2024 IEEE International Conference on Image Processing (ICIP) | 979-8-3503-4939-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIP51287.2024.10648120

PARALLEL TASK-PROMPTS ICM: A VERSATILE FEATURE CODEC FOR MACHINE VISION

Tianma Shen, Ying Liu†

Department of Computer Science and Engineering
Santa Clara University, Santa Clara, CA 95053 USA
Emails: {tshen2, yliu15}@scu.edu

ABSTRACT

Image Coding for Machines (ICM) is developed to compress images with a focus on machine vision tasks rather than human perception. For ICM, It is very important to develop a universal codec adaptable to different machine tasks. In this paper, we propose novel parallel task-prompts that can be easily adapted to various machine vision tasks without necessitating new networks or scratch training. Besides, Our parallel prompts are compatible with mainstream backbones such as transformers and convolutional neural networks, making them widely applicable across different model architectures. In order to fine-tune our task-prompts, we leverage a machine task network as the teacher net, guiding our student ICM network to efficiently compress feature maps for downstream machine tasks. Through extensive experimentation on object detection and segmentation, we demonstrate that our proposed method surpasses traditional image compression techniques and state-of-the-art learning-based feature compression techniques in terms of rate-accuracy performance.

Index Terms— entropy model, image coding for machines, object detection, segmentation, task-prompts, transformer

1. INTRODUCTION

The rapid expansion of machine vision applications [1, 2, 3] has led to an increased demand for efficient image compression techniques. Traditional image codecs [4, 5], while effective for human perception, fall short when it comes to compressing images tailored for machine vision tasks. To overcome the drawbacks of traditional image codecs, Image Coding for Machines (ICM) ensures that crucial information is retained for machine vision tasks. The ICM technologies can be categorized into two main types: compress-then-analyze methods [6, 7, 8, 9] and analyze-then-compress methods [10, 11, 12, 13, 14]. Compress-then-analyze methods involve compressing the image first and then analyzing the compressed representation. Compress-then-analyze methods are often easier to integrate into existing systems. However, this method faces a problem in that machines don't need all feature maps from source images for vision tasks to work, leading to a bigger bit rate. On the contrary, analyze-then-compress methods focus on compressing necessary feature maps as semantic features, which have essential and relevant information for semantic vision tasks.

However, even state-of-the-art (SOTA) ICM methods [12, 13, 14, 15] face challenges that low adaptability for different tasks and quality of datasets. The adaptability refers to the capability of an ICM to dynamically adjust its compression strategies to suit the specific requirements of different machine vision tasks. Traditional, static ICM methods [6] may struggle to address this variability, limiting their effectiveness across a spectrum of tasks. An adaptive ICM [14, 15], on the other hand, can intelligently adjust its compression mechanisms based on specific features of source data, ensuring optimal performance for a wide range of machine vision applications. Furthermore, another challenge is the quality of datasets which plays a pivotal role in the success of any machine learning model. The quality encompasses various aspects such as diversity, resolutions, and number of data. A high-quality dataset [16] ensures that the ICM is exposed to a broad spectrum of scenarios, allowing it to learn robust features and patterns that generalize well to unseen data.

In light of these challenges, our paper proposes a new parallel task-prompt and a new training stage as a solution. Our contributions are summarized as follows:

- We propose novel parallel task-prompts that can be easily adapted to various machine vision tasks, eliminating the need for creating new networks or undergoing training from scratch.
- We designed a two-stage training stage to increase the quality of dataset by proposing an extractor head in the first stage, and leveraging a machine task network as the teacher net to guide our student net in the second stage.
- Experimental studies on two popular computer vi-

†Corresponding author.

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

sion tasks, object detection and segmentation, demonstrate that our proposed method offers state-of-the-art (SOTA) rate-accuracy performance.

2. RELATED WORK

2.1. Image Coding for Machines

The development of state-of-the-art Image Coding for Machines (ICM) has witnessed significant strides in recent years. Early work in this domain, exemplified by end-to-end learnable networks [6], directly concatenates the image code with machine tasks. However, this approach encountered challenges related to balancing competing loss functions. To address the intricacies of loss function optimization, subsequent works emerged [7, 15], offering solutions to the aforementioned problem. These contributions often involved the utilization of pre-trained codecs, followed by fine-tuning the entire network. This strategy aimed to enhance the synergy between image coding and machine tasks while mitigating the risk of unnecessary information causing an increase in bit-rate.

Feature compression works [10, 11] then tackled the issue of unnecessary information by extracting compressible latent representations of feature maps. This allowed the codec to learn semantic features directly relevant to the supervised machine task, improving the efficiency of image compression for machine vision. Building upon these foundations, recent advancements [12, 13] by certain researchers have further refined feature compression methodologies. These improvements focus on compressing feature representations more efficiently through the use of a teacher model and a student model. Additionally, a learnable prior [17, 18] for entropy coding is introduced, enhancing the overall effectiveness of feature compression in the context of machine vision tasks.

2.2. Task-adaptive Methods

Task-adaptive Image Coding for Machines (ICM) represents a pivotal advancement in the field, emphasizing the need for image coding strategies that dynamically adapt to diverse machine vision tasks.

TransTIC [15] proposes instance-specific prompts and task-specific prompts to dynamically adjust the image coding process based on the specific requirements of each task. However, the compress-then-analyze methodology employed by TransTIC introduces unnecessary information, potentially leading to an increase in bit-rate. Prompt-ICM [14] emerges as a response to the limitations of TransTIC, seeking to enhance task-adaptive ICM. The methodology proposed by Prompt-ICM introduces two key elements: the Information Selector (IS) and task-specific prompts. These components play a crucial role in enabling the codec to learn semantic features directly relevant to the supervised machine task.

However, Prompt-ICM exhibits certain limitations, particularly in its efficiency without the guidance of a teacher net. The absence of a teacher model to guide the learning process can impact the overall efficiency of the codec. This limitation prompts further exploration into methodologies that integrate teacher-student models to improve the efficiency of task-adaptive ICM.

2.3. The Quality of Dataset

The success of ICM is significantly influenced by the quality of the datasets they are trained on. Quality, in the context of datasets, encompasses various aspects such as diversity, number, and resolution. The limitations of ICM are often rooted in the data they rely upon. State-of-the-art ICMC [12, 15, 14] typically undergo training and fine-tuning processes using datasets specific to machine vision tasks. Notably, a significant portion of these tasks relies on the Common Objects in COCO 2017 dataset [19]. However, the quality of COCO 2017 has certain constraints. With a training dataset comprising only about 100,000 images, 80 classes, and a resolution of 640 x 480, it falls short in comparison to alternatives like the 7th version of Openimage [16]. The latter boasts a substantial 9 million training images, covering 9,000 classes for object detection and segmentation. This stark contrast highlights the importance of considering dataset quality and diversity in enhancing the performance and capabilities of ICM.

3. PROPOSED METHOD

3.1. The Overall Architecture

In Fig. 1, we illustrate an architecture of our proposed parallel task-prompts feature compression framework tailored for machine vision tasks. Inspired by the entropic student approach [12], this architecture comprises a teacher network (upper pipeline in Fig. 1) and a student network (lower pipeline in Fig. 1). The teacher network, serving as a machine task network, guides the training of the student network. The task head at the end of the teacher network is specialized for addressing specific machine vision tasks, where we've opted for RetinaNet [2] for object detection and Deeplab V3 [3] for semantic segmentation, utilizing them as our teacher networks.

Contrastingly, the student network is composed of two fundamental components, Part 1 and Part 2, as depicted in Fig. 1. Part 1, the feature codec, plays a crucial role in learning and compressing semantic features essential for downstream machine tasks. The decoded semantic feature $\widehat{\mathbf{h}}$ aims to align with the Stage 2 output feature $\mathbf{h}$ of the teacher network. To diminish redundancy in semantic features, we employ a spatial-channel auto-regressive feature context model (SC-AR FCM) [18]. This model captures both spatial and channel correlations within the semantic latent representation, estimating μ and σ for the latent representation's distribution

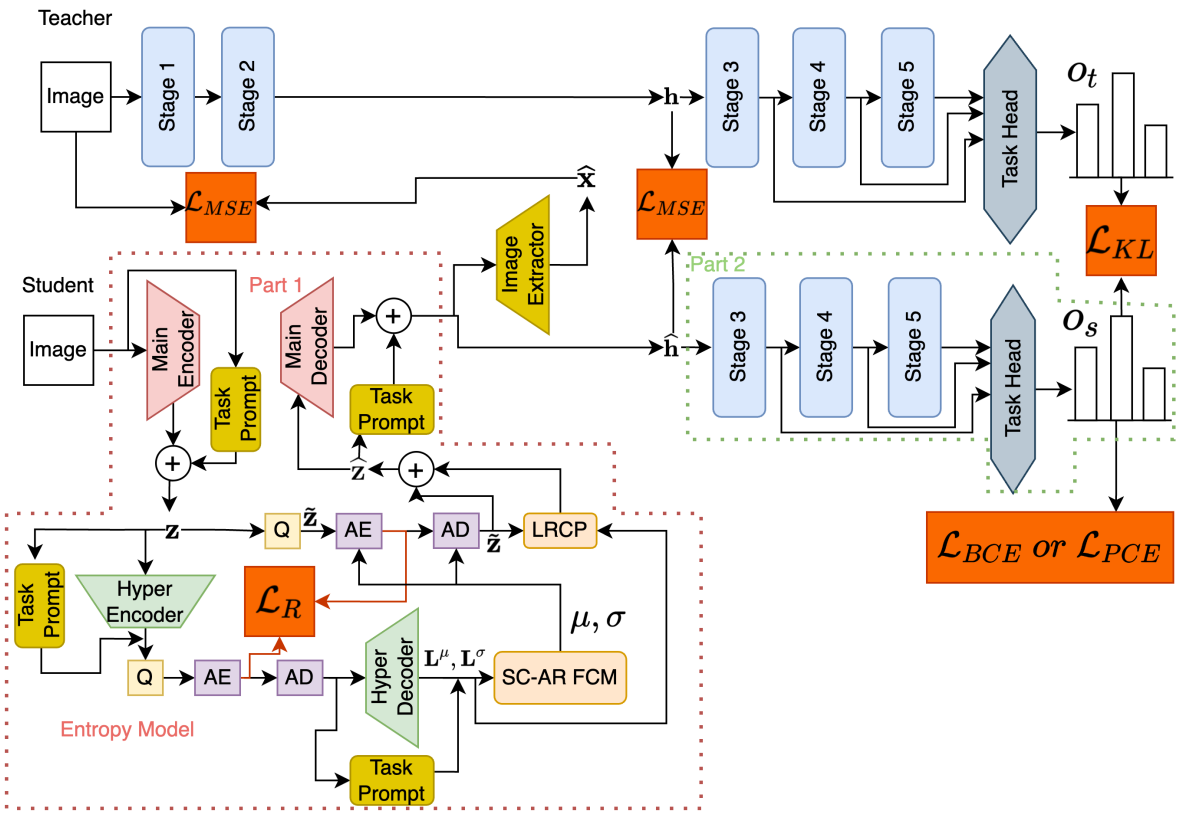

Fig. 1. The overall architecture of the proposed model. AE and AD are the arithmetic encoder and arithmetic decoder. Q represents quantization. SC-AR FCM represents a spatial-channel auto-regressive feature context model. LRCP represents a latent residual cross-attention prediction.

in entropy coding. Subsequently, Part 2 of the student network replicates the remaining pipeline of the teacher network and ultimately produces machine task results, such as object detection bounding boxes and semantic segmentation maps.

3.2. Universal Codec

In order to apply ICM for different vision tasks, it is very important to design an adaptive feature compression system to extract universal features in part 1 of Fig. 1. The ideal universal features should include all information about all vision tasks, so that universal features are suitable for different tasks without training the whole network. To achieve this, we train a universal codec following by image compression in the form of a main encoder/decoder, coupled with a hyper-structure.

Another advantage of using image compression to train our universal codec is that it addresses the issue of low-quality data. We can choose all the image's datasets as training datasets instead of vision tasks' datasets, so that we can improve the quality of datatsets, which have a large number, high diversity, and high-resolution source images. By training on a high-quality dataset such as OpenImages [16], the model can learn more robust and informative universal features.

In our universal codec, we propose an image extractor block to achieve decoded images. This image extractor block just contains 4 convolution layers to upsample the size of decoded semantic feature $\widehat{\mathbf{h}}$. The output of image extractor block is decoded image, which helps to compute the mean squared error (MSE) between $\widehat{\mathbf{x}}$ and $\mathbf{x}$ with the bit rate of source images as the loss function to train our proposed universal codec.

3.3. Task Prompt

Inspired by prompt technique, we proposed parallel task-prompts feature compression in part 1 of Fig. 1. The main idea is to design an extra architecture on an existing codec, which is adjustable for different tasks by only training this extra architecture instead of the whole network. Besides, to extend the applicability of our prompt across different vision task backbones, including CNNs or transformers, we propose a parallel task-prompt architecture. This entails aligning the task-prompts and task's associated backbones in parallel, followed by the summation of the outputs from these two components to form the final output. This design facili-

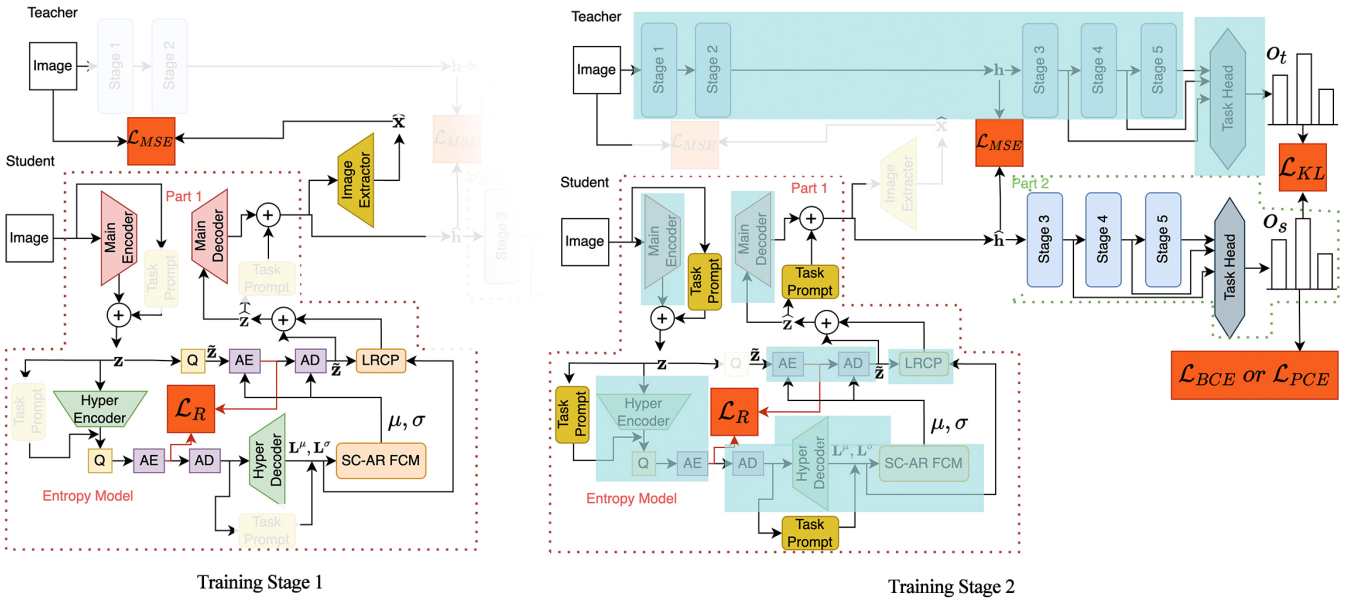

Fig. 2. The training stage of our proposed model. The white half-transparency blocks are not involved in each training stage. The blue transparency blocks are the frozen blocks, which will not update the parameters during the training stage.

tates the integration of parallel task-prompts with mainstream backbones. Notably, the initialization of parameters in our parallel task-prompts is set to zero, ensuring that the prompt outputs are initially zero, thereby having no impact on the primary backbone until the prompts undergo training.

The implementation of our parallel task-prompts involves the utilization of four convolution layers, strategically designed to maintain small prompt sizes relative to the overall model. This ensures that the impact on the model's size is minimal, contributing to efficiency integration with diverse vision tasks. To further mitigate the redundancy of the semantic feature $\widehat{\mathbf{h}}$, we extend the application of the task prompt beyond the main encoder and decoder. Additionally, we incorporate the task prompt into the entropy model for the hyper-structure. This expanded application ensures that the Spatial-Channel Auto-Regressive Feature Context Model (SC-AR FCM) can achieve higher accuracy in estimating μ and σ for the distribution of the latent representation $\widehat{\mathbf{z}}$. By involving the task prompt in multiple components, we enhance the overall effectiveness of our feature compression framework.

3.4. Training Stage

In our model, we have two training stages shown in Fig. 2. In the first stage, the goal is to train a universal codec for universal features. The white half-transparency blocks are not involved in the first stage. The loss function is the bit rate and MSE between the source image and the decoded image. Once the model has been trained, we proceed to the second stage, which is specific to different vision tasks. Remarkably, the first stage only needs to be trained one time for various vision tasks, resulting in significant time savings.

In the second stage, we removed the MSE loss between images and decoded images along with Image Extractor Block shown in Fig. 2. The blue transparency blocks are the frozen blocks, which will not update the parameters during the second training stage. For the semantic segmentation task, our model is trained in an end-to-end manner using the following loss function $\mathcal{L}_{\text{seg}}$ and we choose (2) as the loss function for the object detection task,

$$\mathcal{L}_{seg} = \lambda \cdot \mathcal{L}_R + \mathcal{L}_{MSE} + \mathcal{L}_{KL-seg} + \mathcal{L}_{PCE}, \quad (1)$$

$$\mathcal{L}_{\text{obj}} = \lambda \cdot \mathcal{L}_R + \mathcal{L}_{MSE} + \mathcal{L}_{KL-obj} + \mathcal{L}_{BCE} + \frac{\mathcal{L}_b}{N_b}, \quad (2)$$

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N_h}\|\mathbf{h} - \widehat{\mathbf{h}}\|_2^2, \quad (3)$$

$$\mathcal{L}_{KL-seg} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{M} o_t^{c,n} \log\left(\frac{o_s^{c,n}}{o_t^{c,n}}\right), \quad (4)$$

$$\mathcal{L}_{KL-obj} = -\frac{1}{N_b}\sum_{n=1}^{N_b}\sum_{c=1}^{M} o_t^{c,n} \log\left(\frac{o_s^{c,n}}{o_t^{c,n}}\right), \quad (5)$$

$$\mathcal{L}_{PCE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{M} y_c \log\left(o_s^{c,n}\right), \quad (6)$$

$$\mathcal{L}_{BCE} = -\frac{1}{N_b}\sum_{n=1}^{N_b}\sum_{c=1}^{M} y_c \log\left(o_s^{c,n}\right), \quad (7)$$

$$\mathcal{L}_b = \sum_{n=1}^{N_b}\|\mathbf{p}_n - \widehat{\mathbf{p}}_n\|_2^2. \quad (8)$$

Here, $\mathcal{L}_R$ represents the bit rate of $\widetilde{\mathbf{z}}$ which is measured by bits per pixel (BPP), $\mathcal{L}_{MSE}$ accounts for the mean squared error between the semantic feature of the teacher network $\mathbf{h}$ and that of the student network $\widehat{\mathbf{h}}$, N_h is the number of elements in $\mathbf{h}$, N is the number of pixels in an image, $\mathcal{L}_{KL}$ is the KL divergence between $\mathbf{o}_s$ and $\mathbf{o}_t$, $\mathbf{o}_s$ and $\mathbf{o}_t$ represent the class distribution prediction of student and teacher network. c is the class index in object detection or segmentation, $\mathcal{L}_{PCE}$ corresponds to the pixel-level cross-entropy loss, M is number of classes, and $y_c \in \{0, 1\}$ is the label of class c, using 1-of-M coding, $\mathcal{L}_{BCE}$ corresponds to the bounding box's cross-entropy loss, $\mathcal{L}_b$ corresponds to the regression loss of bounding boxes' coordinates, N_b is the number of bounding boxes, whose IoU is higher than 0.5, $\mathbf{p}$ represents the ground-truth bounding box coordinates, and $\widehat{\mathbf{p}}$ represents the bounding box coordinates predicted by the student network.

4. EXPERIMENTS AND ANALYSIS

4.1. Datasets and Evaluation Metrics

In our experiments, we employed the OpenImages dataset [16] to train our proposed model in the first training stage. For the second stage, we choose the COCO 2017 dataset [19] for both object detection and semantic segmentation tasks.

Throughout the training process, we continuously evaluated our model's performance on the validation dataset. Monitoring the model's loss on this dataset after each epoch, we terminated training if the loss on the validation set ceased to decrease.

For the evaluation of our model's performance, mean average precision (mAP) served as the metric for the object detection task. Calculated based on bounding box (BBox) outputs, mAP considers various Intersection-over-Unions (IoU) thresholds, ranging from 0.5 to 0.95. In the case of the semantic segmentation task, we assessed performance using the mean Intersection-over-Union (mIoU) value. This metric, averaged across 21 distinct segment classes, provided a comprehensive understanding of our model's segmentation accuracy, offering insights into its ability to differentiate and segment objects within the images.

4.2. Object Detection and Segmentation Results

To evaluate the efficacy of our proposed model, we conducted experiments comparing it against the state-of-the-art method, Entropic Student [12], as well as conventional codecs VTM-19.2 [5] and BPG [4]. Notably, VTM-19.2 and BPG fall under the category of compress-then-analyze methods, involving the decompression of source images for machine input. In the second training stage, our model only requires 10 epochs for a specific machine task, a significantly lower number compared to previous work.

Focusing on object detection, our results, depicted in Fig. 3, highlight our model's performance through rate-distortion curves and mean average precision (mAP). For Figure 3, the non-compression baseline is 36.11. Significantly, our model demonstrates superior rate-distortion curves, indicating its ability to efficiently balance compression rates while preserving object detection accuracy.

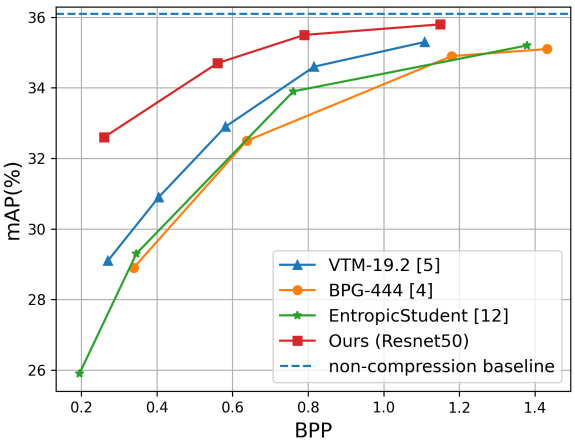

Fig. 3. The results of object detection task.

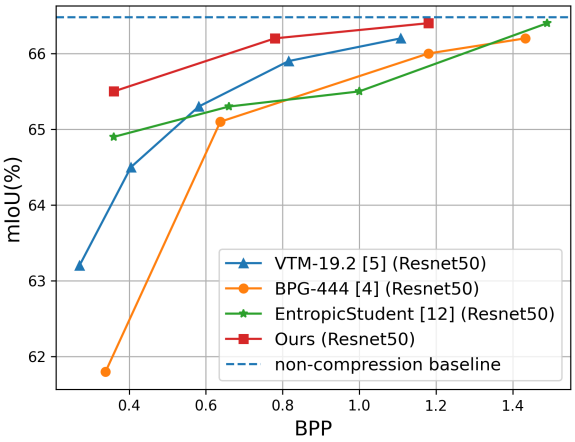

Fig. 4. The results of semantic segmentation task.

For semantic segmentation, as depicted in Fig. 4, we performed a comparative analysis to assess our model's performance against other methods. For Figure 4, the non-compression baseline is 66.44. Our model consistently outperforms alternative approaches, demonstrating a noteworthy superiority over the entropic student method with a substantial 0.27% increase in mIoU. This improvement highlights the robust capability of our approach to excel in semantic segmentation tasks.

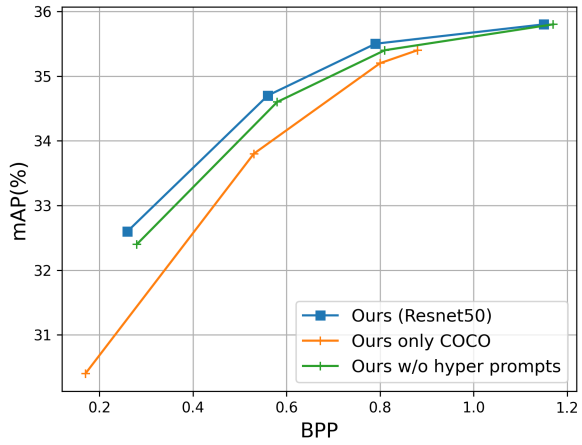

Fig. 5. The ablation study results of object detection task.

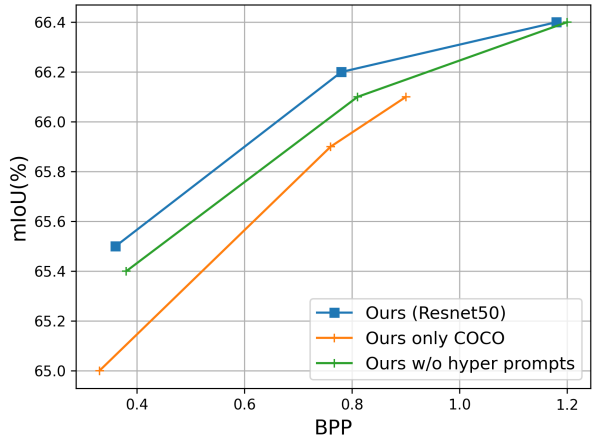

Fig. 6. The ablation study results of semantic segmentation task.

4.3. Ablation Study

In order to ensure our contribution, we did two ablation studies on object detection and segmentation. The orange line in Fig. 5 and Fig. 6, named 'Ours only COCO' is our new proposed model, which is only trained on the COCO2017 dataset in the first training stage. This ablation study aims to make sure the contribution of the high-quality dataset. The gap between the orange and blue lines shows high-quality dataset can improve the performance of codec to compress the semantic feature maps.

The second ablation study aims to determine the necessity of incorporating prompts in the entropy model. Accordingly, the green line depicted in Fig. 5 and Fig. 6, labeled as 'Ours w/o hyper prompts' represents our newly proposed model w/o prompts in the hyper encoder and hyper decoder. Analysis of the results reveals that despite requiring 8 epochs for prompt training, the performance lags behind that of the original model. This underscores the importance of prompts in the entropy model, as their inclusion is evidently crucial for enhancing overall performance and achieving optimal results.

5. CONCLUSION

In conclusion, our work introduces Image Coding for Machines (ICM) as a specialized approach for compressing images tailored to machine vision tasks. The emphasis on developing a universal codec adaptable to diverse machine tasks sets our approach apart. The innovation lies in our proposed parallel task-prompts, offering seamless adaptability to various machine vision tasks without the need for new networks or scratch training. Our extensive experiments in object detection and segmentation showcase that our method outperforms conventional image compression techniques and the state-of-the-art feature compression model, achieving superior rate-accuracy performance. In the future, we will extend our framework to accommodate hybrid machine-human vision tasks.

6. REFERENCES

[1] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Miami, Florida, USA, June 20-25, 2009, pp. 248–255.

[2] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE International Conference on Computer Vision*, Venice, Italy, October 22-29, 2017, pp. 2999–3007.

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

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

[5] 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.

[6] 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.

[7] 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.

[8] C. Gao, D. Liu, L. Li, and F. Wu, "Towards task-generic image compression: A study of semantics-oriented metrics," *IEEE Transactions on Multimedia*, vol. 25, pp. 721–735, 2023.

[9] N. Le, H. Zhang, F. Cricri, R. G. Youvalari, H. R. Tavakoli, and E. Rahtu, "Learned image coding for machines: A content-adaptive approach," in *Proc. IEEE International Conference on Multimedia and Expo*, Shenzhen, China, July 5-9, 2021, pp. 1–6.

[10] Z. Chen, K. Fan, S. Wang, L. Duan, W. Lin, and A. C. Kot, "Toward intelligent sensing: Intermediate deep feature compression," *IEEE Transactions on Image Processing*, vol. 29, pp. 2230–2243, 2020.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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 visualization to machine perception," in *Proc.the IEEE/CVF International Conference on Computer Vision*, Paris,France, October 2-6, 2023, pp. 23 297–23 307.

[16] 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.

[17] D. Minnen, J. Ballé, and G. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Proc. Annual Conference on Neural Information Processing Systems*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Montréal, Canada, December 3-8, 2018, pp. 10 794–10 803.

[18] 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.

[19] 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.