Semantic based generative compression of images for extremely low bitrates

Tom Bordin, Thomas Maugey

Semantic based generative compression of images for extremely low bitrates

Tom Bordin
Inria
Rennes, France
tom.bordin@inria.fr

Thomas Maugey
Inria
Rennes, France
thomas.maugey@inria.fr

***Abstract*—We propose a framework for image compression in which the fidelity criterion is replaced by a semantic and quality preservation objective. Encoding the image thus becomes a simple extraction of semantic, enabling to reach drastic compression ratio. The decoding side is handled by a generative model relying on the diffusion process for the reconstruction of images. We first propose to describe the semantic using low resolution segmentation maps as guide. We further improve the generation, introducing colors map guidance without retraining the generative decoder. We show that it is possible to produce images of high visual quality with preserved semantic at extremely low bitrates when compared with classical codecs.**

I. Introduction

An image is traditionally compressed with the aim of minimizing the error made in the reconstruction. The Mean Squared Error (MSE) naturally comes as a simple and efficient criterion to evaluate the fidelity of the output in terms of distortion [1]. This metric remains widely used to evaluate methods compressing images ranging from high to low bitrates. But what happens when compression is pushed to the extreme ($\sim$0.02 bpp)? While excessive compression on sensitive data such as movies or vacations photos is not desirable, a drastic reduction of storage could be welcomed for the so-called "cold data". This massive amount of data that is stored but almost never accessed is estimated to represent 60% of today's storage while projected to become 80% by 2025 [2]. In that case, compression at extremely low bitrates could be an alternative to deleting potentially useful data or keeping a huge amount of data that might not be used. However, when targeting such bitrates, the relevance of the MSE as an evaluation criterion drops. As showcased by Blau *et al.* [3] there exists a tradeoff between the perceived quality of the output and the fidelity in terms of pixels. They call it the perception-distortion tradeoff. Moreover, Blau and Michaeli [4] show that decreasing the bitrate exacerbates the opposing goals of the two metrics, see Fig.1. Optimization of the MSE then leads to the apparition of numerous compression artifacts, which makes the use of such bitrates really unattractive. Recent neural methods, such as HiFiC [5] and its latest versions [6] [7], propose to alleviate this issue by integrating perceptual metrics to their training. They favorably replace latest standards even at low bitrates (0.2 bpp), but this remains insufficient to cope with the explosion of data. Our focus is on drastically lower bitrates where MSE-driven generative methods would fail to maintain semantic fidelity.

Fig. 1. Illustration of the perception-distortion tradeoff. Pixel (VVC) and semantic fidelity (SGC) compression methods side-by-side at extremely low bitrates (0.022 bpp). VVC which optimizes the MSE loses the semantic, while our approach manages to show semantic and visual quality at the same time.

Discarding the MSE criterion, we instead choose to optimize the realism and semantic fidelity of the output. Following Theis et al. [8], we encode the image by extracting a compact semantic description. Decoding from this representation then requires the use of a generative approach. This paradigm is referred to as *semantic based generative compression (SGC)*.

We propose a new SGC scheme with semantic represented as information on labels and colors. While labels brings information on the content of the image, the colors guarantee a certain level of consistency in the generation, acting as an overview of the image. The quantity of semantic transmitted

This work was funded by the French National Research Agency (MADARE, Project-ANR-21-CE48-0002)

then depends on the precision of the description, *i.e.* the number and precision of the labels. The decoding side is handled using a generative model, which is able to generate images from abstract semantic concepts. For our decoder, we propose to rely on diffusion models. Indeed, they showed great results in guiding the generation using semantic descriptions. The recent models for text to image generation [9] are a good illustration of their performance, sometimes even able to cheat human eyes in art competitions. We choose to rely on the architecture and weights of Latent Diffusion Models (LDM) [10]. Their model is designed to generate images using segmentation map conditioning which allows semantic fidelity. As color guidance is not native in LDMs, we propose a way to guide the generative process with a color map *without retraining their network*. Our method yields better visual results at extremely low bitrates when compared to recent codecs such as VVC while conserving the semantic.

We present the formulation for the generative compression with semantic representation as well as the architecture of the model we rely on in Section II. In Section III we present results of our method and a comparison with recent codecs.

II. Semantic based generative compression

A. Problem formulation

We define the semantic based generative compression framework and illustrate it in Fig. 2. The input image x is encoded into its semantic representation σ using a semantic encoder $\mathcal{E}$. The decoder $\mathcal{D}$, reconstructs the decoded output $\tilde{x}$ using the described semantic requiring a generative approach. Unlike classical decoders Fig. 2(a), our goal is not to minimize $d(x, \tilde{x})$, a pixel error. Instead, we want to preserve the semantic information in decoding while maximizing the realism or perception of the output. In other words, we switch from a pixel fidelity to a semantic fidelity, Fig. 2(b). To measure the quality and fidelity to the input, we define a semantic reconstruction error. We still measure a distance, but we do it in the semantic space. Using a projection Φ from the pixel to the semantic domain, the error can be expressed as $d(\Phi(x), \Phi(\tilde{x}))$. Φ is a nonlinear function extracting the semantic information of an image. Using this metric, two images can be extremely close semantically while being far in terms of MSE. Ideally, such a metric is resistant to rotations, translations or any other operations which do not change the content of the image and only slightly alter the semantic.

We formulate our problem as a maximization of the visual quality under constraints of very low bitrates $R < R_{low}$ and semantic fidelity $d(\Phi(\tilde{x}), \Phi(x)) < d_{min}$.

$$\max_{\mathcal{E},\mathcal{D}} \Psi(\mathcal{D}(\mathcal{E}(x))) \quad (1)$$
$$\text{s.t } R < R_{low} \text{ and } d(\Phi(\tilde{x}), \Phi(x)) < d_{min}$$

where Ψ is an evaluation of the realism (higher is better). In addition to quality metrics (FID) presented by Rombach et al. [10], we use several Image Quality Assessment (IQA) metric as Ψ to evaluate the generation.

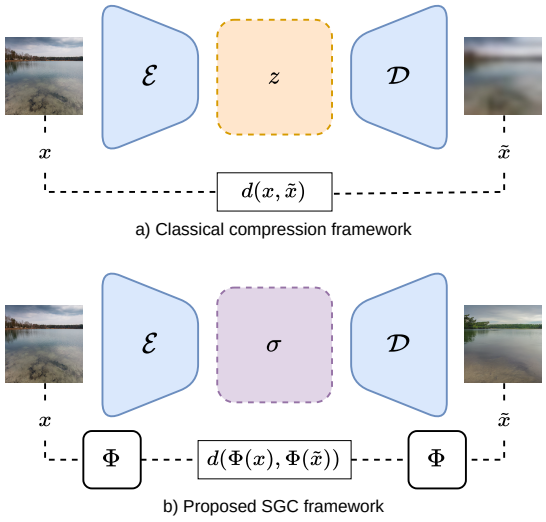

Fig. 2. a) Classical and b) SGC framework for image compression. With an encoder-decoder$(\mathcal{E}, \mathcal{D})$ structure, we propose to use a semantic representation σ. In contrast to classical frameworks, we choose to optimize the semantic fidelity formulated as $d(\Phi(\tilde{x}), \Phi(x))$ rather than the pixel fidelity $d(x, \tilde{x})$.

In the following, we present the choices made for the semantic encoding for $\mathcal{E}$. We then introduce our generative decoder $\mathcal{D}$, which uses the semantic representation as input.

B. Encoding the semantic: $\mathcal{E}$

The choice of the semantic representation defines what should be important to keep inside the image. We propose to describe σ as a combination of a segmentation map and a color map.

1) The segmentation map σ_s: gives information on objects and their position in the image. Using it as an embedding of our image, the encoding $\mathcal{E}(x)$ of an image then becomes the computation and compression of the segmentation map. In our case, it is estimated using the trained model DeepLabV3 [11] on classes of the COCO-Stuff dataset. The segmentation map is then lossly compressed using downsampling operation. Using low resolution of segmentation map σ_s as embedding still allows for a good reconstruction, losing only the small semantic labels in the process. At the decoding side, the segmentation map is upsampled and fed as input to the LDM model.

Images compressed using only segmentation maps present a lot of variability in samples. Indeed, labels only captures one kind of semantic and increasing the number of possible labels comes at the cost of an increased rate. Therefore, a photo during taken in the day will have the same labels as if it were taken during the night. This issue can be observed in the samples presented in Fig. 3(a). To address this, we decide to complete σ with a color map.

2) The color map σ_c: describes roughly the color information of the image. The color map is computed using a series of downsampling and blurring operations. Using a low resolution color map σ_c on its own is not enough to guide the generation.

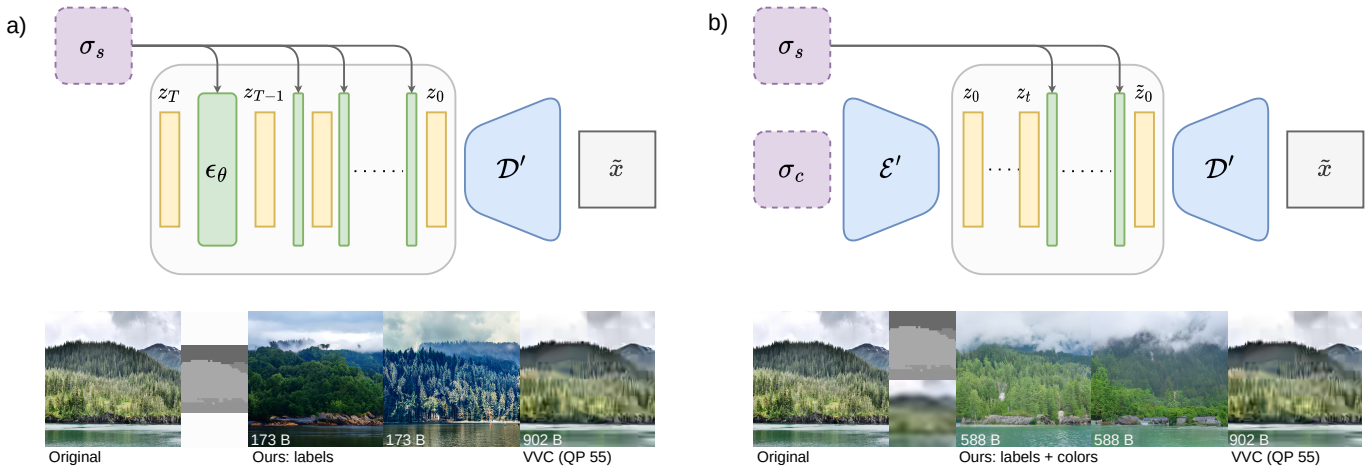

Fig. 3. a) Illustration of the LDM for generation of images conditioned on segmentation map σ_s as input, the 2 center images are results of our method. The generated samples may visually differ a lot from the source, with a lot of variability between samples. b) The proposed alternative for color map σ_c guidance without retraining the model. The samples are then more consistent with respect to the color of the source image. Size is indicated here in bytes(B).

This reveals the same problems as an MSE optimization, *i.e* a blurred blank spot could then either be decoded as a sheep or as a dog. The two semantic descriptions that were chosen are thus complementary. In practice, we use 16×16 color map. This size of color map gives global information on the aspect of the image. Using a higher resolution color map would not necessarily bring more precision on the generation, as the information might be lost when the noise is added at the start of the diffusion process.

To decode an image from the semantic representation, we propose to use a generative decoder, in our case LDM. We modify this model to integrate color guidance without requiring training of the network.

C. Generative decoding with LDM: D

We choose to use Latent Diffusion Models (LDM) [10] as our generative decoder. Diffusion models [12] [13] are trained by maximizing the likelihood that the generated outputs follow the same distribution as the dataset. They repeatedly denoise a randomly initialized vector until convergence to the desired distribution, in our case a realistic image. Rather than doing the diffusion in the pixel domain, LDM's authors propose to do it in the latent space of a separately trained VAE$(\mathcal{E}', \mathcal{D}')$ [14]. The role of the diffusion model ϵ_θ then becomes to generate embeddings of images in this latent space. They present and provide several models trained conditionally on various inputs such as text, segmentation maps or layout.

Following the standard process of diffusion, the generation of samples using an LDM with a conditioned input for image generation is illustrated in Fig. 3(a). A noise z_T is first randomly drawn with gaussian distribution in the latent space and repeatedly guided towards an image embedding z_0 with $z_{t-1} = \epsilon_\theta(z_t, t, \sigma_s)$ for T timesteps. The semantic description σ_s serves as a guide for ϵ_θ and conditions the generation of the semantic. The VAE decoder $\mathcal{D}'$ is then used to reconstruct $\tilde{x}$, more details on the generation are present in their article. We illustrate the generation process on an image of our dataset in the following figures. With samples generated using only the segmentation map, there is a lot of variability in output. We thus added color map to the semantic representation.

In LDMs ϵ_θ does not natively take a color map as input. Rather than retraining an already trained, powerful and large network, we propose an alternative to conditioning for guiding using colors. This method is illustrated in Fig. 3(b). Inspired by previous work [15] we propose to use the color map at the start of the diffusion model. Instead of initializing the diffusion with a random gaussian noise as the first embedding z_T, we use the color map encoded in the latent space using the VAE encoder $\mathcal{E}'(\sigma_c)$ it corresponds to z_0 in Fig. 3(b). Then adding noise, the diffusion process is intercepted at a timestep $t < T$ as if the denoising already started. The rest of the diffusion proceeds normally for t timesteps. This method is possible as the VAE introduced in LDM was trained optimizing an MSE loss. The latent space thus contains information on the pixels of the color map, *i.e.* the colors.

Note that the choice of the timesteps t at which we intercept the diffusion has an impact on the generation. Indeed, as shown [16] the smaller t is the more z_t is close to a realistic sample and the harder it is to redirect the generation. On the other hand, if t is chosen too close to T then z_t does not contain enough information on the colors for the output to respect the color map at generation as too much noise was added. We empirically find that using 70% of T is ideal in our case. Using the same image encoded, we show samples generated using this method in Fig. 3(b). The color adds consistency to generation by transmitting an overview of the original image.

III. Experiments

In this section, we compare our method with standard codecs' compression targeting similar bitrates. Since our method is not deterministic, to avoid cherry-picking, we choose the first generated sample for each image presented. All

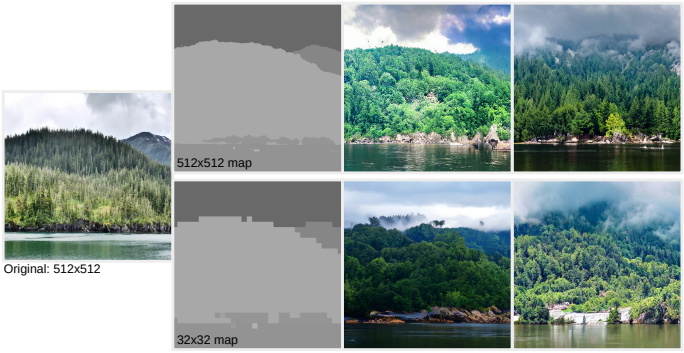

Fig. 4. Using downsampled segmentation maps does not affect neither the quality of generation nor the fidelity to the input in terms of semantic. The network is able to extrapolate and generate correct labels, not strictly inside the given map.

the images present in figures are in 512×512. Reader can find further results on different images in supplementary material available at *https://project.inria.fr/dare/publications/sgc*.

A. Sampling parameters

We use the LDM previously described as our generative decoder for semantic based image generative compression. The architecture and weights of the model [10] which can be found on the LDM's GitHub[1]. We use the conditional semantic generation model trained on the Landscape [14] dataset, fine-tuned for the generation of 512×512 images. Samples are generated using 200 timesteps with a DDIM scheduler and a guidance scale of 2 for the diffusion process. The scale guiding the influence of the condition on the generation.

The segmentation maps were obtained using a pretrained DeepLabV3 [11] model. This segmentation model is used in the encoding task of our framework. In order to further reduce the bitrate and redundancy, the segmentation maps are then downsampled. We show that downsampling the segmentation map up to 16 times does not affect the quality of the generation of images in 512×512, see Fig. 4. The network extrapolates outside the given boundaries to generate more realistic samples. The image is encoded using lossless compression on the downsampled 32×32 segmentation map. At decoder side, the segmentation map is upsampled to fit the input of the model, only smaller area of labels are lost in the process.

B. Discussion

Following the formulation in (1), we evaluate our methods on image quality and semantic similarity. Results of our method are showcased in Fig. 5 along with evaluation metrics in Table I(↑ higher is better, ↓ lower is better). Frechet Inception Distance (FID) evaluation of the chosen architecture are available along more details on LDMs in the original article [10]. We compare our results to VVC, for which we use the latest version of the intra coder(v1.6) with Vvenc implementation [18]. Other learned codecs [5], [7], [19] were not trained for such bitrates. They would require retraining to target similar bitrates for comparison.

[1]https://github.com/CompVis/latent-diffusion

TABLE I
EVALUATION OF SGC

	IQA metrics ↑ Ψ				Φ		R
	DBCNN [20]	**MUSIQ** [21]	**HYPERIQA** [22]	**CLIP-IQA** [23]	**BCE** ↓	**CLIP** ↑ [24]	**bpp** ↑
Input	62.2	67.5	0.58	0.56	-	-	-
VVC(qp 55)	19.8	24.2	0.23	0.32	0.93	0.62	0.0229
SGC(Ours)	49.8	60.6	0.54	0.47	0.40	0.80	0.0209

Images showcased from our method show a good visual quality and preservation of the semantic. The visual results are backed by the corresponding metrics. The perceptual quality Ψ of the image is evaluated using several image quality assessment metrics. We measure the quality over the dataset Landscape, and we systematically achieve better quality than VVC and score closely to images of the dataset on two of those metrics. The semantic similarity Φ is evaluated on two criteria. Firstly, on the labels and their position, using a binary cross entropy between estimated segmentation maps on decoded images and input images. We see from this metric that the segmentation model is not able to predict correctly the classes for the images encoded using VVC. Our method has a better preservation of the semantic when measured on the 182 possible labels. For the second criteria, we use a CLIP alignment metric. CLIP is a model originally used to measure alignment between text and images, however it can also be used to compare two images. Projecting the input and decoded images into CLIP latent using the image encoder, the angle between the two normalized vectors gives an alignment score. We present the measures as the scalar product between the two vectors, with values ranging from 0 for no alignment to 1 for a perfect match. We use the ViT-L/14 version of CLIP. Note that this metric is different from CLIP-IQA which relies on text and image alignment to assess the quality. We notice that even though our method was not trained to optimize these criteria, we still score really close to the input and higher than the VVC encoded image. SGC thus propose an alternative to the distortion optimization, offering quality and semantic preservation at extremely low bitrates.

A significant improvement lies in the absence of compression artifacts when using our method. The reduction in bitrate comes at the cost of a loss of information rather than a loss in image visual quality. Metrics shows that the loss of semantic for classical coding is more important than our methods. With an important constraint on the rate, the precision of the semantic description limits the fidelity we can get. Indeed, using a higher number of labels dedicated to the description of landscapes would bring decoded images closer to originals. This is noticeable in the last two rows of images in Fig. 5, where semantic has been preserved, but resemblance is somewhat lacking, even though both semantic metrics still advantages SGC on those particular examples.

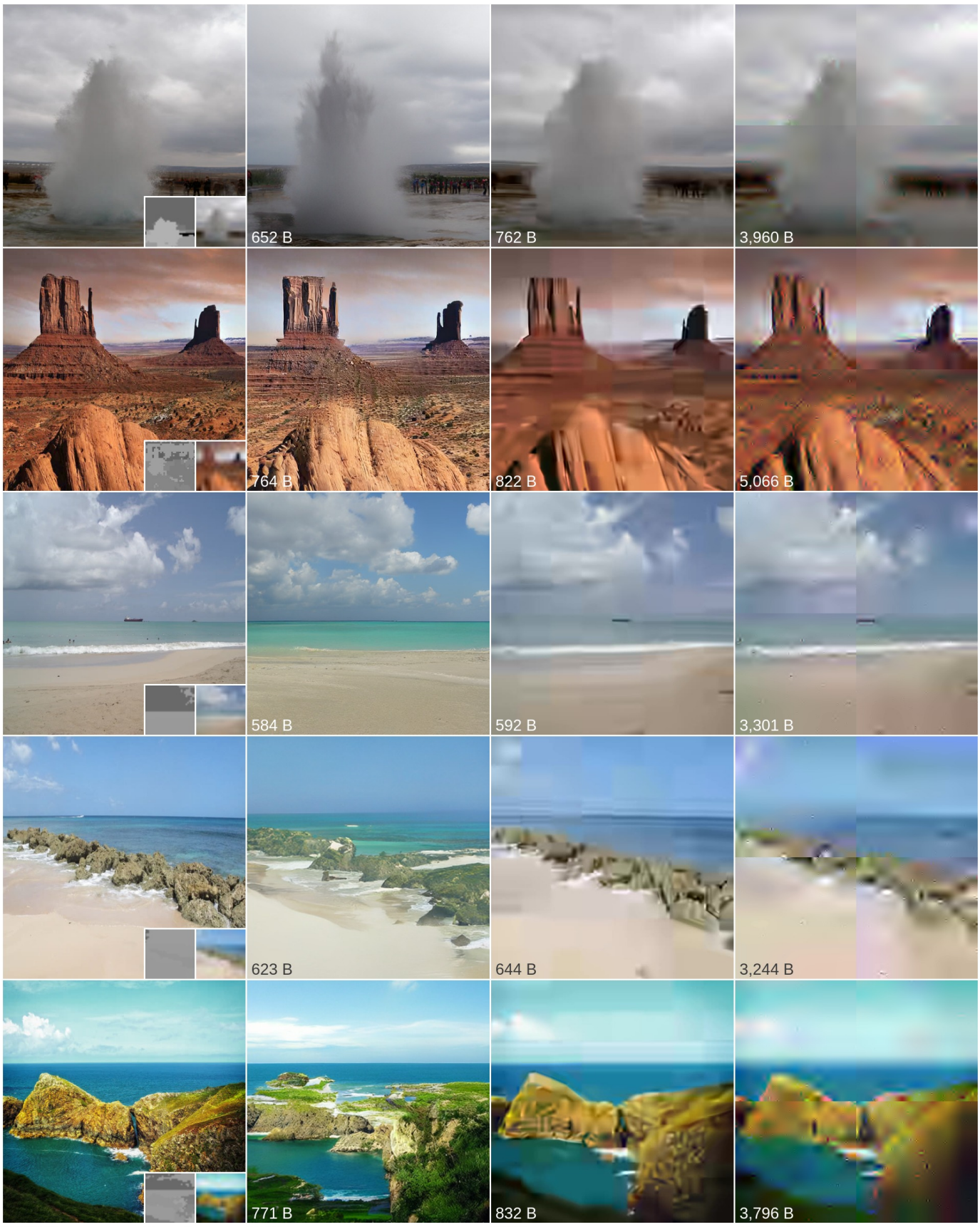

Fig. 5. Illustration of our methods compared to VVC. jpeg2000 [17] does not manage to satisfyingly reach similar bitrates. Our methods yield better image quality while conserving the transmitted semantic.

IV. Conclusion

We proposed a new framework for image compression based on a semantic representation. We proposed to guide the generation using color maps on a model trained model using segmentation maps as input. The decoding relies on a conditional diffusion process to generate images faithful to the semantic. We showed that using a semantic description of the image is enough to produce samples close to the image with high visual quality. Comparing to recent codecs, decoded image are highly detailed through synthetic information. This method could further be applied by doing selective generation, synthesizing only unnecessary information.

An interesting improvement for future work would be to have a fine grain semantic representation to navigate into. This would allow a control on the distribution of the semantic over the image and a better control on the bitrate.

References

[1] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[2] Phil Goodwin. Tape and cloud: Solving storage problems in the zettabyte era o f data. *IDC Corporate, Massachusetts, United States*, 2019.

[3] 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.

[4] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pages 675–685. PMLR, 2019.

[5] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in Neural Information Processing Systems*, 33:11913–11924, 2020.

[6] Emiel Hoogeboom, Eirikur Agustsson, Fabian Mentzer, Luca Versari, George Toderici, and Lucas Theis. High-fidelity image compression with score-based generative models. *arXiv preprint arXiv:2305.18231*, 2023.

[7] Lucas Theis, Tim Salimans, Matthew D Hoffman, and Fabian Mentzer. Lossy compression with gaussian diffusion. *arXiv preprint arXiv:2206.08889*, 2022.

[8] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 221–231, 2019.

[9] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[15] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[16] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[17] David Taubman and Michael Marcellin. *JPEG2000 Image Compression Fundamentals, Standards and Practice: Image Compression Fundamentals, Standards and Practice*, volume 642. Springer Science & Business Media, 2012.

[18] Adam Wieckowski, Jens Brandenburg, Tobias Hinz, Christian Bartnik, Valeri George, Gabriel Hege, Christian Helmrich, Anastasia Henkel, Christian Lehmann, Christian Stoffers, Ivan Zupancic, Benjamin Bross, and Detlev Marpe. Vvenc: An open and optimized vvc encoder implementation. In *Proc. IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–2.

[19] Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. *arXiv preprint arXiv:2209.06950*, 2022.

[20] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018.

[21] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021.

[22] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3667–3676, 2020.

[23] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. *arXiv preprint arXiv:2207.12396*, 2022.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.