Supplementary Material: Gated Multi-Resolution Transfer Network for Burst Restoration and Enhancement

Nancy Mehta[1] Akshay Dudhane[2] Subrahmanyam Murala[1] Syed Waqas Zamir[3]
Salman Khan[2,4] Fahad Shahbaz Khan[2,5]
[1]CVPR Lab, Indian Institute of Technology Ropar [2]Mohamed bin Zayed University of AI
[3]Inception Institute of AI [4]Australian National University [5]Linköping University

This supplementary material contains:

- Detailed explanation of the datasets (§1).
- Additional Ablation Study (§2).
- Difference with prior works (§3).
- More qualitative results (§4).
- Feature Map Visualizations (§5).
- Future Work (§6).

1. Dataset Details

1.1. Burst Super-resolution

(1) **SyntheticBurst dataset** consists of 46,839 and 300 RAW bursts for training and validation, respectively. Each burst contains 14 LR RAW images generated synthetically from a single sRGB image (each of size 48×48 pixels). Each sRGB image is first converted to RAW camera space using the inverse pipeline [1]. Next, the burst is generated with random translations and rotations. Finally, the LR burst is obtained by applying the bilinear downsampling followed by Bayer mosiacking, sampling and noise addition operations.

(2) **BurstSR** dataset [1] consists of 200 real burst images, with each having 14 LR RAW images. LR-HR pair for this dataset has been acquired with Samsung Galaxy S8 smartphone camera and DSLR camera, respectively. From the acquired 200 RAW burst sequences, crops of spatial size 80×80 have been extracted for obtaining a training set consisting of 5405 images and validation dataset comprising of 882 images.

1.2. Burst Denoising

Following the experimental settings of [2], we utilize 20k samples from the Open Images [6] training set to generate the synthetic noisy bursts of burst size 8 and spatial size of 128×128. We evaluate our approach on the grayscale and color burst denoising datasets in [8] and [10]. Both these datasets contains 73 and 100 bursts, respectively. For both these datasets, a burst is generated synthetically by applying random translations to the base image. The shifted images are then corrupted by adding heteroscedastic Gaussian noise with variance $\sigma_r^2 + \sigma_s x$. Here, x denotes the clean pixel value, and (σ_r, σ_s) denotes the read and shot noise parameters, respectively. While training, the noise parameters $(\log(\sigma_r), \log(\sigma_s))$ are sampled uniformly in the log-domain from the range $\log(\sigma_r) \in [-3, -1.5]$ and $\log(\sigma_s) \in [-4, -2]$. The proposed GMTNet is then evaluated on 4 different noise gains (1, 2, 4 and 8) corresponding to the noise parameters $(\log(\sigma_r), \log(\sigma_s)) \rightarrow$ (-2.2, -2.6), (-1.8, -2.2), (-1.4, -1.8), and (-1.1, -1.5), respectively. *Note that the noise parameters for the highest noise gain,* i.e.*, 8 are unseen during training.* Therefore, the performance of this noise level is an indication of the generalization of the network to unseen noise.

1.3. Burst Low-light enhancement

We train and evaluate our model on the See-in-the-Dark (SID) dataset [3], that consists of short exposure burst raw images taken under extremely dark indoor (0.2-5 lux) or outdoor (0.03-0.3 lux) scenes. All these images are acquired with three different exposure times of 1/10, 1/25 and 1/30 seconds, where the corresponding reference images are obtained with 10 or 30 seconds exposures depending on the scene. We evaluate the performance of our models on the Sony subset, that contains 161, 36 and 93 distinct burst sequences for training, validation and testing, respectively. The number of burst images varies from 2-10 for every distinct scene.

2. Additional Ablation Study

2.1. Effect of our resolution transfer and aligned feature enrichment modules

As shown in Figure 1, without the proposed resolution transfer module (RTM) and the aligned feature enrichment (AFE) module, the super-resolution results on real-world photos are blurry.

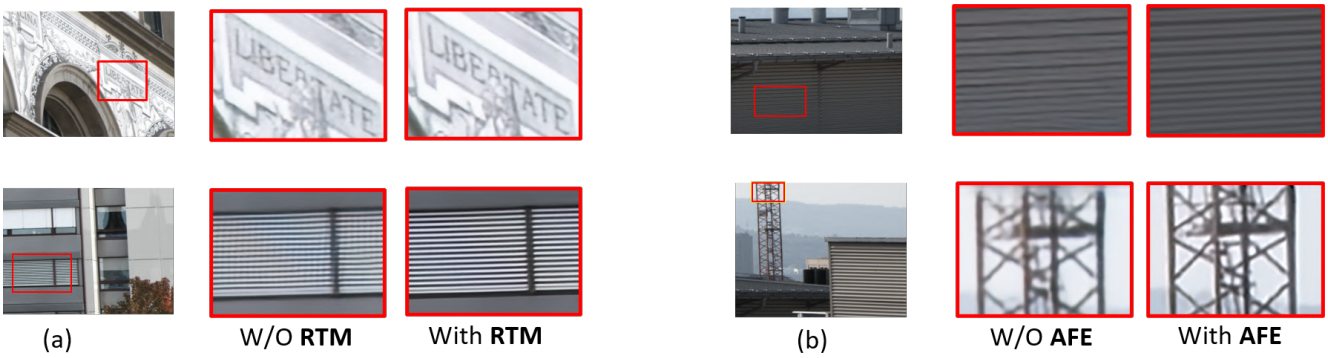

Figure 1. (a) Qualitative Comparison on BurstSR dataset without (W/O) and with Resolution Transfer Merging (RTM) module. Our module benefits from the efficient extraction of features in both low and high-resolution space and makes effective use of information to complete the restoration of the sharp regions, (b) Effect of the proposed aligned feature enrichment (AFE) module in our proposed GMTNet. Images recovered by employing AFE module have lesser artifacts when compared to images obtained without it.

3. Difference with prior works

In this work, we propose a network that jointly performs denoising, demosaicking and super-resolution. We contribute in all the three components of burst processing *i.e.* alignment, fusion and up-sampling. Table 1 highlights the differences between the popular burst restoration approaches [1, 4, 9] and our GMTNet. First, in contrast to the existing alignment modules, the proposed multi-scale burst feature alignment (MBFA) approach denoises and implicitly aligns the burst features at multiple scales using our multi-kernel gated attention (MKGA) and attention-guided deformable alignment (AGDA) modules, respectively. Our MBFA also enriches the aligned burst features through back-projection mechanism and extracts the local and non-local features via encoder-decoder based transformer. As shown in Table 1, existing DBSR [1], EDVR [9] and BIPNet [4] lack some of these properties: feature denoising, multi-scale feature alignment, local-non-local feature extraction, back-projection (feature enrichment), and implicit feature alignment.

Keeping in consideration that different neighboring frames are not equally informative due to occlusion/blurry regions, and misalignment arising from the preceding alignment stage may adversely affect the reconstruction performance, we propose the *transposed attention feature merging* module to adaptively pay attention to the current frame, that is more similar to the reference frame. To capture the similarity between frames, EDVR [9] process the incoming frames in a pyramidal manner through a temporal and spatial attention-based fusion module. This approach [9] focuses more on correlating the frames locally and does not consider the long-range pixel interactions as well as inter-dependencies among different channels. BIPNet [4] propose the pseudo burst mechanism to adaptively merge the frames by considering the inter-dependencies between the channels, but it is quite computationally extensive. The weighted summation-based fusion approach in DBSR [1] applies a softmax function onto the aligned frames, and reference features to attentively assign weights to the aligned and misaligned regions. However, none of the above approaches considers the merits of computing both the **local and non-local**

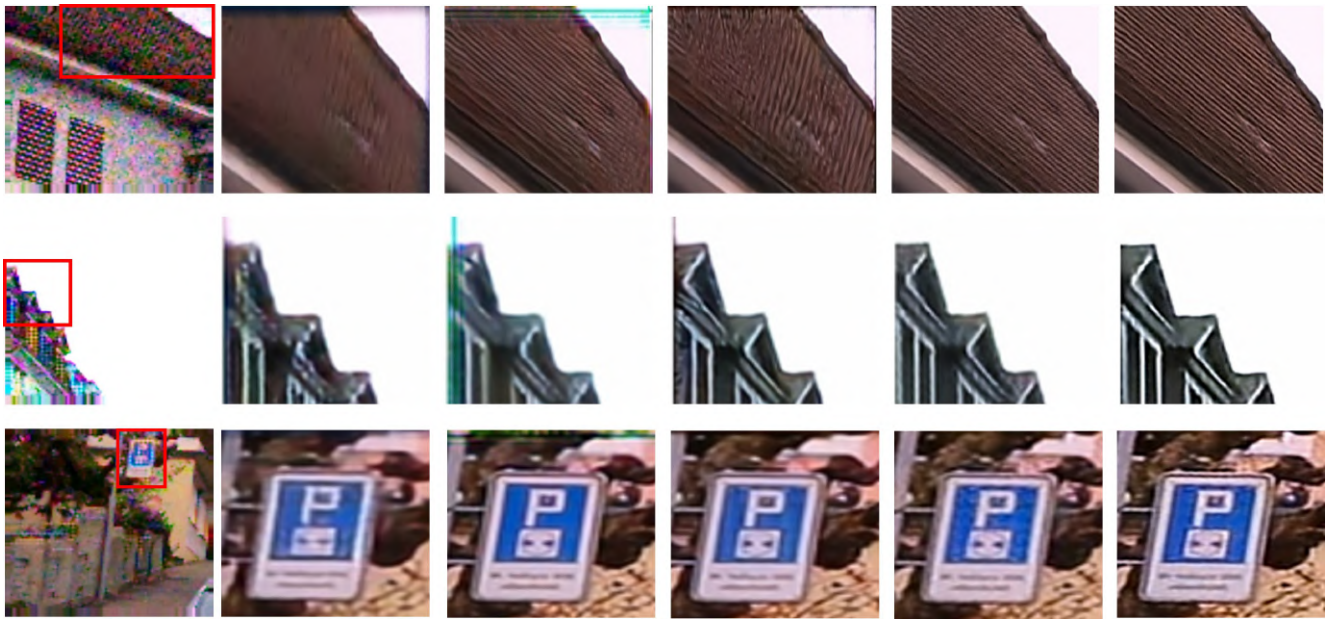

Figure 2. Comparisons for $\times 4$ burst SR on SyntheticBurst [1]. The proposed approach produces more sharper and visually-faithful results than other competing approaches.

correlations among the incoming frames. Our two-stage parallel strategy targets at computing both these relations among the frames by considering **long-range pixel dependencies** between the reference and supporting frames and also within the supporting frames to ease the fusion process.

The up-sampling module generates the output high-resolution image from the fused feature map. Unlike the compared approaches that utilize either **conventional** or **recent up-sampling techniques**, our proposed resolution transfer feature up-sampler (RTFU) considers the merits of utilizing both recent *(slow and more accurate)* and conventional up-samplers *(fast and less accurate)* to get into high-resolution space. Additionally, as shown in the Table 1, the up-sampling strategies in burst restoration methods either use **progressive** up-sampling *(for generating intermediate SR predictions at multiple resolutions)* or **direct** up-sampling, but unlike ours, none of these approaches combines both these strategies in a single network.

Table 1. Comparison between the proposed and existing alignment, fusion and up-sampling modules based on their **properties.**

Tasks	**DBSR [1]**	**EDVR [9]**	**BIPNet [4]**	**Ours**
Alignment	Explicit multi-scale approach, local feature extraction	Implicit multi-scale approach, local feature extraction	Implicit single-scale approach, denoising, local feature extraction, back-projection	Implicit multi-scale approach, denoising, local-non-local feature extraction, back-projection, feature refinement
Fusion	Weighted summation, local feature extraction	Local feature correlation	Inter-frame interaction, local feature extraction	Inter-frame interaction, local-non-local feature correlations
Up-sampling	Direct, pixel-shuffle	Direct, bilinear interpolation	Progressive, transposed convolution	Direct, progressive, pixel-shuffle, bilinear, bicubic interpolation, resolution-transfer

4. Additional Visual Results

We present more images reconstructed by our GMTNet and those of the other competing approaches as qualitative examples for all the considered tasks. The results demonstrated in Figure 2 and Figure 3 clearly show the true potential of our method in successfully recovering fine-grained details from extremely challenging LR images in $\times 4$ burst SR task for synthetic and real images, respectively. Additionally, results in Figure 4 and Figure 5 show that our method performs favorably well on both color [10] and gray-scale [8] noisy images. Particularly, it generates output more closer to the ground-truth compared to the existing SoTA approaches. Further, we provide more visual comparisons for burst low-light enhancement in Figure 6 to show the effectiveness of our model.

Figure 3. Comparisons for $\times 4$ burst super-resolution on Real BurstSR dataset [1]. The proposed approach produces more sharper and cleaner results than other competing approaches.

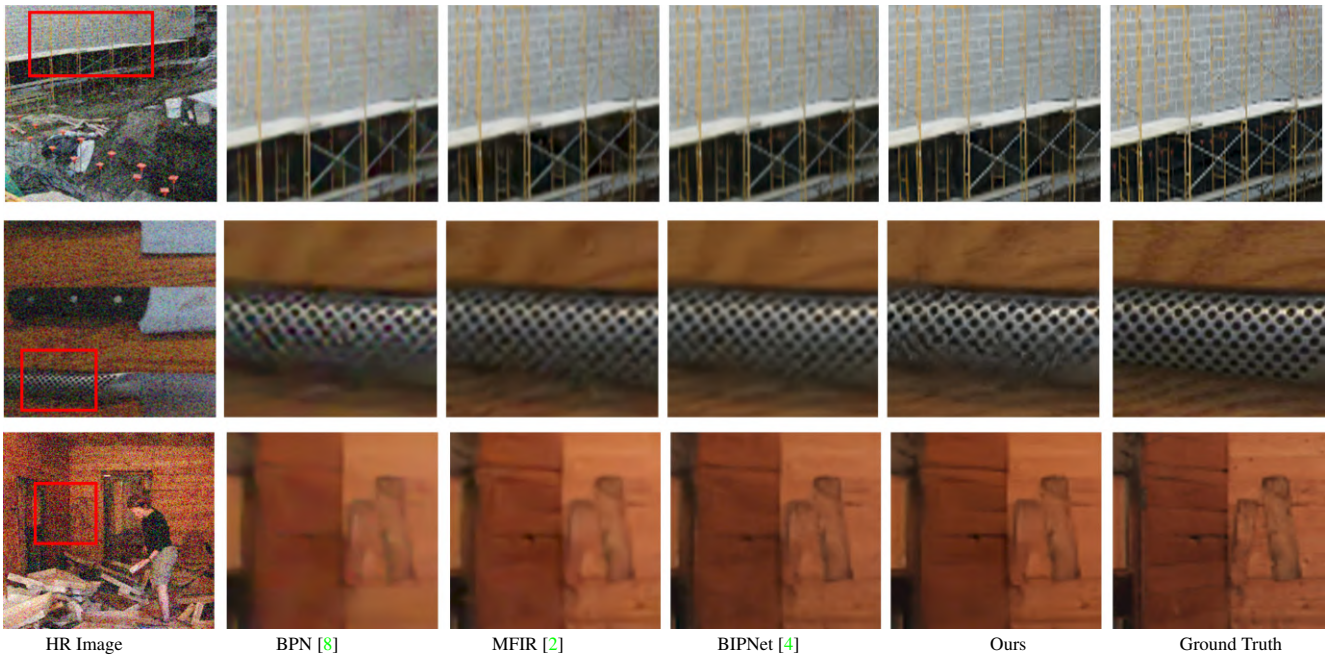

Figure 4. Comparisons for burst de-noising on color dataset [10]. The image reproduction quality of our proposed GMTNet is more faithful to the ground-truth than other competing approaches.

5. Feature Map Visualization

Here, we have visualized the feature maps of all the (14) frames before and after applying the proposed alignment (MBFA) (Fig. 7 and 8). We note that all the unaligned input frames get aligned well with the reference frame, thereby demonstrating the effectiveness of the proposed MBFA module. Similarly, in Figure 9, we provide feature map visualizations before and after applying the resolution transfer merging (RTM) module of the proposed up-sampler (RTFU). It can be seen that for any U_r and U_s pair, the U_s maps have more details than the U_r maps. Our RTM module benefits from the efficient feature extraction in both low and high-resolution spaces. Thus, we can conclude that the proposed modules enhances the feature representations as well as performs the assigned task dedicatedly.

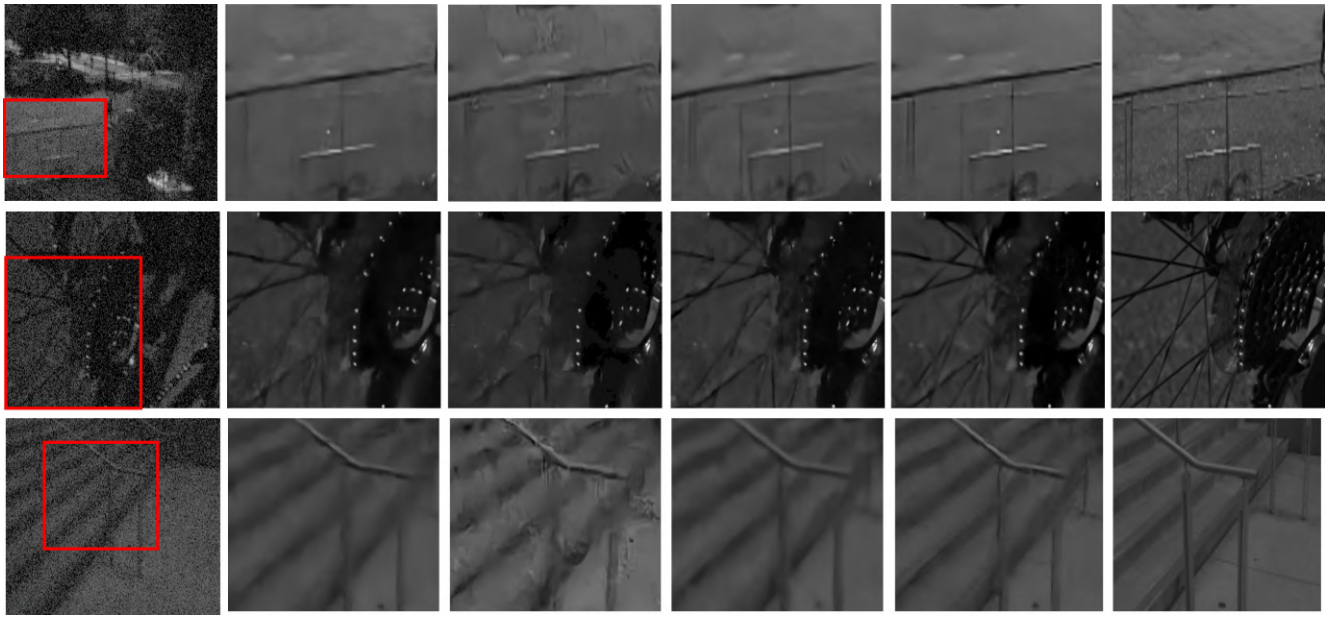

Figure 5. Comparisons for burst de-noising on gray-scale dataset [8]. Our proposed GMTNet preserve the fine image details.

6. Future Work

While GMTNet emerges as a strong backbone architecture across several benchmarks, the proposed modules are extensible and can be well transferred to other burst image restoration applications including de-blurring and satellite imaging. Also, the proposed modules are suitable for majority of video restoration tasks.

Figure 6. Comparisons for burst low-light enhancement on SONY-subset [3]. GMTNet generates sharper result with structural fidelity.

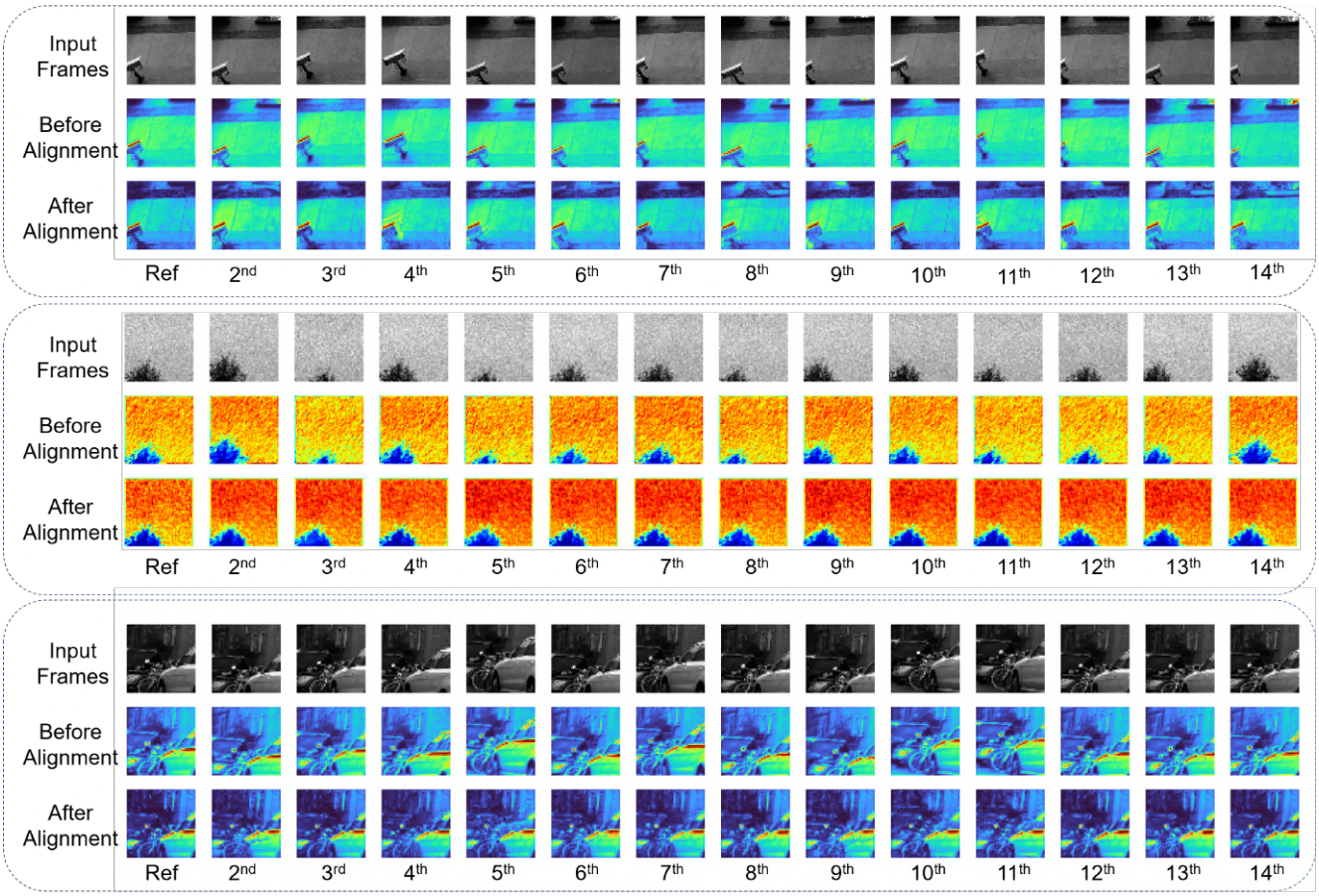

Figure 7. Feature map visualization before and after the proposed multi-scale burst feature alignment (MBFA) module on sample burst images from Synthetic BurstSR dataset. It is clearly observed that the proposed MBFA module aligns the burst frame features implicitly with respect to the reference frame features. **Highlighting point** is that the proposed MBFA module is implicitly trained along with the other modules of proposed network. We have not followed any separate supervision to train the MBFA module for alignment task.

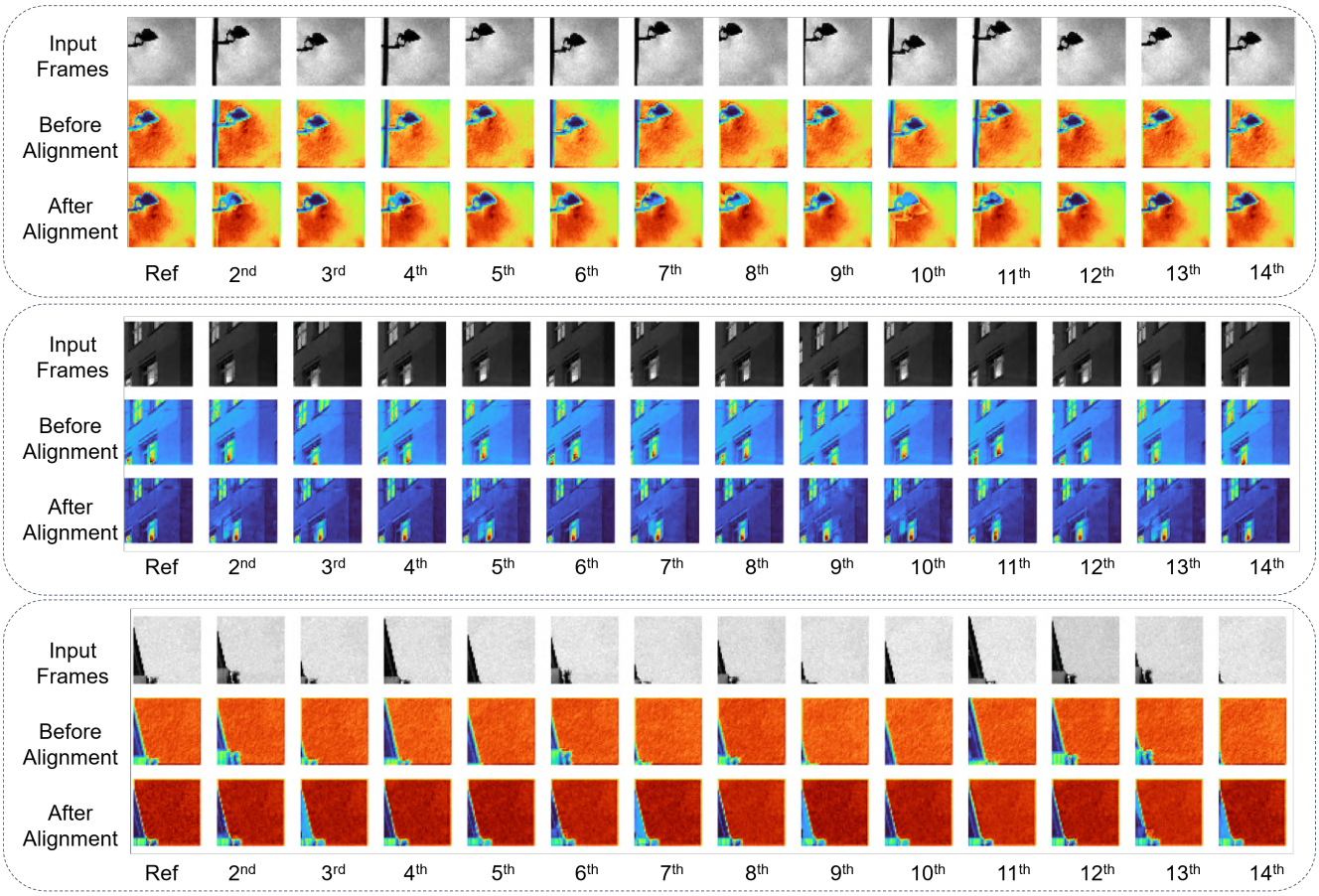

Figure 8. Feature map visualization before and after the proposed multi-scale burst feature alignment (MBFA) module on sample burst images from Synthetic BurstSR dataset. It is clearly observed that the proposed MBFA module aligns the burst frame features implicitly with respect to the reference frame features. **Highlighting point** is that the proposed MBFA module is implicitly trained along with the other modules of proposed network. We have not followed any separate supervision to train the MBFA module for the alignment task.

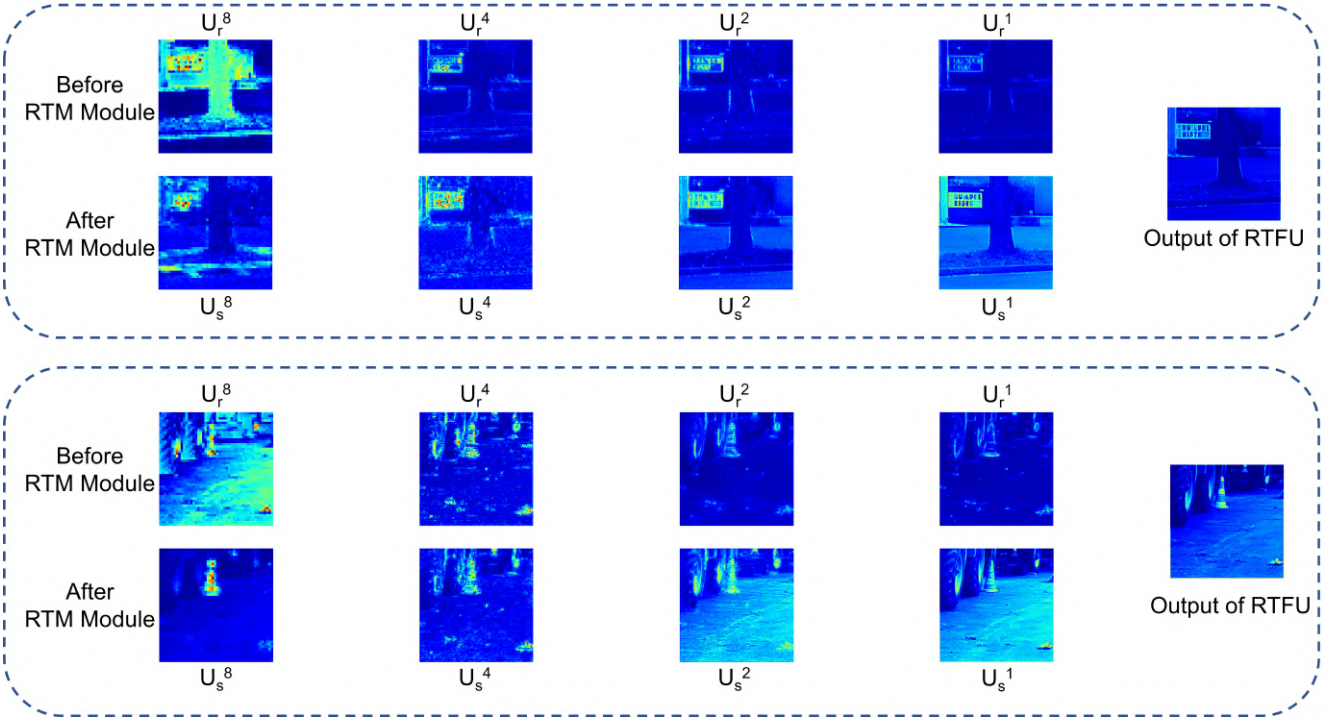

Figure 9. Feature map visualization before and after the proposed resolution transfer merging (RTM) module on sample burst images from Synthetic BurstSR dataset. It is clearly observed that for any pair of U_r and U_s, latter is having more details than former one. The reason behind this is every U_s feature response is obtained by fusion of all the U_r feature responses. Thus, our RTM module benefits from the efficient extraction of features in both low and high resolution space and makes more effective use of information to complete the restoration of sharp regions.

References

[1] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9209–9218, 2021. 1, 2, 3, 4

[2] Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Deep reparametrization of multi-frame super-resolution and denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2460–2470, 2021. 1, 3, 4, 5

[3] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018. 2, 6

[4] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5759–5768, 2022. 2, 3, 4, 5, 6

[5] Ahmet Serdar Karadeniz, Erkut Erdem, and Aykut Erdem. Burst photography for learning to enhance extremely dark images. *IEEE Transactions on Image Processing*, 30:9372–9385, 2021. 6

[6] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2(3):18, 2017. 1

[7] Bruno Lecouat, Jean Ponce, and Julien Mairal. Lucas-kanade reloaded: End-to-end super-resolution from raw image bursts. In *ICCV*, 2021. 3

[8] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2502–2510, 2018. 1, 3, 4, 5

[9] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 3

[10] Zhihao Xia, Federico Perazzi, Michaël Gharbi, Kalyan Sunkavalli, and Ayan Chakrabarti. Basis prediction networks for effective burst denoising with large kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11844–11853, 2020. 1, 3, 4