# PAN-Crafter

## Learning Modality-Consistent Alignment for PAN-Sharpening

Jeonghyeok Do[1] Sungpyo Kim[1] Geunhyuk Youk[1] Jaehyup Lee[2†] Munchurl Kim[1†]

KAIST EE
KAIST ELECTRICAL ENGINEERING

VICLab Video and Image Computing Lab

[1]Korea Advanced Institute of Science and Technology
[2]Kyungpook National University
[†]Co-corresponding Authors



ICCV OCT 19-23, 2025 HONOLULU HAWAII

# Motivation

- PAN-Sharpening (PS)
  - Many of remote sensing imagery tasks *require high-resolution images*
  - Inherent limitations in sensor technologies
    - A single imaging system *cannot simultaneously achieve* high spatial resolution

↳ Modern Earth observation satellites employ dual-sensor systems

- High-resolution panchromatic (*PAN*) sensor
- Low-resolution multi-spectral (*LRMS*) sensor

# Motivation

- PAN-Sharpening (PS)
  - Fuse high-resolution PAN images with low-resolution MS (LRMS) images
  - Produce high-resolution multi-spectral (HRMS)

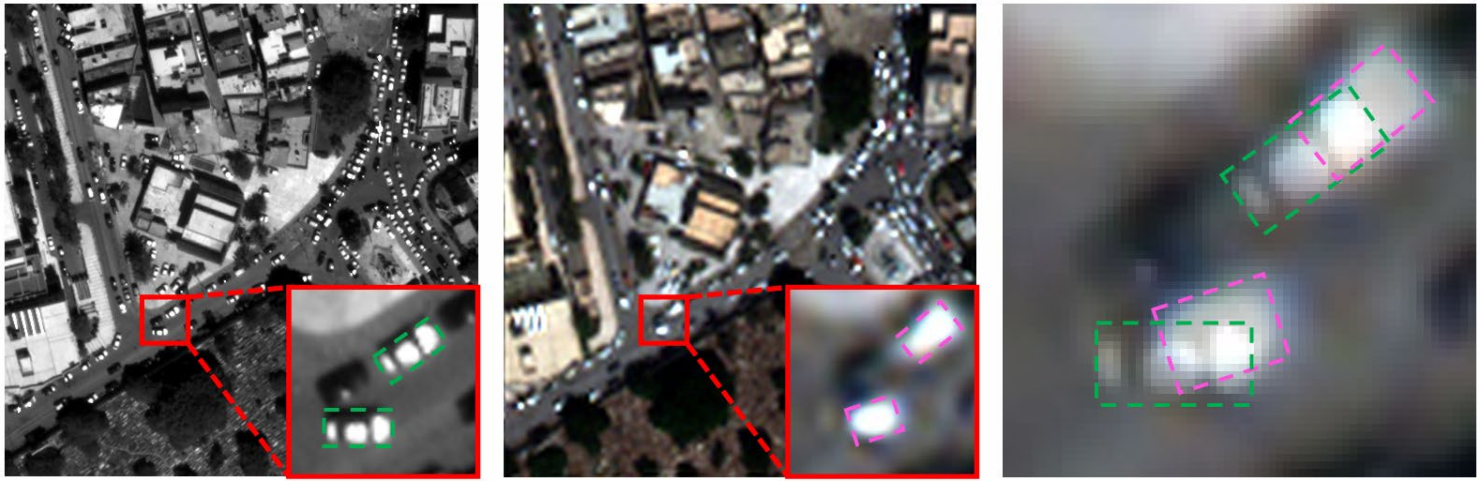
PAN

LRMS

HRMS

# Motivation

- PAN-Sharpening (PS)
  - Fundamental challenge: cross-modality misalignment
    - Arise from *differences in sensor placement*, *acquisition timing*, and *resolution disparity*



(a) PAN image $\mathbf{I}_{\text{pan}}$ (b) LRMS image $\mathbf{I}_{\text{ms}}^{\text{lr}}$ (c) Overlay of $\mathbf{I}_{\text{pan}}$ and $\mathbf{I}_{\text{ms}}^{\text{lr}}$

# Motivation

- Previous PS methods: **most existing** methods
  - Assume perfect pixel-wise alignment
  - Rely on per-pixel reconstruction losses, such as $\ell_1$ and $\ell_2$
    - *Spectral distortion, double edges, and blurring* when misalignment is present

PAN

"Misaligned"

LRMS

HRMS
(output)

# Motivation

- Previous PS methods: **spatial-adaptive convolutional layers**
  - *Explicitly identifies misalignment* as a critical challenge in PS

    (SIPSA, CVPR 2021)
    - Fixed-scale alignment mechanisms fail to capture complex spatial shifts
  - *Self-similarity-based feature aggregation*

    (LAGConv, AAAI 2022) (CANConv, CVPR 2024)
    - Do not explicitly correct geometric discrepancies between PAN and MS images

Align textures and structures across modalities

Ensure modality-consistent feature refinement at multiple spatial scales

# Proposed Method

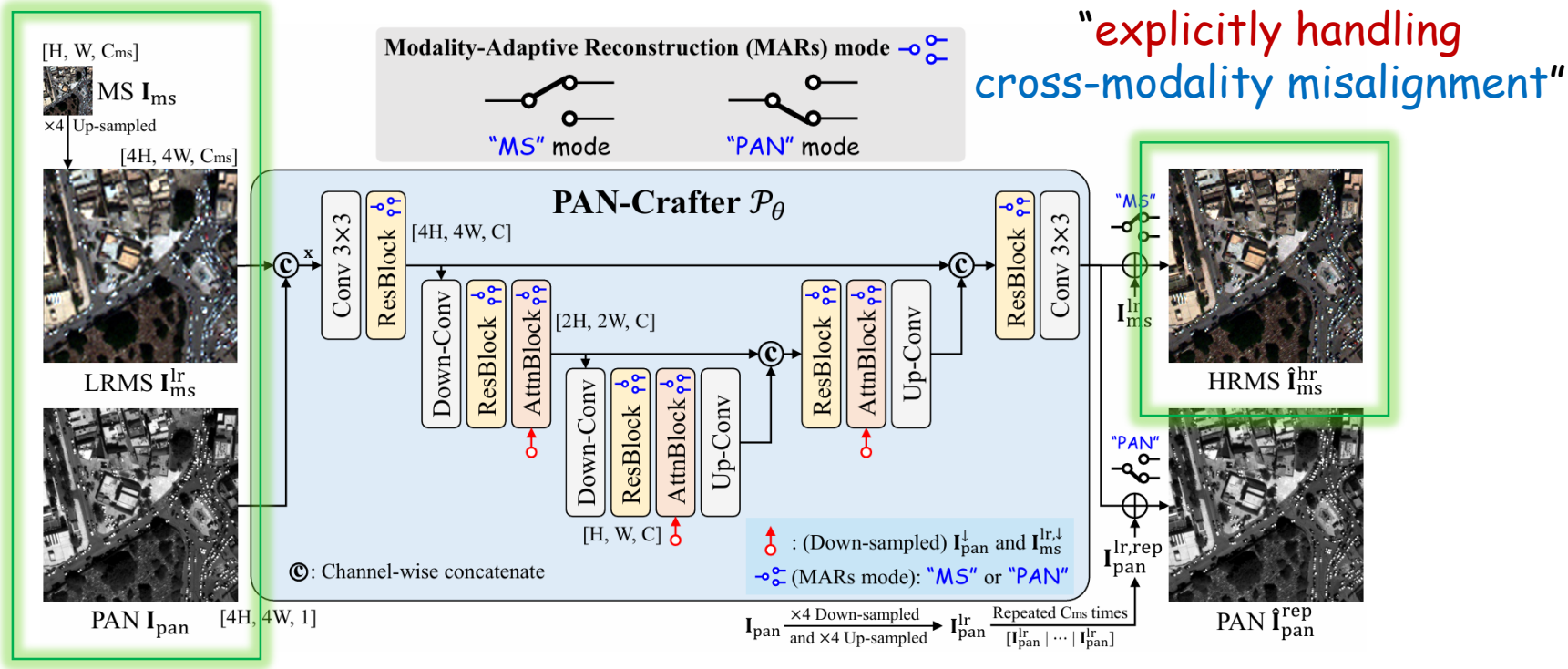
- We propose **PAN-Crafter**, a *modality-consistent alignment framework* designed to handle cross-modality misalignment during the fusion process

  - Modality-Adaptive Reconstruction (MARs)
    - *Unified reconstruction framework*
    - *Robust learning from misaligned PAN-MS image pairs* by dynamically generating both HRMS and PAN images

  - Cross-Modality Alignment-Aware Attention (CM3A)
    - Novel ***alignment mechanism*** that adaptively refines textures and structures *between PAN and MS images*, improving spatial-spectral consistency

# Proposed Method

- Overview of the **PAN-Crafter**: Our goal
  - The network synthesizes a HRMS image from the given PAN and LRMS input



# Proposed Method

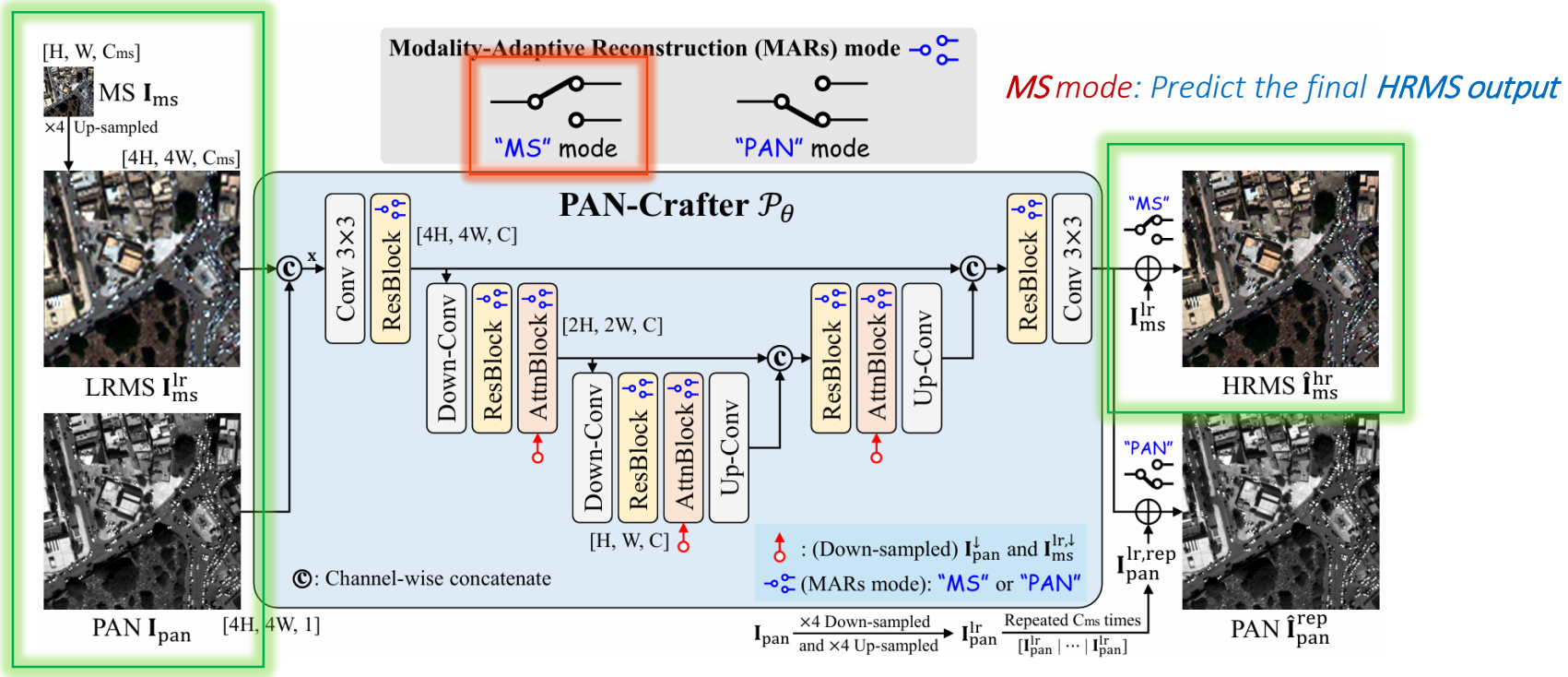
- Overview of the **PAN-Crafter**: Modality Adaptive Reconstruction (MARs)
  - Reconstruct either HRMS or PAN images depending on the *modality selection*

Reconstruct both HRMS and PAN images within a **shared network**

The network effectively incorporates **sharp spatial structures** while maintaining **spectral fidelity** in the HRMS prediction

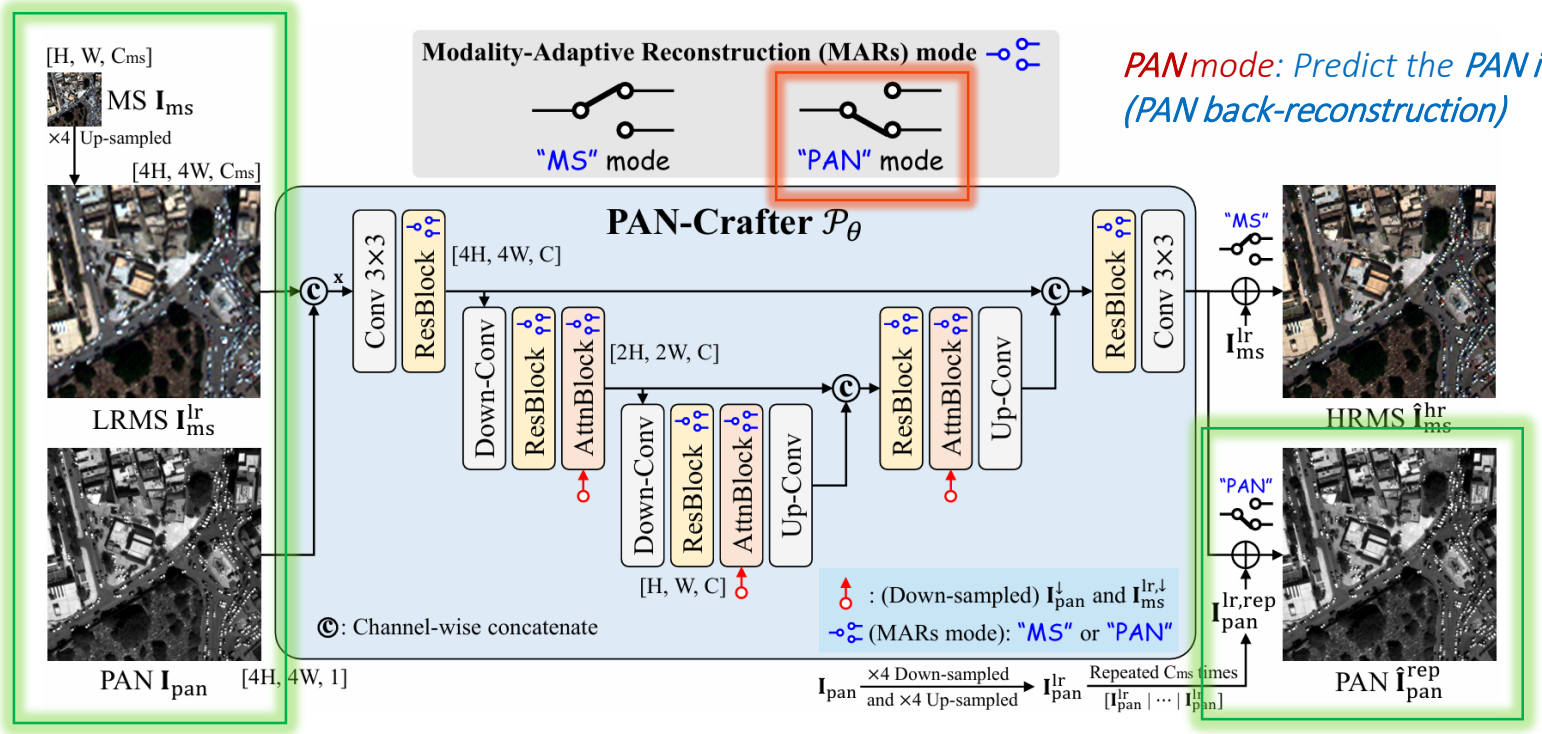
# Proposed Method

- Overview of the **PAN-Crafter**: Modality Adaptive Reconstruction (MARs)
  - Reconstruct either HRMS or PAN images depending on the *modality selection*



# Proposed Method

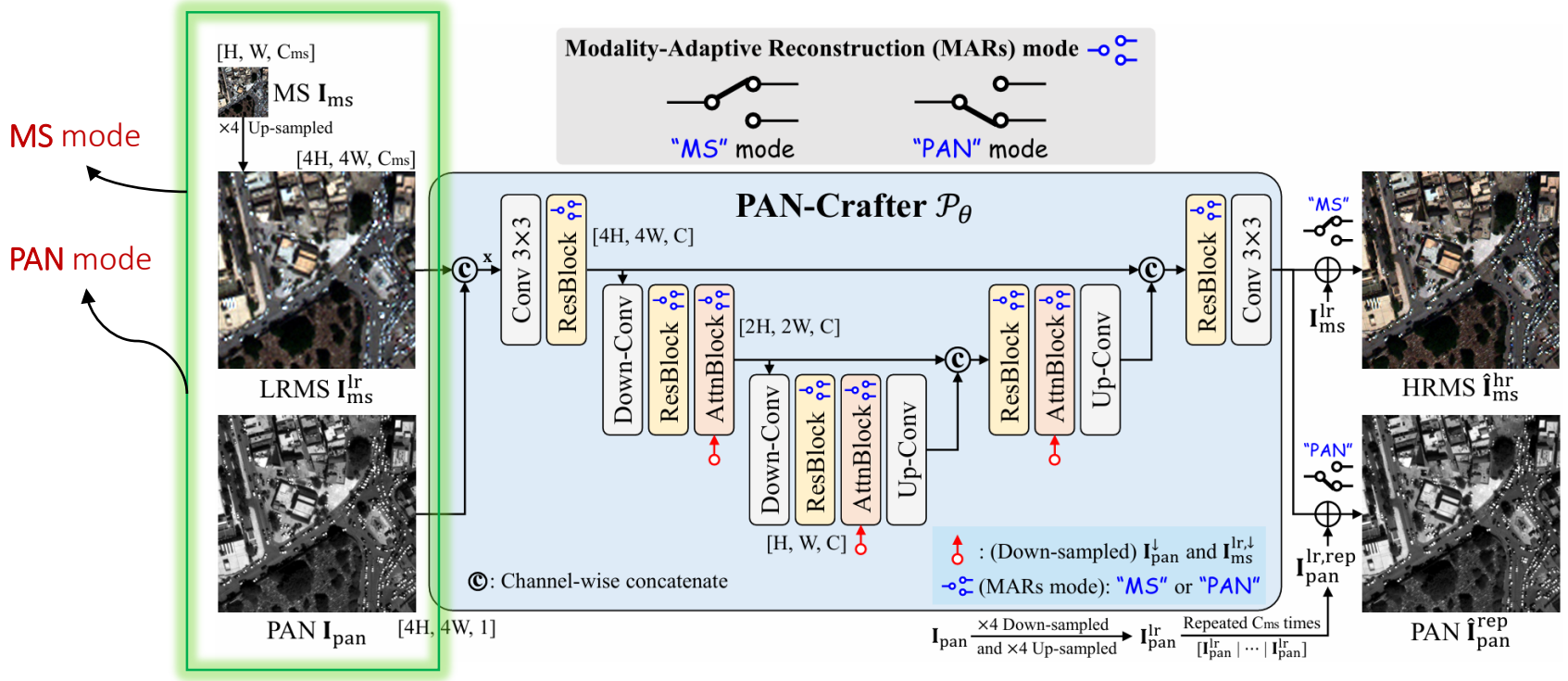
- Overview of the **PAN-Crafter**: Modality Adaptive Reconstruction (MARs)
  - Reconstruct either HRMS or PAN images depending on the *modality selection*




*PAN mode: Predict the PAN input (PAN back-reconstruction)*

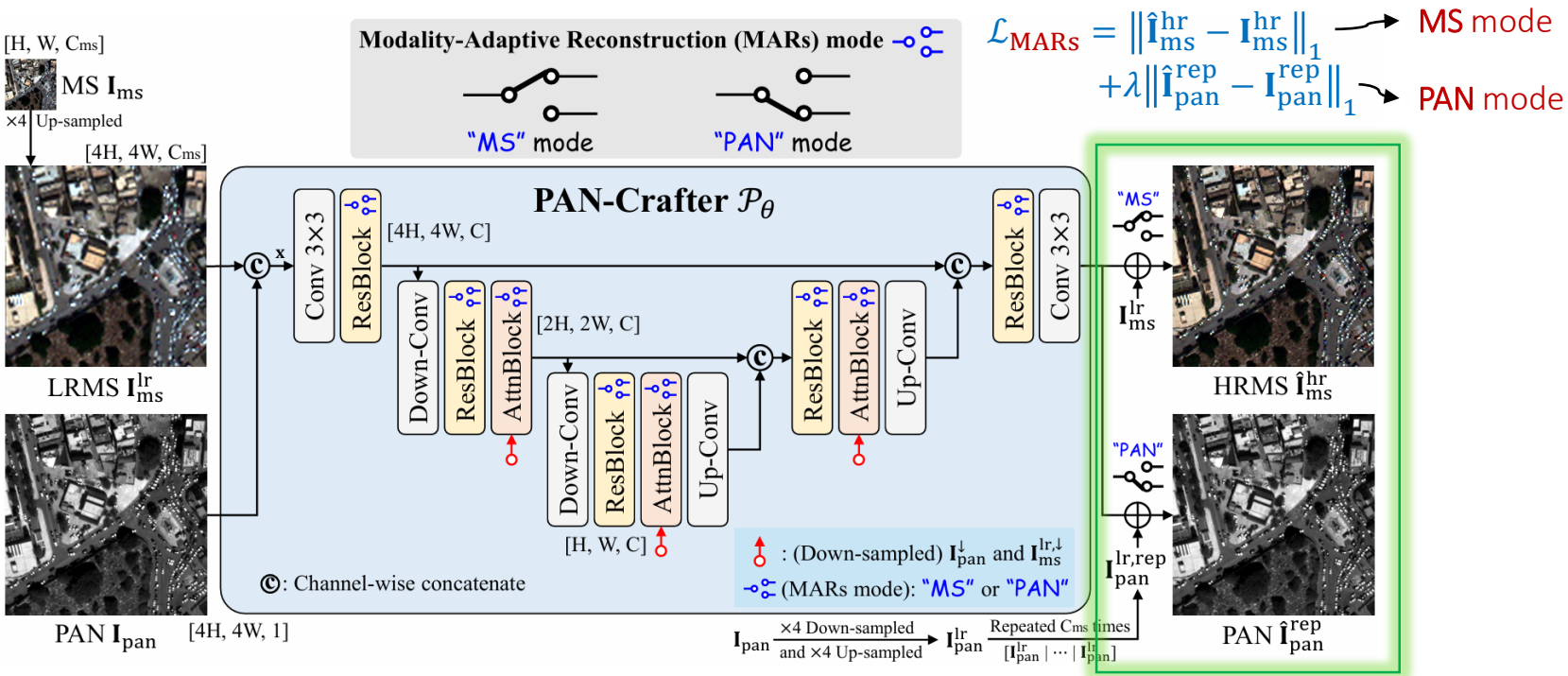
# Proposed Method

- Overview of the **PAN-Crafter**: MARs loss
  - Learn modality-specific features while maintaining consistency across both tasks



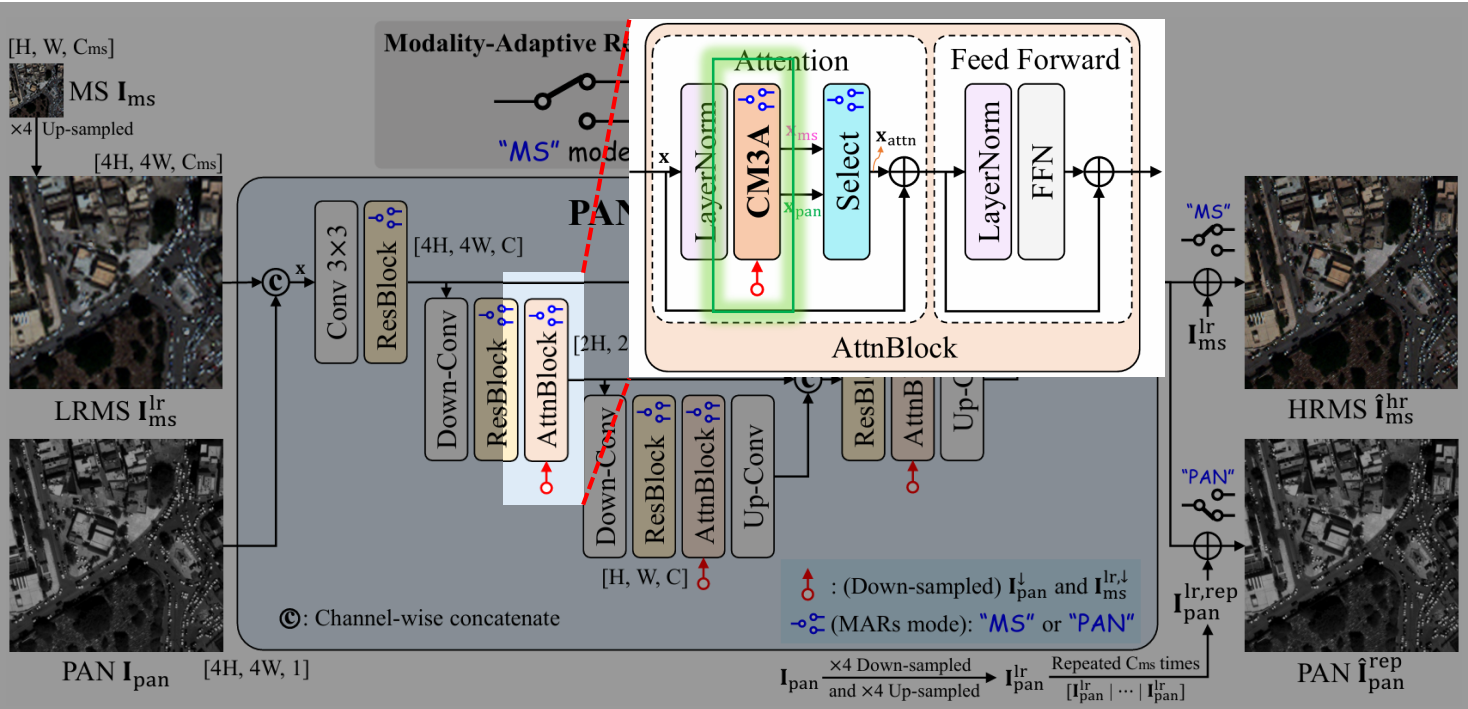
# Proposed Method

- Overview of the **PAN-Crafter**: MARs loss
  - Learn modality-specific features while maintaining consistency across both tasks



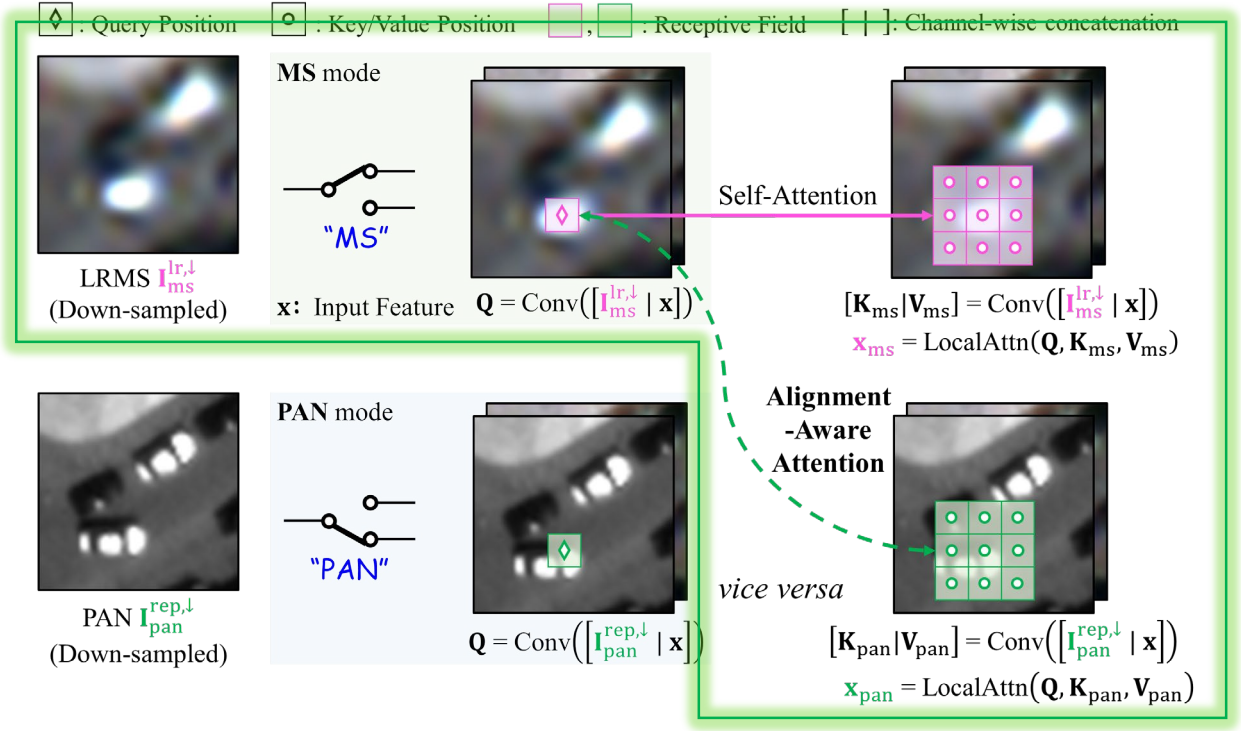
# Proposed Method

- Cross-Modality Alignment-Aware Attention (CM3A)
  - **Dynamically align features** based on the *selected MARs mode*



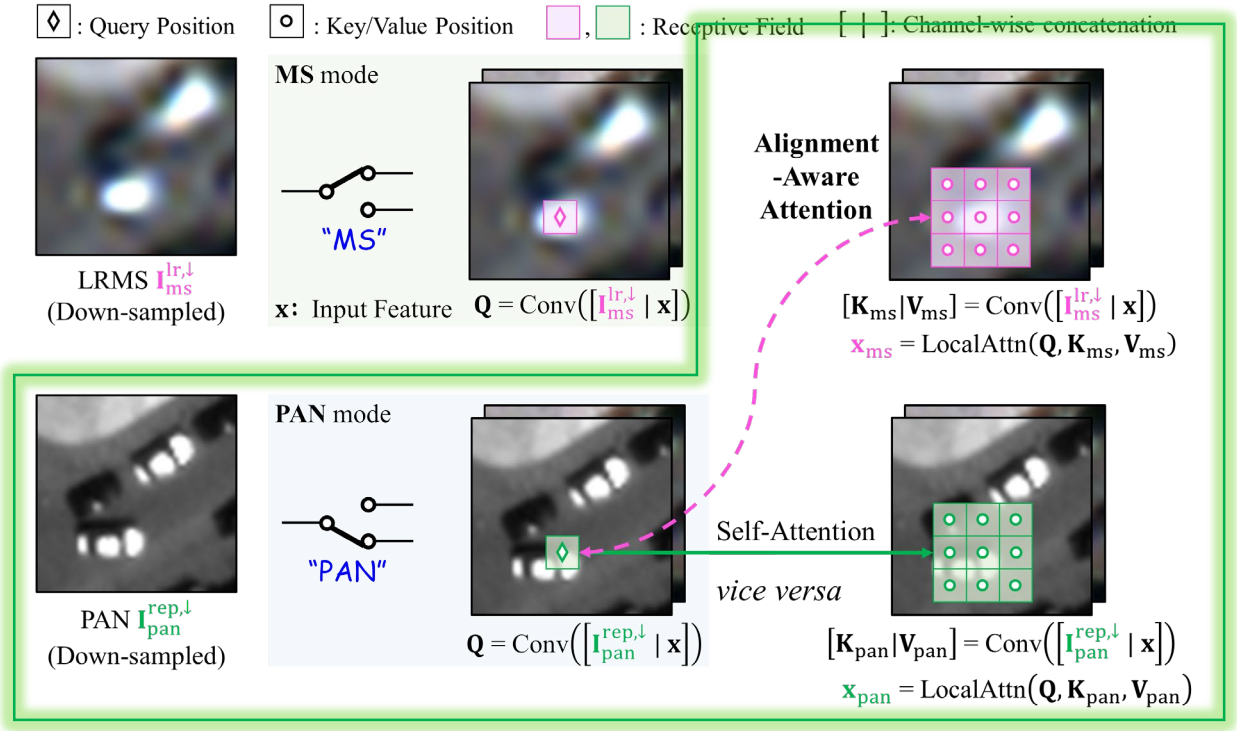
# Proposed Method

- Cross-Modality Alignment-Aware Attention (CM3A)
  - HRMS reconstruction (MS mode): transfer MS texture to PAN structure



# Proposed Method

- Cross-Modality Alignment-Aware Attention (CM3A)
  - PAN back-reconstruction (PAN mode): transfer PAN structure to MS texture



# Experiment Results

- **Quantitative Results** (WV3/GF2/QB datasets)

*Computational complexity:*
*256×256×8 HRMS target at reduced-resolution*

| **WV3** Dataset | | Full-Resolution | | | Reduced-Resolution | | | | | | Inference Time↓ (s) | Memory↓ (GB) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | Publications | HQNR↑ | $D_s$↓ | $D_\lambda$↓ | ERGAS↓ | SCC↑ | SAM↓ | Q8↑ | PSNR↑ | SSIM↑ | | |
| PanNet [57] | ICCV 2017 | 0.918 | 0.049 | 0.035 | 2.538 | 0.979 | 3.402 | 0.913 | 36.148 | 0.966 | - | - |
| MSDCNN [58] | JSTARS 2018 | 0.924 | 0.050 | 0.028 | 2.489 | 0.979 | 3.300 | 0.914 | 36.329 | 0.967 | - | - |
| FusionNet [50] | ICCV 2021 | 0.920 | 0.053 | 0.029 | 2.428 | 0.981 | 3.188 | 0.916 | 36.569 | 0.968 | - | - |
| LAGConv [20] | AAAI 2022 | 0.915 | 0.055 | 0.033 | 2.380 | 0.981 | 3.153 | 0.916 | 36.732 | 0.970 | **0.004** | 3.281 |
| S2DBPN [61] | TGRS 2023 | 0.946 | 0.030 | 0.025 | 2.245 | 0.985 | 3.019 | 0.917 | 37.216 | 0.972 | 0.005 | 2.387 |
| PanDiff [30] | TGRS 2023 | 0.952 | 0.034 | **0.014** | 2.276 | 0.984 | 3.058 | 0.913 | 37.029 | 0.971 | 2.955 | 2.328 |
| DCPNet [62] | TGRS 2024 | 0.923 | 0.036 | 0.043 | 2.301 | 0.984 | 3.083 | 0.915 | 37.009 | 0.972 | 0.109 | 7.213 |
| TMDiff [52] | TGRS 2024 | 0.924 | 0.059 | 0.018 | 2.151 | 0.986 | 2.885 | 0.915 | 37.477 | 0.973 | 9.997 | 9.910 |
| CANConv [13] | CVPR 2024 | 0.951 | 0.030 | 0.020 | 2.163 | 0.985 | 2.927 | 0.918 | 37.441 | 0.973 | 0.451 | 2.713 |
| **PAN-Crafter** | - | **0.958** | **0.027** | 0.016 | **2.040** | **0.988** | **2.787** | **0.922** | **37.956** | **0.976** | 0.009 | **1.711** |

| Methods | **GF2** Dataset | | | | | | **QB** Dataset | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Full-Resolution | | Reduced-Resolution | | | | Full-Resolution | | Reduced-Resolution | | | |
| | HQNR↑ | $D_s$↓ | ERGAS↓ | SCC↑ | SAM↓ | PSNR↑ | HQNR↑ | $D_s$↓ | ERGAS↓ | SCC↑ | SAM↓ | PSNR↑ |
| PanNet [57] | 0.929 | 0.052 | 1.038 | 0.975 | 1.050 | 39.197 | 0.851 | 0.092 | 4.856 | 0.966 | 5.273 | 35.563 |
| MSDCNN [58] | 0.898 | 0.079 | 0.862 | 0.983 | 0.946 | 40.730 | 0.888 | 0.058 | 4.074 | 0.977 | 4.828 | 37.040 |
| FusionNet [50] | 0.865 | 0.105 | 0.960 | 0.980 | 0.971 | 39.866 | 0.853 | 0.079 | 4.183 | 0.975 | 4.892 | 36.821 |
| LAGConv [20] | 0.895 | 0.078 | 0.816 | 0.985 | 0.886 | 41.147 | 0.892 | **0.035** | 3.845 | 0.980 | 4.682 | 37.565 |
| S2DBPN [61] | 0.935 | 0.046 | 0.686 | 0.990 | 0.772 | 42.686 | 0.908 | 0.036 | 3.956 | 0.980 | 4.849 | 37.314 |
| PanDiff [30] | 0.936 | 0.045 | 0.674 | 0.990 | 0.767 | 42.827 | 0.919 | 0.055 | 3.723 | 0.982 | 4.611 | 37.842 |
| DCPNet [62] | 0.953 | 0.024 | 0.724 | 0.988 | 0.806 | 42.312 | 0.880 | 0.073 | 3.618 | 0.983 | **4.420** | 38.079 |
| TMDiff [52] | 0.942 | 0.030 | 0.754 | 0.988 | 0.764 | 41.896 | 0.901 | 0.068 | 3.804 | 0.981 | 4.627 | 37.642 |
| CANConv [13] | 0.919 | 0.063 | 0.653 | 0.991 | 0.722 | 43.166 | 0.893 | 0.070 | 3.740 | 0.982 | 4.554 | 37.795 |
| **PAN-Crafter** | **0.964** | **0.017** | **0.522** | **0.994** | **0.596** | **45.076** | **0.920** | 0.039 | **3.570** | **0.984** | 4.426 | **38.195** |

# Experiment Results

## Quantitative Results (WV3/GF2/QB datasets)

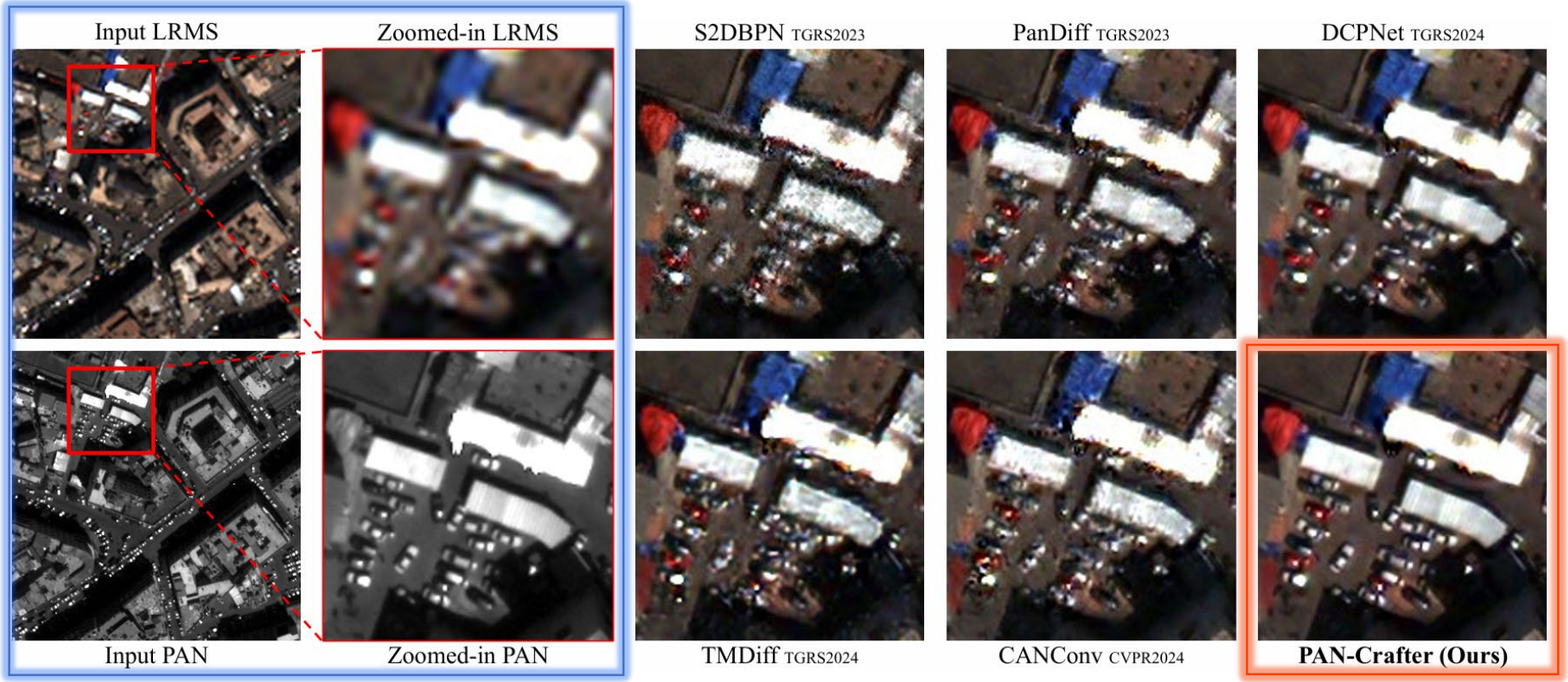
*Computational complexity:*
*256 ×256 ×8 HRMS target at reduced-resolution*

| **WV3** Dataset | | Full-Resolution | | | Reduced-Resolution | | | | | | Inference Time↓ (s) | Memory↓ (GB) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | Publications | HQNR↑ | $D_s$↓ | $D_\lambda$↓ | ERGAS↓ | SCC↑ | SAM↓ | Q8↑ | PSNR↑ | SSIM↑ | | |
| PanNet [57] | ICCV 2017 | 0.918 | 0.049 | 0.035 | 2.538 | 0.979 | 3.402 | 0.913 | 36.148 | 0.966 | - | - |
| MSDCNN [58] | JSTARS 2018 | 0.924 | 0.050 | 0.028 | 2.489 | 0.979 | 3.300 | 0.914 | 36.329 | 0.967 | - | - |
| FusionNet [50] | ICCV 2021 | 0.920 | 0.053 | 0.029 | 2.428 | 0.981 | 3.188 | 0.916 | 36.569 | 0.968 | - | - |
| LAGConv [20] | AAAI 2022 | 0.915 | 0.055 | 0.033 | 2.380 | 0.981 | 3.153 | 0.916 | 36.732 | 0.970 | **0.004** | 3.281 |
| S2DBPN [61] | TGRS 2023 | 0.946 | 0.030 | 0.025 | 2.245 | 0.985 | 3.019 | 0.917 | 37.216 | 0.972 | 0.005 | 2.387 |
| PanDiff [30] | TGRS 2023 | 0.952 | 0.034 | **0.014** | 2.276 | 0.984 | 3.058 | 0.913 | 37.029 | 0.971 | 2.955 | 2.328 |
| DCPNet [62] | TGRS 2024 | 0.923 | 0.036 | 0.043 | 2.301 | 0.984 | 3.083 | 0.915 | 37.009 | 328.33× | 0.109 | 7.213 |
| TMDiff [52] | TGRS 2024 | 0.924 | 0.059 | 0.018 | 2.151 | 0.986 | 2.885 | 0.915 | 37.477 | 1110.78× | 9.997 | 9.910 |
| CANConv [13] | CVPR 2024 | 0.951 | 0.030 | 0.020 | 2.163 | 0.985 | 2.927 | 0.918 | 37.441 | 50.11× | 0.451 | 2.713 |
| **PAN-Crafter** | - | **0.958** | **0.027** | 0.016 | **2.040** | **0.988** | **2.787** | **0.922** | **37.956** | **0.976** | 0.009 | **1.711** |

| Methods | **GF2** Dataset | | | | | | **QB** Dataset | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Full-Resolution | | Reduced-Resolution | | | | Full-Resolution | | Reduced-Resolution | | | |
| | HQNR↑ | $D_s$↓ | ERGAS↓ | SCC↑ | SAM↓ | PSNR↑ | HQNR↑ | $D_s$↓ | ERGAS↓ | SCC↑ | SAM↓ | PSNR↑ |
| PanNet [57] | 0.929 | 0.052 | 1.038 | 0.975 | 1.050 | 39.197 | 0.851 | 0.092 | 4.856 | 0.966 | 5.273 | 35.563 |
| MSDCNN [58] | 0.898 | 0.079 | 0.862 | 0.983 | 0.946 | 40.730 | 0.888 | 0.058 | 4.074 | 0.977 | 4.828 | 37.040 |
| FusionNet [50] | 0.865 | 0.105 | 0.960 | 0.980 | 0.971 | 39.866 | 0.853 | 0.079 | 4.183 | 0.975 | 4.892 | 36.821 |
| LAGConv [20] | 0.895 | 0.078 | 0.816 | 0.985 | 0.886 | 41.147 | 0.892 | **0.035** | 3.845 | 0.980 | 4.682 | 37.565 |
| S2DBPN [61] | 0.935 | 0.046 | 0.686 | 0.990 | 0.772 | 42.686 | 0.908 | 0.036 | 3.956 | 0.980 | 4.849 | 37.314 |
| PanDiff [30] | 0.936 | 0.045 | 0.674 | 0.990 | 0.767 | 42.827 | 0.919 | 0.055 | 3.723 | 0.982 | 4.611 | 37.842 |
| DCPNet [62] | 0.953 | 0.024 | 0.724 | 0.988 | 0.806 | 42.312 | 0.880 | 0.073 | 3.618 | 0.983 | **4.420** | 38.079 |
| TMDiff [52] | 0.942 | 0.030 | 0.754 | 0.988 | 0.764 | 41.896 | 0.901 | 0.068 | 3.804 | 0.981 | 4.627 | 37.642 |
| CANConv [13] | 0.919 | 0.063 | 0.653 | 0.991 | 0.722 | 43.166 | 0.893 | 0.070 | 3.740 | 0.982 | 4.554 | 37.795 |
| **PAN-Crafter** | **0.964** | **0.017** | **0.522** | **0.994** | **0.596** | **45.076** | **0.920** | 0.039 | **3.570** | **0.984** | 4.426 | **38.195** |

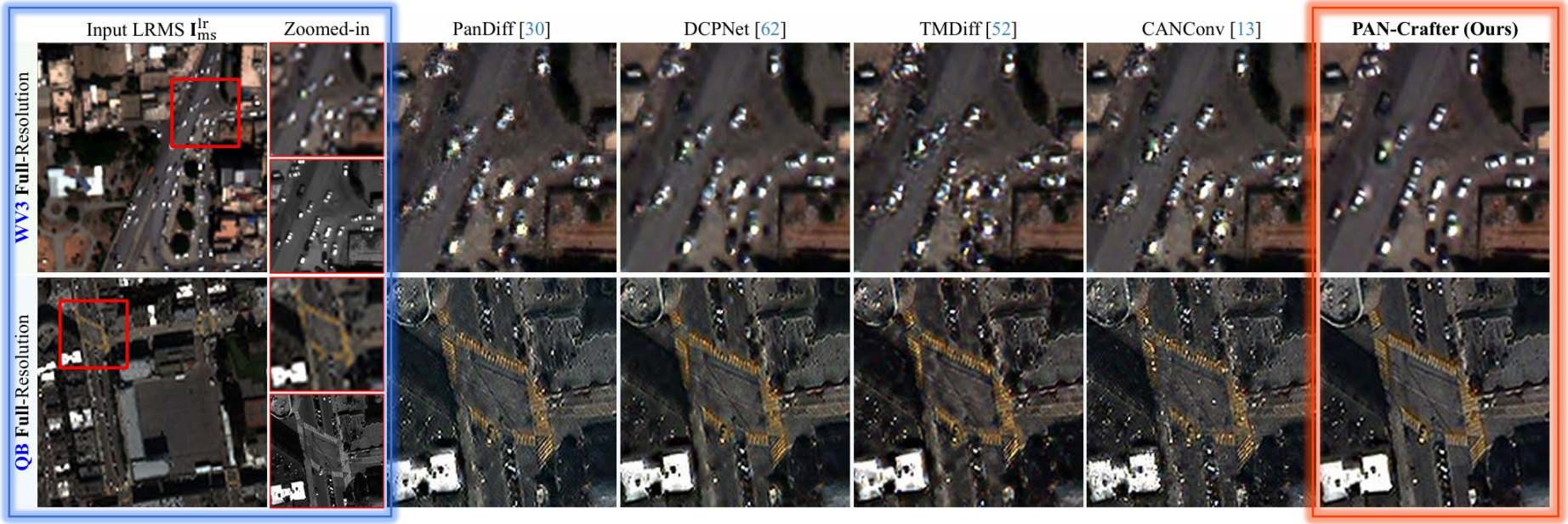
# Experiment Results

- **Qualitative Results** (WV3 dataset, full-resolution)



# Experiment Results

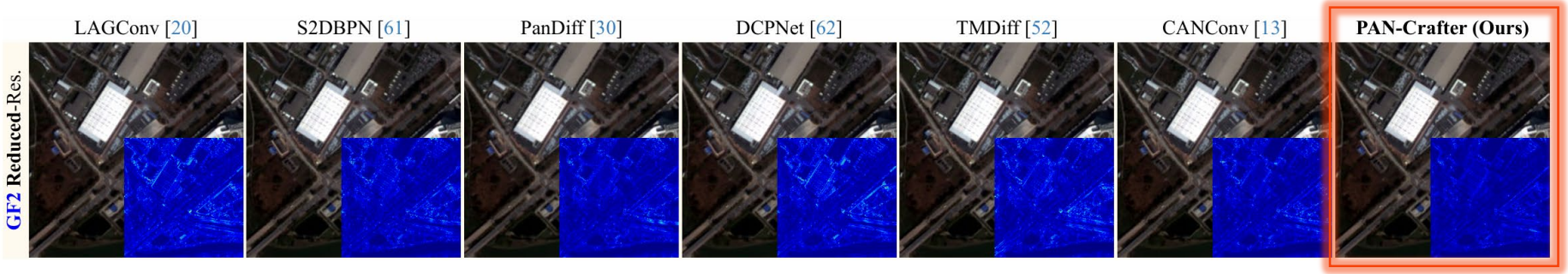
- **Qualitative Results** (WV3/QB datasets)
  - Full-resolution



# Experiment Results

- **Qualitative Results** (GF2 dataset)
  - Reduced-resolution

*Error map:*
*brighter regions indicate*
*higher reconstruction errors*



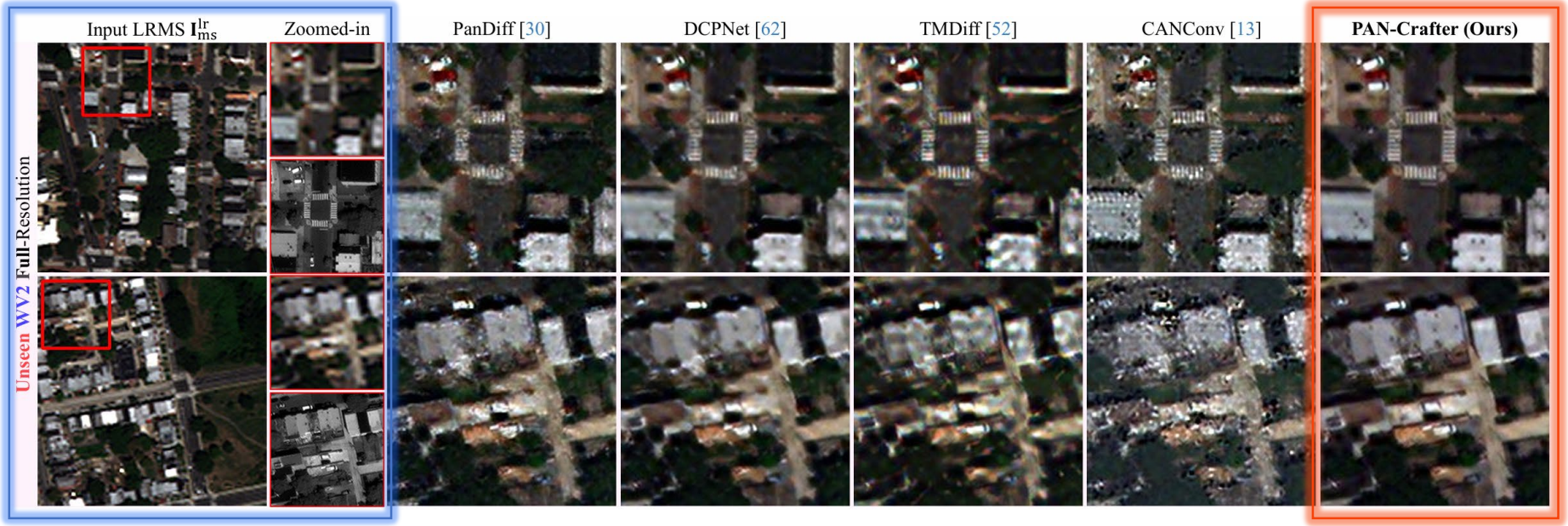
# Experiment Results

- **Generalization on unseen satellite dataset** (WV2 dataset)
  - Conduct fully **zero-shot** evaluations
  - Unseen satellite dataset not included in the training phase

| Methods | **WV2** Dataset (**Unseen** satellite dataset) | | | | |
|---|---|---|---|---|---|
| | HQNR↑ | ERGAS↓ | SCC↑ | SAM↓ | PSNR↑ |
| PanNet [57] | 0.875 | 5.481 | 0.876 | 7.040 | 27.120 |
| MSDCNN [58] | 0.862 | 4.930 | 0.905 | 5.898 | 27.901 |
| FusionNet [50] | 0.862 | 5.100 | 0.902 | 6.118 | 27.616 |
| LAGConv [20] | 0.902 | 5.133 | 0.885 | 6.094 | 27.525 |
| S2DBPN [61] | 0.813 | 5.703 | 0.915 | 7.063 | 26.748 |
| PanDiff [30] | 0.932 | 4.291 | 0.916 | 5.430 | 28.964 |
| DCPNet [62] | 0.797 | 5.507 | **0.931** | 10.174 | 27.063 |
| TMDiff [52] | 0.874 | 5.157 | 0.875 | 6.087 | 27.473 |
| CANConv [13] | 0.876 | 4.328 | 0.918 | 5.481 | 29.005 |
| **PAN-Crafter** | **0.942** | **4.169** | 0.924 | **5.078** | **29.276** |

# Experiment Results

- **Generalization on unseen satellite dataset** (WV2 dataset)
  - Conduct fully **zero-shot** evaluations
  - Unseen satellite dataset not included in the training phase



# Experiment Results

- Ablation Study: CM3A and MARs (WV3 dataset)
  - The combination of both components achieves the best performance

↳ Synergistic effect in jointly refining spatial and spectral consistency

| CM3A | MARs | WV3 Dataset | | | | | |
|---|---|---|---|---|---|---|---|
| | | HQNR↑ | ERGAS↓ | SAM↓ | PSNR↑ | Time↓ | Memory↓ |
| | | 0.948 | 2.232 | 2.980 | 37.245 | 0.006 | 1.537 |
| ✓ | | 0.949 | 2.212 | 2.970 | 37.285 | 0.007 | 1.556 |
| | ✓ | 0.956 | 2.122 | 2.873 | 37.602 | 0.009 | 1.701 |
| ✓ | ✓ | **0.958** | **2.040** | **2.787** | **37.956** | 0.009 | 1.711 |

# PAN-Crafter

## Learning Modality-Consistent Alignment for PAN-Sharpening

For more details, please visit here

# Thank You!

Project Page

KAIST EE
KAIST ELECTRICAL ENGINEERING
VICLab Video and Image Computing Lab

ICCV OCT 19-23, 2025 HONOLULU HAWAII