# DR-SCAN: AN INTERPRETABLE DUAL-BRANCH RESIDUAL SPATIAL AND CHANNEL ATTENTION NETWORK FOR REMOTE SENSING AND GEOSCIENCE IMAGE SUPER-RESOLUTION

**Suraj Neelakantan, Martin Längkvist & Amy Loutfi**
Center for Applied Autonomous Sensor Systems, Örebro University,
Fakultetsgatan 1, 701 82 Örebro, Sweden
{firstname.lastname}@oru.se

## ABSTRACT

High-resolution imaging is essential in remote sensing and geoscience for precise environmental and geological analysis. DR-SCAN (Dual-Branch Residual Spatial and Channel Attention Networks), a neural network architecture for image super-resolution across these domains, is introduced. Evaluated on the UCMerced Land Use and DeepRock-SR datasets, DR-SCAN demonstrates adaptability to diverse remote sensing landscapes and effectiveness in resolving pore-scale geological features. Feature map visualizations highlight the model's ability to prioritize critical spatial features, enhancing interpretability for domain-specific applications.

## 1 INTRODUCTION

High-resolution (HR) images are used in various remote sensing and geoscience tasks ranging from regional geological mapping (Kruse et al., 2003), environmental monitoring (Tucker & Townshend, 2000), and precision farming (Jin et al., 2019), to name a few. However, hardware and environmental constraints limit the resolution of the images collected. Single-image super-resolution (SISR), is a process of reconstruction of an HR image using the low-resolution (LR) counterpart, thus enhancing the use of images for downstream scientific and operational analyses.

In this work, DR-SCAN, a novel dual-branch architecture for SISR tailored to remote sensing and geoscience applications is presented. Moving away from the traditional single-branch neural network architectures that learn local details (e.g. edges, fine grained-textures) and global structures (e.g. terrain features) together in a shared feature space, our DR-SCAN architecture separates these components into two specialized branches and fuses them later on. Specifically, a **shallow branch** dedicated to learn high-frequency details and a **deep branch** intended to capture a broader spatial context. Subsequently, these complementary outputs are fused, producing super-resolution (SR) images. Notably, DR-SCAN also provides a mechanism for a **user-driven adaptation**, allowing the user to modulate how much emphasis should be given to local and global features, an attribute that could be particularly valuable while reconstructing remote-sensing and geoscience images. Furthermore, by visualizing the feature maps of these two branches, an **interpretable evaluation** can be made so that domain experts can better understand the SR process.

Earlier interpolation-based methods (e.g. bicubic) often produce blurry outputs (Wang et al., 2020), while early deep learning (DL) models (e.g. SRCNN (Dong et al., 2014), VDSR (Kim et al., 2016)) and deeper residual networks (e.g. EDSR (Lim et al., 2017)) improved SR but relied only on pixel-wise loss functions (e.g L1, MSE), leading to smooth textures and loss of high-frequency details in some cases (Wang et al., 2022). Later, generative adversarial networks (GANs) were used for SISR (e.g., SRGAN (Ledig et al., 2017), ESRGAN (Wang et al., 2018)), enabling the restoration of sharper details using perceptual and adversarial losses. However, GANs can be challenging to train. Attention-based networks like RCAN (Zhang et al., 2018) further improved feature selection by weighting channel-wise and spatial-wise features. However, without explicitly **disentangling** small- and large-scale features, such architectures may lack adaptability to various remote sensing

tasks. In contrast, DR-SCAN separates and fuses these features, providing a more flexible and interpretable framework for SISR in remote sensing and geoscience.

## 2 METHODOLOGY

Let $\mathbf{I}_{LR} \in \mathbb{R}^{C\times H\times W}$ be the LR input and $\mathbf{I}_{HR} \in \mathbb{R}^{C\times \alpha H\times \alpha W}$ be the HR ground truth, where $C$ is the number of channels, $H, W$ are the spatial dimensions, and $\alpha$ is the scale factor. The goal is to learn a function $f_\theta$ parameterized by $\theta$, mapping $\mathbf{I}_{LR}$ to an estimate $\hat{\mathbf{I}}_{HR} \approx \mathbf{I}_{HR}$. $f_\theta : \mathbf{I}_{LR} \to \hat{\mathbf{I}}_{HR}$.

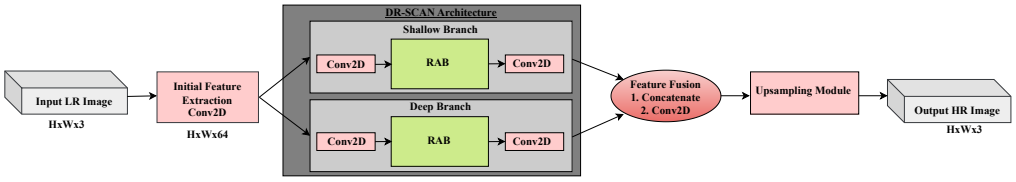


Figure 1: Proposed DR-SCAN architecture showing the dual-branch design: a shallow path with residual attention blocks and a deep path with hierarchical residual groups each containing multiple residual attention blocks.

The proposed DR-SCAN network (Figure 1) consists of a shallow branch $S_\phi$ and a deep branch $D_\psi$, where the shallow branch extracts fine details (e.g. edges, textures), while the deep branch captures global structures: $\mathbf{F}_{\text{shallow}} = S_\phi(\mathbf{I}_{LR})$, $\mathbf{F}_{\text{deep}} = D_\psi(\mathbf{I}_{LR})$. These features are combined using a fusion operator $\mathcal{F}$, yielding: $\mathbf{F}_{\text{fused}} = \mathcal{F}(\mathbf{F}_{\text{shallow}}, \mathbf{F}_{\text{deep}})$. The final HR image is reconstructed using an upsampling module followed by a bicubic-upsampled skip connection: $\hat{\mathbf{I}}_{HR} = \text{Upsample}(\mathbf{F}_{\text{fused}}) + \text{Bicubic}(\mathbf{I}_{LR})$.

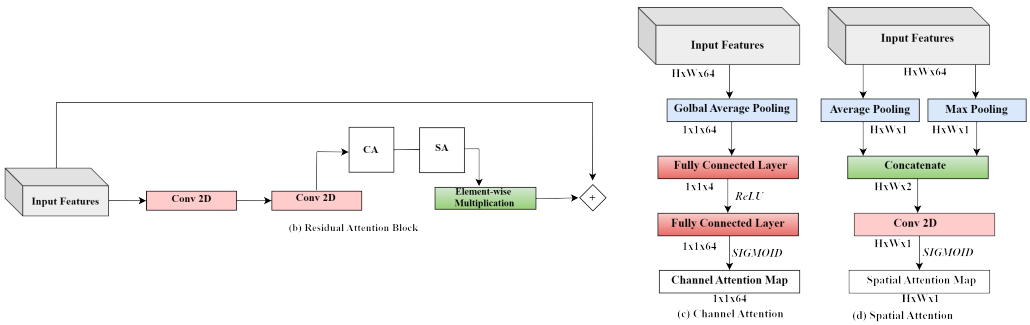


Figure 2: Schematic representation of the residual attention block incorporating the two attention modules: channel attention for inter-channel dependencies via global average pooling and spatial attention for spatial dependencies via concatenated average-max pooling operations.

Each branch consists of ***residual attention block*** (RABs) (Figure 2(a)), where an input feature map $\mathbf{x} \in \mathbb{R}^{C\times H'\times W'}$ is refined to produce $\mathbf{x}'$ using: $\mathbf{x}' = \mathbf{x} + \gamma \cdot \text{AttnConv}(\mathbf{x})$, where $\gamma$ is a learnable scaling factor, and $\text{AttnConv}$ applies channel and spatial attention.

The ***channel attention*** (CA) (Figure 2(b)), emphasizes *what* features are important. Unlike CBAM (Woo et al., 2018), which applies both average and max pooling, only **global average pooling** (GAP) is used to reduce computational cost: $\mathbf{F}_{\text{avg}}^{(c)} = \text{GlobalAvgPool}(\mathbf{x})$, $\mathbf{F}_{\text{avg}}^{(c)} \in \mathbb{R}^C$. This is the passed through a multi-layer perceptron (MLP) with a reduction ratio $r$: $\mathbf{w}_c = \sigma\Big(\mathbf{W}_1\,\delta\big(\mathbf{W}_0\,\mathbf{F}_{\text{avg}}^{(c)}\big)\Big)$, where $\delta$ is ReLU activation function, $\sigma$ is sigmoid activation function, and $\mathbf{W}_0 \in \mathbb{R}^{\frac{C}{r}\times C}$, $\mathbf{W}_1 \in \mathbb{R}^{C\times\frac{C}{r}}$ are learnable weights. The final channel-wise reweighting is applied via: $\mathbf{x}_{\text{ca}} = \mathbf{x} \odot \mathbf{w}_c$.

The ***spatial attention*** (SA) (Figure 2(c)), determines *where* features are important by first applying average and max pooling across channels: $\mathbf{F}_{\text{avg}}^{(s)} = \text{AvgPool}(\mathbf{x}), \mathbf{F}_{\text{max}}^{(s)} = \text{MaxPool}(\mathbf{x})$. These descriptors are then concatenated and passed through a convolution layer:

$\mathbf{M}_s(\mathbf{x}) = \sigma\Big(f_{k\times k}([\mathbf{F}_{\text{avg}}^{(s)}; \mathbf{F}_{\text{max}}^{(s)}])\Big)$, where $f_{k\times k}$ is a convolution layer of kernel size $k \times k$. The resulting spatial attention map is applied to the input feature map as: $\mathbf{x}_{\text{sa}} = \mathbf{x} \odot \mathbf{M}_s(\mathbf{x})$

The fused features from the shallow and the deep branches are then passed through an upsampling module. A final convolutional layer and a skip connection using bicubic upsampling to stabalize learning in deeper architectures to produce $\hat{\mathbf{I}}_{HR}$, thus completing the DR-SCAN methodology.

The UCMerced and DeepRock-SR datasets are used for remote sensing and geoscience experiments, respectively. PSNR and SSIM (Wang, 2004) are the evaluation metrics. Data augmentation includes random rotations and flips. The network is trained using AdamW optimizer following a cosine annealing schedule. All experiments are implemented in PyTorch and trained on NVIDIA A100 GPUs with batch sizes of 32/16.

## 3 Results and Discussions

### 3.1 Remote Sensing - UCMerced dataset

| Scale | Bicubic | SC | SRCNN | FSRCNN | CNN-7 | LGCNet | DCM | DRCM | DR-SCAN(Ours) |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 30.76/0.8789 | 32.77/0.9166 | 32.84/0.9152 | 33.18/0.9196 | 33.15/0.9191 | 33.48/0.9235 | 33.65/0.9274 | 34.37/0.9296 | **35.07/0.9356** |
| 3 | 27.46/0.7631 | 28.28/0.7971 | 28.66/0.8038 | 29.09/0.8167 | 29.02/0.8155 | 29.28/0.8238 | 29.52/0.8394 | 30.26/0.8507 | **30.81**/0.8498 |
| 4 | 25.65/0.6725 | 26.51/0.7152 | 26.78/0.7219 | 26.93/0.7267 | 26.86/0.7264 | 27.02/0.7333 | 27.22/0.7528 | 27.88/0.7707 | 27.86/**0.7885** |

Table 1: Comparison of **PSNR (dB)/ SSIM** for different SR methods across various scaling factors. Bold numbers indicate the highest values.

Table 1 compares the performance of DR-SCAN with other SR methods by different scaling factors using PSNR and SSIM metrics in the UCMerced dataset. DR-SCAN outperforms traditional methods like Bicubic and other CNN-based approaches and achieves the highest PSNR and competitive SSIM scores, particularly for $\times 2$ and $\times 3$ scales. Figure 3 shows an example of a downscaled image of $\times$ 4 and is SR output from DR-SCAN.

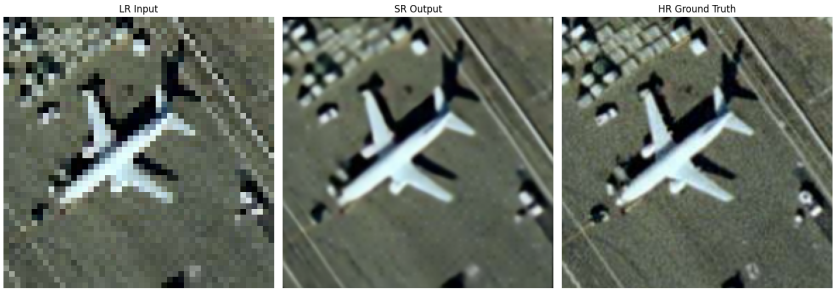


Figure 3: Example of qualitative analysis of DR-SCAN on $\times 4$ downscaled image from UCMerced dataset. From left to right: LR input, SR output using DR-SCAN, and HR ground truth.

### 3.2 Geoscience - DeepRockSR

| Method | Year | Carbonate PSNR | Sandstone PSNR | Carbonate SSIM | Sandstone SSIM |
|---|---|---|---|---|---|
| SRCNN | 2014 | 31.4669 | 34.4151 | 0.8691 | 0.8684 |
| EDSR | 2017 | 31.5464 | 35.1065 | 0.8700 | 0.8697 |
| SRResNet | 2017 | 31.5086 | 34.9821 | 0.8694 | 0.8692 |
| SwinIR-light | 2021 | 31.5030 | 34.7320 | 0.8696 | 0.8689 |
| MAN | 2023 | 31.5519 | 35.1475 | 0.8701 | 0.8703 |
| SAFMN | 2023 | 30.2270 | 33.7771 | 0.8611 | 0.8575 |
| MDBN | 2024 | 31.5415 | 35.0539 | 0.8700 | 0.8696 |
| TDFIF-Net | 2023 | 31.5573 | 35.1636 | 0.8702 | 0.8703 |
| DR-SCAN (Ours) | 2024 | 31.5093 | 35.0830 | 0.8693 | 0.8669 |

Table 2: Comparison on Carbonate2D and Sandstone2D datasets for scales $\times 2$ (Zhang et al., 2024).

Comparing different SR on the DeepRock-SR dataset in Table 2, DR-SCAN shows competitive performance for both carbonate and sandstone samples. The minimal performance differences (≈0.1 dB) among recent methods indicate that current DL approaches have likely reached near-optimal performance for $\times 2$ SR.

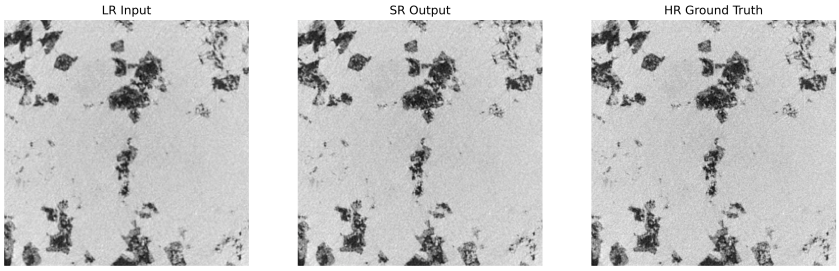


Figure 4: Visual comparison of SR results on a carbonate microstructure image. From left to right: LR input, SR output using DR-SCAN, and HR ground truth.

### 3.3 Interpretability via Dual-Branch Feature Maps

To illustrate where each branch has the highest activation, the final outputs of both branches and compute the average across the channel dimension is extracted. Given a batch of feature maps $F \in \mathbb{R}^{B\times C\times H\times W}$, where $B$ is the batch size, $C$ is the number of channels, and $H, W$ are the spatial dimensions, the average branch activation is computed as: $\text{AvgMap}(F) = \frac{1}{C}\sum_{c=1}^{C} F^{(c)}$.

The shallow and deep branch activations, along with their differences, are then plotted. These plots for two sample outputs from the test set are shown in Figure 5.

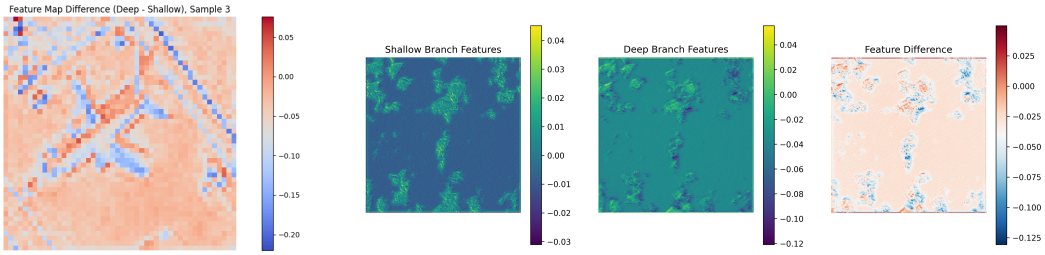


Figure 5: Visualization of feature map difference showing deep vs shallow branch activations: (Left) UC Merced dataset; (Right) DeepRockSR dataset showing each branch feature maps and their difference map.

These visualizations offer interpretability by revealing how the network learns both local features (edges) and higher-level structure within the same network. In Figure 5 area under red color suggest the deep branch has the highest activation, whereas blue colored area around the edges indicate regions dominated by shallow branch activations. Thus, the dual-branch framework provides a more transparent view of the reconstruction process and reinforces the architectural design rationale for SISR.

## 4 Conclusion

In this paper DR-SCAN, a dual-branch network for SISR was introduced. Our quantitative evaluations and visual comparisons demonstrate that DR-SCAN performs competitively against state-of-the-art methods. Through feature visualization, the dual branches split their focus between fine details and structural elements, validating our architectural design. Future work includes dynamic weighting of each branch in the DR-SCAN architecture so that one can adaptively balance the contributions of fine-detail enhancement and structural reconstruction based on image content, potentially improving performance across diverse datasets and applications.

## 5 ACKNOWLEDGEMENTS

This work has been supported by the Industrial Graduate School Collaborative AI and Robotics funded by the Swedish Knowledge Foundation Dnr:20190128 and in collaboration with the industrial partner Orexplore Technologies.

## REFERENCES

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, pp. 184–199. Springer International Publishing, 2014.

Shichao Jin, Yanjun Su, Shang Gao, Fangfang Wu, Qin Ma, Kexin Xu, Tianyu Hu, Jin Liu, Shuxin Pang, Hongcan Guan, et al. Separating the structural components of maize for field phenotyping using terrestrial lidar data and deep convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 58(4):2644–2658, 2019.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1646–1654, 2016.

Fred A Kruse, Joseph W Boardman, and Jonathan F Huntington. Comparison of airborne hyperspectral data and eo-1 hyperion for mineral mapping. *IEEE transactions on Geoscience and Remote Sensing*, 41(6):1388–1400, 2003.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, 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*, pp. 4681–4690, 2017.

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*, pp. 136–144, 2017.

Compton J Tucker and John RG Townshend. Strategies for monitoring tropical deforestation using satellite data. *International Journal of Remote Sensing*, 21(6-7):1461–1471, 2000.

Peijuan Wang, Bulent Bayram, and Elif Sertel. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Science Reviews*, 232:104110, 2022.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

Ying Da Wang, Ryan T Armstrong, and Peyman Mostaghimi. Boosting resolution and recovering texture of 2d and 3d micro-ct images with deep learning. *Water Resources Research*, 56(1): e2019WR026052, 2020.

Z Wang. Image quality assessment: Form error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):604–606, 2004.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Yubo Zhang, Chao Han, Lei Xu, Junhao Bi, Haihua Kong, and Haibin Xiang. Tdfif-net: Two-dimensional feature interaction fusion network for digital core images. *SSRN*, 2024.

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 286–301, 2018.