

# Supplementary Materials: Training Pansharpening Networks at Full Resolution Using Degenerate Invariance

Anonymous Authors

## 1 SUPPLEMENTARY MOTIVATION DETAILS

In this section, we review and outline the development of observation models and the assumption of scale invariance, which are the most relevant work to the proposed method, and are also closely related to our research motivation.

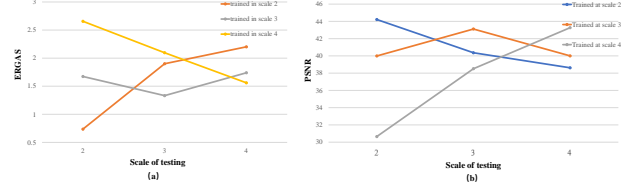
### 1.1 Observation Models

Observation models are the theoretical basis when implementing information transformation from high to low dimensions. In the field of pansharpening, observation models include spatial observation model (Ds fuction in Fig.2 of the paper) and spectral model (Fd fuction in Fig.2 of the paper). The former is a resolution degradation process from HR to LR images, while the latter is a modality degradation procedure from MS to PAN images. Detailed discussions of them are provided below.

In current methods, there are two mainstream assumptions for the spectral observation model to establish the link between MS and PAN images (as Eqs.4 and 5). In other words, these methods consider that the PAN image (or its gradient) can be modeled as a linear combination among all bands (or their gradients) of the MS image. Unfortunately, there is a significant difference in the response characteristics of MS and PAN sensors, as shown in Fig.4. As a result, above linear intensity assumption-based methods are difficult to accurately describe the transformation relationship between MS and PAN images in the intensity domain. Rather than choosing to continue to seek more accurate solutions for Eq.4, subsequent methods prefer to establish the transformation relationship in the gradient domain, as shown in Eq.5.

Since the gradients are relatively sparse, the modality differences between MS and PAN images can be greatly suppressed in the gradient domain, which makes the accurate solution of the transformation relationship in Eq.5 more possible. However, the wavelength range corresponding to PAN images is often wider than any channel of MS images, which means it is problematic to require each channel of the fused image to keep the gradient consistent with the PAN image.

Let us revisit the definition of the spectral observation model, which fundamentally captures the relationship between PAN and MS sensor spectral responses. This connection relates purely to sensor attributes, exhibiting invariance irrespective of scale or representation. However, a simple linear mapping fails to adequately characterize it. Under this conception, the gradient transformation assumed in Eq. 5 appears divergent from original intent. More precisely, the proper course is to pursue more precise intensity domain mapping from MS to PAN, encapsulating their richer dimensional correspondence. Modeling the end-to-end operator supersedes isolated assumptions by holistically optimizing network functions aligned with imaging process nuance. My self-supervised framework circumvents theoretical inconsistencies by faithfully resolving attributes inherently encoded across scales within native



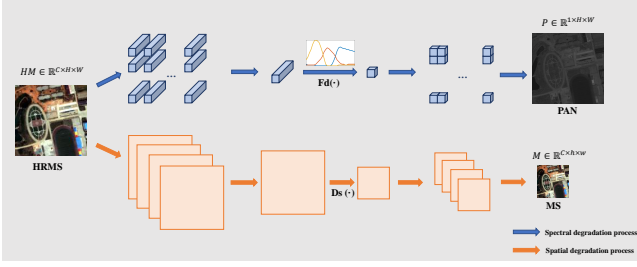
**Figure 1: Significant performance difference of pansharpening due to scale variance. (a) ERGAS indicates the relative dimensionless global error in synthesis, and a smaller value means better performance. (b) PSNR denotes Peak Signal-to-Noise Ratio, and a bigger value also means better performance**

observations. Both quantitative and qualitative results reinforce that this strategy better represents sensor-specific phenomena and facilitates reconstruction of finer landscape details. Continued exploration refining such data-driven techniques holds promise to further strengthen interpretability and fidelity between model interpretations and actual instrument behaviors. The overarching goal remains enabling robust exploitation of remote sensing data’s full high-dimensional potential.

### 1.2 Degenerate Invariance

In the field of computer vision, invariance is a common property prior, which usually refers to some objective laws that do not change within some limits. Wald protocol [1] provides a scheme to evaluate the pansharpening network, that is, using the test results of the generated data at the reduced resolution scale to judge the network performance. Subsequent methods have mostly experimented with this recommendation, thus introducing new scales. In this case, it is natural to discuss the scale invariance of the various steps in the task, in other words, we want to get a pansharpening network with good generalization over the scale range. For this reason, many people have conducted research and proposed countermeasures, but they still face some difficulties. So we propose to change the idea to solve this problem, the extra scale itself is introduced by us, if the training of the network can be limited to the original resolution, the requirement of scale invariance can be overcome naturally.

First, we verify through experiments that the scale change will actually affect the performance of pansharpening network, as shown in Fig.1. Specifically, we follow the degradation framework to perform three consecutive spatial downsampling for original MS2 and PAN1, producing MS and PAN images at different scales. Then, we adopt a representative framework to obtain three pansharpening models at Scales 2, 3, 4 under a supervised learning paradigm. The results show that the performance of the network decreases significantly with the change of scale. We still agree with the wald



**Figure 2: A detailed demonstration of the degradation process. We try to design a degradation process such as graphically coupled, that is, the spatial observation model and the spectral observation model are processed in different dimensions of the image, and such a design satisfies the desired exchange law of the degradation process.**

protocol that the evaluation of network performance can be tested on different scales. However, based on the above analysis, we believe that the experiment under full resolution is more in line with the task requirements.

$$Fd_2(Ds_1(HM)) = Ds_2(Fd_1(HM)). \quad (1)$$

Review several degradation processes under the original scale, It's easy to draw a simple inference as Equ.1. If  $Fd_1 = Fd_2$  holds, we hope the following boundary conditions be satisfied:

- The spatial observation models of MS and PAN images are the same, i.e.,  $Ds_1 = Ds_2$ .
- The spatial and spectral observation models obey the commutative property, i.e.,  $Fd_2(Ds_1(\cdot)) = Ds_2(Fd_1(\cdot))$ .

In fact, from the second commutation law it can be inferred that the degenerate processes are equal, and if we can decouple the two degenerate processes, the commutation law can be satisfied naturally, as shown in Fig.2.

## 2 DATASET DETAILS

In this section, we go through the building process and details of the dataset.

There are two datasets used in our experiments, including WorldView II and GF-2. The spatial resolutions of panchromatic images in these two satellites are 0.5m and 0.8m, respectively, while the spatial resolutions of their corresponding LRMS images are 1.8 m and 3.2 m with four bands including red, green, blue and near-infrared. For each used dataset, it is divided into three parts. Besides, the expansion strategy of tailoring and decomposition is applied for performing data augmentation on training and validation sets. After removing images that produce large black areas due to registration effects, the detail information is shown in the Tables.1 and 2. Notably, the reported numbers of training and validation data is computed after performing data augmentation. We crop the panchromatic and LRMS images into 60,000 image patch pairs of sizes  $128 \times 128$  and  $32 \times 32$ , respectively, and then randomly split them into 90% and 10% as our training data and validation data, respectively.

**Table 1: The basic information for each reduced resolution dataset.**

Datasets	WordView2	GaoFen2
Train/Test	768/80	2720/208
PAN	128×128	128×128
LRMS	32×32×4	32×32×4
HRMS	128×128×4	128×128×4

**Table 2: The basic information for each full resolution dataset.**

Datasets	WordView2	GaoFen2
Train/Test	2779/198	2792/200
PAN	512×512	512×512
LRMS	128×128×4	128×128×4

Specifically, we downsample the original MS image in WorldView-2 by a factor of 2, so as to extend the resolution ratio from the original 4 to 8. Thus, in the simulation dataset, the spatial resolutions of the MS images are 3.68 m, and the spatial resolutions of the MS images are 0.46 m. For each used dataset, it is divided into three parts. Besides, the expansion strategy of tailoring and decomposition is applied for performing data augmentation on training and validation sets. After removing images that produce large black areas due to registration effects, the detail information is shown in the Table 1. Notably, the reported numbers of training and validation data is computed after performing data augmentation

## 3 SUPPLEMENTARY EXPERIMENTAL RESULTS

In this section, we present more experimental results and comparison graphs to more clearly demonstrate the significant promotion of our method in generating high-quality raw scale HRMS images on the network. In Fig.3, we chose an image with obvious noise, in the snowy scene, the presence of rain and fog to form the foreground in front of the actual ground scene. At this time, when the original model trained under reduced resolution is observed, the restoration of the texture structure of the scene is greatly interfered by the noise interference. In contrast, the networks trained using our method show good quality at this point, and the recovery of objects' edges and details appears much clearer.

In addition, we also selected the recovery of texture information and color information of more typical scenes on other data sets. As shown in Fig.4, the position of the red box we identified contains relatively more complex information of various types, and the comparison is more obvious in this scene. As can be seen clearly, in color, the two images on the right show significantly higher contrast compared to the left, and the edges and gaps between the buildings retain more sharp details.

Together with prior quantitative analyses, these visualizations reinforce the value of our self-supervised optimization paradigm. By more authentically representing image formation process attributes, it enables models to better resolve critical high-dimensional content

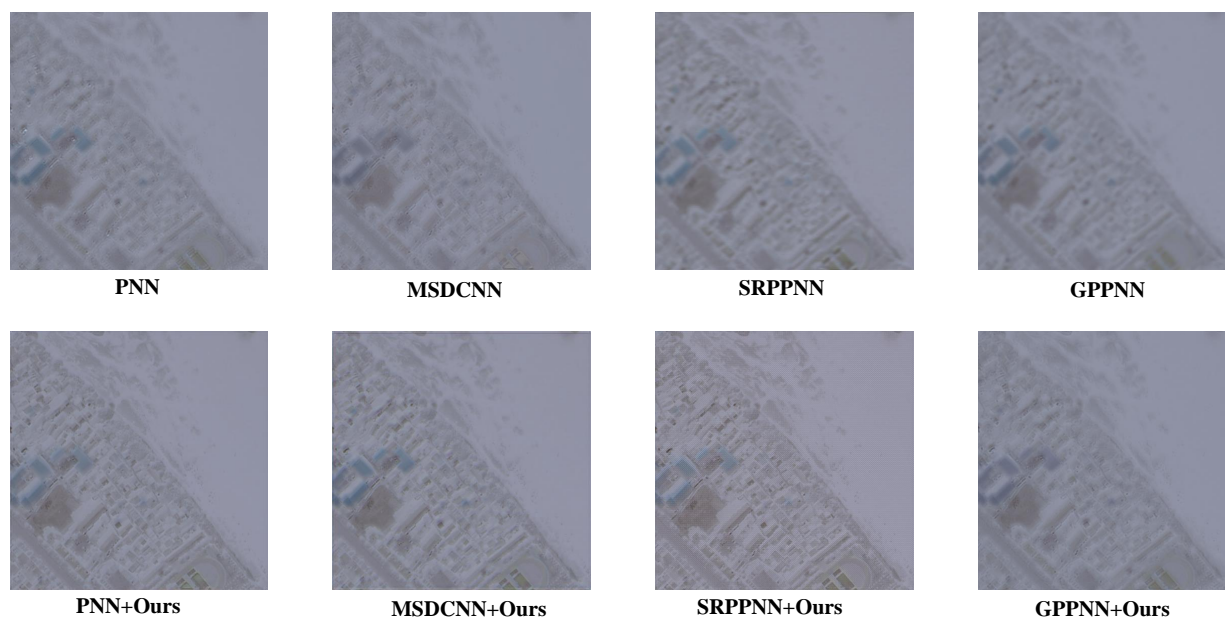


Figure 3: Visualize comparison of one sample image from the GaoFen2 dataset.

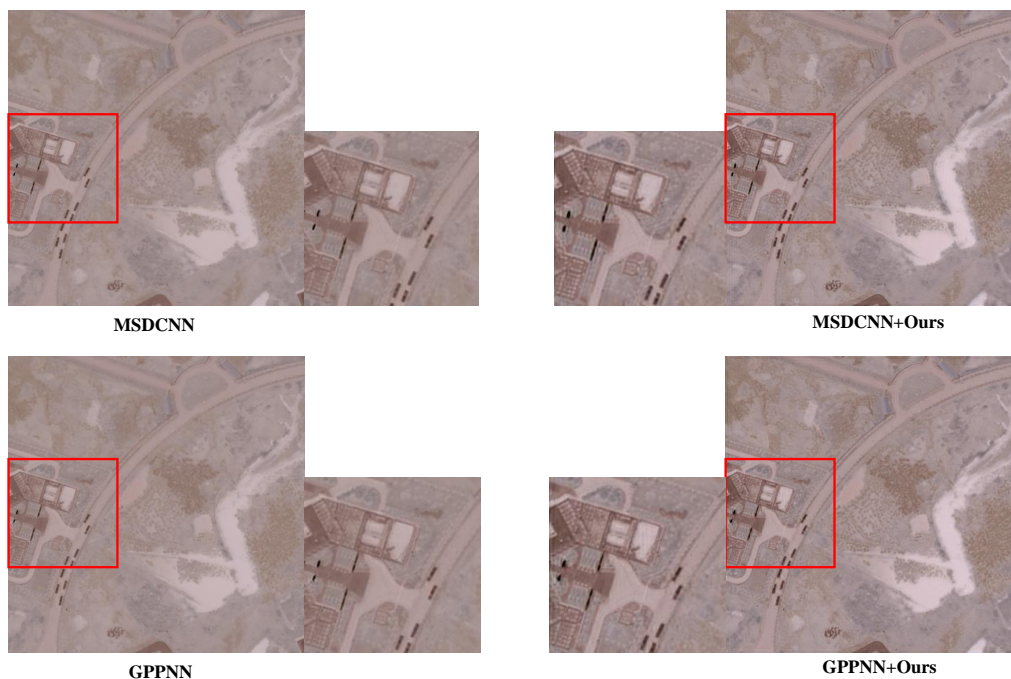


Figure 4: Visualize comparison of one sample image from the WorldView2 dataset.

inherently encoded across scales within remote sensing imagery collections. Continued study exploring this style of resolution-matched training holds promise to further deepen models’ interpretability and real-world alignment. The overarching goal is facilitating robust

exploitation of satellites’ full resolution potentials for applications in precision agriculture, infrastructure assessment and beyond.

REFERENCES

[1] Lucien Wald, Thierry Ranchin, and Marc Mangolini. 1997. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric engineering and remote sensing* 63, 6 (1997), 691–699.