution. In GAN the generator produces high-resolution image and the discriminator is trained to distinguish between patch of the original image and patch which are produced by the generator. Two components are designed to defeat each other in a zero-sum game (Goodfellow et al., 2014). SRGAN (Ledig et al., 2017) is the first one to demonstrate the ability to pay more attention to visual effects, introducing adversarial and perceptual losses. ESRGAN (Wang et al., 2019) goes another step by introducing relativistic discriminator and removing batch normalization layers.

But different than images, depth maps contain piecewise affine regions that have sharp depth discontinuities and no textures (Riegler et al., 2016b). Besides, they are sensitive to artifacts (Xie et al., 2015). In this paper, we focus on the training strategy and framework of super-resolution networks when dealing with real low resolution depth maps.

2.2 Guided Depth Map Super-Resolution

GDSR has become an essential topic in multi-modal image processing and super-resolution. The idea behind it is that there are statistical co-occurrences between the texture edges of RGB images and the discontinuities of depth maps (Xie et al., 2015). Hence, information in RGB images can be utilized to restore low-resolution depth maps. Three categories have been identified recently for traditional methods for GDSR: learning-free, learning-based and hybrid approaches. Early work on GDSR consisted mostly of filter-based and optimization methods which require no training. Filter-based methods focus on preserving sharp depth edges under the guidance of the intensity image. For example,
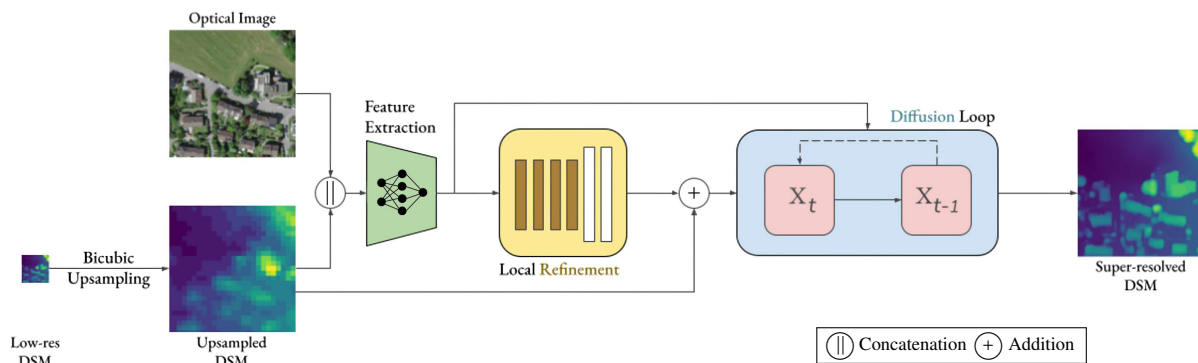
Figure 3: Summary of the proposed architecture. Real-GDSR comprises mainly a two-step process: Initially, high-dimensional features are extracted from both bicubic-upsampled low-resolution DSM and high-resolution optical image by a pre-trained model. Subsequently, a local refinement network refines the upsampled DSM by incorporating residual blocks and upsampling operations, followed by a diffusion network, which iteratively enhancing the refined upsampled DSMs, emphasizing edge features from the high-resolution optical image.

RGB images guide the acquisition of bilateral weights (Kopf et al., 2007). Optimization-based approaches use diverse data priors to construct energy functions, with data-fidelity regularization limiting the solution space. Several other methods (Xie et al., 2015; Diebel and Thrun, 2005) employ random field models. Other than that Liu and Gong (2013) demonstrated an early learning-free application of anisotropic diffusion. Initial learning-based techniques, such as bimodal co-sparse analysis (Kiechle et al., 2013) and joint dictionary learning (Tosic and Drewes, 2014), learn the relationship of RGB and depth information. Deep learning models are introduced to utilize neural networks to learn the mapping from low-resolution to high-resolution images. MSG-Net (Hui et al., 2016b) built on a U-Net deep network architecture and learns the residual errors of bicubic interpolations by embedding the source at the smallest scale. Kim et al. (2021) propose deformable kernel networks (DKN) and its fast implementation (FDKN), that learn sparse and spatially-invariant filter kernels. He et al. (2021) employ a high-frequency guidance module to embed the guide details into the depth map. Deep learning techniques have been used recently by researchers to improve the ability of predicting outputs of given inputs that have not encountered within formal frameworks. Riegler et al. (2016a) train a neural network by unrolling the optimization phases of a first-order primal-dual algorithm, enabling them to train their deep feature extractor throughout the training process. Using a graph-based, MRF-style optimizer, Lutio et al. (2022) apply the implicit function theorem. DADA (Metzger et al., 2023) achieved state-of-the-art performance by adapting the concept of guided anisotropic diffusion with deep convolutional networks.

2.3 DSM Super-Resolution

Research on DSM super-resolution is limited, despite its significance in remote sensing. The majority of research focuses on DEM superresolution. This includes interpolation-based methods like bicubic, and bilinear are used for DEM enhancement which results in smooth terrain models. Xu et al. (2015b) proposed a super-resolution algorithm based upon non-local means. It operates by using a predetermined equation to search for similar patches across the training set. Weights determined by the searching phase are then used to upscale the resolution of the target DEM. Deep learning-based methods also derived from the advances in SISR domain. D-SRCNN (Chen et al., 2016) was proposed based on SRCNN (Dong et al., 2015). Later the same author implemented EDSR (Lim et al., 2017) for the same purpose (Xu et al., 2019). Demiray et al. proposed a DEM super-resolution model,

namely D-SRGAN (Demiray et al., 2021a), with the implementation of SRGAN (Ledig et al., 2017), and EffecientNetV2 (Tan and Le, 2021) for DEM SR (Demiray et al., 2021b).

So far none of the methods targeted specifically for urban DSMs. Because of their characteristics, performing large-scale super-resolution on such data is still a challenge. First, these mentioned models sofar trained only on the data generated by bicubic interpolation which only work well on clean low-resolution data with simple degradations. This is inconsistent with real-world needs, where low-resolution data have more complex degradations. Second, urban DSMs provide more details which even harder to reconstruct. To address this conflict, we proposed a practical solution, where we collect real low and high-resolution DSMs pairs and include guiding information in both training and inference. Optical images of the same scenes are obtained to act as guidance. Other than that, We utilize both local and global context in our approach to handle the characteristics of DSMs. In the following, we explain in more detail how we achieve this.

3. METHODOLOGY

Real-GDSR consists mainly of two steps: First as a pre-step, a pre-trained model is used to extract features from both optical image and low-resolution DSM. Then, we use a residual convolutional neural network (CNN)-based network, namely the local refinement network, whose objective is to take as input high-dimensional features and transform them into a high-resolution DSMs. Second, we use an anisotropic diffusion network to further enhance the refined DSMs focusing on the edge features from the high-resolution optical image. Figure 3 shows an overview of our proposed pipeline. In the following subsections, we describe each component of our pipeline in more details.

3.1 Local Refinement

In prior works in DEM super-resolution, we observe that many existing works often follow a common design concept for image super-resolution, where their networks built of upsampling layers and have very large receptive field, for example, a U-Net like architecture or using multiple dilation convolution or attention layers. In this work, taking inspiration from image inpainting task, we see super-resolution as an image-to-image translation task where we assume that an initial coarse high-resolution DSM of the observed scene has already been generated with existing

traditional method like bicubic and later refined within the training. We highlight that for DSMs a network with small receptive field is enough to fulfill the task. For the local refinement, we create a shallow network with four residual blocks and two upsampling procedures (see middle part of Fig. 3). Therefore, this network has small receptive fields. This architecture eliminates the effect of distant and unsuccessful filling contents and allows for the restoration of missing sections or buildings utilizing local information around them. Additionally, we incorporate a long skip connection that adds the initial DSM straight to the final downsampling layer's output, allowing the network to regress residuals rather than absolute height.

3.2 Edge-Enhancing Diffusion

After the local refinement process, structural features are restored with guidance of surrounding local regions. However, different than images, DSMs are sensitive to outliers. For this purpose, we introduce a diffusion-based edge-enhancement which helps broaden the scope of information captured. Anisotropic diffusion can be understood as an adaptive filtering technique aimed at smoothing while preserving inter-region content such as edges or boundaries, which are crucial for image interpretation. This is achieved by applying inhomogeneous diffusion, where the diffusivity is guided by a scalar function or diffusion tensor derived from the gradients of the evolving image. The concept of calculating diffusivity from a guide image has been investigated in the area of edge enhancement and semantic segmentation.

Inspired by prior work in guided depth super-resolution (Metzger et al., 2023), we implement a similar network, where the diffusion component mirrors traditional optimization approaches solved via an iterative diffusion loop. For each iteration, multiple steps of anisotropic diffusion are conducted. Here, the diffusion weights are influenced by the guide to minimize diffusion at boundaries with high contrast and enhance diffusion within regions that are homogeneous. A convolutional feature extractor is used to set diffusion weights by passing the guide through. Therefore, the process can transfer edge information from the guide to preserve depth discontinuities in the target image.

Given a source image S a guide $G \in H \times W \times C$, where $C = 3$ for RGB images or a larger number for deep features, the first step is to initialize $X_0 \in H \times W$ with an upsampled version of S. The diffusion step is defined as

$$\hat{x}_t^p = x_{t-1}^p + \lambda \cdot \sum_{n \in N_4(p)} \left(x_{t-1}^n - x_{t-1}^p \right) \cdot c(g^p, g^n), \qquad (1)$$

where x_t^p denotes the pixel value of X_t at location p (and similarly for g^p). $N_{4(p)}$ denotes the four-neighbours of pixel p. λ controls the rate of diffusion. For four-neighbours, λ should be set to < 0.25 to ensure numerical stability. Diffusion coefficients for the neighboring pairs of pixels are calculated by function c which formulated based on their values in the guide. A higher diffusion coefficient means that information spreads more freely across neighboring pixels, resulting in stronger smoothing effects. Conversely, a lower diffusion coefficient restricts the spread of information, preserving edges and fine details in the image. This method follows prior work (Perona and Malik, 1990), which defines

$$c(g^p, g^n) = \frac{K^2}{K^2 + \| g^p - g^n \|_2^2} \qquad (2)$$

where K controls the sensitivity to the gradients in G.

We adapt the work of Metzger et al. (2023) by removing the adjustment step. The adjustment step was utilized by the authors to constrain the output of the diffusion to always match the source image when downsampled. This is done to preserve adherence between the input and output. However, in our approach, we opt to forego the adjustment step. This decision is motivated by the recognition that low-resolution DSMs often contain minimal information, and their use in the adjustment step may hinder the diffusion process.

To this end, our proposed network is trained in an end-to-end manner, and the final training loss is the summation of losses of two sub-networks.

4. EXPERIMENTS

4.1 Datasets and Implementation

Imagery and Study Area   In this study, we generate a dataset based on real-world DSMs. We evaluate our method on DSMs acquired over two main cantons of Switzerland: Zurich and Bern. We use low-resolution DSMs with a ground sampling distance (GSD) of 5 m generated from Cartosat-1 stereoscopic satellite as input instead of conventional bicubic-downsampled. We take high-resolution DSMs and their corresponding RGB orthoimage (GSD $= 0.5$ m and 0.1 m) provided by The Federal Office of Topography on the Swisstopo Portal[1]. We downsample the orthoimages to match the resolution of the DSMs using bicubic interpolation. The dataset consists of 2200 patches of size $(256, 256)$. We use 2000 samples for training and 200 for testing. The study area includes widely spaced, detached residential buildings, allotments, and high commercial buildings. We accomplish this by utilizing the Esri global land cover map (Karra et al., 2021) to filter the dataset, specifically focusing on the *Built Area* class.

Implementation Details   We randomly load training patches during training. At inference time, we reconstruct large-scale scenes by applying the learned model in a sliding window. We follow best practices to normalize the data. Every DSM is normalized such that all height points averaged to 0 using its local mean and global standard deviation computed from all training samples. Optical images are normalized with the mean and standard deviation over the intensity values of all training pixels. In all experiments, we use a hidden feature dimension of 64 for the feature extractor and the refinement decoder. ResNet-50 (He et al., 2016) backbone pretrained on ImageNet (Deng et al., 2009) is used as feature extractor. Before extracting the features, low-resolution DSMs are upsampled to the resolution of the ground truth DSMs using bicubic interpolation and concatenated with the guide. For training, we train all methods, including our own, with the $\mathcal{L}_1$ loss. For the diffusion network, we adopt the same setup and strategy outlined in (Metzger et al., 2023), with K and λ set to 0.001 and 0.24. The number of diffusion steps with and without gradients in training phase, are set to 8000 and 1024, respectively. Additionaly in our refinement network and DSRGAN, perceptual loss is added. We employ the ADAM optimizer with a base learning rate of 5×10^{-5} and no weight decay. We set the batch size to 2 for training and 1 for testing. We stop training once the RMSE on the test set have converged. We implemented our model in PyTorch and run it on a NVIDIA Titan RTX GPU.

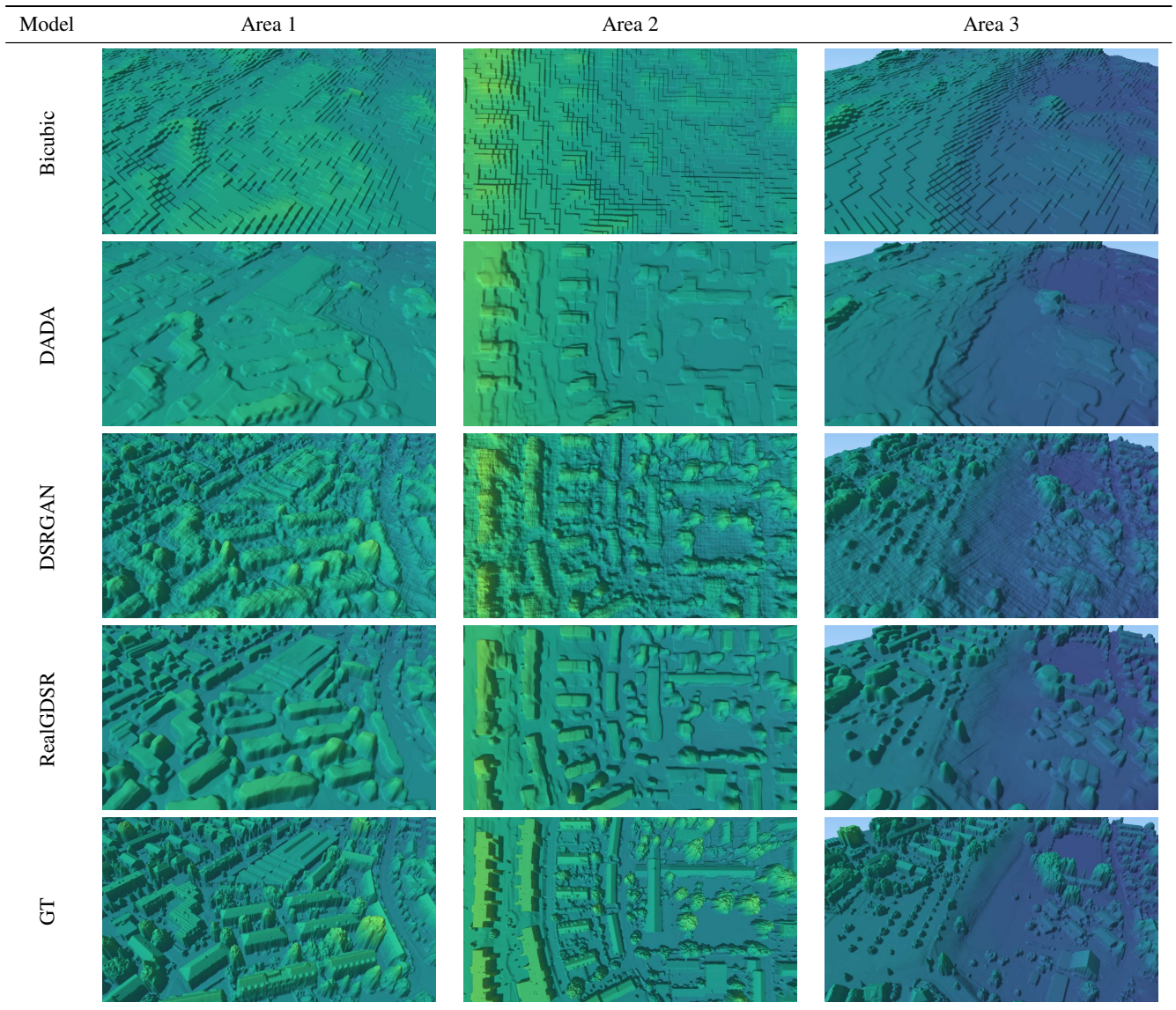[1] https://www.swisstopo.admin.ch/en/geodata.html

Figure 4: Visual comparison of RealGDSR with selected baselines. Our approach demonstrates its accuracy while producing regularized and smooth DSMs. All examples are taken from the test set.

4.2 Baselines

We compare Real-GDSR against the following baselines:

- Bicubic: The upsampled low-resolution DSMs using bicubic-interpolation.

- DADA: Deep Anisotropic Diffusion-Adjustment network is a hybrid framework for guided super-resolution that combines deep feature learning and anisotropic diffusion. The approach achieved edge-enhancing properties from the diffusion boosted by the contextual reasoning capabilities of large pre-trained models and a strict adjustment step guarantees perfect adherence to the source image. Our diffusion network is similar to this approach without the adjustment step.

- D-SRGAN: DEM Super-Resolution with Generative Adversarial Networks implemented ESRGAN for single DSM super-resolution. To compare it with our models we modify the input of the model by concatenating the low-resolution DSM with extracted features from the optical images using the same feature extractor as our model. We make no changes to the GAN.

4.3 Evaluation Metrics

We evaluate the models' performance by examining the root mean square error (RMSE), the normalized median absolute deviation (NMAD), and the median absolute error (MedAE), which are all derived from per-pixel differences between prediction and ground truth.

5. RESULTS

5.1 Comparisons with Prior Works

We start by assessing the performance of Real-GDSR, so as to quantify the impact of our framework. The high-resolution DSM generated by bicubic interpolation serves as the baseline. Due to the limited transformation, the bicubic-upsampled DSMs show grid-like patterns and lack of structural features (see Fig. 4, 1st row), resulting in 5.6 m RMSE, 3.6 m NMAD and 2.3 m MedAE. Applying our model improves the reconstruction significantly. The RMSE is lowered to 3.6 m, a 36% improvement compared to the bicubic-upsampled. Similarly, the NMAD and MedAE are also
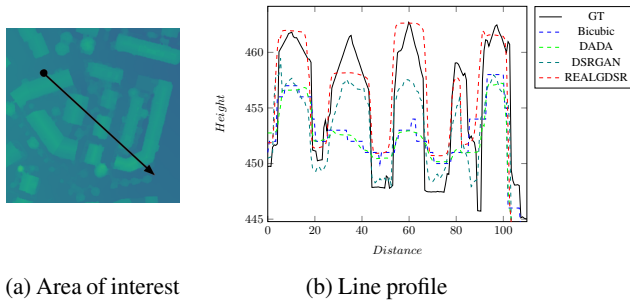
(a) Area of interest (b) Line profile

Figure 5: Line profile analysis of RealGDSR and other baselines.

reduced to $2.6\,\text{m}$ and $1.5\,\text{m}$. Beside the quantitative improvement, visually we can see that the reconstructed 3D geometry is clearly recovered. Buildings have sharp lines, and there are fewer visible artifacts and bumps on the terrain. The most notable finding is the recovery of detailed building structures like a cluster of buildings on Fig. 4, 4th row, 1st column. Furthermore, Real-GDSR reconstructs realistic scenes even in the presence of previously unseen building shapes and arrangement of buildings. The reconstruction of height and arrangement of buildings in urban areas are more accurate, where most of the baselines failed (see Fig. 5). In such cases, the model predict the height using information extracted from the optical image.

Model	RMSE	NMAD	MedAE
Bicubic	5.59	3.62	2.30
DADA	5.55	3.40	2.17
D-SRGAN	4.1	**2.06**	3.08
Real-GDSR	**3.54**	2.65	**1.46**

Table 1: Performance comparison of Real-GDSR and baselines in meters (m). Our model outperforms other methods.

Among all methods Real-GDSR yields the lowest reconstruction errors accross most metrics (Table 1). Compared to other networks it reconstructs more accurate heights and smoother surfaces (Fig. 4, 4th row) where other networks failed (Fig. 4, 2nd row) and produce noises on the terrain and building shapes (Fig. 4, 3rd row). While having higher NMAD, Real-GDSR achieves 14% lower RMSE and 53% lower MedAE compared to the second best model DSRGAN.

5.2 Ablation Study

	Refinement	Diffusion	RMSE	NMAD	MedAE
1	✓		4.54	**1.81**	**1.42**
2		✓	5.53	3.45	1.95
3	✓	✓	**3.54**	2.65	1.46

Table 2: Contribution of each component in our network

Local Residual Refinement Network   One of the main contributions of this work is our local residual refinement step of bicubic-interpolated sample which focusing on local features. To evaluate this we retrained our model removing the diffusion step leaving only refinement network and compared it (Table 2, 1st row) with the modified DSRGAN (Table 1, 3rd row). Although having worse RMSE with $0.4\,\text{m}$ difference, our network has greater variability. This is evidenced by its lower NMAD and MedAE values compared to DSRGAN, emphasizing its ability to handle outliers more effectively.

Refinement and Diffusion Networks   We conduct an evaluation on the contribution of each component in Real-GDSR . Our refinement network outperforms other baselines in NMAD and MedAE with $1.8\,\text{m}$ and $1.4\,\text{m}$, respectively (Table 1, 1st row). Our diffusion network performed better than DADA because of the adjustment-step removal. Adding diffusion component the RMSE is lower but increasing both NMAD and MedAE (Table 1, 3rd row). It indicates that the diffusion network, while effectively regularizing the model's structure and filtering out outliers, may inadvertently remove some of the intricate details captured by the refinement network.

6. CONCLUSION

We have presented Real-GDSR, a practical yet effective approach to guided super-resolution of DSMs. The approach is trying to solve the super-resolution problem in two steps: local refinement and an edge-enhancing diffusion. We highlight that the combination of CNN-based network and diffusion process bring the best of both worlds. Moreover, the local refinement network follows a residual learning strategy, i.e., it is trained to refine an imperfect bicubic-upsampled digital surface model (DSM) by predicting correction to the height, using both the DSMs and optical images as input. Together with the diffusion step, it can leverage information from optical images not only for the local distribution of the heights but also globally preserving the edges effectively. Our approach learns to restore substantial geometry such as sharp building lines and smooth height discontinuities. It also successfully restore missing shape details in the low-resolution DSM with information from the optical images.

In our experiments, Real-GDSR reaches top performance and outperforms state-of-the-art networks, including GANs and hybrid methods. We also found that Real-GDSR is fairly robust in terms of generalization that it can generate accurate high-resolution DSMs of unseen test samples. However, we acknowledge the trade-off between bias and variance in our approach and believe that more diverse training data can be gathered for a more extensive and representative analysis on different section of cities. Hence, our approach can be applied universally across various regions. On a conceptual level, we hope that our work motivates further research on DSM super-resolution.

References

Aktaruzzaman, M. and Schmitt, T. G., 2009. Detailed digital surface model (dsm) generation and automatic object detection to facilitate modelling of urban flooding. ISPRS J. Photogramm. Remote Sens.

Bhat, S. F., Alhashim, I. and Wonka, P., 2021. Adabins: Depth estimation using adaptive bins. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4009–4018.

Bittner, K., d'Angelo, P., Körner, M. and Reinartz, P., 2018. Dsm-to-lod2: Spaceborne stereo digital surface model refinement. Remote Sensing 10(12), pp. 1926.

Bittner, K., Liebel, L., Körner, M. and Reinartz, P., 2020. Long-short skip connections in deep neural networks for dsm refinement. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 43(B2), pp. 383–390.

Cai, J., Zeng, H., Yong, H., Cao, Z. and Zhang, L., 2019. Toward real-world single image super-resolution: A new benchmark and a new model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3086–3095.

Chen, Z., Wang, X., Xu, Z. and Hou, W., 2016. Convolutional neural network based dem super resolution. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 41, pp. 247–250.

Demiray, B. Z., Sit, M. and Demir, I., 2021a. D-srgan: Dem super-resolution with generative adversarial networks. SN Computer Science 2, pp. 1–11.

Demiray, B. Z., Sit, M. and Demir, I., 2021b. Dem super-resolution with efficientnetv2.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee, pp. 248–255.

Diebel, J. and Thrun, S., 2005. An application of markov random fields to range sensing. In: Y. Weiss, B. Schölkopf and J. Platt (eds), Advances in Neural Information Processing Systems, Vol. 18, MIT Press.

Dong, C., Loy, C. C., He, K. and Tang, X., 2015. Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence 38(2), pp. 295–307.

Godard, C., Aodha, O. M., Firman, M. and Brostow, G. J., 2019. Digging into self-supervised monocular depth estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3828–3838.

Godard, C., Mac, O., Gabriel, A. and Brostow, J., 2017. Unsupervised monocular depth estimation with left-right consistency. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 270–279.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. In: Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence and K. Weinberger (eds), Advances in Neural Information Processing Systems, Vol. 27, Curran Associates, Inc.

Gui, S. and Qin, R., 2021. Automated lod-2 model reconstruction from very-high-resolution satellite-derived digital surface model and orthophoto. ISPRS Journal of Photogrammetry and Remote Sensing 181, pp. 1–19.

Ham, B., Cho, M. and Ponce, J., 2017. Robust guided image filtering using nonconvex potentials. IEEE transactions on pattern analysis and machine intelligence 40, pp. 192–207.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

He, L., Zhu, H., Li, F., Bai, H., Cong, R., Zhang, C., Lin, C., Liu, M. and Zhao, Y., 2021. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp. 9225–9234.

Hui, T.-W., Loy, C. C. and Tang, X., 2016a. Depth map super-resolution by deep multi-scale guidance. pp. 353–369.

Hui, T.-W., Loy, C. C. and Tang, X., 2016b. Depth map super-resolution by deep multi-scale guidance. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14, Springer, pp. 353–369.

Izraelevitz, D., 1994. Model-based multispectral sharpening. Vol. 2231, pp. 60–71.

Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J. C., Mathis, M. and Brumby, S. P., 2021. Global land use/land cover with sentinel 2 and deep learning. In: 2021 IEEE international geoscience and remote sensing symposium IGARSS, IEEE, pp. 4704–4707.

Kiechle, M., Hawe, S. and Kleinsteuber, M., 2013. A joint intensity and depth co-sparse analysis model for depth map super-resolution. In: Proceedings of the IEEE international conference on computer vision, pp. 1545–1552.

Kim, B., Ponce, J. and Ham, B., 2021. Deformable kernel networks for joint image filtering. International Journal of Computer Vision 129(2), pp. 579–600.

Kim, J., Lee, J. K. and Lee, K. M., 2016a. Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1646–1654.

Kim, J., Lee, J. K. and Lee, K. M., 2016b. Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1637–1645.

Kopf, J., Cohen, M. F., Lischinski, D. and Uyttendaele, M., 2007. Joint bilateral upsampling. ACM Trans. Graph. 26(3), pp. 96–es.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4681–4690.

Lim, B., Son, S., Kim, H., Nah, S. and Mu Lee, K., 2017. Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 136–144.

Liu, J. and Gong, X., 2013. Guided depth enhancement via anisotropic diffusion. In: Advances in Multimedia Information Processing–PCM 2013: 14th Pacific-Rim Conference on Multimedia, Nanjing, China, December 13-16, 2013. Proceedings 14, Springer, pp. 408–417.

Lutio, R. D., Becker, A., D'Aronco, S., Russo, S., Wegner, J. D. and Schindler, K., 2022. Learning graph regularisation for guided super-resolution. pp. 1979–1988.

Metzger, N., Daudt, R. C. and Schindler, K., 2023. Guided depth super-resolution by deep anisotropic diffusion. pp. 18237–18246.

Partovi, T., Fraundorfer, F., Bahmanyar, R., Huang, H. and Reinartz, P., 2019. Automatic 3-d building model reconstruction from very high resolution stereo satellite imagery. Remote Sensing 11(14), pp. 1660.

Patterson, T. J., Bullock, M. E. and Wada, A. D., 1992. Multispectral band sharpening using pseudoinverse estimation and fuzzy reasoning. Vol. 1693, pp. 170–181.

Perona, P. and Malik, J., 1990. Scale-space and edge detection using anisotropic diffusion. IEEE Transactions on pattern analysis and machine intelligence 12(7), pp. 629–639.

Riegler, G., Ferstl, D., Rüther, M. and Bischof, H., 2016a. A deep primal-dual network for guided depth super-resolution. In: British Machine Vision Conference, The British Machine Vision Association.

Riegler, G., Rüther, M. and Bischof, H., 2016b. Atgv-net: Accurate depth super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14, Springer, pp. 268–284.

Stucker, C. and Schindler, K., 2022. Resdepth: A deep residual prior for 3d reconstruction from high-resolution satellite images. ISPRS Journal of Photogrammetry and Remote Sensing 183, pp. 560–580.

Stucker, C., Ke, B., Yue, Y., Huang, S., Armeni, I. and Schindler, K., 2022. Implicity: City modeling from satellite images with deep implicit occupancy fields. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 2, pp. 193–201.

Tan, M. and Le, Q., 2021. Efficientnetv2: Smaller models and faster training. In: International conference on machine learning, PMLR, pp. 10096–10106.

Tosic, I. and Drewes, S., 2014. Learning joint intensity-depth sparse representations. IEEE Transactions on Image Processing 23(5), pp. 2122–2132.

Voynov, O., Artemov, A., Egiazarian, V., Notchenko, A., Bobrovskikh, G., Burnaev, E. and Zorin, D., 2018. Perceptual deep depth super-resolution. Proceedings of the IEEE International Conference on Computer Vision 2019-October, pp. 5652–5662.

Wang, X., Xie, L., Dong, C. and Shan, Y., 2021a. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 1905–1914.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y. and Loy, C. C., 2019. Esrgan: Enhanced super-resolution generative adversarial networks. In: Computer Vision–ECCV 2018 Workshops: Munich, Germany, September 8-14, 2018, Proceedings, Part V 15, Springer, pp. 63–79.

Wang, Y., Zorzi, S. and Bittner, K., 2021b. Machine-learned 3d building vectorization from satellite imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1072–1081.

Wang, Z., Liu, D., Yang, J., Han, W. and Huang, T., 2015. Deep networks for image super-resolution with sparse prior. In: Proceedings of the IEEE international conference on computer vision, pp. 370–378.

Xie, J., Feris, R. S. and Sun, M.-T., 2015. Edge-guided single depth image super resolution. IEEE Transactions on Image Processing 25(1), pp. 428–438.

Xu, Z., Chen, Z., Yi, W., Gui, Q., Hou, W. and Ding, M., 2019. Deep gradient prior network for dem super-resolution: Transfer learning from image to dem. ISPRS Journal of Photogrammetry and Remote Sensing 150, pp. 80–90.

Xu, Z., Wang, X., Chen, Z., Xiong, D., Ding, M. and Hou, W., 2015a. Nonlocal similarity based dem super resolution. ISPRS Journal of Photogrammetry and Remote Sensing 110, pp. 48–54.

Xu, Z., Wang, X., Chen, Z., Xiong, D., Ding, M. and Hou, W., 2015b. Nonlocal similarity based dem super resolution. ISPRS Journal of Photogrammetry and Remote Sensing 110, pp. 48–54.