Deep Burst Super-Resolution

Supplementary Material

Goutam Bhat Martin Danelljan Luc Van Gool Radu Timofte
Computer Vision Lab, ETH Zurich, Switzerland

We provide additional details and analysis of our approach in this supplementary material. In Section 1, we provide additional details about our network architecture. We analyse the impact of sub-pixel shifts in the input images for MFSR in Section 2, while the impact of training dataset for real world SR is analysed in Section 3. Section 4 provides a qualitative analysis of the impact of our training loss (3) used to train our networks on the BurstSR dataset. Additional qualitative comparison with existing super-resolution approaches are provided in Section 5

1. Network Architecture

Here, we provide additional details about our burst super-resolution network architecture.

Encoder: The encoder module maps the packed RAW image $\tilde{b}_i$ to a 64 dimensional feature embedding using a convolution layer. The resulting feature map is processed by 9 residual blocks, before being passed to another convolution layer which expands the feature dimensionality to 512. An illustration of the Encoder module is provided in Figure 1.

Weight Predictor: The weight predictor module computes the un-normalized element-wise fusion weights for each aligned feature embedding $\tilde{e}_i$. It first projects the feature embeddings $\tilde{e}_i$ and $\tilde{e}_1$ to 64 dimensional feature maps $\tilde{e}_i^p$ and $\tilde{e}_1^p$ respectively, using a convolution layer with shared weights. Additionally, the weight predictor module also extracts flow features $\hat{f}_i$ using the flow vectors f_i. The modulo 1 of the flow vectors, f_i mod 1, is first passed through a convolution layer, followed by a residual block to obtain 64 dimensional flow features $\hat{f}_i$. The flow features $\hat{f}_i$, the projected feature embedding $\tilde{e}_i^p$, and the residual $\tilde{e}_i^p - \tilde{e}_1^p$ are then concatenated along the channel dimension, and passed through a convolution layer. The output 128 dimensional feature map is processed by 3 residual blocks, before being passed to a final convolution layer which predicts raw element-wise fusion weights $\tilde{w}$. An illustration of the weight predictor module is provided in Figure 2.

Decoder: The decoder module projects the merged feature map $\hat{e}$ to a 64 dimensional feature space. The projected features are then passed through 5 residual blocks, before being passed to the sub-pixel convolution layer, which upsamples the feature map by a factor $2s$. The sub-pixel convolution layer first increases the feature dimensionality to $2^2 s^2 32$ using a convolution layer. The feature vectors at each spatial location are then re-arranged into a $2s \times 2s \times 32$ map to obtain a 32 dimensional feature map with $2s$ times higher resolution compared to the input. The upsampled feature map is then processed by 4 residual blocks, before being passed to a convolution layer which predicts the output RGB image. An illustration of the Decoder module is provided in Figure 3.

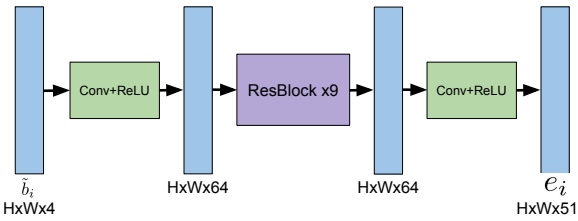

Figure 1. The network architecture employed for the Encoder module E.

	PSNR ↑	LPIPS ↓	SSIM ↑
Ours	**38.61**	**0.084**	**0.941**
No Shifts	37.00	0.106	0.920
Single Image	36.42	0.123	0.913

Table 1. Impact of sub-pixel shifts in the input burst for MFSR. Results are shown on the synthetic test set.

	PSNR ↑	LPIPS ↓	SSIM ↑
Ours	**47.52**	**0.031**	**0.983**
Only Synthetic	44.52	0.081	0.967
Only BurstSR	47.14	0.037	0.981

Table 2. Impact of fine-tuning on real data

2. Impact of input shifts

Here, we investigate the importance of having sub-pixel shifts in the input images for MFSR. We train and evaluate a baseline network **No Shifts** on synthetic bursts generated without any simulated camera motion. That is, all the im-

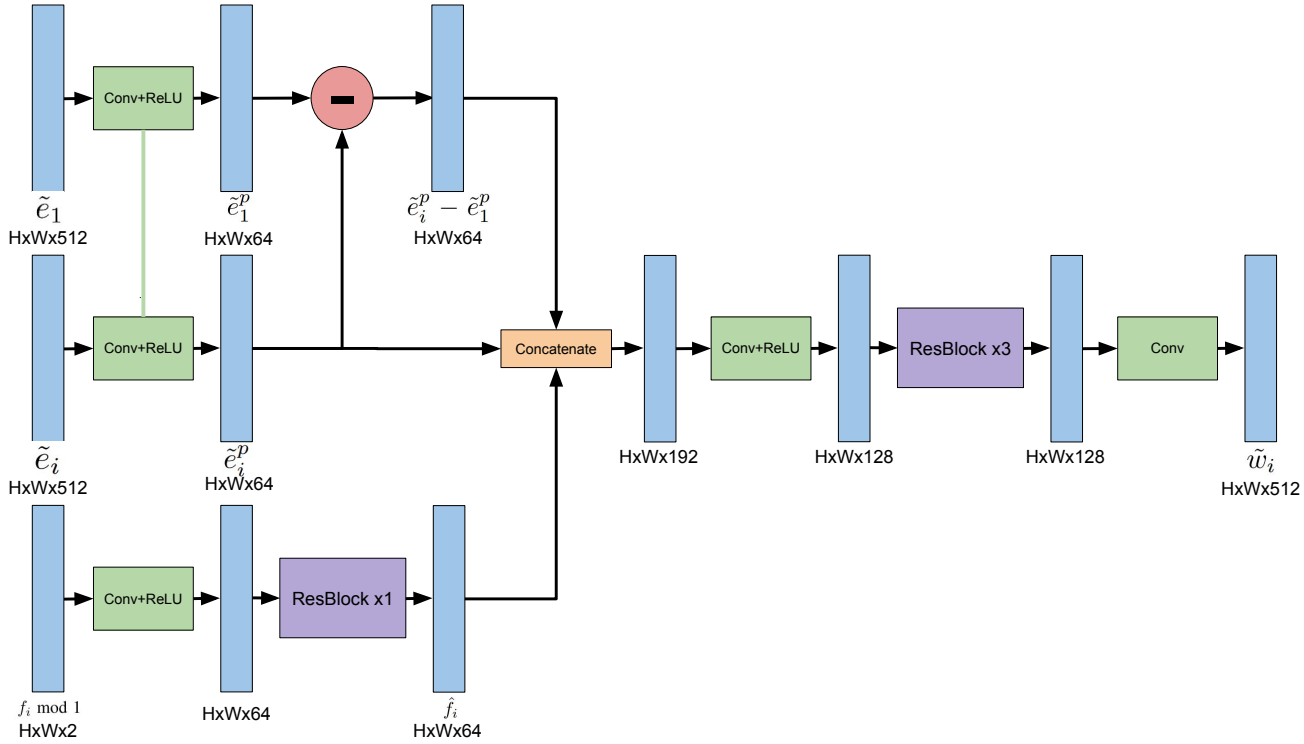

Figure 2. The network architecture employed for the Weight Predictor module W.

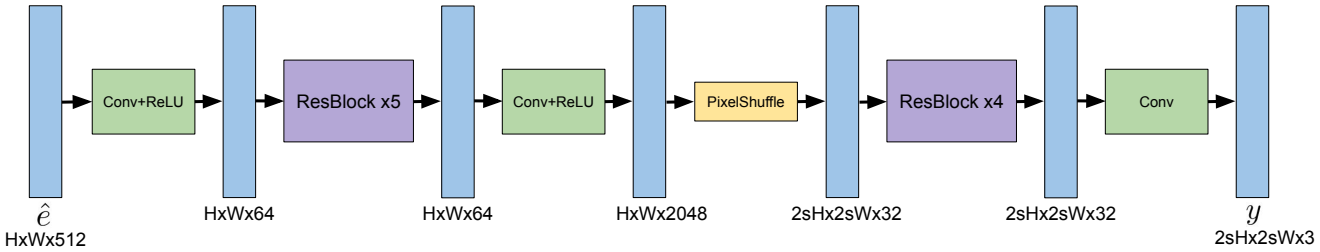

Figure 3. The network architecture employed for the Decoder module D.

ages in the burst are identical except having different independent noise. We also include our SISR baseline for comparison. While the **No Shifts** network can exploit the burst images in order to obtain better denoising, its performance improvement over the SISR baseline is limited to < 0.6 dB (see Table 1). In contrast, our approach obtains a significant improvement of 1.61 dB in PSNR over **No Shifts** when operating on burst with sub-pixel shifts. These results show that the majority of performance gains of our approach over the SISR baseline is obtained by effective fusion of information contained in the different aliased samplings of the scene.

3. Impact of training dataset

We analyse the impact of pre-training our model on the synthetic data, as well as fine-tuning on the real data. We compare our approach with two baselines, i) a network **Only Synthetic** trained using only the synthetic data, and ii) a network **Only BurstSR** trained using only the real-world BurstSR dataset. The results on the BurstSR validation set are shown in Table 2. The network trained only using synthetic data fails to generalize to the real world images, obtaining a PSNR of 44.52 dB. In contrast, the network trained from scratch on BurstSR performs much better with a PSNR of 47.14 dB. The best results are obtained when combining both the strategies: pre-training first using large scale synthetic data, and finetuning the resulting network on real data.

4. Impact of our training loss

In this section, we analyze the impact of our training loss, defined in Eq. (3) in the main paper, which is used

to train our model on the real-world BurstSR dataset. Our loss aligns the network prediction to the ground truth image in order to handle spatial misalignments between the input burst and the ground truth. Furthermore, it also handles the color mismatch between the input-ground truth pair by estimating the color mapping function between the two. We compare the network trained using our loss (3) with a network trained using direct pixel-wise loss without performing any explicit spatial alignment and color space correction. Additionally, we also include a network trained only on synthetic data for comparison. The results of this analysis on the BurstSR validation set are shown in Figure 4. Compared to using direct pixel-wise loss, the network training using our loss (3) can generate sharper images with better details.

5. Qualitative Examples

Here, we provide additional qualitative comparison of our approach with the approaches described in Section 6.2 of the main paper; (i) Single Image baseline, (ii) DeepJoint [2]+RRDB [3], and (iii) HighRes-net [1]. Visual examples from the BurstSR test set are shown in Figure 5. Compared to the other methods, our approach can best reconstruct the high frequency image details with high fidelity to the high-resolution ground truth.

References

[1] Michel Deudon, A. Kalaitzis, Israel Goytom, M. R. Arefin, Zhichao Lin, K. Sankaran, Vincent Michalski, S. Kahou, Julien Cornebise, and Yoshua Bengio. Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery. *ArXiv*, abs/2002.06460, 2020. 3

[2] Michaël Gharbi, Gaurav Chaurasia, Sylvain Paris, and F. Durand. Deep joint demosaicking and denoising. *ACM Transactions on Graphics (TOG)*, 35:1 – 12, 2016. 3

[3] Xintao Wang, K. Yu, Shixiang Wu, Jinjin Gu, Yi-Hao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and X. Tang. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV Workshops*, 2018. 3

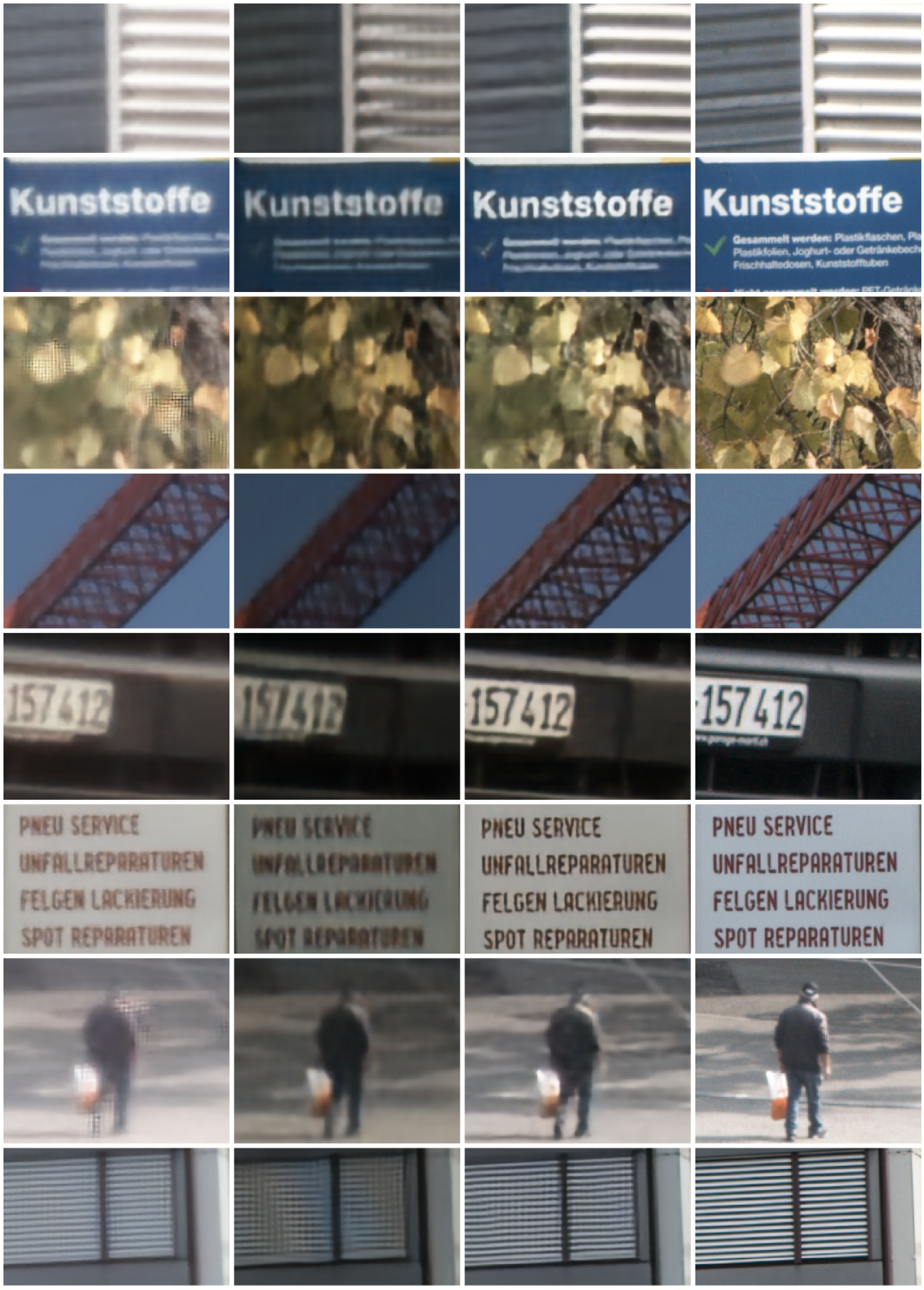

Figure 4. Qualitative comparison of a network trained on BurstSR dataset using our training loss (3) with a network trained using direct pixel-wise loss on the BurstSR validation set. A network trained only on the synthetic dataset is also included for comparison. Note that there is a color shift between the predictions of the networks, as the networks are trained using different output color spaces. Hence, we encourage the reader to focus on image details, *e.g*. sharp edges, presence of artifacts and not on the color space differences.

Figure 5. Qualitative comparison of our approach with existing super-resolution approaches on the BurstSR test set.