Mosaic Super-resolution via Sequential Feature Pyramid Networks

Mehrdad Shoeiby [1], Mohammad Ali Armin [2], Sadegh Aliakbarian [2], Saeed Anwar [2], and Lars petersson [2]

[1]CSIRO-DATA61
[2]Affiliation not available

October 30, 2023

Abstract

Advances in the design of multi-spectral cameras have led to great interests in a wide range of applications, from astronomy to autonomous driving. However, such cameras inherently suffer from a trade-off between the spatial and spectral resolution. In this paper, we propose to address this limitation by introducing a novel method to carry out super-resolution on raw mosaic images, multi-spectral or RGB Bayer, captured by modern real-time single-shot mosaic sensors. To this end, we design a deep super-resolution architecture that benefits from a sequential feature pyramid along the depth of the network. This, in fact, is achieved by utilizing a convolutional LSTM (ConvLSTM) to learn the inter-dependencies between features at different receptive fields. Additionally, by investigating the effect of different attention mechanisms in our framework, we show that a ConvLSTM inspired module is able to provide superior attention in our context. Our extensive experiments and analyses evidence that our approach yields significant super-resolution quality, outperforming current state-of-the-art mosaic super-resolution methods on both Bayer and multi-spectral images. Additionally, to the best of our knowledge, our method is the first specialized method to super-resolve mosaic images, whether it be multi-spectral or Bayer.

Mosaic Super-resolution via Sequential Feature Pyramid Networks

Mehrdad Shoeiby
CSIRO-Data61
mehrdad.shoeiby@data61.edu.au

Mohammad Ali Armin
CSIRO-Data61
ali.armin@data61.edu.au

Sadegh Aliakbarian
Australian NAtional University

Saeed Anwar
CSIRO-Data61

Lars Petersson
CSIRO-Data61

Abstract

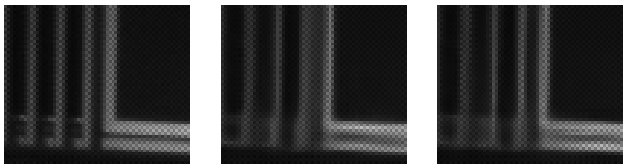
Advances in the design of multi-spectral cameras have led to great interests in a wide range of applications, from astronomy to autonomous driving. However, such cameras inherently suffer from a trade-off between the spatial and spectral resolution. In this paper, we propose to address this limitation by introducing a novel method to carry out super-resolution on raw mosaic images, multi-spectral or RGB Bayer, captured by modern real-time single-shot mosaic sensors. To this end, we design a deep super-resolution architecture that benefits from a sequential feature pyramid along the depth of the network. This, in fact, is achieved by utilizing a convolutional LSTM (ConvLSTM) to learn the inter-dependencies between features at different receptive fields. Additionally, by investigating the effect of different attention mechanisms in our framework, we show that a ConvLSTM inspired module is able to provide superior attention in our context. Our extensive experiments and analyses evidence that our approach yields significant super-resolution quality, outperforming current state-of-the-art mosaic super-resolution methods on both Bayer and multi-spectral images. Additionally, to the best of our knowledge, our method is the first specialized method to super-resolve mosaic images, whether it be multi-spectral or Bayer.

1. Introduction

Real-time snap-shot mosaic image sensors are a category of imaging devices that encompass modern RGB and multi-spectral cameras. In fact, RGB cameras are a sub-category of multi-spectral cameras that are only capable of measuring three spectral channels *i.e*. red, blue and green.

The recent improvements in these devices has given rise to multi-spectral cameras with the performance comparable to modern RGB cameras in terms of size, portability and speed [33].

Despite the great interest in these devices, with applications ranging from astronomy [2] to tracking in autonomous vehicles [30, 24], they suffer from an inherent constraint: a trade-off between the spatial and the spectral *resolution*. The reason is the limited physical space on 2D camera image sensors. A higher spatial resolution (smaller pixel size) reduces the number of possible wavelength channels on the image sensor, and thus creates a limitation in certain applications where the size and portability are essential factors, for instance, on a UAV [4]. A more portable (*i.e*. smaller and lighter) camera suffers more from lower spatial and spectral resolution. The spectral and spatial constraints of mosaic imaging devices motivates the need for super-resolution (SR) algorithms for the type of images that these devices create (mosaic images). Nevertheless, compared to the amount of existing literature on normal RGB (interpolated/demosaiced) RGB images SR, few efforts have been made toward mosaic image super-resolution (SR).



a) b) c)

Figure 1. Comparison between the baseline [26] and our Mosaic super-resolution. a) Ground-truth, b) baseline, and c) our mosaic super-resolution. The baseline smooths out the bayers pattern and fuses the information while our method produces the results similar to the ground-truth

The related literature aiming at the task of SR in RGB domain (discussed in Section 2) is predominantly carried out on interpolated/demosaiced RGB images. However, with most modern RGB cameras, mosaic Bayer images can be obtained instead of demosaiced images. In this regard, various studies [39, 8] have pointed out and discussed that interpolation or demosaicing deteriorates SR performance due to *(1)* removing high-frequency information as interpolation/demosaicing can be viewed as a form of low

pass filtering [14], while SR aims at predicting such high-frequency information; and *(2)* interpolation/demosaicing introduces artifacts [39, 14] which can be either viewed as a signal loss or noise in the input image. The SR literature on mosaic RGB images, despite its importance, is limited to a few recent works [18, 28, 36]. We believe that in most modern SR applications such as microscopy and astronomy, in which having access to high-resolution images is vital, it is counter-intuitive to throw away high-frequency content mosaic images and only rely on the interpolated/demosaiced images.

Surprisingly, despite its significance, mosaic image SR has been less considered in the literature. This, however, motivates us to conduct an in-depth study on how to get benefit of such vital information. Therefore, in this paper, we propose a SR framework for real-time mosaic image sensors (cameras) to bridge the gap identified in the literature. We believe that our approach will benefit many applications, such as in astronomy [21] or microscopy [25] in which high quality super-resolved images are essential. To summarize, our primary contributions are:

- We demonstrate an effective use of the ConvLSTM module to learn the relationship between features from different stages of a CNN with sequentially increasing receptive fields, and achieve state-of-the-art in mosaic SR.
- To the best of our knowledge, our method is the first addressing SR of mosaic images directly. We also demonstrate the different nature of mosaic images compared to demosaiced images by showing that methods specifically designed for mosaic SR does not necessarily perform well on demosaiced/interpolated RGB images.

As a secondary contribution, we experiment with different attention mechanisms and assess their relative performance in our network. Furthermore, investigating the structure of an LSTM module, we discover that elements of it are designed to apply implicit attention. By incorporating our LSTM inspired attention mechanism in our network, we empirically show its superior performance compared to other attention mechanisms.

2. Related Work

The RGB super-resolution methods dominate the SR literature; therefore, we first review the literature that focus on RGB SR and then discuss the more related existing literature on mosaic SR. One of the first works in RGB CNN based SR (SRCNN) [5], although simply composed of three convolutional layers, significantly outperformed the conventional SR algorithms. Following the success of SRCNN [5], many CNN based algorithms [22, 6, 35] were developed. For example, fast SRCNN (FSRCNN) [6] encompassing eight convolutional layers, speeds up the SR process by using as input the original low-resolution patch instead of a bi-cubically upsampled one. They highlight the fact that using interpolation to scale up images deteriorates the SR performance. Note that, extending from the discussion in [6], RGB images are, in fact, interpolated from mosaic Bayer images, and hence, super-resolving directly from the raw mosaic images instead should result in better performance.

Current approaches, similar to ours, use residual connections [15, 22, 1]. For example, [15] introduced very deep SR (VDSR), which has a single global skip-connection from the input to the final output. Similarly, enhanced deep SR *i.e*. EDSR [22] employs residual blocks (RBs) with short skip connections. More recently, a cascading residual network (CARN) [1] is proposed, which also employs a variant of RBs with cascading connections. While CARN [1] is lagging behind EDSR [22] in terms of PSNR, it improves efficiency and speed. More lately, motivated by the success of DenseNet [13], CNN-based SR networks have concentrated on the dense connection model. For example, SRDenseNet [31] uses dense connections to learn compact models, avoiding the problem of vanishing gradients and eases the flow from low-level features to high-level features. Recently, the residual-dense network (RDN) [38] employed dense connections to learn the local representations from the patches at hand. The dense network with multi-scale skip connections has similarities with our proposed method in terms of feature aggregation. However, their method aggregates features, whereas our method aggregates a sequence of features with sequentially increasing receptive fields, and uses a ConvLSTM module [34] to learn these sequential features.

The visual attention [23] concept in SR was introduced by RCAN [36], which models the inter-channel dependencies using a channel attention (CA) mechanism. This process is coupled with a very deep network composed of groups of RBs called RGs (RGs). Following in the footsteps of [36], the residual attention module (SRRAM) by [16], employed spatial attention as well as CA while still lagging behind RCAN [36]. Most recently, Second-Order Attention Network (SAN) [3] was introduced. The authors argue that the task of SR has achieved outstanding performance; however, at the expense of using very deep and wide networks. They argue that maybe a better utilization of intermediate features would help improve the results. Note that better utilization of intermediate features was brought up by RCAN; in fact, this was precisely the incentive behind CA in the RCAN setup. Nevertheless, the SAN authors [3] propose a second-order attention network within a residual in residual structure. The quantitative results are, on average, on par with RCAN, and in terms of the network

size, SAN is only marginally smaller than RCAN (15.7M vs. 16M parameters).

The SR algorithms above focus mainly on super-resolving RGB images even though, as discussed before, the multi-spectral images are comparatively more adversely affected by the resolution constraints. Reviewing the limited multi-spectral/mosaic SR literature, one would realize that these networks are not structurally different *i.e*. they are not fine-tuned for mosaic images by taking into account any spectral correlation of different channels. The scarcity of multi-spectral SR algorithms may be due to the absence of multi-spectral SR benchmarking platforms, as well as the difficulty of accessing suitable SR spectral datasets. For instance, [20] aims to improve the quality of hyperspectral (not multi-spectral) images and is one of the few CNN based spectral SR methods. To the best of our knowledge, the only multi-spectral SR methods [19, 26] were submitted to the PIRM2018 multi-spectral SR challenge [28, 29]. The work in [19] adopted an image completion technique that requires $\times 2$ and $\times 3$ down-sampled images as input to a 12 layer convolutional network. While achieving good results, it addresses the problem of $\times 3$ SR given $\times 2$ and $\times 3$ down-sampled images, rather than single-image SR. It is also not an end-to-end CNN based implementation. The best end-to-end CNN based method in the challenge was proposed by [26], which implicitly employed the RCAN [36], which is the current state-of-the-art in multi-spectral SR.

The main body of the RCAN structure constitutes a sequence of RGs, with the receptive field increasing sequentially, that is, at a deeper RG it sees a larger area of the input image. This can be considered as different levels of an image pyramid. In one of our main contributions, we propose that a convolutional LSTM (ConvLSTM) [34] can learn the sequential relationship between these pyramidal feature levels. A superficially similar idea was used for optical flow estimation. To elaborate, the authors refer to *pyramid convolutional LSTMs* [10] in their structure, in which the pyramid part uses one convolutional LSTM at each semantic level to generate features by taking the number of input frames as the step size for the LSTMs. On the contrary, we only use one Convolutional LSTM with the step size being the number of semantic levels (RGs) that are being considered. Our second contribution is the use of a convolutional LSTM at the input of the network to learn the sequential relationship between different wavelengths of the network. Our third contribution is a self-attention mechanism that is inserted between the RGs of the RCAN structure.

3. Method

Inspired by the success of RCAN [26] in multi-spectral SR, we consider a simplified RCAN as the backbone for our method and develop our framework on top of it. Briefly, the multi-spectral RCAN consists of five RGs (RGs), and each RG has three RBs and each RB has one CA. We empirically observed that the features from each RG can be utilized better if higher level of aggregations are considered. In other word, we found that although processing features in a feed-forward manner has shown promising performance, one can better utilize the intermediate features if the dependencies between different RGs are taken into account. To this end, we treat the output of each RG as a separate representation, processing them in a pyramid Bidirectional ConvLSTM to learn relations between features of various receptive fields. We also observed that, despite the necessity for an attention mechanism, the CA used in the original RCAN cannot effectively learn to highlight the informative part of the feature vectors (as demonstrated in our experiments). Hence, we also design an attention mechanism, inspired by the internal operations of a ConvLSTM, and apply it between different RGs. In this section, we discuss different components of our model in detail.

3.1. Bidirectional Pyramid ConvLSTM

Bidirectional LSTMs have shown promising performance in learning long-range dependencies in a sequence. As the name implies, this class of models is capable of learning such relations in both the forward and backward directions, providing the model with stronger representation compared to its unidirectional counterpart. In our case, we propose to treat the output of each RG in our backbone as a sequence of feature maps. In other words, the features at different receptive fields act as the features at different time-steps. Since our feature maps carry strong spatial information, we utilize a bidirectional ConvLSTM [34].

A ConvLSTM takes a sequence (in our case pyramid receptive fields which are output features of the RCAN RGs X_t) as input and apply the following operation on them:

$$
\begin{aligned}
i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \odot C_{t-1} + b_i) \\
f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \odot C_{t-1} + b_f) \\
C_t &= f_t \odot C_{t-1} + i_t \odot tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \\
o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \odot C_t + b_o) \\
H_t &= o_t \odot tanh(C_t)
\end{aligned}
\tag{1}
$$

where i_t, f_t and o_t are input, forget, and output gate of a ConvLSTM cell. C_t denotes the cell state which is passed to the next ConvLSTM cell and H_t indicates the output features of a ConvLSTM cell. Here $*$, and $\odot$ refers to the convolution operation and Hadamard product. σ is a Sigmoid function. Our Bidirectional ConvLSTM has five steps for the features of 5 RGs, and it maintains two sets of hidden and state cells per unit, for back and forth sequences. This allows Bidirectional ConvLSTM to have access to receptive field contexts in both directions and therefore increases the performance of the proposed network.

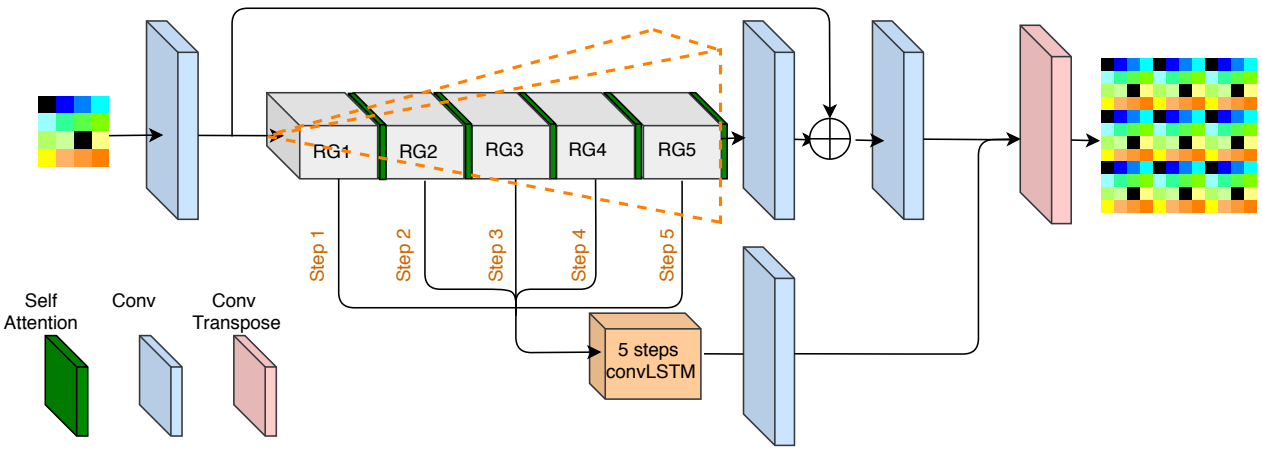


Figure 2. Overview of our proposed SR network. Our model gets as input an LR mosaic image and $\times$3super-resolves it to HR mosaic image. Our network is based on RCAN [26] with five residual groups (RGs) and three residual blocks (RBs).

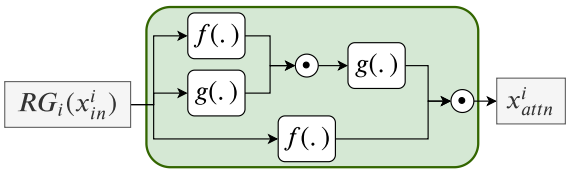


Figure 3. Overview of our self-attention mechanism: Adhering to LSTM intuition, $f(.)$ is a $Sigmoid$ function and $g(.)$ is a $Tanh$ function.

Since the features representing each time-step carry information at different receptive fields (with respect to the input), we consider our design of ConvLSTM as a *Pyramid*, thus naming this component Pyramid ConvLSTM.

3.2. Self attention mechanism

As discussed earlier in Section 3, we expect an attention mechanism to play a tangible role in SR. Following this expectation, the original RCAN [39] architecture employs CAs inside RGs. When delving into the effect of each component in the backbone, we observed an inconsistency in the effect of CA in performance. Depending on the data presented, e.g., in mosaic format or cube format, Bayer or Multi-spectral, the effect of CA varies. Also, CA in RCAN is deployed 3 times in each RG. Therefore, an attention mechanism between the RGs may be exploited as a way to achieve more efficient flow of information. This observation inspired us to investigate what type of attention mechanisms better suits the problem at hand.

To this end, we first investigate existing attention mechanisms [36, 12, 7] that have shown reasonable results on different computer vision problems. The failure of existing attention mechanisms in SR suggests that the nature of our problem is different from the ones cited above. By studying the LSTM structure [11] carefully, we realize that an LSTM by design provides implicit attention to input features and selectively passes more vital information to the next stage. The structure in Figure 3 is equivalent to an LSTM cell with only one step and with zero-initialized hidden and cell states. If we insert this structure between different RGs (the green block in Figure 2) the first $Tanh$ followed by a $Sigmoid$ applies a non-linearity on the input feature map and then performs a gating operation, determining what information is needed to be passed to the next stage. We repeat this process twice to provide additional non-linearity and gating, which follows the intuition behind LSTM operations. To this end, and inspired by the gating inside a convolutional LSTM, our self-attention mechanism gets as input the output of each RG and applies the following function:

$$\begin{aligned} x^i_{attn} = & f(RG_i(x^i_{in})) \odot \\ & g(f(RG_i(x^i_{in}) \odot g(RG_i(x^i_{in})) \end{aligned} \tag{2}$$

where RG_i is the i^{th} RG, x^i_{in} is the input feature map to the i^{th} RG, and x^i_{attn} is the resulting feature map for its corresponding input. To stay with the logic behind LSTMs [11], in our design, f and g are the non-linear functions of $Sigmoid$, and $Tanh$ respectively.

As mentioned before, the attention mechanism shown in Eq. 2, can be considered equivalent to the internal operations of a convolutional LSTM when the cell states carry no information. This is well-suited to our task since we do ignore any relation to other RGs and computing the refined features based only on the output of the current RG, acting as *self*-attention.

3.3. Loss functions

In SR literature, a simple loss function such as L_1 [36] or L_2 [26, 19], or a perceptual loss function such as SSIM [32] is usually utilized to train models. Here, for consistency, we choose L_1 loss as our baseline loss function since an L_1 function is less sensitive to outliers compared to an

L_2 function. We use the PyTorch $SmoothL_1$ implementation, which is a more stable implementation compared to the vanilla L_1. $SmoothL_1$ can be expressed as

$$SmoothL_1(\Theta) = \frac{1}{N}\sum_{i=1}^{M} Z^i \tag{3}$$

where

$$Z^i = \begin{cases} 0.5 \times (DIF)^2 & if|DIF| < 1 \\ |DIF| - 0.5 & otherwise, \end{cases}$$

and $DIF = HR^i_{RGB} - LR^i_{MS}$.

4. Dataset generation

Multi-spectral dataset. We generate HR and LR mosaic images from HR multi-spectral cubes in the StereoMSI dataset [29], by taking the spatial offset of each wavelength in a multi-spectral pixel into account [29]. The HR multi-spectral cubes have a dimension of $14 \times 240 \times 480$, and LR$\times 3$ have a dimension of $14 \times 80 \times 160$. The multi-spectral images have 16 channels and each pixel exhibit a 4×4 pattern [29]. However, as a result of the particular camera that captured these images, two of these channels are spectrally redundant and are not present, leaving us with 14 channels. Following the spatial location provided in [29], we transform this 14 channel cube to a mosaic pattern. For the two redundant wavelengths, we assign zero value. In Figure 2, these two wavelengths are indicated by black pixels. The resulting HR and LR mosaic patterns have dimensions $1 \times 960 \times 1920$, and $1 \times 320 \times 640$ respectively.

Bayer dataset Regarding the Bayer dataset [29], an extended StereoMSI dataset has become available. The size of the images is 1086×2046. To generate LR mosaic images, the HR mosaic was used to build an image cube of size $4 \times 362 \times 682$. The 4 channels correspond to two green, one red, and one blue. The image cubes were down-sampled and used to reconstruct LR mosaic images.

Converting multi-spectral mosaics to zero-padded multi-spectral cubes. In a recent multi-spectral color-prediction work [27], the authors proposed converting multi-spectral mosaics into the format of zero-padded multi-spectral cubes (for simplicity, we refer to this format as *mosaic cubes*) as a way to give the network an extra wavelength dimension and they showed that this data representation helps achieve better performance; We verify that this data representation indeed helps us boost our quantitative results for multi-spectral SR. Please note that we use the cubic data format as the input, and the corresponding, actual, mosaic images are used as ground truth. In the case of Bayer SR, we did not observe any improvements. Although it remains to be demonstrated, the reason could be that Bayer pixels contain two green pixels with identical wavelengths. Hence, for Bayer SR we use simple mosaic images.

5. Experiments

5.1. Settings

Dataset: We evaluate our approach on the PIRM2018 spectral SR challenge dataset [29, 28] as our multi-spectral mosaic SR evaluation. We use their extended Bayer images available for RGB mosaic SR evaluation. With the multi-spectral dataset, we have 350 HR-LR image pairs with 300 images used for training and 30 and 20 images set aside for validation and testing, respectively. For the Bayer dataset, to stay within a comparable dataset size, we have 300 training image pairs, 50 for validation, and 50 for testing.

Evaluation metrics: The 20 (multi-spectral) and 50 (Bayer) test images were super-resolved to a scale of $\times 3$ and evaluated using the Pixel Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) metrics. For the SSIM metric, a window size of 7 is used with the metric being calculated for each channel and then averaged.

Training settings: During training, we performed data augmentation on each mini-batch of 16 images which included random 60×60 cropping of the input image, random rotation by $0°$, $90°$,$180°$, $270°$ with $p = 0.25$, and random horizontal flips with $p = 0.5$. Our model is trained by the ADAM optimizer [17] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The initial learning rate was set to 10^{-4} and then halved every 2500 epochs. To implement our models, we used the PyTorch framework. To test our algorithms, we select the models with the best performance on the validation set and present the test results for those models.

6. Results and Discussion

6.1. Multi-spectral mosaic SR

To assess the effect of using a mosaic vs. mosaic cube format [27], In Table 1, we first train our baseline RCAN on mosaic data and mosaic cubes with LR input (the mosaic format is always used for HR ground truth), with and without CA (the first four rows). As explained before, and according to the results, the CA mechanism improves the performance more when using a mosaic cube data format compared with the mosaic data format. Moreover, the zero-padded cube data format improves the results compared to mosaic data by $0.04dB$. The rest of the experiments in Table 1 are carried out with a zero-padded cube data format as the input, and mosaic data format as the output. In all the tables, best and second-best results are shown in **bold** and underlined fonts, respectively.

The fifth row of Table 1 shows the effect of our Pyramid ConvLSTM structure, which has led to a considerable $0.08dB$ improvement in PSNR. The utilization of our

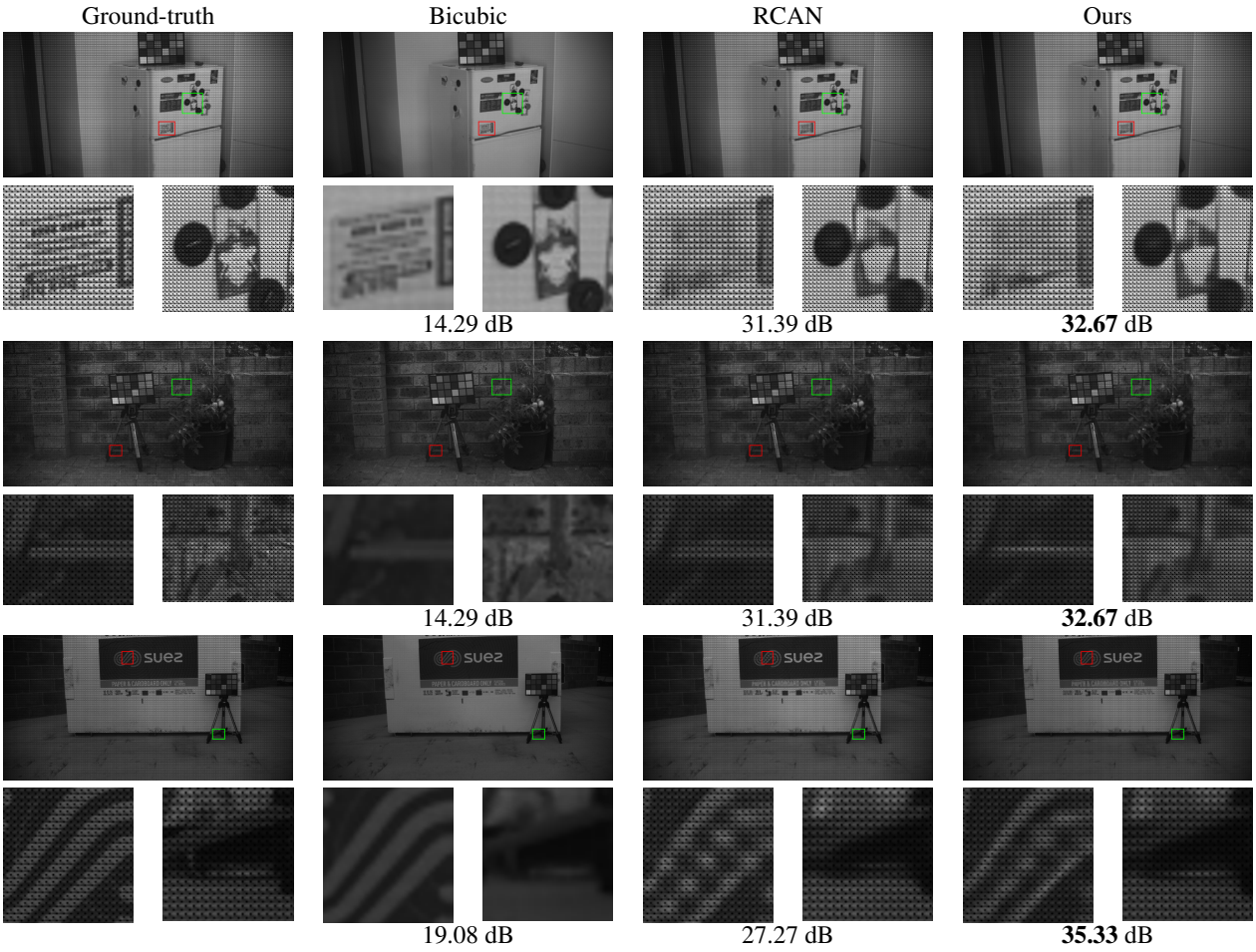


Figure 4. Qualitative results for Bayer SR. Since the images are grayscale by nature, the results are best seen when zoomed in. Note, PSNR results per baseline are provided in the corresponding columns.

proposed ConvLSTM inspired attention module, (lstmA), boosts the results by an additional $0.02dB$. In total, our proposed method boosts the SOTA approaches by $0.10dB$. Taking into account the effect of using mosaic cubes, our improvement adds up to $0.14dB$. Note that compared to the top PIRM2018 algorithms, our algorithm clearly outperforms existing methods. It is worth noting that no self-ensemble algorithm was used in the post-processing stage to achieve further improvements. These results purely demonstrate the effectiveness of our Pyramid ConvLSTM module boosted slightly by our lstmA module. Qualitative results, depicted in figure 4, are also evident of the superiority of our method.

6.2. Bayer SR

As explained earlier, we use the mosaic data format for the Bayer SR task. This is based on our observation in which no improvement is obtained when using the mosaic cube data format. We hypothesize the reason is that Bayer pixels contain 2 green pixels with identical wavelengths, thereby defying the logic behind using mosaic cubes [27]. The results of our experiments are provided in Table 2. The first two rows demonstrate the effect of CA, indicating that the model may not be able to take advantage of CA when uses Bayer data. Overall, our Pyramid ConvLSTM structure, together with the lstmA module, outperforms the baselines in terms of PSNR metric by $0.07dB$. Qualitative results, depicted in figure 5, are also evidence of the superiority of our method. Note, to the best of our knowledge, there are no specialized SR algorithms available on Bayer SR. Hence, we only compare with a bicubic baseline, which is customary in SR literature as well as the RCAN implementation [26] that is SOTA in multi-spectral SR (RCAN [37] is also SOTA in standard RGB SR). The closest algorithm to ours, as mentioned in section 2, is [39], which carries out joint demosaicing, and SR *i.e.* does not produce mosaic images).

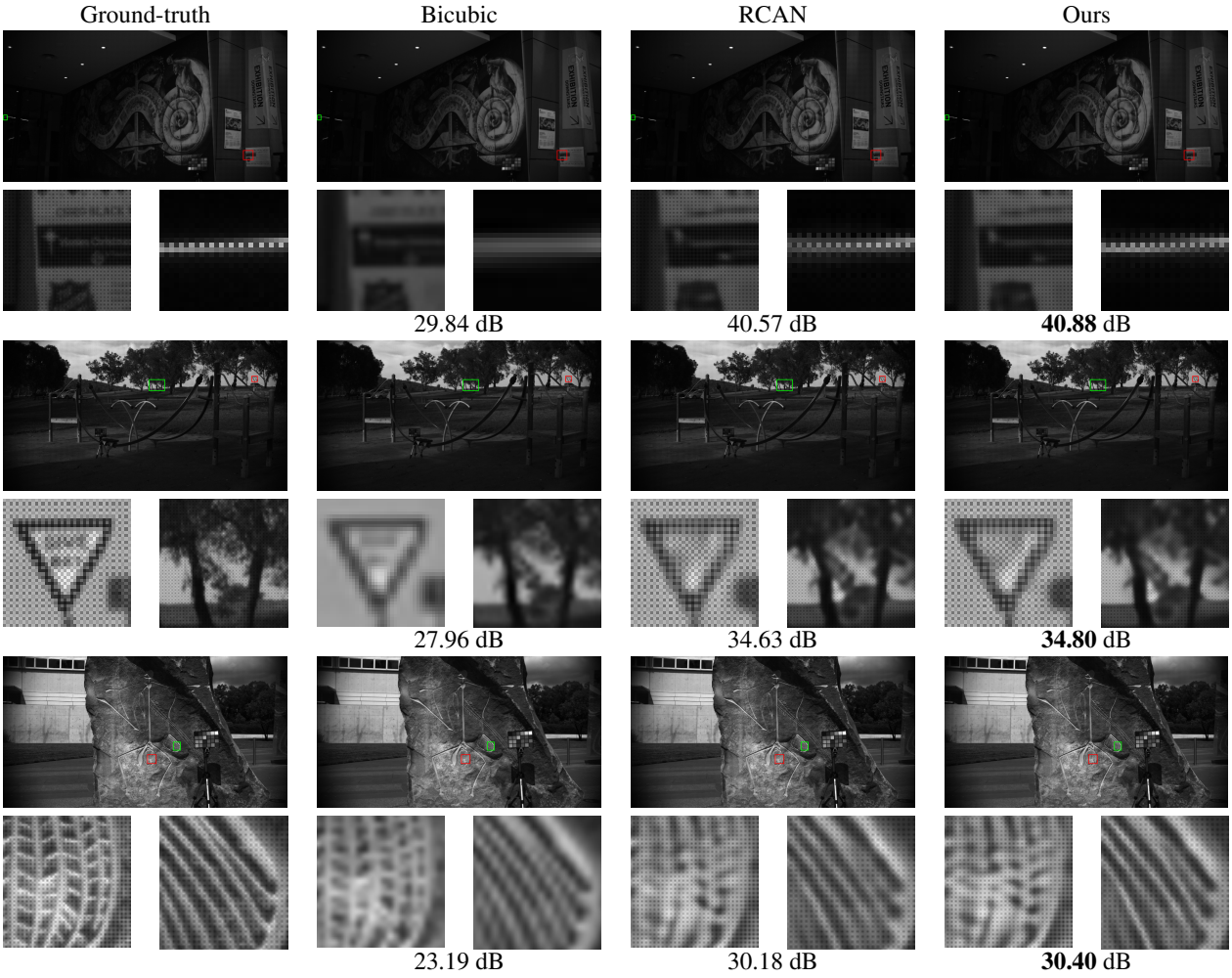


Figure 5. Qualitative results for Bayer SR. Since the images are grayscale by nature, the results are best seen when zoomed in. Note, PSNR results per baseline are provided in the corresponding columns.

6.2.1 Effect of ConvLSTM in PyrRCAN

Here, we aim to assess whether the ConvLSTM is learning the sequence of the concatenated feature pyramid from features from different field of view, or it is merely the effect of reusing the intermediate features. We choose the PyrRCAN$^-$ structure, without our lstmA module, to better isolate the effect of the ConvLSTM module. We remove the ConvLSTM module and instead feed the features into the convolutional layer that follows the ConvLSTM in Fig. 2. The results, presented in Table 3, show the ConvLSTM module is indeed learning additional information from the concatenated features. In fact, without the ConvLSTM module, the results are worse than the baselines.

6.2.2 Effect of attention

As discussed in Section 3, we investigate the effect of different attention mechanisms when placed between the RGs, guided by the intuition that such a mechanism can facilitate more effective information flow between the RGs. The attention mechanisms of our choice are *(i)* the CA that was used in the SOTA multi-spectral SR work [26], *(ii)* the CA mechanism proposed in [12], *(iii)* a bi-linear attention mechanism proposed in [7], *(iv)* our proposed lstmA without the $Sigmoid$ function in the bottom branch (to emulate a mask attention mechanism), and finally *(v)* our proposed lstmA. To assess the effect of all these attention mechanisms more directly, we remove CA from RCAN, so the networks under study only use one type of attention mechanism and only between RGs. The results show that our proposed lstmA mechanism outperforms all the other methods, and it is even marginally superior to BLA proposed in [7]. The results of the ablation study for Bayer mosaic SR follow a more or less similar trend as the results on multi-spectral mosaic SR.

Table 1. Mean and standard deviation (in parenthesis) of PSNR, SSIM on the Multi-spectral dataset. Except where specified, the networks are input with zero-padded cubic mosaics. RCAN$^-$ indicates RCAN without CA.

Method	PSNR	SSIM
Bicubic	28.63 (3.50)	0.4552 (0.0691)
RCAN (Mosaic)	33.17 (3.62)	0.6500 (0.061)
RCAN$^-$ (Mosaic)	33.15 (3.64)	0.6472 (0.0614)
RCAN$^-$	33.16 (3.65)	0.6492 (0.0616)
RCAN	33.21 (3.64)	0.6500 (0.0610)
PyrRCAN	33.293 (3.70)	0.6535 (0.0618)
PyrRCAN + lstmA	**33.31** (0.0625)	**0.6522** (0.0625)

Table 2. Mean and standard deviation (in parenthesis) of PSNR, SSIM for the Bayer dataset. RCAN$^-$ indicates RCAN without CA.

Method	PSNR	SSIM
Bicubic	28.63 (3.50)	0.6398 (0.1364)
RCAN	30.63 (3.65)	0.6560 (0.0998)
RCAN$^-$	30.66 (3.65)	0.6589 (0.0988)
PyrRCAN$^-$ + lstmA	**30.70** (3.65)	**0.6609** (0.1000)

Table 3. Effect of Pyramid ConvLSTM

Method	PSNR	SSIM
Multi-spectral mosaic SR		
PyrRCAN w/o ConvLSTM	33.21	0.6514
PyrRCAN	**33.29**	**0.6535**
Bayer mosaic SR		
PyrRCAN w/o ConvLSTM	30.591	0.655
PyrRCAN	**30.653**	**0.6582**

6.3. Demosaiced RGB images

Our experimental observations indicated that our method is not as effective on demosaiced RGB images, *i.e.*, standard RGB images, as it proved to be for mosaic images. The reason for this was discussed in Section 3. Demosaiced/interpolated images can, in fact, be considered low pass filtered, lacking some crucial information that can be exploited by an SR algorithm. We believe our Pyramid ConvLSTM structure is capable of exploiting such high-frequency information that may be discarded in the process of interpolation. For instance, sub-pixels (wavelengths)

Table 4. Ablation on different attention methods.

Method	PSNR	SSIM
Multi-spectral mosaic SR		
PyrRCAN$^-$ + CA(RCAN)	33.22 (3.6857)	0.6512 (0.06185)
PyrRCAN$^-$ + CA [12]	33.22 (3.67)	0.6511 (0.0622)
PyrRCAN$^-$ + Bi-linear attention [7]	33.24 (3.67)	**0.6517** (0.06277)
PyrRCAN$^-$ + lstm w/o Sigmoid	33.22 (3.712)	0.6475 (0.6475)
PyrRCAN$^-$ + lstmA	**33.26** (3.70)	0.6513 (0.0622)
Bayer mosaic SR		
PyrRCAN$^-$ + CA(RCAN)	30.65 (3.6582)	0.6580 (0.1001)
PyrRCAN$^-$ + CA [12]	30.63 (3.69573)	0.6576 (0.09949)
PyrRCAN$^-$ + Bi-linear attention [7]	30.63 (3.652)	0.6546 (0.0998)
PyrRCAN$^-$ + lstmA w/o Sigmoid	30.67 (3.664)	0.6612 (0.0998)
PyrRCAN$^-$ + lstmA	**30.700** (3.658)	**0.6609** (0.1000)

in mosaic Bayer and multi-spectral pixels display a high-frequency change in intensity, crucial information which is somewhat absent from an interpolated image. Also, high-frequency patterns in either 2×2 or 4 pixels, which seems to appear throughout a mosaic image, contain some intra-wavelength dependencies typical to a multi-spectral or hyper-spectral pixel [9]. A sequential feature pyramid, as we have proposed, is capable of capturing these dependencies throughout a mosaic image.

7. Conclusion

In this paper, we presented an SR algorithm designed explicitly for mosaic super-resolution. Our algorithm exhibit SOTA performance, achieved primarily via constructing a sequential feature pyramid and exploiting a ConvLSTM module to learn the inter-dependencies in the sequence. We also explored different attention modules, replacing CA in RCAN, and observed that an LSTM inspired attention mechanism provides the most significant improvement.

Apart from achieving SOTA and providing structural novelties introduced in this paper, we believe the most important message to convey, based on our experiments, is in regards to the task of Bayer SR. An intuitive observation, verified by our experiments, shows that indeed, mosaic and demosaiced images are different, and algorithms specific to each format need to be developed separately. Also, if a real-life application requires an SR algorithm, it makes more sense to capture Bayer images, which contain more high-frequency information, given that most modern RGB cameras are capable of outputting Bayer patterns. Hence, we believe that it is more beneficial to the computer vision community, with more scientific applications in mind

(*e.g.*, microscopy, astronomy, food monitoring), that more research is dedicated to the task of mosaic super-resolution rather than a one-dimensional focus on standard RGB images. We hope that this work encourages such a research direction.

References

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and, lightweight super-resolution with cascading residual network. *arXiv preprint arXiv:1803.08664*, 2018. 2

[2] James F Bell, Danika Wellington, Craig Hardgrove, Austin Godber, Melissa S Rice, Jeffrey R Johnson, and Abigail Fraeman. Multispectral imaging of mars from the mars science laboratory mastcam instruments: Spectral properties and mineralogic implications along the gale crater traverse. In *AAS/Division for Planetary Sciences Meeting Abstracts*, volume 48, 2016. 1

[3] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 2

[4] D Doering, MR Vizzotto, C Bredemeier, CM da Costa, RVB Henriques, E Pignaton, and CE Pereira. Mde-based development of a multispectral camera for precision agriculture. *IFAC-PapersOnLine*, 49(30):24–29, 2016. 1

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 2016. 2

[6] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *ECCV*, 2016. 2

[7] Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Bilinear attention networks for person retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8030–8039, 2019. 4, 7, 8

[8] Ying Fu, Yinqiang Zheng, Hua Huang, Imari Sato, and Yoichi Sato. Hyperspectral image super-resolution with a mosaic rgb image. *IEEE Transactions on Image Processing*, 27(11):5539–5552, 2018. 1

[9] Alexander FH Goetz. Three decades of hyperspectral remote sensing of the earth: A personal view. *Remote Sensing of Environment*, 113:S5–S16, 2009. 8

[10] Shuosen Guan, Haoxin Li, and Wei-Shi Zheng. Unsupervised learning for optical flow estimation using pyramid convolution lstm. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 181–186. IEEE, 2019. 3

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4, 7, 8

[13] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 2

[14] Sunil Prasad Jaiswal, Lu Fang, Vinit Jakhetiya, Jiahao Pang, Klaus Mueller, and Oscar Chi Au. Adaptive multispectral demosaicking based on frequency-domain analysis of spectral correlation. *IEEE Transactions on Image Processing*, 26(2):953–968, 2016. 2

[15] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 2

[16] Jun-Hyuk Kim, Jun-Ho Choi, Manri Cheon, and Jong-Seok Lee. Ram: Residual attention module for single image super-resolution. *arXiv preprint arXiv:1811.12043*, 2018. 2

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[18] Fayez Lahoud, Ruofan Zhou, and Sabine Süsstrunk. Multimodal spectral image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[19] Fayez Lahoud, Ruofan Zhou, and Sabine Süsstrunk. Multimodal spectral image super-resolution. In *European Conference on Computer Vision*, pages 35–50. Springer, 2018. 3, 4

[20] Yunsong Li, Jing Hu, Xi Zhao, Weiying Xie, and JiaoJiao Li. Hyperspectral image super-resolution using deep convolutional neural network. *Neurocomputing*, 266:29–41, 2017. 3

[21] Zhan Li, Qingyu Peng, Bir Bhanu, Qingfeng Zhang, and Haifeng He. Super resolution for astronomical observations. *Astrophysics and Space Science*, 363(5):92, 2018. 2

[22] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 2

[23] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014. 2

[24] Mohammad Najafi, Sarah Taghavi Namin, and Lars Petersson. Classification of natural scene multi spectral images using a new enhanced crf. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3704–3711. IEEE, 2013. 1

[25] Elias Nehme, Lucien E Weiss, Tomer Michaeli, and Yoav Shechtman. Deep-storm: super-resolution single-molecule microscopy by deep learning. *Optica*, 5(4):458–464, 2018. 2

[26] Zhan Shi, Chang Chen, Zhiwei Xiong, Dong Liu, Zheng-Jun Zha, and Feng Wu. Deep residual attention network for spectral image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 1, 3, 4, 6, 7

[27] Mehrdad Shoeiby, Petersson Lars, Sadegh Aliakbarian, Ali Armin, and Antonio Robles-kelly. Super-resolved chromatic mapping of snapshot mosaic image sensors via atexture sensitive residual network. 5, 6

[28] Mehrdad Shoeiby, Antonio Robles-Kelly, Radu Timofte, Ruofan Zhou, Fayez Lahoud, Sabine Susstrunk, Zhiwei

Xiong, Zhan Shi, Chang Chen, Dong Liu, Zheng-Jun Zha, Feng Wu, Kaixuan Wei, Tao Zhang, Lizhi Wang, Ying Fu, Zhiwei Zhong, Koushik Nagasubramanian, Asheesh K. Singh, Arti Singh, Soumik Sarkar, and Ganapathysubramanian Baskar. PIRM2018 challenge on spectral image super-resolution: Methods and results. In *European Conference on Computer Vision Workshops (ECCVW)*, 2018. 2, 3, 5

[29] Mehrdad Shoeiby, Antonio Robles-Kelly, Ran Wei, and Radu Timofte. Pirm2018 challenge on spectral image super-resolution: Dataset and study. *arXiv preprint arXiv:1904.00540*, 2019. 3, 5

[30] Karasawa Takumi, Kohei Watanabe, Qishen Ha, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. Multispectral object detection for autonomous vehicles. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 35–43. ACM, 2017. 1

[31] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *ICCV*, 2017. 2

[32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[33] Di Wu and Da-Wen Sun. Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A reviewpart i: Fundamentals. *Innovative Food Science & Emerging Technologies*, 19:1–14, 2013. 1

[34] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015. 2, 3

[35] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *CVPR*, 2018. 2

[36] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision*, pages 294–310. Springer, 2018. 2, 3, 4

[37] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. *arXiv preprint arXiv:1807.02758*, 2018. 6

[38] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018. 2

[39] Ruofan Zhou, Radhakrishna Achanta, and Sabine Süsstrunk. Deep residual network for joint demosaicing and super-resolution. In *Color and Imaging Conference*, volume 2018, pages 75–80. Society for Imaging Science and Technology, 2018. 1, 2, 4, 6