"Dynamic Cross Feature Fusion for Remote Sensing Pansharpening": Supplementary Material

Xiao Wu [1], Ting-Zhu Huang [1]*, Liang-Jian Deng [1]*, Tian-Jing Zhang [2]
[1]School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]Yingcai Honors College, University of Electronic Science and Technology of China, Chengdu 611731, China
wxwsx1997@gmail.com; tingzhuhuang@126.com;
liangjian.deng@uestc.edu.cn; zhangtianjinguestc@163.com

Abstract

In this supplementary material, we show the supplementary qualitative and quantitative results of dynamic cross feature fusion network (DCFNet). In Sect. 1, we introduce a curve about training losses on WorldView-3. The experimental part has described the discussion about the effectiveness of the network. In Sect. 2.1, we put explanations on how to visualize DCFNet via Grad-CAM applied to pansharpening. In Sect. 2.2, we show visual qualities about feature maps and class activation maps, as well as the analyses of inter-branch fusions.

1. Convergence and optimal epoch

DCFNet possesses triple parallel branches and fuses multi-scale features via pyramid cross feature transfer (PCFT). DCFNet adopts Adam optimizer with the learning rate being 1×10^{-3}, coefficients of running averages of gradient, and its square being (0.9, 0.999). The batch size is set to 32. We train DCFNet in Pytorch 1.7.0 and a GPU NVIDIA GeForce RTX 3080 with 10GB (more details can be seen in the section of the experiment). From Fig. 1, we can know that DCFNet can fastly achieve better losses compared to other models on the WorldView-3 dataset. When our training losses are less than other losses, our DCFNet has outperformed well on many test datasets. Finally, we spent about 300 epochs training DCFNet to the optimal epoch, which means that for the next 100 epochs, the validation losses will no longer decrease.

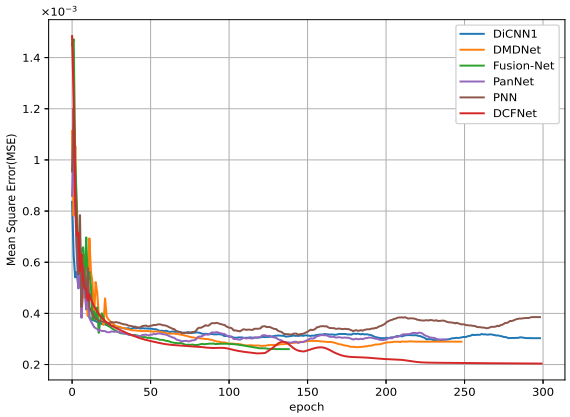


Figure 1: Convergence curves for all the compared CNN methods on WorldView-3 training dataset. According to [1], PNN [2] is trained with 300 epochs and FusionNet [1] is plotted with the first 139 epochs.

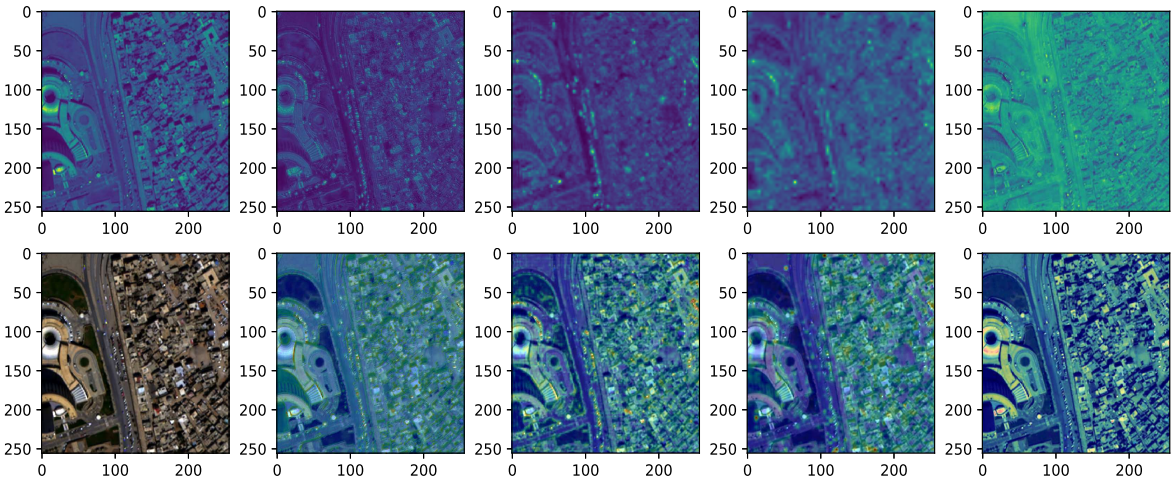
2. Visualization and analysis

2.1. Preliminary

In pansharpening, we fuse a low-resolution multispectral image and a high-resolution panchromatic image to generate a high-resolution multispectral image. In order to further investigate the effects of inter-branch fusions, we adopt Grad-CAM [3] to visualize the proposed DCFNet. Grad-CAM can be summarized as follows:

$$\alpha_k^c = \overbrace{\frac{1}{HW}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}, \tag{1}$$

*Corresponding author.



(a) Output image (b) Main branch (c) Medium branch (d) Low branch (e) Inter-branch fusion

Figure 2: First row is feature maps, whose channels are averaged and second row is class activation maps about the last Conv layer with gradients of cross-scale features concatenation on Tripoli dataset. (WorldView-3). Note that images of different resolutions are shown, so we reserve the axes for easier viewing.

α_k^c denotes the average value of gradients backpropagated from the output of feature map k for a target class c. Then, we apply gradients into feature maps (size: $C \times H \times W$).

$$Y^c = ReLU\left(\sum_k \alpha_k^c A_k\right), \tag{2}$$

A_k represents the feature map k, which is the output of a convolution layer. Y^c is class activation maps. Grad-CAM has one-to-one correspondences between the target class and the predicted class. Hence, we correspond a band to a class in the original paper, $c \in [1, 8]$. In this way, the class activation maps can represent the influences of the corresponding feature maps on the output image. We show class activation maps (see Fig. 2 and Fig. 3) of DCFNet. Figures are RGB images composed of 1, 3, and 5 of the 8 channels.

2.2. Visualization

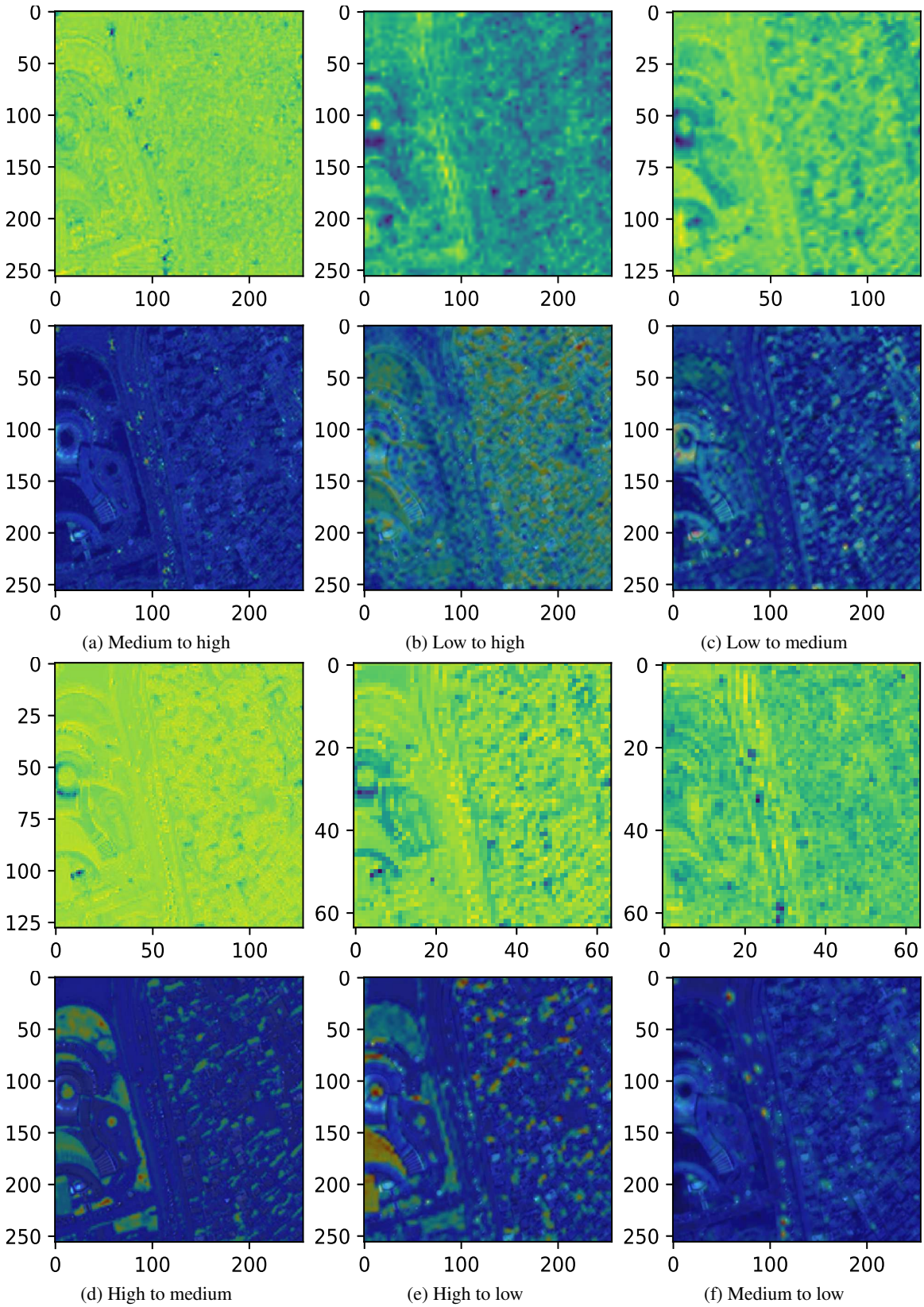
Grad-CAM is an approach of class activation map for intuitive explanations from any CNN-based network. So, we use Grad-CAM to visualize the last inter-branch fusion and the last PCFT, respectively. Visualization results are shown in Fig. 2. They indicate that the main branch is spatial reduction-free and has rich details. Notably, the average values of images in the main branch and images of high-to-low transfers are greater than other branches and low-to-high transfers, which reflects on the fact that information is supplemented between branches. Aiming to transfer the cross-scale feature maps back to high-resolution branches, the main branch obtains the supplementary of high-level semantic information from other branches. Fig. 3 shows the effects of PCFT. Since the main branch represents high-resolution feature maps, we think the main branch focuses attention on the textures of features, which are supplemented to other branches. Lower resolution branches have larger receptive fields, so they can auxiliarily explore complex features. And we can find all branches have similar class activation maps, so parallel branches can cooperate with each other to produce the final results. It also demonstrates PCFT can make it easier for parallel branches to capture globally contextual information.

References

[1] Liang-Jian Deng, Gemine Vivone, Cheng Jin, and Jocelyn Chanussot. Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 2020, DOI: 10.1109/TGRS.2020.3031366. 1

[2] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. Pansharpening by convolutional neural networks. *Remote Sensing*, 8(7):594, 2016. 1

[3] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, pages 618–626, 2017. 1



(a) Medium to high (b) Low to high (c) Low to medium

(d) High to medium (e) High to low (f) Medium to low

Figure 3: Visualizations of Fig 3 about the last PCFT.