Methodological Factors Affecting the Generalizability of Deep Learning-Based Digital Elevation Model Super-Resolution

Zahra Anvarian[1], Farhad Maleki[1]
[1]Department of Computer Science at University of Calgary
*Correspondence: mahsa.anvarian@ucalgary.ca; farhad.maleki1@ucalgary.ca

Abstract

High-resolution digital elevation models (DEMs) are important for terrain-dependent analysis in precision agriculture and remote sensing. Deep learning-based DEM super-resolution (SR) has shown promising results for generating higher-resolution DEMs from lower-resolution inputs; however, reported performance can be strongly affected by how training and test data are selected. This paper presents a methodological evaluation of DEM SR protocols, focusing on spatial leakage, test-region representativeness, and dataset-level terrain diversity, three key factors that can substantially limit the generalizability of DEM SR models. Using 30 m SRTM DEM data, we generate 120 m-to-30 m SR pairs and evaluate the TfaSR (Terrain feature-aware Super-Resolution) model under controlled experimental conditions to isolate the effect of each factor. First, randomly splitting tile-based data into training and test sets introduces spatial leakage that inflates apparent performance: the model achieved a DEM RMSE of 4.52 m on the internal random test set, while evaluation on a terrain-matched but geographically independent test region raised this to 5.02 m, revealing that random-split evaluation overestimates true generalization. Second, a low-relief test region produced substantially lower DEM RMSE than a terrain-representative test region (1.68 vs. 5.02 m), a gap that reflects the simpler terrain structure of flat areas rather than superior model generalization. Third, on a shared terrain-diverse test set, TfaSR trained on a terrain-diverse dataset achieved lower DEM RMSE than the official pretrained TfaSR, whose training data is geographically distributed but low-relief-dominated (5.50 vs. 5.69 m), indicating that terrain diversity in the training data, rather than broad geographic coverage alone, drives generalization to varied terrain. These results show that random tile-based splitting, unrepresentative test-region selection, and limited training-data terrain diversity can lead to misleading conclusions about DEM SR generalization. Credible evaluation should explicitly report spatial partitioning strategies, train-test terrain-feature similarity, and dataset-level terrain composition.

Keywords: Digital elevation models (DEMs), Super-resolution, Geospatial deep learning models, Spatial leakage, Terrain representativeness

1 Introduction

High-resolution digital elevation models (DEMs) are essential geospatial data products for terrain-dependent analysis in precision agriculture and remote sensing, supporting tasks such as hydrological modelling, irrigation and drainage planning, erosion risk assessment, and terrain characterization[1–3]. Many of these tasks depend on fine-scale elevation variation and local terrain structures that coarse-resolution DEMs may not adequately capture, yet global DEM products often lack the required spatial resolution[4,5]. Because acquiring high-resolution DEMs through LiDAR, photogrammetry, or field surveying is often expensive or impractical, there is increasing interest in DEM super-resolution (SR), which reconstructs a higher-resolution DEM from a lower-resolution input[6,7].

Recent deep learning-based DEM SR methods have reported promising results by adapting image super-resolution architectures, in-

cluding convolutional neural networks, residual networks, generative adversarial networks, and transformer-based models[8–14]. In these approaches, DEMs are commonly treated as single-channel raster data, and models are trained to predict high-resolution elevation grids from degraded low-resolution inputs. Performance is typically evaluated using reconstruction metrics such as root mean square error (RMSE), mean absolute error (MAE), or peak signal-to-noise ratio (PSNR). While these metrics are useful, they do not by themselves establish whether a model generalizes reliably to geographically independent terrain. For geospatial raster data, reported performance depends not only on model architecture and loss functions, but also on how training and test samples are selected.

A major concern is spatial leakage caused by random tile-based train/test splitting. Because DEM tiles are sampled from continuous, spatially autocorrelated terrain surfaces, randomly selected test tiles may lie close to training tiles and share highly similar topographic structure[15–17]. Test results obtained under such splits may therefore overstate how well a model generalizes to independent geographic regions, even though tile-based datasets and random partitioning remain common in DEM SR studies[12–14].

A second concern is the representativeness of the selected test region. Even a geographically separated test area can be artificially easy if it is dominated by low-relief terrain, or artificially difficult if it is substantially more rugged than the training data. Reliable evaluation therefore requires characterizing test regions in terms of terrain-feature distributions, such as elevation variation, slope, Terrain Ruggedness Index (TRI), and Topographic Position Index (TPI), rather than relying on geographic separation alone.

A related concern is the terrain diversity of the dataset as a whole. A dataset spanning multiple geographic areas may still represent a narrow terrain domain if most samples come from flat or low-relief landscapes, in which case strong results on a random split only support generalization within that restricted domain. Evaluating dataset-level terrain diversity therefore requires examining the distribution of terrain features across samples, rather than relying on geographic coverage as a proxy for topographic breadth.

This paper presents a methodological evaluation of deep learning-based DEM SR, focusing on evaluation validity rather than architectural novelty. While each of these concerns has been recognized in the broader geospatial machine learning literature, their quantified impact on DEM SR evaluation has not been systematically examined under controlled experimental settings. Using 30 m SRTM (Shuttle Radar Topography Mission) DEM data from Alberta and British Columbia, Canada, we construct 120 m-to-30 m SR pairs and evaluate the TfaSR model[12]. To isolate the effect of evaluation protocol, the model and training configuration are kept fixed across all experiments.

The main contributions of this paper are as follows:

1. We quantify the effect of random tile-based splitting in DEM SR by comparing internal random testing with geographically independent region-based testing under a fixed model configuration.

2. We show that region-based testing alone is insufficient unless the test region is terrain-representative of the training region, and we evaluate representativeness using terrain-feature distributions and similarity measures.

3. We demonstrate with controlled experiments that geographic coverage alone does not ensure terrain diversity: a training dataset collected from multiple regions may still be dominated by low-relief terrain, limiting generalization to more varied topographic conditions.

We argue that credible DEM SR evaluation should move beyond random tile splits and aggregate reconstruction metrics, explicitly reporting spatial partitioning strategies, train-test terrain-feature similarity, and dataset-level terrain di-

versity so that generalization and real-world applicability claims are better supported.

2 Related Work

Deep learning-based DEM SR has advanced by adapting architectures originally developed for image SR, including convolutional networks, residual networks, generative adversarial networks, and transformer-based models[8–11]. DEM-specific methods have further introduced terrain-aware components to better preserve topographic structure. D-SRGAN and D-SRCAGAN applied adversarial learning and channel attention, showing that learned models can outperform traditional interpolation by recovering nonlinear relationships between low- and high-resolution elevation grids[6,18]. TfaSR incorporated deformable convolution and terrain-related loss terms to improve the reconstruction of local terrain features such as ridges and valleys[12]. GISR introduced global-information constraints to improve consistency between local and broader elevation patterns[13], and transformer-based approaches such as DSRT and TTSR aimed to capture local–global dependencies and multiscale terrain heterogeneity[14,19]. These studies demonstrate important progress in DEM SR model design; however, most rely on random tile-based partitioning within a fixed geographic region and evaluate results using aggregate reconstruction metrics, without examining whether reported improvements support reliable generalization to independent terrain.

A central issue in geospatial model evaluation is spatial dependence between samples. Nearby locations often share similar terrain characteristics due to spatial autocorrelation[15], and prior work in remote sensing has shown that random sample partitioning can overestimate predictive performance when spatial dependence is ignored[16,17,20]. This concern applies directly to DEM SR, where tiled datasets are generated from continuous terrain surfaces and random tile-based splitting is the prevailing evaluation practice[12–14]. DEM SR has largely inherited image SR's evaluation conventions, in which random splits are standard and sample independence is assumed — an assumption that does not hold for spatially autocorrelated terrain data.

Beyond spatial separation, the terrain characteristics of the test data also affect the interpretation of DEM SR results. Terrain analysis literature provides established descriptors such as elevation variability, slope, TRI, and TPI for characterizing terrain complexity[1–3,21,22]; however, DEM SR studies typically report performance without explicitly characterizing train-test terrain similarity or dataset-level terrain diversity[6,12–14,18]. Together, these gaps suggest that DEM SR generalization remains understudied as an evaluation problem, independently of the architectural advances being reported.

3 Methodological Evaluation Framework

The evaluation of DEM SR differs from standard image SR because DEM samples are spatially referenced and represent continuous physical terrain surfaces. As a result, the validity of reported performance depends not only on the model and reconstruction metrics, but also on the spatial partitioning strategy, the terrain characteristics of the test region, and the terrain diversity of the dataset, which the following subsections examine in turn.

3.1 Random Tile-Based Splitting and Spatial Leakage

A common evaluation protocol in DEM SR divides tiled DEM samples into training, validation, and test subsets by random sampling, assuming that individual tiles are sufficiently independent. However, DEM tiles are extracted from continuous terrain surfaces, where nearby samples may share similar elevation ranges, slope patterns, ridges, valleys, drainage structures, and surface roughness; randomly selected test tiles may therefore be geographically close to training tiles and contain highly similar terrain patterns.

Figure 1 illustrates this issue using the DEM SR dataset released with TfaSR[12], which is publicly available through Figshare (DOI: `https://doi.org/10.6084/m9.figshare.19225374`). Although the dataset includes multiple geographic regions, random partitioning is applied at the tile level within the sampled areas. As shown

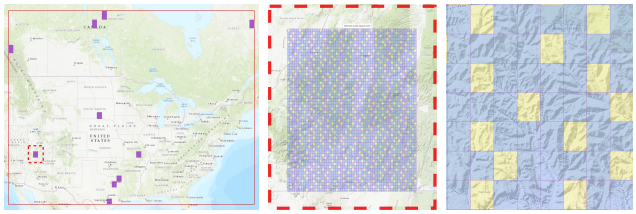

Figure 1: Random tile-based splitting in the publicly released TfaSR DEM SR dataset[12]. Yellow tiles are test samples; blue tiles are training samples. The zoomed panel shows test tiles directly adjacent to training tiles.

in the zoomed panel, some test tiles are directly adjacent to training tiles from the same continuous terrain surface. These test tiles are not duplicate samples, but their spatial proximity to training tiles can introduce leakage because neighbouring DEM patches often share similar topographic structures.

Throughout this paper, spatial leakage refers to an evaluation condition in which test samples are geographically close to, or strongly correlated with, samples used during training. Under this condition, test performance may reflect interpolation within a familiar terrain distribution rather than generalization to an independent region. Random tile-based splitting is not necessarily invalid when the goal is to estimate performance within the same mapped area, but it is insufficient for supporting claims about geographic generalization. To examine this issue, our first experiment compares internal random testing with geographically independent region-based testing under an otherwise fixed training and model configuration.

3.2 Test-Region Representativeness

Region-based testing reduces direct spatial leakage, but geographic separation alone does not guarantee a meaningful evaluation. The difficulty of DEM SR depends strongly on the terrain characteristics of the test region. Low-relief areas often contain limited fine-scale elevation variation and may produce low reconstruction errors, whereas rugged or mountainous areas contain sharper slope changes, ridges, valleys, and local relief that are harder to reconstruct.

We define test-region representativeness as the degree to which a geographically independent test region reflects the terrain-feature distribution of the training region for the intended evaluation objective. This does not require the test region to be identical to the training region, but it should be comparable enough that the evaluation is not artificially easy, artificially difficult, or unrelated to the claimed application domain. We assess representativeness using terrain-feature distributions based on elevation variation, slope, TRI, and TPI. Our second experiment compares a low-relief external test region with a terrain-representative external test region to quantify how test-region selection affects reported DEM SR accuracy.

3.3 Dataset-Level Terrain Diversity

Train-test similarity should also be interpreted in relation to the terrain diversity of the dataset as a whole. A dataset may include multiple geographic areas but still represent a narrow terrain domain if most samples are concentrated in flat or low-relief landscapes. In such cases, a random split may produce training and test sets with similar feature distributions, but the resulting evaluation only supports generalization within that limited terrain domain.

We define dataset-level terrain diversity as the range of terrain conditions represented across the full evaluation dataset. This factor is important because geographic coverage does not necessarily imply topographic diversity. To support broad generalization claims, a DEM SR dataset should contain a sufficient range of terrain complexity, or the scope of the claims should be restricted to the terrain domain represented by the data. Our third experiment examines this factor by comparing a model trained on a geographically distributed dataset with one trained on a terrain-diverse dataset of equal size, evaluated on a shared, geographically independent test set, to test whether geographic coverage alone is sufficient for broad generalization.

4 Terrain-Feature-Based Analysis

To evaluate whether DEM SR test results support meaningful generalization claims, we char-

acterize the terrain conditions of the training and test data at the tile level, using descriptors computed from the reference high-resolution DEM. These descriptors serve both to assess test-region representativeness and to gauge dataset-level terrain diversity.

4.1 Terrain Descriptors

Each DEM tile is represented using a set of terrain descriptors that summarize both elevation variation and local terrain structure. Elevation-based descriptors include mean elevation, standard deviation, elevation range, interquartile range, 5th–95th percentile range, and median absolute deviation. These measures describe the vertical variation within each tile and help distinguish low-relief tiles from tiles with stronger elevation variability. However, elevation statistics alone do not fully characterize terrain complexity because two tiles may have similar elevation ranges but different local surface forms.

To capture local terrain structure, we compute slope, TRI, and TPI[2,21,22]. Slope describes the local rate of elevation change, TRI measures local ruggedness by comparing a cell with its surrounding neighbourhood, and TPI measures whether a cell is locally higher or lower than its neighbouring cells. For a cell with elevation z_c and eight neighbouring elevations z_i, TRI and TPI are defined as:

$$TRI = \sqrt{\sum_{i=1}^{8}(z_c - z_i)^2}, \quad TPI = z_c - \frac{1}{8}\sum_{i=1}^{8} z_i. \quad (1)$$

For each tile, the derived slope, TRI, and TPI layers are summarized using mean, standard deviation, and maximum values. Together, these per-tile elevation and slope/TRI/TPI summaries form the terrain-feature set on which all similarity measures reported in this paper are computed.

4.2 Train-Test Similarity Measures

After computing terrain descriptors for all tiles, we compare the feature distribution of each candidate set against a reference set chosen per experiment (the internal random test set in Experiment 1 and the training set in Experiment 2). A standard scaler is fitted on the reference set only and applied to the compared set, so comparisons are made in a common feature space without using the compared set's information during normalization.

We use complementary similarity measures rather than relying on a single criterion. The mean standardized feature shift measures the difference between the reference and compared feature means after scaling. The Wasserstein distance and Kolmogorov–Smirnov (KS) statistic compare the two distributions feature by feature, where lower values indicate stronger distributional similarity[23,24].

Finally, we use classifier-based separability as a multivariate similarity measure[25]. A logistic regression classifier is trained to distinguish training tiles from test tiles using only terrain descriptors. The resulting area under the receiver operating characteristic curve (AUC) is converted into a distance score:

$$d_{AUC} = 2|AUC - 0.5|. \quad (2)$$

A value near zero indicates that the classifier cannot reliably distinguish training and test tiles, suggesting high terrain-feature similarity. A value near one indicates that the two sets are easily separable and therefore less similar in the selected terrain-feature space.

These measures are interpreted jointly: a representative test region should show low mean feature shift, low distributional distance, and low classifier separability. The same descriptors also support dataset-level analysis of whether the full dataset spans a broad range of terrain conditions or is concentrated within a narrow domain.

5 Experiments and Results

5.1 Experimental Setup

All model-based experiments use 30 m SRTM DEM data and follow a 120 m-to-30 m DEM SR setting. High-resolution (HR) DEM tiles are extracted as non-overlapping 64×64 patches. Low-resolution (LR) inputs are generated by downsampling the HR tiles by a factor of four using nearest-neighbour interpolation. This degradation process ensures spatial alignment between LR inputs and HR reference DEMs.

We use TfaSR as the reference DEM SR model because it is a terrain-aware deep learning architecture designed specifically for DEM SR. To isolate the effect of evaluation protocol and dataset composition, all experiments use the same model architecture, loss function, normalization strategy, and training hyperparameters. The model is trained using the combined content mean squared error (MSE) and slope MSE loss, Adam optimizer, learning rate of 10^{-4}, and batch size of 32. In all experiments, models are trained for 100 epochs, and the final checkpoint is used for evaluation to avoid test-set-based model selection.

Performance is reported using DEM RMSE, slope RMSE, and aspect RMSE. DEM RMSE measures elevation reconstruction error directly, while slope and aspect RMSE provide additional information about errors in terrain derivatives. Experiment 3 additionally reports TRI and TPI RMSE, which measure reconstruction error in the TRI and TPI derivative layers. Aspect RMSE is interpreted with caution in low-relief regions because aspect can become unstable when local slope values are close to zero.

5.2 Experiment 1: Random Tile-Based Testing vs. Region-Based Testing

The source study area is located in Alberta and British Columbia, Canada, and is defined by the bounding box $[-118.6414, 49.5252, -115.0269, 51.9172]$.[1] From this region, 26,954 HR DEM tiles were generated. These tiles were randomly divided into 21,563 training tiles and 5,391 internal test tiles using an 80/20 random split.

As a more appropriate evaluation for this task, we also extracted a separate test region defined by the bounding box $[-120.1849, 52.1301, -117.7734, 53.0521]$. This external region was not used during training and is geographically separated from the source training area. It was subsampled to match the internal random test set in both sample count (5,391 tiles) and terrain-feature distribution (verified in Figure 2 and Table 1), so that performance differences reflect spatial separation rather than terrain difficulty. The same trained model was evaluated on both the internal random test set and the external region-based test set.

Figure 2 compares the terrain-feature distributions of the training set, the internal random test set, and the external region-based test set using elevation range, mean TRI, mean TPI, and mean slope. The internal random test set closely follows the training distribution, as expected from random sampling within the same source region. Importantly, the external region-based test set also shows substantial overlap with the training distribution across these descriptors, while remaining geographically independent. This makes it suitable for evaluating whether performance changes when the test data are spatially separated rather than randomly sampled from the same continuous region.

Table 1 complements this visual comparison with quantitative terrain-feature similarity measures computed relative to the internal random test set, for the external region-based test set, with the training set shown as context. These measures determine whether the external region provides terrain comparable to the internal random test set while remaining geographically independent.

Table 2 summarizes the DEM SR performance under the two evaluation settings. The internal random test set produces lower DEM RMSE than the external region-based test set. Since the training configuration and model are fixed, this difference suggests that random tile-based evaluation can provide a more optimistic estimate of performance than testing on a geographically independent region.

Because Table 1 confirms that the external region is terrain-matched to the internal random test set, this performance gap is associated mainly with spatial independence rather than terrain difference.

[1] Bounding boxes are reported in decimal degrees as [west, south, east, north], that is, minimum longitude, minimum latitude, maximum longitude, and maximum latitude.

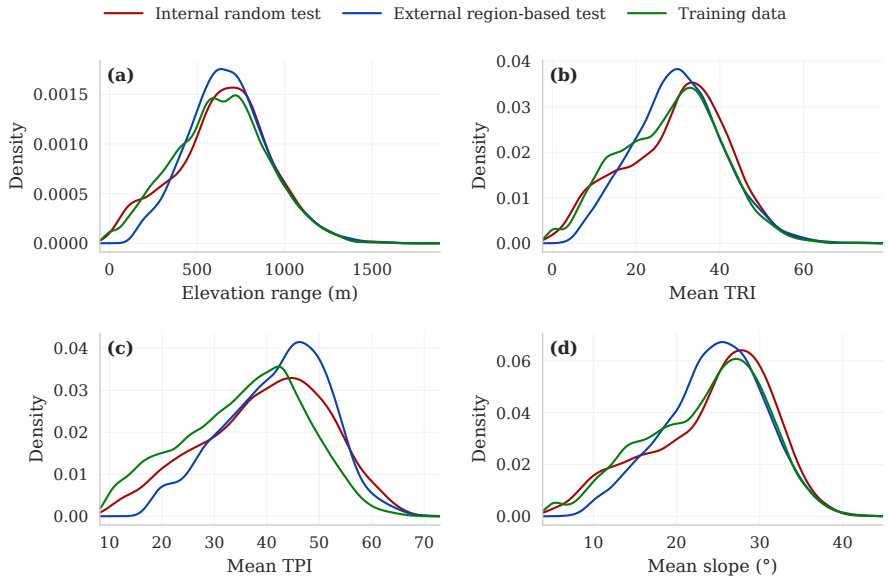

Figure 2: Terrain-feature distributions for Experiment 1: training set, internal random test set, and external region-based test set.

Table 1: Terrain-feature similarity in Experiment 1, relative to the internal random test set (training set shown as context). Lower values indicate stronger similarity.

Compared set	Mean shift	Wasserstein	KS	d_{AUC}
External region-based test	0.088	0.117	0.063	0.218
Training data (context)	0.221	0.186	0.092	0.255

5.3 Experiment 2: Low-Relief vs. Terrain-Representative Test Region

We compare two geographically independent test regions. The first is a low-relief test region defined by the bounding box $[-118.0041, 53.3780, -115.9058, 54.4125]$, which contains limited elevation variation and smoother terrain. The second is the external region-based test set from Experiment 1, which is selected to better match the terrain-feature distribution of the training region.

Both test regions are evaluated using the same trained model from Experiment 1, isolating the effect of test-region terrain complexity. Figure 3 compares the terrain-feature distributions of the training set and the two candidate test regions using elevation range, mean TRI, mean TPI, and mean slope. The low-relief test region is concentrated at substantially lower values across all four descriptors, whereas the terrain-representative test region closely overlaps the training distribution. Table 3 reports the corresponding terrain-feature similarity between the training set and each candidate test region.

Table 4 shows that the low-relief test region produces substantially lower DEM RMSE and slope RMSE than the terrain-representative test region. This result indicates that test-region selection can strongly affect reported accuracy. A low-relief test region may produce favourable error values because the reconstruction task is less complex, not necessarily because the model generalizes robustly across more complex terrain.

The high aspect RMSE in the low-relief region should be interpreted carefully. In flat or nearly flat areas, aspect is often unstable because small elevation differences can cause large angular changes in the estimated aspect direction. Therefore, the low DEM RMSE and slope RMSE are more informative for assessing the relative ease of the low-relief test region.

Table 2: DEM SR performance: internal random vs. external region-based testing.

Evaluation setting	DEM RMSE (m)	Slope RMSE (°)	Aspect RMSE (°)
Internal random test	4.52	2.99	50.51
External region-based test	5.02	3.29	50.34

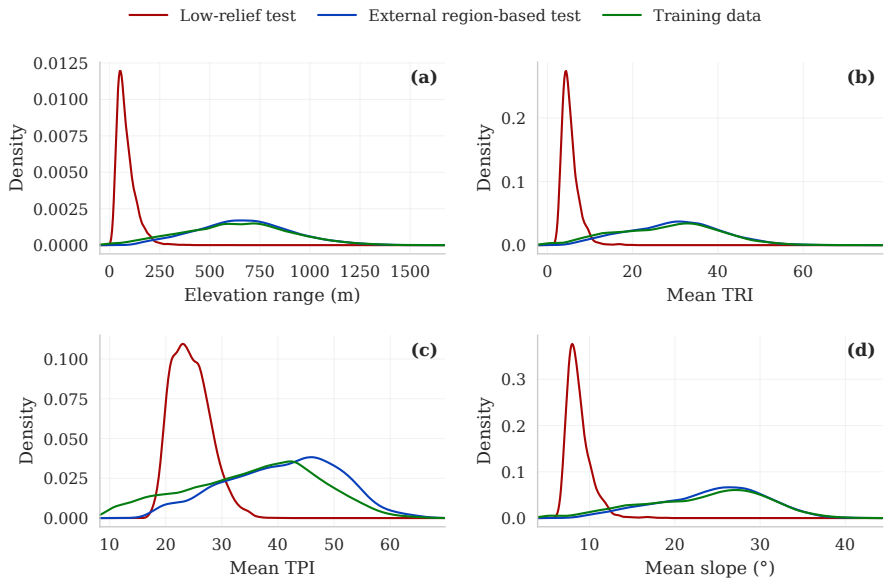

Figure 3: Terrain-feature distributions for Experiment 2: training set, low-relief test region, and terrain-representative test region.

5.4 Experiment 3: Effect of Dataset-Level Terrain Diversity on Generalization

We compare the official pretrained TfaSR model, trained on a geographically distributed but low-relief-dominated dataset, with a TfaSR model trained on a terrain-diverse dataset of equal size.

The terrain-diverse training dataset spans four geographically separated regions covering mountainous, alpine, and moderate hilly terrain, and the test set spans three geographically independent regions that do not overlap with the training regions or with the original TfaSR dataset.[2]

To make the comparison consistent in terms of dataset size, we selected 25,088 training patches and 6,272 test patches, matching the split size of the TfaSR dataset. The terrain-diverse TfaSR model was trained using the same architecture and training configuration as the original TfaSR model, including the same degradation process, tile size, loss function, optimizer, learning rate, batch size, normalization strategy, and number of epochs; only the training dataset was changed. The official pretrained TfaSR model was then evaluated directly on the same terrain-diverse test set. Because the only difference between the two models is the terrain composition of their training data, the comparison isolates the effect of training-data terrain diversity.

Figure 4 compares the terrain-feature distributions of the original TfaSR dataset and the terrain-diverse dataset. The TfaSR dataset shows a strong concentration in low-relief terrain, despite being collected from multiple geographic areas. In contrast, the terrain-diverse dataset covers a wider range of elevation variation, slope, TRI, and TPI values, indicating broader terrain diversity. The terrain-diverse training and test distributions also show substantial overlap, suggesting that the test set is representative of the terrain conditions in the training data while re-

[2]Training bounding boxes: [−118.6414, 49.5252, −115.0269, 51.9172], [85.4901, 27.3000, 87.4786, 28.4304], [87.6544, 27.4400, 88.8739, 28.1117], and [−119.4653, 47.7098, −117.2481, 48.9874]. Test bounding boxes: [−120.1849, 52.1301, −118.20, 53.0521], [94.2435, 30.3515, 95.2213, 30.7586], and [85.0342, 49.1155, 87.9565, 50.9030].

Table 3: Terrain-feature similarity between the training set and each candidate test region in Experiment 2. Lower values indicate stronger similarity.

Test region	Mean shift	Wasserstein	KS	d_{AUC}
Low-relief test region	1.514	1.461	0.769	0.995
Terrain-representative test region	0.266	0.232	0.107	0.375

Table 4: DEM SR performance: low-relief vs. terrain-representative test region.

Test region	DEM RMSE (m)	Slope RMSE (°)	Aspect RMSE (°)
Low-relief test region	1.68	1.44	96.20
Terrain-representative test region	5.02	3.29	50.34

maining geographically independent.

Table 5 reports the performance of both models on the shared terrain-diverse test set. The model trained on the terrain-diverse dataset achieves lower error across all reported metrics. In terms of DEM RMSE, it improves from 5.69 m to 5.50 m compared with the official pretrained TfaSR model. Similar improvements are observed for TRI, TPI, slope, and aspect RMSE. Although the improvement is moderate in absolute terms, it is statistically significant and consistent: a one-sided Wilcoxon signed-rank test on the 6,272 paired test samples yields $p = 2.63 \times 10^{-222}$, with 4,336 samples showing improved DEM RMSE against 1,936 with higher error, and the gain extends across all reported terrain metrics, suggesting that training-data terrain diversity produces a real and broad improvement in transfer to terrain-diverse test regions.

These results support the third methodological claim: geographic diversity alone does not guarantee terrain diversity, and training on terrain-limited data may restrict generalization to more diverse topographic conditions. The comparison should not be interpreted as a critique of the TfaSR architecture itself, since both models use the same architecture and training configuration, differing only in their training data. Instead, it shows that the composition of the training dataset is an important factor in DEM SR generalization.

6 Discussion

DEM SR performance should be interpreted in relation to the evaluation protocol, not only the model architecture. Because nearby DEM tiles often share similar terrain structures, random tile-based testing can give more favourable results than geographically independent testing; random tile splits should therefore not be the only evidence used to claim geographic generalization.

Region-based testing must also be accompanied by terrain-feature analysis: a geographically independent test region can still be artificially easy if dominated by low-relief terrain, or artificially difficult if much more rugged than the training region. Reporting descriptors such as elevation variation, slope, TRI, and TPI clarifies whether reported errors reflect model generalization or test-region difficulty.

At the dataset level, geographic coverage should not be treated as equivalent to terrain diversity. A dataset collected from multiple areas may still cover a narrow terrain domain, which limits the scope of generalization claims. For this reason, DEM SR evaluation should report performance not only overall but also across terrain-complexity classes. This is especially important for applications where errors in rugged or high-relief terrain can affect downstream terrain derivatives and hydrological analyses.

This study has several limitations. The experiments use TfaSR as the reference model, so numerical results may differ for other architectures. The terrain descriptors used here provide a compact representation of terrain complexity but do not capture all possible topographic properties. In addition, the experiments focus on a 120 m-to-30 m SRTM-based SR setting; future work should examine other DEM products, scale factors, and geographic regions.

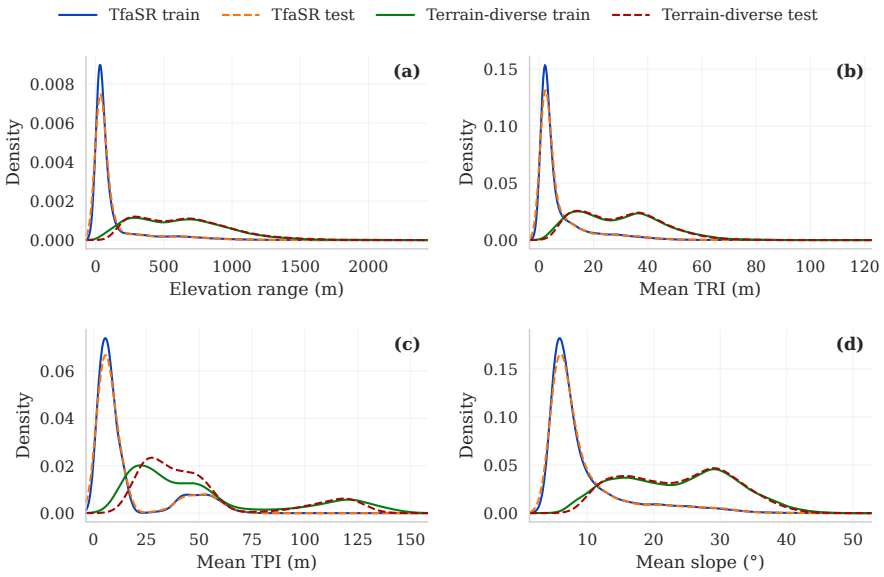

Figure 4: Terrain-feature distributions for Experiment 3. The original TfaSR dataset is concentrated mainly in low-relief terrain, while the terrain-diverse dataset covers a broader range of terrain conditions.

Table 5: Performance of the official pretrained TfaSR model and the TfaSR model trained on the terrain-diverse dataset, both evaluated on the same terrain-diverse test set.

Model	DEM RMSE	TRI RMSE	TPI RMSE	Slope RMSE	Aspect RMSE
Terrain-diverse TfaSR	5.50	7.50	2.96	3.37	57.75
Official pretrained TfaSR	5.69	7.71	3.03	3.48	58.48

7 Conclusion

This paper examined DEM SR evaluation from a methodological perspective. Using controlled experiments with a fixed TfaSR model, we showed that reported performance is affected by random tile-based splitting, test-region terrain representativeness, and dataset-level terrain diversity. These findings suggest that credible DEM SR evaluation should explicitly report spatial partitioning strategies, train-test terrain-feature similarity, dataset-level terrain composition, and class-wise performance across terrain types. Such practices can help ensure that reported improvements reflect meaningful generalization rather than favourable data partitioning or test-region selection.

References

[1] Ian D. Moore, Richard B. Grayson, and Anthony R. Ladson. Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. *Hydrological Processes*, 5(1):3–30, 1991.

[2] John P. Wilson and John C. Gallant, editors. *Terrain Analysis: Principles and Applications*. John Wiley & Sons, New York, NY, USA, 2000.

[3] Paolo Tarolli. High-resolution topography for understanding earth surface processes: Opportunities and challenges. *Geomorphology*, 216:295–312, 2014.

[4] Tom G. Farr, Paul A. Rosen, Edward Caro, Robert Crippen, Riley Duren, Scott Hensley, Mike Kobrick, Mimi Paller, Ernesto Rodriguez, Ladislav Roth, David Seal, Scott Shaffer, Joanne Shimada, Jeffrey Umland, Marian Werner, Michael Oskin, Douglas Burbank, and Douglas Alsdorf. The shuttle radar topography mission. *Reviews of Geophysics*, 45(2):RG2004, 2007.

[5] Google Earth Engine Data Catalog. USGS/SRTMGL1_003: NASA SRTM Digital Elevation 30m. `https://developers.google.com/earth-engine/datasets/catalog/USGS_SRTMGL1_003`, 2024. Accessed: 2026-05-25.

[6] Bekir Z Demiray, Muhammed Sit, and Ibrahim Demir. D-SRGAN: DEM super-resolution with generative adversarial networks. *SN Computer Science*, 2(1):48, 2021.

[7] Bekir Z Demiray, Muhammed Sit, and Ibrahim Demir. Dem super-resolution with efficientnetv2. *arXiv preprint arXiv:2109.09661*, 2021.

[8] Chao Dong, Chen Change Loy, Kaiming He, and

Xiaoou Tang. Image super-resolution using deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 184–199, 2016.

[9] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017.

[10] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017.

[11] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 1833–1844, 2021.

[12] Yifan Zhang, Wenhao Yu, and Di Zhu. Terrain feature-aware deep learning network for digital elevation model superresolution. *ISPRS Journal of Photogrammetry and Remote Sensing*, 189:143–162, 2022.

[13] Xiaoyi Han, Xiaochuan Ma, Houpu Li, and Zhanlong Chen. A global-information-constrained deep learning network for digital elevation model superresolution. *Remote Sensing*, 15(2):305, 2023.

[14] Yi Wang, Shichao Jin, Zekun Yang, Hongcan Guan, Yu Ren, Kai Cheng, Xiaoqian Zhao, Xiaoqiang Liu, Mengxi Chen, Yu Liu, and Qinghua Guo. TTSR: A transformer-based topography neural network for digital elevation model super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–19, 2024.

[15] Waldo R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(sup1):234–240, 1970.

[16] David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017.

[17] Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten F. Dormann, Gérard Cornu, Guillaume Viennois, Nicolas Bayol, Alexei Lyapustin, Sylvie Gourlet-Fleury, and Raphaël Pélissier. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11(1):4540, 2020.

[18] Xiaotong Deng, Weihua Hua, Xiuguo Liu, Siying Chen, Wen Zhang, and Jianchao Duan. D-SRCAGAN : DEM Super-resolution Generative Adversarial Network. *IEEE Geoscience and Remote Sensing Letters*, pages 1–1, 2022.

[19] Zhuoxiao Li, Xiaohui Zhu, Shanliang Yao, Yong Yue, Ángel F. García-Fernández, Eng Gee Lim, and Andrew Levers. A large scale digital elevation model super-resolution transformer. *International Journal of Applied Earth Observation and Geoinformation*, 124:103496, 2023.

[20] Haijun Feng et al. Information leakage in deep learning-based hyperspectral remote sensing image classification tasks. *Remote Sensing*, 15(15):3793, 2023.

[21] Shawn J. Riley, Stephen D. DeGloria, and Robert Elliot. A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences*, 5(1–4):23–27, 1999.

[22] Andrew D. Weiss. Topographic position and landforms analysis. In *Poster Presentation, ESRI User Conference*, San Diego, CA, USA, 2001.

[23] Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

[24] George Marsaglia, Wai Wan Tsang, and Jingbo Wang. Evaluating kolmogorov's distribution. *Journal of Statistical Software*, 8(18):1–4, 2003.

[25] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *Proceedings of the International Conference on Learning Representations*, 2017.