Multi-Scale Feature Fusion using Channel Transformers for Guided Thermal Image Super Resolution

Raghunath Sai Puttagunta [1], Birendra Kathariya [1], Zhu Li [1], George York [2]
[1]University of Missouri-Kansas City & [2] US Air Force Academy

CVPR SEATTLE, WA JUNE 17-21, 2024

Introduction:

- RGB camera sensors are hindered by factors like low lighting, occlusions, and adverse weather conditions due to their dependence on visible light.
- Thermal sensors, capturing heat in the infrared spectrum, offer versatility in challenging conditions.
- The high resolution thermal image sensors are expensive and that hinders it from widespread adaption.
- The goal of this work is to learn a high resolution thermal image from a low resolution thermal image with RGB image as a guide.

Motivation:

- CNN-based methods are widely used in image enhancement and restoration tasks. Although they achieved impressive performance in these tasks they have some limitations
  - ❑ Limited receptive field size preventing them from modeling long-range pixel dependencies.
  - ❑ Static weights at inference and they cannot flexibly adapt to the given input.
- To effectively address these limitations of CNN based methods we propose a transformer model, which has shown to capture the long-range dependencies and dynamically adapts to a given input.

Network Architecture:

- Our architecture is multi-scale inspired by TSFNet [1].
- There are three main components in our network , Shallow Feature Extractor, Fusion Block and Reconstruction Block

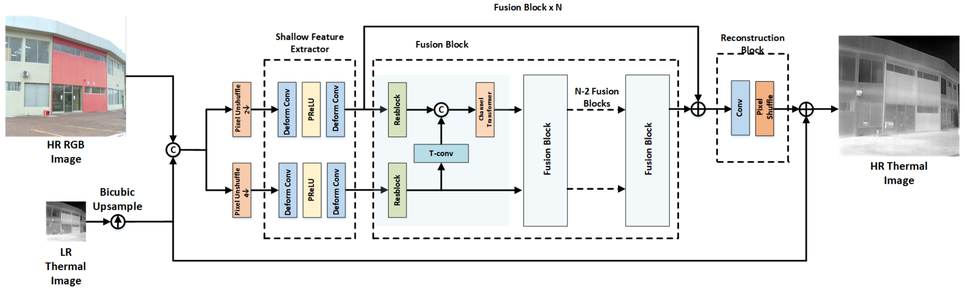

Figure 1- Network Architecture

- The channel attention on our model is inspired by MST++ [2].

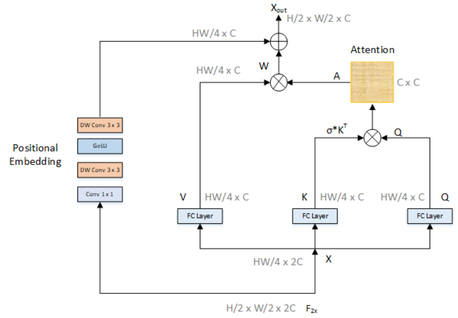

Figure 2 – Channel Attention

Results:

Method	PSNR x8	SSIM x 8	PSNR x 16	SSIM x16	Params (M)	FLOPS (G)
Bicubic	25.17	0.8494	22.04	0.7901	-	-
Restromer	28.72	0.8753	25.39	0.8059	15.08	83.94
AHMF	28.38	0.8676	24.72	0.7790	3.36	11.75
NAFNet	29.16	0.8832	25.50	0.8069	116.34	86.51
MSFFCT (ours)	**29.42**	**0.8879**	**25.90**	**0.8188**	12.17	154.59

Table 1 – Results on PBVS 24 Validation dataset

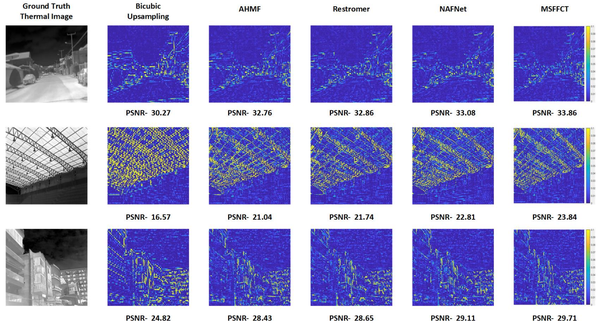

Figure 1 – Qualitative Results on PBVS 24 Validation dataset

Conclusion:

- We address the limitations of CNN- limited long range dependency and dynamically able to adapt to an input by introducing a transformer based model.
- Our transformer is channel-wise self-attention which is computationally very efficient..
- Our work achieved 2nd place in terms of PSNR, SSIM for both x8 and x16 PBVS 24 Guided Thermal Image Super Resolution challenge.

References :

1) B. Kathariya, Z. Li and G. V. d. Auwera, "Joint Pixel and Frequency Feature Learning and Fusion via Channel-Wise Transformer for High-Efficiency Learned In-Loop Filter in VVC," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 4070-4083, May 2024, doi: 10.1109/TCSVT.2023.3323483.

2) Cai, Y., Lin, J., Lin, Z., Wang, H., Zhang, Y., Pfister, H., Timofte, R., & Van Gool, L. (2022). MST++: Multi-stage Spectral-wise Transformer for Efficient Spectral Reconstruction.