Arbitrary Scale Super-Resolution Neural Network Based on Residual Channel-Spatial Attention

JAVIER GURROLA-RAMOS1, TERESA E. ALARCÓN2, AND OSCAR DALMAU1

1Mathematics Research Center, Department of Computer Science, Guanajuato 36023, Mexico
2Department of Computer Science and Engineering, Centro Universitario de los Valles, Ameca 46600, Mexico

Corresponding author: Javier Gurrola-Ramos (francisco.gurrola@cimat.mx)

This work was supported by the Universidad de Guadalajara under Grant PROSNI-2022-UdeG-RO_03.

ABSTRACT In recent years, the performance of convolutional neural networks in single-image super-resolution has improved significantly. However, most state-of-the-art models address the super-resolution problem for specific scale factors. In this paper, we propose a convolutional neural network for arbitrary scale super-resolution. Specifically, given a range of scale factors, the proposed model can generate super-resolution images with non-integer scale factors within the range. The proposed model incorporates a channel-spatial attention block in which the scale factor is also provided. This module recovers the most relevant information from the low-resolution image given the scale factor and enhances the upsampled image before generating the high-resolution target image. This channel attention block allows learning the channel and spatial dependencies. Additionally, we incorporate global residual learning so that the model recovers the details of an upsampled low-resolution image by interpolation. We evaluated the proposed method through extensive experiments on widely used benchmark datasets for single-image super-resolution. In order to assess the performance of the model, we used the peak signal-to-noise ratio and the structural similarity index measure. The proposed model achieves an average of 35.36, 31.78, 29.62 for peak signal-to-noise ratio, and 0.9410, 0.8828, 0.8334 for structural similarity index measure for the standard evaluation scale factors ×2, ×3, ×4, respectively. The experimental results show a better performance of the proposed model over other state-of-the-art models for arbitrary scale super-resolution, and are competitive with models trained for specific scale factors.

INDEX TERMS Single image super-resolution, arbitrary scale, convolutional neural network, channel-spatial attention, image enhancement.

I. INTRODUCTION

Single-image super-resolution (SISR) is an important subject in image processing and computer vision. The aim of SISR is to reconstruct an accurate high-resolution image ($I_{HR}$) given its corresponding low-resolution image ($I_{LR}$). SISR has been widely used in several computer vision applications, including security and surveillance imaging [1], medical imaging [2], object recognition [3], satellite images [4], etc. However, getting an image with super resolution (SR) is an ill-posed inverse problem since various solutions exist for a given $I_{LR}$ input. The SISR problem has been addressed by traditional techniques, including interpolation-based methods [5], reconstruction-based methods [6], and learning-based models [7], [8], [9]. In recent years, convolutional neural networks (CNN) have notably outperformed several computer vision tasks, including image super-resolution [10]. There two main streams in super-resolution algorithms based in CNN: single scale SR and multiple scale SR.

A. SINGLE SCALE SR

Dong et al. [11] proposed the first CNN model for SISR, achieving significant improvement over conventional methods. The model was named super-resolution convolutional neural network (SRCNN). This method first
upsamples the low-resolution image to the target size, and then the model improves the quality of the upsampled image to estimate the $I_{\text{HR}}$ image. Wang et al. in [12] propose a densely connected network based on edge-guided care (EGADNet) for SISR. This model is composed of two branches that obtain the content and the high-frequency information, and a final branch that combines the previous information and returns the super-resolution image.

Haris et al. [13] proposed a back-projection CNN model that exploits iterative upsampling/downsampling layers, providing an error feedback mechanism for projection errors at each stage. In this way, the proposal better examines the mutual dependency of low and high-resolution images. In order to improve the network accuracy, the authors in [13] introduce a dense connection to capture the relationship between $I_{\text{LR}}$ and $I_{\text{HR}}$. This also contributes to implementing a large-scale factor. Nonetheless, the training step of this model needs a very large dataset, which substantially increases the computation time. On the other hand, the proposed network only uses a single-scale convolution kernel, which makes it difficult to extract feature information at different scales [14].

Zhang et al. [15] propose a residual channel attention to SISR by considering interdependence among channels. The model leads to the detection of more important feature maps and also effectively captures the relationships between them. One drawback of the proposal in [15] is the high computational complexity [10]. Chen et al. [16] propose a pre-trained model for low-level computer vision tasks based on image transformers [16], [17]. Their proposal uses different input/output modules trained for specific tasks but keeps the same encoding/decoding transformers. This approach has shown a good performance not only in SISR but also in other image processing tasks. However, it requires a very large training dataset, and a large number of trainable parameters. Chen et al. [18] propose a gradient-based CNN model that takes advantage of the structural information present in $I_{\text{LR}}$. This model uses a dual-branch structure, in which one branch estimates the high-resolution image, and the other uses gradient feature maps to recover the structure of the image. Since the gradient is sensitive to noise, this could affect the performance of this model. Zhang et al. [19] propose the residual dense network (RDN) model, which improves the feature extraction of $I_{\text{LR}}$ images through dense convolutions and local residual learning. Although the good behavior of the model, this kind of schema is not efficient for detecting hierarchical features [20]. The main limitation of the previous works is that we need to train the model for each scale, which is impractical in many situations.

### B. MULTIPLE SCALE SR

Hu et al. [21] propose a meta-upscale module to handle non-integer scale factors. The proposal dynamically estimates the filters used in the upsampling for different scale factors using meta-learning. This module was initially applied to the RDN and the enhanced deep super-resolution network (EDSR) [22] models as the upsampling module. The meta-upscale module receives the scale factor as additional information to estimate the filter values. However, this additional information is not used by the rest of the model, i.e., the backbone parameters of the model are the same for all scale factors for which the model was trained. Fu et al. [23] extend the meta module by combining its original linear input vector with a quadratic and a bicubic encoded vector. This approach leads to a higher complexity of the encoding vector as the degree increases with no necessary a better performance of the proposal.

Shen et al. [24] propose an approach based on applying the Laplacian pyramid recursively to generate arbitrary scale $I_{\text{HR}}$ images. The scale factor used in this method is in the range $(1, 2)$, so in order to generate the desired target size the method should be applied several times. This module also presents a plug-in element for predefined CNN models to extend their capability to handle non-integer scale factors. Despite the good behavior of the method, including a low number of parameters, the pyramidal approach increases the computational complexity and the use of memory.

Kim et al. [25] improve the SRCNN model by adding more convolutional layers and global residual learning so that the model learns the differences between the upsampled $I_{\text{LR}}$ image and the target $I_{\text{HR}}$. The training step of the model is carried out with multiple scales, which improves the performance. However, the architecture considers a bicubic interpolation for the input image, and all subsequent layers work in high resolution. Therefore, the computation time and memory increase [22]. Chen et al. [26] propose the local implicit image function (LIIF) that basically combines the RDN model [19] with a multilayer perceptron which increases the computational cost.

There are also several methods for more realistic applications in which the target scale size is not an integer value. Wang et al. [27] propose a plug-in module for arbitrary scale SISR, which consists of multiple scale-aware feature adaptation blocks and a scale-aware upsampling layer. This module is capable of generating images with different scale factors along the height and width of the image. Feng et al. [28] propose a dynamic kernel generation model that improves the image reconstruction performance. The model consists of two stages. The first one extracts the features from low-resolution images by means of self-organizing feature mapping network, allowing the construction of regularized pattern. The regularized pattern contains information at different locations. The second stage uses a regularized pattern and a meta-learning mechanism to predict the weights of the convolution kernels that correspond the regularized pattern at each location. The deep learning community also addresses the problem of parameter estimation for different scale factors. In this direction, Liu et al. [29] propose a framework based on low-resolution reconstruction, blind kernel estimation, and recursive error compensation. Their approach uses one set of parameters while achieving varying scales and obtaining higher efficiency and satisfactory accuracy.
Most existing methods are trained or designed for a pre-defined degradation process (e.g., bicubic down-sampling) from a $I_{HR}$ to a $I_{LR}$. As a result, these models can struggle with low-resolution images obtained by a degradation process different from the one considered in the training dataset [30]. In recent years, the interest in this area of research, known as blind image super-resolution, has increased. Among the methods designed for blind-image SR are those based on a predefined set of different degradation models [31], based on the estimation of the degradation model by reducing the artifacts iteratively on the reconstructed SR images [30], and based on deep learning [32], [33]. Compared with the commonly used degradation model in synthetic low-resolution images, the actual degradation in real-world scenarios is more complex and varying because it can be affected by several factors such as imaging system and imaging environment.

Another recent research on SISR is real-world single image super-resolution [34]. In this case, the pair of $I_{LR}$ and $I_{HR}$ are generated collecting images from the same scene using different camera configurations for different image resolutions. Cai et al. [35] proposed a real SR dataset, which consists of pairs of images obtained by adjusting the lens of two digital single lens reflex cameras and a progressive image registration framework to achieve pixel-wise registration for the different focal lengths. Wei et al. also proposed a real SR dataset [36] that extends the dataset proposed in [35], but using five digital single lens reflex cameras, and consists of both outdoor and indoor scenes. Zhang et al. proposed another real dataset [37] which is composed of pairs of raw images taken at different levels of optical zoom. Euclidean motion model is applied for image registration to describe the relationship between the images with different resolutions and it is optimized by minimizing the enhanced correlation coefficient. Additionally, the dataset contains both raw sensor data and RGB images for SR from raw data.

In summary, and according to our examination of the reported proposals, many of the models work with low-resolution images previously upsampled through bicubic interpolation to the target size [11], [25]. This strategy allows arbitrary handling of scale factors, but it may generate noise amplification and blurring to the input image. Besides, processing a larger input image increases the computational cost significantly. On the other hand, most CNN models work around this problem by feeding $I_{LR}$ into the model and upsampling it within the network using transposed convolution [38] or sup-pixel upsampling [39] near the end of the model. These upsampling methods limit the models to integer scale factors, reducing their flexibility in applications.

With the aim of addressing the previous limitations, and considering the significant improvement of the CNNs over conventional SR techniques, we propose a residual channel-spatial attention network, namely RC-SAN, for arbitrary scale SISR. The proposal is able to accomplish the super-resolution image without the limitation of integer scale factors and enough attenuation of the noise and blurring as a consequence of scaling. Furthermore, the proposal does not consider the image degradation model that generates the $I_{LR}$, so in principle, it can be used for any degradation model.

We develop an attention block that combines spatial and channel information along the scale factor to preserve or shrink the information generated by residual dense blocks. The model uses bicubic interpolation to perform the upsampling operation before the last section of the model, so most of the process is on the low-resolution image. Using bicubic interpolation allows deal with arbitrary scale factors and perform global residual learning such that the model enhances the previously upsampled $I_{LR}$ and speed-up the learning process.

The main contributions of this paper are summarized as follows:

- We propose a neural network for single image super-resolution with arbitrary scale in a given range (1, 4).
- We extend the proposed model for specific scale factors ($\times2, \times3, \times4$).
- We present an attention block that combines the feature maps and the provided scale factor. This block is designed to learn the channel and spatial dependencies of feature maps.

Additionally, we incorporate, in the last part of the model, a global residual learning using a bicubic interpolation of the low-resolution image. This strategy allows to handle arbitrary scales and speeds up the training of the model. Extensive experiments demonstrate that the proposed method is competitive with state-of-the-art approaches for SISR, even with those trained for specific scale factors.

This paper is organized as follows: Section II explains the proposal and its methodology, Section III describes all the experiments we carried out and the results, Section IV discusses the results obtained, Section V presents some limitations of the proposal and finally in Section VI the conclusions are given.

II. PROPOSED METHOD

In this section, we first introduce the architecture of the proposed RCSAN model and then describe the building blocks of the model.

A. NETWORK ARCHITECTURE

The proposed RCSAN model, see Fig. 1, consists of three modules and a global residual learning (GRL): the first module is a shallow features extractor (block in green), the second module is the core of the model and represents a deep feature extractor based on a series of residual dense convolutional networks with an attention block (block in gray) and the last module is a high-resolution re-constructor block (block in yellow), which is finally combined with a GRL.

Module 1: In order to extract the shallow features from the original low resolution image, $I_{LR}$, we use $3 \times 3$ convolutions,
based on [19], and the PReLU activation function [40]:

\[ F_{LR} = H_{LRFE2}(H_{LRFE1}(I_{LR})) \]  

(1)

where \( H_{LRFE1}(\cdot) \) and \( H_{LRFE2}(\cdot) \) denote the first and second convolution operation respectively, see Fig. 1 block in green.

**Module 2:** After extracting the shallow features \( F_{LR} \), we apply the deep features extractor block, Fig. 1 block in gray, consisting of a sequence of \( I \) residual dense groups (RDG), Fig. 2, section II-B. This sequence can be defined as follows

\[ F_{i+1} = \begin{cases} F_{LR} & \text{if } i = 0 \\ H_{RDG,i}(F_{i}) & \text{if } i = 1, \ldots, I - 1 \end{cases} \]  

(2)

where \( H_{RDG,i} \) is the \( i-th \) RDG block, details of the RDG block are described in Sec. II-B. That is, \( F_{i+1} \) is obtained after applying \( H_{RDG,i} \) to the previous feature map \( F_{i} \). The last element of the sequence \( F_{I} \) is the output of this module.

**Module 3:** This module is formally written as follows:

\[ F_{BI} = B^{\dagger}(F_{I}) \]  

(3)

\[ F_{HR} = H_{HRFE2}(H_{HRFE1}(RDG_{HR}(F_{BI}))) \]  

(4)

where \( B^{\dagger}(\cdot) \) corresponds to a bicubic interpolation, \( RDG_{HR}(\cdot) \) is an RDG module that performs deep feature extraction of the upsampled \( F_{BI} \) image (see Fig. 2), and \( H_{HRFE1}(\cdot) \) and \( H_{HRFE2}(\cdot) \) are \( 3 \times 3 \) convolutions with PReLU activation function. This module is shown in Fig. 1 block in yellow.

We first apply a bicubic interpolation to \( F_{I} \), the output of the deep feature extractor, Eq. (3), in order to upsample the image to the desired scale factor. This upsampling method allows the model to deal with arbitrary scale factors, contrary to other methods, like transposed convolution or sub-pixel upsampling, where the scale factor is fixed. Further, the final convolution \( H_{HRFE2}(\cdot) \) reduces the number of feature maps to match the number of the input image channels, what allows us to introduce a global residual learning.

**Global Residual Learning:** Finally, and based on the previous comment, we perform the global residual learning between the output of the high-resolution re-constructor block, \( F_{HR} \), and the bicubic interpolation of the LR image \( I_{LR} B^{\dagger}(I_{LR}) \):

\[ I_{HR} = F_{HR} + B^{\dagger}(I_{LR}). \]  

(5)

This type of learning allows the model to learn important details and features of the image that the interpolation process cannot recover.

**B. RESIDUAL DENSE GROUP**

The residual dense group (RDG block) is shown in Fig. 2. This module consists of the following elements: one sequence of dense convolutions with an attention block, Fig. 4, one concatenation, a factorization block through asymmetric convolutions and a residual learning. The last factorization block summarizes the feature maps generated by the dense blocks, and the residual learning is applied between the input of the RDG and the output of the factorization. The \((i+1)\)-th deep feature image is generated from the \(i\)-th deep feature image as follows: we first obtain the sequence of feature maps...
\{F_{ij}\} = F_{ij} + \text{DAB}_{ij} \otimes F_{ij,K+1} \quad (13)

where $C_{i,j,K+1}$ corresponds to the $1 \times 1$ convolution with PReLU activation function, $\text{FC}_{i,j,1}$, $\text{FC}_{i,j,2}$ are the fully connected layers, with PReLU and Sigmoid activation functions respectively, and $\otimes$ is the multiplication between the elements generated by $\text{DAB}_{ij}$ and the feature maps of $F_{ij,K+1}$, i.e., every feature map in $F_{ij,K+1}$ is weighted by the corresponding entry in the vector $\text{DAB}_{ij}$.

D. IMPLEMENTATION DETAILS

In the proposed model RCSAN, we use the appropriate zero pad in the $3 \times 3, 3 \times 1$, and $1 \times 3$ convolutions to maintain the size of the image. The convolution of sizes $3 \times 3$ and $1 \times 1$ generate 64 feature maps and the convolutions of sizes $3 \times 1$ and $1 \times 3$ generate 256 feature maps. The fully-connected layers in the attention block generate 64 features. Additionally, the number of residual dense groups in the deep feature extraction section is $J = 4$. The number of dense blocks and dense convolutions for all the residual dense groups are $J = 4$ and $K = 6$, respectively.

E. ABLATION STUDY

Table 1 shows the effects of using global residual learning (GRL), channel-spatial attention block, and the incorporation of the scale factor in the attention block. In this study we use the Set14 dataset, section III-B. We compare the performance of the model incorporating these components after the first 200 training epochs. To consider the scale factor information without using the attention block, the scale factor is concatenated as a feature map before the last $1 \times 1$ convolution of the dense convolutional block, with a negligible increase in the number of trainable parameters. This addition improves about 0.03 and 0.22 of the peak signal-to-noise ratio (PSNR) value compared with the base model. However, the improvement when this information is combined with GRL is little significant. The attention block by itself improves over the use of the scale factor as additional information for the scale factors $\times 2$ and $\times 3$. However, this is the component that adds more parameters to the model (1.39 M). On the other hand, the combination of the attention block and the scale factor information improves the performance for all the scale factors. The GRL modification increases the performance for all the evaluated scale factors and for all the other model components. This may be mainly because the model only has to retrieve the details of the previously scaled low-resolution image instead of estimating the high-resolution image directly.

To visualize the convergence process, Fig. 5 shows the training process for the studied combinations in Table 1. Note that the GRL increases the speed of the training process of the model without incorporating additional parameters.

III. EXPERIMENTS

A. TRAINING DETAILS

The flowchart in Fig. 6 shows the process of training, validation and testing of the RCSAN model. We use the
DIV2K dataset [41] and Flickr2K [42] as training dataset. The DIV2K and Flickr2K training datasets consist of 800 and 2650 high-quality, high-resolution color images, respectively. We use a validation dataset of the DIV2K, consisting of 100 images with similar characteristics to the training dataset.

For training the proposed model, we randomly generate input/output training patches \((x_i, y_i)\) from the DIV2K training partition and the Flickr2K dataset. The \(I_{LR}\) patch size is set to \(48 \times 48\). The scale factor is selected randomly uniformly in the range \([1.1, 4.0]\) with step 0.1 such that the whole training batch has the same scale factor. Additionally, we apply data augmentation consisting of random vertical and horizontal flips and random 90° rotations. In addition, we trained a version of the proposed model for specific scale factors, denoted as RCSAN\(_s\), to assess the model in this scenario and compare it with state-of-the-art models trained for a single scale factor. All the versions of the model were trained from scratch.

In order to optimize the parameters of the RCSAN model, we use the Charbonnier loss function [43] and the AdamW algorithm [44] with parameters \(\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}\), and \(\lambda = 10^{-5}\). The initial learning rate is \(\alpha_0 = 10^{-4}\), which is halved every 200 epochs. This optimization algorithm allows us to work with decoupled loss function and weight decay term. The RCSAN model was trained with a batch size of 16 for 1000 epochs and 1000 iterations of back propagation per epoch. Our model was implemented in Python 3.6 using the PyTorch framework. The training time was 14 days in two Nvidia RTX Titan GPU. The source code and models are available at https://github.com/JavierGurrola/RCSAN.

### B. TESTING DETAILS

The proposed models were tested on five widely used benchmark datasets for SISR: Set5 [45], Set14 [46], BSD100 [47], Urban100 [48], and Manga109 [49]. To assess the results of the model, we first convert the estimated \(I_{HR}\) image into YCrCb space color and then calculate their corresponding PSNR [50] and the structural similarity index measure (SSIM) [51] metrics on the Y channel.

We compare with state-of-the-art models designed for specific scale and arbitrary scale factors. The compared models for specific scale factors are DBPN [13], GFSR [18], IGNN [52], SAN [53], HAN [54], RCAN [15], NLSA [55], FSDN [24], IPT [16], and EGADNet [12]. The compared models trained for arbitrary scale factors are ASDN [24], Meta-RDN [21], ArbRCAN [27], RSAN [23], LIIF [26], fSISR [29], and RPB [28]. The scale factors considered for comparisons with state-of-the-art models are \(\times 2\), \(\times 3\), and \(\times 4\). Based on [56], a self-ensemble strategy is introduced to improve the proposed RCSAN and RCSAN\(_s\), denoted as RCSAN+, and RCSAN\(_s\)+, respectively.
IV. RESULTS
In Table 2, we compare the performance of the proposed RCSAN model with the baseline methods bicubic interpolation and Meta-RDN, for the arbitrary scale SR task. In this experiment we consider the BSD100 dataset. We use scale factors from 1.1 to 4.0 with a step length of 0.1.
The results show that the proposed RCSAN outperforms the Meta-RDN model by 0.135 dB on average and the bicubic interpolation by 2.832 dB on average. Fig. 7 illustrates an example of arbitrary scale SR for the image “86000”, taken from the BSD100 dataset.

Tables 3-5 shows the average PSNR and SSIM metrics of the compared methods on the test datasets for ×2, ×3, and ×4 scale factors. The proposed arbitrary scale RCSAN model is competitive with state-of-the-art designed for arbitrary scale SR and outperforms in most cases. Notice that the proposed model excels significantly for the case of the Urban100 dataset with a difference of 0.14 dB and 0.0023 in PSNR and SSIM values respectively between the RCSAN and the model with the closest value (RSAN) for the ×4 scale factor. The data sets in which the proposed model struggles the most are Set5 and Set14. Although the proposed model also stands out for BSD100 and Manga109 datasets, the gap between the performance of the proposal and the model with the closest performance is lower. Note that, the self-ensemble version outperforms all the compared models. In particular, the performance has a greater increase for the case of Manga109 with 0.25 dB and 0.003 in PSNR and SSIM values respectively.

Fig. 8 shows reconstructed regions from 5 images using seven methods: Bicubic, HAN, ITP, ArbRCAN, meta-SR, RSAN and our proposal. The PSNR and SSIM values are also reported. In all cases, the scale factor applied to the images is ×4. It is worth mentioning that for all the compared methods, we used the source code and pre-trained models provided by the authors. In this case, we used the RCSAN.
FIGURE 8. Comparative results of different methods for SISR with $\times 4$ scale factor. On the left are original images, and on the right are regions of interest that have been predicted by different methods. On the bottom of each region appear the name of the methods and the obtained metrics PSNR (dB) and SSIM. The best results are highlighted in red, and the second-best in blue.
TABLE 5. Quantitative results of different SISR methods on Set5, Set14, BSD100, Urban100 and Manga109 test datasets for ×4 scale factor. The best two results of PSNR (dB) and SSIM are highlighted in red and blue respectively for the models trained for a specific scale factor, and bold and italic for arbitrary scale models respectively.

| Model      | Scale | Set5 [45] | Set14 [46] | BSD100 [47] | Urban100 [48] | Manga109 [49] |
|------------|-------|-----------|------------|-------------|---------------|---------------|
|            |       | PSNR      | SSIM       | PSNR        | SSIM          | PSNR          |
| Bicubic    |       | 28.42     | 0.8104     | 26.00       | 0.7027       | 25.96         |
| DBPN [13]  |       | 32.47     | 0.8980     | 28.82       | 0.7860       | 27.72         |
| GFSR [18]  |       | 32.45     | 0.8980     | 28.73       | 0.7847       | 27.65         |
| IGNN [52]  |       | 32.57     | 0.8998     | 28.85       | 0.7891       | 27.77         |
| SAN [53]   |       | 32.64     | 0.9003     | 28.92       | 0.7888       | 27.78         |
| HAN [54]   |       | 32.64     | 0.9002     | 28.90       | 0.7890       | 27.80         |
| RCAN [15]  | ×4 (specific) | 32.63     | 0.9002     | 28.87       | 0.7889       | 27.77         |
| NLSA [55]  |       | 32.59     | 0.9000     | 28.87       | 0.7891       | 27.78         |
| FSDN [24]  |       | 32.63     | 0.9000     | 28.89       | 0.7890       | 27.79         |
| IPT [16]   |       | 32.64     | -          | 29.01       | -             | 27.82         |
| EOGANet* [12] |       | 32.36     | 0.8964     | 28.66       | 0.7827       | 27.61         |
| RCSAN, (ours) |       | 32.61     | 0.9006     | 28.85       | 0.7891       | 27.82         |
| RCSAN+, (ours) |       | 32.71     | 0.9014     | 28.94       | 0.7905       | 27.86         |
| ASDr [24]  |       | 32.77     | 0.8960     | 28.56       | 0.7840       | 27.63         |
| Meta-RDN [21] |       | 32.55     | -          | 28.87       | -             | 27.76         |
| ArbRCAN [27] |       | 32.62     | 0.9001     | 28.90       | 0.7891       | 27.77         |
| RSAN [23]  |       | 32.50     | 0.8960     | 28.84       | 0.7885       | 27.80         |
| LISF [26]  |       | 32.60     | 0.9000     | 28.76       | 0.7870       | 27.76         |
| RCSISR [29] |       | 32.60     | 0.9000     | 28.86       | 0.7881       | 27.76         |
| RFB [28]   |       | 32.50     | 0.8990     | 28.86       | 0.7881       | 27.76         |
| RCSAN (ours) |       | 32.63     | 0.9006     | 28.84       | 0.7885       | 27.80         |
| RCSAN+ (ours) |       | 32.69     | 0.9012     | 28.90       | 0.7895       | 27.84         |

A. NUMBER OF PARAMETERS AND MODEL COMPLEXITY

Fig. 9 compares the PSRN obtained for the methods in the Set14 dataset in terms of the number of parameters. In this case, we only compare the methods that have reported their number of parameters. It can be observed that the RCSAN outperforms most of the methods with a better trade-off between the performance and the number of parameters. The GFSR is considered a low-weight model; therefore, its performance is expected to be lower. Note that the ITP model requires a large number of parameters to achieve good performance. On the other hand, among the methods with arbitrary scaling: RCSAN, Meta-RDN, and RSAN, the proposed RCSAN has fewer parameters and better performance.

Table 6 shows the execution time (in seconds) for different SISR deep learning models for a 64 × 64 with a ×4 scale factor, the number of multiplication-accumulation operations model trained for arbitrary scale. The first row of Fig. 8 depicts the comparison results using the image “img067” from the Urban100 dataset. Some methods, such as ITP and ArbRCAN, present a loss in the structure and obtain blurred regions. On the contrary, the proposed method can recover more high-frequency detail and less blurry regions compared to the other methods. The second row shows the comparison using the “butterfly” image from the Set5 dataset. The proposed RCSAN method recovers the image with more contour and structure details. In this case, the RSAN method has the best PSNR value, but the proposed RCSAN model has the best SSIM value; nevertheless, the visual results between both methods are very similar. The third row shows the comparison of the “Baboon” image from the Set14 dataset. All methods present difficulties in reconstructing this image, having problems recovering details of the hairs around the nose. We can see that the reconstructed region has a high degree of blurring for all methods. For this case, the RCSAN gets the second-best PSNR and SSIM values. The fourth row shows the comparison of the “Comic” image from the Set14 dataset. In this case, neither method recovers the gray details and pattern present in the upper right of the zoomed region. However, the RCSAN method retrieves the image with less blur and more detail on the edge. The last row shows the comparison of the “MariaSamaNihaNaisyo” image from the Manga109 dataset. For this image, the proposed model is able to recover more texture and details on the borders.
FIGURE 10. Preliminary results obtained by the RCSAN model, using a realistic dataset for super-resolution. From left to right: low resolution, high resolution and reconstructed high resolution images. The numbers below the low resolution images correspond to the obtained PSNR and SSIM measures. The images are taken from ImagePairs testing dataset available at https://www.microsoft.com/applied-sciences/projects/imagepairs.

TABLE 6. Comparison of the execution time, multiplication-accumulation operations (MACs), and number of trainable parameters of several architectures for estimate a 64 × 64 with a ×4 scale factor.

| Model       | GPU time (s) | CPU time (s) | MACs (G) | # Params (M) |
|-------------|--------------|--------------|----------|-------------|
| HAN         | 0.2017       | 0.7028       | 116.09   | 16.07       |
| IPT         | 0.2094       | 2.6325       | 257.23   | 115.63      |
| ArbRCAN     | 0.0700       | 0.4005       | 65.16    | 16.93       |
| Meta-RDN    | 0.0508       | 0.7438       | 119.1    | 22.42       |
| RSAN        | 0.0532       | 0.8106       | 118.09   | 22.61       |
| NLSA        | 0.2109       | 2.5048       | 433.8    | 44.16       |
| RCSAN       | 0.0829       | 2.3523       | 301.07   | 19.69       |

(MACs) required to generate the 256 × 256 images, and the number of parameters of the models. The execution time of the proposed RCSAN is comparable with the ArbRCAN and slightly longer than the Meta-RDN and RSAN for the case of GPU. In the case of the CPU execution time, this is more degraded for the proposed model than for the others. However, the RCSAN is still faster than the IPT and NLSA models. In contrast to the fast GPU execution time, the number of operations for the RCSAN is high in proportion to the low number of trainable parameters. This is mainly because the last section of the model processes the image in high resolution. Thus, the number of required operations increases at this stage.

V. LIMITATIONS OF THE STUDY

Considering that the purpose of the article was to propose an arbitrary scale super-resolution model, the design of the experiment considered only one type of degradation. Therefore, the main limitation of the study was the evaluation of the proposed model in the context of blind or real super-resolution. It is worth mentioning that the proposed model is blind to degradation. It would require training the model using databases with different degradation models and probably adjusting the model for these tasks. Another limitation is that the number of multiplication-accumulation operations is high, while GPU time is low. This suggests improving the last module (module 3) of the model since this part works on the high-resolution image. These limitations will become the future work of this study.

VI. CONCLUSION

In this work, we presented a residual channel-spatial attention network for arbitrary scale single-image super-resolution. We also extended the previous model for specific scale factors. The proposed models incorporate an attention block that takes advantage of the spatial information, channel
information, and the scale factor applied to the image to improve the results. Additionally, we incorporated global residual learning using a bicubic interpolation of the low-resolution image. The use of global residual learning allowed the model to enhance previously upsampled low-resolution images instead of directly estimating the high-resolution image. At the same, it allowed to handle arbitrary scales and speed up the model’s training. Experiments on several datasets showed that the proposed model is competitive compared with state-of-the-art models for arbitrary scale super-resolution, even with those trained for specific scale factors. The extended versions of the proposed model, trained for specific scale factors, also yielded competitive results.

The main limitation of this study is its application to blind or real super-resolution. This limitation can be addressed by training the model on databases with different degradation models and possibly adapting the model for these tasks. For this purpose, new components could also be proposed. The model can also be extended to solve denoising and super-resolution problems simultaneously. All this remains as future work.

VII. ACKNOWLEDGMENT

This work was supported by the program to support the improvement of production conditions for members of the National System of Researchers (SNI) and the National System of Art Creators (SNCA) 2022, the Universidad de Guadalajara (project No. 265147) under Grant PROSNI-2022-UdeGRO_03, and the project “Laboratorio de Supercomputo del Bajo (No. 300832)”.

REFERENCES

[1] P. Rasti, T. Uiboupin, S. Escalera, and G. Anbarjafari, “Convolutional neural network super resolution for face recognition in surveillance monitoring,” in Articulated Motion Deformable Objects, F. J. Perales and J. Kütter, Eds. Cham, Switzerland: Springer, 2016, pp. 175–184.

[2] Y. Huang, L. Shao, and A. F. Frangi, “Simultaneous super-resolution and cross-modality synthesis of 3D medical images using weakly-supervised joint convolutional sparse coding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., Jul. 2017, pp. 5787–5796.

[3] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch, “EnhanceNet: Single image super-resolution through automated texture synthesis,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 4491–4500.

[4] T. Lu, J. Wang, Y. Zhang, Z. Wang, and J. Jiang, “Satellite image super-resolution via multi-scale residual deep neural network,” Remote Sens., vol. 11, no. 13, p. 1588, Jul. 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/13/1588

[5] K. Turkowski, “Filters for common resampling tasks,” Graph. Gems, pp. 147–165, Aug. 1990. [Online]. Available: https://citeseer.nj.nipr.jp/417595.html#40061009712

[6] K. Zhang, X. Gao, D. Tao, and X. Li, “Single image super-resolution with non-local means and steering kernel regression,” IEEE Trans. Image Process., vol. 21, no. 11, pp. 4544–4556, Nov. 2012.

[7] R. Timofte, V. De Smet, and L. Van Gool, “Anchored neighborhood regression for fast example-based super-resolution,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 1920–1927.

[8] T. Peleg and M. Elad, “A statistical prediction model based on sparse representations for single image super-resolution,” IEEE Trans. Image Process., vol. 23, no. 6, pp. 2569–2582, Jun. 2014.

[9] R. Timofte, V. De Smet, and L. Van Gool, “A+-: Adjusted anchored neighborhood regression for fast super-resolution,” in Computer Vision—ACCV (Lecture Notes in Computer Science), D. Cremers, I. Reid, H. Saito, and M.-H. Yang, Eds. Cham, Switzerland: Springer, 2015, pp. 111–126.

[10] S. Anwar, S. Khan, and N. Barnes, “A deep journey into super-resolution: A survey,” ACM Comput. Surveys, vol. 53, no. 3, pp. 1–34, May 2021, doi: 10.1145/3390462.

[11] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in Computer Vision—ECCV (Lecture Notes in Computer Science), D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 184–199.

[12] Z. Wang, Y. Lu, and Q. Shi, “Edge guided attention based densely connected network for single image super-resolution,” in Neural Information Processing (Lecture Notes in Computer Science), T. Mantoro, M. Lee, M. A. Ayu, K. W. Wong, and A. N. Hidayanto, Eds. Cham, Switzerland: Springer, 2021, pp. 632–643.

[13] M. Haris, G. Shakhnarovich, and N. Ukitia, “Deep back-projection networks for super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 1664–1673.

[14] X. Chen and C. Sun, “Multiscale recursive feedback network for image super-resolution,” IEEE Access, vol. 10, pp. 6393–6406, 2022.

[15] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 286–301.

[16] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, “Pre-trained image processing transformer,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 12299–12310.

[17] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, “Image transformer,” in Proc. 35th Int. Conf. Mach. Learn., vol. 80, J. Dy and A. Krause, Eds. Jul. 2018, pp. 4055–4064. [Online]. Available: https://proceedings.mlr.press/v80/parmar18a.html

[18] J. Chen, D. Huang, X. Zhu, and F. Chen, “Gradient-guided and multi-scale feature network for image super-resolution,” Appl. Sci., vol. 12, no. 6, p. 2935, Mar. 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/6/2935

[19] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2472–2481.

[20] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Liao, “MADNet: A fast and lightweight network for single-image super-resolution,” IEEE Trans. Cybern., vol. 51, no. 3, pp. 1443–1453, Mar. 2021.

[21] X. Hu, H. Ma, X. Zhang, Z. Wang, T. Tan, and J. Sun, “Meta-SR: A magnification-arbitrary network for super-resolution,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1575–1584.

[22] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jul. 2017, pp. 136–144.

[23] X. Fu, J. Chen, T. Zhang, and Y. Lin, “Residual scale attention network for arbitrary scale image super-resolution,” Neurocomputing, vol. 427, pp. 201–211, Feb. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220317872

[24] J. Shen, Y. Wang, and J. Zhang, “ASDN: A deep convolutional network for arbitrary scale image super-resolution,” Mobile Netw. Appl., vol. 26, no. 1, pp. 13–26, Feb. 2021, doi: 10.1007/s11036-020-01720-2.

[25] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Oct. 2016, pp. 1646–1654.

[26] Y. Chen, S. Liu, and X. Wang, “Learning continuous image representation with local implicit image function,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 8628–8638.

[27] B. Lim, Y. Wang, Z. Lin, J. Yang, W. An, and Y. Guo, “Learning a single network for scale-arity super-resolution,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 4801–4810.

[28] H. Feng, L. Ma, and J. Tian, “A dynamic convolution kernel generation method based on regularized pattern for image super-resolution,” Sensors, vol. 22, no. 11, p. 4231, Jun. 2022. [Online]. Available: https://www.mdpi.com/1424-8247/22/11/4231

[29] Y. Liu, L. Wan, F. Lyu, and W. Feng, “Fine-grained scale space learning for single image super-resolution,” Vis. Comput., vol. 38, no. 9, pp. 1–13, 2022, doi: 10.1007/s00371-022-05551-w.

[30] J. Gu, H. Lu, W. Zuo, and C. Dong, “Blind super-resolution with iterative kernel correction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1604–1613.
[31] K. Zhang, W. Zuo, and L. Zhang, “Learning a single convolutional super-resolution network for multiple degradations,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3262–3271.

[32] Z. Luo, Y. Huang, S. Li, L. Wang, and T. Tan, “Unfolding the alternating optimization for blind super resolution,” in Adv. Neural Inf. Process. Syst., vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Red Hook, NY, USA: Curran Associates, 2020, pp. 5632–5643. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/3d2d8bcb37f9d77bcb69d1a5b76c3fa-Paper.pdf

[33] L. Wang, Y. Wang, X. Dong, Q. Xu, J. Yang, W. An, and Y. Guo, “Unsupervised degradation representation learning for blind super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 10581–10590.

[34] H. Chen, X. He, L. Qing, Y. Wu, C. Ren, R. E. Sheriff, and C. Zhu, “Real-world single image super-resolution: A brief review,” Inf. Fusion, vol. 79, pp. 124–145, Mar. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521001792

[35] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, “Toward real-world single image super-resolution: A new benchmark and a new model,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 3086–3095.

[36] P. Wei, Z. Xie, H. Lu, Z. Zhan, Q. Ye, W. Zuo, and L. Lin, “Component divide-and-conquer for real-world image super-resolution,” in Computer Vision—ECCV (Lecture Notes in Computer Science), A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham, Switzerland: Springer, 2020, pp. 101–117.

[37] X. Zhang, Q. Chen, R. Ng, and V. Kolotun, “Zoom to learn, learn to zoom,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3762–3770.

[38] C. Dong, C. C. Loy, and X. Tang, “Accelerating the super-resolution convolutional neural network,” in Computer Vision—ECCV (Lecture Notes in Computer Science), B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham, Switzerland: Springer, 2016, pp. 391–407.

[39] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, C. Dong, C. C. Loy, and X. Tang, “Accelerating the super-resolution convolutional neural network for multiple degradations,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 3086–3095.

[40] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, “Second-order attention for single image super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 11065–11074.

[41] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, “Single image super-resolution via a holistic attention network,” in Computer Vision—ECCV (Lecture Notes in Computer Science), A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham, Switzerland: Springer, 2020, pp. 191–207.

[42] Y. Mei, Y. Fan, and Y. Zhou, “Image super-resolution with non-local sparse attention,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 3517–3526.

[43] R. Timofte, R. Rothe, and L. Van Gool, “Seven ways to improve example-based single image super resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 1865–1873.

JAVIER GURROLA-ROMAS received the B.Eng. degree in computer engineering from UAA, Aguascalientes, Mexico, in 2016, and the M.Sc. degree in computer science and industrial mathematics from the Mathematics Research Center (CIMAT), Guanajuato, Mexico, in 2019, and the Ph.D. degree in computer science from the Center for Mathematics Research, Guanajuato, Mexico, in 2007. She received the title of an Engineer in automated systems of management at the Moscow Institute of Direction Sergio Orchonikidze, Russia, in 1989. She is currently an Associate Professor with the Computational Sciences and Engineering Department, University of Guadalajara’s Valley Campus, Ameca, Jalisco. Her field of research interests include digital image processing, including filtering, segmentation, and pattern recognition.

TERESA E. ALARCON received the master’s degree in digital image processing from the José Antonio Echeverria Polytechnic Institute, Cuba, in 1999, and the Ph.D. degree in computer science from the Center for Mathematics Research, Guanajuato, Mexico, in 2007. She received the title of an Engineer in automated systems of management at the Moscow Institute of Direction Sergio Orchonikidze, Russia, in 1989. She is currently an Associate Professor with the Computational Sciences and Engineering Department, University of Guadalajara’s Valley Campus, Ameca, Jalisco. Her field of research interests include digital image processing, including filtering, segmentation, and pattern recognition.

OSCAR DALMAU received the B.Sc.Ed. degree in mathematics from ISP Manzanillo Cuba, in 1989, and the M.Sc. degree in computer science and industrial mathematics and the Ph.D. degree in computer science from the Mathematics Research Center (CIMAT), Guanajuato, Mexico, in 2004 and 2010, respectively. He is currently with CIMAT. His research interests include the areas of machine learning, optimization, image processing, and computer vision.

[44] Q. Huynh-Thu and M. Ghanbari, “Scope of validity of PSNR in image/video quality assessment,” Electron. Lett., vol. 44, no. 13, pp. 800–801, Jun. 2008. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/el_20080522

[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[46] S. Zhou, J. Zhang, W. Zu, and C. C. Loy, “Cross-scale internal graph neural network for image super-resolution,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Red Hook, NY, USA: Curran Associates, 2020, pp. 3499–3509. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/ce3a156092f8c8d6d3237ccf30797777-Paper.pdf

[47] Y. Mei, Y. Fan, and Y. Zhou, “Image super-resolution with non-local sparse attention,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 3517–3526.

[48] R. Timofte, R. Rothe, and L. Van Gool, “Seven ways to improve example-based single image super resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 1865–1873.

[49] Q. Huynh-Thu and M. Ghanbari, “Scope of validity of PSNR in image/video quality assessment,” Electron. Lett., vol. 44, no. 13, pp. 800–801, Jun. 2008. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/el_20080522

[50] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[51] S. Zhou, J. Zhang, W. Zu, and C. C. Loy, “Cross-scale internal graph neural network for image super-resolution,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Red Hook, NY, USA: Curran Associates, 2020, pp. 3499–3509. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/ce3a156092f8c8d6d3237ccf30797777-Paper.pdf

[52] Y. Mei, Y. Fan, and Y. Zhou, “Image super-resolution with non-local sparse attention,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 3517–3526.

[53] R. Timofte, R. Rothe, and L. Van Gool, “Seven ways to improve example-based single image super resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 1865–1873.

OSCAR DALMAU received the B.Sc.Ed. degree in mathematics from ISP Manzanillo Cuba, in 1989, and the M.Sc. degree in computer science and industrial mathematics and the Ph.D. degree in computer science from the Mathematics Research Center (CIMAT), Guanajuato, Mexico, in 2004 and 2010, respectively. He is currently with CIMAT. His research interests include the areas of machine learning, optimization, image processing, and computer vision.

* * *

VOLUME 10, 2022

108709