

Lossless Image Compression through Super-Resolution

Sheng Cao, Chao-Yuan Wu, Philipp Krähenbühl

The University of Texas at Austin

Abstract. We introduce a simple and efficient lossless image compression algorithm. We store a low resolution version of an image as raw pixels, followed by several iterations of lossless super-resolution. For lossless super-resolution, we predict the probability of a high-resolution image, conditioned on the low-resolution input, and use entropy coding to compress this super-resolution operator. Super-Resolution based Compression (SReC) is able to achieve state-of-the-art compression rates with practical runtimes on large datasets. Code is available online.¹

1 Introduction

Mankind captures and shares a collective trillion of new photos annually [32]. Images capture all aspects of our beautifully complex and diverse visual world. Yet, not every arrangement of pixel color values forms a real image. Most are illegible noise. This is the main insight behind lossless image compression. In fact, the Shannon source coding theorem directly links the likelihood of a group of pixels to be a real image, and our ability to compress that image [40]. The main challenge is designing an effective probabilistic model of pixel values.

In this paper, we propose lossless image compression through *super-resolution* (SR). Unlike standard super-resolution, which predicts one single output image, we predict a *distribution* over all possible super-resolved images. Each pixel in a low-resolution image induces an autoregressive distribution over four high-resolution output pixels. This distribution is then entropy-coded using arithmetic coding (AC), yielding a losslessly compressed super-resolution operator. Our overall compression algorithm stores a low-resolution version of the image as raw pixels, and then applies three iterations of losslessly-compressing super-resolution operator, as shown in Figure 1. We train the losslessly compressed super-resolution operator to maximize the log-likelihood of a high-resolution image, conditioned on its downsampled version on standard image datasets.

Compression through super-resolution shares components with existing deep lossless image compression methods [18, 33], yet enjoys several additional advantages. Our neural network is lightweight and efficient. Each set of four output pixels is independent of all other outputs at the same level; hence super-resolution is easily performed in parallel. Furthermore, we show that simple constraints imposed by the super-resolution process, allow 25% of the pixels to be reconstructed

¹ <https://github.com/caoscott/SReC>

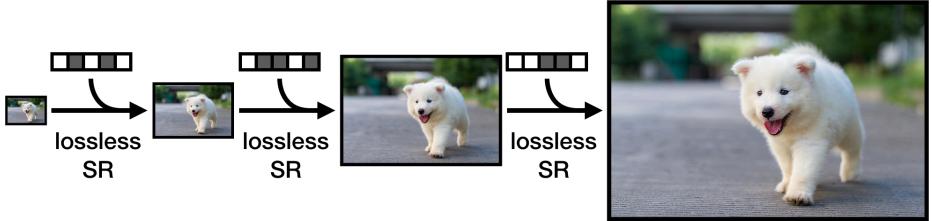


Fig. 1. Model Overview. We propose lossless image compression through super resolution (SR). Our method first encodes a low-resolution image efficiently, and then leverages SR models to efficiently entropy-code the high-resolution images.

“for free”. Finally, we show how the SR setting strictly limits the range of the probability distribution, further reducing the bitrate used in entropy coding.

We evaluate our algorithm on the ImageNet64 [8, 11] and Open Images datasets [25]. Our experiments show that our super-resolution-based image compression algorithm outperforms state-of-the-art lossless image compression algorithms across a varying set of image resolutions and data sources. The runtime of our method is comparable to the fastest prior work, and our models are as small as the most compact prior lossless deep image compression methods.

2 Related Work

Most lossless image compression algorithms rely on entropy coding of commonly repeating image patterns. The two signals most commonly exploited are inter image similarities and natural image statistics.

Hand-designed codecs encompass some of the most popular lossless compression methods. PNG [1] compresses the raw bits of a color image using the DEFLATE algorithm [12]. It exploits bit-level repetitions in the image but often fails to capture large structural similarities or common image statistics. JPEG2000 [42] builds its lossless compression in a wavelet transform, which captures some local image statistics. WebP [2] combines multiple image transformations before entropy coding. Currently, the best performing hand-designed codec is FLIF [43]. It builds on the MANIAC entropy coding algorithm and captures repeating local image patterns in entropy coding. Hand-designed codecs efficiently exploit the local structure of image formation but only capture simple image statistics that are hand-specified.

Entropy coding, such as arithmetic coding (AC) [54] or asymmetric numeral systems (ANS) [14], is able to convert generative models of image formation into a compression algorithm, as long as the model provides a probability estimate of the current image. However, not all generative models are equally efficient.

Compression with Autoregressive Models. Autoregressive models predict a distribution over natural images as a distribution over color values, conditioned

on previously predicted colors. For example, Van Oord *et al.*'s PixelCNN [52] predicts probabilities for subpixels² in a raster scan and RGB order. Each subpixel probability is conditioned on previously seen subpixels. Salimans *et al.* [39] uses a discrete logistic mixture to model the joint distribution of a pixel. Reed *et al.* [35] speeds up these methods by predicting probabilities of multiple pixels at once. Kolesnikov and Lampert [24] use grayscale or downscaled versions of the image as auxiliary variables to PixelCNN to improve model performance. PixelCNN and their variants generally perform well in terms of log-likelihood, but are impractical for compression due to the long runtime. For images of size $W \times H$, the original PixelCNN requires $\mathcal{O}(W+H)$ distinct network evaluations, each predicting a diagonal slice of the image. Reed *et al.* [35] improves this to $\mathcal{O}(\log(W+H))$ using a hierarchical encoding scheme. However, they use shallow PixelCNNs [39] to model dependency between blocks of pixels. In practice, this is still too slow for lossless compression; see [33] for detailed analysis. Our model uses a similar hierarchical structure with a few important differences: We use a simpler factorization of the pixel-wise probabilities, allow different hierarchical level to share a feature embedding, and use a more efficient architecture. Our final compression model is $\sim 60 \times$ faster than Reed *et al.* [35] when used for image compression and comparable to the fastest learned image compression techniques [33].

Latent Vector Models. An efficient alternative to condition unseen pixels on previously seen pixels is through a latent vector. Mentzer *et al.* [33] encode an image x into smaller latent vectors z_1, \dots, z_3 , and entropy code x using $P(x | z_1, \dots, z_3)$ estimated by a network. The latent vectors are discretized such that they are also efficient to store. Integer Discrete Flow (IDF) [18] uses a flow-based deep generative model [36] to invertibly transform the input image into a latent vector. It factorizes probabilities of the latent vectors and compress the latent vectors. During decompression, the latent vectors are decompressed and inverted to obtain the image. IDF has high performance on ImageNet32 and ImageNet64 [8], as it is able to optimize factorized log-likelihood directly. However, it struggles to learn higher resolution models. It learns a discretely parametrized flow, which leads to large approximation errors when many layers are stacked. In addition, current implementation of flow-based methods are relatively inefficient. Our method is about $55 \times$ faster on high-resolution images.

Dataset Compression. Bits-back methods [16] are a family of methods that compress continuous latent vectors at fine discretization levels instead of discrete latent vectors. Bits Back with ANS [49], Bit-Swap [23], and Hierarchical Latent Lossless Compression [50] build on variational auto-encoders [22], while Local Bits-Back [17] is a flow-based method. Bits-back methods yield the best performance on ImageNet32 and ImageNet64 [8] when compressing the entire test set into a *single* vector. However, they are designed as dataset compression algorithms rather than single image compression algorithms. They are currently unable to compress single images efficiently. For example, Local Bits-Back (LBB) [17] requires an initial bit-buffer of 52 bits per subpixel (bpsp) at a dataset

² 1 pixel = 3 subpixels: R, G, and B.

compression rate of 3.63 bpsp. This initial investment is amortized when compressing a large dataset of images. When used for single image compression, LBB would compress a single images at 55 bpsp, which is much worse than the uncompressed BMP at 8 bpsp. Our algorithm on the other hand does not rely on a bit-buffer, but instead entropy codes each image independently at a bitrate close to the best bits-back approaches.

Lossy Compression. Lossy image/video compression, on the other hand, allows for some distortion in the decompressed data in exchange for reduced storage size. Recent deep learning based methods typically directly predict the decompressed output [4, 5, 19, 27, 30, 37, 38, 46, 47, 55], as opposed to predicting a distribution of outputs, as in a lossless compression method.

Super-Resolution. Super-resolution (SR) is a task to construct a high-resolution image given a low-resolution image [6, 10, 13, 20, 26, 28, 29, 34, 44, 45, 48, 53, 56]. Recent works have advanced the state-of-the-art performance with the advances in CNN architecture [13], image generation [26], and the likelihood based methods [10]. Our method leverages recent advances in SR to predict likely high-resolution images for compression. However, unlike standard super-resolution, our algorithm predicts a probability for each high-resolution image, in order to entropy code the image in a lossless manner.

3 Preliminaries

Lossless Compression methods encode an image x into a bitstream b_x using an invertible transformation. The goal of a compression algorithm is to minimize the expected code length $L := \mathbb{E}_{x \sim P} [|b_x|]$ of the bitstream over a distribution $x \sim P$ of natural images. The entropy $H_P = \mathbb{E}_{x \sim P} [-\log_2 P(x)]$ bounds the expected code length L from below following Shannon’s Source Coding Theorem [9, 40].

Arithmetic coding (AC) [54] is a form of entropy encoding that reaches the theoretical lower-bound of the code length within a few bits if it is given access to the distribution of images $P(x)$. For infinite precision numerical computation, AC obtains a code of length $L \leq H_P + 1$. For finite precision implementations, a few bits are wasted due to rounding. AC maps the entire image into an interval within a range $[0, 1]$, where the size of the interval is equivalent to the probability $P(x)$ of that image. The image is then encoded as the shortest integer number in that interval. Intuitively, frequently used images are mapped to a larger interval, and thus require fewer bits to encode.

In our work, we learn a distribution $P_\theta(x)$ over natural images, and then use this distribution for arithmetic coding. The code length induced by our distribution P_θ is bound by the cross entropy between the true natural image distribution P and the learned distribution P_θ : $L \leq \mathbb{E}_{x \sim P} [-\log_2 P_\theta(x)] + 1$. Thus, minimizing the bit-length is equivalent to minimizing the cross-entropy, or negative log-likelihood of the model P_θ under our data distribution P .

4 Method

Let $x^{(0)} \in \{0, \dots, 255\}^{W \times H \times 3}$ be a 3-channel input image with width W and height H . Let $y^{(1)} = \text{avgpool}_2(x^{(0)}) \in \mathbb{R}^{\lceil \frac{W}{2} \rceil \times \lceil \frac{H}{2} \rceil \times 3}$ be a downsampled version of the input, where avgpool_2 denotes average pooling of size 2 and stride 2: four neighboring pixels are averaged into a single output value. Finally, let $x^{(1)} \in \{0, \dots, 255\}^{\lceil \frac{W}{2} \rceil \times \lceil \frac{H}{2} \rceil \times 3}$ be a rounded version of $y^{(1)}$. Any further low-resolution image is then defined recursively $y^{(l+1)} = \text{avgpool}_2(x^{(l)})$ and $x^{(l)} = \text{round}(y^{(l)})$.

Our compression algorithm stores the low resolution image $x^{(3)}$ in its raw form. It also stores the rounding values $r^{(l)} = y^{(l)} - x^{(l)} \in \{-\frac{1}{4}, 0, \frac{1}{4}, \frac{1}{2}\}$, for $l = 1, 2, 3$ raw using two bits per pixel and channel. Rounding is close to uniformly random and contains little compressible information. The super-resolved pixels on the other hand are highly compressible. Our algorithm conditionally encodes the higher resolution image $x^{(l)}$ given a lower resolution image $y^{(l+1)}$ using AC based on probability estimated by a super-resolution network. Section 4.1 describes the network structure and training objective, while Section 5 covers the exact architectural details. Section 4.2 and Section 4.3 describe the encoding and decoding schemes respectively.

4.1 Autoregressive Super-Resolution Network

The goal of the super-resolution network is to predict a distribution over $x^{(l)}$ given $y^{(l+1)}$, so that we can efficiently entropy code $x^{(l)}$ based on $P(x^{(l)} | y^{(l+1)})$. In the following section, we describe our network for one level of super-resolution and omit the superscript for simplicity. Note that since we define $y_{i,j}$ to be the average of 4 pixels, $x_{2i,2j}$, $x_{2i,2j+1}$, $x_{2i+1,2j}$, and $x_{2i+1,2j+1}$, to super-resolve each pixel $y_{i,j}$, we only need to predict a distribution over three pixel values $P(x_{2i,2j}, x_{2i,2j+1}, x_{2i+1,2j} | Y_{i,j})$, where $Y_{i,j}$ is a local image region (receptive field) around pixel $y_{i,j}$. We do not encode the fourth pixel $x_{2i+1,2j+1}$ as it is reconstructed for free using $x_{2i+1,2j+1} = 4y_{i,j} - x_{2i,2j} - x_{2i,2j+1} - x_{2i+1,2j}$.

To leverage the correlation among the four pixels, we factorize their probabilities autoregressively:

$$\begin{aligned} & P(x_{2i,2j}, x_{2i,2j+1}, x_{2i+1,2j} | Y_{i,j}) \\ &= P(x_{2i,2j} | Y_{i,j}) P(x_{2i,2j+1} | Y_{i,j}, x_{2i,2j}) P(x_{2i+1,2j} | Y_{i,j}, x_{2i,2j}, x_{2i,2j+1}). \end{aligned} \quad (1)$$

Each term in the factorization is estimated by a convolutional neural network (CNN), as in Fig. 2. The features from earlier networks are fed into later networks through skip connections. Each network is further provided the true pixel values of previously coded pixels.

We similarly factorize the probability of a single pixel into three autoregressive terms for its three channels x^R, x^G, x^B (omitted in Fig. 2 for simplicity):

$$\begin{aligned} P(x_{ij} | \mathbf{Z}_{ij}) &= P(x_{ij}^R, x_{ij}^G, x_{ij}^B | \mathbf{Z}) \\ &= P(x_{ij}^R | \mathbf{Z}_{ij}) P(x_{ij}^G | \mathbf{Z}_{ij}, x_{ij}^R) P(x_{ij}^B | \mathbf{Z}_{ij}, x_{ij}^R, x_{ij}^G). \end{aligned} \quad (2)$$

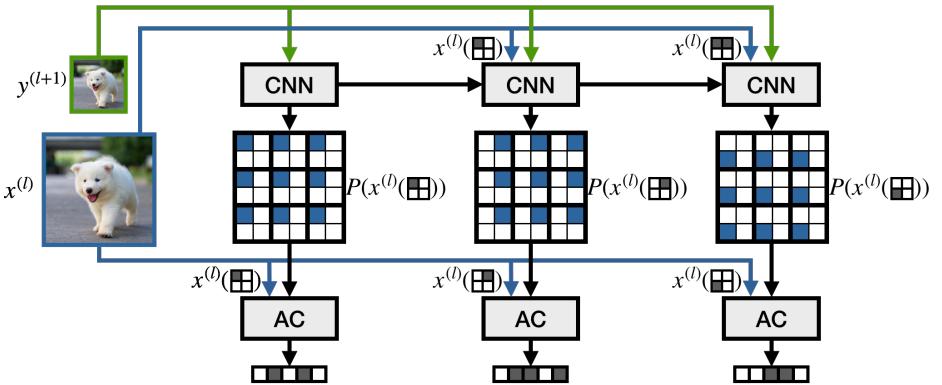


Fig. 2. Encoding $x^{(l)}$ conditioned on a downsampled image $y^{(l+1)}$. Our autoregressive network predicts the probability distribution over the first three pixels in an upsampled block sequentially. An arithmetic coder (AC) then entropy codes each of the pixels based on the estimated probability. (The fourth pixel could be computed given previously decoded pixels, so we do not need to encode it.) We use a block in parentheses to denote indexing. For example, “ (\blacksquare) ” denotes indexing the top-left pixel (i.e., $x_{2i,2j}$) in each block. “ $(\blacksquare\blacksquare)$ ” denotes indexing the top two pixels (i.e., $x_{2i,2j}$ and $x_{2i,2j+1}$) in each block.

\mathbf{Z} denotes the conditioning variables of x introduced by Equation (1). We follow PixelCNN++ [39] and parametrize this probability as a mixture of logistic functions

$$P(x_{ij}^R | \mathbf{Z}_{ij}) = \sum_{k=1}^K w_k \text{logistic}(x_{ij}^R | \mu_{ijk}, s_{ijk}),$$

where the logistic function $\text{logistic}(x|\mu, s) = \sigma\left(\frac{x-\mu+0.5}{s}\right) - \sigma\left(\frac{x-\mu-0.5}{s}\right)$ is the difference of two sigmoid functions. The distributions for x_{ij}^G and x_{ij}^B are defined analogously. We use a total of $K = 10$ mixture components for each. Our deep network produces the mixture weights w_{ijk} , mean μ_{ijk} , and standard deviation s_{ijk} parameters. For the green and blue color values, the mean μ_{ijk} and weight w_{ijk} are a linear function of the previously decoded color values. The linear functions allow for a weak form of conditioning, while keeping inference time low, as all distributional parameters are produced by a *single* network forward pass. See PixelCNN++ [39] for details.

Our overall super-resolution network contains three levels of super-resolution with skip-connections from lower-resolution to higher-resolution layers, see Section 5 for details.

Training Objective. We train our network to minimize the cross entropy between the predicted model probability P_θ and a data distribution P given by samples from an image dataset. As the model contains skip connections between

levels, we train all three super-resolution levels jointly:

$$\ell = \sum_{l=0}^2 -\mathbb{E} \left[\log P_\theta \left(x^{(l)} \mid y^{(l+1)} \right) \right]. \quad (3)$$

This objective tightly bounds to the expected bit length $\ell - 1 \leq L \leq \ell$ (see Section 3 for more discussions).

Both training and evaluation are straightforward and only depend on known quantities, e.g. down-sampled versions of the original image $x^{(l)}$ and $y^{(l)}$. They contain no interdependencies and are performed fully convolutionally in parallel. However, encoding and decoding contain several dependencies, e.g. the probability of the green pixel is not known before the red pixel is decoded. In the next two sections, we highlight how the structure of our model still allows a massively parallel encoding and decoding of the image.

4.2 Encoding

Arithmetic coding has one major drawback. Encoding and decoding are inherently sequential and follow the same fixed order. For encoding this is not a major limitation, as all probability estimates are known ahead of time. However, at decoding time the super-resolution network contains many dependencies that constrain an efficiently parallel probability estimate. In this section, we describe an encoding order which allows for massively parallel decoding in the next section.

Our algorithm first encodes the lowest-resolution level $x^{(3)}$ as raw pixels. It then stores the rounding bits to reconstruct $y^{(3)}$, and it subsequently encodes the arithmetic codes of the super-resolution network and rounding bits for all consecutive levels $x^{(2)}$, $y^{(2)}$, $x^{(1)}$, $y^{(1)}$ and $x^{(0)}$. The algorithm encodes rounding bits as two bits per color channel corresponding to the four rounding values: $\{-\frac{1}{4}, 0, \frac{1}{4}, \frac{1}{2}\}$.

For each super-resolved image $x^{(l)}$, we encode all blocks in raster scan order. Let \blacksquare be the first of four pixels to be super-resolved, \blacksquare the second, \blacksquare the third. Our algorithm first encodes all red values of the first super-resolved pixel $\blacksquare R$ for all lower-resolution pixels. All other channels then follow in order: $\blacksquare R \rightarrow \blacksquare G \rightarrow \blacksquare B$, followed by $\blacksquare R \rightarrow \blacksquare G \rightarrow \blacksquare B$, and finally $\blacksquare R \rightarrow \blacksquare G \rightarrow \blacksquare B$. All values of each channel are then arithmetically encoded. Figure 2 illustrates the process. Since we can predict probabilities for all blocks in parallel and arithmetic coding is computationally light-weight, the whole process is efficient.

4.3 Decoding

Both the network architecture and encoding procedure are chosen as to make decoding as efficient as possible, as shown in Figure 3. Both the low-resolution image $x^{(3)}$ and all rounding parameters are stored raw and can be read directly from disk. The compressed super-resolution operator depends on arithmetic coding and proceeds in three steps using three distinct network passes. The super-resolution network first produces the mixture parameters of the RGB color values

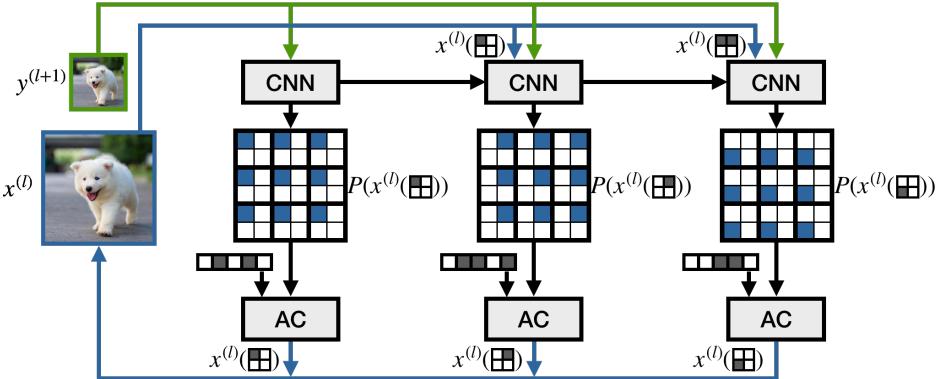


Fig. 3. Decoding $x^{(l)}$ given encoded bits and a downsampled image $y^{(l+1)}$. Our autoregressive network predicts the probability distribution over the pixels in $x^{(l)}$ sequentially given $y^{(l+1)}$. The arithmetic coder (AC) then decodes the encoded bits based on the estimated probability. We use a block in parentheses to denote indexing, similar to Fig. 2.

of the first pixel $\boxed{\square}$. These mixture parameters only depend on the low-resolution input image and can all be computed in parallel. Arithmetic coding then decodes color values one at a time: $\boxed{\square}R \rightarrow \boxed{\square}G \rightarrow \boxed{\square}B$. Note that the mixture components of green and blue depend on the previously decoded values and need to be estimated in that order. However, this can again happen in parallel once an entire color plane is decoded.

Once the first pixel $\boxed{\square}$ is decoded, a second network pass produces the mixture parameters of the second pixel $\boxed{\square}$, which is decoded analogous to the first. A final network pass then produces the third pixel value $\boxed{\square}$. The final pixel $\boxed{\square}$ is reconstructed in closed form: $x_{2i+1,2j+1} = 4y_{i,j} - x_{2i,2j} - x_{2i,2j+1} - x_{2i+1,2j}$. All network passes and parameter computations are performed in parallel over the entire image, allowing for efficient decoding on a GPU. However, decoding is marginally slower than encoding as it forces several GPU synchronizations caused by arithmetic coding.

5 Implementation Details

Network Architecture. Fig. 4 shows the architecture of our autoregressive super-resolution network. It has simple ResNet-like [15] network design, similar to a typical super-resolution architecture [29]. Following Salimans *et al.* [39], we use discretized mixture of logistics for the pixel probability distribution, and adopt the same RGB pixel conditioning scheme. See the Appendix for the exact architecture specification and details.

Handling General Image Sizes. In cases where an image’s width/height is not divisible by 2^3 , we use repeat-padding at right-most column (or bottom

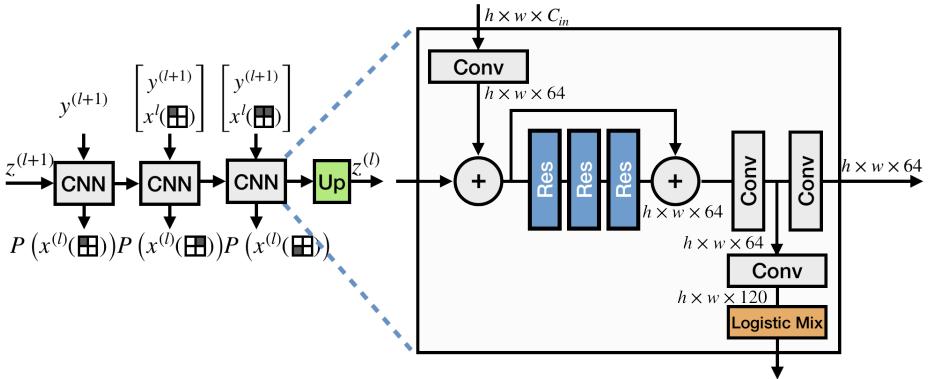


Fig. 4. Network Architecture. We use a simple building block for each of the autoregressive super-resolution steps. Right hand side shows the detailed architecture. It is composed of simple convolutional layers with residual connections to model visual patterns and a discretized-mixture-of-logistic output layer. ‘Res’ denotes a residual block; see the Appendix for details. ‘Up’ denotes an upsampling layer implemented as a PixelShuffle operator [41]. It upsamples the output feature map such that it matches the resolution of the next super-resolution level. Note the autoregressive structure: an output feature map and the pixel value of one step is passed as input to the next step.

row) when downsampling, which is equivalent to pooling only 2 or 1 pixels at the border during down-sampling. When the image has odd number of columns (or rows), at decompression time, we discard the right-most column (or bottom row) to keep the original size. Padded values do not need to be stored during encoding.

Super-Resolution Constraints. One important advantage of SR-based compression is that it naturally imposes range constraints in each block, making predicting the values easier. Specifically, given the definition of average pooling $\frac{1}{4} \left(\sum_{i'=2i}^{2i+1} \sum_{j'=2j}^{2j+1} x_{i',j'}^{(l)} \right) = y_{i,j}^{(l+1)}$, and the range constraint of a pixel $x_{i,j} \in \{0, 1, \dots, 255\}, \forall i, j$, we have

$$\begin{aligned}
 4y_{i,j} - 3 \cdot 255 &\leq x_{2i,2j} \leq 4y_{i,j} \\
 4y_{i,j} - x_{2i,2j} - 2 \cdot 255 &\leq x_{2i,2j+1} \leq 4y_{i,j} - x_{2i,2j} \\
 4y_{i,j} - x_{2i,2j} - x_{2i,2j+1} - 1 \cdot 255 &\leq x_{2i+1,2j} \leq 4y_{i,j} - x_{2i,2j} - x_{2i,2j+1}
 \end{aligned} \tag{4}$$

We can thus exclude impossible ranges during entropy coding, making our probability estimation more accurate. For simple non-factorized models, these constraints significantly improve the performance. For our autoregressive model, these constraints can be easily learned, so there is no need to explicitly impose them, as shown in experiments.

6 Experiments

We present detailed ablation study results in Section 6.1, qualitative analysis in Section 6.2, and qualitative evaluation compared with other state-of-the-art methods in Section 6.3.

Datasets and Protocol. Our evaluation protocol follows Mentzer *et al.* [33]. We evaluate SReC on two datasets, ImageNet64 [8, 11] and Open Images [25]. ImageNet64 consists of downsampled 64×64 images of ImageNet [11]. It contains $\sim 1.28m$ training and 50k validation images. Open Images [25] consists of high-resolution images. We use 2 different versions of Open Images, JPEG ³ and PNG. For fair comparison, we follow the preprocessing steps of Mentzer *et al.* [33]: We downscale the images to 768 pixels on the longer side to reduce artifacts from prior compression. We discard small ($< 1.25 \times$ downsampling) or high-saturation images. For PNG images, we apply random downscaling while keeping the shorter side ≥ 512 pixels. We use Lanczos [51] interpolation instead of bilinear for downscaling. We pick the same set of validation images as Mentzer *et al.* [33]. This process results in $\sim 336k$ training and 500 validation images.

We measure compression rate by bits per subpixel (bpsp). We measure runtime of all methods on the same machine with AMD Ryzen 5 1600 and NVIDIA GTX 1060. All runtime measurements use a single image (batch size of 1).

Training Details. We use Adam [21] with a batch size of 32 and no weight decay. For regularization, we apply gradient norm clipping at 0.5. We train our model for 10 epochs on ImageNet64 [8] and 50 epochs on Open Images [25]. We use a learning rate of 10^{-4} , which is then decreased by a factor of 0.75 every epoch for ImageNet64 [8] and every 5 epochs for Open Images [25].

We apply random horizontal flipping for training. We train with the same crop sizes as L3C for fair comparison, which is 64×64 on ImageNet64 and 128×128 on Open Images.

6.1 Ablation Experiments

We use Open Images [25] (PNG) for extensive ablation studies.

Network Design. We first ablate our network design in Table 1a. We evaluate different designs using log-likelihood in bits per subpixels (bpsp). We start with a baseline which simply replaces the latent factors of a latent factor model by downsampled images (denoted ‘SR’ in table). This model is equivalent to L3C [33] with RGB latent factors, and intermediate supervision on those latent factors. When we impose the range constraints of Equation (4) (denoted ‘SR + constraints’ in table), we can see a large improvement in bpsp ($3.36 \rightarrow 3.03$).

³ In their published results L3C uses JPEG compressed images (with compression artifacts) for Open Images. This mistake was later discovered and fixed after publication. Here, we report results on both the JPEG compressed images, and lossless PNG images for a fair comparison. Note that JPEG compression is much easier to learn for all algorithms.

	bpsp		bpsp
SR	3.36	1-level	3.87
SR + constraints	3.03	2-level	2.90
SR + factorization	2.69	3-level	2.69
SR + factorization + constraints	2.69	4-level	2.68

(a) Network Design.

(b) Compression Scheme.

Table 1. Ablation Study. We perform ablations on the Open Images dataset [25]. Table 1a validates that a super-resolution-based method benefits from the natural constraints imposed by the setting, and our factorized method is able to easily leverage the constraints and performs better than baselines. Table 1b demonstrates that each super-resolution level improves our modeling power, leading to stronger compression.

	time (s)	%		bpsp	%		W×H	enc (s)	dec (s)
$x^{(3)}$	0.0002	0.01	rounding bits	0.656	24.3		32^2	0.036	0.049
$x^{(2)}$	0.077	6.7	metadata	0.002	0.1		64^2	0.045	0.085
$x^{(1)}$	0.230	20.0	$x^{(3)}$ (raw image)	0.125	4.6		320^2	0.277	0.327
$x^{(0)}$	0.842	73.3	$x^{(2)}$	0.126	4.7		640^2	0.977	1.101
Total	1.149		$x^{(1)}$	0.432	16.0		720^2	1.166	1.549
			$x^{(0)}$	1.360	50.3		960^2	2.148	2.373
			Total		2.701				

(a) Speed.

(b) Compression Rate.

(c) Scalability.

Table 2. Detailed Analysis. We present speed and compression-rate analysis on Open Images [25] in Table 2a and Table 2b respectively. We demonstrate that our method scales to high-resolution (up to 960×960) images in Table 2c.

In fact, this latent factor RGB model with constraints performs as well as the full fledged L3C [33] model. This result suggests that by framing lossless compression as super-resolution, we can indeed benefit from the natural constraints, resulting in more accurate predictions. If we use a factorized model ('SR + factorization') to model image structures, we see even better results ($3.03 \rightarrow 2.69$). Note that the factorized model can easily learn these simple range constraints, and is most likely learning an even more complex prior in the color distribution. In fact, adding range constraints to a factorized model does not further improve performance (denoted 'SR + factorization + constraints'). In the following, we thus use our factorized variant without explicit constraints as our default model.

Compression Scheme. Table 1b compares log-likelihoods of different numbers of super-resolution levels. Each additional level improves the performance, saturating at three levels. We thus use a 3-level design as our default choice.

Speed. Table 2a presents detailed decompression runtime analysis with AC. Decompressing $x^{(3)}$ is very efficient (0.01% of total runtime) as they are simply stored as raw pixels. Higher-resolution super-resolution consumes most of the runtime. Overall we observe runtime of each level roughly linear to the number of pixels in the level.



Fig. 5. Super-resolution distribution visualization. We sample images from our network to visualize what it learns. From left to right: original image $x^{(0)}$, $x^{(0)}$ (zoomed in), downsampled image $x^{(3)}$, and samples. We sample images from full $x^{(3)}$, but just present the zoomed-in 100^2 -pixel view for clear presentation. (Best viewed on screen.)

Compression Rate. Table 2b shows detailed analysis of compression rate with AC. We see that while $x^{(3)}$ is simply stored as raw images, it contributes to only a small fraction of the overall bpsp. Finer-level of super-resolution requires more bpsp as expected, as finer-grained details are harder to predict. Rounding bits also have less structure to exploit. We store them uncompressed, taking $\sim 24\%$ of the overall bpsp. Metadata to store width and height is negligible. The rounding in AC causes 0.01 bpsp increase compared to raw log-likelihoods.

Scalability. Table 2c presents the runtime with AC when scaling to high-resolution images. Images $\geq 640^2$ come from the higher-resolution DIV2K dataset [3]. We see that even for high-resolution 960^2 images, both the encoding and decoding time stay practical. Encoding is more efficient than decoding, because it requires fewer CPU-GPU synchronizations.

6.2 Qualitative Analysis

Super-Resolution Distribution Visualization. Our network learns a distribution over possible super-resolutions. In Fig. 6, we present images sampled

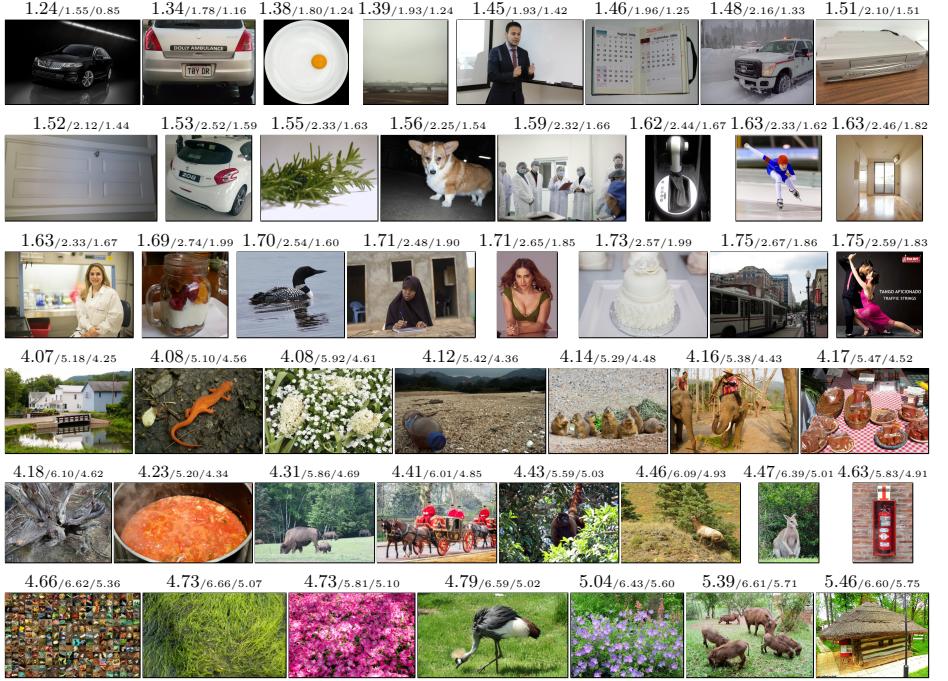


Fig. 6. Case Study. We present the most and the least compressible images (measured by SReC bpsp) in the validation set of Open Images [25]. The numbers above each image is the bpsp of “SReC/PNG/WebP” respectively. Compared to traditional methods, SReC obtains larger performance gain on more challenging cases (lower part of the table). This shows that our network effectively models complicated patterns that hand-engineered methods fail to exploit. (Best viewed on screen.)

from the distribution (conditioned on $x^{(3)}$) to visualize what the network learns. We see that our model learns a wide range of possible super-resolutions.⁴ We conjecture that the diversity is crucial for generalization to unseen images, contributing to the improved compression rate.

Case Study. We analyze what images SReC and other methods can compress more (or less) in Fig. 6. Not surprisingly, images with consistent colors or simple patterns are easier to compress. Images with high frequency changes or fine details are more challenging to compress. Compared to traditional methods, we note that SReC obtains larger performance gain on more challenging (less compressible) cases. This suggests that our network effectively models complicated image patterns that hand-engineered methods struggle at.

⁴ Artifacts are present in some cases, in part due to that the network is not optimized for image generation. The good bpsp suggests that the learned distribution does model the image distribution well.

#params (10^6)	ImageNet64			Open Images			
	encode time (s)	decode time (s)	bpsp	encode time (s)	decode time (s)	bpsp (JPEG) ⁵	bpsp
Reed <i>et al.</i> [35] ⁶	-	~ 4.68	3.70	-	-	-	-
PNG [1]	$1.3 \cdot 10^{-3}$	$8.0 \cdot 10^{-5}$	5.74	0.17	$9.8 \cdot 10^{-5}$	3.78	4.03
WebP [2]	-	0.021	$2.1 \cdot 10^{-4}$	4.64	0.40	$7.0 \cdot 10^{-4}$	2.67
FLIF [43]	-	0.022	0.010	4.54	1.23	0.30	2.47
L3C [33]	5.01	0.031	0.023	4.42	1.33	1.13	2.58
IDF [18]	84.33	1.33	1.02	3.90	57.31	62.33	2.34
SReC	4.20	0.044	0.071	4.29	0.99	1.15	2.29

Table 3. Comparison to prior work. We compare compression performance of SReC vs. other methods on ImageNet64 [8] and Open Images [25] in bpsp, runtime, and number of parameters. We additionally list an efficient PixelCNN variant of Reed *et al.* [35] purely for reference. It is not practical for lossless compression yet due to its long runtime. SReC outperforms all practical algorithms in terms of bpsp, while being efficient and small in size.

6.3 Comparison to Prior Work

In Table 3, we compare SReC with engineered codecs, PNG [1], WebP [2], and FLIF [43], as well as deep learning based methods L3C [33] and IDF [18]. We are not able to train IDF on full resolution due to GPU memory constraints and thus tiled the model over 64×64 crops on Open Images [25]. On Open Images [25], SReC outperforms all prior work, while being efficient — $\sim 55 \times$ faster than the second best performing method, IDF [18]. Engineered codecs, such as PNG or FLIF, are more efficient than deep-network based methods. However, they fail to achieve good compression rates. On 64×64 small images (ImageNet64 [8]), SReC again demonstrates a strong compression rate. It outperforms all methods, except for IDF, which is $30 \times$ slower to encode and $14 \times$ slower to decode. We also list performance of an efficient PixelCNN variant of Reed *et al.* [35] purely for reference. Reed *et al.* [35] achieves a good compression rate, but its runtime is not practical ($66 \times$ slower than SReC). Our method is also parameter efficient — smaller than other deep-learning based methods. For example, SReC is $\sim 20 \times$ smaller than the second-best-performing method.

7 Conclusions

We propose Super-Resolution based Compression (SReC), which relies on multiple levels of lossless super-resolution. We show that lossless super-resolution operators are efficient to store, due to the natural constraints induced by the super-resolution setting. On multiple datasets, we show state-of-the-art compression rates. Overall, our method is simple and efficient. Its model size is small, and its runtime is on par with or faster than other deep-network based compression methods.

⁶ Bpsp results are taken from Reed *et al.* [35]. Timing is extrapolated from 32×32 runtime following [33].

References

1. Portable network graphics, <http://libpng.org/pub/png/libpng.html>
2. Webp image format, <https://developers.google.com/speed/webp/>
3. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: CVPR Workshops (2017)
4. Agustsson, E., Tschannen, M., Mentzer, F., Timofte, R., Gool, L.V.: Generative adversarial networks for extreme learned image compression. In: ICCV (2019)
5. Ballé, J., Laparra, V., Simoncelli, E.P.: End-to-end optimized image compression. In: ICLR (2017)
6. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: ICCV (2019)
7. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
8. Chrabaszcz, P., Loshchilov, I., Hutter, F.: A downsampled variant of ImageNet as an alternative to the CIFAR datasets. arXiv preprint arXiv:1707.08819 (2017)
9. Cover, T.M., Thomas, J.A.: Elements of information theory. John Wiley & Sons (2012)
10. Dahl, R., Norouzi, M., Shlens, J.: Pixel recursive super resolution. In: ICCV (2017)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009)
12. Deutsch, P.: Deflate compressed data format specification version 1.3 (1996)
13. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV (2014)
14. Duda, J.: Asymmetric numeral systems. arXiv preprint arXiv:0902.0271 (2009)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
16. Hinton, G.E., Van Camp, D.: Keeping the neural networks simple by minimizing the description length of the weights. In: COLT (1993)
17. Ho, J., Lohn, E., Abbeel, P.: Compression with flows via local bits-back coding. In: NeurIPS (2019)
18. Hoogeboom, E., Peters, J., van den Berg, R., Welling, M.: Integer discrete flows and lossless compression. In: NeurIPS (2019)
19. Johnston, N., Vincent, D., Minnen, D., Covell, M., Singh, S., Chinen, T., Jin Hwang, S., Shor, J., Toderici, G.: Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In: CVPR (2018)
20. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: CVPR (2016)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
23. Kingma, F., Abbeel, P., Ho, J.: Bit-swap: Recursive bits-back coding for lossless compression with hierarchical latent variables. In: ICML (2019)
24. Kolesnikov, A., Lampert, C.H.: PixelCNN models with auxiliary variables for natural image modeling. In: ICML (2017)
25. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., et al.: Open Images: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from <https://github.com/openimages> (2017)

26. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017)
27. Li, M., Zuo, W., Gu, S., Zhao, D., Zhang, D.: Learning convolutional networks for content-weighted image compression. In: CVPR (2018)
28. Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., Wu, W.: Feedback network for image super-resolution. In: CVPR (2019)
29. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: CVPR Workshops (2017)
30. Lu, G., Ouyang, W., Xu, D., Zhang, X., Cai, C., Gao, Z.: DVC: An end-to-end deep video compression framework. In: CVPR (2019)
31. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: ICML (2013)
32. Meeker, M.: Internet trends. Bond reports (2019)
33. Mentzer, F., Agustsson, E., Tschannen, M., Timofte, R., Gool, L.V.: Practical full resolution learned lossless image compression. In: CVPR (2019)
34. Rad, M.S., Bozorgtabar, B., Marti, U.V., Basler, M., Ekenel, H.K., Thiran, J.P.: Srobb: Targeted perceptual loss for single image super-resolution. In: ICCV (2019)
35. Reed, S.E., van den Oord, A., Kalchbrenner, N., Gómez, S., Wang, Z., Belov, D., de Freitas, N.: Parallel multiscale autoregressive density estimation. In: ICML (2017)
36. Rezende, D.J., Mohamed, S.: Variational inference with normalizing flows. In: ICML (2015)
37. Rippel, O., Bourdev, L.: Real-time adaptive image compression. In: ICML (2017)
38. Rippel, O., Nair, S., Lew, C., Branson, S., Anderson, A.G., Bourdev, L.: Learned video compression. In: ICCV (2019)
39. Salimans, T., Karpathy, A., Chen, X., Kingma, D.P.: PixelCNN++: A PixelCNN implementation with discretized logistic mixture likelihood and other modifications. In: ICLR (2017)
40. Shannon, C.: A mathematical theory of communication. Bell System Technical Journal (1948)
41. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: CVPR (2016)
42. Skodras, A., Christopoulos, C., Ebrahimi, T.: The jpeg 2000 still image compression standard. IEEE Signal processing magazine (2001)
43. Sneyers, J., Wuille, P.: Flif: Free lossless image format based on maniac compression. In: ICIP (2016)
44. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: CVPR (2017)
45. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: CVPR (2017)
46. Toderici, G., O'Malley, S.M., Hwang, S.J., Vincent, D., Minnen, D., Baluja, S., Covell, M., Sukthankar, R.: Variable rate image compression with recurrent neural networks. In: ICLR (2016)
47. Toderici, G., Vincent, D., Johnston, N., Jin Hwang, S., Minnen, D., Shor, J., Covell, M.: Full resolution image compression with recurrent neural networks. In: CVPR (2017)
48. Tong, T., Li, G., Liu, X., Gao, Q.: Image super-resolution using dense skip connections. In: ICCV (2017)

49. Townsend, J., Bird, T., Barber, D.: Practical lossless compression with latent variables using bits back coding. In: ICLR (2019)
50. Townsend, J., Bird, T., Kunze, J., Barber, D.: Hilloc: Lossless image compression with hierarchical latent variable models. In: ICLR (2020)
51. Turkowski, K.: Filters for common resampling tasks. In: Graphics gems (1990)
52. Van Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: ICML (2016)
53. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: EDVR: Video restoration with enhanced deformable convolutional networks. In: CVPR Workshops (2019)
54. Witten, I.H., Neal, R.M., Cleary, J.G.: Arithmetic coding for data compression. Communications of the ACM (1987)
55. Wu, C.Y., Singhal, N., Krähenbühl, P.: Video compression through image interpolation. In: ECCV (2018)
56. Zhou, R., Susstrunk, S.: Kernel modeling super-resolution on real low-resolution images. In: ICCV (2019)

Appendix A Architecture Details

Each level l of super-resolution uses a network to predict a distribution of $x^{(l)}$ given $y^{(l+1)}$. Network weights are not shared across levels. Levels $l = 0, 1$ additionally take the output activation $z^{(l+1)}$ from the previous level as the second input. See left hand side of Fig. 4 for a schematic illustration of the network (of one level). The network for each level predicts the first three pixels in each 4-pixel block in an autoregressive fashion.

The right hand side of Fig. 4 illustrates the exact architecture. All convolutional layers use a kernel size of 3×3 , stride of 1×1 , dilation of 1×1 , padding of 1×1 , and 64 output channels except for the first convolution in each CNN, which differs in using kernel size of 1×1 . Since each image has 3 channels, the inputs, $y^{(l+1)}$, $[y^{(l+1)}, x^{(l)}(\blacksquare)]$, and $[y^{(l+1)}, x^{(l)}(\blacksquare)]$ has the number of input channels $C_{in} = 3, 6$, and 9, respectively.⁷

Residual block (denoted ‘Res’ in Fig. 4) consists of two convolutional layers with a leaky ReLU [31] in between. A skip connection [15] is used to sum the input and the output of this two-convolutional-layer block.

Logistic Mixture Output. The convolutional layer before the discrete logistic mixture output [39] is a stacked atrous convolution operator [7], following the same design in L3C [33]. Following Salimans *et al.* [39], we use 10 mixtures, with each discrete logistic being parameterized by 12 parameters.

Upsampling layer (denoted ‘Up’ in Fig. 4), is implemented as a PixelShuffle operator [41]. It doubles the width and height of the input, while simultaneously shrinks the channel size by a factor of 4. The convolution before PixelShuffle has 256 output channels, such that after applying PixelShuffle, the channel size remains 64.

Additional Training Details. We apply random 128×128 cropping while training on Open Images [25]. We keep ImageNet64 [8] images as 64×64 . We apply random horizontal flipping during training.

⁷ Square brackets denote channel concatenation.