Generative Adversarial Networks for Extreme Learned Image Compression

Eirikur Agustsson* Michael Tschannen* Fabian Mentzer* Radu Timofte Luc Van Gool

aeirikur@vision.ee.ethz.ch mi.tschannen@gmail.com mentzerf@vision.ee.ethz.ch timofter@vision.ee.ethz.ch vangool@vision.ee.ethz.ch

ETH Zürich, Switzerland

Abstract

We present a learned image compression system based on GANs, operating at extremely low bitrates. Our proposed framework combines an encoder, decoder/generator and a multi-scale discriminator, which we train jointly for a generative learned compression objective. The model synthesizes details it cannot afford to store, obtaining visually pleasing results at bitrates where previous methods fail and show strong artifacts. Furthermore, if a semantic label map of the original image is available, our method can fully synthesize unimportant regions in the decoded image such as streets and trees from the label map, proportionally reducing the storage cost. A user study confirms that for low bitrates, our approach is preferred to state-of-the-art methods, even when they use more than double the bits.

1. Introduction

Image compression systems based on deep neural networks (DNNs), or deep compression systems for short, have become an active area of research recently. These systems (e.g. [39, 5, 34, 6, 30]) are often competitive with modern engineered codecs such as WebP [46], JPEG2000 [38] and even BPG [7] (the state-of-the-art engineered codec). Besides achieving competitive compression rates on natural images, they can be easily adapted to specific target domains such as stereo or medical images, and promise efficient processing and indexing directly from compressed representations [42]. However, deep compression systems are typically optimized for traditional distortion metrics such as peak signal-to-noise ratio (PSNR) or multi-scale structural similarity (MS-SSIM) [45]. For very low bitrates (below 0.1 bits per pixel (bpp)), where preserving the full image content becomes impossible, these distortion metrics lose significance as they favor pixel-wise preservation of local (high-entropy) structure over preserving texture and global structure (see [8] and Sec. 4.3). To further advance deep image compression it is therefore of great importance to develop new training objectives beyond PSNR and MS-SSIM. A promising candidate towards this goal are adversarial losses [13] which were shown recently to capture global semantic information and local texture, yielding powerful generators that produce visually appealing high-resolution images from semantic label maps [20, 44].

In this paper, we propose a principled GAN framework for full-resolution image compression and use it to realize

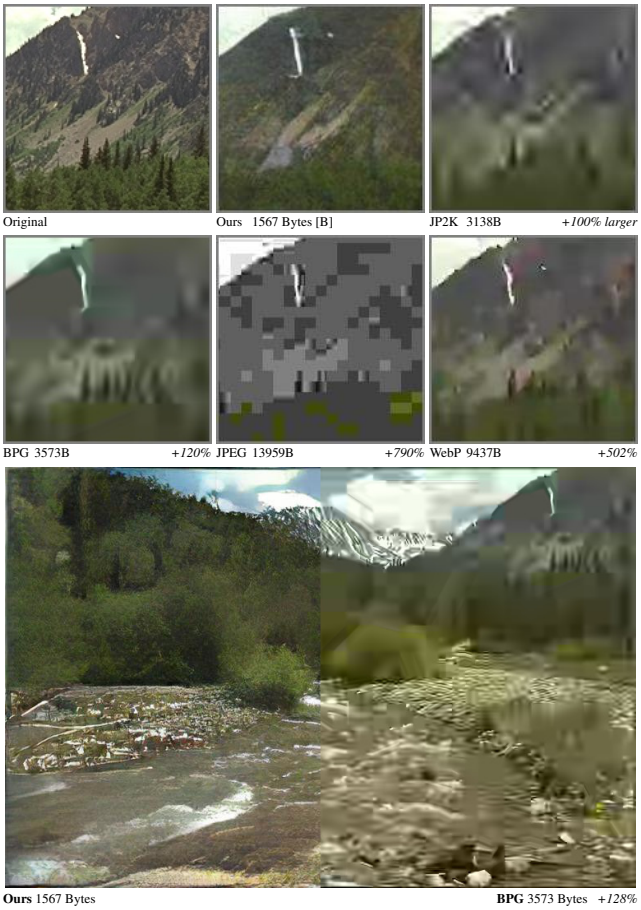


Figure 1. Visual comparison of our result to that obtained by other codecs. Note that even when using more than twice the number of bytes, all other codecs are outperformed by our method visually.

*The first three authors contributed equally.

an extreme image compression system, targeting bitrates below 0.1bpp. Furthermore, in contrast to prior work, we provide the first thorough user study of such a framework in the context of full-resolution image compression.

In our framework, we consider two modes of operation (corresponding to unconditional and conditional GANs [13, 32]), namely

- *generative compression (GC)*, preserving the overall image content while generating structure of different scales such as leaves of trees or windows in the facade of buildings, and
- *selective generative compression (SC)*, completely generating parts of the image from a semantic label map while preserving user-defined regions with a high degree of detail.

We emphasize that GC does not require semantic label maps (neither for training, nor for deployment). A typical use case for GC are bandwidth constrained scenarios, where one wants to preserve the full image as well as possible, while falling back to synthesized content instead of blocky/blurry blobs for regions for which not sufficient bits are available to store the original pixels. SC could be applied in a video call scenario where one wants to fully preserve people in the video stream, but a visually pleasing synthesized background serves the purpose as well as the true background. In the GC operation mode the image is transformed into a bitstream and encoded using arithmetic coding. SC requires a semantic/instance label map of the original image which can be obtained using off-the-shelf semantic/instance segmentation networks, e.g., PSPNet [49] and Mask R-CNN [17], and which is stored as a vector graphic. This amounts to a small, image dimension-independent overhead in terms of coding cost. However, the size of the compressed image is reduced proportionally to the area which is generated from the semantic label map, typically leading to a significant overall reduction in storage cost.

For GC, a comprehensive user study shows that our compression system yields visually considerably more appealing results than BPG [7] (the current state-of-the-art engineered compression algorithm) and the recently proposed autoencoder-based deep compression (AEDC) system [30]. In particular, our GC models trained for compression of general natural images are preferred to BPG when BPG uses up to 95% and 124% more bits than those produced by our models on the Kodak [24] and RAISE1K [11] data set, respectively. When constraining the target domain to the street scene images of the Cityscapes data set [9], the reconstructions of our GC models are preferred to BPG even when the latter uses up to 181% more bits. To the best of our knowledge, these are the first results showing that a deep compression method outperforms BPG on the Kodak data set in a user study—and by large margins.

In the SC operation mode, our system seamlessly combines preserved image content with synthesized content, even for regions that cross multiple object boundaries, while faithfully preserving the image semantics. By partially generating image content we achieve bitrate reductions of over 50% without notably degrading image quality.

In summary, our main contributions are as follows.

- We provide a principled GAN framework for full-resolution image compression and use it to build an extreme image compression system.
- We are the first to thoroughly explore such a framework in the context of full-resolution image compression.
- We set new state-of-the-art in visual quality based on a user study, with dramatic bitrate savings.

2. Related work

Deep image compression has recently emerged as an active area of research. The most popular DNN architectures for this task are to date auto-encoders [39, 5, 1, 27, 42, 31, 6] and recurrent neural networks (RNNs) [40, 41]. These DNNs transform the input image into a bit-stream, which is in turn losslessly compressed using entropy coding methods such as Huffman coding or arithmetic coding. To reduce coding rates, many deep compression systems rely on context models to capture the distribution of the bit stream [5, 41, 27, 34, 30]. Common loss functions to measure the distortion between the original and decompressed images are the mean-squared error (MSE) [39, 5, 1, 27, 6, 42], or perceptual metrics such as MS-SSIM [41, 34, 6, 30]. Some authors rely on advanced techniques including multi-scale decompositions [34], progressive encoding/decoding strategies [40, 41], and generalized divisive normalization (GDN) layers [5, 4].

Generative adversarial networks (GANs) [13] have emerged as a popular technique for learning generative models for intractable distributions in an unsupervised manner. Despite stability issues [35, 2, 3, 29], they were shown to be capable of generating more realistic and sharper images than prior approaches and to scale to resolutions of 1024×1024px [47, 22] for some data sets. Another direction that has shown great progress are conditional GANs [13, 32], obtaining impressive results for image-to-image translation [20, 44, 50, 28] on various data sets (e.g. maps to satellite images), reaching resolutions as high as 1024×2048px [44].

The work of [34] trains and evaluates a deep compression system optimized for the classical MS-SSIM [45] metric. Furthermore, they supplement their method with an adversarial training scheme to reduce compression artifacts. However, it is impossible to assess the benefit of their adversarial scheme since there is no ablation study showing its effect. In contrast, we provide a thorough study of the

benefit of our GAN formulation, compared to optimizing for classical losses such as MSE and MS-SSIM. Additionally, their approach is very different: First, their GAN loss is non-standard, operating on pairs of real/fake images classifying "which one is the real one", whereas ours has a principled interpretation in terms of divergences between probability distributions (as in [13, 33]). Second, their training uses various heuristics to balance the training, such as reweighting losses based on gradient magnitudes and alternating the training of the generator and discriminator based on manually defined thresholds on the losses.

Santurkar *et al*. [36] use a GAN framework to learn a generative model over thumbnail images, which is then used as a decoder for thumbnail image compression. Other works use adversarial training for compression artifact removal (for engineered codecs) [12] and single image super-resolution [26]. Finally, related to our SC mode, spatially allocating bitrate based on saliency of image content has a long history in the context of engineered compression algorithms, see, e.g., [37, 15, 16].

3. Background

Generative Adversarial Networks: Given a data set $\mathcal{X}$, GANs can learn to approximate its (unknown) distribution $p_{\boldsymbol{x}}$ through a generator $G(\boldsymbol{z})$ that tries to map samples $\boldsymbol{z}$ from a fixed prior distribution $p_{\boldsymbol{z}}$ to the data distribution $p_{\boldsymbol{x}}$. The generator G is trained in parallel with a discriminator D by searching (using SGD) for a saddle point of a min-max objective $\min_G \mathcal{L}_{\text{GAN}}$ with

$$\mathcal{L}_{\text{GAN}} := \max_D \mathbb{E}[f(D(\boldsymbol{x}))] + \mathbb{E}[g(D(G(\boldsymbol{z})))], \quad (1)$$

where G and D are DNNs and f and g are scalar functions. Nowozin *et al*. [33] show that for suitable choices of f and g solving $\min_G \mathcal{L}_{\text{GAN}}$ allows to minimize general f-divergences between the distribution of $G(z)$ and p_x. We adapt Least-Squares GAN [29] in this paper, where $f(y) = (y-1)^2$ and $g(y) = y^2$ (which corresponds to the Pearson χ^2 divergence).

Conditional Generative Adversarial Networks: For conditional GANs (cGANs) [13, 32], each data point $\boldsymbol{x}$ is associated with additional information $\boldsymbol{s}$, where $(\boldsymbol{x}, \boldsymbol{s})$ have an unknown joint distribution $p_{\boldsymbol{x},\boldsymbol{s}}$. We now assume that $\boldsymbol{s}$ is given and that we want to use the GAN to model the conditional distribution $p_{\boldsymbol{x}|\boldsymbol{s}}$. In this case, both the generator $G(\boldsymbol{z}, \boldsymbol{s})$ and discriminator $D(\boldsymbol{z}, \boldsymbol{s})$ have access to the side information $\boldsymbol{s}$, leading to the divergence

$$\mathcal{L}_{\text{cGAN}} := \max_D \mathbb{E}[f(D(\boldsymbol{x}, \boldsymbol{s}))] + \mathbb{E}[g(D(G(\boldsymbol{z}, \boldsymbol{s}), \boldsymbol{s}))].$$

Deep Image Compression: To compress an image $\boldsymbol{x} \in \mathcal{X}$, we follow the formulation of [1, 30] where one learns an encoder E, a decoder G, and a finite quantizer q. The encoder E maps the image to a latent feature map $\boldsymbol{w}$, whose values are then quantized to L levels $\mathcal{C} = \{c_1, \dots, c_L\} \subset \mathbb{R}$ to obtain a representation $\hat{\boldsymbol{w}} = q(E(\boldsymbol{x}))$ that can be encoded to a bitstream. The decoder then tries to recover the image by forming a reconstruction $\hat{\boldsymbol{x}} = G(\hat{\boldsymbol{w}})$. To be able to backpropagate through the non-differentiable q, one can use a differentiable relaxation of q, as in [30].

The average number of bits needed to encode $\hat{\boldsymbol{w}}$ is measured by the entropy $H(\hat{\boldsymbol{w}})$, which can be modeled with a prior [1] or a conditional probability model [30]. The so called "rate-distortion" trade-off between reconstruction quality and bitrate to be optimized is then

$$\mathbb{E}[d(\boldsymbol{x}, \hat{\boldsymbol{x}})] + \beta H(\hat{\boldsymbol{w}}). \quad (2)$$

where d is a loss that measures how perceptually similar $\hat{\boldsymbol{x}}$ is to $\boldsymbol{x}$. Given a differentiable estimator of the entropy $H(\hat{\boldsymbol{w}})$, the weight β controls the bitrate of the model. However, since the number of dimensions $\dim(\hat{\boldsymbol{w}})$ and the number of levels L are finite, the entropy is bounded by (see, e.g., [10])

$$H(\hat{\boldsymbol{w}}) \leq \dim(\hat{\boldsymbol{w}}) \log_2(L). \quad (3)$$

It is therefore also valid to set $\beta = 0$ and control the maximum bitrate through the bound (3) (i.e., adjusting L and/or $\dim(\hat{\boldsymbol{w}})$ through the architecture of E). While potentially leading to suboptimal bitrates, this avoids to model the entropy explicitly as a loss term.

4. GANs for extreme image compression

4.1. Generative Compression

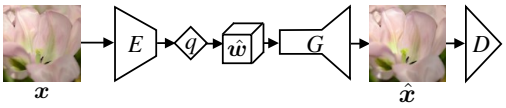


Figure 2. Architecture of our GC network.

The proposed GAN framework for extreme image compression can be viewed as a combination of (conditional) GANs and learned compression as introduced in the previous section. See Fig. 2 for an overview of the architecture. With an encoder E and quantizer q, we encode the image $\boldsymbol{x}$ to a compressed representation $\hat{\boldsymbol{w}} = q(E(\boldsymbol{x}))$. This representation is optionally concatenated with noise $\boldsymbol{v}$ drawn from a fixed prior $p_{\boldsymbol{v}}$, to form the latent vector $\boldsymbol{z}$. The decoder/generator G then tries to generate an image $\hat{\boldsymbol{x}} = G(\boldsymbol{z})$ that is consistent with the image distribution $p_{\boldsymbol{x}}$ while also recovering the specific encoded image $\boldsymbol{x}$ to a certain degree. Using $\boldsymbol{z} = [\hat{\boldsymbol{w}}, \boldsymbol{v}]$, this can be expressed by our saddle-point objective for (unconditional) generative compression,

$$\min_{E,G} \max_D \quad \mathbb{E}[f(D(\boldsymbol{x}))] + \mathbb{E}[g(D(G(\boldsymbol{z}))] + \lambda \mathbb{E}[d(\boldsymbol{x}, G(\boldsymbol{z}))] + \beta H(\hat{\boldsymbol{w}}), \quad (4)$$

where $\lambda > 0$ balances the distortion term against the GAN loss and entropy terms.[1]

Since the last two terms of (4) do not depend on the discriminator D, they do not affect its optimization directly. This means that the discriminator still computes the same f-divergence $\mathcal{L}_{\text{GAN}}$ as in (1), so we can write (4) as

$$\min_{E,G} \quad \mathcal{L}_{\text{GAN}} + \lambda \mathbb{E}[d(\boldsymbol{x}, G(\boldsymbol{z}))] + \beta H(\hat{\boldsymbol{w}}). \tag{5}$$

We note that equation (5) has completely different dynamics than a normal GAN, because the latent space $\boldsymbol{z}$ contains $\hat{\boldsymbol{w}}$, which stores information about a real image $\boldsymbol{x}$.

The bitrate limitation on $H(\hat{\boldsymbol{w}})$ is a crucial element. If we allow $\hat{\boldsymbol{w}}$ to contain arbitrarily many bits (using $\beta = 0$ and $L, \dim(\hat{\boldsymbol{w}})$ large enough), E and G could learn to near-losslessly recover $\boldsymbol{x}$ from $G(\boldsymbol{z}) = G(q(E(\boldsymbol{x})))$, such that the distortion term would vanish. In this case, the divergence between $p_{\boldsymbol{x}}$ and $p_{G(\boldsymbol{z})}$ would also vanish and the GAN loss would have no effect. On the other hand, if $H(\hat{\boldsymbol{w}}) \to 0$ (using $\beta = \infty$ or $\dim(\hat{\boldsymbol{w}}) = 0$), $\hat{\boldsymbol{w}}$ becomes deterministic. In this setting, $\boldsymbol{z}$ is random and independent of $\boldsymbol{x}$ (through the $\boldsymbol{v}$ component) and the objective reduces to a standard GAN plus the distortion term, which then acts as a regularizer.

By constraining the entropy of $\hat{\boldsymbol{w}}$, E and G will never be able to make d fully vanish. In this case, E, G need to balance the GAN objective $\mathcal{L}_{\text{GAN}}$ and the distortion term $\lambda \mathbb{E}[d(\boldsymbol{x}, G(\boldsymbol{z}))]$, which leads to $G(\boldsymbol{z})$ on one hand looking "realistic", and on the other hand preserving the original image. For example, if there is a tree for which E cannot afford to store the exact texture (and make d small) G can synthesize it to satisfy $\mathcal{L}_{\text{GAN}}$, instead of showing a blurry green blob. Thereby, the distortion term stabilizes GAN training and tends to prevent mode collapse (as mode collapse would lead to a very large distortion value). We refer to this setting as *generative compression* (GC).

As for the GANs described in Sec. 3, we can easily extend GC to a conditional case. We consider the setting where the additional information $\boldsymbol{s}$ for an image $\boldsymbol{x}$ is a semantic label map of the scene, but with a twist: Instead of feeding $\boldsymbol{s}$ to E, G and D, we *only give it to the discriminator* D during training. We refer to this setting as "GC (D^+)". We emphasize that *no semantics are needed* to encode or decode images with the trained models, in neither GC nor GC (D^+) (since E, G do not depend on $\boldsymbol{s}$).

Finally, we note that Eq. 5 is similar to classical rate-distortion theory, where $H(\hat{\boldsymbol{w}})$ is the rate/entropy term. Regarding the interaction between the GAN loss and the MSE loss, we observe that the MSE loss stabilizes the training as it penalizes collapse of the GAN.

[1] In this formulation, we need to encode a real image to sample from $p_{\hat{\boldsymbol{w}}}$. However, this is not a limitation, as our goal is compressing real images, not generating completely new ones.

4.2. Selective Generative Compression

For GC and GC (D^+), E, G automatically navigate the trade-off between generation and preservation over the entire image, without any guidance. We also consider a different setting, *selective generative compression* (SC). Here, the network is guided in terms of what should be generated and what should be preserved. An overview of the network structure is given in Fig. 9 in Appendix E.

For simplicity, we consider a binary setting, where we construct a single-channel binary heatmap $\boldsymbol{m}$ of the same spatial dimensions as $\hat{\boldsymbol{w}}$. Regions of zeros correspond to regions that should be fully synthesized, regions of ones should be preserved. However, since our task is compression, we constrain the fully synthesized regions to have the same semantics $\boldsymbol{s}$ as the original image $\boldsymbol{x}$. We assume the semantics $\boldsymbol{s}$ are separately stored, and feed them through a feature extractor F before feeding them to the generator G. To guide the network with the semantics, we mask the (pixel-wise) distortion d, such that it is only computed over the region to be preserved. Additionally, we zero out the compressed representation $\hat{\boldsymbol{w}}$ in the regions that should be synthesized. Provided that the heatmap $\boldsymbol{m}$ is also stored, we then only encode the entries of $\hat{\boldsymbol{w}}$ corresponding to the preserved regions, greatly reducing the bitrate needed to store it. At bitrates where $\hat{\boldsymbol{w}}$ is much larger on average than the storage cost for $\boldsymbol{s}$ and $\boldsymbol{m}$, this approach can result in large bitrate savings.

We consider two different training modes: Random instance (RI) which randomly selects 25% of the instances in the semantic label map and preserves these, and random box (RB) which picks an image location uniformly at random and preserves a box of random dimensions. While the RI mode is appropriate for most use cases, RB can create more challenging situations for the generator as it needs to integrate the preserved box seamlessly into generated content.

4.3. PSNR and MS-SSIM as quality measures

Our model targets realistic reconstructions where texture and sometimes even more abstract image content is synthesized. Common distortion measures such as PSNR and MS-SSIM cannot measure the "realistic-ness", as they penalize changes in local structure rather than assessing preservation of the global image content. This fact was *mathematically* proven recently by [8], showing the existence of a fundamental perception-distortion tradeoff, i.e., low distortion is at odds with high perceptual quality in the context of lossy reconstruction tasks. Intuitively, measuring PSNR between synthesized and real texture patches essentially quantifies the variance of the texture rather than the perceptual quality of the synthesized texture. This becomes apparent by comparing reconstructions produced by our GC model with those obtained by the MSE baseline and BPG in Fig. 3. While our reconstructions clearly look realistic, they have 4.2dB

larger MSE than those of BPG. We therefore rely on human opinions collected in a thorough user study to evaluate our GC models.

5. Experiments

5.1. Architecture, Losses, and Hyperparameters

The architecture for our encoder E and generator G is based on the global generator network proposed in [44], which in turn is based on the architecture of [21]. We present details in Appendix E.

For the entropy term $\beta H(\hat{\boldsymbol{w}})$, we adopt the simplified approach described in Sec. 3, where we set $\beta = 0$, use $L = 5$ centers $\mathcal{C} = \{-2, 1, 0, 1, 2\}$, and control the bitrate through the upper bound $H(\hat{\boldsymbol{w}}) \leq \dim(\hat{\boldsymbol{w}}) \log_2(L)$. For example, for GC, with $C = 2$ bottleneck channels, we obtain 0.0181bpp.[2] We note that this is an upper bound; the actual entropy of $H(\hat{\boldsymbol{w}})$ is generally smaller, since the learned distribution will neither be uniform nor i.i.d, which would be required for the bound to hold with equality. We use an arithmetic encoder to encode the channels of $\hat{\boldsymbol{w}}$ to a bit-stream, storing frequencies for each channel separately (similar to [1]). In our experiments, this leads to 8.8% smaller bitrates compared to the upper bound. We leave the exploration of context models to potentially further reduce the bitrate for future work.

For the distortion term d we adopt MSE with $\lambda = 10$. Furthermore, we adopt the feature matching and VGG perceptual losses, $\mathcal{L}_{\text{FM}}$ and $\mathcal{L}_{\text{VGG}}$, as proposed in [44] with the same weights, which improved the quality for images synthesized from semantic label maps. These losses can be viewed as a part of $d(\boldsymbol{x}, \hat{\boldsymbol{x}})$. However, we do not mask them in SC, since they also help to stabilize the GAN in this operation mode (as in [44]). We refer to Appendix B for training details.

5.2. Evaluation

Data sets: We train GC models (without semantic label maps) for compression of diverse natural images using 188k images from the *Open Images* data set [25] and evaluate them on the widely used Kodak image compression data set [24] as well as 20 randomly selected images from the *RAISE1K* data set [11]. To investigate the benefits of having a somewhat constrained application domain and semantic information at training time, we also train GC models with semantic label maps on the *Cityscapes* data set [9], using 20 randomly selected images from the validation set for evaluation. To evaluate the proposed SC method (which requires semantic label maps for training and deployment) we again rely on the Cityscapes data set. Cityscapes was previously used to generate images form semantic label maps using GANs [20, 50].

[2] $H(\hat{\boldsymbol{w}})/WH \leq \frac{WH}{16 \cdot 16} \cdot C \cdot \log_2(L)/WH = 0.0181$bpp, where W, H are the dimensions of the image and 16 is the downsampling factor to the feature map, see Appendix E.

Figure 3. Visual example of images produced by our GC network with $C = 4$ bottleneck channels along with the corresponding results for BPG, and a baseline model with the same architecture ($C = 4$) but trained for MSE only (MSE bl.), on Cityscapes. We show bitrate in bpp and PSNR in dB. The reconstruction of our GC network is sharper and has more realistic texture than BPG and MSE bl., even though the latter two have higher PSNR. In particular, the MSE bl. produces blurry reconstructions even though it was trained on the Cityscapes data set, demonstrating that domain-specific training alone is not enough to obtain sharp reconstructions at low bitrates.

Baselines: We compare our method to the HEVC-based image compression algorithm BPG [7] (in the 4:2:2 chroma format) and to the AEDC network from [30]. BPG is the current state-of-the-art engineered image compression codec and outperforms other recent codecs such as JPEG2000 and WebP on different data sets in terms of PSNR (see, e.g. [6]). We train the AEDC network (with bottleneck depth $C = 4$) for MS-SSIM on Cityscapes exactly following the procedure in [30] except that we use early stopping to prevent overfitting (note that Cityscapes is much smaller than the ImageNet data set used in [30]). The so-obtained model has a bitrate of 0.07 bpp and gets a slightly better MS-SSIM than BPG at the same bpp on the validation set. To investigate the effect of the GAN term in our total loss, we train a baseline model with an MSE loss only (with the same architecture as GC and the same training parameters, see Sec. B in the Appendix), referred to as "MSE baseline".

User study: Given that classical distortion metrics like PSNR or MS-SSIM are not suited for the task we study here (Section 4.3), we quantitatively evaluate the perceptual quality of our GC models in comparison with BPG and AEDC (for Cityscapes) with a user study on Amazon Mechanical Turk (AMT).[3] We consider two GC models with $C = 4, 8$ bottleneck channels trained on Open Images, three GC (D^+) models with $C = 2, 4, 8$ trained on Cityscapes, and BPG

[3] https://www.mturk.com/

at rates ranging from 0.045 to 0.12 bpp. Questionnaires are composed by combining the reconstructions produced by the selected GC model for all testing images with the corresponding reconstruction produced by the competing baseline model side-by-side (presenting the reconstructions in random order). The original image is shown along with the reconstructions, and the pairwise comparisons are interleaved with 3 probing comparisons of an additional uncompressed image from the respective testing set with an obviously JPEG-compressed version of that image. 20 randomly selected unique users are asked to indicate their preference for each pair of reconstructions in the questionnaire, resulting in a total of 480 ratings per pairing of methods for Kodak, and 400 ratings for RAISE1K and Cityscapes. For each pairing of methods, we report the mean preference score as well as the standard error (SE) of the per-user mean preference percentages. Only users correctly identifying the original image in all probing comparisons are taken into account for the mean preference percentage computation. To facilitate comparisons for future works, we will release all images used in the user studies.

Semantic quality of SC models: The issues with PSNR and MS-SSIM described in Sec. 4.3 become even more severe for SC models as a large fraction of the image content is generated from a semantic label map. Following image translation works [20, 44], we therefore measure the capacity of our SC models to preserve the image semantics in the synthesized regions and plausibly blend them with the preserved regions—the objective SC models are actually trained for. Specifically, we use PSPNet [48] and compute the mean intersection-over-union (IoU) between the label map obtained for the decompressed validation images and the ground truth label map. For reference we also report this metric for baselines that do not use semantic label maps for training and/or deployment.

6. Results

6.1. Generative compression

Fig. 5 shows the mean preference percentage obtained by our GC models compared to BPG at different rates, on the Kodak and the RAISE1K data set. In addition, we report the mean preference percentage for GC models compared to BPG and AEDC on Cityscapes. Example validation images for side-by-side comparison of our method with BPG for images from the Kodak, RAISE1K, and Cityscapes data set can be found in Figs. 1, 4, and 3, respectively. Furthermore, we perform extensive visual comparisons of all our methods and the baselines, presented in Appendix F.

Our GC models with $C = 4$ are preferred to BPG even when images produced by BPG use 95% and 124% more bits than those produced by our models for Kodak and RAISE1K, respectively. Notably this is achieved even though there is a distribution shift between the training and testing set (recall that these GC models are trained on the Open Images data set). The gains of domain-specificity and semantic label maps (for training) becomes apparent from the results on Cityscapes: Our GC models with $C = 2$ are preferred to BPG even when the latter uses 181% more bits. For $C = 4$ the gains on Cityscapes are comparable to those obtained for GC on RAISE1K. For all three data sets, BPG requires between 21% and 49% more bits than our GC models with $C = 8$.

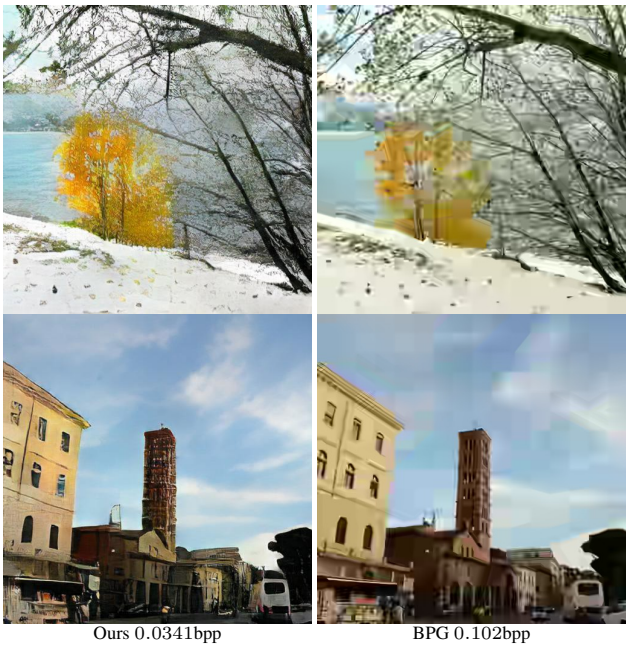


Figure 4. Visual example of an image from RAISE1k produced by our GC network with $C = 4$ compared to BPG.

Discussion: The GC models produce images with much finer detail than BPG, which suffers from smoothed patches and blocking artifacts. In particular, the GC models convincingly reconstruct texture in natural objects such as trees, water, and sky, and is most challenged with scenes involving humans. AEDC and the MSE baseline both produce blurry images.

We see that the gains of our models are maximal at extreme bitrates, with BPG needing 95–181% more bits for the $C = 2, 4$ models on the three data sets. For $C = 8$ gains are smaller but still very large (BPG needing 21–49% more bits). This is expected, since as the bitrate increases the classical compression measures (PSNR/MS-SSIM) become more meaningful—and our system does not employ the full complexity of current state-of-the-art systems:

We give an overview of relevant recent learned compression methods and their differences to our GC method and BPG in Table 1 in Appendix A, where we see that BPG is

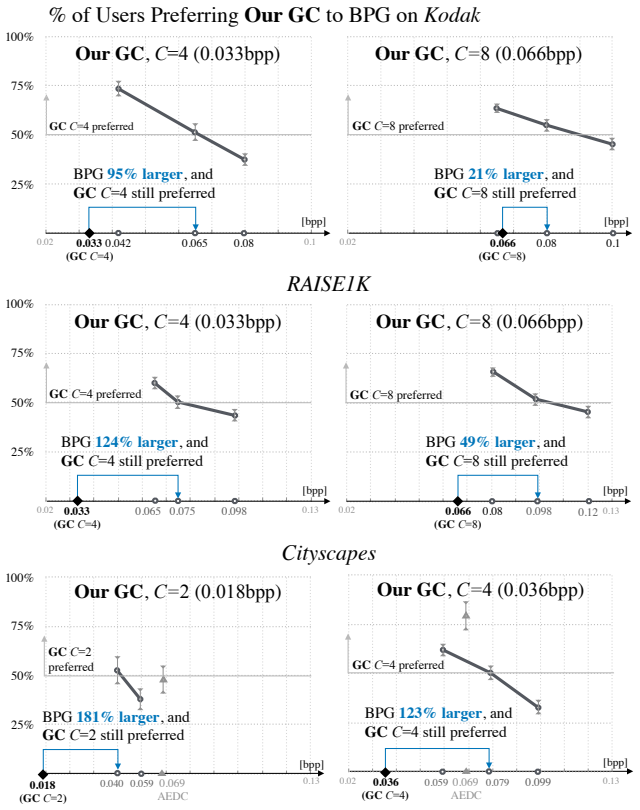


Figure 5. User study results evaluating our GC models on Kodak, RAISE1K and Cityscapes. Each plot corresponds to one of our models. The bitrate of that model is highlighted on the x-axis with a black diamond. The thick gray line shows the percentage of users preferring our model to BPG at that bitrate (bpp). The blue arrow points from our model to the highest-bitrate BPG operating point where more than 50% of users prefer ours, visualizing how many more bits BPG uses at that point. For Kodak and RAISE1K, we use GC models trained on Open Images, without any semantic label maps. For Cityscapes, we used GC (D^+) (using semantic label maps only for D and only during training), and we additionally compared to the AEDC baseline (MS-SSIM optimized).

$\mathcal{U}$ (Open Images) $\mathcal{U}$ (Cityscapes) WGAN-GP (Cityscapes)

Figure 6. Sampling codes $\hat{\boldsymbol{w}}$ uniformly ($\mathcal{U}$, left), and generating them with a WGAN-GP (right).

still visually competitive with the current state-of-the-art.

Given the dramatic bitrate savings we achieve according to the user study (BPG needing 21–181% more bits), and the competitiveness of BPG to the most recent state-of-the-art [31], we conclude that our proposed system presents a

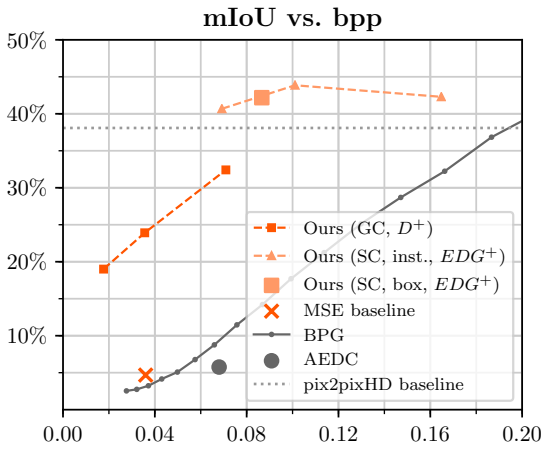


Figure 7. Mean IoU as a function of bpp on the Cityscapes validation set for our GC and SC networks, and for the MSE baseline. We show both SC modes: RI (inst.), RB (box). D^+ annotates models where instance semantic label maps are fed to the discriminator (only during training); EDG^+ indicates that semantic label maps are used both for training and deployment. The pix2pixHD baseline [44] was trained from scratch for 50 epochs, using the same downsampled 1024×512px training images as for our method.

significant step forward for visually pleasing compression at extreme bitrates.

Sampling the compressed representations: In Fig. 6 we explore the representation learned by our GC models (with $C = 4$), by sampling the (discrete) latent space of $\hat{\boldsymbol{w}}$. When we sample uniformly, and decode with our GC model into images, we obtain a "soup of image patches" which reflects the domain the models were trained on (e.g. street sign and building patches on Cityscapes). Note that we should not expect these outputs to look like normal images, since nothing forces the encoder output $\hat{\boldsymbol{w}}$ to be uniformly distributed over the discrete latent space.

However, given the low dimensionality of $\hat{\boldsymbol{w}}$ ($32 \times 64 \times 4$ for 512×1024px Cityscape images), it would be interesting to try to learn the true distribution. To this end, we perform a simple experiment and train an improved Wasserstein GAN (WGAN-GP) [14] on $\hat{\boldsymbol{w}}$ extracted from Cityscapes, using default parameters and a ResNet architecture (only adjusting the architecture to output $32 \times 64 \times 4$ tensors instead of $64 \times 64 \times 3$ RGB images). By feeding our GC model with samples from the WGAN-GP generator, we easily obtain a powerful generative model, which generates sharp 1024×512px images *from scratch*. We think this could be a promising direction for building high-resolution generative models. In Figs. 20–22 in the Appendix, we show more samples, and samples obtained by feeding the MSE baseline with uniform and learned code samples. The latter yields noisier "patch soups" and much blurrier image samples than our GC network.

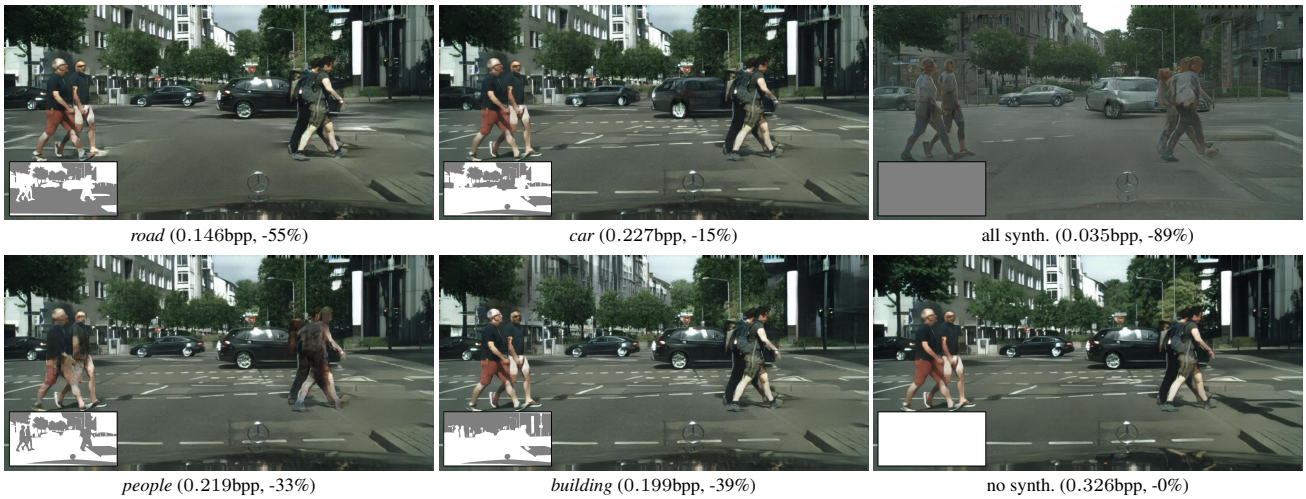


Figure 8. Synthesizing different classes using our SC network with $C = 8$. In each image except for *no synthesis*, we additionally synthesize the classes *vegetation, sky, sidewalk, ego vehicle, wall*. The heatmaps in the lower left corners show the synthesized parts in gray. We show the bpp of each image as well as the relative savings due to the selective generation.

6.2. Selective generative compression

Fig. 7 shows the mean IoU on the Cityscapes validation set as a function of bpp for SC networks with $C = 2, 4, 8$, along with the values obtained for the baselines. Additionally, we plot mean IoU for GC with semantic label maps fed to the discriminator (D^+), and the MSE baseline.

In Fig. 8 we present example Cityscapes validation images produced by the SC network trained in the RI mode with $C = 8$, where different semantic classes are preserved. More visual results for the SC networks trained on Cityscapes can be found in Appendix F.7, including results obtained for the RB operation mode and by using semantic label maps estimated from the input image via PSPNet [49].

Discussion: The quantitative evaluation of the semantic preservation capacity (Fig. 7) reveals that the SC networks preserve the semantics somewhat better than pix2pixHD, indicating that the SC networks faithfully generate texture from the label maps and plausibly combine generated with preserved image content. The mIoU of BPG, AEDC, and the MSE baseline is considerably lower than that obtained by our SC and GC models, which can arguably be attributed to blurring and blocking artifacts. However, it is not surprising as these baseline methods do not use label maps during training and prediction.

In the SC operation mode, our networks manage to seamlessly merge preserved and generated image content both when preserving object instances and boxes crossing object boundaries (see Appendix F.7). Further, our networks lead to reductions in bpp of 50% and more compared to the same networks without synthesis, while leaving the visual quality essentially unimpaired, when objects with repetitive structure are synthesized (such as trees, streets, and sky). In some cases, the visual quality is even better than that of BPG at the same bitrate. The visual quality of more complex synthesized objects (e.g. buildings, people) is worse. However, this is a limitation of current GAN technology rather than our approach. As the visual quality of GANs improves further, SC networks will as well. Notably, the SC networks can generate entire images from the semantic label map only.

Finally, the semantic label map, which requires 0.036 bpp on avg. for downscaled 1024×512px Cityscapes images, represents a relatively large overhead compared to the storage cost of the preserved image parts. This cost vanishes as the image size increases, since the semantic mask can be stored as an image dimension-independent vector graphic.

7. Conclusion

We proposed a GAN-based framework for learned generative compression, and presented the first thorough study of such a framework for full-resolution image compression. Our results show that for low bitrates, such generative compression (GC) can give dramatic bitrate savings compared to previous state-of-the-art methods optimized for classical objectives such as MS-SSIM and MSE, when evaluated in terms of visual quality in a user study. Furthermore, we demonstrated that constraining the application domain to street scene images leads to additional storage savings, and explored (for SC) selectively combining fully synthesized image contents with preserved ones when semantic label maps are available.

Interesting directions for future work are to develop a mechanism for controlling spatial allocation of bits for GC (e.g., to achieve better preservation of faces; possibly using semantic label maps), and to combine SC with saliency information to determine what regions to preserve.

Acknowledgments: This work was supported by the ETH Zurich General Fund, and an Nvidia GPU hardware grant.

References

[1] Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc van Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. In *Advances in Neural Information Processing Systems*, pages 1141–1151, 2017. 2, 3, 5

[2] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017. 2

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*, pages 214–223, 2017. 2

[4] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimization of nonlinear transform codes for perceptual quality. *Picture Coding Symposium (PCS)*, 2016. 2

[5] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. In *International Conference on Learning Representations (ICLR)*, 2017. 1, 2

[6] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 2, 5

[7] Fabrice Bellard. BPG Image format. https://bellard.org/bpg/. 1, 2, 5, 12

[8] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. 1, 4

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 2, 5, 12

[10] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012. 3

[11] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. Raise: a raw images dataset for digital image forensics. In *Proceedings of the ACM Multimedia Systems Conference*, pages 219–224. ACM, 2015. 2, 5, 12

[12] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. Deep generative adversarial compression artifact removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4826–4835, 2017. 3

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 1, 2, 3

[14] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017. 7, 24

[15] Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing*, 19(1):185–198, 2010. 3

[16] Rupesh Gupta, Meera Thapar Khanna, and Santanu Chaudhury. Visual saliency guided video compression algorithm. *Signal Processing: Image Communication*, 28(9):1006–1022, 2013. 3

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. IEEE, 2017. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 12

[19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015. 12

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 1, 2, 5, 6, 12

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. 5

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 12

[24] Kod. Kodak PhotoCD dataset. http://r0k.us/graphics/kodak/. 2, 5, 12

[25] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017. 5, 12

[26] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 3

[27] Mu Li, Wangmeng Zuo, Shuhang Gu, Debin Zhao, and David Zhang. Learning convolutional networks for content-weighted image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3214–3223, 2018. 2

[28] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017. 2

[29] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821. IEEE, 2017. 2, 3

[30] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Conditional probability models for deep image compression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 5, 12

[31] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems*, pages 10771–10780, 2018. 2, 7, 12, 13, 19, 20, 21, 22

[32] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2, 3

[33] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016. 3

[34] Oren Rippel and Lubomir Bourdev. Real-time adaptive image compression. In *Proceedings of the International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 2922–2930, International Convention Centre, Sydney, Australia, 2017. 1, 2, 12, 13, 17, 18

[35] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016. 2

[36] Shibani Santurkar, David Budden, and Nir Shavit. Generative compression. *Picture Coding Symposium (PCS)*, 2018. 3

[37] X Yu Stella and Dimitri A Lisin. Image compression based on visual saliency at individual scales. In *International Symposium on Visual Computing*, pages 157–166. Springer, 2009. 3

[38] David S. Taubman and Michael W. Marcellin. *JPEG 2000: Image Compression Fundamentals, Standards and Practice*. Kluwer Academic Publishers, Norwell, MA, USA, 2001. 1

[39] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszar. Lossy image compression with compressive autoencoders. In *International Conference on Learning Representations (ICLR)*, 2017. 1, 2

[40] George Toderici, Sean M O'Malley, Sung Jin Hwang, Damien Vincent, David Minnen, Shumeet Baluja, Michele Covell, and Rahul Sukthankar. Variable rate image compression with recurrent neural networks. *arXiv preprint arXiv:1511.06085*, 2015. 2

[41] George Toderici, Damien Vincent, Nick Johnston, Sung Jin Hwang, David Minnen, Joel Shor, and Michele Covell. Full resolution image compression with recurrent neural networks. *arXiv preprint arXiv:1608.05148*, 2016. 2

[42] Robert Torfason, Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Towards image understanding from deep compression without decoding. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 2

[43] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016. 12

[44] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 5, 6, 7, 13

[45] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, Nov 2003. 1, 2

[46] Web. WebP Image format. https://developers.google.com/speed/webp/. 1

[47] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5907–5915, 2017. 2

[48] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid Scene Parsing Network. *ArXiv e-prints*, Dec. 2016. 6

[49] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 8, 26

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2223–2232, 2017. 2, 5, 12

A. Comparison with State-of-the-art

We give an overview of relevant recent learned compression methods and their differences to our GC method and BPG in Table 1. [34] were state-of-the-art in MS-SSIM in 2017, while work [31] is the current state-of-the-art in image compression in terms of classical metrics (PSNR and MS-SSIM) when measured on the Kodak data set [24]. Notably, all methods except ours (BPG, Rippel et al., and Minnen et al.) employ adaptive arithmetic coding using context models for improved compression performance. Such models could also be implemented for our system, and have led to additional savings of 10% in [30]. Since Rippel et al. and Minnen et al. have only released a selection of their decoded images (for 3 and 4, respectively, out of the 24 Kodak images), and at significantly higher bitrates, a comparison with a user study is not meaningful. Instead, we try to qualitatively put our results into context with theirs.

In Figs. 13–15 in Sec. F.4, we compare qualitatively to [34]. We can observe that even though Rippel *et al*. [34] use 29–179% more bits, our models produce images of comparable or better quality.

In Figs. 16–19 in Sec. F.5, we show a qualitative comparison of our results to the images provided by the work of [31], as well as to BPG [7] on those images. First, we see that BPG is still visually competitive with the current state-of-the-art, which is consistent with moderate 8.41% bitrate savings being reported by [31] in terms of PSNR. Second, even though we use much fewer bits compared to the example images available from [31], for some of them (Figs. 16 and 17) our method can still produce images of comparable visual quality.

B. Training Details

We employ the ADAM optimizer [23] with a learning rate of 0.0002 and set the mini-batch size to 1. Our networks are trained for 150000 iterations on Cityscapes and for 280000 iterations on Open Images. For normalization we used instance normalization [43], except in the second half of the Open Images training, we train the generator/decoder with fixed batch statistics (as implemented in the test mode of batch normalization [19]), since we found this reduced artifacts and color shift.

C. Data set and Preprocessing Details

To train GC models (which do not require semantic label maps, neither during training nor for deployment) for compression of diverse natural images, we use 200k images sampled randomly from the *Open Images* data set [25] (9M images). The training images are rescaled so that the longer side has length 768px, and images for which rescaling does not result in at least $1.25\times$ downscaling as well as high saturation images (average S > 0.9 or V > 0.8 in HSV color space) are discarded (resulting in an effective training set size of 188k).

We evaluate these models on the Kodak image compression data set [24] (24 images, 768×512px), which has a long tradition in the image compression literature and is still the most frequently used data set for comparisons of learned image compression methods. Additionally, we evaluate our GC models on 20 randomly selected images from the *RAISE1K* data set [11], a real-world image data set consisting of 8156 high-resolution RAW images (we rescale the images such that the longer side has length 768px). To investigate the benefits of having a somewhat constrained application domain and semantic labels at training time, we also train GC models with semantic label maps on the *Cityscapes* data set [9] (2975 training and 500 validation images, 34 classes, 2048×1024px resolution) consisting of street scene images and evaluate it on 20 randomly selected validation images (without semantic labels). Both training and validation images are rescaled to 1024×512px resolution.

To evaluate the proposed SC method (which requires semantic label maps for training and deployment) we again rely on the Cityscapes data set. Cityscapes was previously used to generate images form semantic label maps using GANs [20, 50]. The preprocessing for SC is the same as for GC.

D. Compression Details

We compress the semantic label map for SC by quantizing the coordinates in the vector graphic to the image grid and encoding coordinates relative to preceding coordinates when traversing object boundaries (rather than relative to the image frame). The so-obtained bitstream is then compressed using arithmetic coding.

To ensure fair comparison, we do not count header sizes for any of the baseline methods throughout.

E. Architecture Details

For the GC, the encoder E convolutionally processes the image $\boldsymbol{x}$ and optionally the label map $\boldsymbol{s}$, with spatial dimension $W \times H$, into a feature map of size ${}^{W}\!/_{16} \times {}^{H}\!/_{16} \times 960$ (with 6 layers, of which four have 2-strided convolutions), which is then projected down to C channels (where $C \in \{2, 4, 8\}$ is much smaller than 960). This results in a feature map $\boldsymbol{w}$ of dimension ${}^{W}\!/_{16} \times {}^{H}\!/_{16} \times C$, which is quantized over L centers to obtain the discrete $\hat{\boldsymbol{w}}$. The generator G projects $\hat{\boldsymbol{w}}$ up to 960 channels, processes these with 9 residual units [18] at dimension ${}^{W}\!/_{16} \times {}^{H}\!/_{16} \times 960$, and then mirrors E by convolutionally processing the features back to spatial dimensions $W \times H$ (with transposed convolutions instead of strided ones).

Similar to E, the feature extractor F for SC processes the semantic map $\boldsymbol{s}$ down to the spatial dimension of $\hat{\boldsymbol{w}}$,

	BPG	Rippel et al. (2017)	Minnen et al. (2018)	Ours (GC)
Learned	No	Yes	Yes	Yes
Arithmetic encoding	Adaptive	Adaptive	Adaptive	Static
Context model	CABAC	Autoregressive	Autoregressive	None
Visualized bitrates [bpp][4]	all[5]	0.08–	0.12–	0.033–0.066
GAN	No	Non-standard	No	f-div. based
S.o.t.a. in MS-SSIM	No	No	Yes	No
S.o.t.a. in PSNR	No	No	Yes	No
Savings to BPG in PSNR			8.41%	
Savings to BPG in User Study				17.2–48.7%

Table 1. Overview of differences between [31] (s.o.t.a. in MS-SSIM and PSNR), to BPG (previous s.o.t.a. in PSNR) and [34] (s.o.t.a. in MS-SSIM in 2017, also used GANs).

which is then concatenated to $\hat{\boldsymbol{w}}$ for generation. In this case, we consider slightly higher bitrates and downscale by $8\times$ instead of $16\times$ in the encoder E, such that $\dim(\hat{\boldsymbol{w}}) = {}^{W}/_{8} \times {}^{H}/_{8} \times C$. The generator then first processes $\hat{\boldsymbol{w}}$ down to ${}^{W}/_{16} \times {}^{H}/_{16} \times 960$ and then proceeds as for GC.

For both GC and SC, we use the multi-scale architecture of [44] for the discriminator D, which measures the divergence between $p_{\boldsymbol{x}}$ and $p_{G(\boldsymbol{z})}$ both locally and globally.

We adopt the notation from [44] to describe our encoder and generator/decoder architectures and additionally use `q` to denote the quantization layer (see Sec. 3 for details). The output of `q` is encoded and stored.

- **Encoder GC:** `c7s1-60, d120, d240, d480, d960, c3s1-`C`, q`
- **Encoders SC:**
  - Semantic label map encoder: `c7s1-60, d120, d240, d480, d960`
  - Image encoder: `c7s1-60, d120, d240, d480, c3s1-`C`, q, c3s1-480, d960`

  The outputs of the semantic label map encoder and the image encoder are concatenated and fed to the generator/decoder.
- **Generator/decoder:** `c3s1-960, R960, R960, R960, R960, R960, R960, R960, R960, u480, u240, u120, u60, c7s1-3`

F. Visuals

In the following Sections, F.1, F.2, F.3, we show the first five images of each of the three data sets we used for the user study, next to the outputs of BPG at similar bitrates.

Secs. F.4 and F.5 provide visual comparisons of our GC models with [34] and [31], respectively, on a subset of images form the Kodak data set.

In Sec. F.6, we show visualizations of the latent representation of our GC models.

Finally, Sec. F.7 presents additional visual results for SC.

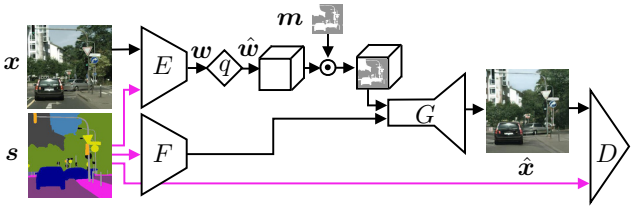


Figure 9. Structure of the proposed SC network. E is the encoder for the image $\boldsymbol{x}$ and the semantic label map $\boldsymbol{s}$. q quantizes the latent code $\boldsymbol{w}$ to $\hat{\boldsymbol{w}}$. The subsampled heatmap multiplies $\hat{\boldsymbol{w}}$ (pointwise) for spatial bit allocation. G is the generator/decoder, producing the decompressed image $\hat{\boldsymbol{x}}$, and D is the discriminator used for adversarial training. F extracts features from $\boldsymbol{s}$.

F.1. Generative Compression on Kodak

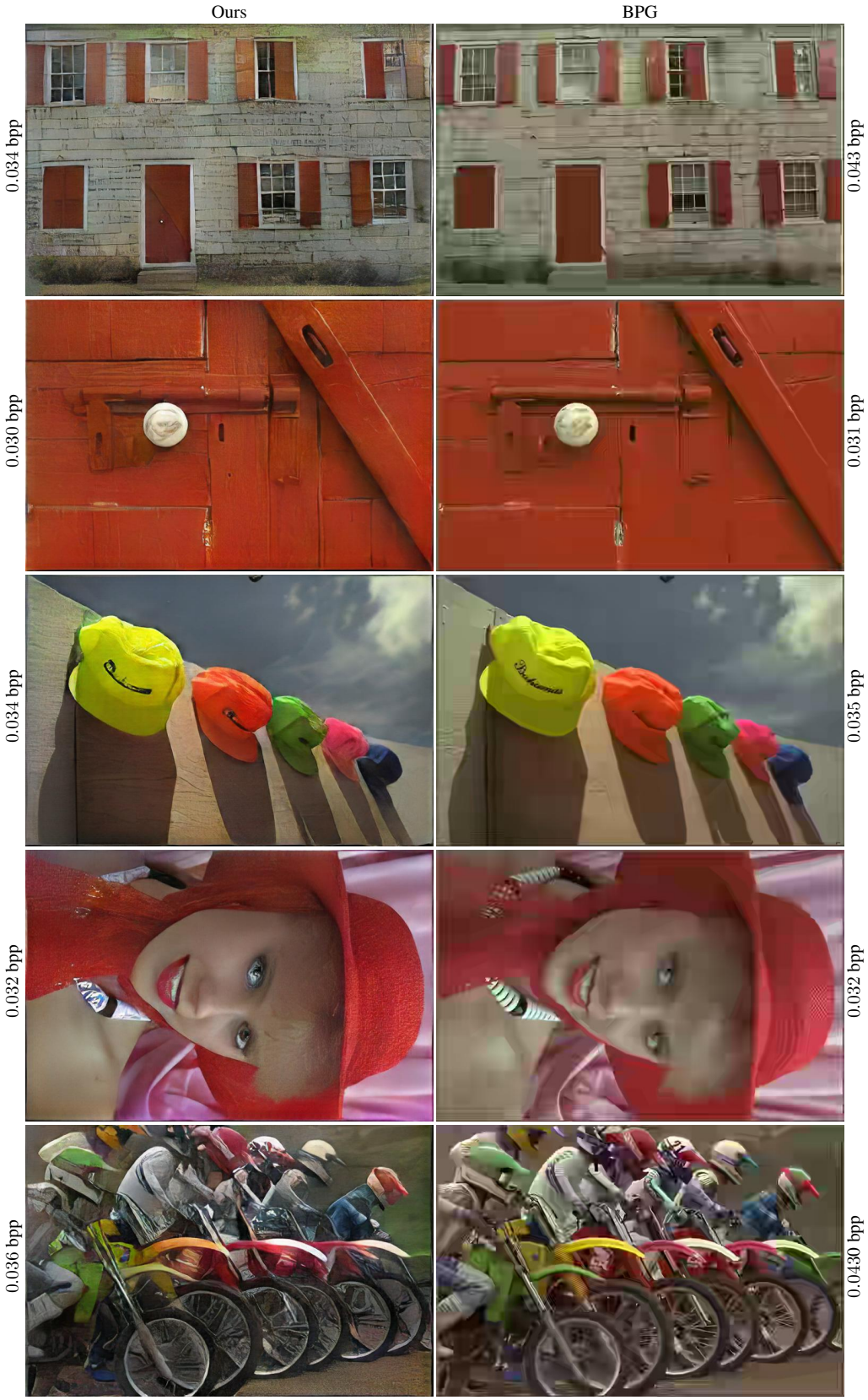


Figure 10. First 5 images of the Kodak data set, produced by our GC model with $C = 4$ and BPG.

F.2. Generative Compression on RAISE1k

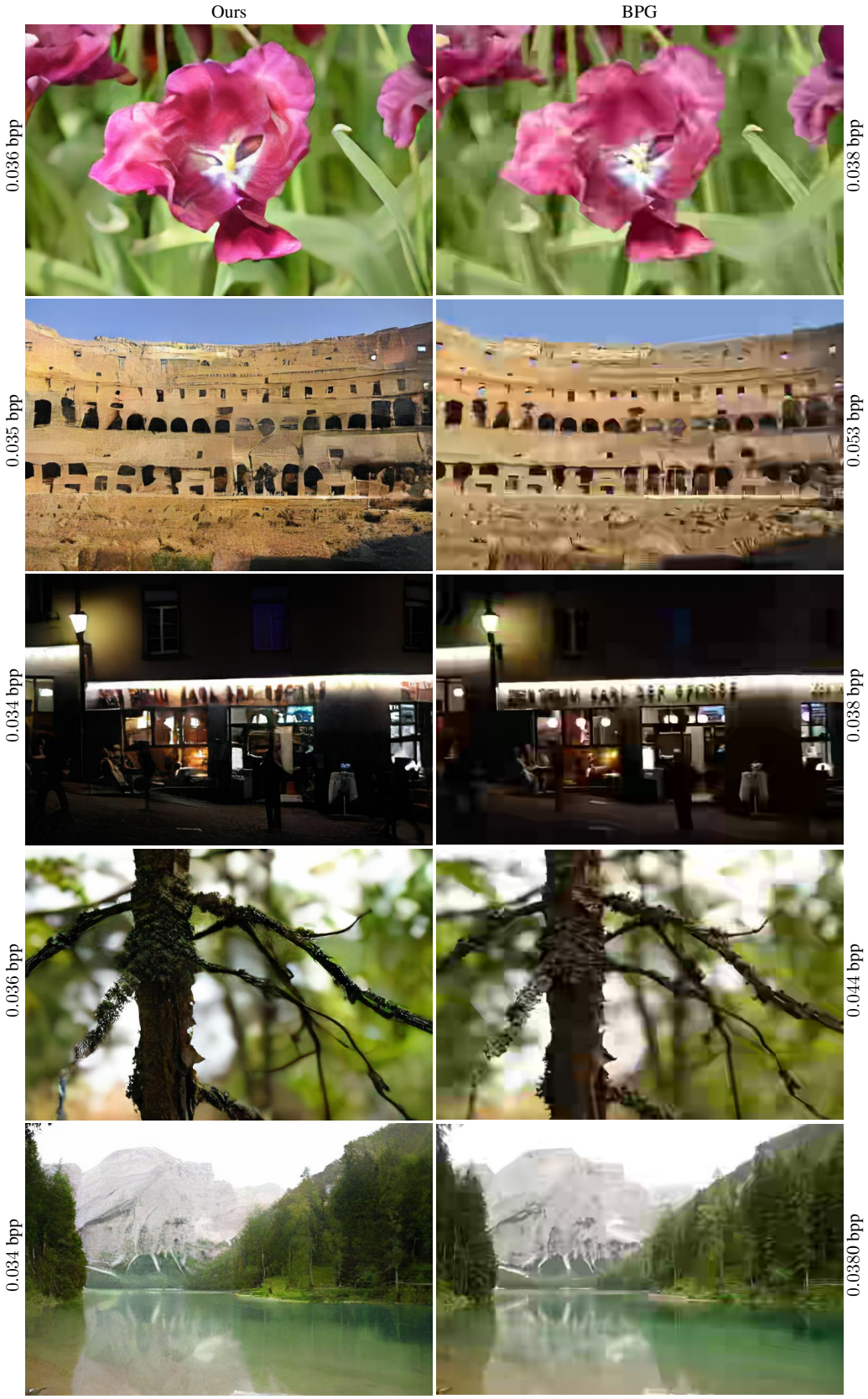


Figure 11. First 5 images of RAISE1k, produced by our GC model with $C = 4$ and BPG.

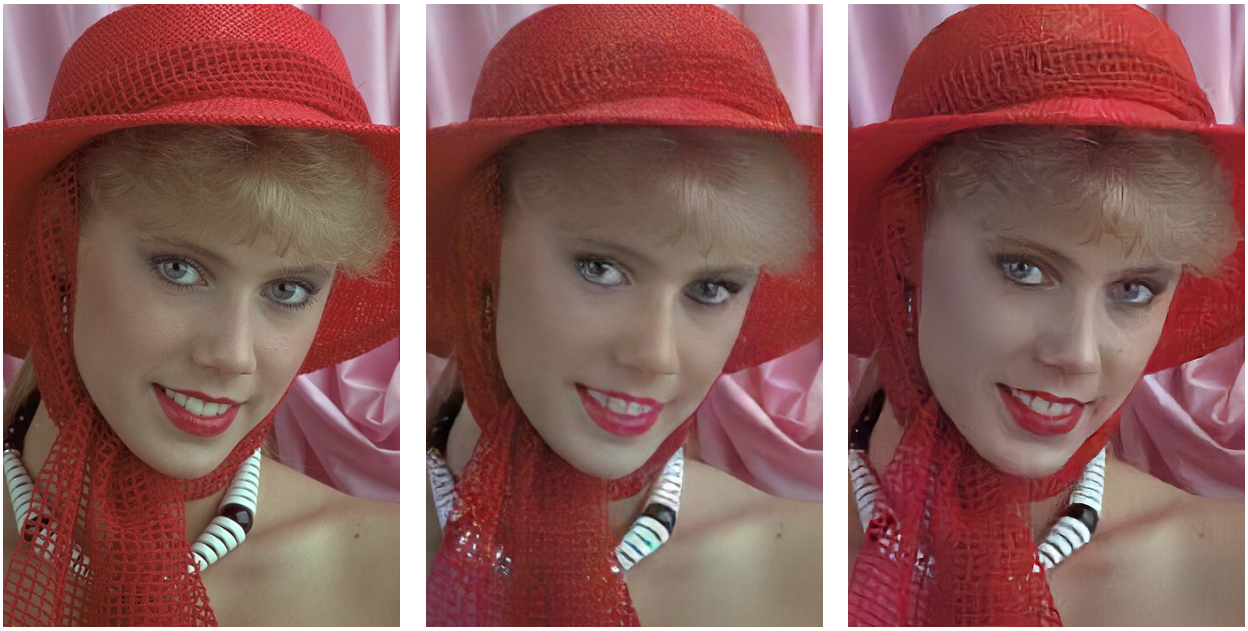
F.3. Generative Compression on Cityscapes

Figure 12. First 5 images of Cityscapes, produced by our GC model with $C = 4$ and BPG.

F.4. Comparison with [34]

Original | Ours, 0.0304bpp | Rippel et al., 0.0828bpp (+*172%*)

Figure 13. Our model loses more texture but has less artifacts on the knob. Overall, it looks comparable to the output of [34], using significantly fewer bits.



Original | Ours, 0.0651bpp | Rippel et al., 0.0840bpp (+*29%*)

Figure 14. Notice that compared to [34], our model produces smoother lines at the jaw and a smoother hat, but provides a worse reconstruction of the eye.

Original | Ours, 0.0668bpp | Rippel et al., 0.0928bpp (+*39%*)

Figure 15. Notice that our model produces much better sky and grass textures than [34], and also preserves the texture of the light tower more faithfully.

F.5. Comparison with [31]

Original

Ours, 0.0668bpp

Minnen et al., 0.221bpp *230% larger*

BPG, 0.227bpp

Figure 16. Notice that our model yields sharper grass and sky, but a worse reconstruction of the fence and the lighthouse compared to [31]. Compared to BPG, Minnen et al. produces blurrier grass, sky and lighthouse but BPG suffers from ringing artifacts on the roof of the second building and the top of the lighthouse.

Original

Ours, 0.0685bpp

Minnen et al., 0.155bpp, *127% larger*

BPG, 0.164bpp

Figure 17. Our model produces an overall sharper face compared to [31], but the texture on the cloth deviates more from the original. Compared to BPG, Minnen et al. has a less blurry face and fewer artifacts on the cheek.

Original

Ours, 0.0328bpp

Minnen et al., 0.246bpp, *651% larger*

BPG, 0.248bpp

Figure 18. Here we obtain a significantly worse reconstruction than [31] and BPG, but use only a fraction of the bits. Between BPG and Minnen et al., it is hard to see any differences.

Original

Ours, 0.03418bpp

Minnen et al., 0.123bpp, *259% larger*,

BPG, 0.119bpp

Figure 19. Here we obtain a significantly worse reconstruction compared to [31] and BPG, but use only a fraction of the bits. Compared to BPG, Minnen et al.has a smoother background but less texture on the birds.

F.6. Sampling the compressed representations

Figure 20. We uniformly sample codes from the (discrete) latent space $\hat{\boldsymbol{w}}$ of our generative compression models (GC with $C = 4$) trained on Cityscapes and Open Images. The Cityscapes model outputs domain specific patches (street signs, buildings, trees, road), whereas the Open Images samples are more colorful and consist of more generic visual patches.

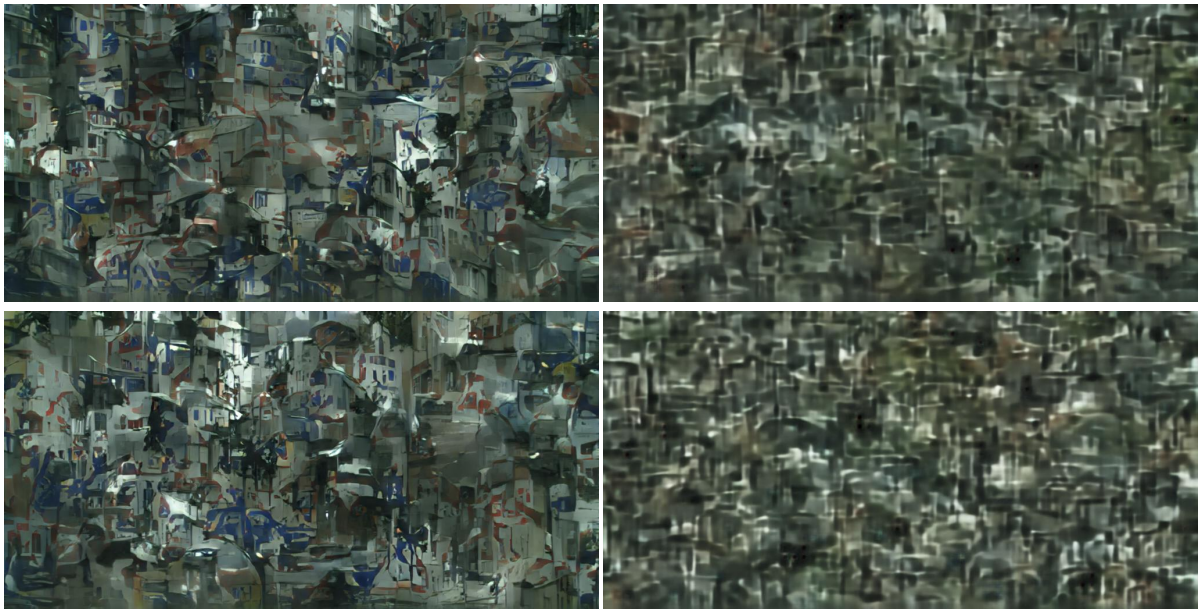


Figure 21. We train the same architecture with $C = 4$ for MSE and for generative compression on Cityscapes. When uniformly sampling the (discrete) latent space $\hat{\boldsymbol{w}}$ of the models, we see stark differences between the decoded images $G(\hat{\boldsymbol{w}})$. The GC model produces patches that resemble parts of Cityscapes images (street signs, buildings, etc.), whereas the MSE model outputs looks like low-frequency noise.

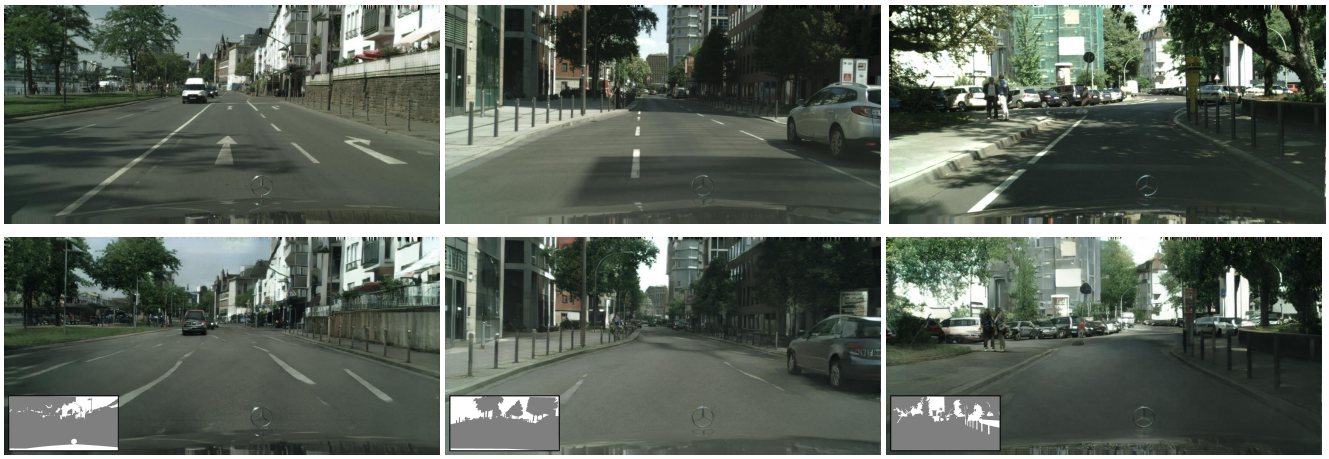
GC model with $C = 4$ MSE baseline model with $C = 4$

Figure 22. We experiment with learning the distribution of $\hat{\boldsymbol{w}} = E(\boldsymbol{x})$ by training an improved Wasserstein GAN [14]. When sampling form the decoder/generator G of our model by feeding it with samples from the improved WGAN generator, we obtain much sharper images than when we do the same with an MSE model.

F.7. Selective Compression on Cityscapes

Figure 23. Synthesizing different classes for two different images from Cityscapes, using our SC network with $C = 4$. In each image except for *no synthesis*, we additionally synthesize the classes *vegetation, sky, sidewalk, ego vehicle, wall*.

Figure 24. Example images obtained by our SC network ($C = 8$) preserving a box and synthesizing the rest of the image, on Cityscapes. The SC network seamlessly merges preserved and generated image content even in places where the box crosses object boundaries.



0.019 bpp 0.021 bpp 0.013 bpp

Figure 25. Reconstructions obtained by our SC network using semantic label maps estimated from the input image via PSPNet [49].