Boosting High-Level Vision with Joint Compression Artifacts Reduction and Super-Resolution

Xiaoyu Xiang  
Purdue University  
West Lafayette, IN 47907, USA  
xiang43@purdue.edu

Qian Lin  
HP Labs  
Palo Alto, CA 94304, USA  
qian.lin@hp.com

Jan P. Allebach  
Purdue University  
West Lafayette, IN 47907, USA  
allebach@purdue.edu

Abstract—Due to the limits of bandwidth and storage space, digital images are usually down-scaled and compressed when transmitted over networks, resulting in loss of details and jarring artifacts that can lower the performance of high-level visual tasks. In this paper, we aim to generate an artifact-free high-resolution image from a low-resolution one compressed with an arbitrary quality factor by exploring joint compression artifacts reduction (CAR) and super-resolution (SR) tasks. First, we propose a context-aware joint CAR and SR neural network (CAJNN) that integrates both local and non-local features to solve CAR and SR in one-stage. Finally, a deep reconstruction network is adopted to predict high quality and high-resolution images. Evaluation on CAR and SR benchmark datasets shows that our CAJNN model outperforms previous methods and also takes 26.2% shorter runtime. Based on this model, we explore addressing two critical challenges in high-level computer vision: optical character recognition of low-resolution texts, and extremely tiny face detection. We demonstrate that CAJNN can serve as an effective image preprocessing method and improve the accuracy for real-scene text recognition (from 85.30% to 85.75%) and the average precision for tiny face detection (from 0.317 to 0.611).

I. INTRODUCTION

Image down-scaling and compression techniques are widely used to meet the limits of hardware storage and data capacity, which sometimes sacrifice the visual effects as well as bringing troubles to visual detection and recognition. Compression artifact reduction (CAR) [1] and single image super-resolution (SISR) [2] have been used in manifold applications, e.g. digital zoom on smartphones [3], video streaming [4] and print quality enhancement [5], [6] to restore a high-quality and high-resolution image. Since Dong [7] first proposed SRCNN that applied a three-layer convolutional neural network (CNN) for the SISR task, more and more works [8], [9], [10] have explored how to make use of the deep neural networks (DNN) to achieve better image quality as measured by PSNR and SSIM [11], or better visual quality [12], [13] as measured by other perceptual metrics [14], [15].

Conventional methods adopt a two-stage pipeline to leverage the quality and resolution of real-world images: first preprocess the user’s photos with a compression artifacts reduction (CAR) algorithm [16], [17], [18], [19], [20], and then conduct a super-resolution (SR) algorithm [2], [7], [8], [9], [10], [21], [22], [23], [24]. However, the CAR step often causes loss of high-frequency information, which results in a lack of detail in reconstructed SR images. Besides, the computation and data transmission between the two models is time-consuming. To deal with these issues, a single-stage method that jointly solves the Compression Artifact Reduction and Super-Resolution (CARS) problems is needed to reach a balance between reducing artifacts while retaining most details for the upscale step with a short run-time.

Both CAR and SR aim to learn the high-frequency information for reconstruction. Thus, instead of simply concatenating two networks together, we design two functional modules in a single-stage network that reduces the model size by simplifying the two reconstruction processes into one, and can directly obtain high-quality SR output without reconstructing the intermediate artifact-free LR images. Towards this end, we propose a context-aware joint CAR and SR neural network (CAJNN) that can make use of the locally related features in low-quality, low-resolution images to reconstruct high-quality, high-resolution images. To train this network, we construct a paired LR-HR training dataset based on modeling the degradation kernels of web images. Our model turns out to be able to reconstruct high-resolution and artifact-free images with high stability for user’s images from a garden variety of web-apps (e.g. Facebook, Instagram, WeChat). Figure 1 illustrates the performance of our proposed algorithm and the benefits of the single-stage joint CARS method compared with previous two-stage methods: our result can reconstruct a more visually appealing output with accurate structures, sharp edges, and significantly fewer compression artifacts. These output images are not only more recognizable for human viewers, but also for off-the-shelf computer vision algorithms.

Fig. 1: Demonstration of the joint CAR and SR task. For a user’s image without ground truth, our joint CAR and SR model (CAJNN) can generate better output with sharper edges and significantly fewer artifacts compared with the two-stage CAR+SR method.

Input  
CAR+SR  
Ours (CAJNN)
In this paper, we demonstrate that our proposed CAJNN can enhance the detection and recognition accuracy of high-level vision tasks by reducing the compression artifacts and increasing the resolution of input images.

To summarize, our contributions are mainly three-fold: (1) We propose a novel CAJNN framework that jointly solves the CAR and SR problems for real-world images, that are from unknown devices with unknown quality factors. Here, we explore ways to represent and combine both local and non-local information to enforce image reconstruction performance without knowing the input quality factor. (2) Our experiments show that CAJNN achieves the new state-of-the-art performance on multiple datasets e.g. Set5 [25], Set14 [26], BSD100 [27], Urban100 [28], etc. as measured by the PSNR and SSIM [11] metrics. Compared with the prior art, it generates more stable and reliable outputs for any level of compression quality factors. (3) We provide a new idea for enhancing high-level computer vision tasks like real-scene text recognition and extremely tiny face detection: by preprocessing the input data with our pretrained model, we can improve the performance of existing detectors. Our model demonstrates its effectiveness on the WIDER face dataset [29] and the ICDAR2013 Focused Scene Text dataset [30].

II. RELATED WORK

CNN-based Single Image Super-Resolution Convolutional Neural Network (CNN) methods have demonstrated a remarkable capability to recover LR images with known kernels after the pioneering work of Dong et al. [7] that adopted a 3-layer CNN to learn an end-to-end mapping from LR images to HR images. The follow-up work FSRCNN [23] established the general structure of most SR networks until today, which conducts most computations in the low-resolution domain and upsamples the image to the required scale at the end of the network. After 2016, more and more works began to explore how to make the network go deeper. EDSR [8] reduces the number of parameters by removing the batch normalization layer, and shares the parameters between the low-scale and high-scale models to achieve better training results. RDN [9, 10] and RRDB [13] employ densely-connected residual groups as the major reconstruction block to reach large depth and to allow sufficient low-frequency information to be bypassed. In the meantime, some useful structures have been introduced to enhance the processing speed or output quality. Shi et al. [24] designed a sub-pixel upscaling mechanism. RCAN [21] introduces a channel attention mechanism to rescale channel-wise features adaptively, and SAN [31] exploits a more powerful feature expression with second-order channel attention.

Compression Artifacts Reduction Lossy compression methods [32, 33] are widely applied in web image transmission due to their higher compression rates. Traditional methods for the CAR problem generally fall into two categories: unsupervised methods, which include removing noise and increasing sharpness [18], and supervised methods like dictionary-based algorithms [34]. After the success of SRCNN on the super-resolution task, Yu et al. [35] directly applied its architecture to compression artifacts suppression. Similar to the development of SR, CNN-based CAR networks also go deeper with the introduction of residual blocks and skip connections [36, 37, 38]. Besides, SSIM loss is employed [17] as a supervision method to obtain better performance than MSE loss. JPEG-related priors are also considered in the network structure design, e.g. DDCN [39] adds a Discrete Cosine Transform (DCT)-domain before the dual networks, and the D³ [40] takes a further step in the practice of dual-domain approaches [34] by converting sparse-encoding approaches into a one-step sparse inference module.

Unlike the above approaches that require reconstructing the intermediate clean LR images, our joint CARSR framework directly obtains artifact-free HR images without prior information of quality factors or explicit CAR supervision in the LR domain.

III. JOINT COMPRESSION ARTIFACTS REDUCTION AND SUPER-RESOLUTION

Given an LR JPEG image $I^{LRLQ}$, our goal is to reconstruct the high resolution, high-quality image $G(I^{LRLQ})$ that approaches the high-resolution, high-quality ground truth $I^{HRHQ}$ with a generator $G$. The CARSR task can be expressed as:

$$\arg\min_{\theta} l(I^{HRHQ}, G(I^{LRLQ}, \theta_{g})),$$

where $l$ is any designated loss function (e.g. MSE, L1, Charbonnier, etc.). $G$ is the function representing our deep neural network with parameters $\theta$, for which we wish $G_{\theta} \approx F^{-1}((I^{LRLQ} \odot k) \uparrow, s)$, where $\odot$ stands for the convolution operation, $k$ is the degradation kernel of downsampling method, and $s$ is the downsampling factor. To effectively handle the CARSR task, we propose a single-stage framework, CAJNN. Our proposed model is end-to-end trainable with $I^{HRHQ}$ and $I^{LRLQ}$ pairs according to the function above.

The CAJNN framework mainly consists of three modules (see Figure 1): the context-aware feature extractor, the reconstruction module, and the upsampling and enhancement module. The context-aware feature extractor captures and assembles both intra- and inter-block information from different receptive fields. The reconstruction module further refines the extracted feature maps. Finally, after the processing of the upsampling and enhancement module, these feature maps are converted to high-resolution outputs.

A. Model

Here we discuss the CAJNN structure in detail. The majority of our network operates in the feature domain. Given $I^{LRLQ}$ ($c \cdot h \cdot w$ in size), a feature extraction layer first turns the image into feature maps ($n_f \cdot h \cdot w$ in size, where $n_f$ denotes the number of feature channels) in the domain for the following process. The feature map will be converted to a high-resolution image ($c \cdot H \cdot W$ in size) after passing the upsampling and enhancement module. To achieve a balance between GPU capacity and output quality, we apply $n_f = 64$
channels to ensure enough information for the reconstruction. We adopt a $3 \times 3$ convolution layer that serves as the initial feature extractor. After this module, the input image $f_{LRLQ}^{L}$ is turned into a $64 \cdot h \cdot w$ tensor $f^{L}$.

1) Context-aware Feature Extractor: The pipeline of JPEG compression involves the following steps: color space transformation (e.g., JPEG, H.264/AVC, H.265/HEVC), downsampling, block splitting, discrete cosine transform (DCT), quantization, and entropy encoding. Some previous research assumed that the quality factors of input images are known, and the original images are well-aligned by the JPEG block boundaries. However, real-world inputs do not always meet such assumptions. In the worst case, the input images might be compressed multiple times and contain sub-blocks or larger blocks, which requires the model to be insensitive, or even blind to the encoding block alignment. Thus, the spatial context information of both intra- and inter-JPEG blocks is essential for designing a CARSR network.

We adopt the ASPP module to extract and integrate multi-scale features with an atrous spatial pyramid pooling (ASPP) [21]. We adjust the dilation rates of each layer in the pyramid to extend the filter’s perceptive field for extracting different ranges of context information, in which the largest field-of-view should cover the $8 \times 8$ block. Besides, we should avoid sampling overlap in different levels of the $3 \times 3$ convolutions. Considering the factors above, we choose 1, 3, 4 as the dilation groups to find a good balance between accurately retrieving local details and assimilating context information between adjacent blocks. The input tensors are sent to 3 layers of the pyramid: a $3 \times 3$ convolutional layer with dilation rate $= 1$, a $3 \times 3$ convolutional layer with dilation rate $= 3$, and a $3 \times 3$ convolutional layer with dilation rate $= 4$. The outputs of these three layers are concatenated and aggregated by a $1 \times 1$ convolution. The process in ASPP can be described by:

$$f^{L} = [C_{3x3,1} \otimes f^{L} | C_{3x3,3} \otimes f^{L} | C_{3x3,4} \otimes f^{L}] \oplus C_{1x1,1},$$  \hspace{1cm} (2)

where $f^{L}$ denotes the output feature ($64 \cdot h \cdot w$ in size), $C_{a \times a, r}$ represents the parameters of $a \times a$ convolution with dilation rate $r$, and $\oplus$ is a concatenation operation.

2) Reconstruction: RRDB (residual-in-residual dense block) [13] is applied as the basic block for the reconstruction trunk. Compared with residual blocks, it densely connects the convolution layers to local groups while removing the batch normalization layer. In our network, the reconstruction module includes 20 RRDBs.

3) Upsampling and Enhancement: After the reconstruction module, the image feature is preprocessed by a $3 \times 3$ convolution layer before the PixelShuffle layer [24] for upsampling. The PixelShuffle layer produces an HR image from LR feature maps directly with one upscaling filter for each feature map. Compared with the upconvolution, the PixelShuffle layer is $\log_2 s^2$ times faster in theory because of applying sub-pixel activation to convert most of the computations from the HR to the LR domain. The feature $f_{LRLQ}^{SR''}$ is turned into a $c \cdot sh \cdot sw$ HR image by the PixelShuffle layer, which can be described by:

$$f_{LRLQ}^{SR'} = PS(W_{L} \otimes f_{LRLQ}^{SR''} + b_{L}),$$  \hspace{1cm} (3)

where $W_{L}$ denotes the convolution weights and $b_{L}$ the bias in the LR domain, $PS$ is a periodic shuffling operator for rearranging the input LR feature tensor $f_{LRLQ}^{SR''} (c \cdot s^2 \cdot h \cdot w)$ to a HR tensor of shape $c \cdot sh \cdot sw$:

$$PS(T)_{x,y,c} = T_{[x/s] \cdot [y/s],c \cdot s \text{ mod}(y,s) + c \cdot s \text{ mod}(x,s)}.$$  \hspace{1cm} (4)

Instead of directly outputting the high-resolution image, we process it through two $3 \times 3$ convolution layers for further enhancement, and get $I_{SR}^{F} = C_{3x3,1} \oplus (C_{3x3,1} \otimes I_{SR}^{SR'})$.

To make the major network focus on learning the high-frequency information in the input image, we bilinearly upsample the input LR image $f_{LRLQ}^{L}$ and add it to the final output $G(I_{LRLQ}^{L}, \theta_{g})$:

$$G(I_{LRLQ}^{L}, \theta_{g}) = I_{LRLQ}^{SR} \uparrow_{s} + I_{SR}^{F}.$$  \hspace{1cm} (5)

This long-range skip connection changes the target of our major network from directly reconstructing a high-resolution image to reconstructing its residual. By letting the low-frequency information of the input bypass the major network, it lowers the difficulty of reconstruction and increases the convergence speed of the network.

IV. EXPERIMENTS AND ANALYSIS

A. Experimental Setup

Training Dataset: In this paper, we choose the DIV2K dataset [22] (800 RGB images of 2k resolution) and Flick2K dataset [23] (2650 RGB images of 2k resolution) as our training set. To get the training pairs that approach the degradation kernels of web images, we first model the downsampling and compression types of popular web platforms. We also discover that adding severely compressed samples to the training set can improve the output quality in terms of PSNR, even for input images compressed with high-quality factors. Based on these pre-experimental results, the images of the training set are downsampled with a scaling factor $s = 4$ and compressed by MATLAB [44] with random quality factors from 10 to 100.
Besides, we perform data augmentation on these images by randomly cropping, randomly rotating by $90^\circ$, $180^\circ$, and $270^\circ$, and randomly horizontal-flipping. As a result, each cropped image patch can have eight augmentation types at maximum.

### Test Datasets
We compare the performance of our model and previous methods on Set5 [25], Set14 [26], BSD100 [27], Urban100 [28] and Manga109 [45]. Each image is downsampled and compressed with quality factors of 10, 20 and 40 to be consistent with previous works.

### Implementation Details
Our network is trained on one Nvidia Titan Xp graphics card. The batch size is 36, and the patch size is 128 for ground truth and 32 for low-resolution input. We use Adam [46] as the optimizer with a cosine annealing learning rate, in which the initial learning rate is $2e^{-4}$, and the minimum learning rate is $1e^{-7}$. The scheduler restarts every $5e5$ iterations. The network is trained for $1e6$ iterations in total.

#### B. Results for Image Quality Assessment

##### Comparison with SOTA on Standard Test Sets
We compare the performance of CAJNN to the previous state-of-the-art (SOTA) methods on the standard test sets as mentioned above. We report the PSNR and SSIM [111] of the Y channels in the test sets to be consistent with previous works. We also show the number of parameters and the inference time on Set5 in Table I. Depending on the workflow for solving the CAR and SR problem, these methods can be categorized into the following three types: (1) SR: directly use pretrained SR models. (2) CAR+SR: the aforementioned two-stage method, which first removes the compression artifacts and then sends the output images to the SR model. (3) Joint CAR & SR: the single-stage method that jointly handles CAR and SR with one model. We report both the direct output and the self-ensembled output of our network.

According to Table I, CAJNN significantly outperforms the existing methods for all QFs, and yielding the highest overall PSNR across all five datasets with Set5. The improvement is consistently observed on SSIM, as well. Moreover, our model is more light-weight than most of the current models, including one-stage and two-stage in summation, which results in faster inference speed on the same hardware (all the tests are conducted on one Nvidia Titan Xp graphics card).

Figure 3 gives a qualitative example of the result of our model, where the input image is *woman* from the Set5 [25] that is downsampled and compressed by a wide range of quality factors from 10 to 100. It is worth noting that compression with very low quality factors causes a significant color shift to the hue and spatial distribution of the original image, which can be seen in the leftmost LR image (QF = 10). Our model is able to correctly restore the color aberrations of RGB images with a high consistency among different QFs.

##### Results on User Images
Besides the above experiments on standard test images, we also conduct experiments on real user images to demonstrate the effectiveness of our model. We mainly focus on the perceptual effect since there are no ground-truth images. Figure 4 shows the CAJNN results on real-world image from the WIDER face dataset [29].

For comparison, RCAN [21] and RRDB [13] are used as representative SR method, ARCNN [16] and DnCNN [37] are used as the representative CAR methods. The real-world images have unknown downsampling kernels and compression mechanisms, depending on the platforms. According to Figure 4, the SR methods generate images with obvious color shift and ringing artifacts. These artifacts are alleviated with two-stage methods. Still, the results are blurry. Compared with the two-stage methods, our CAJNN can provide SR outputs with sharp edges and rich details, which demonstrates the superiority of our proposed single-stage method when applied to real-world CARSR problems.

#### C. Results for Low-Resolution Text Recognition

Comparing the input LR image and our output in Figure 4, the texts become more readable after being processed by our model. Inspired by this observation, we conducted the following experiments to explore our model’s potential to leverage a real-scene text recognition task from low-resolution characters.

We compare the total accuracy of generic text detection on the ICDAR2013 Focused Scene Text dataset [30] with TPS-ResNet-BiLSTM-Attn [27] as the text recognition method. The baseline result is acquired by directly recognizing the original input images. As a comparison with the baseline, we use the CAJNN model as described in previous sections to generate artifact-free SR images from the original images and conduct recognition on the output images.

As can be seen in Table I, the preprocessing of CAJNN improves the recognition accuracy from 85.30% to 85.75%, which indicates that the outputs of our model are not only visually appealing to human viewers, but also include more distinct information for the text recognition network as shown in Figure 5. It is worth noting that our output image is $4 \times$ the size compared with the baseline inputs, and the average detection time is increased from 31.22s to 41.56s. Although the improvement in accuracy demonstrates the positive effect yielded by our model, the rise in computation is hard to ignore. Therefore, we disentangle the influence of SR and...
CAR by bicubic downsampling the CARSR output images and acquire the third recognition result. Since the image size remains the same as that of the original image, the detection time is identical to the baseline. Compared with the baseline, the recognition accuracy still improves 0.27% due to the reduction of compression artifacts, which indicates that our model is capable of extracting and maintaining critical features of input images. This experimental result points out a plausible direction for future text recognition research: the image quality plays a vital role in the recognition accuracy, which can be improved by utilizing the learned priors from a pretrained CARSR model.

TABLE II: Text recognition accuracy on the ICDAR 2013 Focused Scene Text dataset. Compared with the baseline method, the introduction of our CARSR method improves the detection performance by 0.45% (without downsampling) and 0.27% (with downsampling).

D. Results for Extremely Tiny Face Detection

Extremely tiny face detection is another practical, yet challenging task in high-level computer vision. Most of the state-of-the-art (SOTA) face detectors have already taken various scales and distortions into consideration to achieve impressive detection performance.

Fig. 5: Test samples of ICDAR2013 dataset (word_161, word_836). The first column shows original input images, the second column is the CARSR output generated by our method, and the third column is acquired by downsampling the second column. By comparing the detection results in the first and second columns, our method can serve as a supportive method for the recognition of low-resolution texts. Besides, the artifact-free image in the third column can also provide more recognizable features for the baseline model without increasing the image size.

D. Results for Extremely Tiny Face Detection

Extremely tiny face detection is another practical, yet challenging task in high-level computer vision. Most of the state-of-the-art (SOTA) face detectors have already taken various scales and distortions into consideration to achieve impressive detection performance. We proposed a solution to tackle tiny face detection by explicitly restoring an HR face from a small blurry one using a Generative Adversarial Network (GAN).

We experimentally validate the effect achieved by our CA-
JNN on tiny face images in the WIDER FACE dataset [29] by comparing the detection results from the following three types of data: original HR (serves as the baseline), downsampled LR (serves as the extremely tiny face inputs), and CARSR outputs from our model. [52] is applied as the backbone face detector (We use an unofficial PyTorch [53] implementation provided by https://github.com/varunagrawal/tiny-faces-pytorch).

Table III shows the Average Precision (AP) of the downsampled tiny images and our enhanced ones on all the three validation sets (easy, medium, and hard) of WIDER FACE [29]. From Table III, we observe that the data processed by CAJNN dramatically improves the detection of LR inputs from 0.317 to 0.611 in AP on the hard set. The reason is that the baseline detector performs downsampling operations by large strides on the tiny faces. Considering the fact that the tiny faces themselves contain less information than average, the detailed information of face structure is lost after several downsampling convolutions. In contrast, our CAJNN provides an artifact-free SR image, which can boost the detection performance by better utilizing the information of small faces. In Figure 6, the precision-recall curve of our reconstructed image (green line) is close to the ground truth (red line) on the easy and medium subsets. In the hard subset, our CAJNN yields a significant improvement compared to the LR curve. The gap between our output and the GT is due to the irreversible loss of information in extremely tiny faces that happens more frequently in the hard set during the downsampling process.

### E. Ablation Study

**Effect of Multi-scale Information** As discussed in previous sections, both intra- and inter-block context information is important for designing a CARSR network. In other low-level vision tasks, context information at different scales has already been proved to be effective in improving network performance. Inspired by the first convolution layer of the ResNet [54], previous researchers [55] applied $7 \times 7$ convolution to extract the context features for the video frame interpolation task. However, such big kernels bring a tremendous number of parameters to the network, especially when embedded in the feature domain, resulting in higher computational cost. Another way of enlarging the filter’s receptive field is to use a non-local module [56], [57], where the input images are downsampled by convolutional strides and processed at different scales. The non-local module has a rather complex structure and also a large number of parameters. In order to use the context information in a much simpler and lighter representation, our method adopts atrous convolution. By adjusting the dilation rate $r$, the filter can incorporate the context information from a larger receptive field without dramatically increasing the number of parameter as compared to the above methods.

We conduct an ablation study to illustrate the effect of different ways of representing contextual information in Table IV. In Rows 1-3, we compare the performance of the non-local module, ASPP, and sequential atrous pooling. Comparing the base model to Column 1 in Table IV, we can conclude that the introduction of multi-scale information via a non-local module can significantly improve the PSNR by 0.406 dB. This result validates the superiority of aggregating both intra- and inter-block features rather than using a purely local representation for the CARSR task. Furthermore, as seen by comparing Columns 1 and 2, replacing the non-local module by our well-designed ASPP can improve the PSNR by 0.002 dB. Although the improvement is rather small, it is worth noting that the ASPP has fewer convolution layers and parameters, which results in a smaller model size and fewer FLOPs. Remarkably, it can achieve results that are comparable, or even better than that yielded by models with more parameters. By comparing Columns 3 and 4, we also note that the PSNR of ASPP is higher than that of sequential atrous pooling by 0.03 dB, which means that the pyramid-fusion structure is more efficient in representing the multi-scale information.

| Input Data | Easy | Medium | Hard |
|------------|------|--------|------|
| GT         | 0.900| 0.887  | 0.792|
| LR         | 0.824| 0.692  | 0.317|
| LR + Ours  | 0.893| 0.857  | 0.611|

**TABLE IV: Ablation Study on the validation set (Set5).** We report the performance of CAJNN without the long-range skip connection and ASPP as the baseline. Rows 1-3 show the influence of different ways to extract contextual information by replacing ASPP with other network structures. Rows 4-5 compare the effect of two different upsampling methods on PSNR. The combination of the ASPP and Pixelshuffle modules yields the best performance, and thus is adopted in our network architecture.

| Model                  | Base | 1  | 2  | 3  | 4  |
|------------------------|------|----|----|----|----|
| Non-local module       | √    |    |    |    |    |
| ASPP                   | √    | √  | √  |    |    |
| Sequential atrous pooling | √  | √  | √  | √  | √  |
| Upconvolution          | √    | √  |    |    |    |
| Pixelshuffle           | √    | √  | √  | √  | √  |
| PSNR (dB)              | 28.292| 28.274| 28.216| 28.292| 28.262|

**TABLE III: Average precision of three data types in the WIDER FACE validation set [29] with the same face detector [52].** The application of our CARSR method greatly improves the detection performance with LR images on all three subsets.
TABLE V: Ablation Study on joint end-to-end supervision. We introduce the explicit reconstruction loss as a disentanglement mechanism of CAR and SR. By changing the weight of this loss term, we can study the effect of different levels of joint-supervision. Among all the settings, the model trained without the reconstruction loss performs best on our validation set.

| Model | a | b | c | Weight of reconstruction loss $\lambda$ |
|-------|---|---|---|---------------------------------------|
|       | 16 | 1 | 0 |                                       |
| PSNR (dB) | 27.501 | 27.627 | **27.672** |

and use it to explicitly supervise the intermediate CAR output $\hat{G}(f^L)$ after the context-aware module:

$$l^{LR} = l(l^{LQHR}, \hat{G}(f^L)).$$

(7)

Denoting the pixel-wise loss of the final output and ground truth (shown in Equation 1) as $l_{HR}$, the overall training loss becomes:

$$l = l_{HR} + \lambda l^{LR},$$

(8)

by increasing the weight $\lambda$, we can acquire models trained with higher disentanglement levels. We train three models with $\lambda = 0, 1, 16$ while keeping all the other factors the same. The performance of these models on our validation set is shown in Table V. The trend is obvious: the PSNR increases as the entanglement increases, which demonstrates the effectiveness of the joint CARSR method with a single-stage network.

V. CONCLUSION

In this paper, we propose a single-stage network for the joint CARSR task to directly reconstruct an artifact-free high-resolution image from a compressed low-resolution input. To address the CARSR problem, we make use of the contextual information by introducing a specially designed ASPP that integrates both intra- and inter-block features. Our experiments illustrate the effectiveness and efficiency of our method with both standard test images and real-world images. Moreover, the extensive experimental results reveal a high potential for enhancing the performance of current methods for various high-level computer vision tasks, e.g. real-scene resolution text recognition, and extremely tiny face detection.

ACKNOWLEDGMENT

This research is supported by HP Inc., Palo Alto, CA.

REFERENCES

[1] M.-Y. Shen and C.-C. J. Kuo, “Review of postprocessing techniques for compression artifact removal,” Journal of Visual Communication and Image Representation, vol. 9, no. 1, pp. 2–14, 1998.

[2] J. Allebach and P. W. Wong, “Edge-directed interpolation,” in Proceedings of 3rd IEEE International Conference on Image Processing, vol. 3. IEEE, 1996, pp. 707–710.

[3] B. Wronski, I. Garcia-Dorado, M. Ernst, D. Kelly, M. Kramm, C.-K. Liang, M. Levoy, and P. Milanfar, “Handheld multi-frame super-resolution,” arXiv preprint arXiv:1905.03277, 2019.

[4] X. Xiang, Y. Tian, Y. Zhang, Y. Fu, J. P. Allebach, and C. Xu, “Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3370–3379.

[5] Z. Xiao, M. Nguyen, E. Maggard, M. Shaw, J. Allebach, and A. Reibman, “Real-time print quality diagnostics,” in Image Quality and System Performance XIV (Part of IS&T Electronic Imaging 2017), no. 12. Society for Imaging Science and Technology, 2017, pp. 174–179.

[6] X. Xiang, R. Jessome, E. Maggard, Y. Bang, M. Cho, and J. Allebach, “Blockwise based detection of local defects,” in Image Quality and System Performance XVI (Part of IS&T Electronic Imaging 2019), no. 10. Society for Imaging Science and Technology, 2019, pp. 303–1.

[7] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in Proceedings of the European Conference on Computer Vision. Springer, 2014, pp. 184–199.

[8] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, “Enhanced deep residual networks for single image super-resolution,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 136–144.

[9] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2472–2481.

[10] “Residual dense network for image restoration,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

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

[12] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.

[13] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, “Ergan: Enhanced super-resolution generative adversarial networks,” in The European Conference on Computer Vision Workshops, September 2018.

[14] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in Proceedings of the European Conference on Computer Vision. Springer, 2016, pp. 694–711.
