Semantic Importance-Based Deep Image Compression Using a Generative Approach

Xi Gu[1], Yuanyuan Xu[1](✉), and Kun Zhu[2]

[1] Hohai University, Nanjing, China
yuanyuan_xu@hhu.edu.cn
[2] Nanjing University of Aeronautics and Astronautics, Nanjing, China

Abstract. Semantic image compression can greatly reduce the amount of transmitted data by representing and reconstructing images using semantic information. Considering the fact that objects in an image are not equally important at the semantic level, we propose a semantic importance-based deep image compression scheme, where a generative approach is used to produce a visually pleasing image from segmentation information. A base-layer image can be reconstructed using a conditional generative adversarial network (GAN) considering the importance of objects. To ensure that objects with the same semantic importance have similar perceptual fidelity, a generative compensation module has been designed, considering the varying generative capability of GAN. The base-layer image can be further refined using residuals, prioritizing regions with high semantic importance. Experimental results show that the reconstructed images of the proposed scheme are more visually pleasing compared with relevant schemes, and objects with a high semantic importance achieve both good pixel and semantic-perceptual fidelity.

Keywords: Image Compression · Semantic image coding · Scalable coding

1 Introduction

Image compression is an essential and fundamental topic for efficient image storage and transmission across modern networks. Traditional image codecs, such as WebP [8], JPEG [23], JPEG2000 [22] and BPG [4] use a hybrid image coding framework, where basic coding blocks of an image are transformed, quantized and entropy coded. Block-based coding introduces blocking effects, especially at a low bitrate. Complicate dependency among coding modules makes the codec difficult to optimize as a whole. Due to the strong representation capabilities of deep learning, learned image compression [3] based on deep neural networks, or deep image compression, has received widespread attention in recent years. The learned image compression framework can be optimized end-to-end, showing competitive performance compared with traditional codecs.

To further reduce the amount of transmitted data, image semantic coding [11,12] can be performed. As traditional and learned image codecs, even the

S. Rudinac et al. (Eds.): MMM 2024, LNCS 14555, pp. 70–81, 2024.
https://doi.org/10.1007/978-3-031-53308-2_6

lightfiled image codec [14], are typically optimized for pixel fidelity, such as MSE (Mean Square Error) and PSNR (Peak Signal to Noise Ratio), semantic image coding aims at maintaining semantic-perceptual fidelity [11]. According to [21], perceptual fidelity metrics refer to objective metrics of human viewing experience, while semantic fidelity is the semantic difference between original and reconstructed image such as the difference in object detection accuracy. A good choice to optimize towards semantic-perceptual fidelity is to use generative adversarial network (GAN) [7], since GAN can capture both global semantic information and local texture and is powerful to produce a visually pleasing image even with semantic labels only. In recent years, GAN has been used to improve image coding efficiency [1,2,11,12,17].

Regarding pixel fidelity, the work in [2] has proposed a generative image compression framework based on deep semantic segmentation (DSSLIC), where a down-sampled version of the image and a semantic segmentation map need to be transmitted and used to guide the image generation. Residual can be sent to refine the generated image as well. Similar as DSSLIC, the work in [15] utilizes semantic segmentation map to compress image and the redundancy in the spatial dimension of feature map is addressed by using octave convolution. To avoid using extra bits for segmentation map, the work [10] trained a segmentation network from the up-sampled version of the down-sampled image to obtain an accurate segmentation map at both encoder and decoder.

Targeting for perceptual quality, an image compression system based on GANs has been designed in [1] for extreme low bitrates, where quantized extracted features are fed to GANs. In [17], a neural compression scheme with a GAN has been proposed, which can be optimized to yield reconstructions with high perceptual fidelity that are visually close to the input. The above works optimize towards fidelity for the whole image. However, objects in the same image are different in terms of semantic importance. For example, when viewing portrait photos, humans are more important than buildings semantically. In [1], a selective generative compression scheme has been proposed as well only using GAN to generate unimportant regions, while user-defined regions were preserved with fine details. This scheme can only support two levels of semantic importance. The semantic image coding scheme proposed by Huang *et al.* [11] quantized features of different semantic concepts adaptively, where bit allocation among semantic concepts was determined using a reinforcement learning algorithm. GAN was used at the decoder to reconstruct images from quantized features. The scheme was optimized in terms of semantic-perceptual loss without considering pixel fidelity.

Pixel fidelity and perceptual fidelity are both important, especially for regions with high semantic importance. Optimizing for both pixel and perceptual loss, Huang *et al.* [12] proposed a deep image semantic coding scheme which uses both quantized extracted features and the segmentation map for image generation, and residuals were transmitted to restore fine details. However, varying importance of different regions are not considered. In this work, we propose a generative image compression that can support any levels of semantic importance, and both pixel-level and semantic-perceptual distortion are evaluated in the training of the proposed work. In the proposed approach, objects with higher semantic importance are evaluated weighting more on pixel-level accuracy, while the unimportance regions are mainly concerned with perceptual accuracy. The proposed scheme is a scalable coding scheme as the work [24], where the base layer consists of segmentation map and quantized features, and the enhancement layer comprises coded residuals. Like multiple description coding [13,16,26] that combats transmission losses via source coding, a scalable image coding scheme can provide a certain degree of robustness tolerating the loss of enhancement layer.

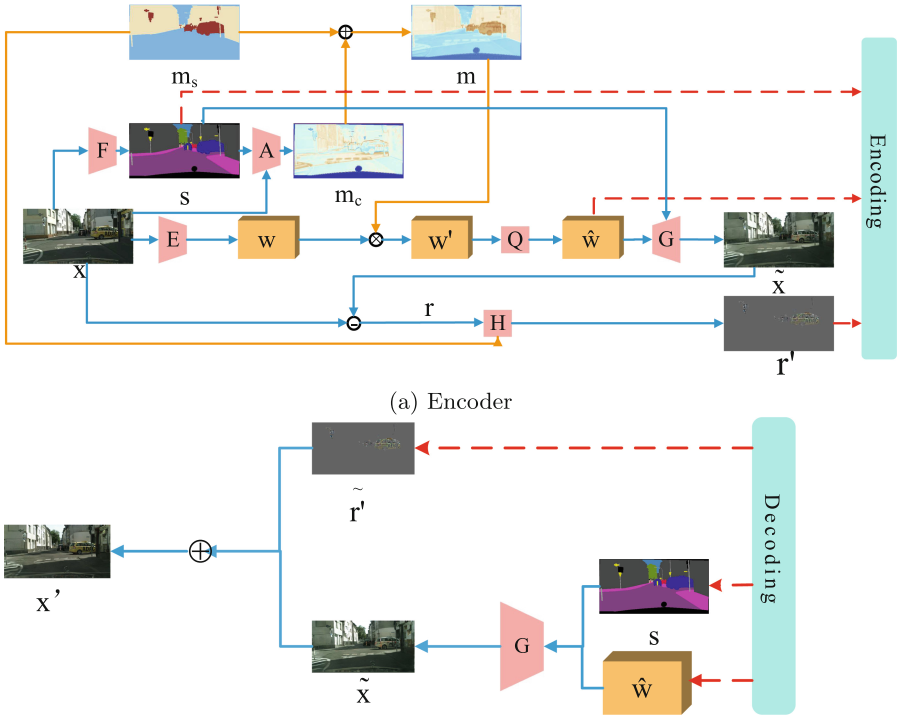
The main contributions of this work are summarized as follows:

- We propose a generative deep image compression framework considering varying semantic importance of objects. Guided by an importance map of an image of objects, a base-layer image can be reconstructed using a conditional GAN using the extracted feature and segmentation map. Prioritized residual coding is then performed that can be used to refine important regions of the base-layer image.
- Considering the varying generative capability of GAN for objects with different characteristics, a generative compensation module has been designed to ensure that objects with the same semantic importance have similar percentual fidelity.
- The proposed framework has been optimized and evaluated using both the weighted pixel-level distortion and adversarial loss concerning perceptual loss. As a result, objects with a high semantic importance can achieve both good pixel and semantic-perceptual fidelity, and regions with lower importance are visually pleasing.

2 Semantic Importance-Based Deep Image Compression

Illustration of the proposed framework is shown in Fig. 1. Encoder E extracts the latent representation w from the input image x. w is scaled according to the importance map m and quantized by a quantizer Q. The generator G uses the scaled and quantized latent code $\hat{w}$, and the segmentation map s from the segmentation network F, to produce a base-layer reconstructed image, $\widetilde{x}$. In the importance map generation module, a pre-defined semantic importance map, m_s, and a generative compensation map, m_c, are fused to form a final map, m. m_c is generated by the generative-compensation network, A, using s and x as the input. In the residual coding module, a residual processor H conducts prioritized processing for residual, $r = x - \widetilde{x}$, according to m_s. The processed residual r' is then compressed and transmitted.

As shown in Fig. 1(a), the image is layered coded. The base layer consists of lossless compressed $\hat{w}$ and s, while the enhancement layer includes lossy compressed r'. In Fig. 1(b), the generator G can use decoded $\hat{w}$ and s to generate an image $\hat{x}$. With an available enhancement layer, the decoded residual $\hat{r}$ is added to $\hat{x}$ to obtain an enhanced image $x' = \widetilde{r} + \widetilde{x}$.



(a) Encoder

(b) Decoder

Fig. 1. The proposed framework. (a) The encoder consists of segmentation network F, encoder network E, generative-compensation network A, quantizer Q, generator G, and residual processor H. (b) The decoder generates a base-layer image using the segmentation map and quantized latent code, while an enhanced image can be obtained combining the residual.

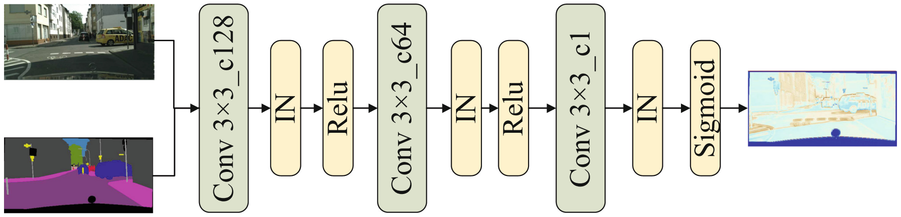


Fig. 2. The network architecture of generative-compensation module.

2.1 Model Learning

In this framework, a pretrained segmentation network, PSPNet [25] is used, and a fixed quantizer maps features to the nearest quantization centers which has a soft quantization version during back propagation as in [12]. The conditional GAN adopts the same architecture as the work in [12]. The generator G takes quantized features $\hat{w}$ form a distribution of $p_{\hat{W}}$ and segmentation map s as the input, and tries to generate an image that can cheat the discriminator D minimizing $L_G = E_{\hat{w} \sim p_{\hat{W}}}[-log(D(G(\hat{w}, s), s))]$ [12]. The discriminator G tries to differentiate the generated images from the original images, minimizing $L_D = E_{\hat{w} \sim p_{\hat{W}}}[-log(1 - D(G(\hat{w}, s), s))] + E_{x \sim p_{X|s}}[-log(D(x, s))]$ [12]. Besides fixed segmentation network and quantizer, the residual processor H is rule-based. Therefore, we need to jointly train the models of the encoder network E, the generative-compensation network A, and the generator G.

The loss function for training needs to consider both pixel fidelity and semantic-perceptual fidelity, weighting more on pixel fidelity for semantic important areas. If an image can be divided into N regions according to semantic importance, a region C_k $(k = 1, 2, ..., N)$ has a semantic weight of w_k $(w_k \in [0, 1])$ with $w_1 \geq w_2 \geq ... \geq w_k$. A larger weight means it is more important. The overall loss function can be expressed as follows:

$$L = \lambda_1 WMSE + \lambda_1 d(G(\hat{w}, s), x) + L_G + L_D + \lambda_2 R(\hat{w}), \tag{1}$$

where $x(i, j)$, $G(\hat{w}, s)(i, j)$, $d()$, $R()$, λ_1 and λ_2 are the pixel values at position (i, j) in the original image and reconstructed base-layer image, the feature-level distortion function using VGG network, rate function, weights for pixel-perceptual distortion and rate, respectively. Weighted MSE (WMSE) is defined as

$$WMSE = \sum_{k=1}^{N} w_k \cdot \frac{1}{|C_k|} \sum_{p(i,j) \in C_k} [x(i, j) - G(\hat{w}, s)(i, j)]^2, \tag{2}$$

where $p(i, j) \in C_k$ means that the pixel at position (i, j) is in region C_k, and $|C_k|$ are the number of pixles in C_k.

Since the segmentation network is pre-trained, and the optimization is conducted for the base-layer, the rates of s and r' are not included in the loss function. Besides the adversarial loss, the feature-level distortion function using VGG network is introduced as in [12] as another semantic fidelity metric. Combining weighted MSE and unweighted feature loss, the semantic important regions are impacted more by the pixel fidelity.

2.2 Importance Guided Bitrate Allocation

The semantic importance distribution of an image changes for different applications. For example, vehicles are most semantic important for vehicle detection but not for pedestrian detection. Therefore, we assume a predefined semantic importance map m_s is available. Guided by m_s, a bitrate allocation scheme is then designed.

Generative Compensation. Originally, we only use m_s to guide bitrate allocation. However, we observe that objects with the same semantic weight in the generative image do not have the same perceptual quality, due to varying generative capability of GAN. For example, sky and road in an image can have the same semantic weight, but the perceptual quality of sky with simple texture is better than that of a road. Therefore, a generative-compensation module has been added to compensate the generative capability.

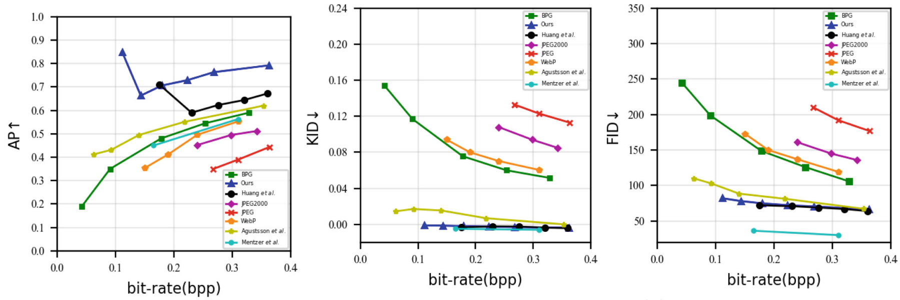
The network architecture of the generative-compensation module is shown in Fig. 2. This generative-compensation module, A, is a network with three convolutional layers generating a pixel-level importance map m_c, whose values are within $[0, 1]$. The input of the model is the original image and segmentation map. Including semantic labels can make m_c reflect the semantic and edge information of the image better.

Fused Importance Map. Two maps, m_s and m_c can be fused into one map m to guide the bitrate allocation of different regions, with

$$m = \lambda * m_s + (1 - \lambda) * m_c, \tag{3}$$

where λ is a weight factor. The value range for m is within $[0, 1]$ as well.

For latent code w with C feature maps with a dimension of $H * W$, m is rescaled to $H*W$. Each element in a feature map is multiplied by a corresponding important weight in m. The multiplication is performed at the element level for each channel. By scaling the latent code with a weight between $[0, 1]$ followed by quantization with fixed quantization centers, the coefficients with a small weight are represented with a lower quantization precision and thus a lower bitrate. In this way, bitrate is allocated according to importance.



(a) Average precision

(b) Kernel Inception Distance

(c) Fréchet Inception Distance score

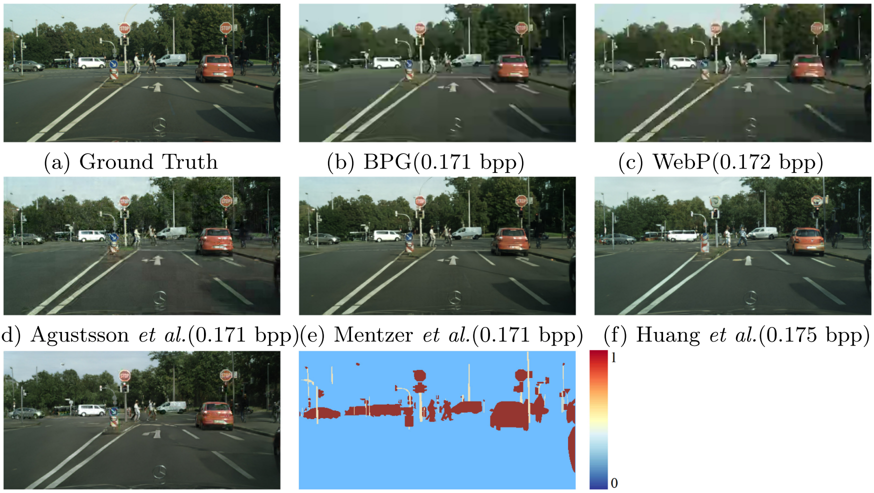
Fig. 3. Comparsion of JPEG, JPEG2000, WebP, BPG, Huang *et al.* [12], Agustsson*et al.* [1], and our model in terms of objective metrics. The arrow (↓) on the y-axis indicates that a lower value is better, and (↑) indicates that a higher value is better.

2.3 Prioritized Residual Coding

Residual coding is conducted considering prioritizes as well. A residual processor skips part of residuals according to m_s yielding a processed residual r'. For example, with a low bit rate, we only keep the residual of most important area. If the ω is the semantic weight threshold for residual skipping, r' can be expressed as

$$r'(i,j) = \begin{cases} r(i,j), & if\ x(i,j) \in C_k\ AND\ w_k \geq \omega \\ 0, & otherwise \end{cases} \tag{4}$$

where $r(i,j)$ and $r'(i,j)$ are the values of residual and processed residual at position (i,j), respectively. Different ω is used for varying targeting bitrates. The residual image r' is encoded by the lossy BPG encoder with different quality factors for varying bitrate.



(a) Ground Truth (b) BPG(0.171 bpp) (c) WebP(0.172 bpp)

(d) Agustsson *et al.*(0.171 bpp)(e) Mentzer *et al.*(0.171 bpp) (f) Huang *et al.*(0.175 bpp)

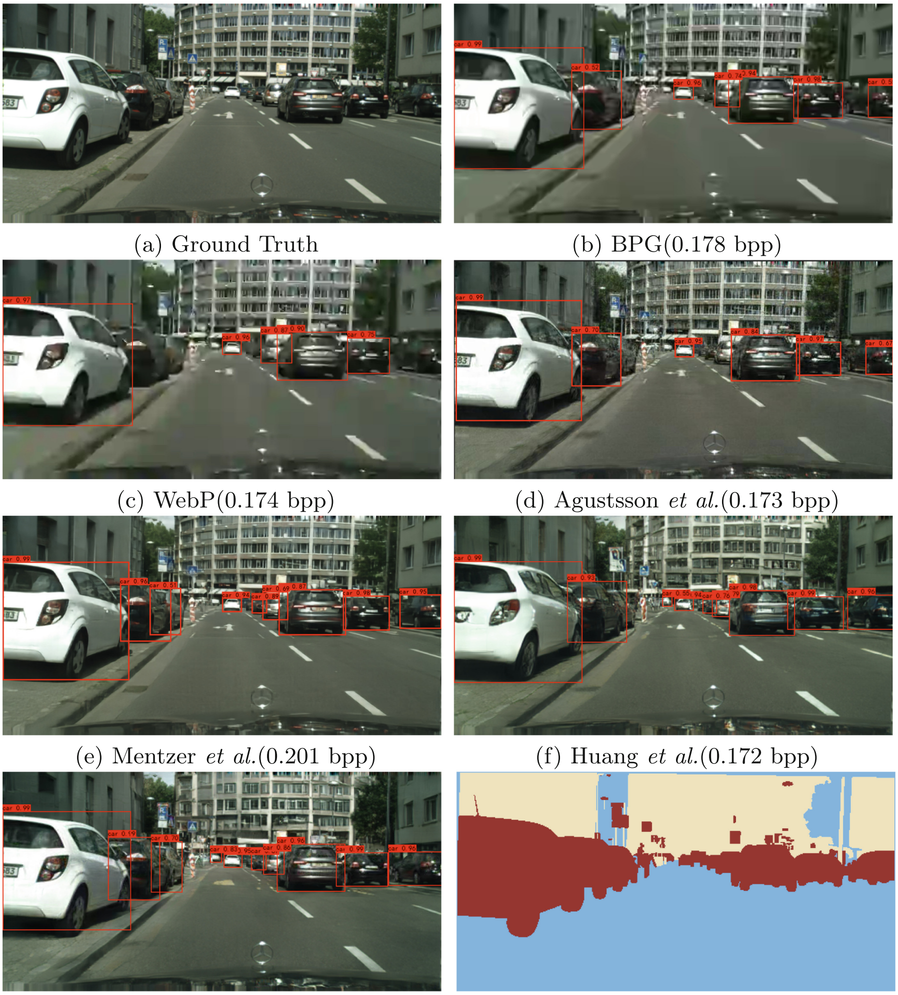
(g) Ours(0.169 bpp) (h) Semantic importance map

Fig. 4. Visual comparison of reconstructed images of WebP, BPG,Agustsson *et al.* [1], Mentzer *et al.* [17], Huang *et al.* [12] and our model.

3 Experiment

3.1 Experiment Settings

The proposed model is trained with the Cityscapes dataset [6] with images downsampled to $256 * 512$. The training set consists of 2975 images, and our performance is evaluated using the test set. The latent representation w is represented using 128 channels. The semantic importance maps are generated according to



(a) Ground Truth

(b) BPG(0.178 bpp)

(c) WebP(0.174 bpp)

(d) Agustsson *et al.*(0.173 bpp)

(e) Mentzer *et al.*(0.201 bpp)

(f) Huang *et al.*(0.172 bpp)

(g) Ours(0.171 bpp)

(h) Semantic importance map

Fig. 5. Example of using reconstructed images of WebP, BPG,Agustsson *et al.* [1], Mentzer *et al.* [17], Huang *et al.* [12] and our model for vehicle detection.

annotations. Three levels of semantic importance are considered, whose values are set to 1, 0.67, and 0.33, respectively. The first level includes vehicles, objects, and humans, while the second one contains constructions and flats. The remaining belong to the third level. λ_1 and λ_2 in the loss function of (1) are set to 10 and 1, respectively. λ for the importance map in (3) is set to 0.6. The model is trained with an initial learning rate of $3 * 10^{-4}$ which gradually reduces to 0 after 50 epochs.

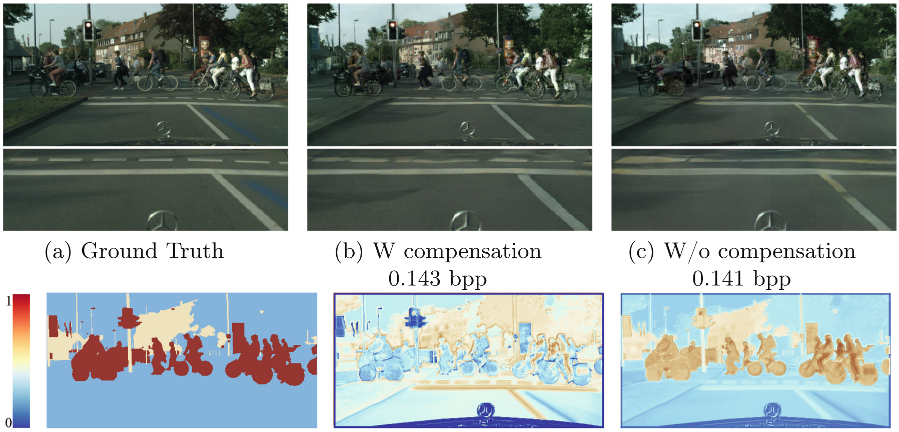
The segmentation map and latent code are lossless compressed using vector graph and arithmetic coding consuming 0.01bpp and 0.1bpp, respectively. ω in

(4) is set to 0.9, 0.6, 0.3 for bitrates that are below 0.2bpp, $[0.2, 0.4]$bpp, and above 0.4bpp, respectively. Residual is compressed by the lossy BPG encoder under quality factors of $\{40, 35\}$ and $\{40, 35, 30\}$ for cases with a bitrate below and above 0.2bpp, respectively. The proposed scheme is compared with JPEG, JPEG2000, WebP, and BPG, Agustsson *et al.* [1], Mentzer *et al.* [17] and Huang *et al.*'s generative semantic image compression scheme in [12] which considers both pixel and perceptual fidelity.

3.2 Results

Objective Evaluation. For semantic fidelity, the proposed scheme is compared with the latest traditional codecs, which are JPEG2000, WebP and BPG, JPEG, and learning based codecs including Agustsson *et al.* [1] and Mentzer *et al.* [17] for performance on the object detection task. The simulation was conducted on an NVIDIA GeForce RTX 3080, based on the PyTorch [19] platform. A yolo-v3 network [20] pre-trained on COCO dataset is used to detect vehicles in reconstruction images of different methods. Average Precision (AP) [18] is used to measure the performance of vehicle detection at different bitrates, where the IOU threshold is set to 0.5. Figure 3 show performance comparison. From Fig. 3, the proposed scheme achieves the highest AP at all bit-rates, since the proposed scheme is designed to support image compression with different semantic importance. Vehicles in the images are assigned with the highest semantic weight, resulting in both good pixel and perceptual fidelities. The highest AP is obtained at the lowest bitrate, because the unimportant regions are fully generated at the lowest bitrate without residual coding which further differentiates salient objects. The AP drops due to the blurry effect of BPG residual coding, and increases with more bitrate used for residual coding.

The semantic-perceptual quality comparison for the whole images is presented as well. Kernel Inception Distance (KID) [5] evaluates the feature space distortion, while Fréchet Inception Distance score (FID) [9] is used to measure the quality of GAN generation. As shown in Fig. 3, the generative schemes have better performance than the traditional codec and Agustsson *et al.* [1] in terms of KID and FID, since these GAN-based approaches are optimized considering semantic-perceptual quality. The performance of the proposed and the scheme in [12] are almost the same. Mentzer *et al.* [17] performs better on FID than our method, since the proposed method focuses on maintaining fidelity for important regions only which does not show competitive generating capability for the whole image.



(a) Ground Truth (b) W compensation 0.143 bpp (c) W/o compensation 0.141 bpp

(d)Original importance (e)Compensation (f) Fused map

Fig. 6. Visual comparison with/without the generative compensation module.

Subjective Evaluation. Except the AP, the objective evaluation above do not differentiate regions with varying importance, and thus subjective evaluation is conducted as well. For perceptual quality, Fig. 4 shows the visual comparison. It can be observed that, even with a lower bitrate, the reconstructed images of the generative approaches are more visual pleasing than those of traditional codecs, Agustsson*et al.* [1] and Mentzer *et al.* [17]. Compared with Huang *et al.*'s method, the regions with high semantic importance in the proposed scheme exhibit a lower level of degradation from the original content, such as the stop sign. In our scheme, semantically unimportant area, such as trees in the background, deviates more from the original content, but is still visually pleasing. For semantic quality, an example for vehicle detection is present in Fig. 5. As shown in the semantic importance map in Fig. 5(h), regions of vehicles have high semantic importance in the proposed scheme. These regions are coded with more bitrate resulting in both good pixel and semantic-percentual fidelity. Compared with other schemes, more correct vehicles are detected using the reconstructed image of the proposed scheme with a lower bitrate.

Effectiveness of Generative Compensation Module. Reconstructed images of the proposed scheme with and without the generative compensation module are shown in Fig. 6. In the original importance map in Fig. 6(d), sky and road have the same semantic weight. However, generative capability of GAN for these objects are different, where sky is more visually pleasing than the road without compensation. With compensation module, the regions with more texture details are compensated as shown in Fig. 6(e), achieving similar perceptual quality for regions with the same semantic importance.

4 Conclusion

In this paper, a generative image compression framework based on semantic importance has been proposed. A predefined semantic importance map and a map from the generation compensation module are combined to guide the compression of the latent representation extracted by the encoder. Both adversarial loss, perceptual loss and semantically weighted pixel-level loss are considered for end-to-end training. The goal is to improve coding efficiency while maintaining a good pixel and semantic-percentual fidelity for regions with high semantic importance, and a reasonable perceptual fidelity for the less important regions. The experimental results verify the effectiveness of the proposed scheme.

References

1. Agustsson, E., Tschannen, M., Mentzer, F., Timofte, R., Gool, L.V.: Generative adversarial networks for extreme learned image compression, pp. 221–231 (2019)
2. Akbari, M., Liang, J., Han, J.: DSSLIC: deep semantic segmentation-based layered image compression. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2042–2046 (2019)
3. Balle, J., Laparra, V., Simoncelli, E.P.: End-to-end optimization of nonlinear transform codes for perceptual quality. In: Picture Coding Symposium (PCS), pp. 1–5. IEEE, Nuremberg, Germany (2016). https://doi.org/10.1109/PCS.2016.7906310
4. Bellard., F.: BPG Image format
5. Binkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. ArXiv:1801.01401 (2018)
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. IEEE (2016)
7. Goodfellow, I., et al.: Generative adversarial nets. In: Neural Information Processing Systems (2014)
8. Google: WebP Image format (2010). https://developers.google.com/speed/webp/
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems 30 (2017)
10. Hoang, T.M., Zhou, J., Fan, Y.: Image compression with encoder-decoder matched semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 160–161 (2020)
11. Huang, D., Gao, F., Tao, X., Du, Q., Lu, J.: Towards semantic communications: deep learning-based image semantic coding. IEEE J. Selected Areas Commun. **41**(1), 55–71 (2022)
12. Huang, D., Tao, X., Gao, F., Lu, J.: Deep learning-based image semantic coding for semantic communications. In: IEEE Global Communications Conference (GLOBECOM), pp. 1–6 (2021)
13. Liu, M., Zhu, C., Wu, X.: Index assignment design for three-description lattice vector quantization. In: 2006 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 4-pp. IEEE (2006)
14. Liu, Y.Y., Zhu, C., Mao, M.: Light field image compression based on quality aware pseudo-temporal sequence. Electron. Lett. **54**(8), 500–501 (2018)

15. Liu, Z., Meng, L., Tan, Y., Zhang, J., Zhang, H.: Image compression based on octave convolution and semantic segmentation. Knowl.-Based Syst. **228**, 107254 (2021)
16. Meng, L., Li, H., Zhang, J., Tan, Y., Ren, Y., Zhang, H.: Convolutional autoencoder based multiple description coding network. KSII Trans. Internet and Inform. Syst. (TIIS) **14**(4), 1689–1703 (2020)
17. Mentzer, F., Toderici, G.D., Tschannen, M., Agustsson, E.: High-fidelity generative image compression. Adv. Neural. Inf. Process. Syst. **33**, 11913–11924 (2020)
18. Padilla, R., Netto, S.L., Silva, E.: A survey on performance metrics for object-detection algorithms. In: International Conference on Systems, Signals and Image Processing (IWSSIP) (2020)
19. Paszke, A., et al.: Automatic differentiation in pytorch (2017)
20. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv e-prints (2018)
21. Shi, J., Chen, Z.: Reinforced bit allocation under task-driven semantic distortion metrics. In: IEEE International Symposium on Circuits And Systems (ISCAS), pp. 1–5 (2020)
22. Skodras, A., Christopoulos, C., Ebrahimi, T.: The JPEG 2000 still image compression standard. IEEE Signal Process. Mag. **18**(5), 36–58 (2001)
23. Wallace, Gregory, K.: The JPEG still picture compression standard. Communications ACM **34**(4), 30–44 (1991)
24. Zhang, D., et al.: Exploring resolution fields for scalable image compression with uncertainty guidance. IEEE Trans. Circ. Syst. Video Technolpp. (2023). https://doi.org/10.1109/TCSVT.2023.3307438
25. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
26. Zhao, L., Bai, H., Wang, A., Zhao, Y.: Multiple description convolutional neural networks for image compression. IEEE Trans. Circuits Syst. Video Technol. **29**(8), 2494–2508 (2018)