# Towards super resolution in the compressed domain of learning-based image codecs

Evgeniy Upenik, Michela Testolina, and Touradj Ebrahimi

Multimedia Signal Processing Group (MMSPG)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland

## ABSTRACT

Learning-based image coding has shown promising results during recent years. Unlike the traditional approaches to image compression, learning-based codecs exploit deep neural networks for reducing dimensionality of the input at the stage where a linear transform would be typically applied previously. The signal representation after this stage, called latent space, carries the information in such a way that it can be interpreted by other deep neural networks without the need of decoding it. One of the tasks that can benefit from the above-mentioned possibility is super resolution. In this paper, we explore the possibilities and propose an approach for super resolution that is applied in the latent space. We focus on the fixed compression model, where the encoder part of the network is frozen and an enhanced decoder is learned. Additionally, we assess the performance of the proposed approach.

**Keywords:** Image processing, super resolution, learning-based image compression, deep learning

## 1. INTRODUCTION

Learning-based image coding has shown promising results during recent years. Unlike the traditional approaches for image compression, learning-based codecs exploit deep neural networks for reducing dimensionality of the input at the stage where a linear transform would be typically applied previously. The signal representation after this stage, called latent space, carries the information in such a way that it can be interpreted by other deep neural networks without the need of decoding. That is to say, various image processing tasks can be performed in the compressed domain of leaning-based image codecs. One of the tasks that can benefit from the above-mentioned approach is super resolution. It has been shown that super resolution techniques based on deep learning significantly outperform the deterministic interpolation algorithms.

In this paper, we explore different scenarios and propose an approach for super resolution that is applied in the latent space. There exist two types of architectures: fixed compression model and enhanced compression model. In the former case, the encoder part of the network is frozen and an enhanced decoder is learned. This makes it easier to train for many different tasks (one at a time) at the expense of a latent representation that is optimized only for the fidelity of its visual reconstruction when compared to the input image. In the latter case, all modules are trained together in an end-to-end fashion; this may result in a more flexible latent representation that benefits the entire network. In this paper, we investigate the first approach.

Finally, we assess the performance of the proposed approach. Two anchors are used for benchmarking: 1) original high-fidelity images are downscaled for the experiment and then used for evaluating the results of super resolution; 2) super resolution is applied to reconstructed images and the result is used for evaluation of super resolution applied in the compressed domain.

## 2. RELATED WORK

In this section, we provide information about the state of the art on related topics of single image super resolution and end-to-end learning-based compression.

Further author information: (Send correspondence to T.E.)
T.E.: E-mail: touradj.ebrahimi@epfl.ch

## 2.1 Single image super resolution

Super resolution is a category of techniques and methods for up-scaling raster images by a factor of two or more. Single image super resolution only takes into account one isolated image. Thus, as opposed to multi-view or video super resolution, it cannot benefit from correlation between subsequent images in order to improve visual quality of the result. Super resolution, in general, is the evolution of image re-sampling methods, such as bilinear, bicubic, and Lanczos filtering, with the latter considered to be the best among the conventional methods.

In recent years, thanks to the advancements in the field of deep learning, super resolution methods have achieved impressive performance in terms of visual quality for up-scaling by factors of four and higher. Further in this subsection, we will review a number of learning-based super resolution methods that will be used in our work.

- **Photo-realistic single image super-resolution using a generative adversarial network (SRGAN)**[1] by Ledig et al. is a super resolution model that uses GAN[2] with deep residual networks that diverge from using Mean Square Error (MSE) as the main optimization target. It differs from previous super resolution methods as it defines a new perceptual loss using high level feature maps of the VGG[3] network. Traditional GANs, as first defined by Goodfellow in 2014,[2] take random noise as an input to the generator. In SRGAN the generator accepts a lower resolution image as an input to GAN. The discriminator, however, operates in a traditional way. The main difference is in the loss function, which, rather than optimizing the MSE between the generated image and the original high resolution image, minimizes the euclidean distance of the feature representations of the reconstructed image and original image obtained from the pre-trained VGG19 network. This results in generated images that are more faithful to a natural manifold rather than to a pixel wise comparison.

- **Enhanced deep residual networks for single image super-resolution (EDSR)**[4] by Lim et al. is a super resolution residual model, scoring first and second place at the NTIRE 2017 competition. It is based on SRResNet[1] with an improved architecture for faster computing and better performance results. The main difference between this architecture and the previous SRResNet, is the increase in the number of feature channels of the convolutional layers. The main idea behind this method is that in a general convolutional neural network architecture the memory occupied is of complexity $BF$ ($B$ being the number of layers, $F$ being the feature channels) while parameters have a complexity of $O(BF^2)$. Therefore, increasing the features rather than the layers can increase the capacity with less computational resources. Another change when compared to SRResNet is the deletion of ReLU activation layers outside the residual blocks.

- **Wide activation for efficient and accurate image super-resolution (WDSR)**[5] by Yu et al. is a super resolution residual model, scoring first place at the NTIRE 2018 competition, based on EDSR[4] with an improved architecture for faster computing and better performance results. The main differences are the following: the number of convolutional filters is reduced to 32 in the residual blocks, while the first layer of this same block is wider by a factor of 2 or 4. We can also see that an additional branch is added to the network, because the convolutional layers outside of the residual blocks are computationally expensive. Thus, a single convolution layer is extracted from the low resolution image and passes through an additional convolution with kernel of size 5, up-sampled and added at the end of the network.

- **Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN)**[6] by Wang et al. obserevd that in the SRGAN[1] architecture unrealistic visual artifacts, sometimes referred as hallucinations, may be very annoying. In order to enhance the visual quality, the authors improved the network architecture, adversarial loss and perceptual loss. They introduced a Residual-in-Residual Dense Block without batch normalization as the basic network building unit. The perceptual loss is alternated by using the features before the activation, providing stronger brightness consistency and texture supervision.

## 2.2 Learning-based image compression

Learning-based image methods for end-to-end coding have emerged as powerful tools in the context of image compression, and are able, in some cases, to outperform the conventional methods.[7] Different approaches for such problem have been explored in the state of the art:

- Among the first works, a compression framework based on Recurrent Neural Networks (RNN) was proposed by Toderici et al.[8,9] for thumbnails and image compression.
- A different approach, adopting autoencoder architectures employing convolutional neural networks (CNNs), has been explored in many recent works.[10–13]
- More recently, methods that take advantage of GANs[2] have been explored[14,15] to generate images with a higher level of details.

Among the cited methods, the architecture proposed by Ballè et al.[12] lately became well-known in the learning-based image compression research community, and is even considered a groundbreaking work for its remarkable performance in terms of visual quality of the decompressed images. For this reason, this architecture has been selected as a base end-to-end codec in this work.

## 3. SUPER RESOLUTION IN COMPRESSION SCENARIO

Nowadays, almost all the images that are captured by modern cameras, disseminated over communication networks, or stored, are compressed with lossy codecs, at the cost of reducing their visual quality. Moreover, an additional step of decompression is typically required in order to perform image processing tasks, e.g. super resolution, on such images. Thus, even though our goal is to investigate compressed domain super resolution, in this section, we establish and prepare a benchmark for assessing the performance of super resolution in a scenario where lossy compression is a part of the pipeline for pixel-domain image processing.

### 3.1 Compression scenario benchmark and anchors

For the purpose of benchmarking super resolution in a compression workflow scenario, we propose to establish two anchors:

1. **Original anchor**: Super resolution is applied to the original high-resolution images, before any compression, and the performance is assessed by comparing the results of the super resolution with the corresponding high-resolution originals using PSNR and MS-SSIM objective visual quality metrics.
2. **Decoded anchor**: Original high-resolution images are down-scaled by a factor of four using bicubic interpolation. Then, the resulting low-resolution samples are compressed using the selected learning-based codec (bmshj2018-hyperprior[12]). As a next step, the super resolution x4 is applied to fully decoded images in the pixel domain. Finally, the results of the super resolution are compared to the corresponding high-resolution originals using PSNR and MS-SSIM objective visual quality metrics.

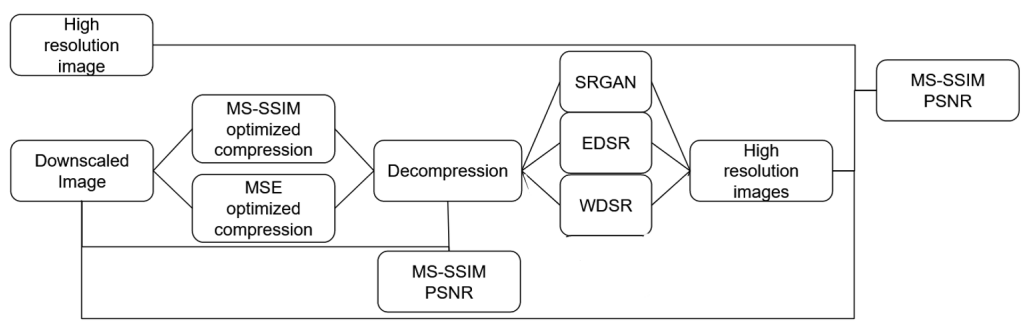



Figure 1. Workflow of the evaluation of super resolution methods in a compression scenario.

Figure 1 shows the workflow of the evaluation of super resolution methods in a compression scenario for the Decoded Anchor described in this subsection. The *Downscaled image* first undergoes a step of lossy compression performed by means of the variational image compression with a scale hyperprior[12] optimized for both MSE

and MS-SSIM, followed by a step of decompression in order to reconstruct the image in the pixel domain. Then four different super resolution methods, namely, SRGAN, EDSR, and WDSR, are applied to the resulting decompressed images with an up-scaling factor of four. Finally, the results of super resolution are objectively compared to the original *high resolution image*.

In our experiment, a subset of five images from the DIV2K dataset was used for benchmarking. One can find the selected images in Figure 2.

### 3.2 Results of benchmarking


Figure 2. Five images from the DIV2K dataset used for evaluation of selected super resolution methods in a compression scenario.

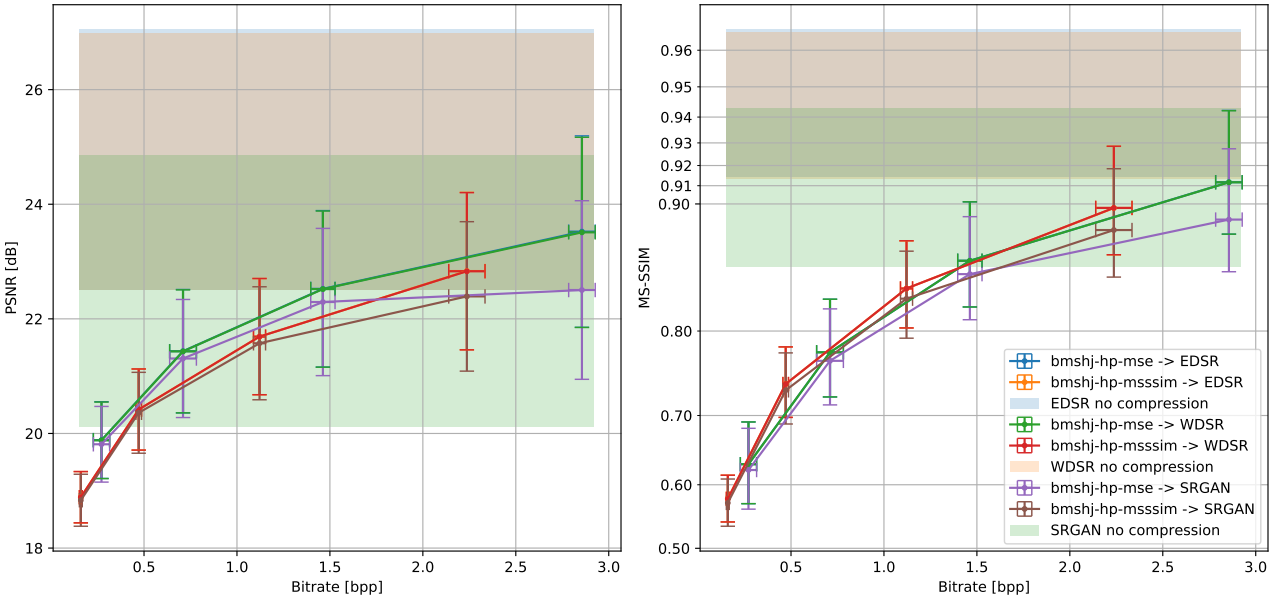



Figure 3. Average rate-distortion plots of the anchors for the PSNR and MS-SSIM .

Figure 3 presents the results of the benchmarking of the state-of-the-art super resolution methods. The plots on the left and on the right show, respectively, the average PSNR and MS-SSIM values for five images from the DIV2K dataset (Figure 2) compressed at different bitrates. The vertical error bars on both plots indicate the standard deviations of the quality metric values. The horizontal error bars represent a standard deviation from the target bitrates among different images. The areas filled with transparent colors show the standard deviation for the Original Anchor, i.e. when super resolution is applied to uncompressed images.

Additional plots for each image can be found in the Appendix A.

## 4. COMPRESSED-DOMAIN SUPER RESOLUTION

As it was already mentioned earlier in this paper, compressed domain image processing in general and super resolution in particular may improve computational complexity and possible visual quality in many modern

imaging workflows. In this, section we propose an adaptation of a state-of-the-art super resolution method that allows performing this image processing task in the compressed domain of a learning-based image codec by applying the processing directly to the latent representation of an autoencoder.

### 4.1 Procedure used for compressed domain super resolution

A typical learning-based image codec consists of an autoencoder, possibly quantization step, and an entropy codec. Autoencoder, here, plays the role of a non-linear transform as opposed to a linear transforms, such as Discrete Cosine Transform or Wavelet Transform, used in hand-engineered compression algorithms e.g. JPEG or JPEG 2000.

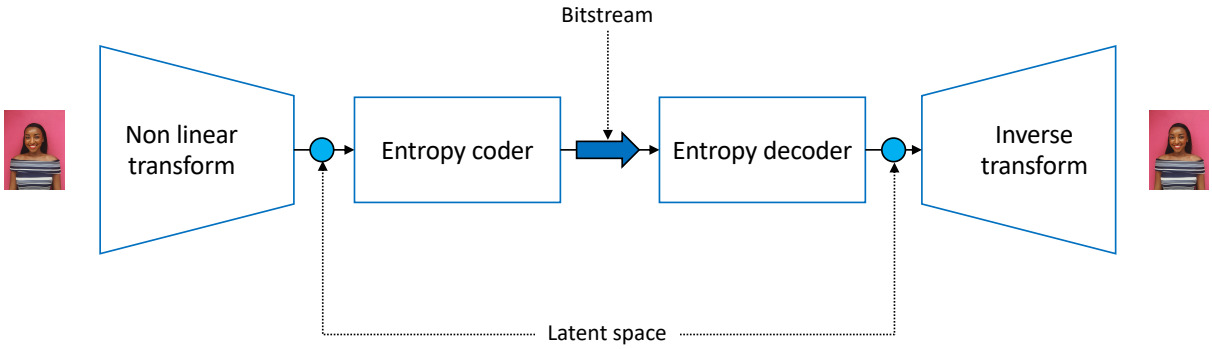



Figure 4. Typical learning-based image compression codec. Note: for the sake of simplicity the quantization step is not shown.

Figure 4 presents block diagram of a typical learning-based image compression codec. For the sake of simplicity the quantization step is not shown in this figure. The data between the entropy encoder and decoder is called a bitstream. The bitstream is the information that is actually transmitted or stored. The blue circles between the transforms and the entropy coding steps indicate the points of the so called latent-space representation of the image.

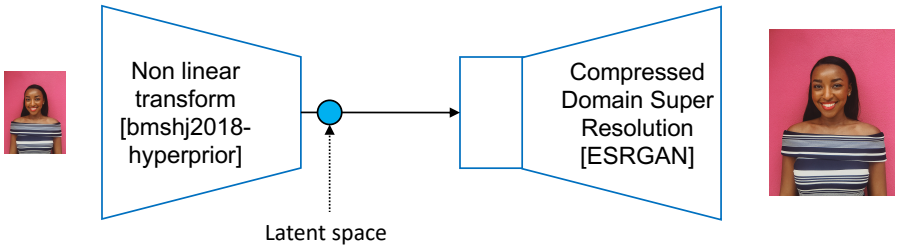



Figure 5. Coupling of a learning-based image codec with a super-resolution DNN.

Figure 5 shows how the coupling of the leaning-based codec *bmshj2018-hyperprior*[12] and the super resolution network ESRGAN[6] is performed. The input of the super resolution network is directly connected to the output of the compression network before entropy coding.

The training of the coupled system is performed following the same procedure as in[6] with the difference that before feeding the training images to ESRGAN they undergo a forward propagation through a pre-trained

*bmshj2018-hyperprior* model for a corresponding quality without the entropy coding step. The implementation details and the description of the training procedure is publicly available*.

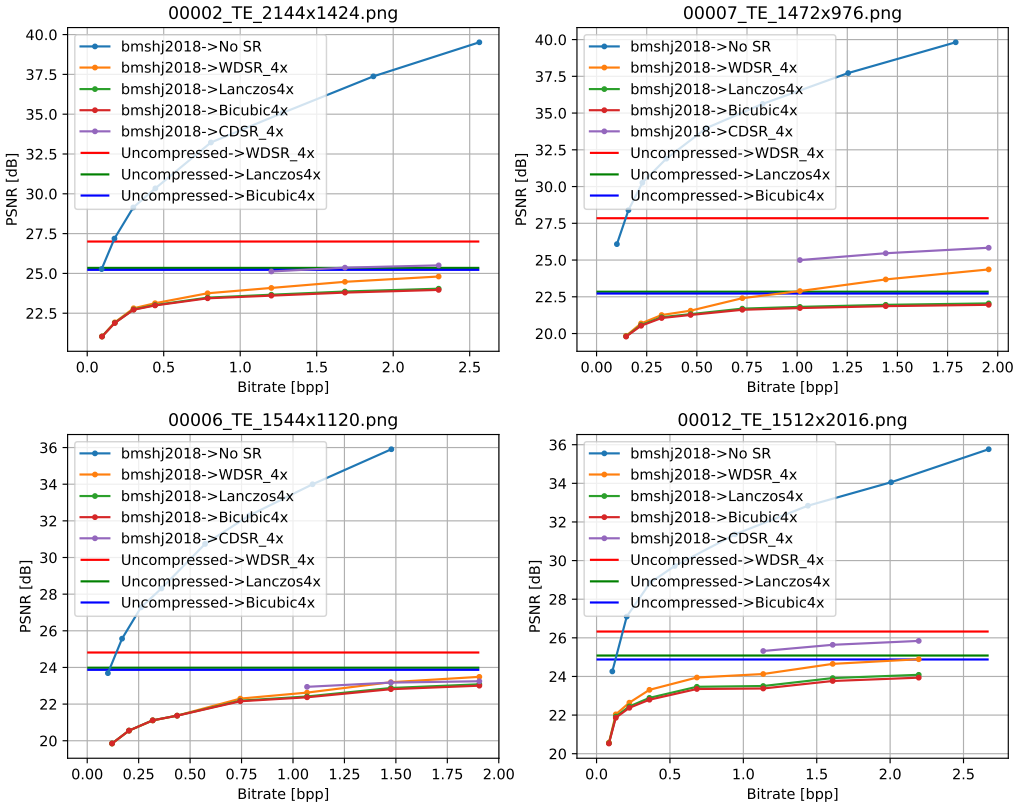
## 4.2 Results and discussion

Image ID 0002 | Image ID 0006 | Image ID 0007 | Image ID 0012

Figure 6. Four images from the JPEG AI dataset used for evaluation of compressed domain super resolution.

Figure 6 shows four images selected from the JPEG AI dataset for assessing the performance of the compressed domain super resolution method proposed in this paper.

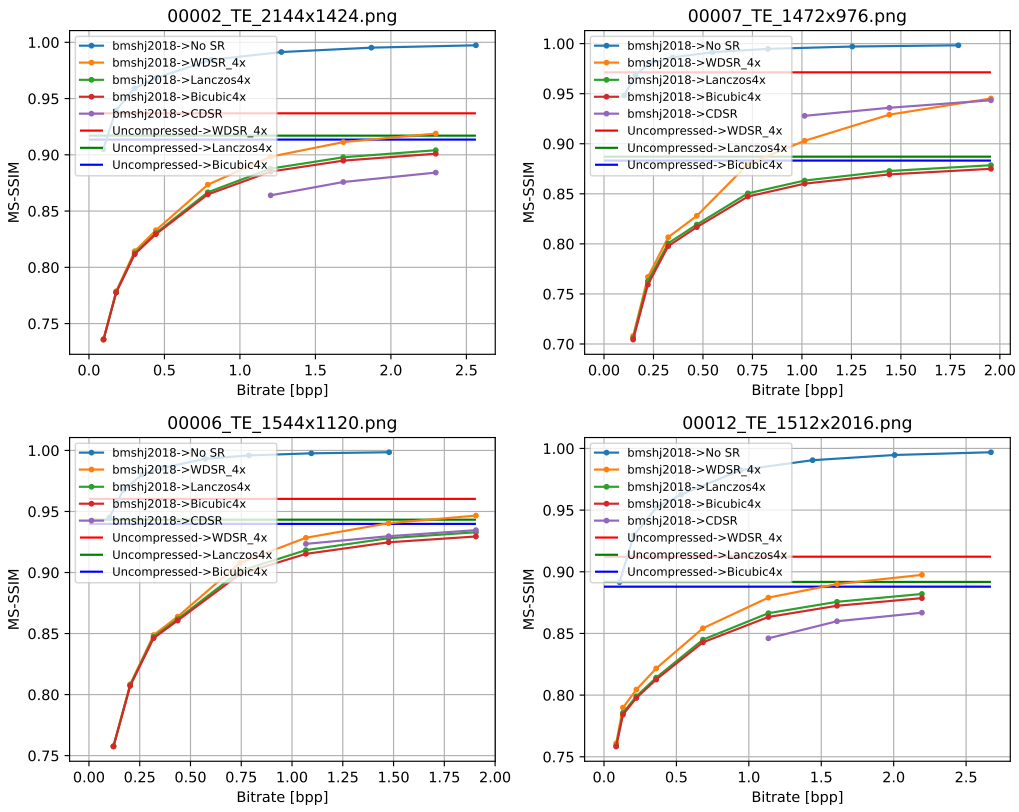
Figures 7a and 7b show the results of the evaluation of the proposed compressed domain super resolution (CDSR) method benchmarked against the anchors described in the Section 3.1.




(a) PSNR

Figure 7. Results of the evaluation of the proposed compressed domain super resolution (CDSR).

*https://github.com/mmspg/cdsr




(b) MS-SSIM

Figure 7. Results of the evaluation of the proposed compressed domain super resolution (CDSR).


Figure 8. Qualitative results of compressed-domain super resolution. From left to right: bicubic 4x-upscaling performed on a fully decompressed low resolution image, Lanczos 4x-upscaling performed on a fully decompressed low resolution image, compressed-domain super resolution (CDSR) applied to the latent space of a leaning-based codec. In all three cases, the image was compressed with bmshj2018-hyperprior at a quality level 6 using pre-trained model provided by the authors of the codec.

Figure 8 shows the qualitative results of compressed-domain super resolution (CDSR) applied to the latent space of a leaning-based codec compared to bicubic 4x-upscaling and Lanczos 4x-upscaling both performed on a fully decompressed low resolution image. In all three cases, the image was compressed with *bmshj2018-hyperprior* at a quality level 6 using pre-trained model provided by the authors of the codec.

## 5. CONCLUSION AND FUTURE WORK

In this paper, we evaluated the available state-of-the-art super resolution methods in compression workflow scenario. We proposed an adapted version of a super resolution network retrained to work with images in compressed domain and assessed its performance by benchmarking it to the anchors. The results show promising performance in terms of visual quality.

The future work may include the investigation of different loss functions for the compress domain super resolution in order to compensate inconsistencies in the results from different objective visual quality metrics. One may also investigate other SR models for compressed domain and additional learning based compression models.

## APPENDIX A. ADDITIONAL RESULTS

This section presents additional results in the form of rate-distortion plots for each image from the selected subset of the DIV2K dataset (Figure 2).

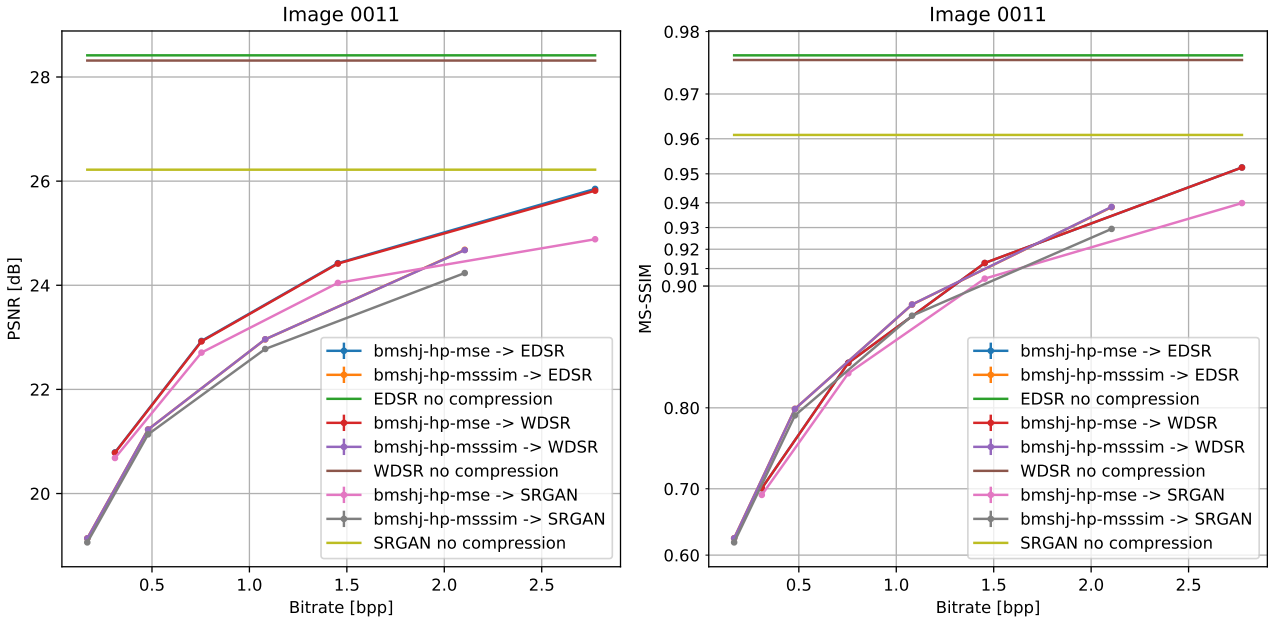



Figure 9. Rate-distortion plots for Image 0011

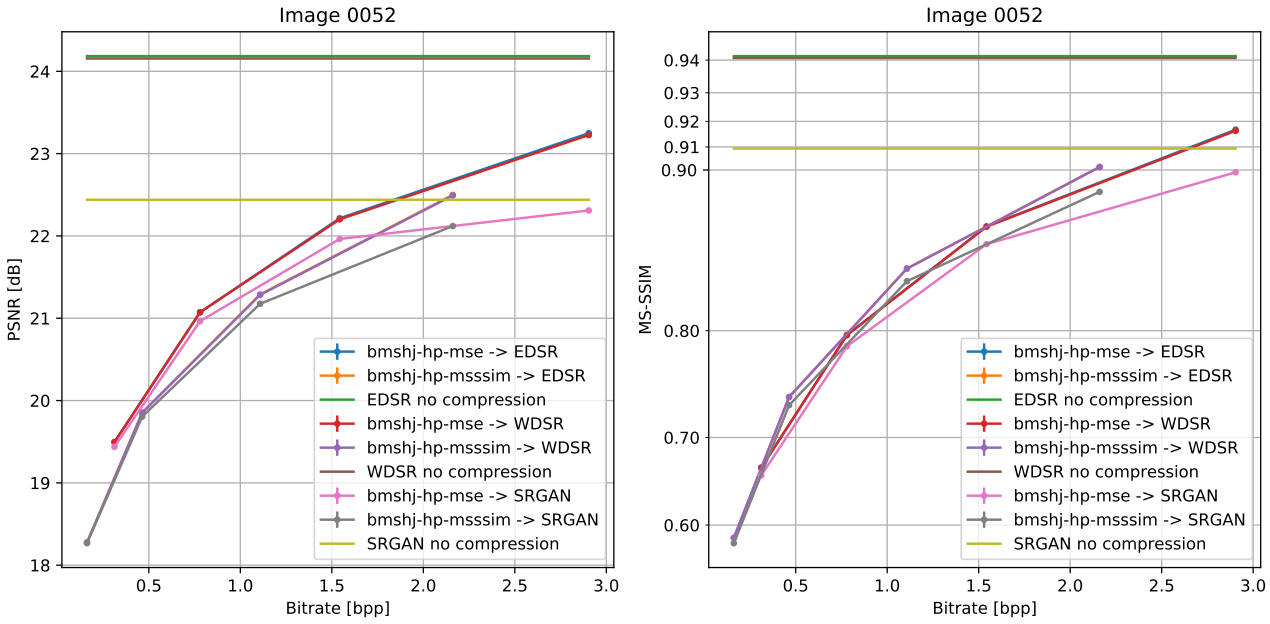



Figure 10. Rate-distortion plots for Image 0052

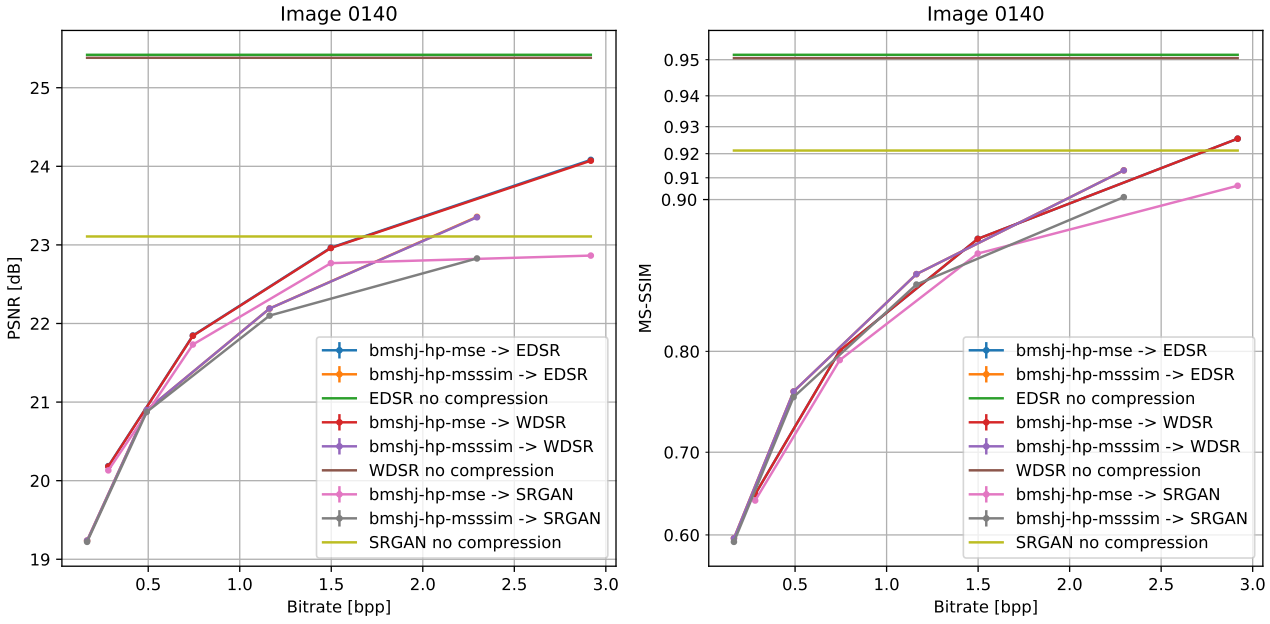



Figure 11. Rate-distortion plots for Image 0140

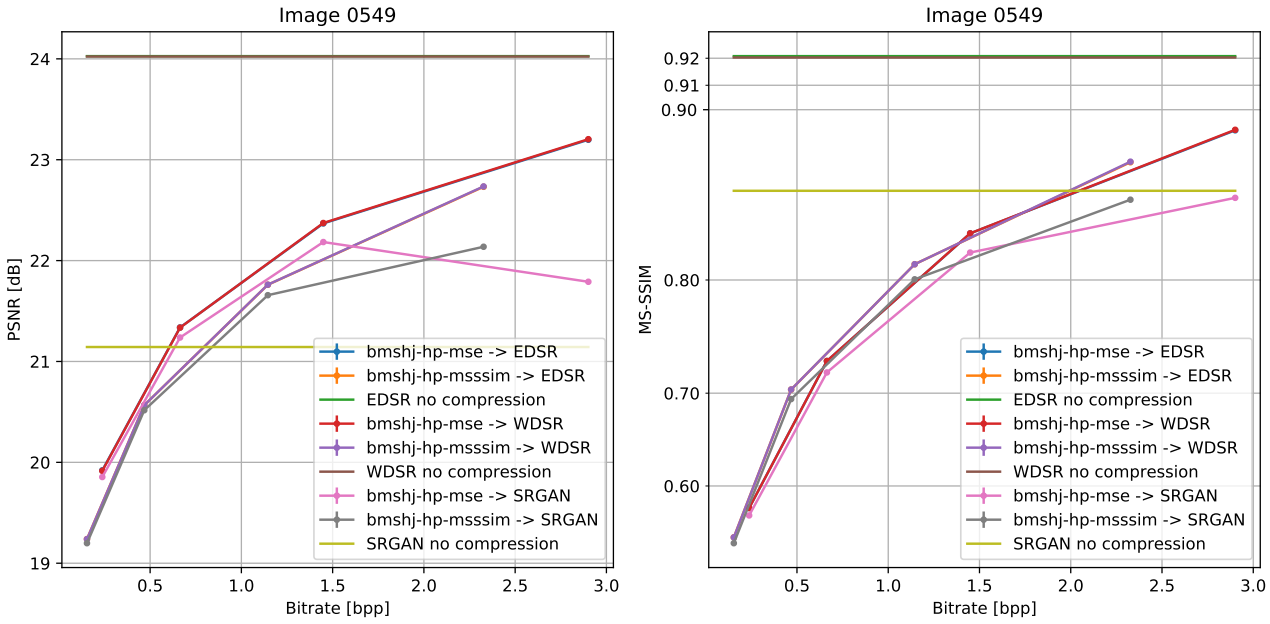



Figure 12. Rate-distortion plots for Image 0549

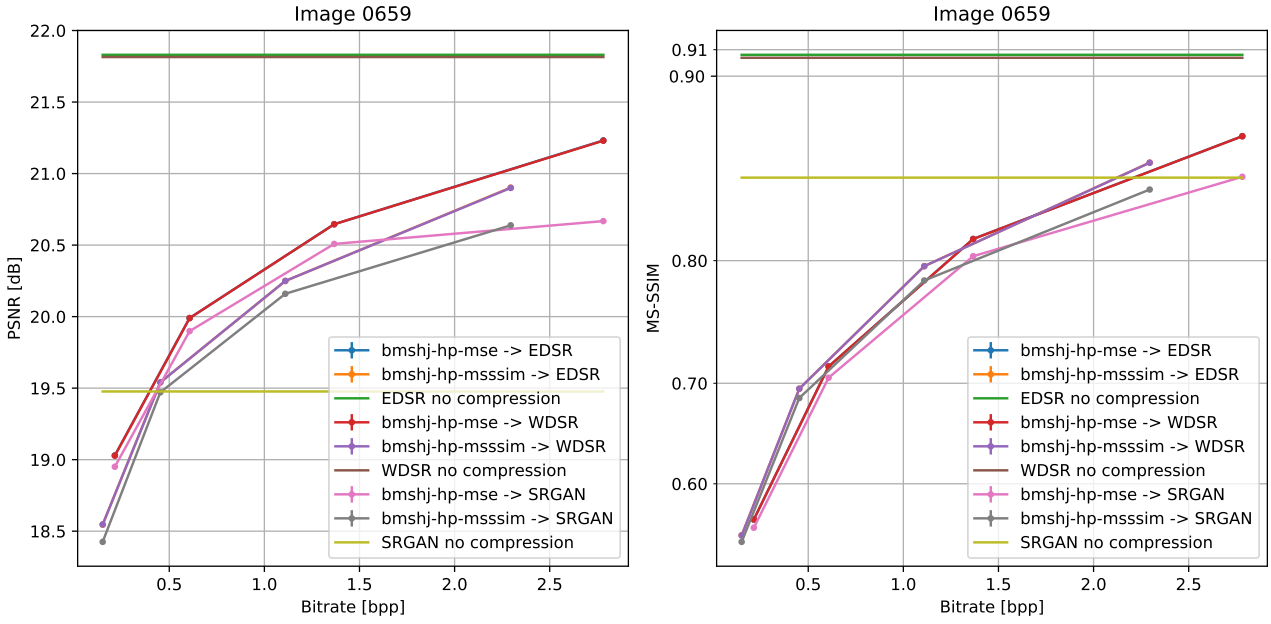



Figure 13. Rate-distortion plots for Image 0659

## ACKNOWLEDGMENTS

Authors would like to acknowledge contributions from the H2020 Innovation Action "Advanced Mixed Realities (AdMiRe)" under grant agreement 952027. In addition, authors would like to thank Chady Moukel for the contribution to the state of the art and anchors benchmarking.

## REFERENCES

[1] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W., "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in [*2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], (July 2017).

[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative adversarial nets," *Advances in neural information processing systems* **27** (2014).

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

[4] Lim, B., Son, S., Kim, H., Nah, S., and Lee, K. M., "Enhanced Deep Residual Networks for Single Image Super-Resolution," in [*2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*], (July 2017).

[5] Yu, J., Fan, Y., Yang, J., Xu, N., Wang, Z., Wang, X., and Huang, T., "Wide Activation for Efficient and Accurate Image Super-Resolution," in [*arXiv:1808.08718 [cs]*], (Dec. 2018).

[6] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Loy, C. C., "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," in [*Computer Vision – ECCV 2018 Workshops*], 63–79, Springer International Publishing (2019).

[7] Testolina, M., Upenik, E., and Ebrahimi, T., "Comprehensive assessment of image compression algorithms," in [*Applications of Digital Image Processing XLIII*], **11510**, 1151020, SPIE International Society for Optics and Photonics (Aug. 2020).

[8] Toderici, G., O'Malley, S. M., Hwang, S. J., Vincent, D., Minnen, D., Baluja, S., Covell, M., and Sukthankar, R., "Variable rate image compression with recurrent neural networks," in [*International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico*], (2016).

[9] Toderici, G., Vincent, D., Johnston, N., Jin Hwang, S., Minnen, D., Shor, J., and Covell, M., "Full resolution image compression with recurrent neural networks," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 5306–5314 (2017).

[10] Theis, L., Shi, W., Cunningham, A., and Huszár, F., "Lossy image compression with compressive autoencoders," in [*International Conference on Learning Representations, ICLR 2017, Toulon, France*], (2017).

[11] Ballé, J., Laparra, V., and Simoncelli, E. P., "End-to-end optimized image compression," in [*International Conference on Learning Representations, ICLR 2017, Toulon, France*], (2017).

[12] Ballé, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N., "Variational image compression with a scale hyperprior," in [*International Conference on Learning Representations*], (2018).

[13] Minnen, D., Ballé, J., and Toderici, G. D., "Joint autoregressive and hierarchical priors for learned image compression," *Advances in Neural Information Processing Systems* **31**, 10771–10780 (2018).

[14] Agustsson, E., Tschannen, M., Mentzer, F., Timofte, R., and Gool, L. V., "Generative adversarial networks for extreme learned image compression," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 221–231 (2019).

[15] Mentzer, F., Toderici, G. D., Tschannen, M., and Agustsson, E., "High-fidelity generative image compression," *Advances in Neural Information Processing Systems* **33** (2020).