HIGH-RESOLUTION MULTI-SPECTRAL IMAGE GUIDED DEM SUPER-RESOLUTION USING SINKHORN REGULARIZED ADVERSARIAL NETWORK

Subhajit Paul, Ashutosh Gupta
Space Applications Centre
Indian Space Research Organization
Ahmedabad, GJ 380015, India
{subhajitpaul, ashutoshg}@sac.isro.gov.in

ABSTRACT

Digital Elevation Model (DEM) is an essential aspect in the remote sensing domain to analyze and explore different applications related to surface elevation information. In this study, we intend to address the generation of high-resolution (HR) DEMs guided by HR multi-spectral (MX) satellite imagery as prior. To promptly regulate this process, we utilize the discriminator activations as spatial attention for the MX prior, and also introduce a Densely connected Multi-Residual Block (DMRB) module to assist in efficient gradient flow. Further, we present the notion of using Sinkhorn distance with traditional GAN to improve the stability of adversarial learning. In this regard, we provide both theoretical and empirical substantiation of better performance in terms of vanishing gradient issues and numerical convergence. We demonstrate both qualitative and quantitative outcomes with available state-of-the-art methods. Based on our experiments on DEM datasets of Shuttle Radar Topographic Mission (SRTM) and Cartosat-1, we show that the proposed model performs preferably against other benchmark methods. We also generate and visualize several high-resolution DEMs covering terrains with diverse signatures to show the performance of our model.

1 INTRODUCTION

The Digital Elevation Model (DEM) is a digital representation of any three-dimensional surface. It is immensely useful in precision satellite data processing, geographic information systems (Trevisani et al., 2012), hydrological studies (Li & Wong, 2010), urban planning (Priestnall et al., 2000), and many other key applications. Due to its diverse applications, the accuracy and resolution of DEM have a substantial impact in different fields of operations (Sørensen & Seibert, 2007; Kim et al., 2019). The major sources of high-resolution (HR) elevation models are terrestrial and airborne systems with restricted coverage and they also typically suffer from several issues and systematic errors Fisher & Tate (2006); Liu (2008). Hence, accurate HR DEM products are expensive, as they require special acquisition and processing techniques. As an alternative, enhancing the resolution (super-resolution) of existing DEMs can be seen as the most optimal strategy to address the shortfall.

Research on DEM super-resolution (SR) is limited despite its significance in remote sensing applications. Generally, traditional methods like linear, and bicubic interpolation are widely used for DEM SR, but they tend to produce smoothed outputs at high-frequency regions (He & Siu, 2011). Reconstruction-based methods like steering kernel regression (SKR) (Takeda et al., 2007) or non-local means (NLM) (Protter et al., 2009), have been proposed to tackle this, however, they still under-perform at a large magnification factor. After the first introduction of SR using Convolutional Neural Network (SRCNN) (Dong et al., 2014), its variant D-SRCNN was proposed by Chen et al. (2016) to address the DEM super-resolution problem which attains performance gain over the traditional methods. Later, with the introduction of Genrative Adversarial Networks (GANs) (Goodfellow et al., 2016) and its variants in SR applications like SR using GANs (SRGAN) (Ledig et al., 2017), Demiray et al. (2020) proposed a DEM super-resolution model, namely D-SRGAN, and later they suggested another model based on EffecientNetV2 (Demiray et al., 2021) for DEM SISR.

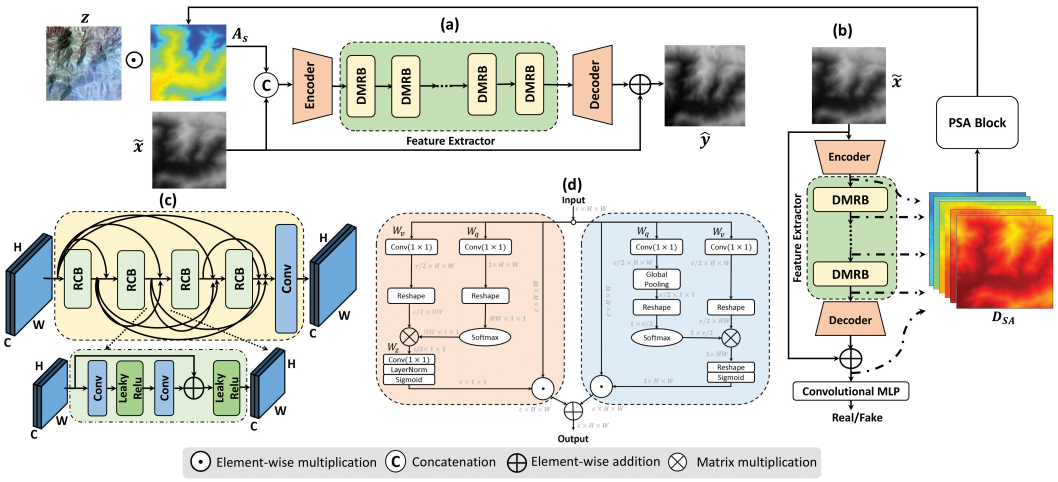

Figure 1: Overview of the proposed adversarial framework. (a) The generator $\mathbf{G}$ takes discriminative spatial attention from (b) discriminator $\mathbf{D}$ as conditional input, paased via a (d) Polarized Self-Attention (PSA) block. Both $\mathbf{G}$ and $\mathbf{D}$ constitute of (c) Densely connected Multi Residual Blocks (DMRBs) with residual convolution block (RCB) as the building block.

However, monocular depth SR using natural HR images as guide is an emerging research field in computer vision. Some of the pioneering works in this domain are Deformable Kernel Networks (DKN) and Faster DKN (FDKN) (Kim et al., 2021), Fast Depth map SR (FDSR) (He et al., 2021), and Deep Anisotropic Diffusion Adjustment (DADA) (Metzger et al., 2023). Inspired by these, we propose a DEM SR framework that effectively utilizes information from an HR multi-spectral (MX) image guide by conditioning it with a discriminative spatial self-attention. In this regard, we also propose a new adversarial learning framework, namely SIRAN (Sinkhorn Regularized Adversarial Network), as well as generate our own dataset by using realistic coarse resolution data instead of bicubic downsampled. In next section, we briefly discuss the methodology.

2 Methodology

In Figure 1, we have illustrated a detailed architectural overview of our framework. The generator $\mathbf{G}$ operates on upsampled coarser resolution DEM $\tilde{x}$, MX image prior z, consisting of false color composite (FCC) of NIR (R), R (G) and G (B) bands, and polarized self-attention (PSA) (Liu et al., 2021) of discriminator spatial feature maps, A_s as conditional input. Let $z \sim \mathbb{P}_Z$, where $z \in \mathbb{R}^{H \times W \times 3}$ with $\mathbb{P}_Z$ being the joint distribution of FCC composition and $\tilde{x} \sim \mathbb{P}_{\tilde{x}}$, where $\mathbb{P}_{\tilde{x}}$ constitute of upsampled coarser resolution DEM with $\tilde{x} \in \mathbb{R}^{H \times W}$. Let $\hat{y} \sim \mathbb{P}_{\mathbf{G}_\theta}$, where $\hat{y} = \mathbf{G}(\tilde{x}, z \odot A_s)$, where $\odot$ denotes the element-wise multiplication and $\mathbb{P}_{\mathbf{G}_\theta}$ denotes the generator distribution parameterized by $\theta \in \Theta$. Let $y \sim \mathbb{P}_y$ with $\mathbb{P}_y$ represent the target HR DEM distribution. The discriminator $\mathbf{D}$ classifies y and $\hat{y}$ to be coming from real or fake sample space and is assumed to be parameterized by $\psi \in \Psi$.

2.1 Network Architecture

We design both $\mathbf{G}$ and $\mathbf{D}$ models based on ResNet (He et al., 2015) and DenseNet (Huang et al., 2016). By combining the idea of skip and dense connections, we design a building block, namely a Densely connected Multi-Residual Block (DMRB) for our overall framework. Each DMRB block is constituted of multiple densely connected Residual Convolution Blocks (RCBs) as shown in Figure 1 (c). DMRB enables efficient context propagation and also stable gradient flow throughout the network. We kept the overall design of discriminator $\mathbf{D}$ with a similar configuration as $\mathbf{G}$ with an encoder followed by six DMRBs and finally a decoder to unravel the generated samples properly as shown in Figure 1 (a) and (b), respectively. The discriminator also adds a Multi-Layer Perceptron (MLP) layer to map its latent features into required shape. Another reason behind the design of $\mathbf{D}$ is to extract dense discriminative latent space features as they can be viewed as spatial attention to MX guide. Since, discriminators perform binary classification for a given input, apparently, in latent space, it captures the discriminative features that will help the generator focus on salient parts of the MX guide. Emami et al. (2019) introduced this concept of transferring domain-specific

latent knowledge of discriminator as spatial attention to the generator. Therefore, utilizing similar concept, $\mathbf{D}$ in our proposed framework has two functional branches which are classification and to approximate the spatial attention maps. We use upsampled coarse resolution DEM $\tilde{x}$ to estimate these attentions. This choice is motivated by the fact that unlike image-to-image translation proposed by Emami et al. (2019), during the test phase, we do not have high-resolution samples in the target domain. For this reason, we use the concept of domain adaptation loss (Rout et al., 2020). These attention maps are passed through a PSA (Liu et al., 2021) block to exclude redundant features while highlighting significant areas by extracting dense features in both channel and spatial dimension as shown in Figure 1 (d). The main reason behind choosing PSA is due of its capability to retain the high internal resolution compared to other self-attention modules. In the following subsection, we will briefly explain our objective formulation.

2.2 Formulation of objective function

Our whole framework is set-up based on adversarial learning. In this regard, WGAN and its variants (Arjovsky et al., 2017; Gulrajani et al., 2017) serve the purpose in most of the applications due to their prompt ability to resolve the problems of conventional GAN. However, as they are designed to solve the Kantarovich formulation of OT problems to minimise the Wasserstein distance, they suffer from curse of dimensionality due to their sample complexity of $\mathcal{O}(n^{-2/d})$ (Genevay et al., 2019), given a sample size n with a dimension d. Another key concern of utilizing these adversarial objectives is the vanishing gradient problem near the optimal point. This leads the generator to converge to a sub-optimal solution and results in a partially aligned generated distribution with respect to the true distribution. Therefore, we regularize the objective function of $\mathbf{G}$ with Sinkhorn loss (Genevay et al., 2018) as defined below.

$$\mathcal{S}_{C,\varepsilon} = \mathcal{W}_{C,\varepsilon}(\mu_\theta, \nu) - \frac{1}{2}\mathcal{W}_{C,\varepsilon}(\mu_\theta, \mu_\theta) - \frac{1}{2}\mathcal{W}_{C,\varepsilon}(\nu, \nu), \tag{1}$$

where, $\mu_\theta \in \mathbb{P}_{\mathbf{G}_\theta}$ and $\nu \in \mathbb{P}_y$ are measure of generated and true distribution with support included in a compact bounded set $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, respectively, and $\mathcal{W}_{C,\varepsilon}$ is Entropic OT (Aude et al., 2016a). Hence, it is not only computationally efficient due to its favorable sample complexity $\mathcal{O}(n^{-1/2})$ Genevay et al. (2019), equation 1 interpolates between OT loss and MMD loss as ε varies from 0 to ∞. Therefore, by properly tuning ε we can leverage the concurrent advantage of non-flat geometric properties of OT loss and, high dimensional rigidity and energy distance properties of MMD loss. Apart from this, the selection of ε also affects the smoothness of Sinkhorn loss (see proposed Theorem 1 in Appendix §A) which manipulates the overall gradients of $\mathbf{G}$, resulting in better prevention of vanishing gradient problems near the optimal region (see proposed Theorem 2 in Appendix §B), and also provides tighter iteration complexity (Rout, 2020) (see proposed Theorem 2 in Appendix §C) resulting in faster convergence compared to other GAN setups. Due to these advantages, we regularize the generator loss with Sinkhorn distance and refer it as $\mathscr{L}_{OT}$ that is estimated by Sinkhorn AutoDiff Algorithm (Genevay et al., 2018) utilizing ε and the Sinkhorn iterations T as the major parameters. As Sinkhorn loss also minimizes the Wasserstein distance, it serves the purpose of WGAN to resolve the issues of original GAN more effectively. Hence, we stick to classical GAN objective function ($\mathscr{L}_{ADV}$) for the generator to establish adversarial learning set-up while regularized with Sinkhorn loss. The generator loss also comprises of pixel loss ($\mathscr{L}_P$) and SSIM loss ($\mathscr{L}_{str}$). Therefore, the overall generator loss is defined as,

$$\lambda_P \mathscr{L}_P + \lambda_{str}\mathscr{L}_{str} + \lambda_{ADV}\mathscr{L}_{ADV} + \lambda_{OT}\mathscr{L}_{OT}, \tag{2}$$

where λ_P, λ_{str}, λ_{ADV} and λ_{OT} represent the weight assigned to pixel loss, SSIM loss, adversarial loss, and Sinkorn loss respectively. Similarly, the discriminator objective function can be defined as,

$$\min_{\mathbf{D}} -\mathbb{E}_{y\sim\mathbb{P}_y}[\log(\mathbf{D}(y)))] - \mathbb{E}_{\hat{y}\sim\mathbb{P}_{\mathbf{G}_\theta}}[\log(1-\mathbf{D}(\hat{y}))] + \lambda_{DA}\mathscr{L}_{DA}, \tag{3}$$

where λ_{DA} is the assigned weight for domain adaptation loss ($\mathscr{L}_{DA}$) (Rout et al., 2020) which is used to enforce $\mathbf{D}$ to mimic the latent features of the HR DEM and sharpen the spatial-attention maps provided an upsampled coarse DEM. The details of all losses are described in Appendix §D.

3 Experiments and results analysis

In this section, we describe our experimental set-up followed by qualitative and quantitative comparison of our model with bicubic as well as other learning-based state-of-the-art methods such as ESRGAN (Demiray et al., 2020) and EffecientNetV2 (Demiray et al., 2021), which also includes recent baseline models for image-guided depth super-resolution like DKN and FDKN (Kim et al., 2021), DADA (Metzger et al., 2023), and FDSR (He et al., 2021).

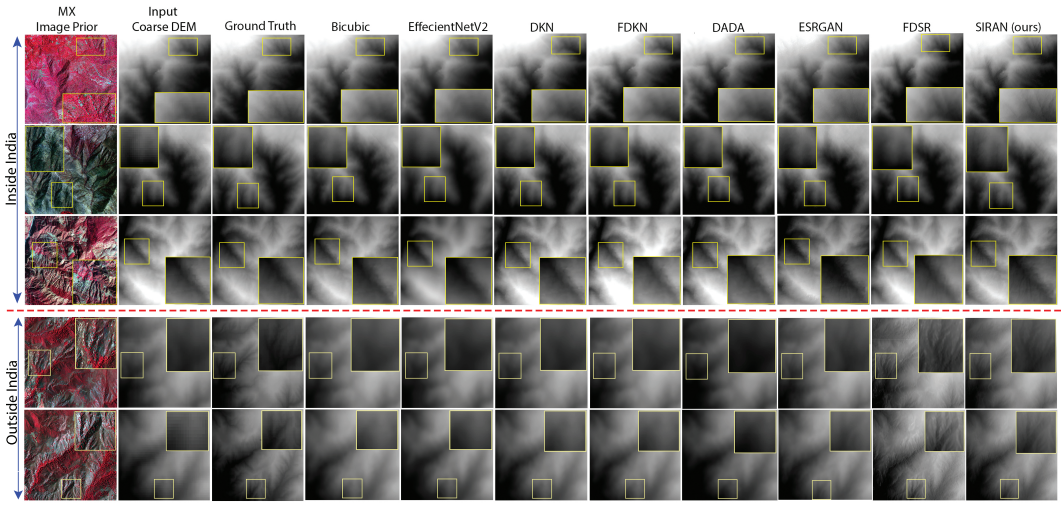

Figure 2: Results for DEM super-resolution (better viewed at 200%) for both inside and ouside India data and comparisons with other baseline methods.

3.1 Dataset and implementation details

Due to lack of realistic DEM SR dataset, we generate our own dataset. For practical application, we opt to use real coarse resolution SRTM DEM with a ground sampling distance (GSD) of 30m as input instead of conventional bicubic downsampled. The HR Indian DEM (GSD=10m) from Cartosat-1 stereoscopic mission is considered our ground truth while the HR MX data (GSD=1.6m) from the Cartosat-2S mission serves as the image guide for our DEM SR task. The DEMs are upsampled to the resolution of MX images using bicubic interpolation to generate a paired dataset. This also helps in increasing the training samples as well as assists the model to learn dense HR features from the prior. The dataset consists of 72,000 paired patches of size $(256, 256)$, featuring diverse landscapes such as vegetation, mountains, and, water regions. We use 40,000 samples for training, 20,000 for cross-validation, and 12,000 for testing, including 10,000 patches from the Indian subcontinent and the rest outside India. As ground truth DEM data is only available for Indian regions, our model is trained on limited landscapes. To check its generalization capability, we test our model on data from the Fallbrook region, US, where Cartosat DEM data is unavailable. For them, we validate our result using available 10m DEM data of the 3D Elevation Program (3DEP) (, USGS). For fair comparison, the guided SR models (DKN, FDKN, DADA and FDSR) are trained from scratch with our generated dataset with MX guide image while ESRGAN and EffecientNetV2 are trained without any guide, as they lack provisions for guide usage. Other implementation details are explained in Appendix §E.

3.2 Results analysis

Table 1: Quantitative comparison with state-of-the-art methods for both patches of inside and outside India. First and second methods are highlighted in red and green, respectively.

Method	**RMSE (m)**		**MAE (m)**		**SSIM(%)**		**PSNR**	
Dataset	**Inside**	**Outside**	**Inside**	**Outside**	**Inside**	**Outside**	**Inside**	**Outside**
Bicubic	15.25	23.19	12.42	22.04	71.27	66.49	30.07	27.79
ENetV2 Demiray et al. (2021)	20.35	30.53	18.72	28.36	69.63	60.04	31.74	25.58
DKN Kim et al. (2021)	12.89	21.16	11.18	19.78	73.59	68.45	32.09	28.22
FDKN Kim et al. (2021)	13.05	21.93	11.34	20.41	74.13	66.83	32.46	27.68
DADA Metzger et al. (2023)	37.49	40.89	32.17	37.74	73.32	69.86	27.94	26.78
ESRGAN Demiray et al. (2020)	31.33	*20.45*	25.56	*18.34*	*82.48*	*75.67*	29.88	*29.05*
FDSR He et al. (2021)	*12.98*	30.58	*10.87*	25.28	81.49	59.81	*33.77*	25.59
SIRAN (ours)	**9.28**	**15.74**	**8.51**	**12.25**	**90.59**	**83.90**	**35.06**	**31.56**

Figure 2 demonstrates the qualitative comparison of our proposed method, where SIRAN highlights key features and retains the perceptual quality with respect to ground truth showcasing its generalization capability. Although ESRGAN and FDSR perform well, ESRGAN tends to produce artifacts and noise while FDSR inpaints unnecessary image details in the generated HR DEMs in

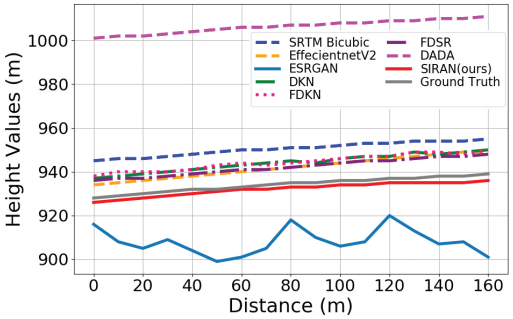

Figure 3: Line profile analysis of SIRAN and other baselines.

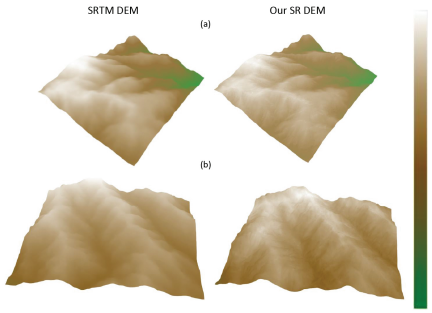

Figure 4: 3-D visualization of Super-resolved and SRTM DEM (better at 200%)

outside India data due to low generalization capability. The quantitative comparison presented with respect to RMSE (m), MAE (m), PSNR and SSIM in table 1 supports these observations. For both inside and outside India patches, our method captures HR structural details resulting in improvement of more than 24% in RMSE and MAE, 8% in SSIM, and 2dB in PSNR. Hence, both quantitative and qualitative analysis suggest that our proposed method SIRAN not only generate high-resolution DEM to the closest proximity of ground truth but also has superior generalization capability for out-of-domain samples. We further validate this by line-profile comparisons as shown in Figure 3 where we show SIRAN's low bias and close adherence to true elevation values. Additionally, in Figure 4, we demonstrate 3-D visualization of generated DEMs, corresponding to a region, where ground truth is unavailable. We compare it with available SRTM DEM, and clearly, our topographic view of generated DEM captures sharper features in mountainous regions as well as in the tributaries of the water basin area. Ablation studies related choice of different modules and loss functions are carried out in Appendix §F.

4 Conclusion

In this paper, we demonstrate an effective approach for DEM super-resolution using realistic coarse data samples. Unlike regular SISR, the proposed method uses high-resolution MX images as prior in a specially designed architectural framework consisting of spatial attention maps from discriminator, PSA, and DMRBs. We also develop a new GAN set-up based on optimization of Sinkhorn distance regularized adversarial learning. We provide theoretical and empirical evidence to show how this choice stabilizes the training of our adversarial model and improves its convergence. We perform quantitative and qualitative analysis by generating and comparing DEMs related to different signatures along with investigating generalization capability by testing out-of-domain samples. We also analyse how each proposed module affects the outcomes. Our method achieves favorable results compared to other state-of-the-art methods.

Limitation: The generated dataset contains samples from only Indian subcontinent regions with limited landscapes and signatures. Therefore, models trained on this dataset may suffer from lack of generalization in presence of certain signatures like urban area.

Dataset: A subset of our generated dataset will be available at: `https://github.com/subhaISRO/DEM-Super-resolution.git`. The details about the dataset will be provided in the corresponding `README.md` file.

5 Acknowledgement

The authors express their sincere gratitude to Shri. Nilesh M Desai, Director, Space Application Centre (SAC) Ahmedabad, Shri. Debajyoti Dhar, Deputy Director, Signal and Image Processing Area (SIPA) and Shri. S Devakanth Naidu, Head, High-resolution Data Processing Division (HDPD) for the encouragement; and other HDPD members for their constant support. We also thank National Remote Sensing Centre (NRSC), ISRO for providing necessary datasets.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Genevay Aude, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport, 2016a.

Genevay Aude, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport, 2016b.

Zixuan Chen, Xuewen Wang, Zekai Xu, and Hou Wenguang. Convolutional neural network based dem super resolution. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B3:247–250, 06 2016. doi: 10.5194/isprsarchives-XLI-B3-247-2016.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances. *Advances in Neural Information Processing Systems*, 26, 06 2013.

Bekir Z. Demiray, Muhammed Ali Sit, and Ibrahim Demir. D-SRGAN: DEM super-resolution with generative adversarial networks. *CoRR*, abs/2004.04788, 2020. URL https://arxiv.org/abs/2004.04788.

Bekir Z Demiray, Muhammed Sit, and Ibrahim Demir. Dem super-resolution with efficientnetv2, 2021.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38, 12 2014. doi: 10.1109/TPAMI.2015.2439281.

Ishan P. Durugkar, Ian Gemp, and Sridhar Mahadevan. Generative multi-adversarial networks. *CoRR*, abs/1611.01673, 2016. URL http://arxiv.org/abs/1611.01673.

Hajar Emami, Majid Moradi Aliabadi, Ming Dong, and Ratna Babu Chinnam. SPA-GAN: spatial attention GAN for image-to-image translation. *CoRR*, abs/1908.06616, 2019. URL http://arxiv.org/abs/1908.06616.

Peter Fisher and Nicholas Tate. Causes and consequences of error in digital elevation models. progress in physical geography. *Progress in Physical Geography*, 30, 08 2006. doi: 10.1191/0309133306pp492ra.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In Amos Storkey and Fernando Perez-Cruz (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1608–1617. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/genevay18a.html.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences, 2019.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017. URL http://arxiv.org/abs/1704.00028.

He He and Wan-Chi Siu. Single image super-resolution using gaussian process regression. In *CVPR 2011*, pp. 449–456, 2011. doi: 10.1109/CVPR.2011.5995713.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Lingzhi He, Hongguang Zhu, Feng Li, Huihui Bai, Runmin Cong, Chunjie Zhang, Chunyu Lin, Meiqin Liu, and Yao Zhao. Towards fast and accurate real-world depth super-resolution: Benchmark dataset and baseline. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9225–9234, 2021. URL https://api.semanticscholar.org/CorpusID:233219880.

Antoine Houdard, Arthur Leclaire, Nicolas Papadakis, and Julien Rabin. On the existence of optimal transport gradient for learning generative models, 2021.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. URL http://arxiv.org/abs/1608.06993.

Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable kernel networks for joint image filtering. *International Journal of Computer Vision*, 129, 02 2021. doi: 10.1007/s11263-020-01386-z.

Dong-Eon Kim, Philippe Gourbesville, and Shie-Yui Liong. Overcoming data scarcity in flood hazard assessment using remote sensing and artificial neural network. *Smart Water*, 4:1–15, 2019. doi: 10.1186/s40713-018-0014-5.

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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, 2017. doi: 10.1109/CVPR.2017.19.

Jing Li and David Wong. Effect of dem sources on hydrologic applications. *Computers, Environment and Urban Systems*, 34:251–261, 05 2010. doi: 10.1016/j.compenvurbsys.2009.11.002.

Huajun Liu, Fuqiang Liu, Xinyi Fan, and Dong Huang. Polarized self-attention: Towards high-quality pixel-wise regression. *CoRR*, abs/2107.00782, 2021. URL https://arxiv.org/abs/2107.00782.

Xiaoye Liu. Airborne lidar for dem generation: Some critical issues. progress in physical geography. *Progress in Physical Geography - PROG PHYS GEOG*, 32:31–49, 02 2008. doi: 10.1177/0309133308089496.

Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Guided depth super-resolution by deep anisotropic diffusion. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Gary Priestnall, Jad Jaafar, and A. Duncan. Extracting urban features from lidar digital surface models. *Computers, Environment and Urban Systems*, 24:65–78, 03 2000. doi: 10.1016/S0198-9715(99)00047-2.

Matan Protter, Michael Elad, Hiroyuki Takeda, and Peyman Milanfar. Generalizing the nonlocal-means to super-resolution reconstruction. *IEEE Transactions on Image Processing*, 18(1):36–51, 2009. doi: 10.1109/TIP.2008.2008067.

Litu Rout. Understanding the role of adversarial regularization in supervised learning. *CoRR*, abs/2010.00522, 2020. URL https://arxiv.org/abs/2010.00522.

Litu Rout, Indranil Misra, S Manthira Moorthi, and Debajyoti Dhar. S2a: Wasserstein gan with spatio-spectral laplacian attention for multi-spectral band synthesis, 2020.

Rasmus Sørensen and Jan Seibert. Effects of dem resolution on the calculation of topographical indices: Twi and its components. *Journal of Hydrology*, 347:79–89, 12 2007. doi: 10.1016/j.jhydrol.2007.09.001.

Hiroyuki Takeda, Sina Farsiu, and Peyman Milanfar. Kernel regression for image processing and reconstruction. *IEEE Transactions on Image Processing*, 16(2):349–366, 2007. doi: 10.1109/TIP.2006.888330.

Sebastiano Trevisani, Marco Cavalli, and Lorenzo Marchi. Surface texture analysis of a high-resolution dtm: Interpreting an alpine basin. *Geomorphology*, s 161–162:26–39, 08 2012. doi: 10.1016/j.geomorph.2012.03.031.

U.S. Geological Survey (USGS). 1/3rd arc-second digital elevation models (dems)- usgs national map 3dep downloadable data collection. 2019. URL URLhttps://www.usgs.gov/the-national-map-data-delivery.

A APPENDIX: SMOOTHNESS OF SINKHORN LOSS

Theorem 1 (Smoothness of Sinkhorn loss)**.** *Consider the Sinkhorn loss $\mathcal{S}_{C,\varepsilon}(\mu_\theta, \nu)$ between two measures μ_θ and ν on $\mathcal{X}$ and $\mathcal{Y}$ two bounded subsets of $\mathbb{R}^d$, with a $\mathcal{C}^\infty$, L_0-Lipschitz, and L_1-smooth cost function C. Then, for $(\theta_1, \theta_2) \in \Theta$, one has,*

$$\begin{aligned} &\mathbb{E}||\nabla_\theta \mathcal{S}_{C,\varepsilon}(\mu_{\theta_1}, \nu) - \nabla_\theta \mathcal{S}_{C,\varepsilon}(\mu_{\theta_2}, \nu)|| \\ &= \mathcal{O}(L(L_1 + \frac{2L_0^2 L}{\varepsilon(1 + Be^{\frac{\kappa}{\varepsilon}})}))||\theta_1 - \theta_2||, \end{aligned} \tag{4}$$

where L is the Lipschitz in θ corresponding to $\mathbf{G}$, $\kappa = 2(L_0|\mathcal{X}| + ||C||_\infty)$, $B = d.\max(||m||, ||M||)$ with m amd M being the minimum and maximum values in set $\mathcal{X}$. Let Γ_ε represent the smoothness mentioned above, then we get the following asymptotic behavior in ε:

1. *as $\varepsilon \to 0$, $\Gamma_\varepsilon \to \mathcal{O}(\frac{2L_0^2 L^2}{B\varepsilon e^{\frac{\kappa}{\varepsilon}}})$*
2. *as $\varepsilon \to \infty$, $\Gamma_\varepsilon \to \mathcal{O}(LL_1)$*

Proof. We will define some of the terminologies, which are necessary for this proof. From equation 6 of the main paper, the entropic optimal transport (Aude et al., 2016a) can be defined as,

$$\begin{aligned} &\mathcal{W}_{C,\varepsilon}(\mu_\theta, \nu) = \inf_{\pi \in \Pi(\mu_\theta, \nu)} \int_{\mathcal{X}\times\mathcal{Y}} [C(\mathbf{G}_\theta(x), y)] d\pi(\mathbf{G}_\theta(x), y) + \varepsilon I_\pi(\mathbf{G}_\theta(x), y), \\ &\text{where } I_\pi(\mathbf{G}_\theta(x), y)) = \int_{\mathcal{X}\times\mathcal{Y}} [\log\left(\frac{\pi(\mathbf{G}_\theta(x), y)}{\mu_\theta(\mathbf{G}_\theta(x))\nu(y)}\right)] d\pi(\mathbf{G}_\theta(x), y), \\ &\text{s.t.} \int_{\mathcal{X}} \pi(\mathbf{G}_\theta(x), y)\, dx = \nu(y), \int_{\mathcal{Y}} \pi(\mathbf{G}_\theta(x), y)\, dy = \mu_\theta(\mathbf{G}_\theta(x)) \;\&\; \pi(\mathbf{G}_\theta(x), y) \geq 0. \end{aligned} \tag{5}$$

The formulation in equation 5 corresponds to the primal problem of regularized OT and, this allows us to express the dual formulation of regularized OT as the maximization of an expectation problem, as shown in equation 6 (Aude et al., 2016b).

$$\begin{aligned} \mathcal{W}_{C,\varepsilon}(\mu_\theta, \nu) = &\sup_{\phi,\psi \in \Phi} \int_{\mathcal{X}} \phi(\mathbf{G}_\theta(x))\; d\mu_\theta(\mathbf{G}_\theta(x)) + \int_{\mathcal{Y}} \psi(y)\; d\nu(y) \\ &- \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} e^{\left(\frac{\phi(\mathbf{G}_\theta(x)) + \psi(y) - C(\mathbf{G}_\theta(x), y)}{\varepsilon}\right)} d\mu_\theta(\mathbf{G}_\theta(x))\, d\nu(y) + \varepsilon \end{aligned} \tag{6}$$

where $\Phi = \{(\phi, \psi) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y})\}$ is set of real valued continuous functions for domain $\mathcal{X}$ and $\mathcal{Y}$ and they are referred as dual potentials. Now, given optimal dual potentials $\phi^*(\cdot)$, and $\psi^*(\cdot)$, the optmal coupling $\pi^*(\cdot)$ as per (Aude et al., 2016b) can be defined as

$$\pi^*(\mathbf{G}_\theta(x), y) = \mu_\theta(\mathbf{G}_\theta(x))\, \nu(y)\, e^{\frac{\phi^*(\mathbf{G}_\theta(x)) + \psi^*(y) - C(\mathbf{G}_\theta(x), y)}{\varepsilon}}. \tag{7}$$

To prove **Theorem 1**, we need an important property regarding its Lipschitz continuity of the dual potentials, which is explained in the following **Lemma**.

Lemma A.1. *If $C(\cdot)$ is L_0 Lipschitz, then the dual potentials are also L_0 Lipschitz.*

Proof. Assuming $\hat{y} = \mathbf{G}_\theta(x)$, then $C(\hat{y}, y)$ is L_0-Lipschitz in $\hat{y}$. As, the entropy $I_\pi(\cdot)$ is selected as Shannon entropy, according to Cuturi (2013) using the softmin operator, the optimal potential $\phi^*(\cdot)$ satisfy the following equation

$$\phi^*(\hat{y}) = -\varepsilon \ln\left[\int_{\mathcal{Y}} \exp\left(\frac{\psi^*(y) - C(\hat{y}, y)}{\varepsilon}\right) dy\right] \tag{8}$$

Now, to estimate the Lipschitz of ϕ^*, we have to find the upper bound of $||\nabla_{\hat{y}} \phi^*(\hat{y})||$. Hence, taking the gradient of equation 8 with respect to $\hat{y}$, the upper-bound of its norm can be written as,

$$||\nabla_{\hat{y}}\phi^*(\hat{y})|| = \frac{||\int_{\mathcal{Y}} \exp\left(\frac{\psi^*(y)-C(\hat{y},y)}{\varepsilon}\right)\nabla_{\hat{y}}C(\hat{y},y)\,dy||}{||\int_{\mathcal{Y}} \exp\left(\frac{\psi^*(y)-C(\hat{y},y)}{\varepsilon}\right)dy||} \tag{9}$$

Now due to Lipschitz continuity of $C(\hat{y}, y)$, we can say $\nabla_{\hat{y}}||C(\hat{y}, y)|| \leq L_0$. Hence, using Cauchy-Schwarz inequality we will get,

$$||\nabla_{\hat{y}}\phi^*(\hat{y})|| \leq ||\nabla_{\hat{y}}C(\hat{y},y)||\frac{||\int_{\mathcal{Y}} \exp\left(\frac{\psi^*(y)-C(\hat{y},y)}{\varepsilon}\right)dy||}{||\int_{\mathcal{Y}} \exp\left(\frac{\psi^*(y)-C(\hat{y},y)}{\varepsilon}\right)dy||} = L_0. \tag{10}$$

This completes the proof of the lemma. An alternative proof is provided by Houdard et al. (2021) in Proposition 4. Similarly, it can be proved for the other potential term. □

For any θ_1, $\theta_2 \in \Theta$ will result in different coupling solutions π_i^*, for $i = 1, 2$. Now, based on Danskins' theorem for optimal coupling $\pi^*(\theta)$, we can write

$$\nabla_\theta \mathcal{W}_{C,\varepsilon}(\mu_\theta, \nu) = \mathbb{E}_{\mathbf{G}_\theta(x), y\sim\pi^*(\theta)}\left[\nabla_\theta C(\mathbf{G}_\theta(x), y)\right] \tag{11}$$

Therefore, for any θ_1 and θ_2, we can write,

$$\begin{aligned}
&||\nabla_\theta \mathcal{W}_{C,\varepsilon}(\mu_{\theta_1}, \nu) - \nabla_\theta \mathcal{W}_{C,\varepsilon}(\mu_{\theta_2}, \nu)|| \leq \\
&||\mathbb{E}_{\mathbf{G}_{\theta_1}(x), y\sim\pi_1^*}\left[\nabla_\theta C(\mathbf{G}_{\theta_1}(x), y)\right] - \mathbb{E}_{\mathbf{G}_{\theta_1}(x), y\sim\pi_2^*}\left[\nabla_\theta C(\mathbf{G}_{\theta_1}(x), y)\right]|| \\
&+ ||\mathbb{E}_{\mathbf{G}_{\theta_1}(x), y\sim\pi_2^*}\left[\nabla_\theta C(\mathbf{G}_{\theta_1}(x), y)\right] - \mathbb{E}_{\mathbf{G}_{\theta_2}(x), y\sim\pi_2^*}\left[\nabla_\theta C(\mathbf{G}_{\theta_2}(x), y)\right]|| \\
&\leq L_0 L||\pi_1^* - \pi_2^*|| + L_1 L||\theta_1 - \theta_2||
\end{aligned} \tag{12}$$

Now with respect to different θ_i, for $i = 1, 2$ with different pair of dual potentials, the $||\pi_1^* - \pi_2^*||$ can be written as below. For simplicity we denote $\mu_\theta \equiv \mu_\theta(\mathbf{G}_\theta(x))$ and $\nu \equiv \nu(y)$.

$$\begin{aligned}
||\pi_1^* - \pi_2^*|| &= ||\mu_{\theta_1}\nu \exp\left(\frac{\phi^*(\mathbf{G}_{\theta_1}(x)) + \psi^*(y) - C(\mathbf{G}_{\theta_1}(x), y)}{\varepsilon}\right) \\
&\quad - \mu_{\theta_2}\nu \exp\left(\frac{\phi^*(\mathbf{G}_{\theta_2}(x)) + \psi^*(y) - C(\mathbf{G}_{\theta_2}(x), y)}{\varepsilon}\right)|| \\
&\leq ||\nu \exp\left(\frac{\phi^*(\mathbf{G}_{\theta_1}(x)) + \psi^*(y) - C(\mathbf{G}_{\theta_1}(x), y)}{\varepsilon}\right)(\mu_{\theta_1} - \mu_{\theta_2})|| \\
&\quad + ||\mu_{\theta_2}\nu\left[\exp\left(\frac{\phi^*(\mathbf{G}_{\theta_1}(x)) + \psi^*(y) - C(\mathbf{G}_{\theta_1}(x), y)}{\varepsilon}\right)\right. \\
&\quad \left. - \exp\left(\frac{\phi^*(\mathbf{G}_{\theta_2}(x)) + \psi^*(y) - C(\mathbf{G}_{\theta_2}(x), y)}{\varepsilon}\right)\right]||
\end{aligned} \tag{13}$$

From Genevay et al. (2019), we know, as the dual potentials are L_0-Lipschitz, $\forall \mathbf{G}_\theta(x) \in \mathcal{X}$, we can write, $\phi^*(\mathbf{G}_\theta(x)) \leq L_0|\mathbf{G}_\theta(x)|$. And from property of c-transform, for $\forall y \in \mathcal{Y}$ we can also write $\psi^*(y) \leq \max_{\mathbf{G}_\theta(x)} \phi^*(\mathbf{G}_\theta(x)) - C(\mathbf{G}_\theta(x), y)$. We assume $\mathcal{X}$ to be a bounded set in our case, hence, denoting $|\mathcal{X}|$ as the diameter of the space, at optimality, we can get that $\forall \mathbf{G}_\theta(x) \in \mathcal{X}$, $y \in \mathcal{Y}$

$$\begin{aligned}
&\Rightarrow \phi^*(\mathbf{G}_\theta(x)) + \psi^*(y) \leq 2L_0|\mathcal{X}| + ||C||_\infty \\
&\Rightarrow \exp\left(\frac{\phi^*(\mathbf{G}_\theta(x)) + \psi^*(y) - C(\mathbf{G}_\theta(x), y)}{\varepsilon}\right)) \leq \exp\left(2\frac{L_0|\mathcal{X}| + ||C||_\infty}{\varepsilon}\right)
\end{aligned} \tag{14}$$

Hence, the exponential terms in equation 13 are bounded, and we can assume it has a finite Lipschitz constant L_{exp}. Taking $\kappa = 2(L_0|\mathcal{X}| + ||C||_\infty)$, and using Cauchy-Schwarz, we can rewrite equation 13 as,

$$\begin{aligned}
||\pi_1^* - \pi_2^*|| &\leq \exp\left(\frac{\kappa}{\varepsilon}\right)||\nu||.||\mu_{\theta_1} - \mu_{\theta_2}|| \\
&\quad + L_{exp}||\mu_{\theta_2}||.||\nu||.||\frac{(\phi^*(\mathbf{G}_{\theta_1}(x)) - \phi^*(\mathbf{G}_{\theta_2}(x))) - (C(\mathbf{G}_{\theta_1}(x), y) - C(\mathbf{G}_{\theta_2}(x), y))}{\varepsilon}|| \\
&\leq \exp\left(\frac{\kappa}{\varepsilon}\right)||\nu||.||\mu_{\theta_1} - \mu_{\theta_2}|| + 2\frac{L_{exp}L_0 L}{\varepsilon}||\mu_{\theta_2}||.||\nu||.||\theta_1 - \theta_2||
\end{aligned} \tag{15}$$

Now, as the input space $\mathcal{X}$ and output space $\mathcal{Y}$ are bounded, the corresponding measures μ_θ and ν will also be bounded. We assume, $||\mu_\theta|| \leq \lambda_1$ and $||\nu|| \leq \lambda_2$. If we apply equation 14 in equation 7, to get the upper bound of the coupling function, we will get $||\pi_1^* - \pi_2^*|| \leq \exp\left(\frac{\kappa}{\varepsilon}\right)||\nu||.||\mu_{\theta_1} - \mu_{\theta_2}||$ which is less than the bound in equation 15. Then, we can find some constant upper bound of $||\pi_1^* - \pi_2^*||$, using the assumed bounds of measures and can write $||\pi_1^* - \pi_2^*|| \leq \exp\left(\frac{\kappa}{\varepsilon}\right)||\nu||.||\mu_{\theta_1} - \mu_{\theta_2}|| \leq K$, such that,

$$K \leq \exp\left(\frac{\kappa}{\varepsilon}\right)||\nu||.||\mu_{\theta_1} - \mu_{\theta_2}|| + 2\frac{L_{exp}L_0L}{\varepsilon}||\mu_{\theta_2}||.||\nu||.||\theta_1 - \theta_2||$$

Then using the marginal condition as shown in in equation 5, we can write equation 15 as,

$$\begin{aligned} K &\leq \lambda_1 \exp\left(\frac{\kappa}{\varepsilon}\right)||\int_{\mathcal{X}} \pi_1^* dx - \int_{\mathcal{X}} \pi_2^* dx|| + 2\lambda_1\lambda_2\frac{L_{exp}L_0L}{\varepsilon}||\theta_1 - \theta_2|| \\ &\leq \lambda_1 \exp\left(\frac{\kappa}{\varepsilon}\right)\int_{\mathcal{X}} ||\pi_1^* - \pi_2^*||.|dx| + 2\lambda_1\lambda_2\frac{L_{exp}L_0L}{\varepsilon}||\theta_1 - \theta_2|| \\ &\leq \lambda_1 \exp\left(\frac{\kappa}{\varepsilon}\right)K\int_{\mathcal{X}} |dx| + 2\lambda_1\lambda_2\frac{L_{exp}L_0L}{\varepsilon}||\theta_1 - \theta_2|| \end{aligned} \tag{16}$$

The input set is a compact set such that $\mathcal{X} \subset \mathbb{R}^d$. So, assuming m and M to be the minimum and maximum value in set $\mathcal{X}$ and considering the whole situation in discrete space, equation 16, can be rewritten as,

$$\begin{aligned} K &\leq \lambda_1 \exp\left(\frac{\kappa}{\varepsilon}\right)K\sum_{x\in\mathcal{X}} |x| + 2\lambda_1\lambda_2\frac{L_{exp}L_0L}{\varepsilon}||\theta_1 - \theta_2|| \\ &\leq \lambda_1 \exp\left(\frac{\kappa}{\varepsilon}\right)Kd\max\left(||M||, |||m|\right) + 2\lambda_1\lambda_2\frac{L_{exp}L_0L}{\varepsilon}||\theta_1 - \theta_2||, \end{aligned} \tag{17}$$

Now, taking $B = d\max\left(||M||, |||m|\right)$, and doing necessary subtraction and division on both sides of equation 17, it can be rewritten as

$$\begin{aligned} K &\leq \frac{2\lambda_1\lambda_2 L_{exp}L_0L}{\varepsilon\left(1 - \lambda_1 B\exp\left(\frac{\kappa}{\varepsilon}\right)\right)}||\theta_1 - \theta_2|| \\ &\leq \frac{2\lambda_1\lambda_2 L_{exp}L_0L}{\varepsilon\left(1 + \lambda_1 B\exp\left(\frac{\kappa}{\varepsilon}\right)\right)}||\theta_1 - \theta_2|| \end{aligned} \tag{18}$$

Equation 18, satisfies because $\frac{\kappa}{\varepsilon} \geq 0$. As, $||\pi_1^* - \pi_2^*|| \leq K$, from equation 18, it can be written as

$$||\pi_1^* - \pi_2^*|| \leq \frac{2\lambda_1\lambda_2 L_{exp}L_0L}{\varepsilon\left(1 + \lambda_1 B\exp\left(\frac{\kappa}{\varepsilon}\right)\right)}||\theta_1 - \theta_2|| \tag{19}$$

Substituting equation 19 in equation 12, we will get,

$$\begin{aligned} ||\nabla_\theta \mathcal{W}_{C,\varepsilon}\left(\mu_{\theta_1}, \nu\right) - \nabla_\theta \mathcal{W}_{C,\varepsilon}\left(\mu_{\theta_2}, \nu\right)|| &\leq L_0L||\pi_1^* - \pi_2^*|| + L_1L||\theta_1 - \theta_2|| \\ &\leq \left(L_1L + \frac{2\lambda_1\lambda_2 L_{exp}L_0^2L^2}{\varepsilon\left(1 + \lambda_1 B\exp\left(\frac{\kappa}{\varepsilon}\right)\right)}\right)||\theta_1 - \theta_2|| \end{aligned} \tag{20}$$

So, the EOT problem defined in equation 5 has $\hat{\Gamma}_\varepsilon$ smoothness in θ with $\hat{\Gamma}_\varepsilon = L_1L + \frac{2\lambda_1\lambda_2 L_{exp}L_0^2L^2}{\varepsilon\left(1+\lambda_1 B\exp\left(\frac{\kappa}{\varepsilon}\right)\right)}$. From this, we can derive the smoothness of Sinkhorn loss defined in equation 1. Note that only the first two terms in equation 1 are θ dependent. Therefore, they only contribute to the gradient approximation and both of them will satisfy the same smoothness condition as defined in equation 20. So, if Sinkhorn loss has smoothness Γ_ε, it will satisfy, $\Gamma_\varepsilon = \frac{3}{2}\hat{\Gamma}_\varepsilon$. In general, we can define the smoothness of Sinkhorn loss with $(\theta_1, \theta_2) \in \Theta$ as,

$$||\nabla_\theta S_{C,\varepsilon}\left(\mu_{\theta_1}, \nu\right) - \nabla_\theta S_{C,\varepsilon}\left(\mu_{\theta_2}, \nu\right)|| \leq \mathcal{O}\left(L_1L + \frac{2L_0^2L^2}{\varepsilon\left(1 + B\exp\left(\frac{\kappa}{\varepsilon}\right)\right)}\right)||\theta_1 - \theta_2|| \tag{21}$$

This completes the statement of **Theorem 1** □

B Appendix: Upper-bound of expected gradient in SIRAN set-up

Theorem 2. *Let $l(\cdot)$, $g(\cdot)$ and $\mathcal{S}_{C,\varepsilon}(\cdot)$ be the objective functions related to supervised losses, adversarial loss and Sinkhorn loss respectively, and θ^* and ψ^* be the parameters of optimal*

generator $\mathbf{G}$ *and discriminator* $\mathbf{D}$. *Let us suppose* $l(p, y)$, *where* $p = \mathbf{G}_\theta(x)$, *is* β*-smooth in* p. *If* $||\theta - \theta^*|| \leq \epsilon$ *and* $||\psi - \psi^*|| \leq \delta$, *then* $||\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{X}\times\mathcal{Y}}[l(\mathbf{G}_\theta(x), y) + \mathcal{S}_{C,\varepsilon}(\mu_\theta(\mathbf{G}_\theta(x)), \nu(y)) - g(\psi; \mathbf{G}_\theta(x))]|| \leq L^2\epsilon(\beta + \Gamma_\varepsilon) + L\delta$, *where* Γ_ε *is the derived smoothness of Sinkhorn loss in equation 4.*

Proof. This proof is inspired by Rout (2020). Assuming $\Gamma = \mathcal{O}\left(L_1 + \frac{2L_0^2}{\varepsilon\left(1+B\exp\left(\frac{\kappa}{\varepsilon}\right)\right)}\right)$ be the smoothness in p for Sinkhorn loss $S_{C,\varepsilon}\left(\mu_\theta\left(p\right), \nu\left(y\right)\right)$, where $p = \mathbf{G}_\theta\left(x\right)$. For simplicity, we use a common set for inputs and outputs as $\mathcal{P}$. Hence, to approximate the gradient of Sinkhorn loss, using Jensen's inequality, we can write,

$$\begin{aligned}
&||\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}[S_{C,\varepsilon}\left(\mu_\theta\left(\mathbf{G}_\theta\left(x\right)\right), \nu\left(y\right)\right)]|| \\
&\leq \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[||\nabla_\theta S_{C,\varepsilon}\left(\mu_\theta\left(\mathbf{G}_\theta\left(x\right)\right), \nu\left(y\right)\right)||\right] \\
&\leq \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\underbrace{||\nabla_p S_{C,\varepsilon}\left(\mu_\theta\left(p\right), \nu\left(y\right)\right)||.||\nabla_\theta \mathbf{G}_\theta\left(x\right)||}_{\text{Cauchy-Schwarz inequality}}\right] \\
&\leq L\mathbb{E}_{(x,y)\sim\mathcal{P}}\left[||\nabla_p S_{C,\varepsilon}\left(\mu_\theta\left(p\right), \nu\left(y\right)\right)||\right]
\end{aligned} \tag{22}$$

Say, for optimized parameter θ^*, $t = \mathbf{G}_{\theta^*}(x)$. Since, $||\theta - \theta^*||$, we can write using the smoothness of sinkhorn loss and Lipschitz of model parameters,

$$\begin{aligned}
&||\nabla_p S_{C,\varepsilon}\left(\mu_\theta\left(p\right), \nu\left(y\right)\right)|| - ||\nabla_t S_{C,\varepsilon}\left(\mu_{\theta^*}\left(t\right), \nu\left(y\right)\right)|| \\
&\leq ||\nabla_p S_{C,\varepsilon}\left(\mu_\theta\left(p\right), \nu\left(y\right)\right) - \nabla_t S_{C,\varepsilon}\left(\mu_{\theta^*}\left(t\right), \nu\left(y\right)\right)|| \\
&\leq \Gamma||p - t|| = \Gamma||\mathbf{G}(\theta)(x) - \mathbf{G}(\theta^*(x))|| \\
&\leq \Gamma L||\theta - \theta^*|| \leq \Gamma L\epsilon
\end{aligned} \tag{23}$$

At optimal condition, $||\nabla_t S_{C,\varepsilon}\left(\mu_{\theta^*}\left(t\right), \nu\left(y\right)\right)|| = 0$ as the distributions of y and $t = \mathbf{G}_{\theta^*}(x)$ are aligned for optimal θ^*. So, by substituting equation 23 in equation 22, we will get

$$||\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[S_{C,\varepsilon}\left(\mu_\theta\left(\mathbf{G}_\theta\left(x\right)\right), \nu\left(y\right)\right)\right]|| \leq L^2\Gamma\epsilon \tag{24}$$

From Lemma 1 of Rout (2020), we get,

$$||\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[l(\mathbf{G}_\theta(x), y)\right]|| \leq L^2\beta\epsilon \tag{25}$$

Similarly, from Lemma 2 of Rout (2020), we get

$$||-\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[g(\psi; \mathbf{G}_\theta(x))\right]|| \leq L\delta \tag{26}$$

Here, ψ is parameters of discriminator $\mathbf{D}$. So using equations 24, 25, and 26, for the combination of losses we will get,

$$\begin{aligned}
&||\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}[l(\mathbf{G}_\theta(x), y) + S_{C,\varepsilon}\left(\mu_\theta\left(\mathbf{G}_\theta\left(x\right)\right), \nu\left(y\right)\right) \\
&- g(\psi; \mathbf{G}_\theta(x))]|| \leq ||\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[l(\mathbf{G}_\theta(x), y)\right]|| \\
&+ ||\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[S_{C,\varepsilon}\left(\mu_\theta\left(\mathbf{G}_\theta\left(x\right)\right), \nu\left(y\right)\right)\right]|| \\
&+ ||-\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[g(\psi; \mathbf{G}_\theta(x))\right]|| \\
&\leq L^2\beta\epsilon + L^2\Gamma\epsilon + L\delta = L^2\epsilon(\beta + \Gamma) + L\delta
\end{aligned} \tag{27}$$

This completes the proof. □

In a regular adversarial set-ups like WGAN and their variants $\epsilon \to 0$ suggests reductions of δ, which eventually results in a vanishing gradient near the optimal region as shown by Rout (2020). However, in our Sinkhorn regularized adversarial framework, the upper bound is also dependent on Γ_ε as exhibited in **Theorem2**. From equation 4, Γ_ε is exponentially variable with respect to the choice of ε. Hence, the selection of adequate ε will have a profound impact in mitigating the vanishing gradient problem suffered by regular adversarial setup. Experimental results to support this claim is shown in §G.

C Appendix: Iteration complexiety of SIRAN

Theorem 3. *Suppose the supervised loss* $l(\theta)$ *is lower bounded by* $l^* > -\infty$ *and it is twice differentiable. For some arbitrarily small* $\zeta > 0$, $\eta > 0$ *and* ϵ_1-*stationary point with* $\epsilon_1 > 0$,

let $||\nabla g(\psi; \mathbf{G}_\theta(x))|| \geq \zeta$, $||\nabla \mathcal{S}_{C,\varepsilon}(\mu_\theta(\mathbf{G}_\theta(x)), \nu(y))|| \geq \eta$ *and* $||\nabla l(\mathbf{G}_\theta(x), y)|| \geq \epsilon_1$, *with conditions* $\delta \leq \frac{\sqrt{2\epsilon_1\zeta}}{L}$, *and* $\Gamma_\varepsilon < \frac{\sqrt{2\epsilon_1\eta}}{L^2\epsilon}$, *then the iteration complexity in Sinkhorn regularized adversarial framework is upper bounded by* $\mathcal{O}(\frac{(l(\theta_0)-l^*)\beta_1}{\epsilon_1^2+2\epsilon_1(\zeta+\eta)-L^2(\delta^2+L^2\Gamma_\varepsilon^2\epsilon^2)})$, *assuming* $||\nabla^2 l(\theta)|| \leq \beta_1$.

Proof. This proof also follows the steps of Theorem 3 from Rout (2020). In the sinkhorn regularized adversarial setup, the parameters θ are updated using fixed step gradient descent. They iterate as,

$$\theta_{t+1} = \theta_t - h_t \nabla (l(\theta_t) + S_{C,\varepsilon}\left(\mu_{\theta_t}\left(\mathbf{G}_{\theta_t}\left(x\right)\right), \nu\left(y\right)\right) - g(\psi; \mathbf{G}_{\theta_t}(x))). \tag{28}$$

For simplicity, we denote $S_{C,\varepsilon}\left(\mu_{\theta_t}\left(\mathbf{G}_{\theta_t}\left(x\right)\right), \nu\left(y\right)\right) \equiv S_{C,\varepsilon}(\mu_{\theta_t}, \nu)$. Using Taylor's expansion,

$$l(\theta_{t+1}) = l(\theta_t) + \nabla l(\theta_t)(\theta_{t+1} - \theta_t) + \frac{1}{2}(\theta_{t+1} - \theta_t)^T \nabla^2 l(\theta_t)(\theta_{t+1} - \theta_t) \tag{29}$$

Now, substituting $\theta_{t+1} - \theta_t$ from equation 28, and using triangle inequality and Cauchy-Schwarz inequality, equation 29 can be rewritten as,

$$\begin{aligned} l(\theta_{t+1}) \leq & l(\theta_t) - h_t||\nabla l(\theta_t)||^2 - h_t||\nabla l(\theta_t)||.||\nabla S_{C,\varepsilon}(\mu_{\theta_t}, \nu)|| - h_t||\nabla l(\theta_t)||.||g(\psi; \mathbf{G}_{\theta_t}(x))|| \\ & + h_t^2||\nabla(l(\theta_t) + S_{C,\varepsilon}(\mu_{\theta_t}, \nu) - g(\psi; \mathbf{G}_{\theta_t}(x)))||^2 \frac{||\nabla^2 l(\theta_t)||}{2}. \end{aligned} \tag{30}$$

Taking into account the assumptions in **Theorem 3** and utilizing Minkowski's inequality, equation 30 can be rewritten as,

$$\begin{aligned} l(\theta_{t+1}) \leq & l(\theta_t) - h_t||\nabla l(\theta_t)||^2 - h_t||\nabla l(\theta_t)||\eta - h_t||\nabla l(\theta_t)||\zeta \\ & + h_t^2(||\nabla(l(\theta_t)||^2 + ||S_{C,\varepsilon}(\mu_{\theta_t}, \nu)||^2 + ||g(\psi; \mathbf{G}_{\theta_t}(x)))||^2)\frac{\beta_1}{2}. \end{aligned} \tag{31}$$

Using $h_t = \frac{1}{\beta_1}$, from equation 31, we can write,

$$\begin{aligned} l(\theta_{t+1}) &\leq l(\theta_t) - \frac{h_t||\nabla l(\theta_t)||^2}{2} - h_t||\nabla l(\theta_t)||\eta - h_t||\nabla l(\theta_t)||\zeta \\ &\quad + \frac{h_t||S_{C,\varepsilon}(\mu_{\theta_t}, \nu)||^2}{2} + \frac{h_t||g(\psi; \mathbf{G}_{\theta_t}(x)))||^2}{2} \\ &\leq l(\theta_t) - \frac{h_t\epsilon_1^2}{2} - h_t\epsilon_1\eta - h_t\epsilon_1\zeta + \frac{h_t L^4\Gamma^2\epsilon^2}{2} + \frac{h_t L^2\delta^2}{2}. \end{aligned} \tag{32}$$

Assuming T iterations to reach this ϵ_1-stationary point, then for $t \leq T$, doing telescopic sum over t,

$$\begin{aligned} \sum_{t=0}^{T-1} l(\theta_{t+1}) - l(\theta_t) &\leq \frac{-T(\epsilon_1^2 + 2\epsilon_1(\zeta+\eta) - L^2(\delta^2 + L^2\Gamma^2\epsilon^2))}{2\beta_1} \\ \Rightarrow T &\leq \frac{2(l(\theta_0) - l^*)\beta_1}{(\epsilon_1^2 + 2\epsilon_1(\zeta+\eta) - L^2(\delta^2 + L^2\Gamma^2\epsilon^2))} \end{aligned} \tag{33}$$

Therefore, using the iteration complexity definition of Rout (2020), we obtain,

$$\sup_{\theta_0 \in \{\mathbb{R}^{h\times d_x}, \mathbb{R}^{d_y \times h}\}, l \in \mathscr{L}} \mathcal{T}_{\epsilon_1}(A_h[l, \theta_0], l) = \mathcal{O}\left(\frac{(l(\theta_0) - l^*)\beta_1}{\epsilon_1^2 + 2\epsilon_1(\zeta+\eta) - L^2(\delta^2 + L^2\Gamma_\varepsilon^2\epsilon^2)}\right). \tag{34}$$

This completes the proof of **Theorem 3**.

C.1 Simplified Theorem 3

Corollary 1. *Using first order Taylor series, the upper bound in* ***Theorem 3*** *becomes* $\mathcal{O}(\frac{l(\theta_0)-l^*}{\epsilon_1^2+\epsilon_1(\zeta+\eta)})$.

Using the similar arguments of **Theorem 3**, and taking first-order Taylor's approximation, we get

$$\begin{aligned} l(\theta_{t+1}) &= l(\theta_t) - h_t||\nabla l(\theta_t)||^2 - h_t||\nabla l(\theta_t)||.||\nabla S_{C,\varepsilon}(\mu_{\theta_t}, \nu)|| - h_t||\nabla l(\theta_t)||.||g(\psi; \mathbf{G}_{\theta_t}(x))|| \\ &\leq l(\theta_t) - h_t\epsilon_1^2 - h_t\epsilon_1\eta - h_t\epsilon_1\zeta \end{aligned} \tag{35}$$

Taking telescopic sum over t for $t \leq T$, we get

$$\sum_{t=0}^{T-1} l(\theta_{t+1}) - l(\theta_t) \leq -Th_t(\epsilon_1^2 + \epsilon_1(\zeta + \eta)) \tag{36}$$

So, using the definition of iteration complexity, we get,

$$\sup_{\theta_0 \in \{\mathbb{R}^{h \times d_x}, \mathbb{R}^{d_y \times h}\}, l \in \mathscr{L}} \mathcal{T}_{\epsilon_1}(A_h[l, \theta_0], l) = \mathcal{O}\left(\frac{l(\theta_0) - l^*}{\epsilon_1^2 + \epsilon_1(\zeta + \eta)}\right) \tag{37}$$

This completes the proof. □

Since the denominator of the derived upper bound in **Theorem 3** is greater than the one mentioned in Theorem 3 by Rout (2020), we can infer that our proposed learning framework has tighter iteration complexity compared to the regular adversarial setup. This is only true when $\Gamma_\varepsilon < \frac{\sqrt{2\epsilon_1 \eta}}{L^2 \epsilon}$, and it can be easily satisfied by appropriate choice of ε. It can also be verified using a simpler setup as shown in **Corollary 1**, as it increases the convergence rate from $\mathcal{O}((\epsilon_1^2 + \epsilon_1\zeta)^{-1})$ (Rout, 2020) to $\mathcal{O}((\epsilon_1^2 + \epsilon_1(\zeta + \eta))^{-1})$ which effects considerably in the training iterations. This also suggests that our setup is equivalent to applying two discriminator operations without involving additional computations which helps in effectively leveraging the advantage of using multiple discriminators (Durugkar et al., 2016). Experiment result to support this proposed theorem is carried out in Appendix §G.

D Appendix: Details of used losses

Besides classification, discriminator $\mathbf{D}$ has another major functional branch, ($\mathbf{D_{SA}}$), to approximate the spatial attention maps. For any input m, $\mathbf{D_{SA}}$ is used to estimate its corresponding normalized spatial feature maps, $D_{SA} : \mathbb{R}^{H \times W} \rightarrow [0, 1]^{H \times W}$. Let $\mathbf{D}$ consists of k DMRBs and a_i be the activation maps after i^{th} DMRB with c output channels, such that $a_i \in \mathbb{R}^{H \times W \times c}$. Since at different depths, the discriminator focuses on different features, we select k different attention maps from the layers in the latent space. Eventually, attention coefficients are calculated as in Emami et al. (2019):

$$\mathbf{D}_{SA}(m) = \sum_{i=1}^{k} \sum_{j=1}^{c} |a_{ij}(m)| \tag{38}$$

Then the domain adaptation loss ($\mathscr{L}_{DA}$) mentioned in §2.2,is defined as,

$$\mathscr{L}_{DA} = \mathbb{E}_{\tilde{x} \sim \mathbb{P}_{\tilde{x}}, y \sim \mathbb{P}_y} \left[\|\mathbf{D}_{SA}(\tilde{x}) - \mathbf{D}_{SA}(y)\|_2^2 \right]. \tag{39}$$

where, y is ground truth DEM and $\tilde{x}$ is bicubic interpolated coarse SRTM DEM as mentioned in §2. The pixel loss ($\mathscr{L}_P$) and SSIM loss ($\mathscr{L}_{str}$) and adversarial loss ($\mathscr{L}_{ADV}$) are defined as,

$$\begin{aligned} \mathscr{L}_P &= \mathbb{E}_{\tilde{x} \sim \mathbb{P}_{\tilde{x}}, z \sim \mathbb{P}_Z, y \sim \mathbb{P}_y} \left[\|y - \mathbf{G}(\tilde{x}, z \odot A_s(\tilde{x}))\|_2^2 \right], \\ \mathscr{L}_{str} &= \mathbb{E}_{\tilde{x} \sim \mathbb{P}_{\tilde{x}}, z \sim \mathbb{P}_Z, y \sim \mathbb{P}_y} - \log(\mathbf{SSIM}(\mathbf{G}(\tilde{x}, z \odot A_s(\tilde{x})), y)), \\ \mathscr{L}_{ADV} &= \mathbb{E}_{\tilde{x} \sim \mathbb{P}_{\tilde{x}}, z \sim \mathbb{P}_Z} - \log(\mathbf{D}(\mathbf{G}(\tilde{x}, z \odot A_s(\tilde{x})))). \end{aligned} \tag{40}$$

where, $A_s(\tilde{x}) = PSA(\mathbf{D}_{SA}(\tilde{x}))$ with PSA being polarized self-attention as discussed in §2

E Appendix: Implementation Details

Every experiment is conducted under identical environments. We use 3×3 convolution kernel and leaky ReLU activation except in global skip connection where 1×1 kernel is used without activation. Each DMRB has 64 convolution operations. We use ADAM optimizer with a fixed learning rate of 0.0001. During adversarial training, we update the critic once every single update in the generator. We set $\lambda_{DA} = 0.1$, $\lambda_P = 100$, $\lambda_{str} = 1$, $\lambda_{ADV} = 1$ and $\lambda_{OT} = 0.01$. For estimating $\mathscr{L}_{OT}$, we set $T = 10$ and $\varepsilon = 0.1$. The entire framework is developed using PyTorch

F Appendix: Ablation study

In this section, we analyze the efficacy of our four proposed modules, image prior, spatial attention from discriminator, PSA, and Sinkhorn loss-based adversarial learning. In the absence of any prior,

Table 2: Quantitative analysis for the effect of introducing different modules for DEM super-resolution

Image Prior	Spatial Attention	PSA	Sinkhorn loss	RMSE (m)	MAE (m)	SSIM (%)	PSNR
✗	✗	✗	✗	16.54	13.63	72.27	30.25
✓	✗	✗	✗	29.32	25.41	78.29	28.25
✓	✓	✗	✗	20.76	18.29	81.68	31.08
✓	✓	✓	✗	18.76	15.13	85.04	32.21
✓	✓	✓	✓	**9.28**	**8.51**	**90.49**	**35.06**

it leads to outcomes similar to bicubic interpolation. This results in low RMSE and MAE, compared to other cases as shown in Table 2. However, that affects its SSIM score as shown in Table 2. Using high-resolution MX image priors resolves this issue to a certain extent by increasing SSIM by more than 6%. Yet, this degrades the performance of the model in terms of RMSE and MAE. To improve upon this situation, we introduce spatial feature maps from the discriminator. Figure 5 shows how individual attention maps after each DMRB prioritizes certain features at different labels. However, the mean attention weights are approximately uniform as shown in Figure 6. PSA handles this matter by emphasizing key features. Addressing both these attentions improves the RMSE and MAE by more than 50% as well as enhances SSIM by almost 7% as shown in Table 2. The introduction of Sinkhorn distance regularization enhances the evaluation parameters further with both MAE and RMSE below 10 m and SSIM more than 90% as shown in Table 2. Apart from this, sinkhorn loss also contributed to a near 2.5X faster convergence rate for the pixel loss as shown in Figure 7. This also supports our argument in **Theorem 3**. We also perform ablation study related to different

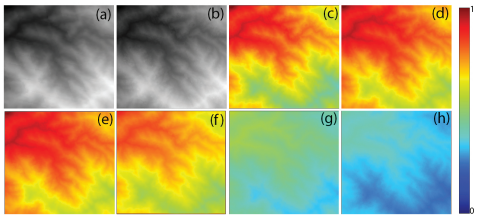

Figure 5: (a) Source, (b) Target, (c)-(h) Discriminator spatial attention after each DMRB from top to bottom.

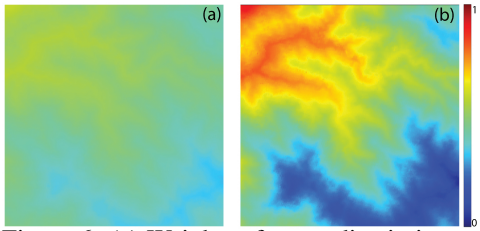

Figure 6: (a) Weights of mean discriminator spatial attention ($\mathbf{D}_{SA}$), (b) weights after passing $\mathbf{D}_{SA}$ through PSA block)

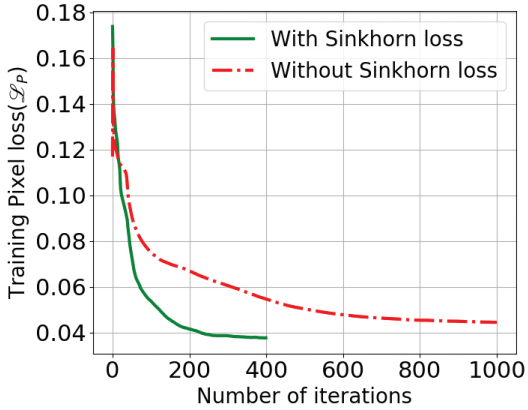

Figure 7: Effect of sinkhorn loss in training convergence for our model

loss functions as shown in Table 3. This can also be visualized from 8. Clearly, introduction of Sinkhorn loss is pivotal in SR performance (3 dB gain). Other losses also contribute to the overall performance.

Table 3: Quantitative analysis on the effect of different losses

Pixel loss	SSIM loss	Adversarial Loss	Sinkhorn loss	RMSE (m)	MAE (m)	SSIM (%)	PSNR
✓	✗	✗	✗	20.68	19.03	74.43	31.07
✓	✓	✗	✗	25.76	22.83	81.55	31.41
✓	✓	✓	✗	18.76	15.13	85.04	32.21
✓	✓	✓	✓	**9.28**	**8.51**	**90.49**	**35.06**

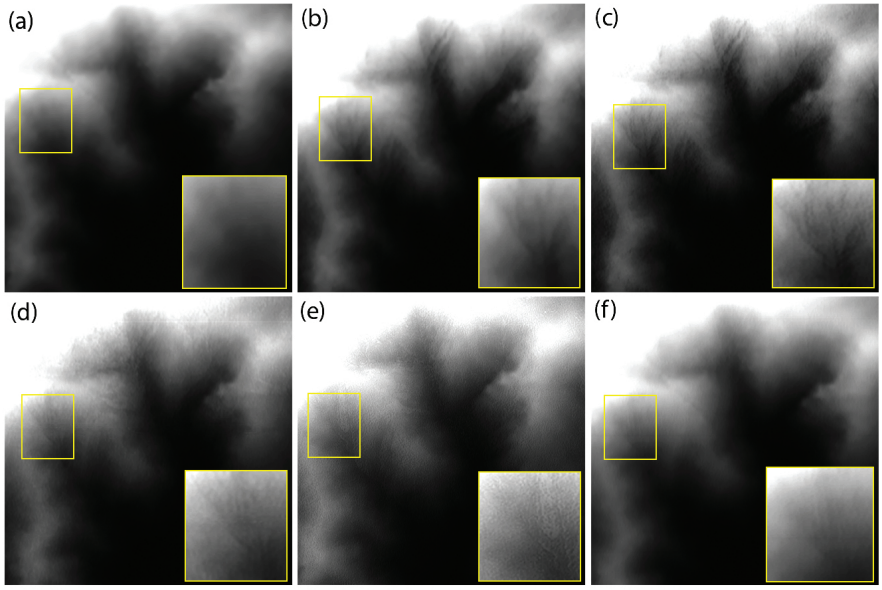

Figure 8: (a) Bicubic LR DEM, (b) HR GT DEM, predicted HR DEM corresponding to (c) all losses, (d) pixel + SSIM + adversarial loss, (e) pixel + SSIM loss, and (f) only pixel loss.

G Appendix: Empirical results related to Theorem 2 and Theorem 3

We perform experiments to answer the proposed claims. There are two main aspects we want to investigate, firstly, how the choice of ε affects the overall training of the model, and secondly, how it performs compared to other state-of-the-art learning methods like WGAN, WGAN+GP, and DCGAN. In both these cases, we analyze the claims of mitigating vanishing gradients in the near-optimal region and fast convergence rate.

G.1 Experiment set-up

In this setting, we are performing a denoising operation on the MNIST dataset. For this 60000 samples of size 28×28 are used during training, while 10000 are used for testing. The convergence criterion is set to be the mean square error of 0.04 or a maximum of 500 epochs. During training, we randomly add Gaussian noise to the training samples to perform the denoising task. The generator is designed as a simple autoencoder structure with an encoder and decoder each having 2 convolutional layers. In practice, we notice that a discriminator with shallow layers is usually sufficient to offer a higher convergence rate. Therefore, we choose, a three-layer fully connected network with 1024 and 256 hidden neurons. All the layers are followed by ReLu activation except the output layer. For optimization, ADAM is utilized with a learning rate of 0.001 with a batch size of 64, and the discriminator is updated once for every single update of the generator.

G.2 Result analysis

Figure 9, shows how changing the value of ε affects the overall iteration complexity. According to this figure, the instances ε are very small and very large, and the learning behavior of the model becomes close to regular adversarial setup which ultimately results in more time requirement for convergence. This is because, as $\varepsilon \to 0$ and $\varepsilon \to \infty$, the smoothness of sinkhorn loss tends to become independent of ε as depicted in **Theorem 1**, which makes the overall setup similar to the regular adversarial framework. This also affects the capability of mitigating the vanishing gradient

problem as shown in Figure 10 and 11. The gradients are approximated using spectral norm and they are moving averaged for better visualization. From Figure 10, in the case of the first layer, as ε varies, the estimated gradients are similar near the optimal region. However, From Figure 11, we can see for the case of the hidden layer, gradient approximation is definitely affected by the choice of ε, and we can see as $\varepsilon \to 0$ and $\varepsilon \to \infty$, the gradients near-optimal region become smaller. However, using $\varepsilon = 0.1$ tends to have higher gradients even if near the optimal region. Therefore, this model will have more capability of mitigating the vanishing gradient problem. Hence, we use this model to compare with other state-of-the-art learning methods.

We compare the rate of convergence and capability of handling the vanishing gradient of SIRAN with WGAN (Arjovsky et al., 2017), WGAN+GP (Gulrajani et al., 2017), and DCGAN. Figure 12 clearly visualizes how our proposed framework has tighter iteration complexity than others, and reaches the convergence faster. This is consistent with the theoretical analysis presented in **Theorem 3**. Figure 13 and 14 also provides empirical evidence of the vanishing gradient issue presented in **Theorem 2**. Both for the first layer and hidden layer, as shown in Figure 13 and 14, the approximated gradients are higher comparatively than others near the optimal region. This results in increasing the effectiveness of SIRAN in handling the issue of the vanishing gradient problem as discussed in above theorems.

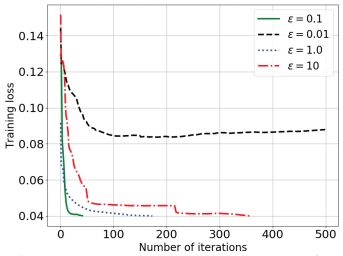

Figure 9: Training Loss for variation of ε

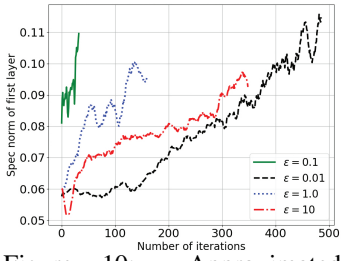

Figure 10: Approximated Spectral norm of gradients of first layer for different values of ε

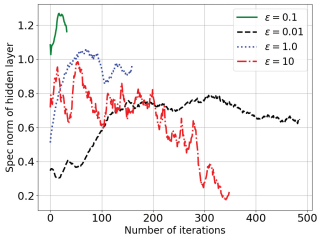

Figure 11: Approximated Spectral norm of gradients of hidden layer for different values of ε

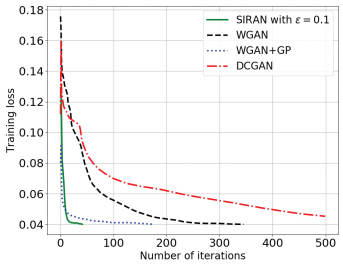

Figure 12: Training Loss for different learning methods

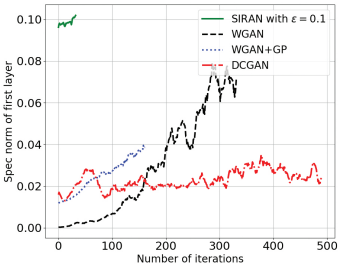

Figure 13: Approximated Spectral norm of gradients of first layer for different learning methods

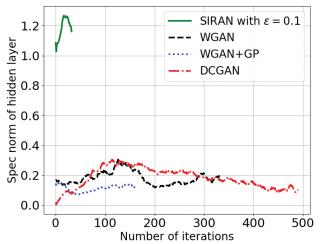

Figure 14: Approximated Spectral norm of gradients of hidden layer for different learning methods