THOUSAND TO ONE: SEMANTIC PRIOR MODELING FOR CONCEPTUAL CODING

Jianhui Chang[1,4], *Zhenghui Zhao*[2], *Lingbo Yang*[2], *Chuanmin Jia*[2], *Jian Zhang*[1,4], *Siwei Ma*[2,3]

[1]Peking University Shenzhen Graduate School, Shenzhen, China
[2]Institute of Digital Media, Peking University, China
[3]Information Technology R&D Innovation Center of Peking University, Shaoxing 312000, China
[4]Peng Cheng Laboratory, China
{jhchang, zhzhao, lingbo, cmjia, zhangjian.sz, swma}@pku.edu.cn

ABSTRACT

Conceptual coding has been an emerging research topic recently, which encodes natural images into disentangled conceptual representations for compression. However, the compression performance of the existing methods is still suboptimal due to the lack of comprehensive consideration of rate constraint and reconstruction quality. To this end, we propose a novel end-to-end semantic prior modeling based conceptual coding scheme towards extremely low bitrate image compression, which leverages semantic-wise deep representations as a unified prior for entropy estimation and texture synthesis. Specifically, we employ semantic segmentation maps as structural guidance for extracting deep semantic prior, which provides fine-grained texture distribution modeling for better detail construction and higher flexibility in subsequent high-level vision tasks. Moreover, a cross-channel entropy model is proposed to further exploit the inter-channel correlation of the spatially independent semantic prior, leading to more accurate entropy estimation for rate-constrained training. The proposed scheme achieves an ultra-high 1000× compression ratio, while still enjoying high visual reconstruction quality and versatility towards visual processing and analysis tasks.

Index Terms— Semantic prior, conceptual coding, low bitrate, cross-channel entropy model

1. INTRODUCTION

With the progress in deep generative models, conceptual coding [1, 2, 3] has emerged as a new paradigm for image compression beyond traditional signal-based image codecs, such as JPEG [4], JPEG2000 [5], HEVC [6] and other learning-based codecs [7, 8]. Aiming at extracting decomposed conceptual representation from input visual data, conceptual coding not only achieves significant bitrate reduction over traditional codecs at comparable reconstruction quality, but also supports more flexible vision tasks. Despite the achieved rapid progress, finding an efficient prior modeling and feature compression scheme under extreme conditions (*e.g.*, 1000× compression ratio) still remain a considerable challenge.

More precisely, current image codecs typically involve entropy estimation through variational entropy-constrained training, where spatial dependencies in signal-based latent codes are often exploited for bitrate reduction [7, 9]. As an approximation of bitrate, entropy can only be minimized properly if statistical dependencies over the compressed domain are properly captured. However, the entropy modeling techniques are still under-explored for conceptual coding. Specifically, existing methods [2, 3] typically adopt a structure-texture dual-layered framework, yet the acquired conceptual codes are often compressed without effective rate optimization. Furthermore, a single latent vector is usually leveraged to model global texture distribution of multiple semantic regions, where the intra-region similarity and cross-region independencies of texture codes are not fully exploited for entropy-constrained training. In consequence, state-of-the-art conceptual coding schemes [3] still exhibit inferior rate-distortion performance against current learning-based codecs (*e.g.*, [10]).

In this paper, we propose a novel semantic prior modeling based conceptual coding approach for ultra-low bitrate image compression by incorporating semantic-wise deep representations as a unified prior for both texture synthesis and entropy estimation. As shown in Fig. 1, instead of assuming a global texture code, we employ semantic segmentation maps as structural guidance to extract deep semantic prior within each individual semantic region. On the one hand, the deep semantic prior models texture distributions semantic-wisely for finer texture representation and synthesis. On the other hand, by taking advantage of semantic correlation, the conceptual representations are presented in a spatially independent form, which benefits more accurate entropy estimation. Moreover, we propose a cross-channel entropy module with a hyperprior model to exploit inter-channel dependencies in se-

This work was supported in part by the National Natural Science Foundation of China (62072008, 61931014), the China Postdoctoral Science Foundation (2020M680238), PKU-Baidu Fund (2019BD003) and High-performance Computing Platform of Peking University, which are gratefully acknowledged.

978-1-6654-3864-3/21/$31.00 ©2021 IEEE

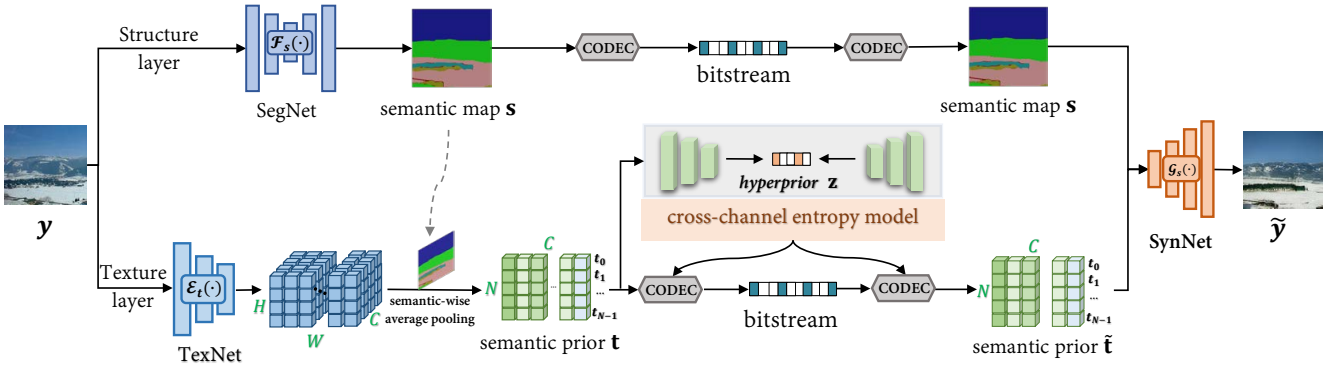

Fig. 1. **Overview of our proposed semantic prior modeling based conceptual coding approach, consisting of a structure layer and a texture layer.** The structure layer is represented by semantic map, which is lossless coded and employed as guidance for extracting deep semantic prior and preserving structural information. The texture layer is modeled by semantic prior and its entropy is estimated by a cross-channel entropy model. The received semantic prior and map are integrated to synthesize the decoded image on the decoder side.

mantic prior distribution for more accurate entropy estimation and higher bitrate reduction. The proposed conceptual coding scheme is end-to-end trainable with entropy-constrained rate-distortion objectives, and is capable of achieving high reconstruction quality at extreme settings (*e.g.*, 1000× compression ratio). Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to propose an end-to-end semantic prior modeling based conceptual coding scheme by extracting semantic-wise deep representations as a unified prior for both texture synthesis and entropy estimation, leading to significantly increased reconstruction quality and flexibility in content manipulation.
- We propose a cross-channel entropy model for effective hyperprior estimation and channel dependency reduction of semantic prior, allowing a more effective rate-distortion optimization with regard to entropy constraint.
- Extensive experiments demonstrate that the proposed method can achieve perceptually convincing reconstructions at extremely low bitrate (0.02-0.03 bpp, $\sim$ 1000× compression ratio), as well as better support for various image analysis and manipulation tasks.

2. PROPOSED METHOD

2.1. Semantic Prior Modeling

In this paper, we adopt the structure-texture layered decomposition form to realize the conceptual coding framework. Considering rate constraint and reconstruction quality comprehensively, we propose to model a semantic prior for texture representation and further entropy estimation. In particular, as shown in Fig. 1, input image $\mathbf{y}$ is processed into two basic visual features separately: 1) the structure layer characterized with the semantic segmentation map $\mathbf{s}$ which contains versatile information including structure layout, semantic category, location and shape, obtained by image segmentation networks $\mathcal{F}_s(\cdot)$ (SegNet, *e.g.*, PSPNet [11]); and 2) the texture representations $\mathbf{t}$ modeled by semantic-wise deep prior extracted with a convolutional neural network (CNN) based feature extractor $\mathcal{E}_t(\cdot)$ (TexNet, *e.g.*, feature encoder in [12]) with the guidance of semantic map $\mathbf{s}$. On the decoder side, the target image $\tilde{\mathbf{y}}$ is reconstructed by integrating the decoded semantic prior $\tilde{\mathbf{t}}$ and lossless semantic map $\mathbf{s}$ by $\mathcal{G}_s(\cdot)$ (*SynNet*, *e.g.*, generator in [13]), *i.e.*, $\tilde{\mathbf{y}} = \mathcal{G}_s(\mathbf{s}, \tilde{\mathbf{t}})$.

The process of extracting semantic prior is shown in Fig. 1. A CNN-based feature extractor first transforms input images into intermediate features with the shape $C \times H \times W$, where C, H, W correspond to channels, height and width, respectively. Then a semantic-wise average pooling layer is utilized to compute spatially average features under the guidance of semantic map, obtaining aggregated latent vectors corresponding to each semantic region as semantic prior. The shape of semantic prior is $C \times N$, where N denotes the number of semantic class. The latent vectors $\{\mathbf{t}_0, \mathbf{t}_1, ..., \mathbf{t}_{N-1}\}$ characterize the prior of semantic region $0, 1, ..., N-1$ correspondingly. By taking advantage of semantic structure and average pooling, source latent feature maps reduce spatial dependencies and present as entropy modeling friendly semantic prior. To further address internal channel dependencies in the latent vector of each semantic region, we propose a cross-channel entropy model and incorporate a hyperprior to model channel correlation for accurate entropy estimation as introduced in Sec. 2.2. Combining reconstruction and entropy estimation tasks in training, our proposed semantic prior could effectively model texture distribution in an entropy modeling friendly form, which benefits both bitrate saving and reconstruction quality.

2.2. Cross-channel Entropy Model

Plenty of entropy models [14, 7, 9] have been introduced for joint rate-distortion optimization in learned image codecs. Typically, Ballé *et al.* developed a school of entropy mod-

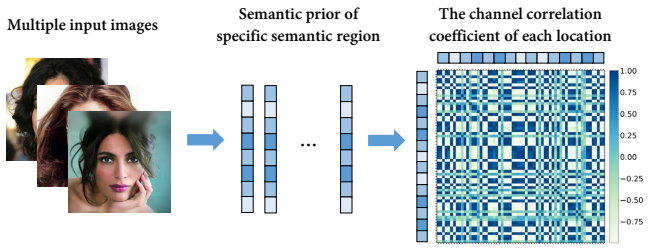

Fig. 2. **The channel correlation coefficient.** Extracting semantic prior from a image set, the latent vectors of specific semantic region (*e.g.*, hair) are separated to calculate the Pearson correlation coefficient [15] among channels. Higher score (blue) shows higher positive correlation and lower score (white) shows higher negative correlation. The visualizations demonstrate high correlation among channels.

els from a simple fully factorized model [8], to conditional Gaussian mixture model incorporating hyperprior [14] and context model [7]. The improvement of entropy model relies on the constantly further exploitation of dependencies in latent codes. However, different from latent codes obtained in signal-based nonlinear transform where spatial dependencies are mainly considered, conceptual representations demonstrate different correlation characteristics, urging the scheme of our matching cross-channel entropy model.

Due to the responsibility for providing accurate semantic location guidance for reconstruction, semantic maps are lossless transmitted. Thus, the proposed entropy model aims to model the probability distribution of semantic prior adaptively in training for the bit-saving purpose along with high reconstruction quality. In essence, the learned semantic prior is presented in a spatial independent form by taking advantage of semantic correlation in the extraction process shown in Fig. 1, leaving internal dependencies at channel dimension to further exploit. To quantitatively analyze the correlation, we extract the semantic prior from random 100 images and separate the latent vectors of specific semantic region (*e.g.*, hair) to calculate the Pearson correlation coefficient [15] matrix as shown in Fig. 2, where the darker blue indicates the positive correspondence across channels in latent vectors and the example results demonstrate a high channel-wise correlation. To this end, we propose a cross-channel entropy model, which incorporates a hyper-encoder to learn a *cross-channel hyperprior* $\mathbf{z}$ to capture channel dependencies by three spatially invariant and channel-wise reduced convolutional layers, and a hyper-decoder to produce statistical parameters to conditional Gaussian mixture model for probability estimation. As side information, the entropy of hyperprior is estimated by an independent density model as [8]. By fully exploiting the statistical redundancy of deep semantic prior, the proposed cross-channel entropy model could effectively reduce the bitrate in training.

2.3. Optimization Objectives

In this paper, we introduce the rate-distortion optimization into conceptual coding [16]. As shown in Fig. 1, input image $\mathbf{y}$ is encoded into semantic maps $\mathbf{s}$ and semantic prior $\mathbf{t}$ respectively. For entropy-constrained training, a cross-channel entropy model $\mathcal{P}_f(\cdot)$ is proposed to incorporate hyperprior $\mathbf{z}$ to estimate semantic prior entropy where the rate of quantized $\tilde{\mathbf{z}}$ is estimated with factorized entropy model [7] $\mathcal{P}_z(\cdot)$. The quantization is simulated with uniform noise as [8] in training and applies rounding algorithm directly in test. The trainable rate constraint can be obtained as,

$$r(\tilde{\mathbf{t}}) = E_{\mathbf{t}\sim p_T}\left\{-log_2(\mathcal{P}_f(\tilde{\mathbf{t}}))\right\} + E_{\mathbf{z}\sim p_Z}\left\{-log_2(\mathcal{P}_z(\tilde{\mathbf{z}}))\right\}. \tag{1}$$

On the decoder side, the decoded texture representation $\tilde{\mathbf{t}}$ and lossless semantic map $\mathbf{s}$ are integrated by $\mathcal{G}_s(\cdot)$ to reconstruct the target image $\tilde{\mathbf{y}}$, *i.e.*, $\tilde{\mathbf{y}} = \mathcal{G}_s(\mathbf{s}, \tilde{\mathbf{t}})$. Since conceptual compression pursues appreciable visual reconstruction quality under extremely low bitrate rather than signal fidelity, and the pixel-wise similarity metrics prove to reduce signal distortion but impair perceptual quality [17], we employ the perceptual loss [18] d_P and feature matching loss [12] d_{FM} to form our distortion:

$$d(\mathbf{y}, \tilde{\mathbf{y}}) = \lambda_P d_P + \lambda_{FM} d_{FM}. \tag{2}$$

Furthermore, the conditional generative adversarial models (GANs [19]) are also employed to learn the distribution mapping from semantic map and semantic prior pair $\{\mathbf{s}, \tilde{\mathbf{t}}\}$ to decoded image $\{\tilde{\mathbf{y}}\}$ under the condition $\mathbf{s}$, where the discriminator $\mathcal{D}$ is applied for the adversarial training. Additionally, the latent regression loss [3] $\mathcal{L}_r$ is utilized as a regularization term to improve the semantic disentanglement of texture representations which can be verified with experiments. With parameterized models $\mathcal{E}_t, \mathcal{G}_s, \mathcal{D}, \mathcal{P}_f, \mathcal{P}_z$ and α, β as hyper-parameters for weight control, the loss objectives for rate-distortion and discriminator are shown as follows:

$$L_{\mathcal{E}_t,\mathcal{G}_s,\mathcal{P}_f} = E_{y\sim p_Y}[\lambda r(\tilde{\mathbf{t}}) + d(\mathbf{y}, \tilde{\mathbf{y}}) + \alpha\mathcal{L}_r - \beta log(\mathcal{D}(\tilde{\mathbf{y}}, \mathbf{s}))], \tag{3}$$

$$L_{\mathcal{D}} = E_{y\sim p_Y}[-log(1 - \mathcal{D}(\tilde{\mathbf{y}}, \mathbf{s}) + (-log(\mathcal{D}(\mathbf{y}, \mathbf{s})))]. \tag{4}$$

3. EXPERIMENTS

3.1. Experimental Settings

Networks. The proposed hyper-encoder and hyper-decoder employ three 1×1 convolutional layers respectively to learn the $16\times$ downscaled channel prior and corresponding mean and scale parameters. Besides, for the convenience of preliminary experiments, the *TexNet*, *SynNet* and discriminator are built upon [20, 13]. Note that we remove the Tanh activation and apply instance and spectral normalization in the decoder and discriminator.

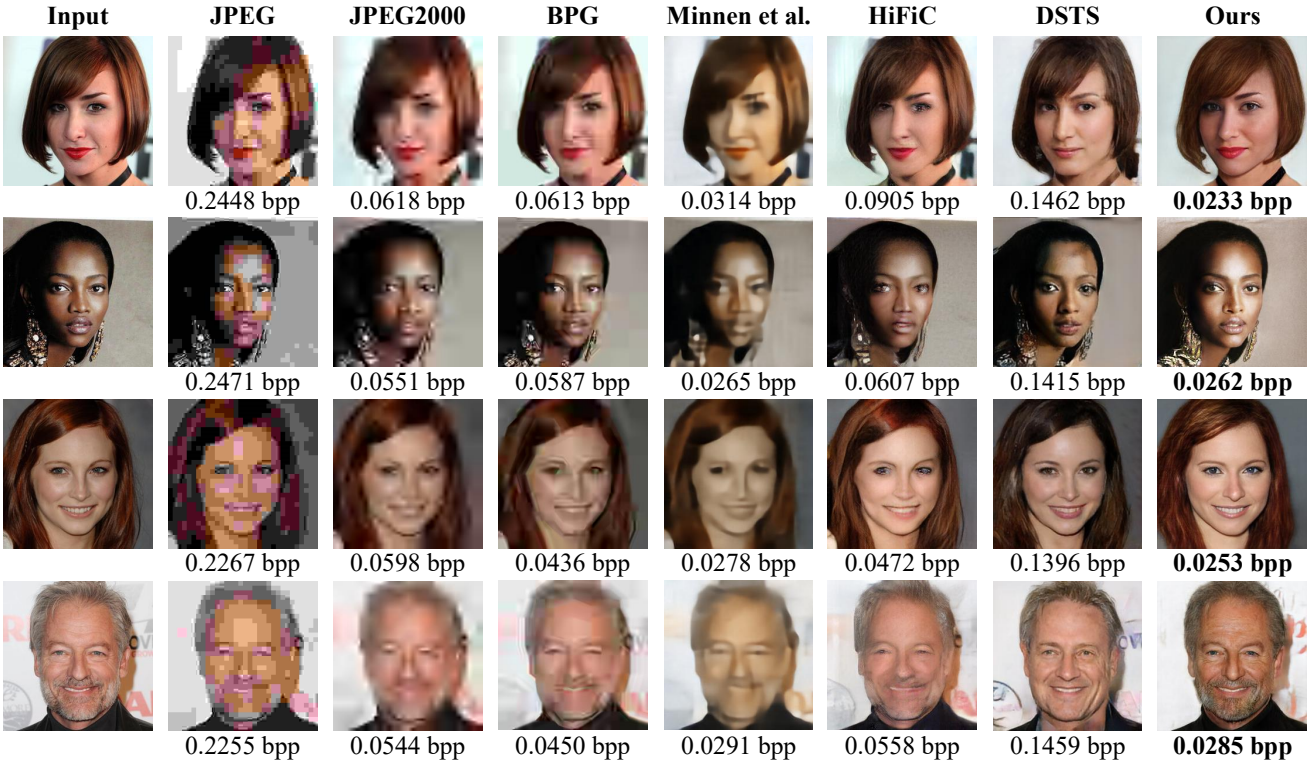

Fig. 3. **Qualitative comparisons with baselines under low bitrate.** The ***bpp*** of each image is reported under the corresponding image. The results show that our method outperforms others by achieving higher reconstruction quality with lower bitrate.

Dataset. The proposed method is mainly evaluated on CelebAMask-HQ[1] containing 19 semantic categories and 30,000 paired images of size 256×256, with 24183 as training set, 2824 as testing set and 2993 as validation set. Besides, ADE20K [21] is also utilized as additional dataset for discussion.

Other settings. The semantic map is lossless coded using FLIF[2]. The channel dimension of texture representations is set to 64 and the quantization scale is set to 0.01 empirically for comparison. We set the learning rate to 0.0001 and the Adam optimizer [22] with default settings is used for training. The parameters in Eq. (3) are set as follows: $\alpha = 1, \beta = 1, \lambda_P = 10.0, \lambda_{FM} = 10.0$. The experiments are conducted on two NVIDIA Tesla V100 GPUs.

3.2. Compression Performance Comparison

Baseline. We compare the compression performance with following typical approaches. For traditional codecs, widely used JPEG, JPEG2000 and HEVC-based BPG[3] are utilized for comparison. For exemplar learned image compression methods, we compare the proposed scheme with Minnen *et al.* [7] and HiFiC [10] which are the state-of-the-art methods optimized without GANs and with GANs, respectively. At last, our method is also compared to the state-of-the-art conceptual compression (named as DSTS) [3] and model without cross-channel hyperprior as variants for ablation study.

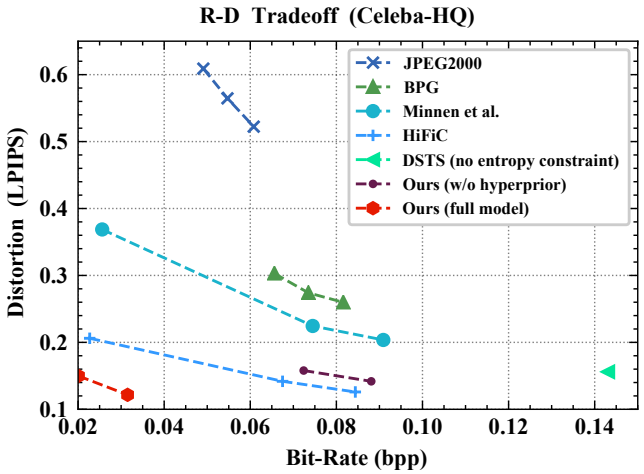

Fig. 4. **Rate-distortion curves aggregated over the Celeba-HQ under extremely low bitrate range.** Lower LPIPS score demonstrates better visual quality.

Qualitative results. The qualitative comparison results are shown in Fig. 3. Note that the bitrate of JPEG, BPG, Minnen *et al.* [7] almost reach the lowest within the limit. It can be seen that the proposed method achieves higher visual reconstruction quality and fidelity under extremely lower bitrate (average 0.0241 bpp) compared to baselines. In particular, compared with our models, the traditional codecs demonstrate severe degraded visual quality at higher bitrates (JPEG 9.9×, JPEG2000 2.5×, BPG 2.7×). Moreover, the decoded results from Minnen *et al.* [7] show over-smoothing and se-

[1]https://github.com/switchablenorms/CelebAMask-HQ
[2]https://github.com/FLIF-hub/FLIF
[3]https://bellard.org/bpg

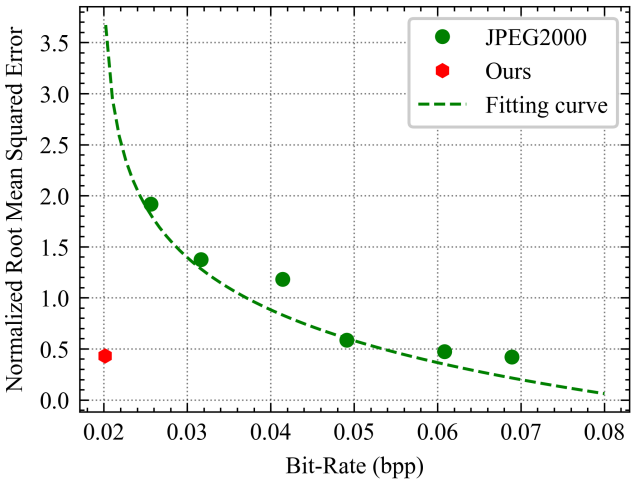

Fig. 5. The normalized root mean square error on facial landmark detection and bitrate of JPEG2000 [5] and proposed method. Our method achieves 65.1% bits saving at similar analysis accuracy.

vere distortion at similar bitrate. Despite cooperated with adversarial training and LPIPS [23] as perceptual distortion metric, at ultra-low bitrate range ($<$0.1 bpp), the reconstruction results appear apparent visual degradation and artifacts, leading to a less competitive model compared to ours.

Quantitative results. Fig. 4 shows RD curves over the publicly available Celeba-HQ[1] dataset by using LPIPS as the visual distortion metric at the ultra-low bitrate range ($<$0.1 bpp). The rate-distortion (R-D) graphs compare our model to existing representative compression schemes. Even though using LPIPS as distortion loss brings a comparison advantage to HiFiC, the results clearly show our model achieves the best perceptual quality score of LPIPS while reaching an extremely low bitrate ever than before, outperforming other state-of-the-art methods.

Ablation Study. As the ablation study for the proposed model, we also show the RD performance of the model which replaces the proposed cross-channel entropy model with an independent Gaussian density model, and the model from DSTS [3] without entropy constraint in Fig. 4. With fixed average bitrate 0.014 bpp of lossless coded structure layer, the results show incorporating proposed cross-channel hyperprior into entropy model obtains an average bits-saving of 62.5% over the non-hyperprior entropy model at similar visual quality, validating the effectiveness of the proposed cross-channel entropy model. Due to lacking entropy constraint, the rate of DSTS is almost fixed at an average 0.1413 bpp. For the texture layer, although the data volume for all semantic regions is 19 times of it in DSTS, the actual bitrate for encoding them is only 4 times than that in DSTS.

On the whole, our model achieves higher efficiency coding and better reconstruction by taking advantage of finer texture modeling and entropy-constrained training.

3.3. Advantages for Vision Tasks

The advantages of proposed method in support of vision tasks can be presented in following two aspects. On the one hand, under "compression then analysis" scenarios, our higher efficiency coding could benefit follow-up vision tasks performed on decoded images. For instance, we perform facial landmark detection on the decoded images from JPEG2000 and the proposed method and calculate the average normalized root mean squared error (NRMSE). As illustrated in Fig. 5, our method outperforms JPEG2000 by achieving a lower NRMSE of 0.432 under a lower bitrate of average 0.021 bpp. Particularly, our method can achieve 65.1% bits saving at similar analysis accuracy, which demonstrates the superiority of the proposed method towards vision tasks.

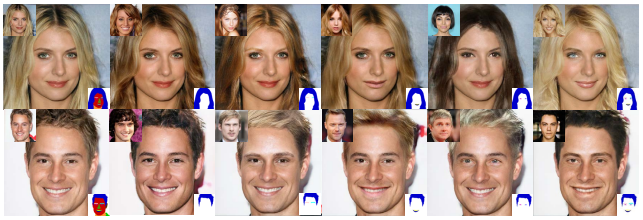

Fig. 6. The semantic-wise image manipulation results. Reference images are shown at the top left corner and manipulated semantic regions are presented at the bottom right corner. The original synthesized images are shown in first column and manipulated results in last columns.

On the other hand, benefited from the visual feature representations, conceptual coding has essential advantages over joint vision tasks in the compressed domain, corresponding to "analysis then compression" scenarios. In our approach, various visual features including structure, texture and semantic information can be applied to the analysis and content manipulation tasks directly without decoding, allowing higher efficiency and effectiveness. Furthermore, compared to previous conceptual coding [3], besides providing direct semantic labels, our method can perform finer content manipulation with semantic prior as shown in Fig. 6.

3.4. Generalization Discussion

So far we have demonstrated outstanding compression performance and versatility of proposed method on facial dataset. Fig. 7 shows the example cases of reconstruction results on ADE20K [21] dataset which consists of 150 semantic classes under the same training settings. The scenarios in ADE20K contains much more complex texture and semantic information, validating the generalization and advantages of the proposed joint semantic prior conceptual coding model. In essence, as a data-driven feature-based coding, our method can achieve better performance on domain-specific scenarios which appear structured visual characteristics.

4. CONCLUSION

This paper proposes a novel semantic prior modeling based conceptual coding approach which extracts semantic-wise deep representations to model texture distributions in an entropy modeling friendly form. Furthermore, we propose a cross-channel entropy model which exploits inter-channel

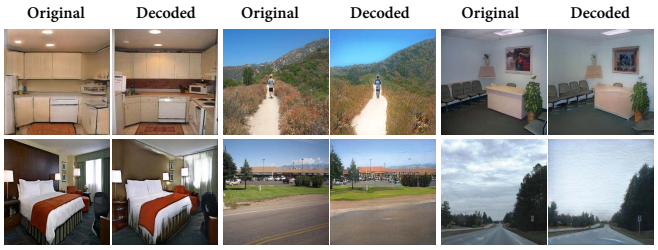

Fig. 7. The reconstruction cases of ADE20K dataset validate the generalization of proposed method.

correlation for accurate entropy estimation of semantic prior, leading to a high efficiency trainable model with entropy constraint. Qualitative and quantitative results demonstrate our method can perform extremely low bitrate image compression with high reconstruction quality and outperform the state-of-the-arts methods. The advantages of the proposed method over visual processing and understanding tasks are also analyzed and verified in our explorative experiments. As a future direction, we would like to investigate more efficient and versatile algorithms for general scenes and video coding.

5. REFERENCES

[1] Karol Gregor, Frederic Besse, Danilo Jimenez Rezende, Ivo Danihelka, and Daan Wierstra, "Towards conceptual compression," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[2] Jianhui Chang, Qi Mao, Zhenghui Zhao, Shanshe Wang, Shiqi Wang, Hong Zhu, and Siwei Ma, "Layered conceptual image compression via deep semantic synthesis," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, 2019.

[3] Jianhui Chang, Zhenghui Zhao, Chuanmin Jia, Shiqi Wang, Lingbo Yang, Jian Zhang, and Siwei Ma, "Conceptual compression via deep structure and texture synthesis," *arXiv preprint arXiv:2011.04976*, 2020.

[4] William B Pennebaker and Joan L Mitchell, *JPEG: Still image data compression standard*, Springer Science & Business Media, 1992.

[5] Majid Rabbani, "JPEG2000: Image compression fundamentals, standards and practice," *Journal of Electronic Imaging*, vol. 11, no. 2, pp. 286, 2002.

[6] Vivienne Sze, Madhukar Budagavi, and Gary J Sullivan, "High efficiency video coding (HEVC)," *Integrated Circuit and Systems, Algorithms and Architectures. Springer*, vol. 39, pp. 40, 2014.

[7] David Minnen, Johannes Ballé, and George D Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[8] Johannes Ballé, Valero Laparra, and Eero Simoncelli, "End-to-end optimized image compression," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2017.

[9] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack, "Context-adaptive entropy model for end-to-end optimized image compression," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2018.

[10] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson, "High-fidelity generative image compression," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[11] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka, "Sean: Image synthesis with semantic region-adaptive normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[14] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston, "Variational image compression with a scale hyperprior," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2018.

[15] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*, pp. 1–4. Springer, 2009.

[16] Thomas M Cover, *Elements of information theory*, John Wiley & Sons, 1999.

[17] Yochai Blau and Tomer Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *Proceedings of International Conference on Machine Learning (ICML)*, 2019.

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of European Conference on Computer Vision (ECCV)*. Springer, 2016.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[20] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2015.

[23] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.