Explicit Residual-Based Scalable Image Coding for Humans and Machines

Yui Tatsumi
Graduate School of FSE,
Waseda University
Tokyo, Japan
yui.t@fuji.waseda.jp

Ziyue Zeng
Graduate School of FSE,
Waseda University
Tokyo, Japan
zengziyue@fuji.waseda.jp

Hiroshi Watanabe
Graduate School of FSE,
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—Scalable image compression methods that serve both machine and human vision (ICMH) have gained increasing attention due to their applicability in various scenarios such as traffic monitoring. While prior studies have made significant strides in this field, many existing models are optimized for specific recognition tasks, which may limit their adaptability. A recent approach, ICMH-FF, addresses this issue by incorporating a task-agnostic codec for machines and a separate codec for additional information required for human-oriented reconstruction. However, its architecture does not incorporate encoder-side interaction and relies on the encoder to implicitly infer residuals, which can pose challenges for interpretability and rate-distortion performance. In this paper, we propose a residual-based ICMH framework that explicitly models the additional information required for human visual perception. Specifically, we present two complementary methods: Feature Residual-based Scalable Coding (FR-ICMH) and Pixel Residual-based Scalable Coding (PR-ICMH). These methods aim to enhance coding efficiency and interpretability without modifying the task-agnostic machine-oriented codec. Moreover, the proposed framework provides flexibility to choose between encoder complexity and compression performance, making it adaptable to diverse application requirements. Experimental results demonstrate the effectiveness of our proposed methods, with PR-ICMH achieving up to 29.57% BD-rate savings over ICMH-FF.**

***Index Terms*—Image coding for machines, Learned image compression, Residual information, Scalable image coding**

I. Introduction

With the rapid advancement of deep learning, recognition models have become integral to a wide range of real-world applications, including traffic monitoring, agricultural management, camera surveillance, and industrial machine vision. In such scenarios, images are primarily analyzed by recognition models with occasional human inspection. To accommodate such dual-purpose use cases, image compression techniques that can simultaneously support both machine and human vision need to be investigated.

To support the needs of human visual perception, numerous Learned Image Compression (LIC) methods have been developed [1] - [6]. More recently, image codecs tailored specifically for machine vision tasks, which is referred to as Image Coding for Machines (ICM), have also been explored [7] - [16]. Unlike LIC, ICM focuses on preserving recognition-related features while discarding human-perceptual details. Therefore, these two types of codecs serve fundamentally different objectives. To achieve a scalable framework for machine and human vision, bridging the gap between these two approaches has become a critical research challenge.

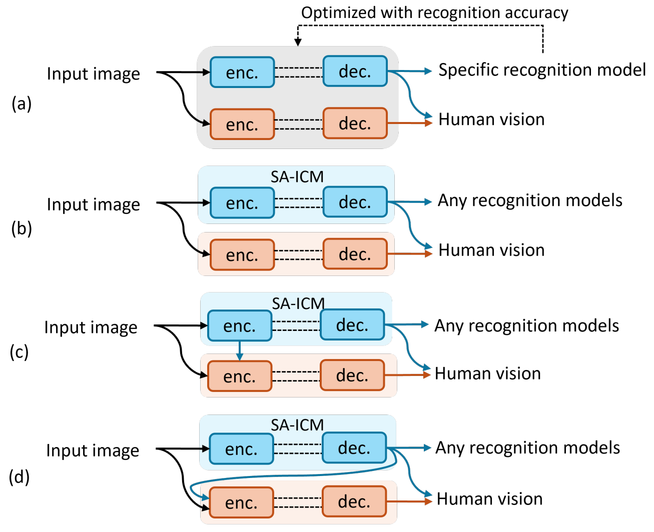


Fig. 1. Overview of scalable image coding pipelines. (a) Conventional task-specific method, (b) ICMH-FF using decoder-side feature fusion, (c) FR-ICMH introducing encoder-side feature residual modeling, (d) PR-ICMH employing pixel-level residual compression.

Many scalable coding methods have been proposed to address this need [17] - [36]. Although they provide efficient structures for scalability, most existing approaches are designed only for specific machine vision tasks and require independent training for each task, as shown in Fig. 1(a). This task-dependency may reduce their flexibility in real-world applications where multiple recognition tasks are involved.

A notable attempt to address this challenge has been proposed in ICMH-FF [17], which introduces a task-agnostic scalable coding architecture. The overview of the framework is illustrated in Fig.1(b). ICMH-FF is based on the idea that the information in the image required for humans includes the information needed for machine vision. Specifically, it employs a machine-oriented codec that is independent of downstream tasks and a separate LIC model to compress additional information necessary for human vision. On the decoder side, their features are merged to reconstruct human-oriented images. However, its design relies heavily on the encoder of the additional LIC to implicitly infer the residual information needed for human viewing without direct access to the original image. This implicit mechanism creates black-box behavior and may result in sub-optimal compression efficiency as well as limited interpretability.

In this paper, we propose a residual-based scalable image coding framework that explicitly models the additional information required for human-oriented reconstruction. In particular, we propose two complementary methods: Feature Residual-based ICMH (FR-ICMH) and Pixel Residual-based ICMH (PR-ICMH). The overview of the pipelines for these proposed methods are shown in Fig.1(c) and (d), respectively. In FR-ICMH, the residual between machine-oriented and human-required features is explicitly computed on the encoder side. This approach not only improves rate-distortion performance, but also enhances interpretability by eliminating the dependence on the encoder's feature selection ability. PR-ICMH extends this idea to the pixel domain by directly compressing the difference image between the reconstructed machines-oriented image and the original image. While FR-ICMH is suitable for scenarios that require a lightweight encoder design, PR-ICMH achieves improved rate-distortion performance by utilizing a more detailed residual representation through additional encoder-side processing. This tradeoff between encoder complexity and reconstruction quality enables the proposed framework to flexibly adapt to a wide range of application constraints without affecting the recognition pipeline.

II. Related Work

A. LIC and ICM

Most state-of-the-art LIC methods build on the frameworks of J. Ballé *et al.* [1] and D. Minnen *et al.* [2]. To improve the decoding speed, channel-wise autoregressive model (Ch-ARM) [3] was proposed, which slices the latent representation along the channel dimension and predicts entropy parameters using previously decoded slices. Building upon Ch-ARM, LIC-TCM [4] further enhances the compression performance.

Existing ICM approaches include ROI-based [7] - [9], Task-Loss-based [10] - [12], and Region-Learning-based [13] - [15]. The first two depend on task-specific prior analysis or loss functions, thus require separate training for each recognition task. In contrast, Region-Learning-based methods aim to achieve task-agnostic compression by preserving recognition-relevant spatial regions. For instance, SA-ICM [13] leverages the Segment Anything model [37] to learn to retain only object boundaries while discarding textures during compression. This also provides privacy protection by removing most facial details. The model is trained using the following loss function:

$$\mathcal{L} = \mathcal{R}(y) + \mathcal{R}(z) + \lambda \cdot mse(x \odot m, \hat{x} \odot m). \tag{1}$$

In (1), y and z denote the output of the encoder and hyperprior-encoder of the LIC model, respectively. $\mathcal{R}(y)$ and $\mathcal{R}(z)$ represent the estimated bitrates of y and z. x is the original image, and $\hat{x}$ is the reconstructed image. mse denotes the mean squared error function. λ is a weighting parameter that controls the trade-off between bitrate and distortion. m is the object mask obtained by the Segment Anything model. The model architecture is based on LIC-TCM with Ch-ARM for entropy modeling.

B. Scalable Image Coding for Humans and Machines

Scalable coding for human and machine vision has emerged as an important direction in recent research. H. Choi *et al.* [18] proposed a dual-stream framework that splits the latent representation extracted from the input image into machine- and human-oriented components. These two streams are transmitted separately, and jointly decoded to reconstruct a human-oriented image. More recently, Adapt-ICMH [19] introduced a plug-and-play adaptation module called the spatial-frequency modulation adapter. In this approach, a human-oriented LIC is employed as a base model, and the adapter is inserted and trained to convert it into a machine-oriented codec. Notably, the adapter can be integrated into any LIC regardless of their model structures. Although these methods are effective in terms of scalability and compression performance, they are trained to optimize recognition accuracy and therefore remain task-specific.

To address this limitation, ICMH-FF has been proposed. In this framework, SA-ICM is utilized for machine vision, while the additional LIC model provides complementary information required for human-oriented reconstruction. These two models are integrated through a feature fusion network implemented on the decoder side. Leveraging the channel-wise structure of Ch-ARM, this network performs channel-wise addition for overlapping slices and directly forwards the remaining slices from the SA-ICM stream. Through this architecture, ICMH-FF achieves scalability, adaptability to various recognition tasks, as well as effective compression performance. However, as the two models are connected only on the decoder side, the current integration scheme leaves room for further reconsideration.

C. Compression Methods Using Residual Connection

Residual connection-based coding has long been one of the core components of efficient image and video compression [20] - [23], [38] - [40]. In traditional standards such as HEVC [38] and VVC [39], the residual between the input frame and its predicted frame is encoded to achieve high compression efficiency. G. Lu *et al.* [40] extended this principal to learning-based video compression framework.

In the context of scalable coding for both human and machine vision, A. Harell [20] proposed VVC+M, where a preview image is reconstructed from machine-oriented features in the base layer, and the residual between the preview and the original image is compressed using the inter mode of VVC. This approach is compatible with a wide range of ICM methods and leverages the efficiency of standard video codec. At the same time, there remains potential to explore end-to-end optimization within LIC-based framework. Furthermore, W. Shi *et al.* [21] introduced a scalable coding method with dual-layer architecture. For vision tasks, intermediate semantic

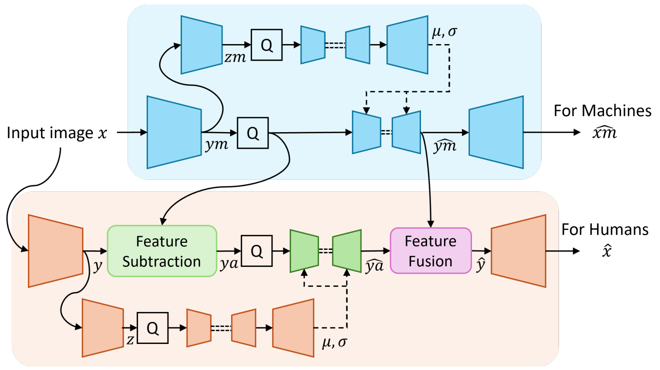


Fig. 2. Overall architecture of the proposed FR-ICMH, which computes and encodes feature-level residuals between the SA-ICM (upper row) and an additional LIC (lower row). The residuals are subsequently fused with the quantized features from SA-ICM to reconstruct human-oriented images.

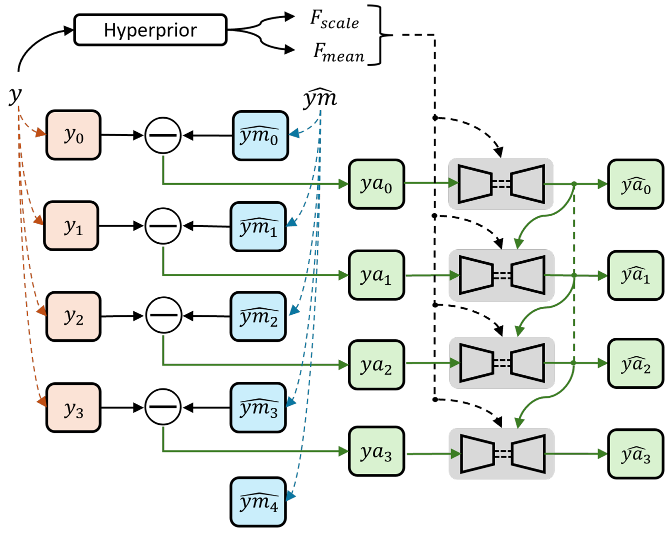


Fig. 3. Structure of the feature subtraction network in FR-ICMH, which computes residuals between features from SA-ICM and LIC for each slice.

features from a pretrained recognition model are encoded. A LRP module is then utilized to generate human-oriented features, and their difference is compressed. This method achieves superior performance in both object detection and image reconstruction, though it depends on prior analysis of the recognition model and must be retrained for each model.

III. Proposed Method

A. Overview

To overcome the structural limitations of ICMH-FF while enhancing both the transparency and efficiency of the encoding process, we propose a residual-based scalable image coding framework that explicitly models the additional information required for human-oriented reconstruction. Our key idea is to avoid the encoder's implicit feature selection by directly providing the residual information as the compression target. By operating on clearly defined residual signals, the proposed approach improves compression efficiency, interpretability, and adaptability to diverse content.

We propose two complementary methods:

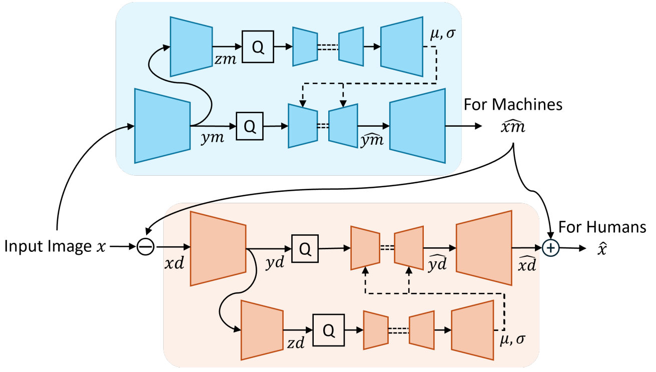


Fig. 4. Overall architecture of the proposed PR-ICMH, which computes pixel-level residuals between the machine-oriented output of SA-ICM (upper row) and the original image. The residuals are encoded by an additional LIC (lower row) and added back to reconstruct human-oriented images.

- **Feature Residual-based Scalable Coding (FR-ICMH)**: Computes and compresses the difference between human-oriented and machine-oriented features by utilizing a feature subtraction network.
- **Pixel Residual-based Scalable Coding (PR-ICMH)**: Directly compresses the pixel-level difference between the original image and the machine-oriented image.

These methods offer distinct trade-offs: PR-ICMH prioritizes compression efficiency at the cost of higher encoder complexity, while FR-ICMH reduces complexity with a slight loss in performance. This enables the framework to adapt to various application scenarios.

B. Feature Residual-based Scalable Coding

The architecture of FR-ICMH is shown in Fig.2. In this method, SA-ICM is utilized as the ICM model, while LIC-TCM is utilized as the additional LIC model for residual information. SA-ICM first encodes the input image into latent features ym, which are then quantized. Simultaneously, the additional LIC encoder processes the input image to produce y. The latent features from SA-ICM and the additional LIC model are divided along the channel dimension into N_m and N_a slices, respectively, denoted as $\{ym_1, ym_2, \ldots, ym_{N_m}\}$ and $\{y_1, y_2, \ldots, y_{N_a}\}$, where $1 \leq N_a \leq N_m$. By setting the number of slices N_a smaller than N_m, the number of channels dedicated to residual information is decreased and computational cost can be reduced.

To support residual computation between features with different numbers of slices, a feature subtraction network is employed to ensure compatibility. As shown in Fig. 3, in this network, the residual features ya are computed slice-by-slice between corresponding slices of y and quantized ym as:

$$ya_k = y_k - \hat{ym}_k, \quad \text{where } k = 1, 2, \ldots, N_a. \tag{2}$$

In (2), ya_k represents a slice of residual features. These residual slices ya_k are then compressed using Ch-ARM-based entropy modeling, maintaining the inherent decoding efficiency and scalability of the slice-wise structure. The decoder reconstructs the full latent representation by adding

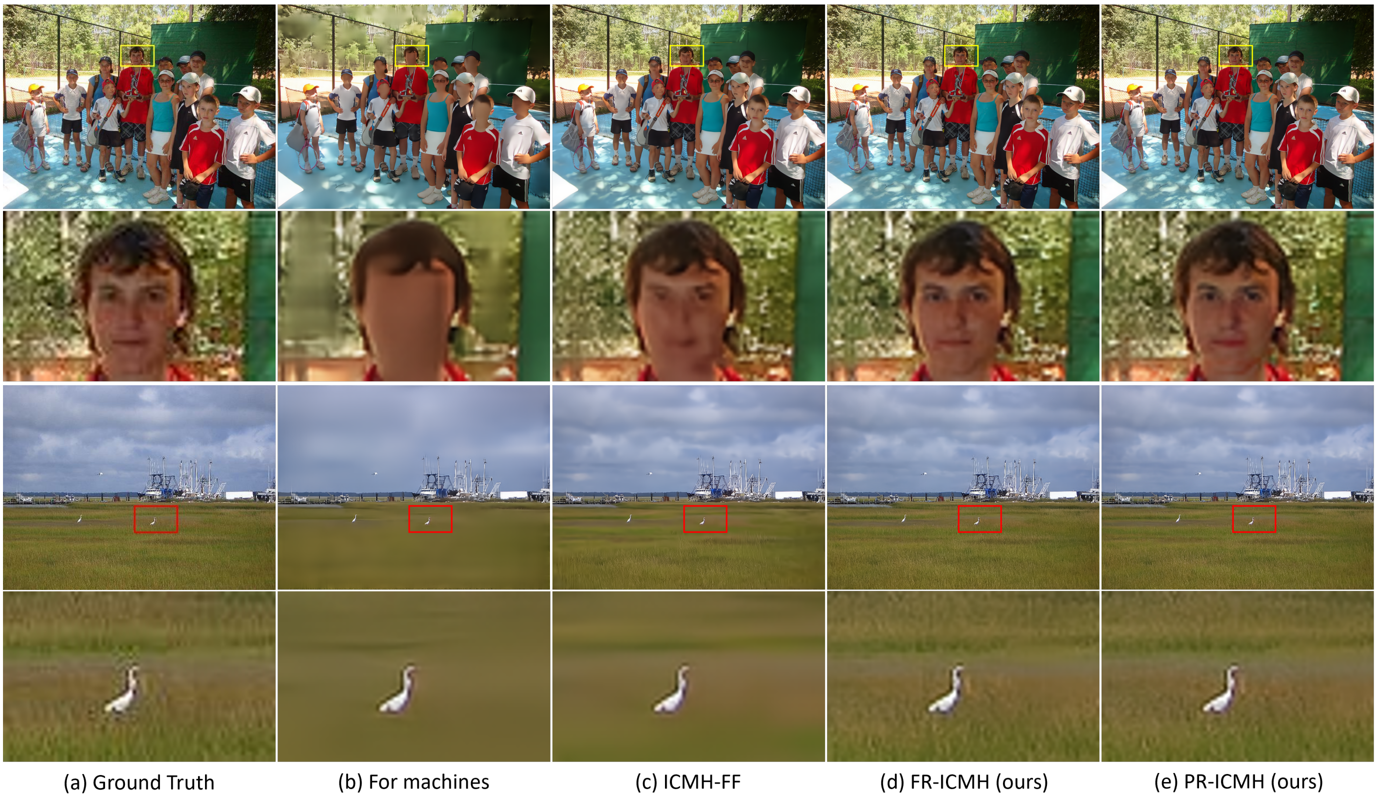


Fig. 5. Examples of reconstructed images for machine and human vision. (a) Ground truth, (b) Machine-oriented reconstruction by SA-ICM ($N_m = 5$), (c) Human-oriented reconstruction by ICMH-FF ($N_a = 4$), (d) Human-oriented reconstruction by the proposed FR-ICMH ($N_a = 4$), and (e) Human-oriented reconstruction by the proposed PR-ICMH ($N_a = 4$). All reconstructions are obtained with $\lambda = 0.05$.

the decoded residual features $\widehat{ya}$ and $\widehat{ym}$ using the feature fusion network proposed in ICMH-FF. The feature fusion function is shown below:

$$\hat{y_k} = \begin{cases} \widehat{ym}_k + \widehat{ya}_k & (1 \leq k \leq N_a) \\ \widehat{ym}_k & (N_a < k \leq N_m) \end{cases} \tag{3}$$

$$\hat{y} = conc(\hat{y_1}, \hat{y_2}, \ldots, \hat{y_{N_m}}). \tag{4}$$

In (3) and (4), $\hat{y}$ denotes the input to the main decoder of additional LIC model. $conc$ stands for the concatenate function. Only the additional LIC model for residual information is trained using the following loss function:

$$\mathcal{L} = \mathcal{R}(ya) + \mathcal{R}(z) + \lambda \cdot mse(x, \hat{x}). \tag{5}$$

In (5), ya and z represent the outputs of the encoder and hyperprior-encoder of the additional LIC model.

C. Pixel Residual-based Scalable Coding

The architecture of PR-ICMH is illustrated in Fig.4. Similar to FR-ICMH, SA-ICM is employed as the ICM model, and LIC-TCM serves as the additional LIC model for encoding the residual information. In this method, the pixel-level difference image xd is computed by directly subtracting the machine-oriented reconstructed image $\widehat{xm}$ from the original image x. This difference image is then compressed using an additional LIC model, which employs Ch-ARM for entropy modeling. Same as the architecture in FR-ICMH, the latent features are divided into N_a slices, denoted as $\{yd_1, yd_2, \ldots, yd_{N_a}\}$. By setting N_a to a smaller value, the number of channels used to represent xd is reduced, thereby lowering the computational cost. At the decoder side, the final human-oriented reconstruction $\hat{x}$ is obtained by adding the machine-oriented image $\widehat{xm}$ and the decoded difference image $\widehat{xd}$. During the training, only the additional LIC model for difference image is trained with the following loss function:

$$\mathcal{L} = \mathcal{R}(yd) + \mathcal{R}(zd) + \lambda \cdot mse(xd, \widehat{xd}). \tag{6}$$

In (6), yd and zd denotes the outputs of encoder and hyperprior-encoder of LIC for difference image, respectively.

IV. Experiment

A. Performance of Image Compression for Humans

We evaluate the compression performance for humans of our proposed methods, FR-ICMH and PR-ICMH. For ICM, we utilize SA-ICM model pre-trained with $\lambda = 0.05$ according to the loss function defined in (1). For the additional LIC model for residual information, LIC-TCM is utilized. In both methods, only the additional LIC is trained. Since the parameters of SA-ICM are fixed, its compression performance for machines is the same as that of SA-ICM. We train both FR-ICMH and PR-ICMH using their respective loss

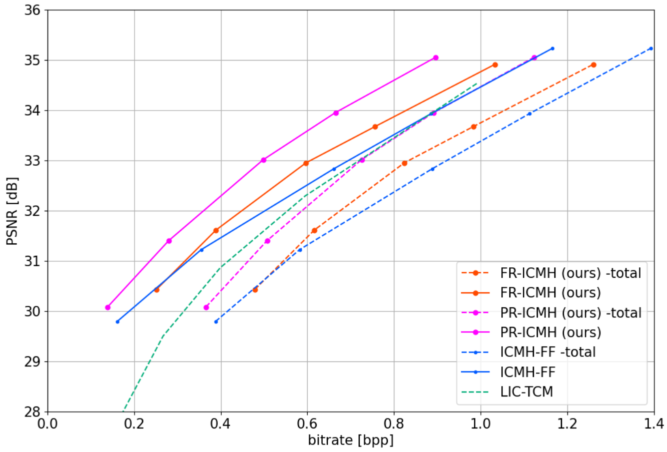
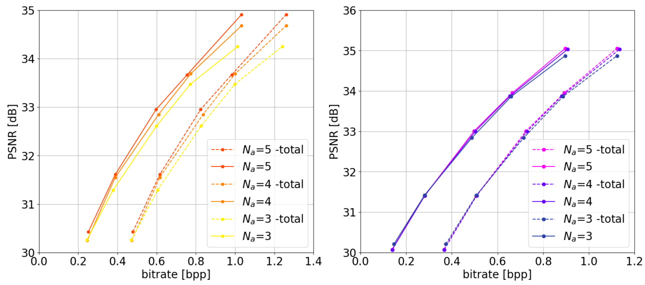


Fig. 6. Rate-distortion curves for human-oriented image reconstruction with N_a = 5.



(a) FR-ICMH (b) PR-ICMH

Fig. 7. Rate-distortion performance of (a) FR-ICMH and (b) PR-ICMH with different numbers of residual slices ($N_a = 3, 4, 5$).

functions, (5) and (6). During the training, five different λ values $\{0.005, 0.01, 0.02, 0.03, 0.05\}$ are utilized. COCO-train dataset is utilized for training and COCO-val is utilized for evaluation [41]. The number of feature slices for SA-ICM, N_m, is fixed to 5, while the additional LIC is trained with $N_a = \{3, 4, 5\}$ to evaluate the effect of number of slices of residual features. Fig. 5 shows that the proposed FR-ICMH and PR-ICMH provide clearer reconstructions of critical visual details, such as facial components and textures in grass, which are less distinct in ICMH-FF.

The rate-distortion curves for human-oriented images with $N_a = 5$ are presented in Fig. 6. The solid lines indicate the bitrate of additional information only, while dashed lines represent the total bitrate (machine-oriented + additional). We compare our proposed methods with ICMH-FF and LIC-TCM. FR-ICMH outperforms ICMH-FF especially at high bitrate regions, while PR-ICMH consistently outperforms across all bitrate ranges. Notably, PR-ICMH achieves performance that is even close to LIC-TCM at high bitrate levels, despite its scalable structure. The performance degradation of FR-ICMH at low bitrates is primarily due to subtracting the high-quality machine-oriented features of SA-ICM ($\lambda = 0.05$) from the human-oriented features. This subtraction yields a very small residual, and results in limited information recovery and lower PSNR.

TABLE I
BD-RATE COMPARISON WITH ICMH-FF FOR DIFFERENT VALUES OF THE RESIDUAL FEATURE SLICES N_a

Method	BD-rate(%)		
	$N_a = 3$	$N_a = 4$	$N_a = 5$
ICMH-FF	0.00	0.00	0.00
FR-ICMH (ours)	-2.43	-4.23	-7.93
PR-ICMH (ours)	-29.57	-26.06	-26.78

TABLE II
COMPARISON BETWEEN THE SIZE OF RESIDUAL INFORMATION COMPRESSION MODEL AND THE NUMBER OF SLICES N_a

N_a	3	4	5
Number of channels for residual information	192	256	350
Number of parameters (M)	58.7	67.0	76.6

The BD-rate of our proposed methods compared to ICMH-FF for different number of slices, N_a, is shown in Table I. For each value of N_a, the corresponding ICMH-FF model is utilized as the baseline. As shown in the table, both proposed methods outperform ICMH-FF across all configurations. In particular, PR-ICMH demonstrates consistently large gains regardless of the number of slices, achieving the best performance with a BD-rate reduction of 29.57% when $N_a = 3$.

B. Effect of Reduction in Number of Slices

We further evaluate our proposed methods by investigating the effect of parameter reduction in the additional LIC model for residual information. Increasing the number of slices N_a leads to a larger number of intermediate features, which in turn increases the number of model parameters. Fig. 7 illustrates how the number of residual feature slices N_a affects the rate-distortion performance of FR-ICMH and PR-ICMH. While both methods show improved performance with higher N_a, the gain is marginal, especially for PR-ICMH. Table II shows the corresponding increase in model parameters. These results indicate that a smaller model can be achieved by reducing the number of slices without significantly sacrificing compression performance.

V. CONCLUSION

In this paper, we propose a residual-based scalable image coding framework for both human visual perception and machine vision. We introduce two complementary ICMH methods, Feature Residual-based (FR-ICMH) and Pixel Residual-based (PR-ICMH). Both proposed methods explicitly model the residual information between machine-oriented and human-required representations, while offering flexible trade-offs between computational cost and compression performance across a wide range of practical applications. Experimental results show that our proposed methods significantly outperform the prior scalable method, ICMH-FF, with PR-ICMH achieving up to 29.57% BD-rate reduction. Future work includes extending our residual-based framework to video coding and supporting variable bitrate control for greater deployment adaptability.

ACKNOWLEDGMENT

The results of this research were obtained from the commissioned research (JPJ012368C05101) by National Institute of Information and Communications Technology (NICT), Japan.

REFERENCES

[1] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, “Variational Image Compression with a Scale Hyperprior,” International Conference on Learning Representations (ICLR), 2018, pp. 1-10.

[2] D. Minnen, J. Ballé, and G. Toderici, “Joint Autoregressive and Hierarchical Priors for Learned Image Compression,” 32nd Conference on Neural Information Processing Systems (NeurIPS), 2018, pp.10794-10803.

[3] D. Minnen and S. Singh, “Channel-Wise Autoregressive Entropy Models for Learned Image Compression,” 2020 IEEE International Conference on Image Processing (ICIP), 2020, pp. 3339-3343.

[4] 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.

[5] J. Balle, V. Laparra, and E. P. Simoncelli, “End-to-end optimized image compression,” arXiv preprint arXiv:1611.01704, 2016.

[6] 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.

[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] J. I. Ahonen, N. Le, H. Zhang, F. Cricri and E. Rahtu, “Region of Interest Enabled Learned Image Coding for Machines,” 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP), 2023, pp. 1-6.

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

[11] N. Le, H. Zhang, F. Cricri, R. Ghaznavi-Youvalari, H. R. Tavakoli and E. Rahtu, “Learned Image Coding for Machines: A Content-Adaptive Approach,” 2021 IEEE International Conference on Multimedia and Expo (ICME), 2021, pp. 1-6.

[12] X. Shen, H. Ou and W. Yang, “Image Coding For Machine Via Analytics-Driven Appearance Redundancy Reduction,” 2024 IEEE International Conference on Image Processing (ICIP), 2024, pp. 1883-1889.

[13] 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.

[14] 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.

[15] T. Shindo, T. Watanabe, Y. Tatsumi and H. Watanabe, ”Delta-ICM: Entropy Modeling with Delta Function for Learned Image Compression,” 2025 IEEE International Conference on Consumer Electronics (ICCE), 2025, pp. 1-6.

[16] R. Feng *et al.*, “Image Coding for Machines with Omnipotent Feature Learning,” Computer Vision - ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol. 13697, 2022, pp 510-528.

[17] 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.

[18] 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.

[19] H. Li *et al.*, “Image Compression for Machine and Human Vision with Spatial-Frequency Adaptation,” Computer Vision – ECCV 2024. ECCV 2024. Lecture Notes in Computer Science, vol. 15109, 2024, pp. 382-399.

[20] 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 (ICMEW), 2023, pp. 200-205.

[21] W. Shi, W. Yin, F. Tao and Y. Wen, “Semantic Prior-Guided Scalable Image Coding,” 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1-5.

[22] A. de Andrade, A. Harell, Y. Foroutan and I. V. Bajić, “Conditional and Residual Methods in Scalable Coding for Humans and Machines,” 2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2023, pp. 194-199.

[23] S. Wang *et al.*, “Towards Analysis-Friendly Face Representation With Scalable Feature and Texture Compression,” in IEEE Transactions on Multimedia, vol. 24, pp. 3169-3181, 2022.

[24] S. Wang, P. AN, C. Yang, K. Huang and X. Huang, “STSIC: Swin-Transformer-based Scalable Image Coding for Human and Machine,” Journal of Visual Communication and Image Representation, vol. 98, 2024.

[25] J. Wei *et al.*, “Layered and scalable image coding with semantic features for human and machine,” Engineering Applications of Artificial Intelligence, vol. 155, 2025.

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

[27] 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 (ICMEW), 2023, pp. 200-205.

[28] Y. Wu, P. An, C. Yang and X. Huang, “Scalable image coding with enhancement features for human and machine,” Multimedia Systems, vol. 30, no. 77, 2024.

[29] 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.

[30] R. Feng, Y. Gao, X. Jin, R. Feng and Z. Chen, “Semantically Structured Image Compression via Irregular Group-Based Decoupling,” 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17191-17201.

[31] Y. -H. Chen *et al.*, “TransTIC: Transferring Transformer-based Image Compression from Human Perception to Machine Perception,” 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23240-23250.

[32] H. Hadizadeh and I. V. Bajic, “Learned scalable video coding for humans and machines,” arXiv preprint arXiv:2307.08978, 2023.

[33] L. Liu, Z. Hu, Z. Chen and D. Xu, “ICMH-Net: Neural Image Compression Towards both Machine Vision and Human Vision,” in Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 8047-8056.

[34] 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.

[35] S. Sun, T. He and Z. Chen, “Semantic Structured Image Coding Framework for Multiple Intelligent Applications,” in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 9, pp. 3631-3642, 2021.

[36] N. Yan *et al.*, “SSSIC: Semantics-to-Signal Scalable Image Coding With Learned Structural Representations,” in IEEE Transactions on Image Processing, vol. 30, pp. 8939-8954, 2021.

[37] A. Kirillov *et al.*, “Segment Anything,” 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3992-4003.

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

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

[40] G. Lu *et al.*, “DVC: An End-To-End Deep Video Compression Framework,” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10998-11007.

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