Learning a Convolutional Neural Network for Image Compact-Resolution

Yue Li, Dong Liu, *Member, IEEE,* Houqiang Li, *Senior Member, IEEE,* Li Li, *Member, IEEE,* Zhu Li, *Senior Member, IEEE,* and Feng Wu, *Fellow, IEEE*

***Abstract*—We study the dual problem of image super-resolution (SR), which we term image compact-resolution (CR). Opposite to image SR that hallucinates a visually plausible high-resolution image given a low-resolution input, image CR provides a low-resolution version of a high-resolution image, such that the low-resolution version is both visually pleasing and as informative as possible compared to the high-resolution image. We propose a convolutional neural network (CNN) for image CR, namely CNN-CR, inspired by the great success of CNN for image SR. Specifically, we translate the requirements of image CR into operable optimization targets for training CNN-CR: the visual quality of the compact-resolved image is ensured by constraining its difference from a naively down-sampled version, and the information loss of image CR is measured by up-sampling/super-resolving the compact-resolved image and comparing that to the original image. Accordingly, CNN-CR can be trained either separately, or jointly with a CNN for image SR. We explore different training strategies as well as different network structures for CNN-CR. Our experimental results show that the proposed CNN-CR clearly outperforms simple bicubic down-sampling, achieves on average 2.25 dB improvement in terms of the reconstruction quality on a large collection of natural images. We further investigate two applications of image CR, i.e. low-bit-rate image compression and image retargeting. Experimental results show that the proposed CNN-CR helps achieve significant bits saving than High Efficiency Video Coding (HEVC) when applied to image compression, and produce visually pleasing results when applied to image retargeting.**

***Index Terms*—Compact-resolution (CR), convolutional neural network (CNN), down-sampling, High Efficiency Video Coding (HEVC), image compression, image retargeting, super-resolution (SR), up-sampling.**

I. Introduction

Changing the resolution of a digital image is a ubiquitous requirement. Consider to display a given image on a given device, once the native resolutions of the image and the device are not matched, we need the resolution change. This can be performed by re-sampling together with simple interpolation (e.g. bicubic), which enjoys high computational efficiency but usually suffers from kinds of visual artifacts in the re-sampled image, such as aliasing and blurring. Developing advanced methods for changing image resolution is then an intensively studied topic in the literature.

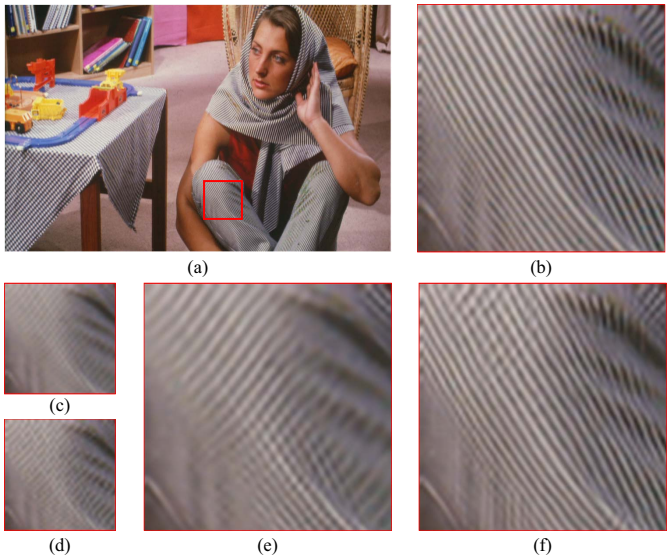

Fig. 1. (a) The original high-resolution image (Barbara). (b) A cropped region of (a). (c) Bicubic down-sampled ($\times 2$) result. (d) CNN-CR^{Joint} ($\times 2$) result. (e) CNN-based SR result of (c) (PSNR = 28.65dB). (f) CNN-based SR result of (d) (PSNR = 34.69dB). Note the quality difference between (c) and (d), and between (e) and (f).

On the one hand, increasing the resolution of image has been investigated for a long while, under the name of single image super-resolution (SR) in low-level computer vision. The existing image SR methods can be categorized into interpolation-based, reconstruction-based, and learning-based [1]. Recently, learning-based image SR using convolutional neural network (CNN) achieves significant improvement and competes favorably than the other methods [2]–[5], along with the great success of CNN in computer vision researches.

On the other hand, decreasing the resolution of image is not well studied yet. Related researches are distributed in several different domains: decreasing the resolution for image compression [6]–[8], decreasing the resolution for display device [9]–[12], perceptual quality-oriented down-scaling [13]–[15], and image retargeting [16]–[19]. In these researches, the process of decreasing image resolution is usually designed for each specific task rather than being generic.

Decreasing and increasing image resolution can be regarded

Date of current version September 13, 2018. This work was supported by the National Program on Key Basic Research Projects (973 Program) under Grant 2015CB351803, by the Natural Science Foundation of China (NSFC) under Grants 61772483, 61390512, and 61425026.

Y. Li is with the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei 230027, China. He was with University of Missouri-Kansas City, 5100 Rockhill Road, Kansas City, MO 64111, USA, as a visiting student (e-mail: lytt@mail.ustc.edu.cn).

D. Liu, H. Li, and F. Wu are with the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei 230027, China (e-mail: dongeliu@ustc.edu.cn; lihq@ustc.edu.cn; fengwu@ustc.edu.cn).

L. Li and Z. Li are with University of Missouri-Kansas City, 5100 Rockhill Road, Kansas City, MO 64111, USA (e-mail: lil1@umkc.edu; zhu.li@ieee.org).

as dual problem of each other. Decreasing resolution almost always incurs loss of information, while increasing resolution usually tries to recover the lost information. Thus, it is beneficial to consider the two problems jointly. In addition, both problems are not well defined so that appropriate regularization is helpful.

In this paper, we study the generic problem of decreasing image resolution. The problem is termed *image compact-resolution* (CR) to highlight its duality with image SR. Specifically, the purpose of image CR is to generate a low-resolution version of a given high-resolution image. As there are multiple ways to generate low-resolution version, we here impose two requirements on image CR to better define it. First, the generated low-resolution version, termed compact-resolved image, should be visually pleasing. Second, the compact-resolved image should be as informative as possible compared to the original high-resolution image; in other words, the image CR process should preserve as much information as possible.

Inspired by CNN-based image SR, we propose a learning-based approach for image CR using CNN, i.e., we train a CNN-CR for natural images. To that end, we translate the above two requirements of image CR into operable optimization targets for training CNN-CR. First, the visual quality of the compact-resolved image is ensured by constraining its difference from a naively down-sampled version. Second, the information loss of image CR is measured by up-sampling/super-resolving the compact-resolved image and comparing that to the original image. According to how the information loss is measured, CNN-CR can be trained either separately, or jointly with a CNN for image SR.

We explore different training strategies as well as different network structures for CNN-CR. And we evaluate the proposed CNN-CR on a large collection of natural images. Our experimental results show that CNN-CR clearly outperforms bicubic down-sampling, improves both the visual quality of the low-resolution image, and the objective and subjective quality of the high-resolution reconstruction (Fig. 1).

We further investigate two applications of image CR, i.e. low-bit-rate image compression and image retargeting. Experimental results demonstrate the effectiveness using CNN-CR in both applications, which leads to significantly better or comparable results than simple down-sampling.

The remainder of this paper is organized as follows. In Section II, we discuss related work about decreasing and increasing image resolution. Section III elaborates the formulation of learning-based image CR and the network structure of the designed CNN-CR. Section IV presents the training strategies for CNN-CR. Applications of the proposed CNN-CR in image compression and image retargeting are discussed in Section V. Section VI presents the experimental results, followed by conclusions in Section VII.

II. Related Work

A. Decreasing Image Resolution

Decreasing Resolution for Image Compression: It is a widely adopted strategy to decrease resolution before encoding and to increase resolution after decoding when the available bandwidth for transmission is limited [6]–[8]. Tsaig *et al.* proposed a variable projection method for optimizing the down- and up-sampling filters used in low-bit-rate image compression [8], where the filters are linear and their near-optimality is claimed with the premise of stationary signal. Jiang *et al.* proposed an end-to-end image compression framework based on CNN [6], where the down- and up-sampling filters are replaced with trained CNN models.

Decreasing Resolution for Display Device: Another series of works had been done to improve the quality of down-sampled images on certain display devices such as liquid crystal display (LCD). Because each pixel on a color LCD is actually composed by three individual sub-pixel stripes (i.e. RGB), the number of individual reconstruction points in LCD increases to three times when considering sub-pixels, which leads to the so-called sub-pixel rendering techniques that present small text better [20]–[23]. Accordingly, taking sub-pixel rendering into account during image down-sampling leads to improvement of the visual quality [9]–[12]. For example, Fang *et al.* proposed a sub-pixel based image down-sampling method to achieve down-sampled images with more details but less color fringing artifacts [12].

Perceptual Quality-Oriented Down-Scaling: Several researches were devoted to image down-scaling oriented for perceptual quality [13]–[15]. Kopf *et al.* proposed a filter-based down-scaling method by optimizing the shape and location of the down-sampling filters, making them better aligned with local image features [13]. This method significantly increases the details in the down-scaled images, but the optimization target correlates poorly with human perception. To overcome this, another method is proposed in [14], where the optimization target becomes the structural similarity (SSIM) [24] between the original and down-scaled images. Recently, Liu *et al.* further proposed a novel down-scaling method, in which the optimization is driven by two L_0-regularized priors [15].

Image Retargeting: Image retargeting consists of changing the aspect ratio as well as the resolution of images in order to best suit for different displays. While it is possible to down-sample an image by different factors along the horizontal and vertical directions, this can easily incur large distortion of the shape of objects. Advanced, content-aware methods have been developed for image retargeting [16]–[19]. For example, the seam carving method proposes to remove pixels in the least noticeable fashion to achieve decreasing of resolution [16].

Limitations: In the aforementioned researches, the process of decreasing image resolution is usually tailored for a specific task rather than being generic. In this paper, we consider the generic problem of image CR. We propose two requirements on image CR, i.e. the quality of the compact-resolved image and the information preservation during compact-resolution. On contrary, the previous researches for low-bit-rate compression did not consider the quality of the down-sampled image, while the previous researches for display device, for perceptual quality optimization, and for retargeting did not consider the information preservation during down-sampling (except for [14], where the SSIM between the original and down-scaled images is considered). In addition, there are few works about using deep learning-based methods

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2872876, IEEE Transactions on Image Processing

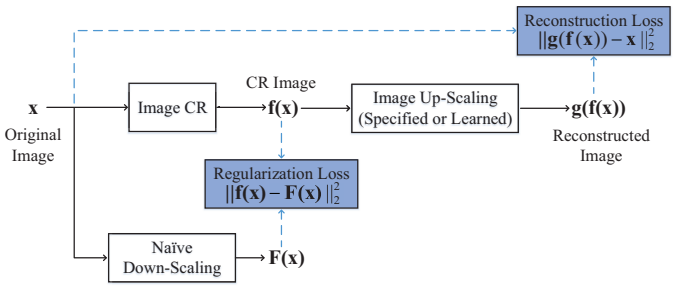

Fig. 2. A general framework for learning-based image CR.

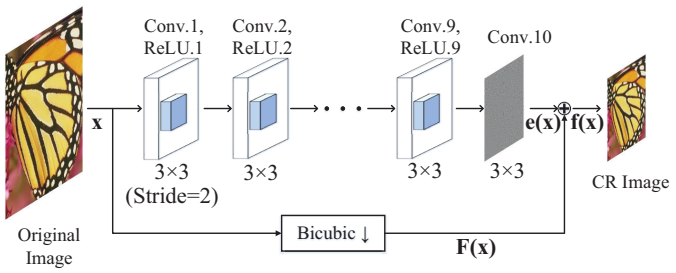

Fig. 3. The network structure of our designed CNN-CR.

for decreasing image resolution (except for [6]), which we investigate in this paper.

B. Increasing Image Resolution

Different from decreasing image resolution that is not well studied yet, increasing image resolution as a generic problem, also known as single image SR, has been studied intensively in the literature. Existing image SR methods can be classified into interpolation-based, reconstruction-based, and learning-based ones [1]. Learning-based image SR using CNN receives great attention of researchers since the pioneering work of Dong *et al.* [2]. Its superior performance can be attributed to the joint optimization of the three separate steps of the previous sparse coding-based SR methods [1]. Subsequent studies on CNN based SR methods further boost the performance by utilizing more elaborate network structures [5] and/or more efficient training methods [3], [4]. For example, Lim *et al.* proposed the so-called EDSR network for image SR [4], which achieves the state-of-the-art performance in the task of single image SR and won NTIRE2017 Image Super-Resolution Challenge [25].

Since increasing and decreasing image resolution have very different implications, the CNN trained for SR is not directly applicable for CR. Besides, the data for training CNN for SR are easy to achieve but how to get data for training CNN for CR is not straight-forward. Since decreasing and increasing image resolution are dual problem of each other, it may be beneficial to consider them jointly. This idea motivates the work of this paper.

III. CNN for Learning-Based Image CR

In this section, we first discuss the formulation of learning-based image CR problem, and then propose a network as CNN-CR.

A. Learning-Based Image Compact-Resolution

In this paper, we impose two requirements on image CR. When using the learning approach for image CR, we need to translate the requirements into operable optimization targets. Thus, we propose a general framework for learning-based image CR, which is shown in Fig. 2. In essence, as we have no "ground-truth" for the compact-resolved image, we define two loss functions, known as the *reconstruction* loss and the *regularization* loss, respectively, to evaluate the compact-resolved image indirectly.

1) Reconstruction Loss: Our proposed reconstruction loss is to evaluate the information loss during image CR. Theoretically speaking, the information loss incurred by down-sampling is related to the high-frequency energy of the signal. However, natural images are too complicated to be characterized by an analytical frequency decomposition. Thus, we want to seek a tractable loss function. For example, we can compare the original image and the compact-resolved image, and evaluate their difference. As images of different spatial resolutions are not directly comparable, we may either down-scale the original image or up-scale the compact-resolved image, before comparison. Since down-scaling may lose information, we up-scale the compact-resolved image and compare against the original image.

Denote the original image as $\mathbf{x}$, the mapping function of image CR as $\mathbf{f}$ and the mapping function of up-scaling as $\mathbf{g}$, then the reconstruction loss is defined as

$$\mathbf{J}_{rec} = ||\mathbf{g}(\mathbf{f}(\mathbf{x})) - \mathbf{x}||_2^2 \tag{1}$$

It is worth noting that in a previous research [14], an optimization target similar to our reconstruction loss was proposed, but used SSIM as the metric to compare the original and reconstructed images. On contrary, we use mean-squared error in (1) that is easier to compute.

In addition, when our objective is to learn the image CR ($\mathbf{f}$ in (1)), we also need to specify the up-scaling ($\mathbf{g}$ in (1)) before we can perform learning. It is also possible that we do not predefine $\mathbf{g}$, instead, we can learn $\mathbf{f}$ and $\mathbf{g}$ jointly. In the following Section IV, we will discuss these two cases: the first is that we specify $\mathbf{g}$ and learn $\mathbf{f}$ alone, the second is that we learn $\mathbf{f}$ and $\mathbf{g}$ jointly. For simplicity, we call the first case *separate* learning and the second case *joint* learning.

2) Regularization Loss: Our proposed regularization loss is to ensure the visual quality of the compact-resolved image. Let $\mathbf{F}$ be the ideal low-pass and decimation function, which should be able to faithfully preserve all the low-frequency information of the input image. The low-resolution image generated by $\mathbf{F}$ should be smooth and have no aliasing. Given the function $\mathbf{F}$, we define the regularization loss as

$$\mathbf{J}_{reg} = ||\mathbf{f}(\mathbf{x}) - \mathbf{F}(\mathbf{x})||_2^2 \tag{2}$$

To better explain (2), we decompose the compact-resolved image $\mathbf{f}(\mathbf{x})$ into the smooth component $\mathbf{F}(\mathbf{x})$ and the detail component $\mathbf{e}(\mathbf{x})$,

$$\mathbf{f}(\mathbf{x}) = \mathbf{F}(\mathbf{x}) + \mathbf{e}(\mathbf{x}) \tag{3}$$

Then (2) can be rewritten as,

$$\mathbf{J}_{reg} = ||\mathbf{e}(\mathbf{x})||_2^2 \tag{4}$$

That is, the regularization loss is essentially the energy of $\mathbf{e}(\mathbf{x})$. We define this regularization loss due to the following two motivations:

- By restricting the energy of $\mathbf{e}(\mathbf{x})$, we ensure that the low frequency component of $\mathbf{f}(\mathbf{x})$ does not differ too much from $\mathbf{F}(\mathbf{x})$. Then the compact-resolved image will still be smooth, and the high-frequency information provided by $\mathbf{e}(\mathbf{x})$ can present more details although at the risk of introducing aliasing.
- Though the energy of $\mathbf{e}(\mathbf{x})$ is restricted, $\mathbf{e}(\mathbf{x})$ is still allowed to have a flexible distribution so as to allow the compact-resolved image to be as informative as possible. The high frequency information provided by $\mathbf{e}(\mathbf{x})$ should carry as much hint as possible for reconstruction.

It is worth noting that an ideal low-pass filter is not practical. In this paper, we use bicubic down-sampling to approximate the function $\mathbf{F}$. Other filters such as Gaussian can be alternatives.

3) Combined Loss: We combine (1) and (2) to achieve the entire objective function for learning

$$\mathbf{J} = \mathbf{J}_{rec} + \lambda \cdot \mathbf{J}_{reg} = ||\mathbf{g}(\mathbf{f}(\mathbf{x})) - \mathbf{x}||_2^2 + \lambda||\mathbf{f}(\mathbf{x}) - \mathbf{F}(\mathbf{x})||_2^2 \tag{5}$$

where λ is a parameter that controls the relative weight of the regularization loss. When λ is larger, the compact-resolved image will be more smooth and contain less high-frequency information, and also preserve less information of the original image accordingly. We empirically set λ as 0.7 in our final results. Experimental results in Section VI-B4 will give further analyses on different λ values.

B. The Proposed CNN-CR

The structure of the proposed CNN-CR is illustrated in Fig. 3. The designed CNN-CR consists of several convolutional layers, all of which except the first and the last are of the same configuration: 64 filters with kernel size 3×3, followed by rectified linear unit (ReLU) as nonlinear activation function. The first layer operates on the input image and serves as a resolution decreasing layer. For example, in the case of 2× down-sizing, the filters in the first layer will be equipped with stride = 2. The last layer is used for generating the compact-resolved image, thus contains a single filter with kernel size 3×3. Note that the resolution decreasing layer is placed at the very beginning of our CNN-CR. This reduces the computational cost of CNN-CR as the following layers are working on smaller feature maps.

1) Network Depth: To the best of our knowledge, we are the first to address the issue of training a CNN for image CR. As pointed out in [26], the depth of the network should match the inherent complexity of the given task. Therefore, we pay our first attention to the depth of CNN-CR when designing the network. Indeed, recent works show that deeper networks seem to always increase the final performance of CNN models, such as for image recognition [27], [28] and for image SR [3], [29]. However, deeper network will inevitably incur higher computational complexity.

When determining the depth of our CNN-CR, we first try a 5-layer network, then a noticeable improvement on reconstruction quality is observed by increasing the depth from 5-layer to 10-layer. Further increasing the depth to 15-layer brings a little performance gain but incurs much higher complexity. Based on this observation, we choose the depth of 10-layer as an appropriate compromise between efficiency and complexity for the image CR task. Experimental results related to network depth are presented in Section VI-B5.

2) Down-Sizing Operation: In the CNN models for image recognition, pooling is usually adopted to decrease the resolution of feature maps. It aggregates the feature values within a local window into a single feature value, loses the spatial resolution meanwhile. Pooling works well because it provides dimensionality reduction and kinds of local invariance, which are useful in the image recognition task. However, the situation becomes different for the image CR task. Since we want the compact-resolved image to be as informative as possible, we care the spatial resolution more than the local invariance. Thus, we adopt convolution with a stride, as shown in Fig. 3, to implement the down-sizing operation in our CNN-CR. This way, the down-sized feature maps correspond to the original image at fixed pixel locations, which can facilitate the reconstruction. Experimental results, which verify the advantage of convolution with stride than pooling, will be presented in Section VI-B5.

3) Residual Learning: Residual learning in CNN is proposed by He *et al.*, which introduces skip-layer connections to achieve both faster convergence in training and better performance [28]. We also adopt residual learning in our CNN-CR and have observed better performance. As shown in Fig. 3, the bicubic down-sampled version of the input image is directly added to the output of CNN-CR to produce the compact-resolved image. After explicitly giving rise to the residual learning, the CNN-CR only learns the detail component $\mathbf{e}(\mathbf{x})$, whose energy is limited according to (4). The empirical benefit of residual learning will be reported in Section VI-B5.

IV. Training of CNN-CR

As mentioned before, to learn the image CR ($\mathbf{f}$ in (1)), we may either specify the up-scaling ($\mathbf{g}$ in (1)), or leave the up-scaling to be learned.

A. Separate Learning of CNN-CR

Shall we specify the up-scaling, we are able to train the CNN-CR separately. To facilitate the training, especially using gradient descent like algorithms, the up-scaling operator is better to be simple and differentiable. Thus, in this paper, we use bilinear as the specified up-scaling operator. Accordingly, we solve the following optimization problem numerically,

$$\begin{aligned}\theta_{\mathbf{f}}^{\star} &= \arg\min_{\theta_{\mathbf{f}}} \sum_{i=1}^{n} \mathbf{J}(\mathbf{x}_i; \theta_{\mathbf{f}}) \\ &= \arg\min_{\theta_{\mathbf{f}}} \sum_{i=1}^{n} (||\mathbf{g}(\mathbf{f}(\mathbf{x}_i; \theta_{\mathbf{f}})) - \mathbf{x}_i||_2^2 \\ &\quad + \lambda||\mathbf{f}(\mathbf{x}_i; \theta_{\mathbf{f}}) - \mathbf{F}(\mathbf{x}_i)||_2^2)\end{aligned} \tag{6}$$

where n is the number of training samples, $\mathbf{x}_i$ is a sample, $\theta_{\mathbf{f}}$ is the parameter set of the CNN-CR mapping function $\mathbf{f}$, and $\mathbf{g}$ denotes the bilinear up-sampling operator.

B. Joint Learning of CNN-CR and CNN-SR

There are at least two reasons that motivate us to study the joint learning of CNN-CR and CNN-SR. First, in some applications, such as down-/up-sampling-based image compression, CR and SR are both used in an integrated system, so it would be better to consider them jointly. Second, it has been generally acknowledged that CNN-based image SR outperforms traditional image SR methods (especially simple bilinear/bicubic) significantly. So we reasonably claim that CNN-SR may better exploit the information in the low-resolution image. Accordingly, when using CNN-SR, the benefit of preserving more information in the compact-resolved image (as a requirement of image CR) can be better unleashed, which may lead to a better trained CNN-CR.

When we want to train CNN-CR and CNN-SR jointly, we do not need to specify the CNN-SR parameters, but we still need to specify the network structure. In this paper, we revise the EDSR baseline structure [4] for our study, i.e., we replace the convolution-shuffle layer in the EDSR baseline structure with a deconvolution layer. We choose to revise the EDSR baseline structure due to its simplicity than the EDSR network, while the EDSR network represents the state-of-the-art of CNN-SR. It is worth noting that other CNN structures for SR can be used as well. In addition, the CNN-SR here is used only for training; after being well trained, CNN-CR can be used jointly with other SR networks as well (e.g. in Section VI-B6).

One may notice that our end-to-end network for joint learning of CNN-CR and CNN-SR is analogous to an auto-encoder [30]. However, it is worth noting that the normal auto-encoder scheme does not impose constraints on the encoded features, corresponding to the compact-resolved image in Fig. 2. Moreover, in the normal auto-encoder, the encoding and decoding parts are symmetric, but we do not require CNN-CR and CNN-SR to be symmetric. More discussions on the symmetric/asymmetric design will be presented in Section VI-B6.

Following the idea in [30] that good initialization is beneficial for training an auto encoder, we conduct a progressive training including three steps for the optimization of our end-to-end network.

First, we try to seek a good initialization for the SR part by separately training the CNN-SR. Following the common practice in training CNN for image SR, we use bicubic down-sampled version as input to CNN-SR and train the CNN to minimize the following loss

$$\theta_{\mathbf{g}}^0 = \arg\min_{\theta_{\mathbf{g}}} \sum_{i=1}^{n} ||\mathbf{g}(\mathbf{y}_i;\theta_{\mathbf{g}}) - \mathbf{x}_i||_2^2 \tag{6}$$

where $\mathbf{y}_i$ is the bicubic down-sampled version of the high-resolution training image $\mathbf{x}_i$, and $\theta_{\mathbf{g}}$ is the parameter set of the up-sampling mapping function $\mathbf{g}$, i.e. CNN-SR here.

Second, we fix the parameters of CNN-SR, and train the CNN-CR to minimize the reconstruction loss, i.e.

$$\theta_{\mathbf{f}}^0 = \arg\min_{\theta_{\mathbf{f}}} \sum_{i=1}^{n} ||\mathbf{g}(\mathbf{f}(\mathbf{x}_i;\theta_{\mathbf{f}});\theta_{\mathbf{g}}^0) - \mathbf{x}_i||_2^2 \tag{7}$$

where $\theta_{\mathbf{f}}$ is the parameter set of the CNN-CR mapping function $\mathbf{f}$.

Third, we fine-tune the parameters of the entire end-to-end network to minimize the joint loss:

$$\begin{aligned}\{\theta_{\mathbf{f}}^\star, \theta_{\mathbf{g}}^\star\} &= \arg\min_{\theta_{\mathbf{f}},\theta_{\mathbf{g}}} \sum_{i=1}^{n} \mathbf{J}(\mathbf{x}_i;\theta_{\mathbf{f}},\theta_{\mathbf{g}}) \\ &= \arg\min_{\theta_{\mathbf{f}},\theta_{\mathbf{g}}} \sum_{i=1}^{n} (||\mathbf{g}(\mathbf{f}(\mathbf{x}_i;\theta_{\mathbf{f}});\theta_{\mathbf{g}}) - \mathbf{x}_i||_2^2 \\ &\qquad + \lambda||\mathbf{f}(\mathbf{x}_i;\theta_{\mathbf{f}}) - \mathbf{F}(\mathbf{x}_i)||_2^2)\end{aligned} \tag{9}$$

The previously achieved parameters $\{\theta_{\mathbf{f}}^0, \theta_{\mathbf{g}}^0\}$ are used for initialization. With the progressive training, we can achieve a better trained model, as will be shown in Section VI-B5.

V. Applications of the Proposed CNN-CR

To evaluate the potential benefits of the proposed CNN-CR, we investigate how to use it in two applications including image retargeting and low-bit-rate image compression. For retargeting, we may use either the separately trained model or the jointly trained model; while for image compression, as mentioned before, it is better to use the jointly trained CNN-CR and CNN-SR.

A. Application in Image Retargeting

Retargeting in general refers to the task of changing resolution to suit for different display devices. Since the proposed CNN-CR ensures the quality of the compact-resolved images, CNN-CR can be used in image retargeting. The only issue is how to provide arbitrary resolution in CNN-CR, which can be solved by replacing the first layer of CNN-CR with a differentiable re-sampling layer [31].

Different from most of the retargeting methods, CNN-CR is not content-aware and thus may incur structural distortion if the aspect ratio of image is to be changed. Nonetheless, when the aspect ratio is not changing, using CNN-CR provides the benefit of keeping the content faithfully. In addition, CNN-CR is friendly to reconstruction and thus favorable when retargeting is combined with transmission, e.g. the retargeted image is transmitted to the target device and user may want to enlarge the retargeted image for viewing.

B. Application in Image Compression

The increasing popularity of ultra high definition (UHD) television [32] raises a great challenge to image and video compression, especially when the available transmission bandwidth is limited. However, though image capturing and displaying devices support higher and higher resolution, such resolution is not necessary to carry the important visual information in nature images. Thus, the common wisdom is to decrease image resolution before encoding and increase image

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2872876, IEEE Transactions on Image Processing

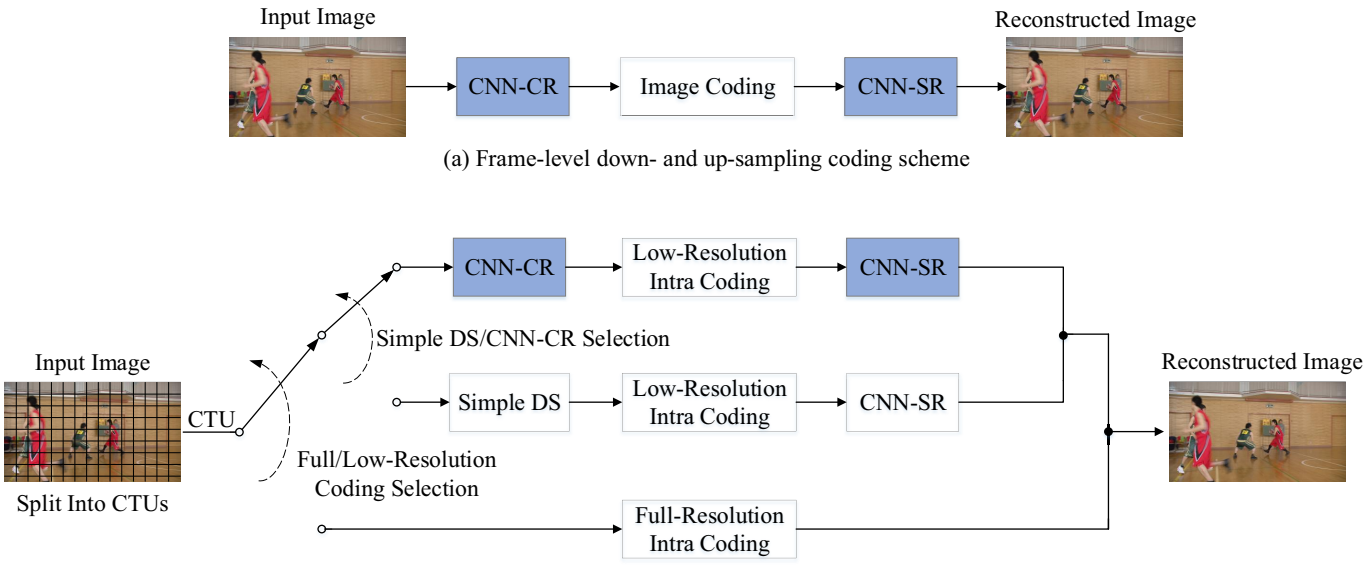

Fig. 4. The studied image compression schemes that perform frame-level down- and up-sampling, and block-level adaptive down- and up-sampling, respectively.

resolution after decoding so as to meet limited bandwidth [33]–[40].

Most of the existing down/up-sampling-based compression schemes adopt fixed, hand-crafted filters to decrease resolution, which loses much information of the original image and thus limits the capability of such schemes. As our CNN-CR is designed to keep as much information as possible, we expect it to outperform the fixed down-sampling filters. Specifically, we design two compression schemes that are shown in Fig. 4 and known as frame-level down- and up-sampling scheme and block-level adaptive down- and up-sampling scheme, respectively.

1) Frame-Level Down- and Up-Sampling Scheme: In Fig. 4 (a), CNN-CR is adopted to decrease the image resolution, while the smaller image is then compressed by normal intra coding methods, such as JPEG, JPEG2000, and HEVC intra coding. The decoded image is then super-resolved by a CNN to reconstruct the image at its native resolution. This scheme is universal as any existing image/intra coding methods can be adopted herein.

One may easily notice the similarity between the scheme in Fig. 4 (a) and the end-to-end structure in Fig. 2 (when up-scaling is learned), except for the intra coding module. It is then interesting to consider, whether it is possible to insert the intra coding module into Fig. 2 for end-to-end training? Unfortunately this seems difficult as the intra coding module is usually *not* a differentiable computation unit. We leave this problem to future work.

It is also worth noting that, since the down- and up-sampling strategy is usually adopted at low bit-rates, the intra coding in Fig. 4 (a) works at low bit-rates and causes significant distortion to the down-sized image. Thus, it is beneficial to deal with such distortion during the CNN-based SR process. During our experiments, we retrain the CNN-SR with compressed low-resolution images as input and original high-resolution images as target. We also observe that it is more beneficial to train a separate model for each quality level of intra coding. During the retraining, it is possible to use a different CNN structure for SR. In other words, we expect the trained CNN-CR model not to depend on a specific CNN-SR. So we intentionally change the CNN-SR structure in our experiments, as will be detailed in Section VI-E1.

2) Block-Level Adaptive Down- and Up-Sampling Scheme: While the frame-level scheme is simple, its capability may be constrained as discussed in [41]: since image content is locally variable, using a fixed down-sampling ratio may be over-simplified to deal with different kinds of content. A block-level down- and up-sampling scheme is preferred when we want to enable adaptability. For example, when dividing an image into multiple blocks that are compressed one by one, each block can be down-sampled by a different ratio, or using a different down-sampling filter, or using a different quality level of coding, or using a different up-sampling filter, etc.

Extended from the scheme proposed in our previous work [41], the designed block-level adaptive scheme using CNN-CR is shown in Fig. 4 (b). The scheme is based on HEVC intra coding, and the basic block is coding tree unit (CTU)[1]. In the scheme, we explore two kinds of adaptability. First, each CTU can be either down-sampled and coded, or directly coded at native resolution. Second, if coded at low resolution, either CNN-CR or simple down-sampling filter can be used for down-sizing. The simple down-sampling filter comes from [42], which was adopted in scalable video coding schemes. Low-resolution coded CTUs are super-resolved by trained CNN models for reconstruction. Similar to the frame-level scheme, we retrain the CNN-SR to cope with the compression

[1]In HEVC, a CTU is recursively divided into smaller coding units (CUs) with a quadtree structure. We did not consider CU-level adaptive down- and up-sampling, since it may incur very high computational complexity for mode decision. We plan to investigate this issue in the future.

artifacts in the low-resolution coded CTUs. And we intentionally change the CNN-SR structure during retraining. For CNN-CR down-sizing and for simple down-sampling, we train different CNN-SR models to match. We also train different models for different quality levels of intra coding.

Note that low-resolution coding usually results in less bit-rate but larger distortion (due to loss of information during down-sampling) compared to full-resolution coding, if using the same coding parameter (e.g. quantization parameter in HEVC). Such bit-rate shift is not friendly to the mode decision, when we want to compare the rate-distortion costs of low- and full-resolution coding. To alleviate the bit-rate shift, we intentionally lower the quantization parameter of low-resolution coding, as we did in our previous work [41].

VI. Experimental Results

In this section, we present the results about image CR, the comparisons between image CR and perceptual down-scaling, and applications of image CR in image retargeting and image compression.

A. Results of Separately Learned CNN-CR

1) Settings: We use the DIV2K dataset [25], which is a newly released high-quality image dataset for studying image SR, to generate training data. The DIV2K dataset consists of 800 training images, 100 validation images, and 100 test images. We use all of them for training. As there is no benchmark dataset for evaluating image CR methods, we select four well-known benchmark datasets, which are previously used for image SR, as testing datasets to verify the performance of our CNN-CR. The four testing datasets are known as Set5 [43], Set14 [44], Urban100 [45] and BSD100 [46]. The training of CNN-CR has the following configurations: ADAM optimizer [47] with hyper-parameters $\beta_1 = 0.9$, $\beta_2 = 0.9$, and $\epsilon = 10^{-8}$, minibatch size is 16, learning rate is initialized as 10^{-4} and halves after every 10 epochs. In this subsection, the CNN-CR is learned separately, and the resulting model is named CNN-CR^{Sep}.

2) Information Loss of CR: Since there is no "ground-truth" for compact-resolved images, the quantitative evaluation of image CR methods is an open problem. In this paper, we propose to utilize the reconstruction results to evaluate information loss of image CR quantitatively. That is, we use different methods to decrease image resolution and then use different methods to increase image resolution back, and evaluate the quality of the reconstructed images with respect to the original images. The reconstruction quality results are summarized in Table I.

It can be observed that, CNN-CR^{Sep} outperforms bicubic down-sampling, and achieves on average 1.25 dB improvement in terms of reconstruction quality over all the four testing datasets, when using bilinear up-sampling for reconstruction. This result is expected, since the CNN-CR^{Sep} was trained together with bilinear up-sampling. We further test the cases of other up-sampling filters, like bicubic and lanczos. As shown in Table I, CNN-CR^{Sep} + bicubic (resp. lanczos) up-sampling also performs better than bicubic down-sampling + bicubic (resp. lanczos) up-sampling, in the sense that on average 0.38 (resp. 0.36) dB improvement is achieved over the four testing datasets. These results confirm that the preserved information introduced by CNN-CR^{Sep} can boost the reconstruction quality.

In addition to the quantitative evaluation, we also inspect the visual quality of the reconstruction results. As shown in Fig. 5, CNN-CR^{Sep} plus bilinear up-sampling can better recover the characters in the reconstructed image, which is attributed to the preserved information in the compact-resolved image.

3) Quality of CR Images: Recall one of the two requirements of image CR is that the compact-resolved image should be visually pleasing. In the proposed methods, we constrain the difference between the compact-resolved image and the bicubic down-sampled image so as to meet this requirement. Accordingly, we calculate the PSNR between the compact-resolved image and the bicubic down-sampled image, which is also shown in Table I. The PSNR reaches 30 dB on average for the testing datasets, which indicates that certain differences exist between compact-resolved images and bicubic down-sampled images. Given close inspection of the visual quality, such as the results shown in Fig. 5, one can notice that the compact-resolved images contain more image details and look more vivid than the bicubic down-sampled images.

4) Regularization: We also train CNN-CR^{Sep} in the absence of the regularization loss, and evaluate the reconstruction quality. Table II summarizes these results. By comparing the results in Table II and Table I, we can find that, when discarding the regularization loss and using bilinear up-sampling, the reconstruction quality becomes better, but when using bicubic (lanczos) up-sampling, the reconstruction quality is much worse. Note that the bilinear up-sampling was used in training. These results show that the regularization loss is vital for the trained CNN-CR^{Sep} to be generalizable, which is the reason why the loss is termed "regularization."

B. Results of Jointly Learned CNN-CR

1) Settings: We adopt the same settings as in the previous subsection, except that the joint learning of CNN-CR and CNN-SR is completed in the three steps (Section IV-B). The trained CNN-CR model is named CNN-CR^{Joint}.

2) Information Loss of CR: The reconstruction quality results are summarized in Table III. It can be observed that, CNN-CR^{Joint} brings even higher gain than CNN-CR^{Sep}, and leads to on average 2.25 dB improvement compared to bicubic down-sampling in terms of reconstruction quality over all the four testing datasets when the scaling factor is 2. It is worth noting that, for our CNN-CR^{Joint} and bicubic down-sampling, we train a different CNN-SR model to match, so as to ensure fair comparison. Table III also includes the results of bicubic down-sampling plus EDSR, which are cited from [4] to represent the state-of-the-art results of single image SR. It is obvious that our CNN-CR^{Joint} plus CNN-SR can outperform the results in [4] by a considerable margin. Also note that when using bicubic down-sampling, our adopted CNN-SR performs worse than EDSR, since our adopted CNN-SR has less layers and less capability than EDSR. Such results again

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2872876, IEEE Transactions on Image Processing

Fig. 5. (a) The original high-resolution image (ppt3 from Set14 dataset). (b) The difference between the bicubic down-sampled ($\times 2$) result and the CNN-CR^{Sep} ($\times 2$) result, difference is scaled and normalized for display. (c) Cropped regions of (a). (d) Bicubic down-sampled results, and then bilinear up-sampled results (PSNR = 25.80dB). (e) CNN-CR^{Sep} results, and then bilinear up-sampled results (PSNR = 27.52dB).

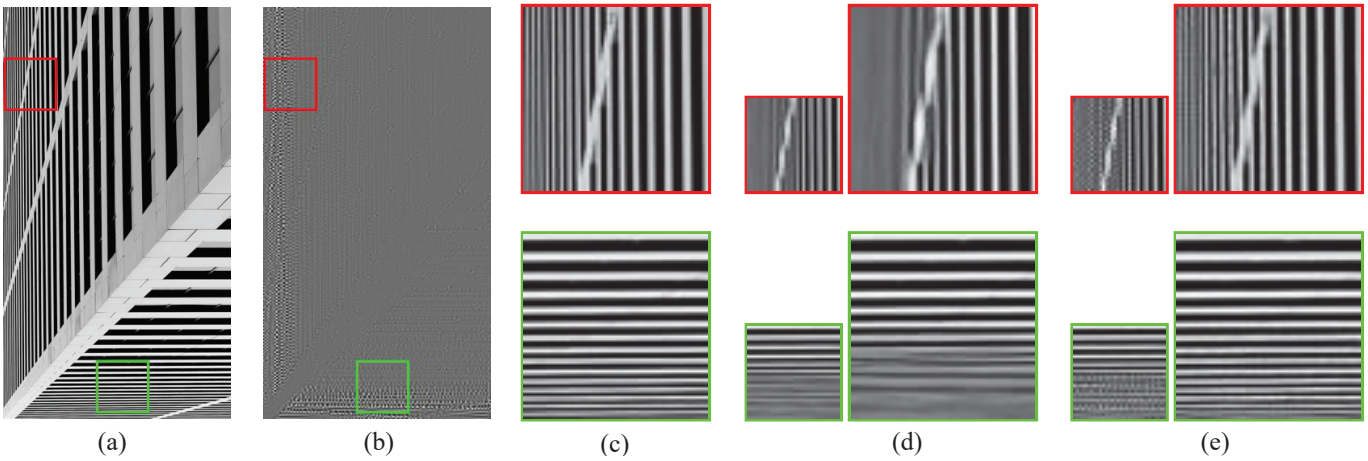

Fig. 6. (a) The original high-resolution image (Img011 from Urban100 dataset). (b) The difference between the bicubic down-sampled ($\times 2$) result and the CNN-CR^{Joint} ($\times 2$) result, difference is scaled and normalized for display. (c) Cropped regions of (a). (d) Bicubic down-sampled results, and then CNN-based SR results (PSNR = 20.61dB). (e) CNN-CR^{Joint} results, and then CNN-based SR results (PSNR = 29.96dB).

TABLE I

RECONSTRUCTION QUALITY (PSNR) OF USING DIFFERENT METHODS TO DECREASE IMAGE RESOLUTION ($\downarrow \times 2$) AND THEN TO INCREASE IMAGE RESOLUTION ($\uparrow \times 2$). CNN-CR^{Sep} IS A SEPARATELY TRAINED CNN-CR MODEL (I.E. WITHOUT CNN-SR). RESULTS ENCLOSED IN PARENTHESES ARE THE PSNR VALUES BETWEEN THE COMPACT-RESOLVED IMAGE AND THE BICUBIC DOWN-SAMPLED IMAGE. GAIN IS CALCULATED BETWEEN CNN-CR^{Sep} + BILINEAR/BICUBIC/LANCZOS UP-SAMPLING AND BICUBIC DOWN-SAMPLING + BILINEAR/BICUBIC/LANCZOS UP-SAMPLING, RESPECTIVELY.

	Bicubic $\downarrow$ Bilinear $\uparrow$	CNN-CR^{Sep} $\downarrow$ Bilinear $\uparrow$	Gain	Bicubic $\downarrow$ Bicubic $\uparrow$	CNN-CR^{Sep} $\downarrow$ Bicubic $\uparrow$	Gain	Bicubic $\downarrow$ Lanczos $\uparrow$	CNN-CR^{Sep} $\downarrow$ Lanczos $\uparrow$	Gain
Set5	32.26	33.55 (33.58)	1.29	33.67	33.36	−0.31	33.72	33.37	−0.35
Set14	28.97	30.27 (32.00)	1.30	30.01	30.30	0.29	30.05	30.32	0.27
BSD100	28.70	29.80 (31.95)	1.10	29.57	29.86	0.29	29.61	29.87	0.26
Urban100	25.60	26.99 (28.92)	1.39	26.56	27.09	0.53	26.60	27.11	0.51
Average	27.38	28.63 (30.61)	1.25	28.32	28.70	0.38	28.36	28.72	0.36

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2872876, IEEE Transactions on Image Processing

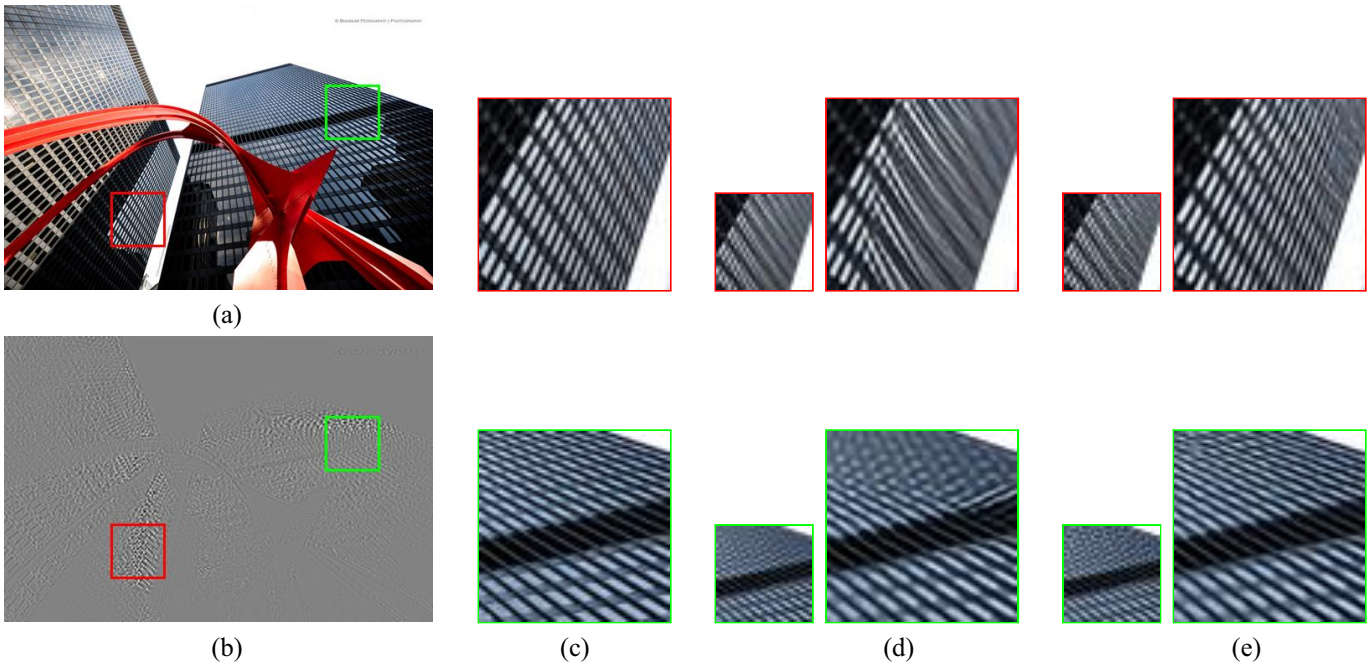

Fig. 7. (a) The original high-resolution image (Img062 from Urban100 dataset). (b) The difference between the bicubic down-sampled ($\times 2$) result and the CNN-CR^{Joint} ($\times 2$) result, difference is scaled and normalized for display. (c) Cropped regions of (a). (d) Bicubic down-sampled results, and then CNN-based SR results (PSNR = 25.47dB). (e) CNN-CR^{Joint} results, and then CNN-based SR results (PSNR = 29.07dB).

TABLE II
RECONSTRUCTION QUALITY (PSNR) OF CNN-CR^{SepNoReg} ($\downarrow \times 2$) + DIFFERENT UP-SAMPLING METHODS, WHERE CNN-CR^{SepNoReg} IS TRAINED IN THE ABSENCE OF THE REGULARIZATION LOSS. RESULTS ENCLOSED IN PARENTHESES ARE THE PSNR VALUES BETWEEN THE COMPACT-RESOLVED IMAGE AND THE BICUBIC DOWN-SAMPLED IMAGE.

	Bilinear ↑	Bicubic ↑	Lanczos ↑
Set5	33.46 (29.24)	31.80	31.74
Set14	30.38 (27.33)	29.17	29.14
BSD100	29.91 (27.03)	28.78	28.73
Urban100	27.23 (24.11)	26.06	26.02
Average	28.80 (25.77)	27.63	27.59

demonstrate that the proposed CNN-CR^{Joint} preserves more information during decreasing image resolution compared to bicubic down-sampling, and the preserved information can boost the performance of image SR significantly.

We also compare CNN-CR^{Joint} with bicubic down-sampling when both are equipped with bicubic up-sampling. It turns out that CNN-CR^{Joint} incurs a little drop of reconstruction quality, since the preserved high frequency information introduced by CNN-CR^{Joint} is not well exploited by simple bicubic up-sampling. This is contrary to the fact in the previous subsection, that the preserved information introduced by CNN-CR^{Sep} can be well exploited by different up-sampling methods, such as bicubic and lanczos. We then recommend the usage of advanced image SR methods (like CNN-SR) together with CNN-CR^{Joint}.

We also inspect the visual quality of the reconstruction results. As shown in Fig. 1, Fig. 6, and Fig. 7, the proposed CNN-CR^{Joint} plus CNN-SR successfully recover the very thin edges and fine textures in the reconstructed images. Note that recovering such image details is very challenging for image SR methods if using bicubic down-sampling.

3) Quality of CR Images: We calculate the PSNR between the compact-resolved images and the bicubic down-sampled images, and show the results in Table III. The PSNR reaches more than 40 dB on average for the testing datasets, indicating that the compact-resolved images achieved by CNN-CR^{Joint} are very similar to the bicubic down-sampled images. This partially ensures the visual quality of the compact-resolved images.

Given close inspection of the visual quality, such as the results shown in Fig. 1, Fig. 6, and Fig. 7, one can notice the following characteristics of the compact-resolved images generated by CNN-CR^{Joint}. On the one hand, the compact-resolved images preserve more high frequency information to facilitate a more accurate reconstruction, then look more vivid than the bicubic down-sampled images. On the other hand, preserving high frequency information introduces the risk of aliasing. For example in Fig. 6 (e), the compact-resolved image looks "noisy" in the regions that have extremely abundant high frequency information in the original image. In short, the CNN-CR tries to preserve high frequency information, which is a double-edged sword for visual quality. It then emphasizes the importance of the weight parameter λ that controls the energy of the preserved high frequency information. More results about the weight will be provided in Section VI-B4.

4) On the Regularization Weight: We perform experiments to observe the effect of the weight parameter, i.e. λ, which controls the relative significance of the regularization loss. Quantitative results are summarized in Table IV, where columns under "LR" show the PSNR between the compact-resolved images and the bicubic down-sampled images, and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2872876, IEEE Transactions on Image Processing

TABLE III
RECONSTRUCTION QUALITY (PSNR) OF USING DIFFERENT METHODS TO DECREASE IMAGE RESOLUTION (↓) AND THEN TO INCREASE IMAGE RESOLUTION (↑). CNN-CR^{Joint} IS A JOINTLY TRAINED CNN-CR MODEL (I.E. WITH CNN-SR). RESULTS ENCLOSED IN PARENTHESES ARE THE PSNR VALUES BETWEEN THE COMPACT-RESOLVED IMAGE AND THE BICUBIC DOWN-SAMPLED IMAGE. GAIN IS CALCULATED BETWEEN CNN-CR^{Joint} + CNN-SR AND BICUBIC DOWN-SAMPLING + CNN-SR.

	Scale	Bicubic ↓ Bicubic ↑	CNN-CR^{Joint} ↓ Bicubic ↑	Bicubic ↓ EDSR ↑ [4]	Bicubic ↓ CNN-SR ↑	CNN-CR^{Joint} ↓ CNN-SR ↑	Gain
Set5	×2	33.67	33.41	38.11	37.69	**38.88** (47.20)	1.19
	×3	30.40	29.89	34.65	33.95	**35.13** (41.64)	1.18
Set14	×2	30.01	29.81	33.92	33.13	**35.40** (43.60)	2.27
	×3	27.34	26.95	30.52	29.67	**31.33** (38.87)	1.66
BSD100	×2	29.57	29.36	32.32	32.05	**33.92** (42.49)	1.87
	×3	27.22	26.91	29.25	28.94	**30.26** (39.02)	1.32
Urban100	×2	26.56	26.42	32.93	31.01	**33.68** (40.83)	2.67
	×3	24.01	23.72	28.80	26.83	**28.81** (36.05)	1.98
Average	×2	28.32	28.14	32.83	31.77	**34.02** (41.91)	2.25
	×3	25.83	25.52	29.25	28.14	**29.78** (37.71)	1.64

TABLE IV
RESULTS OF USING DIFFERENT WEIGHTS IN TRAINING CNN-CR^{Joint}. COLUMNS UNDER "LR" SHOW THE PSNR BETWEEN THE COMPACT-RESOLVED IMAGE AND THE BICUBIC DOWN-SAMPLED IMAGE. COLUMNS UNDER "RECON" SHOW THE PSNR OF THE RECONSTRUCTED IMAGE (ACHIEVED BY CNN-SR) COMPARED WITH THE ORIGINAL IMAGE.

	Lambda = 0		Lambda = 0.1		Lambda = 0.7		Lambda = 2.0	
	LR	Recon	LR	Recon	LR	Recon	LR	Recon
Set5	19.18	39.23	40.40	38.92	47.20	38.88	49.66	38.80
Set14	16.19	35.53	36.76	35.45	43.60	35.40	46.01	35.28
BSD100	15.61	34.12	35.73	33.97	42.49	33.92	44.73	33.85
Urban100	12.75	33.66	33.33	33.68	40.83	33.68	43.30	33.62
Average	14.42	34.12	34.81	34.05	41.91	34.02	44.27	33.95

Fig. 8. The CNN-CR^{Joint} (×2) results corresponding to different weights: 1, 2, 3, 4 correspond to the weight λ equal to 0, 0.1, 0.7, 2.0.

TABLE V
RECONSTRUCTION QUALITY (PSNR) OF CNN-CR^{Joint} + CNN-SR, USING DIRECT TRAINING OR THE PROPOSED PROGRESSIVE TRAINING, RESPECTIVELY.

	Set5	Set14	BSD100	Urban100
Direct	38.63	35.07	33.65	33.38
Progressive	38.88	35.40	33.92	33.68

TABLE VI
RECONSTRUCTION QUALITY (PSNR) OF CNN-CR^{Joint} + CNN-SR, USING NON-RESIDUAL OR RESIDUAL LEARNING, RESPECTIVELY.

	Set5	Set14	BSD100	Urban100
Non-residual	38.49	35.00	33.53	33.28
Residual	38.88	35.40	33.92	33.68

TABLE VII
RECONSTRUCTION QUALITY (PSNR) OF CNN-CR^{Joint} + CNN-SR, USING MAX-POOLING (2×2) OR CONVOLUTION WITH STRIDE = 2 IN CNN-CR^{Joint}, RESPECTIVELY.

	Set5	Set14	BSD100	Urban100
Max pooling	38.79	35.02	33.60	33.50
Conv stride = 2	38.88	35.40	33.92	33.68

columns under "Recon" show the PSNR of the reconstructed images (achieved by CNN-SR) compared with the original images. Also, Fig. 8 presents the compact-resolved images corresponding to different weights. When $\lambda = 0$, no constraint

TABLE VIII
RECONSTRUCTION QUALITY (PSNR) OF CNN-CR^{Joint} + CNN-SR, USING DIFFERENT NUMBERS OF LAYERS IN CNN-CR^{Joint}, RESPECTIVELY.

	Set5	Set14	BSD100	Urban100
Depth = 5	38.41	34.73	33.27	33.06
Depth = 10	38.88	35.40	33.92	33.68
Depth = 15	38.96	35.46	33.93	33.80

is put on the compact-resolved image, thus it can be quite different from a natural image (Fig. 8 (a) is actually normalized for display). A small λ (e.g. 0.1) can work well to produce visually natural image. As λ increases, the compact-resolved image becomes more and more smooth, and the reconstruction quality decreases slightly. We choose $\lambda = 0.7$ to produce the final results in this paper, yet we have observed that different weights can be applied for different images to achieve better tradeoff between the visual quality of the compact-resolved image and the final reconstruction quality.

5) Verification Studies: To verify the network design issues and training strategies, we perform the following set of verification experiments. Table V provides comparative results of using the proposed progressive training and using direct end-to-end training from scratch. As expected, partial pre-training and joint fine-tuning work better. Table VI provides results of using residual learning or not, showing the advantage of residual learning, whose convergence is also faster than non-residual learning. Table VII provides results of using max pooling or using convolution with stride = 2 in CNN-CR^{Joint}, where the latter is better as expected. Table VIII compares the results of CNN-CR^{Joint} with different numbers of layers, which show the performance boost when increasing depth from 5-layer to 10-layer, and the marginal gain when further increasing depth. Thus, 10-layer network is used to produce the final results in this paper.

6) Interoperability between CNN-CR^{Joint} and CNN-SR: When training CNN-CR^{Joint}, we use an auto-encoder-like network that involves a CNN for image SR. It is then worth investigating whether the trained CNN-CR^{Joint} model is dependent on the specific CNN-SR used in training. For this investigation, we perform the following experiments. First, we design a new CNN-SR structure, namely CNN-SR^{Dual}, whose structure is symmetric to CNN-CR (illustration provided in the supplementary material). We consider a symmetric structure because in auto-encoders the encoding and decoding parts are usually symmetric [30]. Next, we use the previously trained CNN-CR^{Joint} model to decrease resolution and train the CNN-SR^{Dual} to increase resolution, whose results are presented in Table IX. We can observe that CNN-CR^{Joint} still outperforms bicubic down-sampling significantly, as the preserved information introduced by CNN-CR^{Joint} can be well exploited by learning-based SR methods. In addition, we replace the CNN-SR in Fig. 2 with the CNN-SR^{Dual}, and retrain CNN-CR. The trained model is denoted as CNN-CR^{Joint2}. We then test the combination of CNN-CR^{Joint2} and CNN-SR^{Dual}, whose results are also presented in Table IX. There is only little difference between CNN-CR^{Joint} and CNN-CR^{Joint2}, although the latter is jointly trained with CNN-SR^{Dual} but the former is not. Thus, we can claim that the trained CNN-CR^{Joint} model does not depend on a specific CNN-SR to work well. Indeed, different SR networks can influence the reconstruction quality significantly, as can be observed by comparing CNN-CR^{Joint} + CNN-SR^{Dual} (Table IX) and CNN-CR^{Joint} + CNN-SR (Table III). Note that CNN-SR is deeper than CNN-SR^{Dual}.

7) Results of Different CR Ratios: We also conduct experiments by setting the CR ratio to 3, and the related results are shown in Table III to be compared with the results of CR ratio equal to 2. The gain achieved by CNN-CR^{Joint} compared with bicubic down-sampling is still significant, and reaches on average 1.64 dB in terms of reconstruction quality over the four testing datasets. As CR ratio increases, the gain provided by CNN-CR^{Joint} becomes less, since larger CR ratio inevitably incurs more loss of information, which is difficult to recover. In addition, as CR ratio increases, the PSNR between compact-resolved image and bicubic down-sampled image also decreases, which is reasonable since CNN-CR^{Joint} needs to add more details into the compact-resolved image.

TABLE IX
RECONSTRUCTION QUALITY (PSNR) OF USING DIFFERENT METHODS TO DECREASE IMAGE RESOLUTION ($\downarrow \times 2$) AND THEN TO INCREASE IMAGE RESOLUTION ($\uparrow \times 2$). CNN-CR^{Joint2} IS A JOINTLY TRAINED CNN-CR MODEL TOGETHER WITH CNN-SR^{dual}, WHICH IS A DIFFERENT CNN FOR SR AND WHOSE STRUCTURE IS SYMMETRIC TO CNN-CR.

	Bicubic $\downarrow$ CNN-SR^{Dual} $\uparrow$	CNN-CR^{Joint} $\downarrow$ CNN-SR^{Dual} $\uparrow$	CNN-CR^{Joint2} $\downarrow$ CNN-SR^{Dual} $\uparrow$
Set5	37.49	38.57	38.57
Set14	32.90	34.77	34.79
BSD100	31.91	33.63	33.58
Urban100	30.53	32.78	32.77

TABLE X
RECONSTRUCTION QUALITY (PSNR) OF LIU'S METHOD [15] + CNN-SR AND CNN-CR^{Joint} + CNN-SR.

	Liu's method [15] $\downarrow$ CNN-SR $\uparrow$	CNN-CR^{Joint} $\downarrow$ CNN-SR $\uparrow$
Set5	34.41	38.88
Set14	31.26	35.40
BSD100	30.63	33.92
Urban100	29.62	33.68

C. Comparison with Perceptual Quality-Oriented Down-Scaling

We experimentally compare CNN-CR with a state-of-the-art image down-scaling method proposed by Liu *et al.* [15]. As mentioned above, such image down-scaling methods concern the perceptual quality of the down-scaled images, but ignore the issue how much information is preserved in the down-scaled images. Accordingly, when comparing the visual quality of the down-scaled images by CNN-CR and by Liu's method (provided in the supplementary material), the latter is usually better as it hallucinates more details in the down-scaled images. But when we super-resolve the down-scaled images (using CNN-SR trained separately) and compare the reconstruction quality, CNN-CR performs much better than Liu's method, as shown in Table X.

D. Results of the Application in Image Retargeting

1) Settings: We consider the case of down-sizing an image by a factor of 2 without changing aspect ratio, which is a seemingly easy case in retargeting. Our trained CNN-CR models, including CNN-CR^{Sep} and CNN-CR^{Joint}, can be directly applied herein. For comparative study, we use the benchmark

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2872876, IEEE Transactions on Image Processing

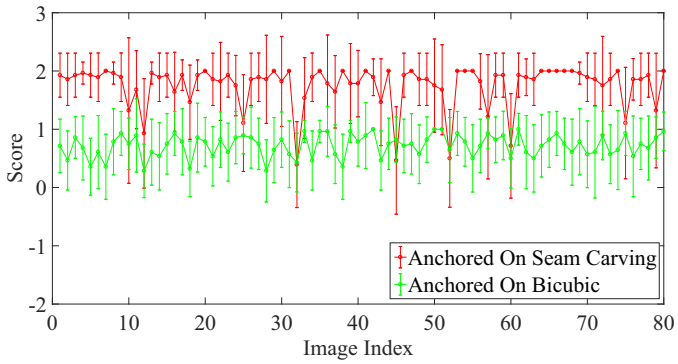

Fig. 9. Mean and standard deviation of subjective scores on the 80 test images, where CNN-CR^{Sep} is compared to seam carving and bicubic down-sampling, respectively.

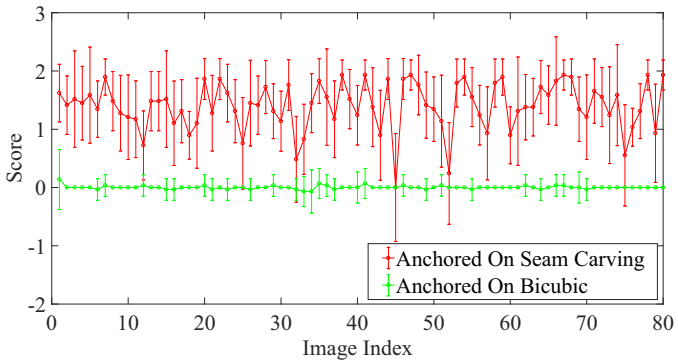

Fig. 10. Mean and standard deviation of subjective scores on the 80 test images, where CNN-CR^{Joint} is compared to seam carving and bicubic down-sampling, respectively.

dataset proposed in [17], which contains 80 natural images. We compare our CNN-CR models with a representative retargeting method known as *seam carving* [16], and also with the simple bicubic down-sampling. It is known that objective evaluation is not well-defined in the retargeting task, and thus we perform subjective evaluation as suggested by [17]. Specifically, we adopt the stimulus-comparison method [48] to perform subjective experiments. 30 subjects, all university students, participate in the experiments. Every subject is asked to observe a series of pairs of retargeted images and give a score for each pair. Score has 5 discrete levels: -2, -1, 0, 1, 2, standing for better, slightly better, indistinguishable, slightly worse, and worse, respectively. Our method is paired with seam carving and bicubic, respectively. After experiments, we calculate the mean and standard deviation of subjective scores per image, which are shown in Figs. 9 and 10.

From Fig. 9 and Fig. 10, it can be observed that both CNN-CR^{Sep} and CNN-CR^{Joint} performs better than the seam carving method. To understand the reason, we inspect the retargeted images as shown in Fig. 11 for example. Since the seam carving method was originally designed for retargeting with changing aspect ratio, it intentionally changes the layout of the image, but at the same time may incur structural distortion. Thus, it is not wise to apply such methods for retargeting if not to change the aspect ratio. When compared with bicubic down-sampling, CNN-CR^{Sep} achieves better subjective quality, while CNN-CR^{Joint} leads to comparable quality, for most of the test images. Thus, we recommend the usage of CNN-CR^{Sep} for retargeting.

E. Results of the Application in Image Compression

1) Settings: We directly apply the jointly learned CR model CNN-CR^{Joint}, as achieved in Section VI-B, into image compression. As for the CNN-SR, we intentionally use another network structure and retrain CNN-SR, as mentioned in Section V-B. Here, the used CNN-SR structure is the same as that in [41], because it is light-weight but achieves satisfactory results. Our intention is to demonstrate that our CNN-CR^{Joint} does not depend on a specific SR network. The down- and up-sampling ratio is 2 in this subsection. Other sampling ratios will be studied in our future work.

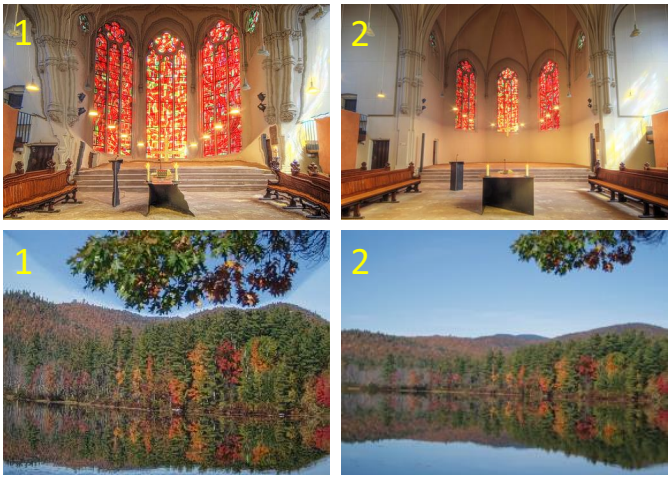

Fig. 11. Example results of image retargeting. Left: seam carving results; Right: CNN-CR^{Joint} results (more results can be found in the supplementary material).

Our frame-level and block-level schemes are implemented upon the HEVC reference software–HM version 12.1[2], using only its intra coding tools. Caffe [49] is integrated into HM for the CNN implementation. Considering down/up-sampling-based coding is a useful tool especially at low bit rates, the QP of HEVC is set to {32, 37, 42, 47}. For our frame-level scheme, we manually adjust QP for each test sequence to make the resulting bitrate as aligned to HEVC as possible. For our block-level scheme, we simply set a constant delta QP (-6) for low-resolution coding [41]. For evaluation, we adopt BD-rate [50], where for the quality metric we use both PSNR and structural similarity (SSIM) [24], as the latter is believed to be more consistent with subjective quality. The test sequences include 20 video sequences that can be divided into Classes A, B, C, D, E [51], and 5 sequences at 4K (3840×2160) resolution from the SJTU dataset [52].

2) Frame-Level Results: The BD-rate results are summarized in Table XII. We compare our scheme with HEVC that directly compresses the image at its original resolution. Results show that our scheme achieves on average 1.3% BD-rate reduction for the HEVC test sequences and 14.9% BD-rate

[2]https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-12.1/

TABLE XI
COMPUTATIONAL TIME OF OUR BLOCK-LEVEL SCHEME COMPARED TO THE HEVC ANCHOR (HEVC IS 1) USING CPU

Class	Encoding Time	Decoding Time
Class A	14.64	540.08
Class B	15.47	395.92
Class C	14.11	397.20
Class D	14.43	182.00
Class E	16.22	534.00
Class UHD	16.26	633.41
Average	15.68	540.70

reduction for the UHD test sequences, when using PSNR as quality metric. When using SSIM, the coding gains are even higher, i.e. 18.6% and 18.0% BD-rate reduction for HEVC and UHD test sequences, respectively. Such results show that to some extent, the proposed method is more friendly to the subjective quality at low bit-rates.

We also compare with a down-/up-sampling-based scheme without CNN-CR^{Joint}, i.e. using simple down-sampling [42] and CNN-SR for up-sampling. Results show that, our proposed method brings on average 7.0% and 3.1% BD-rate reduction for HEVC and UHD test sequences, respectively. It demonstrates that, due to preserving more information, CNN-CR^{Joint} is favorable than simple down-sampling for low-bit-rate image compression.

Fig. 12 shows the rate-distortion (R-D) curves of some typical sequences. It can be observed that, at extremely low bit-rates, the frame-level down- and up-sampling scheme always surpasses the HEVC. But as the bit-rate increases, the gain becomes less. After a certain (switching) bit-rate, HEVC is better than the frame-level scheme. It demonstrates that down-/up-sampling-based strategy is useful for low bit-rates. It is also worth noting that the switching bit-rates for different sequences are diverse. Such R-D curves can interpret why, in Table XII, the frame-level scheme incurs loss for several test sequences. Block-level adaptive scheme helps resolve the loss, as noted below.

3) Block-Level Results: The BD-rate results are also summarized in Table XII, where we compare our block-level adaptive down- and up-sampling scheme with HEVC and with the scheme in [41].

Compared to HEVC, our scheme improves the coding efficiency significantly, leading to on average 6.9% BD-rate reduction for the HEVC test sequences. For the UHD test sequences, our scheme achieves even higher coding gain, i.e. on average 10.4% BD-rate reduction.

Compared to the scheme in [41], our block-level scheme also leads to obvious BD-rate reduction. While the scheme in [41] is also block-level adaptive down- and up-sampling, it uses merely the simple down-sampling filter while adopting CNN for up-sampling. In this paper, our scheme further introduces CNN-CR^{Joint} as an alternative for block down-sampling, and still adopts CNN for up-sampling. Thus, the coding gain comes from the benefit provided by CNN-CR^{Joint} that outperforms the simple down-sampling filter.

Typical R-D curves of the block-level scheme are also shown in Fig. 12. Compared to the frame-level scheme, the block-level scheme performs similarly at extremely low bit-rates, but better at higher bit-rates.

To evaluate the performance of the block-level scheme at higher bit-rates, we also test the QPs 22 and 27 in addition to the QPs {32, 37, 42, 47}. All the BD-rate results are summarized in Table XV. When QP increases, the BD-rate reduction becomes more and more significant, which again demonstrates the advantage of down-/up-sampling-based coding at low bit-rates. Moreover, compared to [41], the scheme in this paper consistently performs better at all bit-rates.

As the benefit of block-level scheme is to provide the adaptability, we also analyze the proportion of blocks that are compressed with different modes. Some symbols are defined as shown in Table XIII, and the hitting ratios are calculated as follows,

$$P_{LR} = \frac{\#C_{LR}}{\#C_{Total}}, P_{CNN-CR} = \frac{\#C_{CNN-CR}}{\#C_{LR}},$$

where # denotes counting the number. Table XIV presents the calculated hitting ratios. P_{LR} is on average 74%, 72%, 54%, 49%, 74%, 87% for Classes A, B, C, D, E and UHD, respectively. Considering the resolutions of these videos, it is obvious that the hitting ratio becomes higher with the increase of image resolution. Among the blocks choosing low-resolution coding, around half of them choose CNN-CR for down-sampling. It shows that both the simple down-sampling filter and the proposed CNN-CR are useful in image compression, but for different video content.

Compared to the computational units in HEVC, CNN-based methods have the drawback of much higher computational complexity [41], [53]. Currently, our implementation is not optimized for computational efficiency, and we conduct experiments using CPU. The computational time comparison is presented in Table XI, which shows much increased encoding/decoding time of our scheme. Note that the increased decoding time varies much across different classes, since it depends on the amount of CTUs that chose CNN-based up-sampling. Moreover, if using GPU, we anticipate the computational time can be reduced significantly.

4) Comparison between Frame- and Block-Level Results: As shown in Table XII, for most of the test sequences, block-level scheme outperforms frame-level scheme in terms of BD-rate reduction measured by PSNR; however, for high-resolution test sequences, especially the UHD sequences, frame-level scheme is better. This can be interpreted by observing Table XIV, where P_{LR} reaches as high as 87% for UHD sequences, i.e., a majority of blocks actually chose the low-resolution coding mode, so it suffices to down-sample the entire frame. Note that the block-level scheme needs to encode two flags for each block, one flag signaling low-resolution or full-resolution coding, and the other flag signaling which down-sampling, such overhead is omitted in the frame-level scheme that then saves more bits. On the contrary, for Class D sequences that natively have lower resolution, frame-level scheme performs worse, while block-level scheme providing the adaptability is better.

In addition, with observation of the R-D curves shown in Fig. 12, when the bit-rate increases, the frame-level scheme

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2018.2872876, IEEE Transactions on Image Processing

TABLE XII
BD-RATE RESULTS OF ALL TEST SEQUENCES (DS STANDS FOR DOWN-SAMPLING)

Class	Sequence	Frame-level scheme				Block-level adaptive scheme			
		Anchored on HEVC		Anchored on simple DS		Anchored on HEVC		Anchored on simple DS [41]	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Class A	Traffic	–12.4%	–17.1%	–3.5%	–4.8%	–11.8%	–14.8%	–1.6%	–2.1%
	PeopleOnStreet	–9.5%	–14.4%	–4.5%	–4.1%	–11.7%	–14.5%	–2.1%	–1.7%
	Nebuta	13.0%	–8.4%	–1.4%	–3.7%	–1.8%	–4.8%	0.2%	–0.3%
	SteamLocomotive	2.6%	–13.1%	–1.2%	–6.1%	–2.3%	–7.6%	–0.6%	–1.4%
Class B	Kimono	–13.0%	–14.5%	–3.7%	–3.7%	–9.0%	–10.8%	–1.4%	–1.3%
	ParkScene	–8.8%	–15.5%	–2.9%	–5.4%	–8.3%	–13.0%	–1.3%	–1.5%
	Cactus	–7.1%	–20.0%	–4.9%	–6.4%	–8.5%	–12.9%	–1.6%	–2.6%
	BQTerrace	6.2%	–19.7%	–9.6%	–13.9%	–4.8%	–11.3%	–1.2%	–1.9%
	BasketballDrive	7.0%	–20.8%	–13.1%	–9.2%	–8.0%	–13.0%	–1.7%	–2.8%
Class C	BasketballDrill	–10.7%	–24.8%	–12.2%	–5.4%	–7.5%	–10.5%	–2.7%	–1.8%
	BQMall	17.2%	–23.4%	–12.1%	–11.1%	–3.9%	–8.0%	–0.7%	–1.1%
	PartyScene	8.9%	–26.2%	–10.6%	–15.6%	–1.9%	–6.5%	–0.6%	–2.9%
	RaceHorsesC	–6.8%	–18.5%	–5.3%	–7.1%	–8.2%	–13.0%	–1.4%	–2.3%
Class D	BasketballPass	6.5%	–22.0%	–12.4%	–7.9%	–4.6%	–9.1%	–2.9%	–4.7%
	BQSquare	7.8%	–26.6%	–12.8%	–8.0%	–1.8%	–3.6%	–0.6%	–1.4%
	BlowingBubbles	3.8%	–18.8%	–8.0%	–7.2%	–4.2%	–8.9%	–1.2%	–4.2%
	RaceHorses	–13.5%	–17.7%	–4.7%	–5.6%	–13.0%	–18.0%	–2.6%	–3.2%
Class E	FourPeople	–3.9%	–18.3%	–5.9%	–4.9%	–9.1%	–14.5%	–1.6%	–3.7%
	Johnny	–8.1%	–12.9%	–6.0%	–7.9%	–10.2%	-11.8%	–1.9%	–0.7%
	KristenAndSara	–0.7%	–20.0%	–6.8%	–4.3%	–7.9%	–14.0%	–0.9%	–1.8%
Class UHD	Fountains	–5.4%	–14.1%	–3.4%	–5.6%	–4.9%	–8.7%	–1.0%	–1.3%
	Runners	–13.2%	–14.9%	–1.5%	–3.4%	–11.9%	–13.1%	–0.7%	–0.5%
	Rushhour	–14.6%	–16.5%	–3.8%	–3.5%	–10.1%	–11.5%	–1.6%	–1.2%
	TrafficFlow	–16.0%	–16.6%	–2.8%	–3.8%	–14.7%	–14.8%	–2.6%	–2.6%
	CampfireParty	–25.4%	–27.6%	–3.8%	–4.6%	–10.4%	–11.9%	–1.9%	–2.1%
Average of Classes A-E		–1.1%	–18.6%	–7.1%	–7.1%	–6.9%	–11.0%	–1.4%	–2.2%
Average of Class UHD		–14.9%	–18.0%	–3.1%	–4.2%	–10.4%	–12.0%	–1.6%	–1.5%

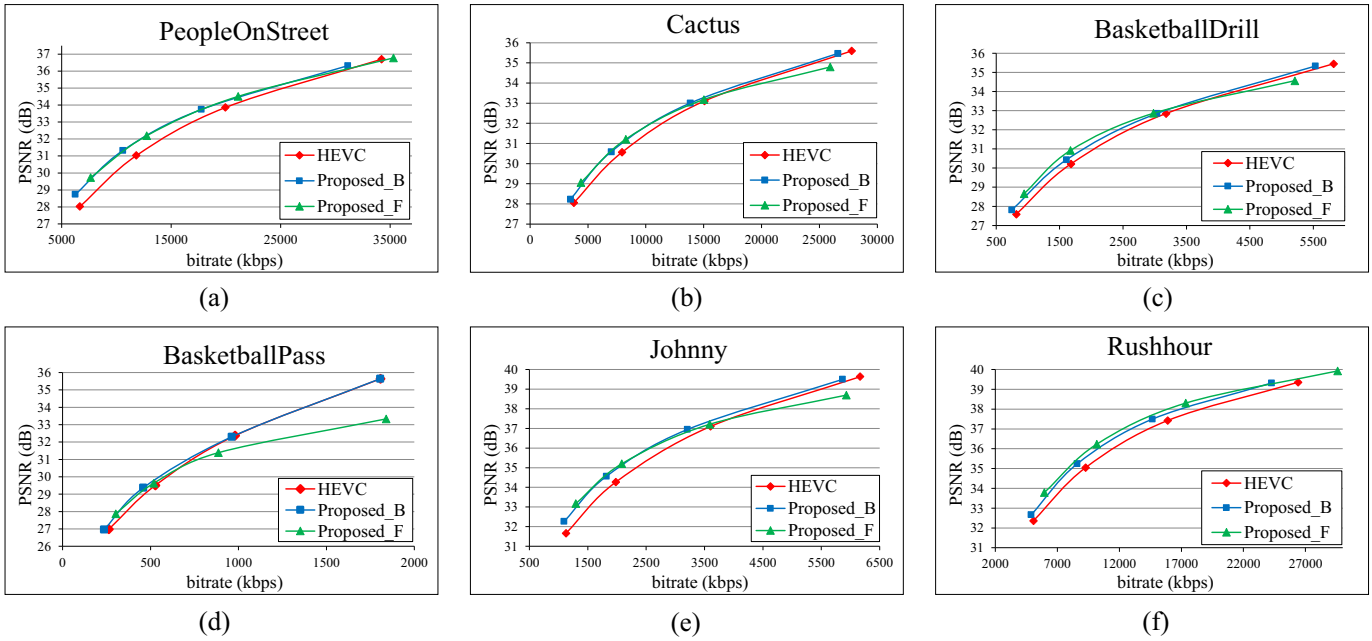

Fig. 12. Rate-distortion (R-D) curves of several typical sequences: (a) PeopleOnStreet, (b) Cactus, (c) BasketballDrill, (d) BasketballPass, (e) Johnny, and (f) Rushhour. "Proposed_B" stands for the block-level adaptive scheme, "Proposed_F" stands for the frame-level scheme, both based on CNN-CR^{Joint}.

TABLE XIII
SYMBOLS FOR CTUS THAT CHOOSE DIFFERENT MODES IN THE BLOCK-LEVEL ADAPTIVE SCHEME

Symbol	Remark
C_{Total}	All CTUs in a frame
C_{LR}	CTUs selecting the mode of low-resolution coding
C_{CNN-CR}	Low-resolution coded CTUs, whose luma component is down-sampled using CNN-CR^{Joint}

TABLE XIV
HITTING RATIO RESULTS ON DIFFERENT CLASSES OF TEST SEQUENCES OF OUR BLOCK-LEVEL ADAPTIVE SCHEME

Class	Class A	Class B	Class C	Class D	Class E	Class UHD
P_{LR}	74%	72%	54%	49%	74%	87%
P_{CNN-CR}	40%	51%	54%	49%	53%	48%

performs worse than HEVC, but block-level scheme is still comparable to HEVC. This is attributed to the built-in switching of full-resolution and low-resolution coding in the block-level scheme.

Furthermore, in Table XII, it can be observed that in terms of BD-rate reduction measured by SSIM, the frame-level scheme usually performs much better than the block-level scheme. The underlying reason is that, the block-level scheme adopts rate-distortion cost to decide whether using full-resolution or using low-resolution coding, where distortion is measured by mean-squared-error. In other words, the decision is optimal if measured by PSNR, but may not be optimal if measured by other metrics like SSIM. It further reveals that, at low bit-rates, down-sampling-coding may improve the visual quality more than improve the PSNR.

VII. CONCLUSION

We have proposed a learning approach for image compact-resolution using convolutional neural network (CNN-CR). The problem of image CR is formulated as to jointly minimize the reconstruction loss and the regularization loss. CNN-CR can be trained either separately, or jointly with a CNN for image SR. We investigate network structures and training strategies used for CNN-CR. Our experimental results show that the proposed CNN-CR outperforms simple down-sampling with a noticeable margin in terms of the reconstruction quality. The compact-resolved images look visually pleasing thanks to the proposed regularization loss. We also investigate applications of CNN-CR in low-bit-rate image compression and image retargeting, and results demonstrate the effectiveness of our method.

Regarding the problem of image CR, one most important open problem is how to evaluate the quality of the compact-resolved images either objectively or subjectively. We plan to investigate this problem in the future. In addition, we plan to extend image CR to video CR, and to explore other applications of image CR, such as converting from YUV 4:4:4 format to YUV 4:2:0 format.

REFERENCES

[1] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[2] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*, pp. 184–199, Springer, 2014.

[3] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1646–1654, 2016.

[4] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

[5] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883, 2016.

[6] F. Jiang, W. Tao, S. Liu, J. Ren, X. Guo, and D. Zhao, "An end-to-end compression framework based on convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[7] W.-S. Lu and A.-M. Sevcenco, "Design of optimal decimation and interpolation filters for low bit-rate image coding," in *Circuits and Systems, 2006. APCCAS 2006. IEEE Asia Pacific Conference on*, pp. 378–381, IEEE, 2006.

[8] Y. Tsaig, M. Elad, P. Milanfar, and G. H. Golub, "Variable projection for near-optimal filtering in low bit-rate block coders," *IEEE transactions on circuits and systems for video technology*, vol. 15, no. 1, pp. 154–160, 2005.

[9] S. Daly, "47.3: Analysis of subtriad addressing algorithms by visual system models," in *SID Symposium Digest of Technical Papers*, vol. 32, pp. 1200–1203, Wiley Online Library, 2001.

[10] S. J. Daly and R. R. K. Kovvuri, "Methods and systems for improving display resolution in images using sub-pixel sampling and visual error filtering," Aug. 19 2003. US Patent 6,608,632.

[11] L. Fang and O. C. Au, "Subpixel-based image down-sampling with min-max directional error for stripe display," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 2, pp. 240–251, 2011.

[12] L. Fang, O. C. Au, K. Tang, X. Wen, and H. Wang, "Novel 2-d mmse subpixel-based image down-sampling," *IEEE transactions on circuits and systems for video technology*, vol. 22, no. 5, pp. 740–753, 2012.

[13] J. Kopf, A. Shamir, and P. Peers, "Content-adaptive image downscaling," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, p. 173, 2013.

[14] A. C. Öztireli and M. Gross, "Perceptually based downscaling of images," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, p. 77, 2015.

[15] J. Liu, S. He, and R. W. Lau, "$l\_\{0\}$-regularized image downscaling," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1076–1085, 2018.

[16] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," in *ACM Transactions on graphics (TOG)*, vol. 26, p. 10, ACM, 2007.

[17] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, "A comparative study of image retargeting," in *ACM transactions on graphics (TOG)*, vol. 29, p. 160, ACM, 2010.

[18] V. Setlur, S. Takagi, R. Raskar, M. Gleicher, and B. Gooch, "Automatic image retargeting," in *Proceedings of the 4th international conference on Mobile and ubiquitous multimedia*, pp. 59–68, ACM, 2005.

[19] D. Vaquero, M. Turk, K. Pulli, M. Tico, and N. Gelfand, "A survey of image retargeting techniques," in *Proc. SPIE*, vol. 7798, p. 779814, 2010.

[20] Y. Amano, "A flat-panel tv display system in monochrome and color," *IEEE Transactions on Electron Devices*, vol. 22, no. 1, pp. 1–7, 1975.

[21] T. L. Benzschawel and W. E. Howard, "Method of and apparatus for displaying a multicolor image," Aug. 23 1994. US Patent 5,341,153.

[22] L.-M. Chen and S. Hasegawa, "Influence of pixel-structure noise on image resolution and color for matrix display devices," *Journal of the Society for Information Display*, vol. 1, no. 1, pp. 103–110, 1993.

[23] M. A. Klompenhouwer and G. Haan, "Subpixel image scaling for color-matrix displays," *Journal of the Society for Information Display*, vol. 11, no. 1, pp. 99–108, 2003.

[24] Z. 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.

[25] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang, B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, *et al.*, "Ntire 2017 challenge on single image super-resolution: Methods and results," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pp. 1110–1121, IEEE, 2017.

TABLE XV
BD-RATE RESULTS AT DIFFERENT QPS OF OUR BLOCK-LEVEL ADAPTIVE SCHEME

Class	BD-Rate (QP 22–37)				BD-Rate (QP 27–42)				BD-Rate (QP 32–47)			
	Anchored on HEVC		Anchored on [41]		Anchored on HEVC		Anchored on [41]		Anchored on HEVC		Anchored on [41]	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Class A	–0.9%	–3.0%	–0.5%	–1.2%	–3.2%	–6.7%	–0.9%	–1.4%	–6.9%	–10.4%	–1.0%	–1.4%
Class B	–2.0%	–4.2%	–0.6%	–1.3%	–4.6%	–8.6%	–1.1%	–1.9%	–7.7%	–12.2%	–1.4%	–2.0%
Class C	–0.4%	–1.1%	–0.2%	–0.6%	–2.1%	–4.8%	–0.6%	–1.4%	–5.4%	–9.5%	–1.4%	–2.0%
Class D	–0.5%	–2.3%	–0.2%	–0.8%	–2.3%	–6.6%	–0.9%	–2.3%	–5.9%	–10.2%	–1.8%	–2.9%
Class E	–1.7%	–4.7%	–0.7%	–1.6%	–4.9%	–9.8%	–1.2%	–2.2%	–9.1%	–13.4%	–1.5%	–2.1%
Avg. Classes A-E	–1.1%	–2.9%	–0.4%	–1.1%	–3.4%	–7.1%	–0.9%	–1.9%	–6.9%	–11.0%	–1.4%	–2.2%
Class UHD	–3.0%	–5.4%	–0.9%	–1.4%	–6.8%	–9.3%	–1.3%	–1.6%	–10.4%	–12.0%	–1.6%	–1.5%

[26] Y. Bengio *et al.*, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[29] C. Ledig, L. Theis, F. Huszár, 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," *arXiv preprint arXiv:1609.04802*, 2016.

[30] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[31] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, pp. 2017–2025, 2015.

[32] ITU, "Parameter values for ultrahigh definition television systems for production and international program exchange," *ITU Recommendation BT.2020*, 2012.

[33] A. M. Bruckstein, M. Elad, and R. Kimmel, "Down-scaling for better transform compression," *IEEE Transactions on Image Processing*, vol. 12, no. 9, pp. 1132–1144, 2003.

[34] K. Takahashi, T. Naemura, and M. Tanaka, "Rate-distortion analysis of super-resolution image/video decoding," in *IEEE International Conference on Image Processing*, pp. 1629–1632, IEEE, 2011.

[35] R. Molina, A. Katsaggelos, L. Alvarez, and J. Mateos, "Toward a new video compression scheme using super-resolution," in *Electronic Imaging 2006*, pp. 607706–607706, International Society for Optics and Photonics, 2006.

[36] D. Barreto, L. Alvarez, R. Molina, A. K. Katsaggelos, and G. Callico, "Region-based super-resolution for compression," *Multidimensional Systems and Signal Processing*, vol. 18, no. 2-3, pp. 59–81, 2007.

[37] M. Shen, P. Xue, and C. Wang, "Down-sampling based video coding using super-resolution technique," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 6, pp. 755–765, 2011.

[38] J. Wu, Y. Xing, G. Shi, and L. Jiao, "Image compression with downsampling and overlapped transform at low bit rates," in *IEEE International Conference on Image Processing*, pp. 29–32, IEEE, 2009.

[39] W. Lin and L. Dong, "Adaptive downsampling to improve image compression at low bit rates," *IEEE Transactions on Image Processing*, vol. 15, no. 9, pp. 2513–2521, 2006.

[40] V.-A. Nguyen, Y.-P. Tan, and W. Lin, "Adaptive downsampling/upsampling for better video compression at low bit rate," in *IEEE International Symposium on Circuits and Systems*, pp. 1624–1627, IEEE, 2008.

[41] Y. Li, D. Liu, H. Li, L. Li, F. Wu, H. Zhang, and H. Yang, "Convolutional neural network-based block up-sampling for intra frame coding," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[42] T. Davies. "Resolution switching for coding efficiency and resilience," Document JCTVC-F158, Turin, Italy, July 2011.

[43] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.

[44] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*, pp. 711–730, Springer, 2010.

[45] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5197–5206, 2015.

[46] C.-Y. Yang and M.-H. Yang, "Fast direct super-resolution by simple functions," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 561–568, 2013.

[47] D. Kinga and J. B. Adam, "A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[48] B. Series, "Subjective methods for the assessment of stereoscopic 3dtv systems," 2015.

[49] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 675–678, ACM, 2014.

[50] G. Bjontegaard, "Calcuation of average PSNR differences between RD-curves," *Doc. VCEG-M33 ITU-T Q6/16, Austin, TX, USA, 2-4 April 2001*, 2001.

[51] F. Bossen *et al.*, "Common test conditions and software reference configurations," *Joint Collaborative Team on Video Coding (JCT-VC), JCTVC-F900*, 2011.

[52] L. Song, X. Tang, W. Zhang, X. Yang, and P. Xia, "The SJTU 4K video sequence dataset," in *QoMEX*, pp. 34–35, 2013.

[53] J. Li, B. Li, J. Xu, R. Xiong, and W. Gao, "Fully connected network-based intra prediction for image coding," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3236–3247, 2018.