Adaptive Rectangular Convolution for Remote Sensing Pansharpening

Supplementary Material

Abstract

The supplementary material provides a detailed description of the experimental setup using the ARConv module, covering several key aspects. It includes an overview of the dataset composition, followed by the configuration of the training process. Additionally, the material offers a brief introduction to the benchmark methods and outlines the specifics of the convolution kernel replacement experiment. Finally, it presents further result comparisons and visualizations to support the findings.

7. Details on Experiments

7.1. Datasets

The experimental data used in this study is captured by three different sensors: WorldView3 (WV3), QuickBird (QB), and Gao-Fen2 (GF2). A downsampling process is used to simulate and build our dataset, which includes three training sets corresponding to the three sensors. Each training set is paired with both reduced-resolution and full-resolution test sets, enabling comprehensive model evaluation across different image qualities. The training sets consist of PAN/LRMS/GT image pairs, with dimensions of 64×64, $64 \times 64 \times C$, and $64 \times 64 \times C$, respectively. The WV3 training set contains 9,714 PAN/LRMS/GT image pairs (C = 8), the QB training set contains 17,139 pairs (C = 4), and the GF2 training set contains 19,809 pairs (C = 4). The corresponding reduced-resolution test sets for these three training sets each consist of 20 PAN/LRMS/GT image pairs, with dimensions of 256×256, $256 \times 256 \times C$, and $256 \times 256 \times C$, respectively. The full-resolution dataset includes 20 RAN/LRMS image pairs, with dimensions of 512×512, $512 \times 512 \times C$. These datasets are publicly available through the PanCollection repository [8].

7.2. Training Details

This section provides a detailed description of the training details for all our experiments, focusing on aspects such as the loss function, optimizer, batch size, number of training epochs, exploratory phase epochs, convolution kernel height and width learning range, initial learning rate, and learning rate decay methods. In all experiments, the loss function used is $l_1$loss, the optimizer is Adam optimizer [17], the batch size is 16, the initial learning rate is 0.0006, the learning rate decays by a factor of 0.8 every 200 epochs and the exploratory phase consists of 100 epochs. The purpose of the exploratory phase is to address the challenge of convergence when selecting the number of convolution kernel sampling points based on the average learned height and width of the kernels. After the exploratory phase, we randomly select a set of convolution kernel sampling point combinations and keep them fixed during the subsequent training process. The remaining configuration differences are shown in the Tab. 9.

7.3. Benchmark Methods

In the main text, we provide a detailed comparison between the proposed method and several established approaches. To facilitate this comparison, Tab. 12 presents a concise overview of the benchmark methods used in our study. The table is divided into two parts by a horizontal line, with traditional methods listed above the line and deep learning methods below the line.

7.4. Replacing Convolution Experiment

In FusionNet, the original architecture consists of four standard residual blocks. In AR-FusionNet, we replace the convolution layers in the two middle residual blocks with our proposed ARConv. This modification results in a total of four ARConv layers in the network, which enhances its ability to capture more complex features. Similarly, in LAGNet, which has five standard residual blocks, we replace the convolution layers in the second and fourth blocks with ARConv. This strategic placement allows us to evaluate ARConv in a deeper network structure, providing a comparison with other models. ARNet and CANNet are constructed similarly, with each replacing the standard convolution modules in the U-Net architecture [23, 35] with their respective proposed convolution modules. Specifically, in CANNet, all standard convolutions are replaced with ARConv, thus transforming it into ARNet. This provides a natural comparison between the two networks, offering valuable insights into the impact of the different convolution techniques. The training set for all three experiments is WV3, other training details can be found in Tab. 8.

7.5. More Results

Tab. 10 and 11 present the performance benchmarks on the full-resolution QB and GF2 datasets, evaluating the effectiveness of various methods. Among the three metrics, D_λ measures the network's ability to capture spectral information, while D_s reflects the network's capacity to preserve spatial details. The metric HQNR $= (1 - D_\lambda)(1 - D_s)$ provides a comprehensive evaluation of the network's overall performance and is considered the most critical metric for assessing methods on full-resolution datasets. Fig. 7

Table 8. The different configurations for replacing convolution experiment. The first three columns represent the experiment name, the number of training epochs, and the convolution kernel height and width learning range. The subsequent columns, Layer1-10, represent the final number of sampling points for each of the ten convolution layers.

Experiment	Epochs	Range	Layer1-2	Layer3-4	Layer5-6	Layer7-8	Layer9-10
AR-FusionNet	400	$1-9$	$3\times5, 7\times7$	$7\times3, 5\times5$	–	–	–
AR-LAGNet	220	$1-9$	$5\times3, 5\times3$	$5\times3, 7\times7$	–	–	–
AR-CANNet	600	$1-18$	$3\times3, 3\times3$	$7\times5, 3\times5$	$3\times3, 3\times3$	$3\times3, 5\times5$	$3\times5, 3\times3$

Table 9. The different configurations for all experiments except replacing convolution experiment which is detailed in Sec. 7.4. The first three columns represent the experiment name, the number of training epochs, and the convolution kernel height and width learning range, where "HWR" stands for Height and Width Range. The subsequent columns, Layer1-10, represent the final number of sampling points for each of the ten convolution layers in ARNet. The names of the first three experiments correspond to their respective training datasets, while all subsequent experiments use the WV3 dataset for training.

Experiment	Epochs	Range	Layer1-2	Layer3-4	Layer5-6	Layer7-8	Layer9-10
WV3	600	$1-18$	$3\times3, 3\times3$	$7\times5, 3\times5$	$3\times3, 3\times3$	$3\times3, 5\times5$	$3\times5, 3\times3$
QB	200	$1-9$	$3\times3, 3\times5$	$5\times7, 3\times3$	$5\times3, 3\times3$	$3\times3, 7\times7$	$3\times3, 3\times3$
GF2	630	$1-18$	$3\times3, 3\times3$	$3\times7, 3\times5$	$3\times3, 3\times3$	$3\times3, 5\times3$	$3\times3, 3\times3$
Ablation study (a)	600	$3-3$	$3\times3, 3\times3$	$3\times3, 3\times3$	$3\times3, 3\times3$	$3\times3, 3\times3$	$3\times3, 3\times3$
Ablation study (b)	600	$1-18$	$3\times3, 3\times3$	$3\times3, 3\times3$	$3\times3, 3\times3$	$3\times3, 3\times3$	$3\times3, 3\times3$
Ablation study (c)	600	$1-18$	$3\times3, 3\times3$	$5\times5, 3\times3$	$3\times3, 3\times3$	$3\times3, 3\times3$	$3\times3, 3\times3$
HWR$1-3$	600	$1-3$	$3\times3, 3\times3$	$3\times3, 3\times3$	$3\times3, 3\times3$	$3\times3, 3\times3$	$3\times3, 3\times3$
HWR$1-9$	600	$1-9$	$5\times3, 3\times3$	$3\times3, 3\times3$	$5\times3, 5\times3$	$3\times3, 3\times3$	$5\times3, 3\times3$
HWR$1-18$	600	$1-18$	$3\times3, 3\times3$	$7\times5, 3\times5$	$3\times3, 3\times3$	$3\times3, 5\times5$	$3\times5, 3\times3$
HWR$1-36$	600	$1-36$	$3\times3, 3\times3$	$3\times3, 3\times3$	$3\times3, 5\times5$	$5\times3, 3\times3$	$3\times3, 3\times3$
HWR$1-63$	600	$1-63$	$3\times3, 3\times3$	$5\times5, 5\times5$	$5\times5, 5\times5$	$3\times5, 5\times5$	$3\times3, 3\times3$
Comparison with DCNv2	600	$1-18$	$3\times3, 3\times3$	$3\times3, 3\times3$	$3\times3, 3\times3$	$3\times3, 3\times3$	$3\times3, 3\times3$

Table 10. Performance benchmarking on the QB dataset using 20 full-resolution samples. Best in bold; second best underlined.

Methods	$D_\lambda \downarrow$	$D_s \downarrow$	HQNR↑
EXP [1]	0.0436±0.0089	0.1502±0.0167	0.813±0.020
TV [21]	0.0465±0.0146	0.1500±0.0238	0.811±0.034
MTF-GLP-FS [31]	0.0550±0.0142	0.1009±0.0265	0.850±0.037
BDSD-PC [29]	0.1975±0.0334	0.1636±0.0483	0.672±0.058
CVPR19 [12]	0.0498±0.0119	0.0783±0.0170	0.876±0.023
LRTCFPan [37]	**0.0226±0.0117**	0.0705±0.0351	0.909±0.044
PNN [19]	0.0577±0.0110	0.0624±0.0239	0.844±0.030
PanNet [38]	0.0426±0.0112	0.1137±0.0323	0.849±0.039
DiCNN [15]	0.0947±0.0145	0.1067±0.0210	0.809±0.031
FusionNet [6]	0.0572±0.0182	0.0522±0.0088	0.894±0.021
DCFNet [36]	0.0469±0.0150	0.1239±0.0269	0.835±0.016
LAGConv [16]	0.0859±0.0237	0.0676±0.0136	0.852±0.018
HMPNet [27]	0.1832±0.0542	0.0793±0.0245	0.753±0.065
CMT [24]	0.0504±0.0122	**0.0368±0.0075**	0.915±0.016
CANNet [9]	0.0370±0.0129	0.0499±0.0092	0.915±0.012
Proposed	0.0384±0.0148	0.0396±0.0090	**0.924±0.0191**

Table 11. Performance benchmarking on the GF2 dataset using 20 full-resolution samples. Best in bold; second best underlined.

Methods	$D_\lambda \downarrow$	$D_s \downarrow$	HQNR↑
EXP [1]	0.0180±0.0081	0.0957±0.0209	0.888±0.023
TV [21]	0.0346±0.0137	0.1429±0.0282	0.828±0.035
MTF-GLP-FS [31]	0.0553±0.0430	0.1118±0.0226	0.839±0.044
BDSD-PC [29]	0.0759±0.0301	0.1548±0.0280	0.781±0.041
CVPR19 [12]	0.0307±0.0127	0.0622±0.0101	0.909±0.017
LRTCFPan [37]	0.0325±0.0269	0.0896±0.0141	0.881±0.023
PNN [19]	0.0317±0.0286	0.0943±0.0224	0.877±0.036
PanNet [38]	**0.0179±0.0110**	0.0799±0.0178	0.904±0.020
DiCNN [15]	0.0369±0.0132	0.0992±0.0131	0.868±0.016
FusionNet [6]	0.0350±0.0124	0.1013±0.0134	0.867±0.018
DCFNet [36]	0.0240±0.0115	0.0659±0.0096	0.912±0.012
LAGConv [16]	0.0284±0.0130	0.0792±0.0136	0.895±0.020
HMPNet [27]	0.0819±0.0499	0.1146±0.0126	0.813±0.049
CMT [24]	0.0225±0.0116	**0.0433±0.0096**	**0.935±0.014**
CANNet [9]	0.0194±0.0101	0.0630±0.0094	0.919±0.011
Proposed	0.0189±0.0097	0.0515±0.0099	0.931±0.012

to 14 display the outputs of ARNet compared to various benchmark methods on both reduced-resolution and full-resolution test sets from the WV3, QB, and GF2 datasets. Additionally, residual maps between the outputs and the ground truth are provided for the reduced-resolution datasets. These figures and tables strongly demonstrate the robustness of our proposed method across multiple datasets.

7.6. More Visualizations

In ARNet, there are a total of five AR-Resblocks, each containing two ARConv layers. For our analysis, we select one ARConv from each block and visualize the heatmaps corresponding to the height and width of its learned convolution kernel. The heatmaps, shown in Fig. 15 to 18, provide valuable insight into the relationship between the kernel shapes and the object sizes present in the feature maps. This adaptability highlights the flexibility of our approach in handling various object scales and offers compelling evidence of the effectiveness of our method in dynamically adjusting to different input characteristics, making it a powerful tool for tasks that require precise and scalable feature extraction.

Table 12. A brief introduction to various benchmark methods.

Method	Year	Introduction
EXP [1]	**2002**	Upsamples the MS image.
MTF-GLP-FS [31]	**2018**	Focuses on a regression-based approach for pansharpening, specifically for the estimation of injection coefficients at full resolution.
TV [21]	**2014**	Uses total variation to regularize an ill-posed problem in a widely used image formation model.
BSDS-PC [29]	**2019**	Addresses the limitations of the traditional BDSD method when fusing multispectral images with more than four spectral bands.
CVPR2019 [12]	**2019**	Proposes a new variational pan-sharpening model based on local gradient constraints to improve spatial preservation.
LRTCFPan [37]	**2023**	Proposes a novel low-rank tensor completion (LRTC)-based framework for multispectral pansharpening.
PNN [19]	**2016**	Adapts a simple three-layer architecture for pansharpening.
PanNet [38]	**2017**	Deeper CNN for pansharpening, incorporating domain-specific knowledge to preserve both spectral and spatial information.
DiCNN [15]	**2018**	Proposes a new detail injection-based convolutional neural network framework for pansharpening.
FusionNet [6]	**2021**	Introduces the use of deep convolutional neural networks combined with traditional fusion schemes for pansharpening.
DCFNet [36]	**2021**	Addresses the limitations of single-scale feature fusion by considering both high-level semantics and low-level features.
LAGConv [16]	**2022**	Employs local-context adaptive convolution kernels with global harmonic bias.
HMPNet [27]	**2023**	An interpretable model-driven deep network for fusing hyperspectral, multispectral, and panchromatic images.
CMT [24]	**2024**	Integrates a signal-processing-inspired modulation technique into the attention mechanism to effectively fuse images.
CANNet [9]	**2024**	Incorporates non-local self-similarity to improve the effectiveness and reduce redundant learning in remote sensing image fusion.

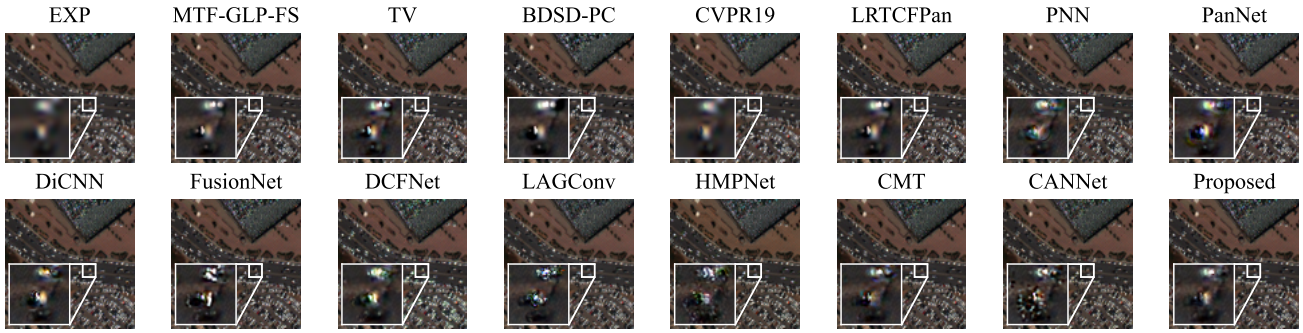

Figure 7. Comparison of qualitative results among benchmark methods on WV3 full-resolution dataset. The first row displays the RGB outputs, and the second row shows the residual relative to the ground truth. **Zoom in for best view.**

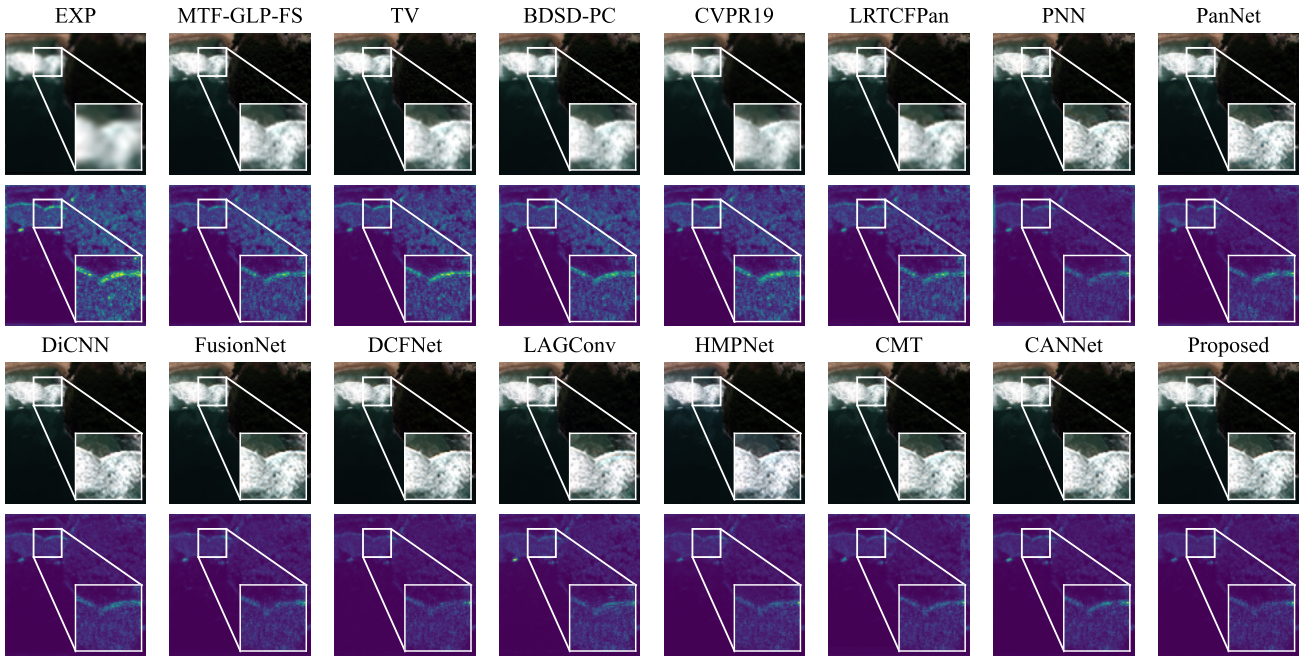

Figure 8. Comparison of qualitative results among benchmark methods on WV3 reduced-resolution dataset. The first row displays the RGB outputs, and the second row shows the residual relative to the ground truth. **Zoom in for best view.**

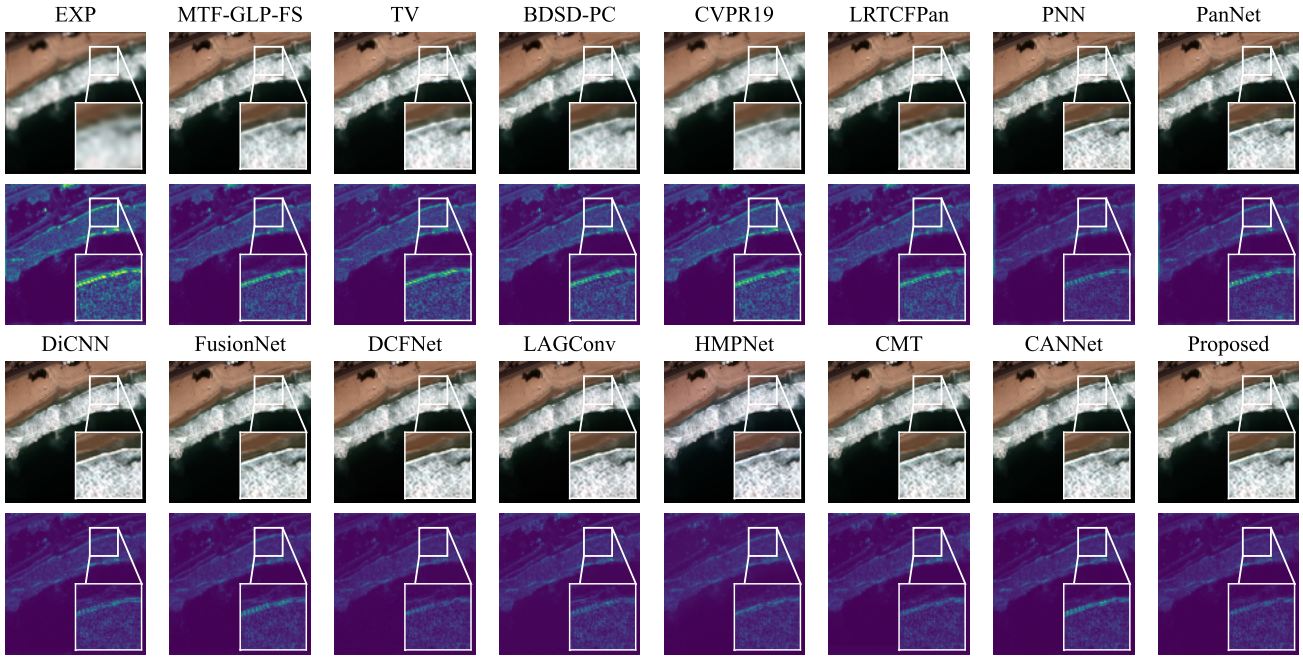

Figure 9. Comparison of qualitative results among benchmark methods on WV3 reduced-resolution dataset. The first row displays the RGB outputs, and the second row shows the residual relative to the ground truth. **Zoom in for best view.**

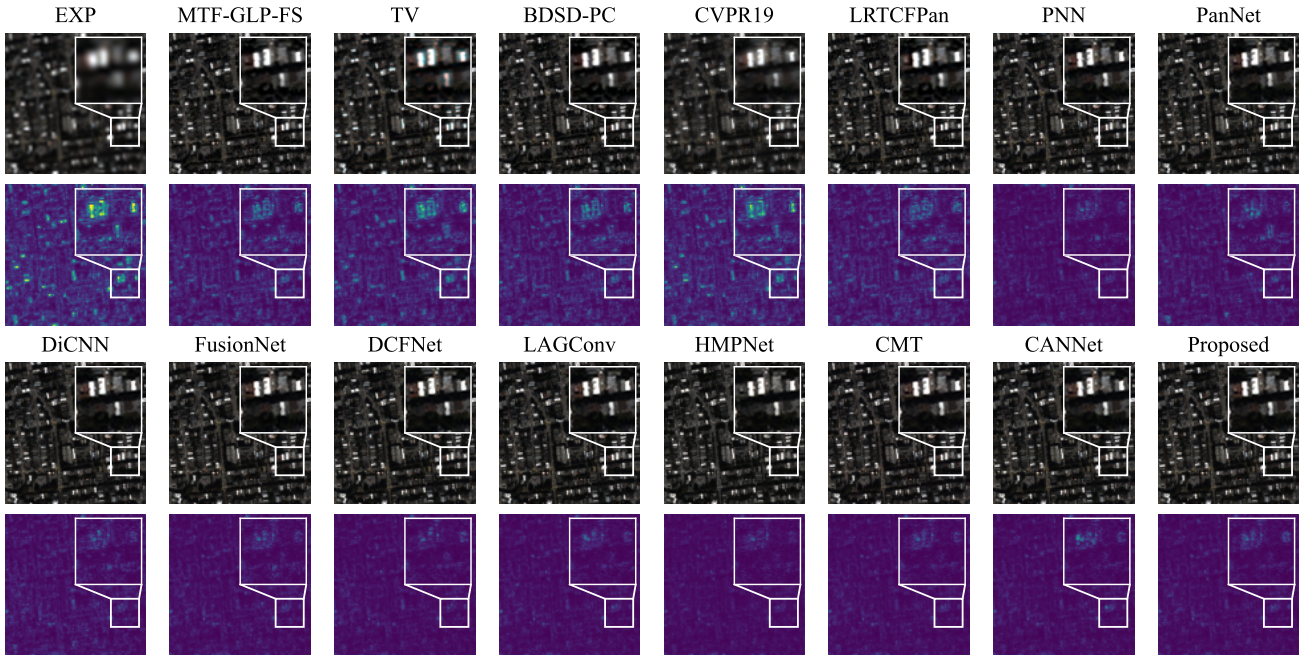

Figure 10. Comparison of qualitative results among benchmark methods on QB reduced-resolution dataset. The first row displays the RGB outputs, and the second row shows the residual relative to the ground truth. **Zoom in for best view.**

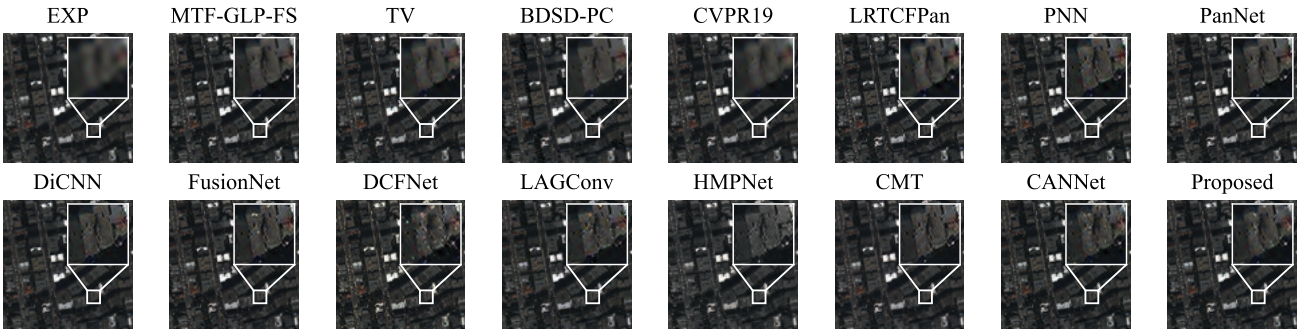

Figure 11. Comparison of qualitative results among benchmark methods on QB full-resolution dataset. The first row displays the RGB outputs, and the second row shows the residual relative to the ground truth. **Zoom in for best view.**

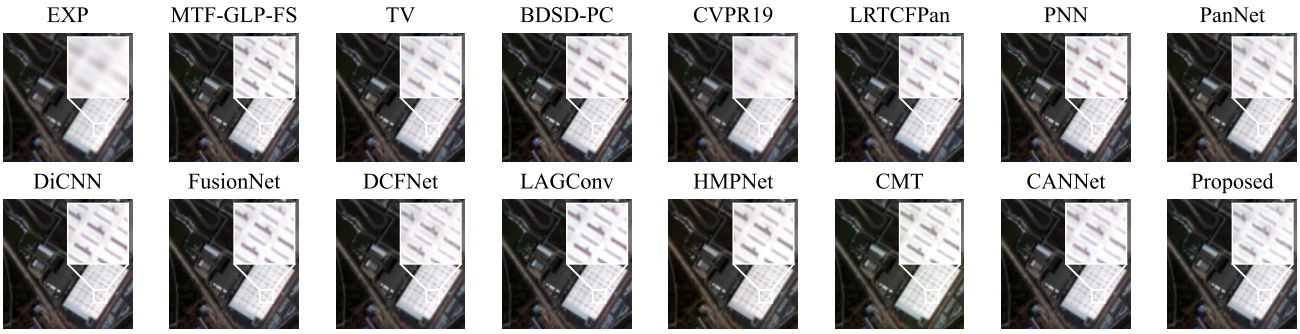

Figure 12. Comparison of qualitative results among benchmark methods on GF2 full-resolution dataset. The first row displays the RGB outputs, and the second row shows the residual relative to the ground truth. **Zoom in for best view.**

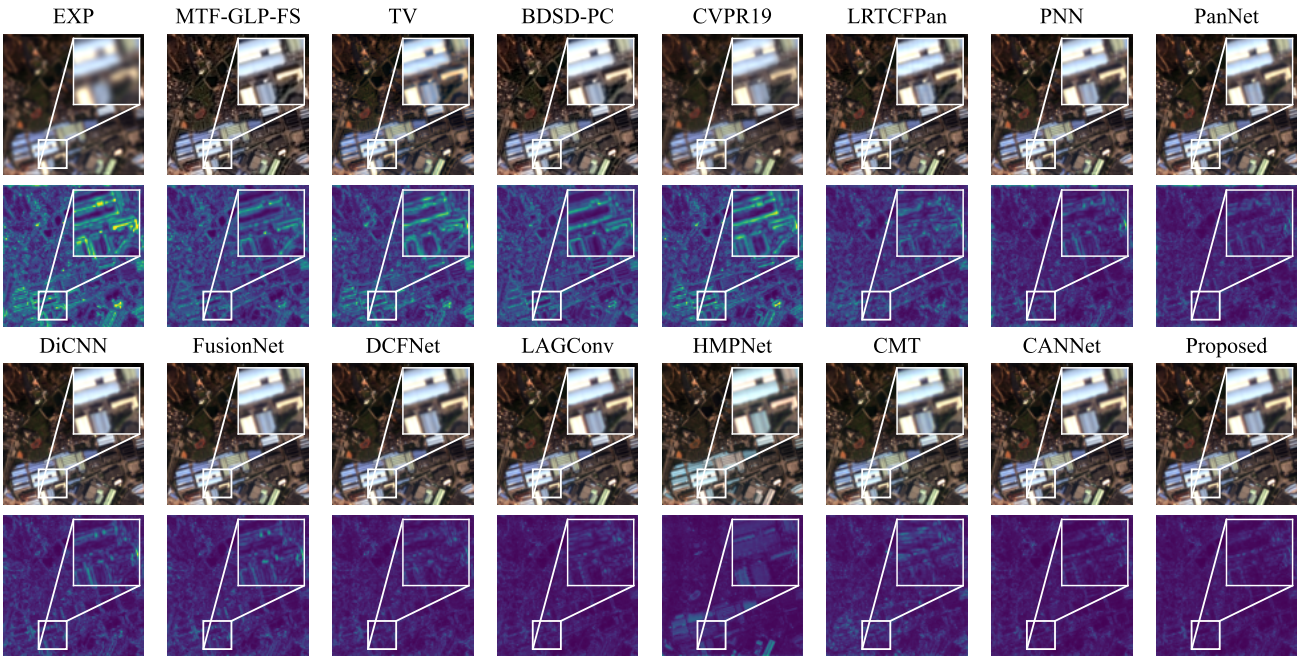

Figure 13. Comparison of qualitative results among benchmark methods on GF2 reduced-resolution dataset. The first row displays the RGB outputs, and the second row shows the residual relative to the ground truth. **Zoom in for best view.**

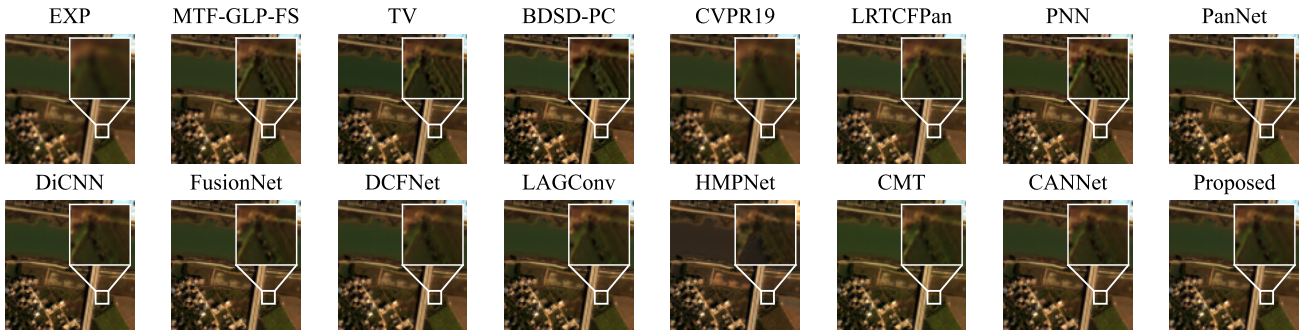

Figure 14. Comparison of qualitative results among benchmark methods on GF2 full-resolution dataset. The first row displays the RGB outputs, and the second row shows the residual relative to the ground truth. **Zoom in for best view.**

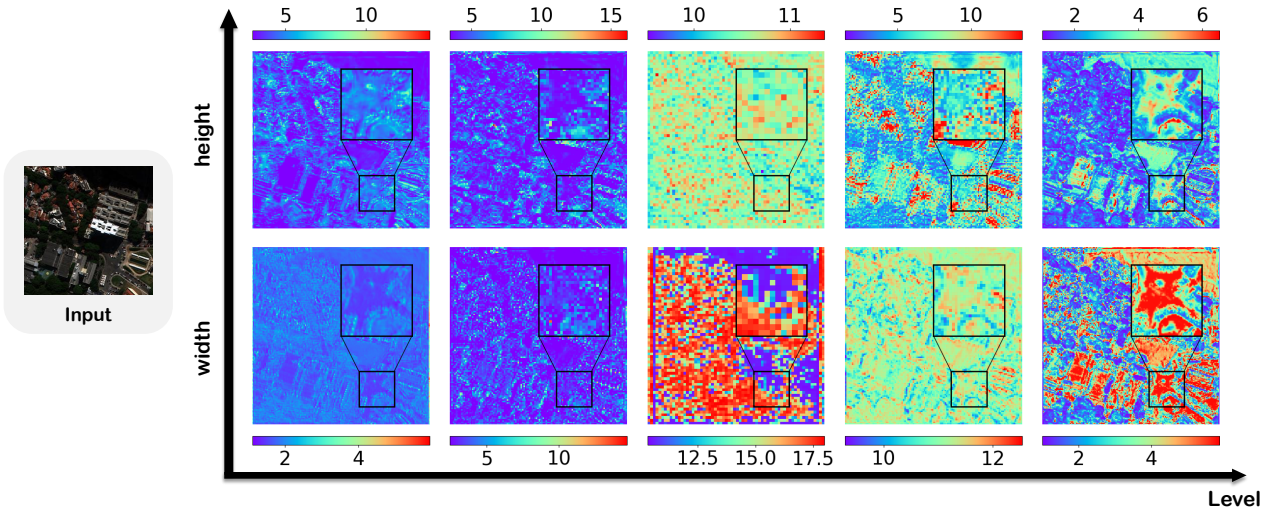

Figure 15. Heatmaps of the heights and widths learned at each pixel by convolutional kernels at different layers. The input image is a sample from the WV3 dataset.

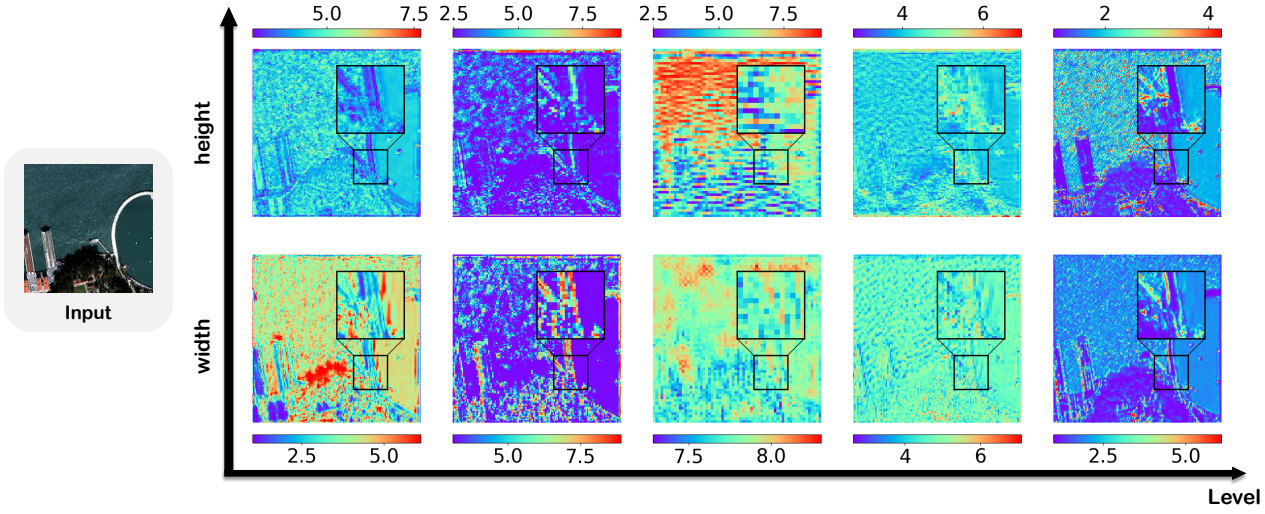

Figure 16. Heatmaps of the heights and widths learned at each pixel by convolutional kernels at different layers. The input image is a sample from the QB dataset.

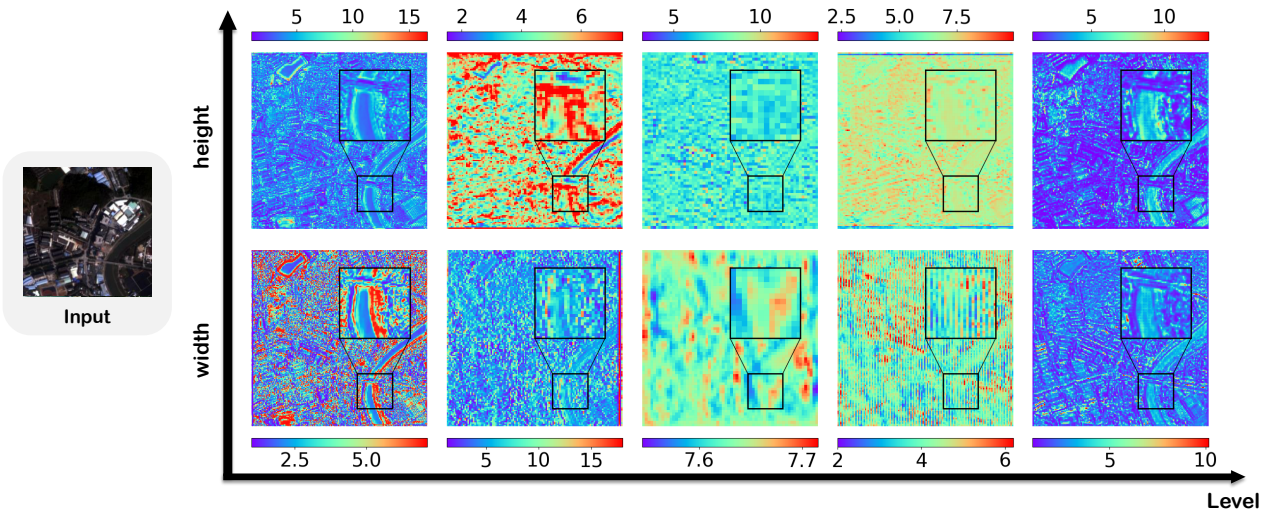

Figure 17. Heatmaps of the heights and widths learned at each pixel by convolutional kernels at different layers. The input image is a sample from the GF2 dataset.

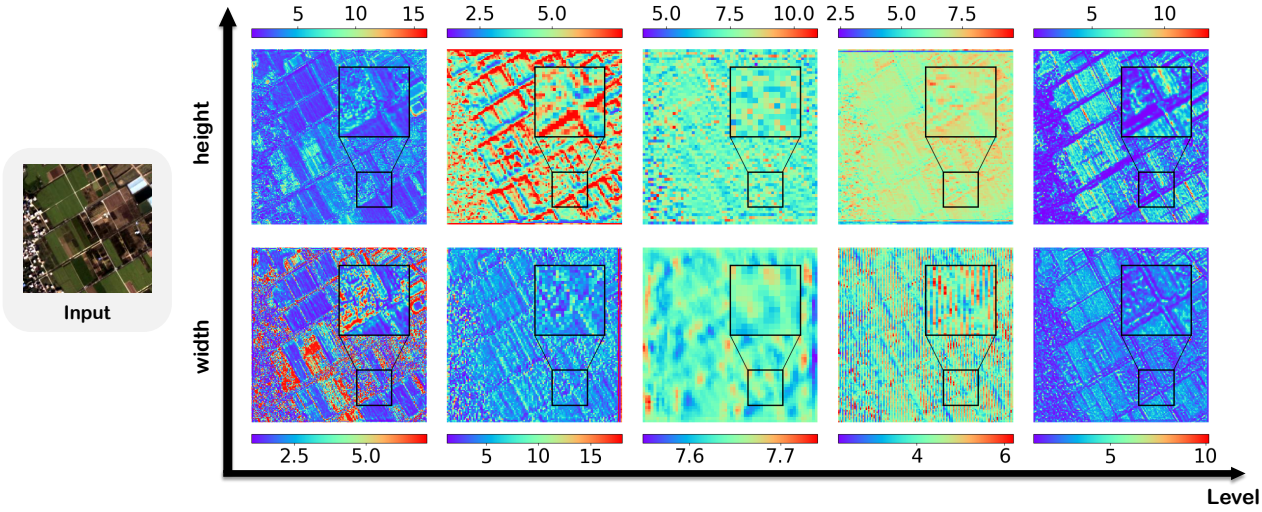

Figure 18. Heatmaps of the heights and widths learned at each pixel by convolutional kernels at different layers. The input image is a sample from the GF2 dataset.