Dual-Layer Image Codec with Text Guidance for Hybrid Visual Perception

Tianma Shen*, Joseph Quan† and Ying Liu*
*Department of Computer Science and Engineering, Santa Clara University, Santa Clara, CA 95053, USA
†Canyon Crest Academy, 5951 Village Center Loop Road, San Diego, CA 92130, USA
Email: {tshen2, yliu15}@scu.edu, labdabcr@gmail.com

***Abstract*—Image compression for humans and machines (ICHM) requires balancing machine vision performance with human visual quality. Recent methods introduce text-assisted compression to improve perceptual quality, but they often lack explicit optimization for downstream machine vision tasks. To address this, we propose a novel dual-layer codec that jointly considers semantic accuracy and visual fidelity. The base layer is optimized for object detection and encodes semantic-aware features, while the enhancement layer refines visual quality by leveraging both the base-layer output and text guidance during encoding—without requiring text at decoding. Experiments on COCO2017 demonstrate that our method achieves superior rate-accuracy trade-offs for both object detection and image reconstruction compared to prior ICHM methods.**

***Keywords*—conditional coding, image coding for machines, image compression, scalable coding, vision-language model.**

I. Introduction

Traditional image codecs maintain high visual quality for human perception but fail to achieve high recognition accuracy for for machine tasks such as object detection [1] and instance segmentation [2]. Since machine vision tasks only require task-specific semantic information rather than complete image reconstruction, image coding for machines (ICM) [3]–[10] has been developed to compress images with lower bit rates while preserving essential features for accurate analysis. However, many real-world scenarios still require image compression for both machine and human interpretation, such as traffic monitoring [11], human–machine interaction [12], and autonomous underwater vehicles [13]. Therefore, image coding frameworks for humans and machines (ICHM) are developed to address this dual requirement.

Recent ICHM work falls into three categories: one encoder-multiple decoders [14]–[16], adaptive prompts [17]–[19], and scalable coding [20]–[24]. One encoder-multiple decoders methods compress images into unified representations processed by task-specific decoders. Adaptive prompt methods freeze a backbone codec and add prompts for other tasks, saving model parameters. Scalable coding compresses images into layered bitstreams: base layers for basic tasks and enhancement layers for sophisticated tasks.

Recent advances in text-guided image compression show great potential for enhancing reconstructed image perceptual quality [25], [26]. Modern text-guided generative models can synthesize realistic images from text prompts, typically leveraging pre-trained large vision-language models like CLIP [27].

This work is supported in part by the National Science Foundation under Grant ECCS-2138635 and the NVIDIA Academic Hardware Grant.

In this paper, we propose a dual-layer text-guided image codec supporting both human and machine vision tasks. Our framework combines scalable coding with text-guided compression, in which a base layer optimized for object detection provides semantic conditional information to a text-assisted enhancement layer for image reconstruction. Experiments demonstrate that our approach outperforms existing methods in object detection accuracy, as well as objective and perceptual image reconstruction quality. The rest of the paper is organized as follows: Section II reviews related work, Section III presents the proposed method, Section IV provides experimental results and analysis, and Section V concludes the paper.

II. Related Work

A. Image Coding for Humans and Machines (ICHM)

The one encoder-multiple decoders architecture approach, exemplified by VNVC [16], encodes video frames into latent representations that are subsequently decoded by task-specific decoders for human perception, super-resolution, and machine analysis. This approach faces two primary limitations: difficulty in learning unified intermediate features that effectively serve multiple downstream tasks, and the complexity of joint training where combined loss functions make it challenging to balance performance across tasks [14], [15].

In contrast, adaptive prompt methods modify existing architectures by injecting learnable parameters. TransTIC [17] injects learnable tokens into Swin Transformer input sequences, while Prompt-ICM [18] applies similar token-based adaptation to classification models. While these methods achieve parameter efficiency, they cannot share compressed bitstreams across different tasks, resulting in bit rate redundancy when multiple tasks require the same input data.

Scalable coding approaches address multi-task scenarios through layered compression. HMI-IC [20] compresses feature residuals between enhancement and base layer representations using residual coding techniques. Similarly, SICHM [23] shares encoders between compression layers but encounters similar multi-task optimization challenges as the one encoder-multiple decoders approaches.

B. Text-Guided Image Compression

Text-guided image compression methods can be categorized based on where textual information is integrated within the compression pipeline: encoder-side or decoder-side approaches.

Encoder-side integration incorporates text prompts during the encoding phase to guide feature extraction and compression. TACO [25] exemplifies this approach by using text

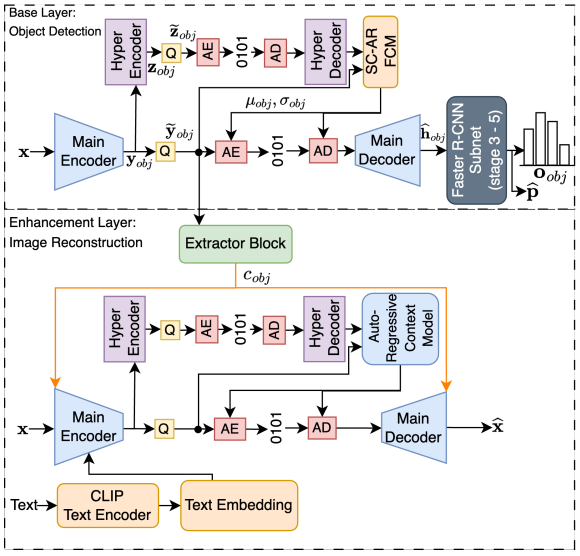

Fig. 1. The overall architecture of the proposed model. AE and AD are the arithmetic encoder and arithmetic decoder. Q represents quantization. SC-AR FCM represents a spatial-channel auto-regressive feature context model.

prompts to inform the encoder's representation learning, enabling the creation of text-aware compressed features. A key advantage of this method is that it enables text-free decoding—once the text-informed features are compressed, no additional textual information needs to be transmitted, eliminating bit rate overhead while preserving image reconstruction quality through the learned text-conditioned representations.

Decoder-side integration [26], conversely, treats text prompts as conditional inputs during image reconstruction. These methods typically employ diffusion models or similar generative approaches to synthesize high-quality images from both the compressed visual features and accompanying text descriptions. While this approach can achieve superior reconstruction quality by leveraging powerful text-to-image generation capabilities, it introduces two significant drawbacks: additional bit rate costs for transmitting textual information alongside compressed features, and increased computational requirements due to the need for large language models to process text embeddings during decoding.

III. The Proposed Method

A. Overall Architecture

Figure 1 illustrates the architecture of our proposed codec, which employs a dual-layer hierarchical structure: a base layer optimized for object detection tasks and the enhancement layer designed for high-quality image reconstruction for human perception. The base layer (top pipeline in Fig. 1) compresses the input image $\mathbf{x}$ using a main encoder that generates a compact latent representation $\mathbf{y}_{obj}$. The corresponding main decoder reconstructs intermediate features $\widehat{\mathbf{h}}_{obj}$, which are subsequently fed into an intermediate layer of the Faster R-CNN [1] architecture to perform object detection. The enhancement layer (bottom pipeline in Fig. 1) focuses on reconstructing high-quality images for human observation. This layer incorporates two key conditioning mechanisms: (1) a conditional signal $\mathbf{x}^P_{obj}$ extracted from the base layer that

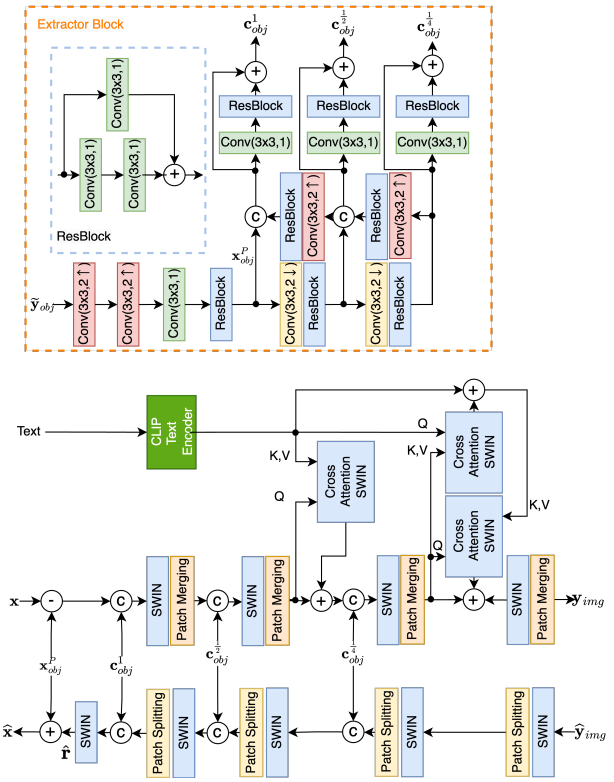

Fig. 2. The top figure is the architecture of the extractor block. ⓒ : channel concatenation, Conv$(3 \times 3, s)$: convolution with a 3×3 kernel and a stride of s, ↑: upsampling, ↓: downsampling, ResBlock: a residual block consisting of two convolutional layers with a skip connection. The bottom figure is the architecture of the text-guided conditional coding as the enhancement layer for image reconstruction.

provides semantic guidance to both the encoder and decoder, and (2) textual prompts in the form of image captions that guide the main encoder during the compression of $\mathbf{x}$. This dual conditioning approach enables the enhancement layer to leverage both visual semantic information from the base layer and textual descriptions to achieve superior reconstruction quality.

B. Base Layer for Object Detection

Our base layer employs a teacher-student network structure [9]. The teacher network is a Faster R-CNN with ResNet50 backbone that performs object detection on uncompressed images. The student network has two components: Part 1 compresses semantic features relevant to object detection, and Part 2 replicates the teacher network's remaining pipeline to generate object class probabilities $\mathbf{o}_{obj}$ and bounding box coordinates $\widehat{\mathbf{p}}$.

The feature codec in Part 1 comprises a main encoder/decoder pair, hyper encoder/decoder pair, and a spatial-channel auto-regressive feature context model (SC-AR FCM) [28]. The base layer is optimized by minimizing the following composite loss function $\mathcal{L}_{\text{obj}}$,

$$\mathcal{L}_{\text{obj}} = \lambda \cdot \mathcal{L}_R + \mathcal{L}_{MSE} + \mathcal{L}_{KL} + \mathcal{L}_{BCE} + \mathcal{L}_b, \qquad (1)$$

where $\mathcal{L}_R$ represents the bit rate of the quantized main and hyper latent representations $\widetilde{\mathbf{y}}_{obj}$ and $\widetilde{\mathbf{z}}_{obj}$. $\mathcal{L}_{MSE}$ is the mean

squared error between the teacher network's semantic feature $\mathbf{h}$ and the student network's decoded feature $\widehat{\mathbf{h}}_{obj}$. $\mathcal{L}_{KL}$ is the Kullback-Leibler divergence (KL) between the object class probabilities predicted by the teacher network and the student network. $\mathcal{L}_{BCE}$ is the cross-entropy loss of the predicted object class probability. $\mathcal{L}_b$ is the MSE between the ground-truth bounding box coordinates $\mathbf{p}_n$ and the predicted bounding box coordinates $\widehat{\mathbf{p}}_n$.

C. Enhancement Layer for Image Reconstruction

The enhancement layer reconstructs high-quality images for human perception by leveraging both conditional information from the base layer and textual guidance during encoding. The extractor block (left panel of Fig. 2) derives conditional information from the base layer by first processing the base layer features $\widetilde{\mathbf{h}}_{obj}$ to obtain an initial image prediction $\mathbf{x}^P_{obj} \in \mathbb{R}^{H\times W\times 3}$. A U-Net architecture then processes $\mathbf{x}^P_{obj}$ to generate multi-scale conditional features $\mathbf{c}^1_{obj}$, $\mathbf{c}^{\frac{1}{2}}_{obj}$, and $\mathbf{c}^{\frac{1}{4}}_{obj}$, where the superscripts denote full, half, and quarter spatial resolutions of the input image $\mathbf{x}$, respectively.

The conditional encoding strategy begins by computing the residual $\mathbf{x} - \mathbf{x}^P_{obj}$ and concatenating it with the full-resolution conditional feature $\mathbf{c}^1_{obj}$ along the channel dimension. This representation is processed through successive Swin Transformer blocks, with corresponding conditional features $\mathbf{c}^{\frac{1}{2}}_{obj}$ and $\mathbf{c}^{\frac{1}{4}}_{obj}$ fused at each scale to provide multi-scale guidance. To incorporate semantic information, CLIP text embeddings are integrated into the encoder backbone through three cross-attention layers, enabling the injection of textual knowledge into intermediate visual features. The cross-attention mechanism operates through an alternating query-key-value assignment pattern: the first cross-attention block uses text latent features as both key and value while visual features serve as the query, enabling visual features to attend to relevant textual semantics. The second cross-attention block reverses this configuration, with text features as the query and visual features as both key and value, allowing textual information to selectively focus on relevant visual content. The third cross-attention block returns to the original assignment. This bidirectional attention mechanism facilitates comprehensive information exchange between textual and visual modalities.

The decoder operates without textual input, eliminating transmission overhead, and progressively upsamples the quantized latent features $\widehat{\mathbf{y}}_{img}$. At each stage, upsampled features are concatenated with corresponding conditional information from $\mathbf{x}^P_{obj}$, facilitating accurate image reconstruction guided by base layer semantic information.

We train the enhancement layer with the following loss function $\mathcal{L}_{\text{img}}$:

$$\mathcal{L}_{\text{img}} = \lambda \cdot \mathcal{L}_R + \mathcal{L}_{MSE} + \mathcal{L}_{LPIPS} + \mathcal{L}_{CLIP}, \tag{2}$$

where $\mathcal{L}_{MSE}$ accounts for the MSE between the input image $\mathbf{x}$ and the decoded image $\widehat{\mathbf{x}}$. The LPIPS loss $\mathcal{L}_{LPIPS}$ measures perceptual similarity by computing feature distances in a pre-trained deep network, typically AlexNet. Specifically, it is computed as:

$$\mathcal{L}_{LPIPS} = \sum_l w_l \cdot \frac{1}{H_l W_l} \sum_{h,w} \left\| \frac{\mathbf{f}^l_{h,w}(\mathbf{x}) - \mathbf{f}^l_{h,w}(\widehat{\mathbf{x}})}{\|\mathbf{f}^l_{h,w}(\mathbf{x})\|_2 + \epsilon} \right\|_2^2, \tag{3}$$

where $\mathbf{f}^l_{h,w}(\cdot)$ represents the feature activations at layer l and spatial location (h, w) of the pre-trained network, w_l are learned linear weights for layer l, H_l and W_l are the spatial dimensions at layer l, and ϵ is a small constant for numerical stability. The CLIP loss $\mathcal{L}_{CLIP}$ measures the semantic similarity between the original and reconstructed images in the CLIP feature space:

$$\mathcal{L}_{CLIP} = \|\mathbf{f}_{\text{CLIP}}(\mathbf{x}) - \mathbf{f}_{\text{CLIP}}(\widehat{\mathbf{x}})\|_2^2, \tag{4}$$

where $\mathbf{f}_{\text{CLIP}}(\cdot)$ represents the feature vector extracted by the pre-trained CLIP vision encoder. This loss ensures that the reconstructed image maintains semantic consistency with the original image in the high-level feature space learned by CLIP.

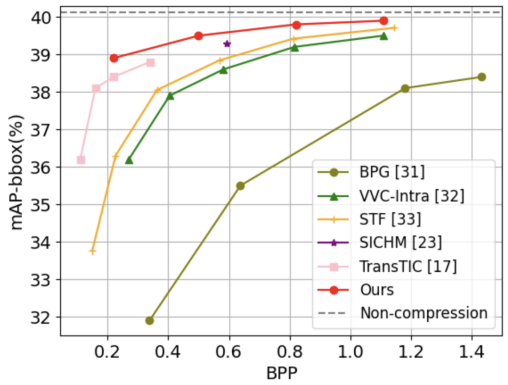

Fig. 3. The rate-accuracy performance of object detection. "Non-compression" represents directly feeding uncompressed images into Faster R-CNN.

IV. Experimental Results and Analysis

A. Datasets and Model Training

In the first stage, we train and test the proposed base layer of Fig. 1 for object detection using the COCO2017 dataset [29]. We adopt Faster R-CNN as the teacher network, which remains frozen during the training process. In the second stage, we train the enhancement layer for image reconstruction on the OpenImages dataset [30] using loss function (2), and test it on the COCO2017 test set. Both the base layer and the CLIP text encoder are frozen.

B. Object Detection Results

For object detection, we use the mean average precision of object bounding boxes (mAP-bbox) to evaluate the performance. First, we measure the average precision (AP) for object class c is then calculated as $\text{AP}_c = \int_0^1 p_c(r)\, dr$ where $p_c(r)$ is the precision-recall curve plotted by varying IoU threshold τ. Finally, the mAP-bbox is computed as the mean of AP_c over all C object classes:

$$\text{mAP-bbox} = \frac{1}{C} \sum_{c=1}^{C} \text{AP}_c. \tag{5}$$

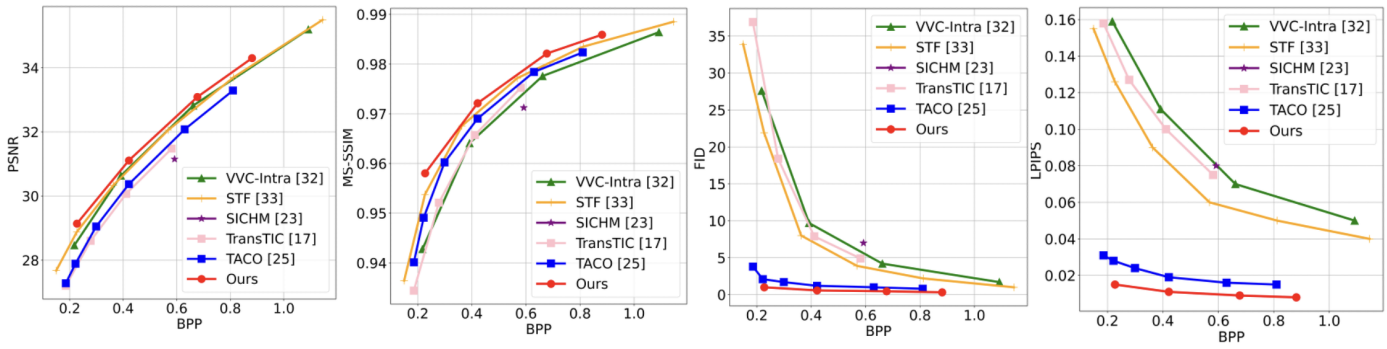

Fig. 4. The rate-distortion PSNR (↑), MS-SSIM (↑), FID (↓) and LPIPS (↓) curves of image coding for image reconstruction on the COCO2017 test dataset.

Fig. 3 presents the base layer's object detection performance compared to conventional codecs (BPG [31], VVC-Intra [32]), learned compression (STF [33], TACO [25]), and ICHM methods (TransTIC [17], SICHM [23]). For conventional and learned baselines, decompressed images are processed by Faster R-CNN for object detection. Our method consistently outperforms all baselines, achieving the highest mAP-bbox values.

C. Image Reconstruction Results

Fig. 4 presents the image reconstruction quality, showing PSNR, MS-SSIM, FID, and LPIPS curves across varying bit rates. Our image reconstruction layer consistently achieves the highest PSNR and MS-SSIM values among all compared methods, demonstrating superior pixel-level accuracy and structural similarity preservation. Furthermore, our model attains the lowest FID and LPIPS scores, indicating exceptional performance in both distributional alignment and perceptual quality metrics. Although STF achieves PSNR and MS-SSIM values comparable to ours, its FID and LPIPS scores are significantly worse. Similarly, while TACO yields FID scores similar to ours, its PSNR, MS-SSIM, and LPIPS metrics are noticeably inferior. These results indicate that our proposed image reconstruction layer achieves a better balance between objective fidelity and perceptual quality.

Figure 5 displays enlarged regions from one representative images, enabling detailed visual comparison of reconstruction quality across methods. Our approach demonstrates clear superiority in preserving fine-grained details and textural information while maintaining sharp edges and minimizing compression artifacts.

V. Conclusion

In this work, we propose a novel text-assisted dual-layer image codec that supports both machine vision and human perception tasks. Our approach consists of a base layer optimized for object detection using a teacher-student framework and an enhancement layer for high-quality image reconstruction. The enhancement layer integrates textual guidance through cross-attention mechanisms while leveraging conditional information from the base layer, enabling text-guided encoding without transmission overhead. Experimental results on the COCO2017 dataset demonstrate state-of-the-art performance in both object detection and image reconstruction.

Fig. 5. The visual result of image reconstruction on the COCO2017 test set.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[3] N. Le, H. Zhang, F. Cricri, R. G. Youvalari, and E. Rahtu, "Image coding for machines: an end-to-end learned approach," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, ON, Canada, June 6-11, 2021, pp. 1590–1594.

[4] L. D. Chamain, F. Racapé, J. Bégaint, A. Pushparaja, and S. Feltman, "End-to-end optimized image compression for machines, a study," in *Proc. Data Compression Conference*, Snowbird, UT, USA, March 23-26, 2021, pp. 163–172.

[5] M. Yang, F. Yang, L. Murn, M. G. Blanch, J. Sock, S. Wan, F. Yang, and L. Herranz, "Task-switchable pre-processor for image compression for multiple machine vision tasks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 6416–6429, 2024.

[6] S. Singh, S. Abu-El-Haija, N. Johnston, J. Ballé, A. Shrivastava, and G. Toderici, "End-to-end learning of compressible features," in *Proc. IEEE International Conference on Image Processing*. Abu Dhabi, United Arab Emirates: IEEE, October 25-28, 2020, pp. 3349–3353.

[7] Y. Matsubara, R. Yang, M. Levorato, and S. Mandt, "Supervised compression for resource-constrained edge computing systems," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*. Waikoloa, HI, USA: IEEE, January 3-8, 2022, pp. 923–933.

[8] Z. Duan and F. Zhu, "Efficient feature compression for edge-cloud systems," in *Picture Coding Symposium (PCS)*. San Jose, CA, USA: IEEE, Dec 2022, pp. 187–191.

[9] T. Shen and Y. Liu, "An effective entropy model for semantic feature compression," in *2024 Picture Coding Symposium (PCS)*. IEEE, 2024, pp. 1–5.

[10] G. Lu, X. Ge, T. Zhong, Q. Hu, and J. Geng, "Preprocessing enhanced image compression for machine vision," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[11] A. A. Verma, B. Chakravarthi, A. Vaghela, H. Wei, and Y. Yang, "etram: Event-based traffic monitoring dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 637–22 646.

[12] J. Yang, Y. Liu, and P. L. Morgan, "Human-machine interaction towards industry 5.0: Human-centric smart manufacturing," *Digital Engineering*, p. 100013, 2024.

[13] K. Hasan, S. Ahmad, A. F. Liaf, M. Karimi, T. Ahmed, M. A. Shawon, and S. Mekhilef, "Oceanic challenges to technological solutions: A review of autonomous underwater vehicle path technologies in biomimicry, control, navigation and sensing," *IEEE Access*, 2024.

[14] X. Fang, Y. Duan, Q. Du, X. Tao, and F. Li, "Sketch assisted face image coding for human and machine vision: a joint training approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 6086–6100, 2023.

[15] J. Cao, X. Yao, H. Zhang, J. Jin, Y. Zhang, and B. W.-K. Ling, "Slimmable multi-task image compression for human and machine vision," *IEEE Access*, vol. 11, pp. 29 946–29 958, 2023.

[16] X. Sheng, L. Li, D. Liu, and H. Li, "Vnvc: A versatile neural video coding framework for efficient human-machine vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[17] Y.-H. Chen, Y.-C. Weng, C.-H. Kao, C. Chien, W.-C. Chiu, and W.-H. Peng, "Transtic: Transferring transformer-based image compression from human perception to machine perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 23 297–23 307.

[18] R. Feng, J. Liu, X. Jin, X. Pan, H. Sun, and Z. Chen, "Prompt-icm: A unified framework towards image coding for machines with task-driven prompts," *arXiv preprint arXiv:2305.02578*, 2023.

[19] T. Shen and Y. Liu, "Parallel task-prompts icm: A versatile feature codec for machine vision," in *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2024, pp. 3695–3701.

[20] Z. Wang, F. Li, J. Xu, and P. C. Cosman, "Human–machine interaction-oriented image coding for resource-constrained visual monitoring in iot," *IEEE Internet of Things Journal*, vol. 9, no. 17, pp. 16 181–16 195, 2022.

[21] Z. Fang, L. Shen, M. Li, Z. Wang, and Y. Jin, "Priors guided extreme underwater image compression for machine vision and human vision," *IEEE Journal of Oceanic Engineering*, vol. 48, no. 3, pp. 888–902, 2023.

[22] Y. Hu, S. Yang, W. Yang, L.-Y. Duan, and J. Liu, "Towards coding for human and machine vision: A scalable image coding approach," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[23] H. Choi and I. V. Bajić, "Scalable image coding for humans and machines," *IEEE Transactions on Image Processing*, vol. 31, pp. 2739–2754, 2022.

[24] S. Li, S. Ma, W. Dai, N. Kan, F. Cheng, C. Li, J. Zou, and H. Xiong, "Task-adapted learnable embedded quantization for scalable human-machine image compression," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[25] H. Lee, M. Kim, J.-H. Kim, S. Kim, D. Oh, and J. Lee, "Neural image compression with text-guided encoding for both pixel-level and perceptual fidelity," in *International Conference on Machine Learning*, 2024.

[26] M. Careil, M. J. Muckley, J. Verbeek, and S. Lathuilière, "Towards image compression with perfect realism at ultra-low bitrates," in *The Twelfth International Conference on Learning Representations*, 2023.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[28] T. Shen and Y. Liu, "Learned image compression with transformers," in *Proc. Big Data V: Learning, Analytics, and Applications*, vol. 12522. Florida, US: SPIE, 2023, pp. 10–20.

[29] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *Proc. Computer Vision European Conference*, vol. 8693, Zurich, Switzerland, September 6-12, 2014, pp. 740–755.

[30] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, and A. Kolesnikov, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1956–1981, Jul. 2020.

[31] F. Bellard, "BPG image format," http://bellard.org/bpg/ (Accessed: 2022-1-18).

[32] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the versatile video coding (vvc) standard and its applications," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, Aug. 2021.

[33] R. Zou, C. Song, and Z. Zhang, "The devil is in the details: Window-based attention for image compression," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, Oct. 2022, pp. 17 471–17 480.