Guided Diffusion for the Extension of Machine Vision to Human Visual Perception

Takahiro Shindo Yui Tatsumi Taiju Watanabe Hiroshi Watanabe
Waseda University

***Abstract*—Image compression technology eliminates redundant information to enable efficient transmission and storage of images, serving both machine vision and human visual perception. For years, image coding focused on human perception has been well-studied, leading to the development of various image compression standards. On the other hand, with the rapid advancements in image recognition models, image compression for AI tasks, known as Image Coding for Machines (ICM), has gained significant importance. Therefore, scalable image coding techniques that address the needs of both machines and humans have become a key area of interest. Additionally, there is increasing demand for research applying the diffusion model, which can generate human-viewable images from a small amount of data to image compression methods for human vision. Image compression methods that use diffusion models can partially reconstruct the target image by guiding the generation process with a small amount of conditioning information. Inspired by the diffusion model's potential, we propose a method for extending machine vision to human visual perception using guided diffusion. Utilizing the diffusion model guided by the output of the ICM method, we generate images for human perception from random noise. Guided diffusion acts as a bridge between machine vision and human vision, enabling transitions between them without any additional bitrate overhead. The generated images then evaluated based on bitrate and image quality, and we compare their compression performance with other scalable image coding methods for humans and machines.**

***Index Terms*—Image Coding for Machines, Scalable Image Coding, Guided Diffusion**

I. Introduction

Humans and machines process information in images to support their respective vision. Image compression is a technology designed to identify the critical information needed for these visions and to represent images using the minimal amount of data necessary. This technology is vital for efficient image transmission and storage. In recent years, the rapid advancements in deep learning have increased the demand for image compression methods tailored not only for human perception, as traditionally emphasized, but also for image recognition. This shift is driven by the growing use of models for image recognition tasks like object detection and segmentation. To address this need, the field of Image Coding for Machines (ICM) has emerged as an important area of research [1]–[4].

Studies have shown that low-frequency components in images play a significant role in human perception, which is why high-frequency components are often discarded during image compression. For example, methods like JPEG [5] and JPEG2000 [6] use Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT), respectively, to process images in the frequency domain and eliminate unnecessary frequency components. This approach reduces the amount of data required to represent the images. In contrast, research has indicated that object regions and edges within images are critical for image recognition. Consequently, machine-oriented image compression methods have been developed that discard irrelevant information while preserving these key features in the image [7]–[10]. These differences result in distinct decoded images depending on whether they are optimized for human or machine vision. Specifically, decoded images intended for machine vision are typically unsuitable for human vision. Therefore, the development of scalable image coding methods that address these differences—allowing for optimized decoding for both human and machine use—is a promising area of research [11]–[19]. Such methods could effectively meet the diverse demands of human and machine image processing.

In scalable image coding, bridging the gap between human and machine vision is crucial. Decoded images for machine vision are typically evaluated based on image recognition accuracy. In contrast, decoded images for human perception are assessed using metrics like PSNR and SSIM, which measure the fidelity of pixel value reproduction, as well as LPIPS [20] and FID [21], which correlate more closely with human subjective evaluation. To address these dual objectives, scalable image coding methods have been developed to achieve high image recognition accuracy with faithful image reconstruction. For human-oriented decoding, image quality is evaluated in terms of PSNR, aiming to reproduce the original image as closely as possible [22]–[27]. However, this approach requires a substantial amount of information to reconstruct the original image, leading to a significant increase of data in the extension from machine vision to human vision.

In this paper, we propose a novel method to eliminate this additional information requirement, providing a seamless bridge between human and machine vision. Drawing inspiration from the success of diffusion models in recent image compression techniques, we integrate diffusion models [28] into image coding frameworks for both human perception and image recognition models. Specifically, the diffusion model, guided by the decoded image for machine vision, generates images suitable for human perception without the need for extra data. Through experiments, we evaluate the bitrate and image quality of the generated images and demonstrate their effectiveness for human perception, offering a promising solution to unify the needs of both human and machine vision

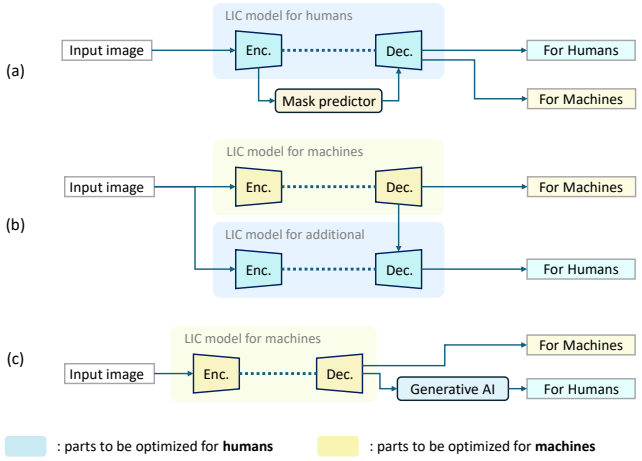

Fig. 1. Image processing in scalable image compression methods: (a) ICMH-Net, (b) ICMH-FF, (c) Ours.

in scalable image coding.

II. Related Works

A. Image Coding for Human Perception and Machine Vision

There are two primary types of scalable image coding methods designed for both human and machine use. The first approach builds on image compression methods optimized for human perception. When decoding images for machine vision, this removes image information that is unnecessary for machines. Examples of this approach include ICMH-Net [22] and Adapt-ICMH [23]. These methods leverage Learned Image Compression (LIC) [29]–[32], a neural network-based image compression framework. By incorporating components optimized for image recognition into a LIC model tailored for human perception, they enable efficient image compression for machine vision. The image processing flow of ICMH-Net is illustrated in Figure 1-(a). A mask predictor is employed to identify which image features are necessary for image recognition and which are not. By masking and removing the unnecessary features, the image can be decoded for machine vision using less information than is required for human perception.

The second approach uses image compression methods optimized for image recognition as a foundation and incorporates additional information when decoding images for human perception. This framework includes ICM method and techniques for compressing supplementary information. VVC+M is a scalable image coding method based on ICM model [26]. In this method, the image is first reconstructed from the decoded features intended for machine vision. This reconstructed image contains only the image information necessary for image recognition. The difference between this reconstructed image and the original image is then compressed using the VVC-inter codec [33], effectively treating the missing information for human perception as additional data. ICMH-FF, on the other hand, combines an ICM method called SA-ICM [34], [35] with a LIC model for compressing the additional information [24]. The image processing flow of this method is shown in Figure 1-(b). In this figure, the top layer represents the image compression process for machine vision, while the bottom layer shows the process for compressing the additional information. To evaluate the performance of image compression for machine vision, VVC+M uses an object detection model, while ICMH-FF utilizes both an object detection model and a segmentation model. For evaluating image compression performance aimed at human perception, both methods use PSNR as the metric, focusing on faithfully reproducing the original image.

B. Diffusion-based Image Compression

Diffusion models [28] have gained significant attention as a new approach to image generation, offering a promising alternative to Generative Adversarial Networks [36]. Among these, Stable Diffusion [37] has become widely adopted due to its user-friendly design. By using text prompts or semantic maps as input conditions, the output image can be effectively controlled. The introduction of ControlNet [38] has further advanced this capability by allowing the diffusion process to be guided with arbitrary conditions, making it even easier to generate desired images. This progress in controlling the image generation process has spurred a growing interest in applying diffusion models to image compression tasks [39].

PIC [40] is an image compression method that leverages a diffusion model to generate images, using a text prompt as the condition. The bitrate required to decode the image corresponds to the amount of information contained in the text prompt. PICS [40] expands on this approach by incorporating sketch images alongside text to control the diffusion process. The spatial constraints introduced by the sketch images enable precise control over the shapes and positions of objects and backgrounds. In PICS, the image bitrate is determined by the combined information in the text prompt and sketch image. Additionally, other studies have proposed methods for controlling the generated image's characteristics, such as using a color map to define the color scheme or low resolution image of the original image to guide the generation process.

Diffusion-based image compression methods are not well-suited for faithfully reproducing the original image and have been reported to perform worse than LIC and traditional rule-based codecs like JPEG, HEVC [41], and VVC [33] in metrics such as PSNR and SSIM. However, numerous studies have demonstrated that these methods outperform existing image compression techniques in evaluation indices like LPIPS and FID, which are believed to align more closely with human subjective perception. Therefore, diffusion-based image coding methods are expected to be an image compression method for human perception.

III. Proposed Method

A. Generative processes conditioned on ICM method.

We use the decoded image designed for machine vision as a condition to generate an image suitable for human vision. We adopt SA-ICM [34] as the image compression

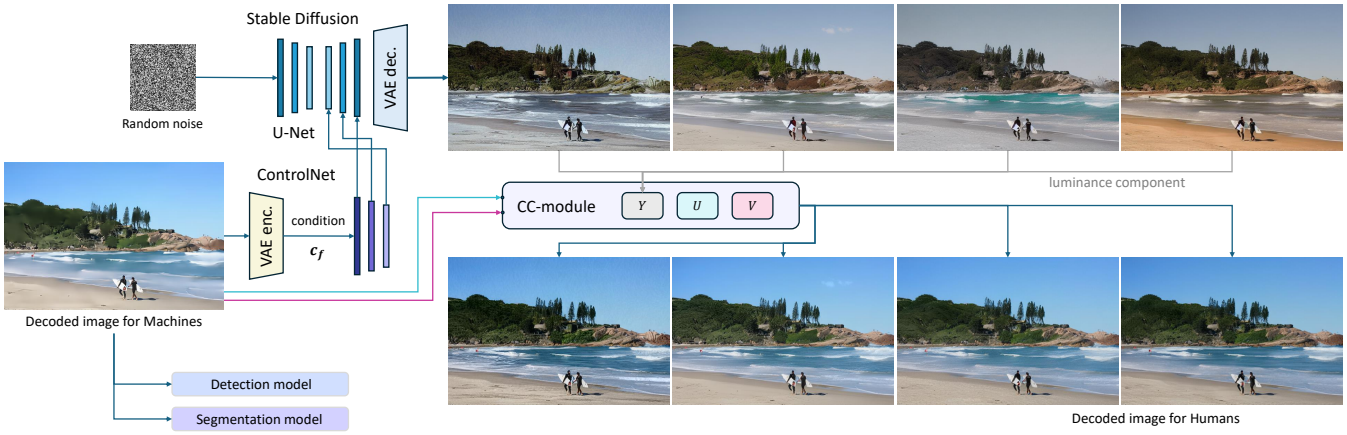

Fig. 2. Image processing flow in the proposed method. The ICM method serves as a condition to generate images optimized for human perception. The generated image is then input into the CC-module, which restores color elements approximating the original image.

method for machine vision, which is an image compression method for object detection and segmentation models. This method decodes the contours of objects within an image while discarding other textures. The resulting decoded image retains the position and size of objects and the background but lacks detailed textures. An example of such a decoded image is shown in Figure 3-(b). In the decoded image intended for machines, for instance, several elephants are recognizable in the center of the image, but its detailed patterns are not visible.

To restore the textures lost during compression, we leverage generative AI. The proposed image processing flow is shown in Figure 2. The generative AI in our proposed method combines Stable Diffusion [37] and ControlNet [38]. Stable Diffusion serves as the diffusion model, while ControlNet is used to guide the generative process. The ControlNet is fed with the SA-ICM decoded image and trained to regain the texture lost due to compression. The loss function for training ControlNet is defined as follows:

$$\mathcal{L} = \mathbb{E}_{z_0,\mathbf{t},\mathbf{c_t},\mathbf{c_f},\epsilon\sim\mathcal{N}(0,1)}[\|\epsilon - \epsilon_\theta(z_t,\mathbf{t},\mathbf{c_t},\mathbf{c_f})\|_2^2]. \quad (1)$$

In (1), $\mathbf{t}$ represents the time step, z_0 denotes the original latent, and z_t represents the latent after adding noise at step $\mathbf{t}$. $\mathbf{c_t}$ and $\mathbf{c_f}$ refer to the text prompt and image condition, respectively. In the training process of the proposed method, $\mathbf{c_t}$ is an empty string, and $\mathbf{c_f}$ is derived from the features of the decoded image from SA-ICM. By predicting the noise present in the latent, it becomes possible to generate images starting from random noise. Training ControlNet to transform images designed for recognition models into images suitable for human perception allows seamless extension from machine vision to human vision without relying on additional information. The only input needed to generate images for human perception is the decoded image intended for image recognition. Using this decoded image as a guide, the image for human visual perception is generated from random noise through the diffusion model. In other words, the information required to decode an image for humans is identical to the amount needed to decode an image for machine vision.

B. Color control of generated images.

The images generated for human visual perception, conditioned on the ICM method, successfully reproduce the positions and locations of objects and backgrounds in the images. However, they are not well-suited for accurate color reproduction. In the SA-ICM decoded images, much of the original texture is lost. To make these images suitable for human use, they need to be enhanced with various colors to restore texture. In other words, the generative AI adds textures in diverse colors based on the SA-ICM decoded images. This process, however, can result in colors that differ from the original image, leading to inconsistencies in color reproduction.

To address this, we apply a Color Controller (CC) module to the generated images. The CC-module adjusts the color components of the images produced by the diffusion model. Specifically, the color components in the generated image are replaced with those from the SA-ICM decoded image, while the luminance component is preserved from the generative AI output. This approach helps the generated image achieve colors that closely match the original.

IV. Experiment

A. Experimental Details and Evaluation Methods.

We conducted experiments to evaluate the effectiveness of the proposed method in extending machine vision to human vision. First, we trained a generative AI model to bridge the gap between these two types of vision. For this training process, we prepare decoded images from the ICM method along with their corresponding original images. To obtain these decoded images, we utilize the official implementation of SA-ICM [34] for image decoding, tailored for image recognition tasks. An example of a decoded image is presented in Figure 3-(b). These machine-oriented decoded images serve as the input for ControlNet, which is trained to reproduce the original images. The loss function for this training is defined in (1). The COCO dataset [42] serve as the benchmark for our

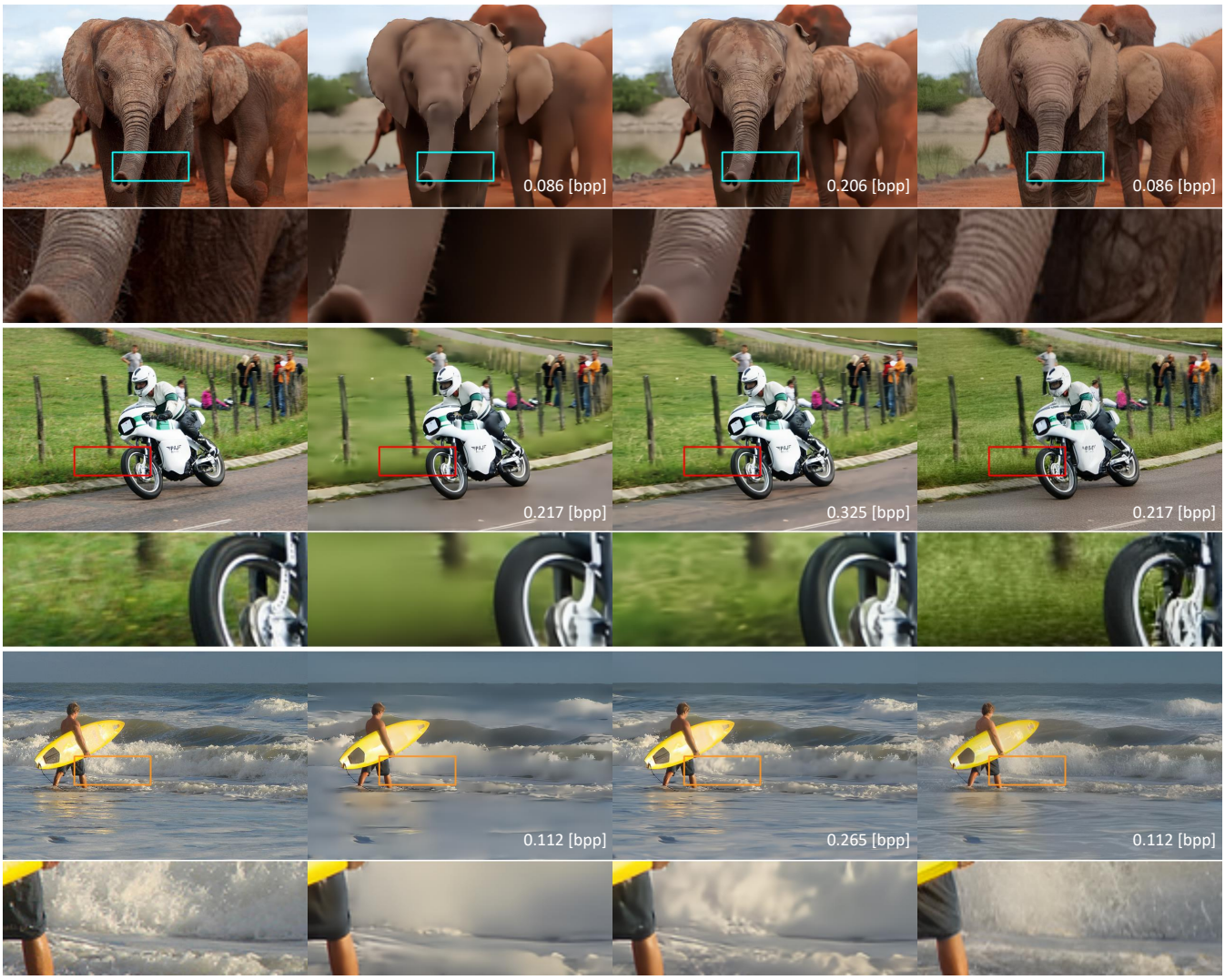

Fig. 3. Examples of original and decoded images: (a) Original image, (b) Decoded image for machine vision using SA-ICM, (c) Decoded image for human vision using ICMH-FF, (d) Decoded image for human perception using the proposed method.

experiments, utilizing 118,287 COCO-train images for training and 5,000 COCO-val images for testing. The text prompt $\mathbf{c_t}$ is left as an empty string during both the training and testing phases.

In this experiment, the proposed method and comparative approaches are primarily evaluated based on bitrate and image quality for human perception. The comparative method, ICMH-FF, is a scalable image coding technique designed for both humans and machines [24]. To ensure a fair comparison, the same ICM method is employed in the implementation of both ICMH-FF and the proposed method. Additionally, common image compression methods for human perception, such as VVC [33] and TCM [44], are included as benchmarks. VVC is a rule-based video compression standard, while TCM is a LIC model that represented the state of the art in 2023. The evaluation of image quality is conducted using metrics such as PSNR, SSIM, LPIPS [20], FID [21], and KID [43].

B. Experimental Results.

Figures 3-(c) and 3-(d) showcase examples of decoded images for human perception using the comparative and proposed methods, respectively. The bitrate required to decode each image is also indicated in the figure. Qualitative evaluation reveals that the proposed method successfully recovers textures lost during the compression process of the ICM method. Unlike conventional approaches, our method effectively depicts fine details, such as elephant's wrinkles and wave splashes, even without relying on additional information.

For the quantitative evaluation, we analyzed the relationship between five image quality indicators and the bitrate, as illustrated in Figures 4-(a) to 4-(e). The blue and green points represent the compression performance of the proposed method with and without the CC-module, respectively. These bitrates indicate only the amount of information needed for

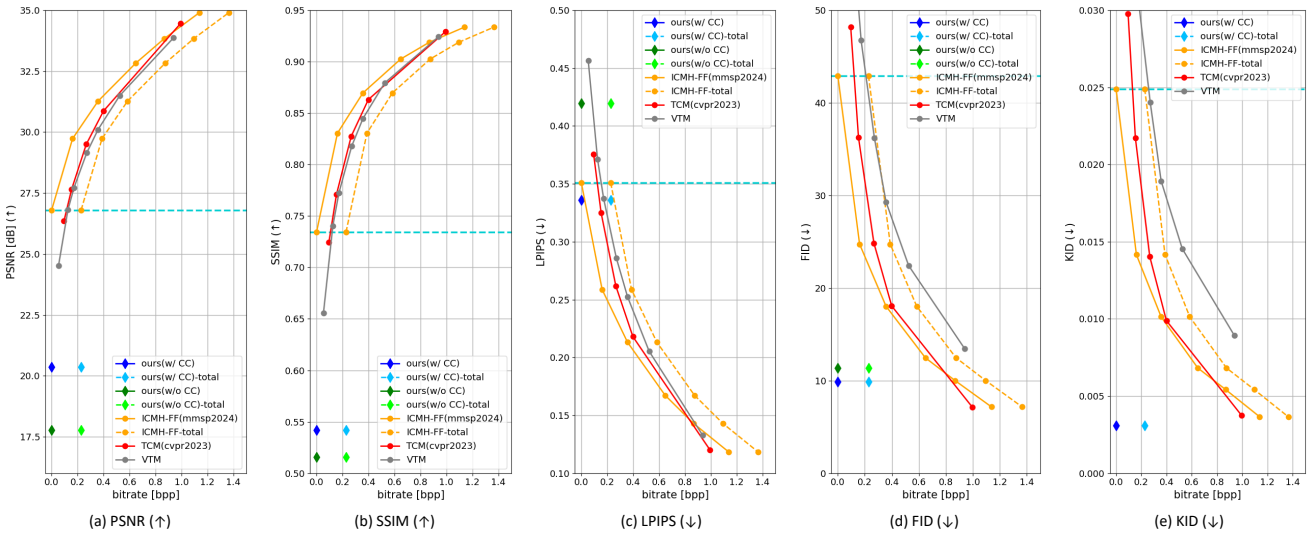

Fig. 4. Image compression performance of the proposed and comparative methods. Five metrics are used to evaluate image quality for human visual perception: (a) PSNR(↑), (b) SSIM(↑), (c) LPIPS(↓), (d) FID(↓), and (e) KID(↓).

transitioning from machine vision to human vision, which remains at 0 [bpp] in all graphs. The light blue and lime green dots represent the total amount of information required by the proposed method to decode images for human perception, including the information necessary for decoding images suitable for machines. The orange curve represents the compression performance of ICMH-FF, where the solid line indicates the amount of additional information, and the dotted line reflects the total of this additional information and the bitrate of the image for machines. The red and gray curves show the compression performance of TCM and VTM, respectively. Figures 4-(a) and 4-(b) reveal that the proposed method is less effective than conventional methods in terms of accurately reproducing the original image's pixel values, as measured by distortion metrics. However, Figures 4-(d) and 4-(e) demonstrate that our method significantly surpasses conventional methods in perceptual metrics like FID and KID, which are closely aligned with human subjective evaluations. These findings indicate that while the proposed vision enhancement method may not excel in fidelity-based decoding, it is highly effective at generating perceptually meaningful images for human interpretation.

Table I compares the bitrates of the methods at equivalent LPIPS and FID scores, while Table II presents the compression performance of each method in image compression for machine vision. For the image recognition model, Yolov5 [45] is used for the object detection and Mask R-CNN [46] for the segmentation. In Table I, the bit rates for both the proposed method and ICMH-FF reflect the total bitrate, which includes the information required for machine decoding. In Table II, since both methods utilize the same ICM method, their image recognition performance is identical. These comparisons highlight that the proposed method, when combined with the ICM method, forms an effective approach to scalable image compression for both human and machine vision. Notably, the proposed method enables the transformation of decoded images for machines into human-perceptible images with no additional information, serving as a valuable tool to bridge the gap between these two types of visual processing.

TABLE I
COMPARISON OF IMAGE COMPRESSION PERFORMANCE FOR HUMANS BETWEEN THE PROPOSED AND CONVENTIONAL METHODS.

method	LPIPS	FID
TCM [44]	0.137 (±0.0[%])	0.710 (±0.0[%])
ICMH-FF [24]	0.235 (+71.53[%])	1.016 (+43.10[%])
Ours	0.227 (+65.69[%])	0.227 (-68.03[%])

TABLE II
COMPARISON OF IMAGE COMPRESSION PERFORMANCE FOR MACHINES BETWEEN THE PROPOSED AND CONVENTIONAL METHODS.

method	detection [45]	segmentation [46]
TCM [44]	0.356 (±0.0[%])	0.342 (±0.0[%])
ICMH-FF [24]	0.227 (-36.23[%])	0.227 (-36.23[%])
Ours	0.227 (-36.23[%])	0.227 (-36.23[%])

V. CONCLUSION

In this paper, we propose a novel method for converting images for machine vision into images for human visual perception. Our experiments demonstrated that by guiding the generation process of generative AI under the constraints of the ICM method, it is possible to create images suitable for humans from machine-targeted images. This approach serves as a bridge, seamlessly connecting machine vision and human perception. Future research should explore applying the proposed vision extension method to other ICM techniques to further validate its effectiveness.

REFERENCES

[1] C. Gao, D. Liu, L. Li and F. Wu, "Towards Task-Generic Image Compression: A Study of Semantics-Oriented Metrics," in IEEE Transactions on Multimedia, vol. 25, pp. 721-735, 2023.

[2] N. Le, H. Zhang, F. Cricri, R. Ghaznavi-Youvalari, and E. Rahtu, "Image Coding For Machines: an End-To-End Learned Approach," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 1590-1594.

[3] N. Le *et al*., "Learned Image Coding for Machines: A Content-Adaptive Approach," 2021 IEEE International Conference on Multimedia and Expo (ICME), 2021, pp. 1-6.

[4] T. Shindo, T. Watanabe, K. Yamada and H. Watanabe, "Accuracy Improvement of Object Detection in VVC Coded Video Using YOLO-v7 Features," 2023 IEEE International Conference on Artificial Intelligence in Engineering and Technology, 2023, pp. 247-251.

[5] G. K. Wallace, "The JPEG still picture compression standard," IEEE Transactions on Consumer Electronics, vol. 38, no. 1, pp. xviii-xxxiv, Feb. 1992.

[6] M. Rabbani and R. Joshi, "An overview of the JPEG 2000 still image compression standard," Signal processing: Image communication, 2002.

[7] H. Choi and I. V. Bajic, "High Efficiency Compression for Object Detection," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 1792-1796.

[8] Z. Huang, C. Jia, S. Wang and S. Ma, "Visual Analysis Motivated Rate-Distortion Model for Image Coding," 2021 IEEE International Conference on Multimedia and Expo (ICME), 2021, pp. 1-6.

[9] B. Li *et al*., "ROI-Based Deep Image Compression with Swin Transformers," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1-5.

[10] T. Shindo, T. Watanabe, K. Yamada and H. Watanabe, "Image Coding for Machines with Object Region Learning," 2024 IEEE 21st Consumer Communications & Networking Conference (CCNC), 2024, pp. 1040-1041.

[11] N. Yan, C. Gao, D. Liu, H. Li, L. Li and F. Wu, "SSSIC: Semantics-to-Signal Scalable Image Coding With Learned Structural Representations," in IEEE Transactions on Image Processing, vol. 30, pp. 8939-8954, 2021.

[12] H. Choi and I. V. Bajić, "Scalable Image Coding for Humans and Machines," in IEEE Transactions on Image Processing, vol. 31, pp. 2739-2754, 2022.

[13] H. Choi and I. V. Bajić, "Scalable Video Coding for Humans and Machines," 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP), 2022, pp. 1-6.

[14] H. Hadizadeh, I. V. Bajić, "Learned Scalable Video Coding For Humans and Machines," arXiv preprint arXiv: 2307.08978, 2023.

[15] P. Zhang, S. Wang, M. Wang, J. Li, X. Wang and S. Kwong, "Rethinking Semantic Image Compression: Scalable Representation With Cross-Modality Transfer," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 8, pp. 4441-4445, 2023.

[16] S. R. Alvar and I. V. Bajić, "Scalable Privacy in Multi-Task Image Compression," 2021 International Conference on Visual Communications and Image Processing (VCIP), 2021, pp. 1-5.

[17] S. Wang, Z. Wang, S. Wang and Y. Ye, "End-to-End Compression Towards Machine Vision: Network Architecture Design and Optimization," in IEEE Open Journal of Circuits and Systems, vol. 2, pp. 675-685, 2021.

[18] F. Codevilla, J. G. Simard, R. Goroshin and C. Pal, "Learned image compression for machine perception," arXiv preprint arXiv:2111.02249, 2021.

[19] S. Wang, S. Wang, W. Yang, X. Zhang, S. Wang, S. Ma and W. Gao, "Towards Analysis-Friendly Face Representation With Scalable Feature and Texture Compression," in IEEE Transactions on Multimedia, vol. 24, pp. 3169-3181, 2022.

[20] R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 586-595.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in Advances in Neural Information Processing Systems, pp. 6626-6637, 2017.

[22] L. Liu, Z. Hu, Z. Chen and D. Xu, "Icmh-net: Neural image compression towards both machine vision and human vision," Proceedings of the 31st ACM International Conference on Multimedia. 2023.

[23] H. Li *et al*., "Image compression for machine and human vision with spatial-frequency adaptation," European Conference on Computer Vision. Springer, 2025.

[24] T. Shindo, T. Watanabe, Y. Tatsumi and H. Watanabe, "Scalable Image Coding for Humans and Machines Using Feature Fusion Network," 2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP), 2024, pp. 1-6.

[25] T. Shindo, Y. Tatsumi, T. Watanabe and H. Watanabe, "Refining Coded Image in Human Vision Layer Using CNN-Based Post-Processing," 2024 IEEE 13th Global Conference on Consumer Electronics (GCCE) 2024, pp. 166-167.

[26] A. Harell, Y. Foroutan and I. V. Bajić, "VVC+M: Plug and Play Scalable Image Coding for Humans and Machines," 2023 IEEE International Conference on Multimedia and Expo Workshops, 2023, pp. 200-205.

[27] N. Le *et al*., "Bridging the Gap Between Image Coding for Machines and Humans," 2022 IEEE International Conference on Image Processing (ICIP), 2022, pp. 3411-3415.

[28] J. Ho, A. Jain and P. Abbeel, "Denoising diffusion probabilistic models," Advances in neural information processing systems 33 (2020), pp. 6840-6851.

[29] J. Ballé, V. Laparra and E. P. Simoncelli, "End-to-end Optimized Image Compression," in Proc. International Conference on Learning Representations (ICLR), 2017, pp. 1-27.

[30] J. Ballé, D. Minnen, S. Singh, S. J. Hwang and N. Johnston, "Variational image compression with a scale hyperprior," in Proc. International Conference on Learning Representations (ICLR), 2018, pp. 1-10.

[31] D. Minnen, J. Ballé and G. Toderici, "Joint autoregressive and hierarchical priors for learned image compression." Advances in neural information processing systems 31 (2018).

[32] Z. Cheng, H. Sun, M. Takeuchi and J. Katto, "Learned Image Compression With Discretized Gaussian Mixture Likelihoods and Attention Modules," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 7936-7945.

[33] Versatile Video Coding, Standard ISO/IEC 23090-3, ISO/IEC JTC 1, Jul. 2020.

[34] T. Shindo, K. Yamada, T. Watanabe and H. Watanabe, "Image Coding For Machines With Edge Information Learning Using Segment Anything," 2024 IEEE International Conference on Image Processing (ICIP), 2024, pp. 3702-3708.

[35] T. Shindo, T. Watanabe, Y. Tatsumi and H. Watanabe, "Delta-ICM: Entropy Modeling with Delta Function for Learned Image Compression," arXiv preprint arXiv: 2410.07669, 2024.

[36] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial networks," Communications of the ACM, 2020, pp. 139-144.

[37] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10674-10685.

[38] L. Zhang, A. Rao and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3813-3824.

[39] L. Jin and H. Watanabe, "Perceptual Image Compression via Stable Diffusion at Low Bitrate," 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval, 2024, pp. 626-629.

[40] E. Lei, Y. B. Uslu, H. Hassani and S. S. Bidokhti, "Text+ sketch: Image compression at ultra low rates," in ICML 2023 Workshop on Neural Compression: From Information Theory to Applications, Jul. 2023.

[41] High Efficiency Video Coding, Standard ISO/IEC 23008-2, ISO/IEC JTC 1, Apr. 2013.

[42] T. Y. Lin *et al*., "Microsoft COCO: Common Objects in Context," Computer Vision - ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. 2014, pp 740-755.

[43] M. Binkowski, D. J. Sutherland, M. Arbel and A. Gretton, "Demystifying MMD GANs," in International Conference on Learning Representations (ICLR), 2018.

[44] J. Liu, H. Sun and J. Katto, "Learned Image Compression with Mixed Transformer-CNN Architectures," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14388-14397.

[45] G. Jocher *et al*., "ultralytics/yolov5: v7.0-yolov5 sota realtime instance segmentation," Zenodo, Nov., 2022.

[46] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," Proceedings of the IEEE international conference on computer vision (ICCV), 2017, pp. 2961-2969.