Improving Super-Resolution Performance using Meta-Attention Layers

Matthew Aquilina®, Christian Galea®, Member IEEE, John Abela®, Member IEEE, Kenneth P. Camilleri®,
Senior Member IEEE, Reuben A. Farrugia®, Senior Member, IEEE

Abstract—Convolutional Neural Networks (CNNs) have achieved impressive results across many super-resolution (SR) and image restoration tasks. While many such networks can upscale low-resolution (LR) images using just the raw pixel-level information, the ill-posed nature of SR makes it difficult to accurately super-resolve an image which has undergone multiple different degradations. Additional information (metadata) describing the degradation process (such as the blur kernel applied, compression level, etc.) can guide networks to super-resolve LR images with higher fidelity to the original source. Previous attempts at informing SR networks with degradation parameters have indeed been able to improve performance in a number of scenarios. However, due to the fully-convolutional nature of many SR networks, most of these metadata fusion methods either require a complete architectural change, or necessitate the addition of significant extra complexity. Thus, these approaches are difficult to introduce into arbitrary SR networks without considerable design alterations. In this paper, we introduce meta-attention, a simple mechanism which allows any SR CNN to exploit the information available in relevant degradation parameters. The mechanism functions by translating the metadata into a channel attention vector, which in turn selectively modulates the network’s feature maps. Incorporating meta-attention into SR networks is straightforward, as it requires no specific type of architecture to function correctly. Extensive testing has shown that meta-attention can consistently improve the pixel-level accuracy of state-of-the-art (SOTA) networks when provided with relevant degradation parameters. Despite average memory/runtime overheads of less than \( \approx 2.6\% / 0.025 \text{ seconds} \) for the datasets and models considered, meta-attention improves the performance for both PSNR and SSIM; for PSNR, the gain on blurred/downsampled \((\times 4)\) images is of 0.2969 dB (on average) and 0.3320 dB for SOTA general and face SR models, respectively. The coding framework used for this paper is available at: https://github.com/um-dsrg/Super-Resolution-Meta-Attention-Networks.

Index Terms—super-resolution, image restoration, convolutional neural networks, channel attention, metadata fusion.

I. INTRODUCTION & RELATED WORK

The task of image super-resolution (SR) is considered to be an ill-posed problem, as any given low-resolution (LR) image could be upscaled into many distinct high-resolution (HR) counterparts. Any distortions (e.g., blurring, noise injection, compression, etc.) that may affect an image, as is often the case in real-life scenarios, are typically challenging to identify correctly and thus reverse, further increasing the difficulty in yielding super-resolved images that are of satisfactory quality. Convolutional Neural Networks (CNNs), despite being very successful at SR tasks due to their extraordinary capability for learning complex representations [1], [2], can only extract a finite amount of information from pixel-level LR data, and thus struggle to tackle multiple degradation operations without further guidance [3].

However, additional information (metadata) such as image attributes or degradation parameters (e.g., the blur kernel, quantity of noise or the level of compression) could aid the SR process by narrowing down the vast HR space. Previous works have proposed CNN networks which accept blur kernels/noise levels as false image channels [3]–[5] or as additional network inputs [6], [7], and thus allow their models to adapt, and respond to, the degradations used for a particular LR image. For super-resolving face images (face SR), various models have been implemented which either incorporate facial metadata (gender, age, facial features, etc.) into their networks as additional inputs [8]–[11], or use these to construct mathematical face representations [12], each with reported improvements in SR quality. Despite these successes, most attribute-fusion networks are difficult to include in generic SR CNNs. Often, attribute-fusion is proposed as either part of a stand-alone framework, or within modules that require significant design changes (e.g., placement of encoder/decoder blocks [8]) or introduce large amounts of extra complexity. Even in blind SR, where the goal is often to predict the unknown degradations used to generate an LR image, most methods do not attempt to introduce any degradation metadata into their networks [13], despite metadata injection having been shown to be successful in several blind SR models [4], [6], [14].

In this paper, we propose a new mechanism for introducing metadata into SR models, hereinafter referred to as meta-attention. Inspired by channel attention [15], meta-attention condenses image attributes into an attention vector, and uses it to specifically modulate network feature maps based on the information present in available metadata. Lightweight fully-connected (FC) layers are used to produce the attention vector, requiring a minimal number of extra parameters to introduce meta-attention into an SR CNN. Placement of meta-attention within a network is straightforward and can thus easily fit into any CNN architecture, without the need for extra branches or complex structural changes. To validate this claim, we have incorporated meta-attention into a variety of state-of-the-art
(SOTA) SR networks, with differing architectures. With just \( \approx 3.2\% \) or less extra parameters, the introduction of meta-attention has achieved an average Peak Signal-to-Noise Ratio (PSNR)/Structural Similarity Index Measure (SSIM) [16] gain of 0.2969 dB/0.0070 over the original SR models, and a PSNR/SSIM gain of 0.3320 dB/0.0024 over a SOTA face SR model, both on blurred and downsamled \((\times 4)\) datasets.

The contributions of this work are thus twofold: we present meta-attention, a technique for the introduction of metadata into any CNN network, requiring minimal changes other than the addition of meta-attention blocks within the feed-forward path of the network. Secondly, we show that meta-attention can help improve the performance of several SOTA networks with differing architectures, across both face and general SR.

The rest of this paper is structured as follows: Section II discusses the implementation of the proposed meta-attention. Section III follows up with the technical details and results of our evaluation of various SOTA networks upgraded with meta-attention. Finally, Section IV presents the main conclusions drawn in this paper, and provides scope for future work.

II. METHODOLOGY

A. Meta-Attention Module

The proposed meta-attention mechanism is based on the concept of channel attention. Channel attention was originally conceived as a means to capture and exploit global spatial relationships in a CNN feature map, and use this to guide a network to focus on those feature maps which contain the most useful information [15]. The mechanism works by modulating the magnitude of each feature map based on an attention score, thus encouraging the downstream convolutional layers to focus on those feature maps having increased magnitudes. Meta-attention attempts to achieve the same modulation effect, but instead based entirely on the provided metadata information (such as the blur kernel, noise level, and compression quantization level). This means that the metadata is helping to steer the network towards features with increased importance for the particular degradation provided. To achieve this effect, given metadata vector \( \mathbf{m} \) (details on how to obtain \( \mathbf{m} \) are given in Section III-A) for the image being super-resolved, we first pass \( \mathbf{m} \) through two FC layers, separated by the Rectified Linear Unit (ReLU) nonlinearity, \( f(x) \). The final layer is set up to contain as many output units as there are channels in the corresponding network. Then, we use the sigmoid function to scale the output of each unit in the range \([0, 1]\) to attain the attention vector \( \mathbf{w} \). Finally, \( \mathbf{w} \) is used to scale the corresponding feature maps. The entire operation can be expressed as in Equation 1, with the corresponding block diagram provided in the dotted box of Fig. 1:

\[
\mathbf{w} = S(W_2 f(W_1 \mathbf{m}))
\]

where \( f(x) = \max(0, x) \) corresponds to the ReLU function, \( S(x) = 1/(1+\exp(-x)) \) is the sigmoid function, and \( W_1, W_2 \) are the weights of the first and second FC layers, respectively.

B. Network & Metadata Compatibility

Meta-attention was designed with the goal of the mechanism being as unobtrusive and lightweight as possible, making it straightforward for inclusion within any SR CNN framework. When adding meta-attention, module placement and quantity could be optimised for each individual network, but we opted to place a meta-attention module within all residual blocks of each network considered for the sake of consistency (as shown in Figure 1). Empirical evaluations show that this architecture is compatible with all residual networks considered, both for general and face SR (as elaborated in Section III).

Previous works have focused almost entirely on blur kernels for their meta-fusion systems. Apart from blur kernels, we also show that our system can accept and utilise compression metadata, both alone and in tandem with blur kernel information (Section III-C). Otherwise, adding further metadata can be achieved simply by concatenating it with the 1D-vector \( \mathbf{m} \).

III. EXPERIMENTS & RESULTS

A. Network Training, Datasets & Metadata Encoding

In order to validate our claim that meta-attention can be inserted into any SR network, and provide a tangible benefit, we inserted meta-attention into both general and face SR SOTA networks. All models (original and modified) were trained and evaluated on the following datasets and degradations, with all
results computed using an NVIDIA GeForce RTX 3090 GPU (except for bicubic interpolation, which is computed on CPU):

**General SR** - DIV2K [20] & Flickr2K [21]: These datasets contain 800 and 2650 high-resolution training images, respectively, with varied subject matter. We used the combined dataset (3450 images) as our training set for general SR, and applied the 100-image validation set of DIV2K for model selection. After the best performing models were selected, test results were computed on the standard Set5 [22], Set14 [23], and Urban100 [26] datasets.

**SRMD** [3] and **SFTMD** [4] were used for each network/meta-network pair to ensure fair recommendations provided by the authors. Identical hyperparameters were selected as a compromise between performance and practicality. We selected RCAN [15], EDSR [17], SAN [19] and HAN [18] as our reference SOTA networks, and constructed meta-versions of each network by inserting meta-attention blocks (ref. Section II). Meta-attention FC layers were initialized with the board, regardless of the architecture of each baseline model. Meta-SPARNet was built by inserting meta-attention within each of SPARNet’s residual blocks (QPI) was selected from the range [20, 40] (based on typical values used by security cameras [28]). This parameter was fixed learning rate, as recommended in [35]. Meta-RCAN has highly evident that the addition of blur kernel information is needed to improve SR results, but specialised networks which are tailored for super-resolving facial images are also available. Thus, we selected a representative general SR model, RCAN [15], and a SOTA face SR model, SPARNet [35], to determine whether face SR can also benefit from metadata insertion. RCAN models were trained using a cosine annealing scheduler as for general SR, while SPARNet models were trained with a fixed learning rate, as recommended in [35]. Meta-RCAN has the same architecture used for general SR but is trained from scratch on CelebA-HQ images. Meta-SPARNet was built by inserting meta-attention within each of SPARNet’s residual blocks, including those within the encoder/decoder layers. Whole LR images were fed to networks during both training and testing. Test results presented correspond to the models with the best validation PSNR after 1300 epochs; this epoch count was selected as a compromise between performance and practicality.

We selected RCAN [15], EDSR [17], SAN [19] and HAN [18] as our reference SOTA networks, and constructed meta-versions of each network by inserting meta-attention blocks (ref. Section II). Meta-attention FC layers were initialized with the board, regardless of the architecture of each baseline model. Meta-SPARNet was built by inserting meta-attention within each of SPARNet’s residual blocks (QPI) was selected from the range [20, 40] (based on typical values used by security cameras [28]). This parameter was fixed learning rate, as recommended in [35]. Meta-RCAN has the same architecture used for general SR but is trained from scratch on CelebA-HQ images. Meta-SPARNet was built by inserting meta-attention within each of SPARNet’s residual blocks, including those within the encoder/decoder layers. Whole LR images were fed to networks during both training and testing. Test results presented correspond to the models with the best validation PSNR after 50 epochs.

**B. Meta-Attention in General SR**

Using blurred images as input and the corresponding blur kernel PCA vectors as the only source of metadata, SR test image quality was objectively measured using PSNR and SSIM [16], results of which are shown in Table I. It is highly evident that the addition of blur kernel information significantly boosts the performance of each network across the board, regardless of the architecture of each baseline method. This result is further reinforced by Figure 2, which

### Table I

| Model       | Trainable Parameters | Average Runtime (s) | PSNR   | SSIM   | PSNR   | SSIM   | PSNR   | SSIM   | PSNR   | SSIM   | PSNR   | SSIM   | PSNR   | SSIM   |
|-------------|----------------------|---------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| Bicubic     | -                    | 0.0054              | 23.5140 | 0.7276 | 23.3000 | 0.6215 | 23.6628 | 0.5864 | 22.0783 | 0.7113 | 20.7192 | 0.5731 |
| SRMD [3]    | 1.546,288            | 0.0289              | 29.8551 | 0.8637 | 26.3506 | 0.7425 | 25.8852 | 0.6975 | 27.9152 | 0.8731 | 23.7010 | 0.7060 |
| SFTMD [4]   | 4.234,691            | 0.0139              | 30.3419 | 0.8727 | 26.6556 | 0.7515 | 26.1175 | 0.7057 | 28.9294 | 0.8894 | 24.2790 | 0.7530 |
| EDSR [17]   | 13.809,923           | 0.0087              | 30.2476 | 0.8741 | 26.5603 | 0.7507 | 26.1604 | 0.7059 | 28.9113 | 0.8887 | 24.3036 | 0.7564 |
| Meta-EDSR   | 44.191,747           | 0.0173              | 30.6688 | 0.8776 | 26.8903 | 0.7579 | 26.2681 | 0.7116 | 29.6419 | 0.8986 | 24.7200 | 0.7695 |
| RCAN [15]   | 15.392,355           | 0.0680              | 30.3981 | 0.8744 | 26.7267 | 0.7530 | 26.1959 | 0.7072 | 29.3058 | 0.8935 | 24.6354 | 0.7646 |
| Meta-RCAN   | 16.085,155           | 0.0886              | 30.6943 | 0.8779 | 26.9303 | 0.7586 | 26.2834 | 0.7121 | 29.7563 | 0.8998 | 24.8770 | 0.7739 |
| HAN [18]    | 16.071,745           | 0.0695              | 30.5187 | 0.8762 | 26.7571 | 0.7534 | 26.1957 | 0.7068 | 29.3308 | 0.8945 | 24.6262 | 0.7641 |
| Meta-HAN    | 16.564,545           | 0.0901              | 30.7304 | 0.8784 | 26.9333 | 0.7589 | 26.2883 | 0.7123 | 29.7654 | 0.9000 | 24.8677 | 0.7742 |
| SAN [19]    | 15.860,488           | 0.3853              | 30.5128 | 0.8757 | 26.6965 | 0.7515 | 26.1644 | 0.7063 | 29.1047 | 0.8942 | 24.4775 | 0.7590 |
| Meta-SAN    | 16.353,288           | 0.4615              | 30.7435 | 0.8786 | 26.9277 | 0.7592 | 26.2916 | 0.7126 | 29.7877 | 0.9003 | 24.9094 | 0.7749 |
shows that meta-aware networks are capable of correctly super-resolving obscure patterns which are typically muddled by the original models (additional examples available in the Supplementary Material). Furthermore, as can be observed in Table I, the overhead in terms of memory (given an average increase of approximately 2.97%) and computational cost (with a runtime increase of approximately 0.0315 seconds per image on average) of the proposed architecture is generally marginal. It is expected that memory requirements and running times could be further reduced with optimized code. Table I also shows how EDSR, which performs worse than SFTMD (a SOTA SR CNN specifically designed for metadata-aware SR) for several of the datasets considered, can be made to outperform SFTMD with the addition of meta-attention.

C. Insertion of Multiple Types of Metadata

Apart from blur kernels, it can also be shown that the proposed framework is able to successfully exploit other types of metadata such as the QPI value representing the degree of image compression used. Table II shows the results of adding different combinations of metadata to RCAN with meta-attention. While adding QPI information alone to RCAN only shows weak improvements, adding both blur kernels and QPI values produces the best results, indicating that the proposed network is capable of not just handling multiple types of meta information, but also extracting and exploiting useful complementary information provided by each input.

D. Meta-Attention in Face SR

Table III shows similar trends for face SR to those observed previously, with meta-attention providing a substantial benefit to PSNR and SSIM values for both models despite an average parameter/runtime increase of just 2.16%/0.0123 seconds.

IV. CONCLUSIONS

This paper has presented meta-attention, a novel framework for introducing metadata into any SR network, regardless of architecture and application, with minimal complexity increase. Results have shown that meta-attention provides substantial performance improvements in terms of both PSNR and SSIM on a variety of architectures, when fed with blur kernel or compression metadata (or both). Moreover, performance gains were also observed not solely for general SR, but also for face SR which has important real-world applications such as in security and law enforcement [36]. We envision that this framework can allow SR researchers to more easily introduce metadata into their networks, especially in applications where this could be readily available (such as extracting the QPI from video bitstreams) or in blind SR, where methods attempt to estimate the attributes of unknown degradations. We also hypothesise that this framework could accept additional image attributes other than degradation metadata (e.g. CelebA facial features), which will be investigated as part of future work.

ACKNOWLEDGEMENTS

This research forms part of the Deep-FIR project, which is financed by the Malta Council for Science & Technology (MCST), for and on behalf of the Foundation for Science & Technology, through the FUSION: R&I Technology Development Programme.
REFERENCES

[1] Z. Wang, J. Chen, and S. C. H. Hoi, “Deep Learning for Image Super-resolution: A Survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–11, 2020.

[2] S. Anwar, S. Khan, and N. Barnes, “A Deep Journey into Super-resolution: A Survey,” ACM Comput. Surv., vol. 53, no. 3, May 2020.

[3] K. Zhang, W. Zuo, and L. Zhang, “Learning a Single Convolutional Super-Resolution Network for Multiple Degradations,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2018.

[4] J. Gu, H. Lu, W. Zuo, and C. Dong, “Blind super-resolution with iterative kernel correction,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2019.

[5] Y.-S. Xu, S.-Y. R. Tseng, Y. Tseng, H.-K. Kuo et al., “Unified dynamic convolutional network for super-resolution with variational degradations,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2020.

[6] V. Cornillère, A. Djelouah, W. Yifan, O. Sorkine-Hornung et al., “Blind Image Super-Resolution with Spatially Variant Degradations,” ACM Trans. Graph., vol. 38, no. 6, Nov. 2019.

[7] K. Zhang, L. V. Gool, and R. Timofte, “Deep Unfolding Network for Image Super-resolution,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2020.

[8] X. Yu, B. Fernando, R. Hartley, and F. Porikli, “Super-Resolving Very Low-Resolution Face Images With Supplementary Attributes,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2018.

[9] M. Li, Z. Zhang, J. Yu, and C. W. Chen, “Learning Face Image Super-resolution Through Facial Semantic Attribute Transformation and Self-Attentive Structure Enhancement,” IEEE Transactions on Multimedia, vol. 23, pp. 468–483, 2021.

[10] C.-H. Lee, K. Zhang, H.-C. Lee, C.-W. Cheng et al., “Attribute Augmented Convolutional Neural Network for Face Hallucination,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Jun. 2018.

[11] J. Xin, N. Wang, X. Gao, and J. Li, “Residual Attribute Attention Network for Face Image Super-resolution,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 9054–9061, Jul. 2019.

[12] J. Xin, N. Wang, X. Jiang, J. Li et al., “Facial Attribute Capsules for Noise Face Super-resolution,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 12476–12483, Apr. 2020.

[13] A. Lugmayr, M. Danelljan, and R. Timofte, “NTIRE 2020 Challenge on Real-World Image Super-resolution: Methods and Results,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Jun. 2020.

[14] L. Wang, Y. Wang, X. Dong, Q. Xu et al., “Unsupervised degradation representation learning for blind super-resolution,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2021, pp. 10 581–10 590.

[15] Y. Zhang, K. Li, K. Li, L. Wang et al., “Image Super-resolution Using Very Deep Residual Channel Attention Networks,” in Proceedings of the European Conference on Computer Vision (ECCV), Sep. 2018.

[16] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.

[17] B. Lim, S. Son, H. Kim, S. Nah et al., “Enhanced Deep Residual Networks for Single Image Super-resolution,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Jul. 2017.

[18] B. Niu, W. Wen, W. Ren, X. Zhang et al., “Single Image Super-resolution via a Holistic Attention Network,” in Computer Vision – ECCV 2020, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 191–207.

[19] T. Dai, J. Cai, Y. Zhang, S.-T. Xia et al., “Second-Order Attention Network for Single Image Super-resolution,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2019.

[20] E. Agustsson and R. Timofte, “NTIRE 2017 Challenge on Single Image Super-resolution: Dataset and Study,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Jul. 2017, pp. 1122–1131.

[21] R. Timofte, E. Agustsson, L. V. Gool, M.-H. Yang et al., “NTIRE 2017 Challenge on Single Image Super-resolution: Methods and Results,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Jul. 2017, pp. 1110–1121.