GSFA-ICM: Image Compression for Machines via Gated Spatial-Frequency Adaptation

Zixuan Gao[1], Meiqin Liu[1], Yuxiao Sun[1], Chao Yao[2], Jian Jin[3], Weisi Lin[3], and Yao Zhao[*,1]
[1]Beijing Jiaotong University, Beijing, China
[2]University of Science and Technology Beijing, Beijing, China
[3]Nanyang Technological University, Singapore
{24125189, mqliu, 22110081}@bjtu.edu.cn, yaochao@ustb.edu.cn, jian.jin@ntu.edu.sg, wslin@ntu.edu.sg, yzhao@bjtu.edu.cn

***Abstract*—Image Compression for Machines (ICM) aims to compress images to support machine vision tasks instead of human vision systems. However, its practical application is limited by the high training costs of existing paradigms. To address this challenge, inspired by Parameter-Efficient Fine-Tuning (PEFT), we propose a lightweight Gated Spatial-Frequency Adapter (GSFA) to repurpose pre-trained base codecs by updating only the GSFA, thereby significantly reducing training costs. Specifically, to minimize task-irrelevant spatial redundancy, we introduce a Spatial-Context Adapter (SCA) that employs a hierarchical learning mechanism to exploit context and align semantics between the codec and machine vision tasks. For efficient task-relevant frequency representation, the Frequency-Spectrum Adapter (FSA) is introduced to synergize FFT-based and Wavelet-based adaptations, effectively disentangling global-local frequency components. Finally, the Gated Fusion Adapter (GFA) dynamically balances the influence of spatial and frequency adaptation. Moreover, the GSFA is plug-and-play and compatible with multiple existing neural image compression models. Experiments demonstrate that our framework saves more than 20% of parameters compared to existing SOTA methods while maintaining competitive rate-task performance in image classification, object detection, and instance segmentation.**

***Index Terms*—Parameter-Efficient Fine-Tuning, Neural Image Compression, Image Compression for Machines**

I. Introduction

In real-world applications, image compression plays an important role in reducing storage and transmission costs, particularly when massive amounts of data are offloaded from edge devices to the cloud for machine vision tasks. However, existing image codecs, especially Neural Image Compression (NIC) models [1]–[3], are typically optimized for human vision systems and often retain semantically irrelevant redundancy that can interfere with the analysis of machine vision tasks and lead to significant bitrate waste. Consequently, Image Compression for Machines (ICM) has emerged as a promising paradigm aimed at maximizing coding efficiency specifically for machine vision systems.

In recent years, some existing ICM methods [4]–[6] focus on task-specific codecs, which compress and reconstruct either images or features tailored for individual machine vision tasks.

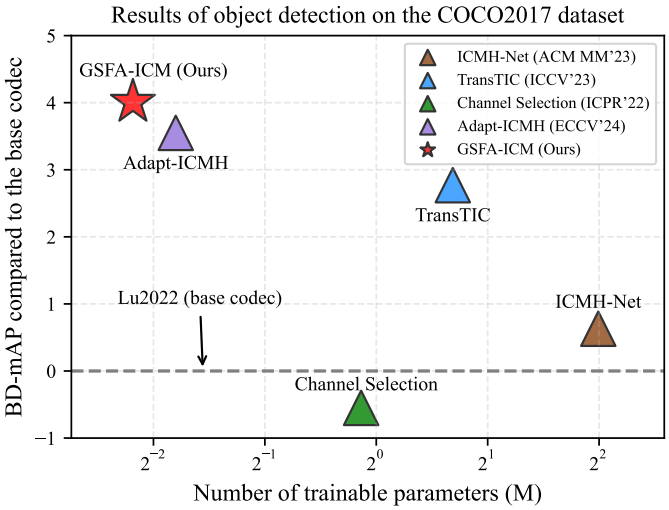


Fig. 1. Rate-task performance and parameters comparison of object detection on the COCO2017 dataset. Experimental results demonstrate that our GSFA-ICM outperforms existing methods with the fewest trainable parameters.

However, this paradigm requires separate optimization for each task, which results in poor scalability and increasing training costs as the number of downstream tasks grows. Other works [7], [8] explore multi-task unified codecs, aiming to develop a general codec applicable to multiple machine vision tasks. Yet, to accommodate diverse task demands, these frameworks incur high parameter overhead due to the complex architectures required to generalize across tasks. Both paradigms suffer from a common bottleneck: they require training the entire framework, which inevitably leads to high training costs.

Inspired by PEFT [9], we propose the Gated Spatial-Frequency Adapter (GSFA) to repurpose pre-trained base codecs for various machine vision tasks. By freezing the base codec and training only the GSFA, our approach significantly reduces training costs. However, simply applying PEFT is insufficient. To enhance rate-task performance, the adapter must be structurally optimized to eliminate task-irrelevant redundancy often present in both spatial and frequency domains.

To realize this, the proposed GSFA comprises three key components. First, to circumvent the parameter overhead and redundancy inherent in stacking deep convolutional networks, we introduce the Spatial-Context Adapter (SCA). By employing a hierarchical learning mechanism, the SCA captures

*Corresponding author
This work is supported by the National Natural Science Foundation of China (62120106009, 62372036, U24B20179, U22A2022, 62332017).

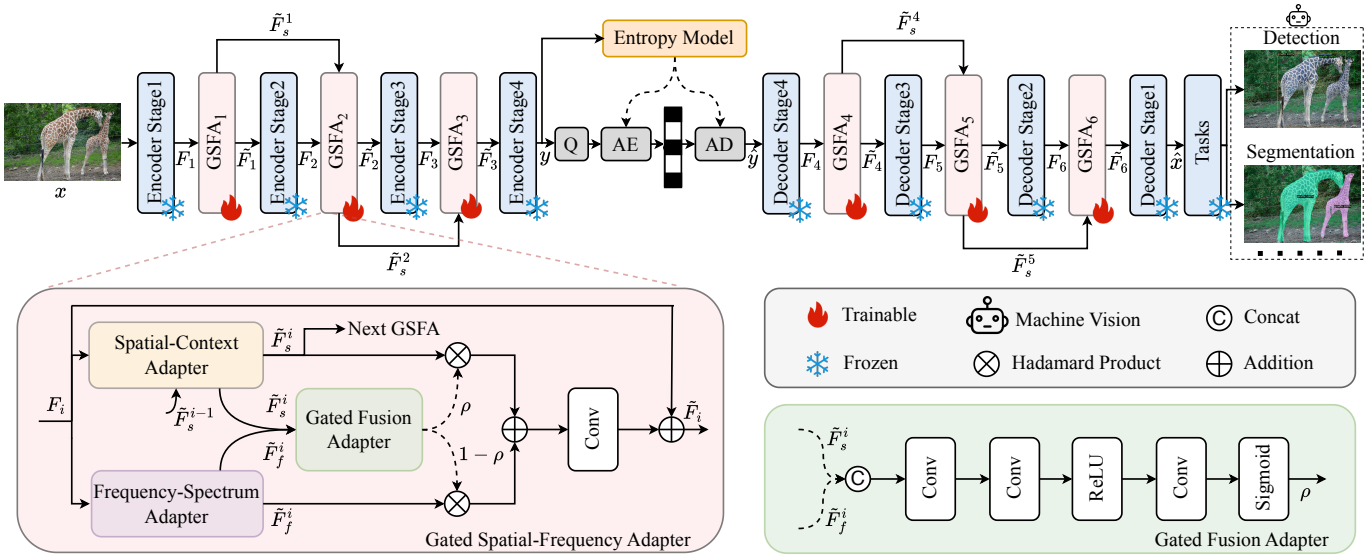


Fig. 2. Overview of our GSFA-ICM framework. The proposed GSFA is plug-and-play and compatible with multiple base codecs. Multiple Gated Spatial-Frequency Adapters (GSFAs) are plugged into both the encoder and decoder of the base codec for PEFT. During the training process, the base codec is frozen and only GSFAs are trainable.

rich contextual dependencies and aligns semantics between the base codec and machine vision tasks, thereby minimizing task-irrelevant redundancy in the spatial domain. Second, recognizing that single transforms (e.g., FFT or Wavelet alone) are often inadequate for efficiently disentangling global-local frequency components, we introduce the Frequency-Spectrum Adapter (FSA). This module synergizes FFT-based and Wavelet-based adaptations, effectively extracting global information and fine-grained local details to eliminate task-irrelevant redundancy in the frequency domain. Finally, unlike naive summation, which ignores the varying importance of different domains, our Gated Fusion Adapter (GFA) dynamically balances the contributions from the SCA and FSA, achieving superior rate-task performance. Notably, GSFA is plug-and-play, making it compatible with a wide range of NIC models for various machine vision tasks. As shown in Fig. 1, our GSFA-ICM achieves the best rate-task performance with the fewest trainable parameters for object detection on the COCO2017 dataset. Furthermore, we validate the effectiveness of our framework on multiple downstream tasks, including image classification, object detection, and instance segmentation.

Our main contributions are summarized as follows:

- We propose a Gated Spatial-Frequency Adapter (GSFA) that adaptively modulates intermediate representations in both spatial and frequency domains via a Gated Fusion Adapter (GFA), effectively balancing the trade-off between task performance and bitrate savings.
- We introduce a Spatial-Context Adapter (SCA) with a hierarchical learning mechanism that exploits context and aligns semantics between the codec and machine vision tasks to minimize task-irrelevant spatial redundancy.
- We introduce a Frequency-Spectrum Adapter (FSA) that synergizes FFT-based and Wavelet-based adaptations to disentangle global-local frequency components, effectively eliminating task-irrelevant frequency redundancy.

II. Proposed Method

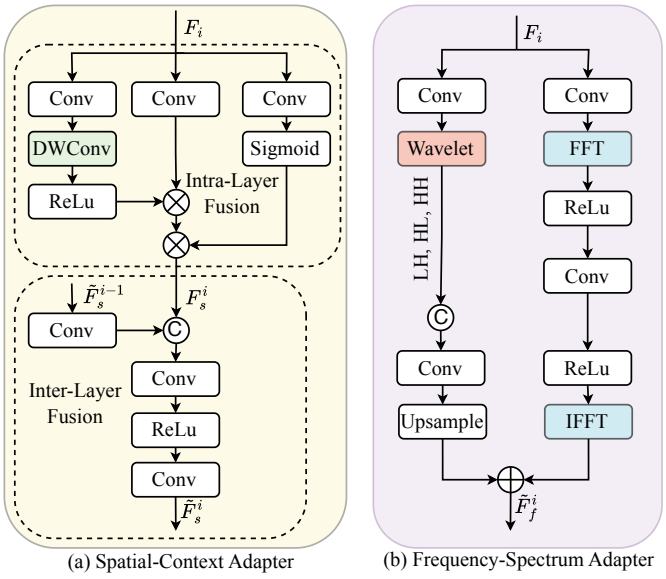
A. Overview

Our GSFA-ICM framework aims to adapt existing NIC models for machine vision tasks, as illustrated in Fig. 2. Specifically, the base codec is composed of multiple sequential stages in both the encoder and decoder. Given an input image x, it is processed through multiple sequential stages, where an intermediate feature F_i is produced by each stage i for $i \in \{1, \ldots, 6\}$. To enable task-specific adaptation, a Gated Spatial-Frequency Adapter (GSFA) is inserted after the corresponding codec stage i, denoted as $GSFA_i$. Consequently, the original feature F_i is fed into the $GSFA_i$ to obtain the adapted feature $\tilde{F}_i$. Through this adaptive feature extraction, a latent representation y is obtained, which is quantized and processed via entropy coding to reconstruct $\hat{y}$. Then, $\hat{y}$ is processed through similar adaptive stages, where intermediate features are refined by inserted GSFAs, to reconstruct the image $\hat{x}$. Finally, $\hat{x}$ is fed into a downstream task model to perform the corresponding machine vision task, such as object detection. In addition, during training, the base codecs are frozen, and only the GSFAs are optimized via PEFT.

As the core component enabling this adaptation, the GSFA comprises three key components: a Spatial-Context Adapter (SCA), a Frequency-Spectrum Adapter (FSA), and a Gated Fusion Adapter (GFA). By integrating these components, the GSFA dynamically modulates features in both domains. The output of GSFA, namely the adapted feature $\tilde{F}_i$ can be formulated as:

$$\begin{aligned} \tilde{F}_i &= GSFA_i(F_i) \\ &= F_i + GFA_i(SCA_i(F_i), FSA_i(F_i)), \end{aligned} \tag{1}$$

where $SCA_i(\cdot)$, $FSA_i(\cdot)$, and $GFA_i(\cdot)$ represent the Spatial-Context Adapter, Frequency-Spectrum Adapter, and Gated Fusion Adapter at the i-th stage, respectively, which will be detailed in the following sections.



(a) Spatial-Context Adapter (b) Frequency-Spectrum Adapter

Fig. 3. Architecture of the Spatial-Context Adapter (SCA) and the Frequency-Spectrum Adapter (FSA). (a) SCA includes the intra-layer and inter-layer fusion. (b) FSA includes the FFT-based and the Wavelet-based adaptation.

B. Spatial-Context Adapter

To minimize task-irrelevant redundancy in the spatial domain, we introduce a Spatial-Context Adapter (SCA) with a hierarchical learning mechanism. This mechanism comprises two complementary components: hierarchical spatial fusion and hierarchical semantic alignment.

1) Hierarchical Spatial Fusion: The first strategy is implemented by the Spatial-Context Adapter (SCA), which captures rich contextual information through intra-layer and inter-layer fusion. This process guides the codec to focus on meaningful regions, thereby reducing task-irrelevant spatial redundancy without stacking deep networks. As shown in Fig. 3(a), the input feature F_i is first refined via the intra-layer fusion to produce an enhanced representation F_s^i. Then, F_s^i is fused with the spatial feature $\tilde{F}_s^{i-1}$ generated by the previous SCA_{i-1}, using the inter-layer fusion to produce the final spatial representation $\tilde{F}_s^i$. The process can be formulated as:

$$\begin{aligned} F_s^i &= Fuse_i^{intra}(F_i), \\ \tilde{F}_s^i &= Fuse_i^{inter}(F_s^i, \tilde{F}_s^{i-1}), \end{aligned} \tag{2}$$

where $Fuse_i^{intra}(\cdot)$ and $Fuse_i^{inter}(\cdot)$ represent the intra-layer fusion and the inter-layer fusion process, respectively. Notably, $i \in \{2, 3, 5, 6\}$. When $i \in \{1, 4\}$, the first SCA_i has no SCA_{i-1} to provide the spatial feature $\tilde{F}_s^{i-1}$.

2) Hierarchical Semantic Alignment: The second strategy involves a constraint to ensure that the contextual information captured by the SCA is aligned with downstream needs. The motivation is that conventional codecs tend to preserve pixel-level fidelity rather than task-relevant semantics, which may lead to suboptimal performance. By aligning the hierarchical intermediate representations of the codec with those of the machine vision task model, our framework encourages the codec to retain semantic information critical for task performance. To this end, we employ a hierarchical semantic alignment loss $\mathcal{L}_{HSA}$. Formally, the $\mathcal{L}_{HSA}$ is defined as:

$$\mathcal{L}_{HSA} = \sum_{i=1}^{2} \mathcal{D}(f_{GSFA}^i, f_{task}^i), \tag{3}$$

where $\mathcal{D}$ is defined as the Mean Squared Error (MSE). f_{GSFA}^i denotes the output features of the GSFA at the i-th stage. f_{task}^i denotes the features extracted from the original input image x using a fixed task model. For image classification, f_{task}^i is extracted from the layers of a fixed ResNet-50 [10]. For object detection and instance segmentation, f_{task}^i refers to the features obtained via the Feature Pyramid Network (FPN) structure in the Faster R-CNN [11] and Mask R-CNN [12]. Furthermore, we use channel pooling operations to align the dimensions of f_{GSFA}^i and f_{task}^i.

C. Frequency-Spectrum Adapter

To eliminate task-irrelevant frequency redundancy, we introduce a Frequency-Spectrum Adapter (FSA) to adapt and enhance feature representations in the frequency domain. It comprises two complementary adaptations. The first adaptation is based on the Fast Fourier Transform (FFT), which performs frequency-domain processing by applying learnable modulation to the amplitude spectrum while preserving the phase. The second adaptation utilizes the Wavelet Transform, which is more sensitive to local discontinuities such as edges, fine textures, and noise bursts. By extracting features at multiple scales, the Wavelet-based adaptation effectively complements the FFT-based adaptation by capturing local variations that the FFT may fail to represent adequately. As shown in Fig. 3(b), the feature F_i is processed through two adaptations to produce the final output $\tilde{F}_f^i$. This process can be defined as follows:

$$\begin{aligned} \tilde{F}_f^i &= FSA_i(F_i) \\ &= Adpt_i^F(F_i) + Adpt_i^W(F_i), \end{aligned} \tag{4}$$

where $Adpt_i^F(\cdot)$ and $Adpt_i^W(\cdot)$ represent the FFT-based and the Wavelet-based adaptations at the i-th stage. Notably, in the Wavelet-based adaptation, we apply a level-1 Haar wavelet decomposition [13] and use only the high-frequency subbands (LH/HL/HH) to emphasize high-frequency details.

D. Gated Fusion Adapter

The Gated Fusion Adapter (GFA) dynamically balances the contributions of spatial and frequency adaptations by adaptively adjusting their weights according to the semantic relevance derived from both domains. First, the features $\tilde{F}_s^i$ and $\tilde{F}_f^i$ are concatenated, then passed sequentially through two convolutional layers followed by a ReLU activation, a final convolutional layer, and a sigmoid function to produce the gating map ρ. Building upon Eq. (1), the gating mechanism of the GFA can be further expanded as follows:

$$\tilde{F}_i = F_i + Conv(\rho \cdot \tilde{F}_s^i + (1 - \rho) \cdot \tilde{F}_f^i), \tag{5}$$

where $\tilde{F}_s^i$ and $\tilde{F}_f^i$ represent the outputs of $SCA_i(\cdot)$ and $FSA_i(\cdot)$, respectively.

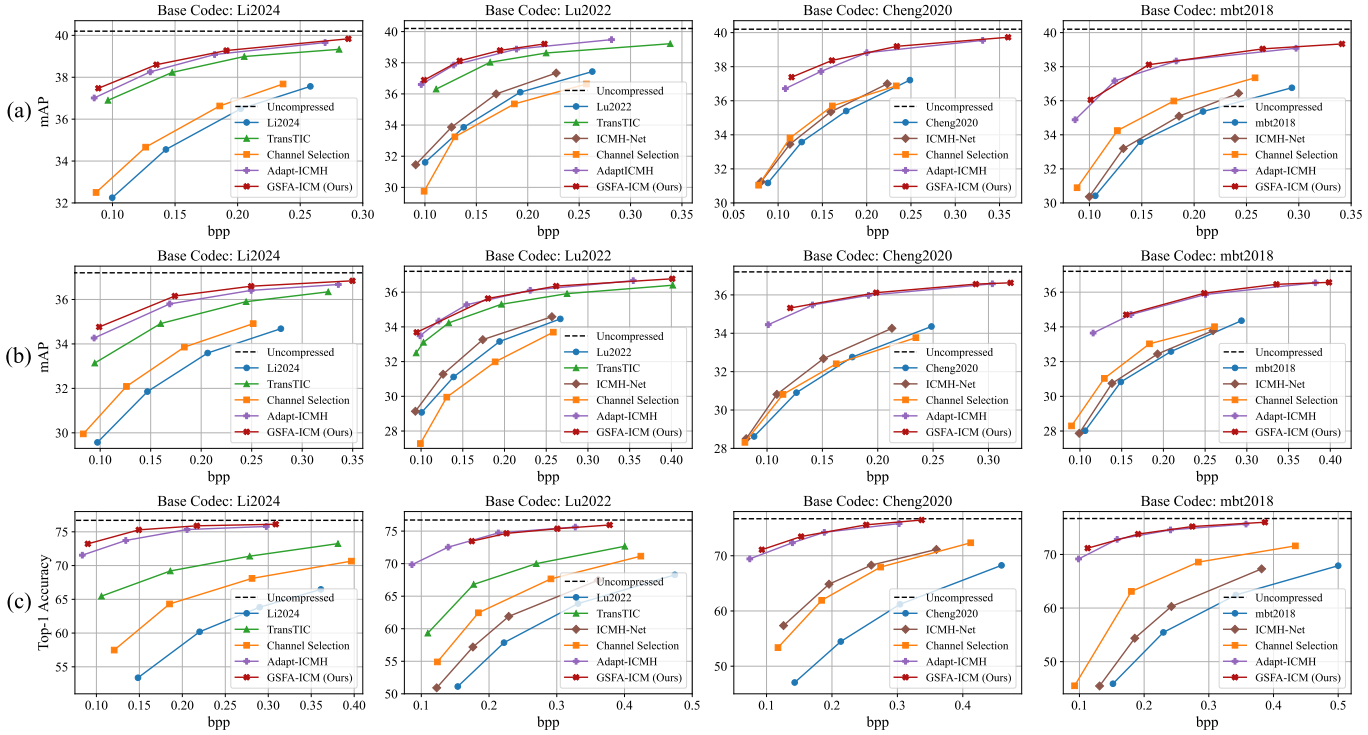


Fig. 4. The rate-task performance across various existing codecs on three tasks: (a) object detection, (b) instance segmentation, and (c) image classification.

E. Training Loss Function

To jointly optimize the compression efficiency, semantic alignment, and perceptual quality, the overall loss function $\mathcal{L}_{total}$ is formulated as:

$$\mathcal{L}_{total} = \mathcal{R} + \lambda_1 \cdot \mathcal{L}_{HSA} + \lambda_2 \cdot \sum_{i=1}^{5} \mathcal{D}(f^i_{task}, \hat{f}^i_{task}), \tag{6}$$

where $\mathcal{R}$ denotes the overall estimated bitrate [14]. λ_1 and λ_2 are trade-off terms to balance the task performance and bitrate savings. Notably, the $\hat{f}^i_{task}$ derived from the reconstructed image $\hat{x}$ is also extracted using a fixed task model.

III. Experiments

A. Experimental Settings

1) Training Details and Datasets: To validate the generalization capability of our framework, four representative NIC models are adopted as pretrained base codecs from two paradigms: the CNN-based (mbt2018 [22] and Cheng2020 [21]) and Transformer-based methods (Lu2022 [19] and Li2024 [15]). Our framework is evaluated on three tasks: image classification, object detection, and instance segmentation.

Our GSFA-ICM is trained utilizing the loss function defined in Eq. (6). Specifically, the image classification task is trained on the ImageNet-*train* dataset [23] for 8 epochs with a batch size of 16, while both object detection and instance segmentation tasks are trained on the COCO2017-*train* dataset [24] for 40 epochs with a batch size of 8. The Adam optimizer is employed with a learning rate of $1e{-}4$. Images are randomly cropped and resized to 256×256 for training machine vision tasks. For the task-specific perceptual distortion, we follow the method [18], employing pretrained ResNet-50 [10], Faster R-CNN [11], and Mask R-CNN [12] for image classification, object detection, and instance segmentation, respectively. Notably, we only retrain the GSFA for each task while keeping the base codec and the task network frozen. More details are shown in the supplementary materials.

2) Evaluation: Bits per pixel (bpp) is utilized to evaluate compression bitrates. For image classification, the pretrained ResNet-50 [10] model from the torchvision library is adopted as the evaluation backbone, with top-1 accuracy used as the quality metric on the ImageNet-*val* dataset [23]. Regarding object detection and instance segmentation, the pretrained Faster R-CNN [11] and Mask R-CNN [12] models from the detectron2 library are employed as the evaluation backbones, respectively. Specifically, all comparative methods are evaluated on the COCO2017-*val* dataset [24], utilizing mean Average Precision (mAP) as the performance metric for both object detection and instance segmentation tasks.

3) State-Of-The-Art Methods: The proposed GSFA-ICM is compared with State-Of-The-Art (SOTA) methods in terms of rate-task performance, including Adapt-ICMH [18], TransTIC [16], ICMH-Net [20], and Channel Selection [17]. It is noteworthy that all methods employ the pretrained base codec as a backbone, which remains frozen during training. The rate-task performance of the original backbone codecs is also evaluated to serve as a baseline for comparison. Additionally, since TransTIC [16] is applicable only to Transformer-based codecs, comparisons involving this method are restricted to Lu2022 [19] and Li2024 [15].

TABLE I
COMPARISON OF RATE-TASK PERFORMANCE AND NUMBER OF TRAINABLE PARAMETERS OF DIFFERENT METHODS WITH MULTIPLE BASE CODECS.

Base Codec	Methods	Image Classification		Object Detection		Instance Segmentation		Trainable
		BD-rate↓	BD-acc↑	BD-rate↓	BD-mAP↑	BD-rate↓	BD-mAP↑	Params ↓(M)
Li2024 [15]	Base Codec	0	0	0	0	0	0	70.97(100%)
	TransTIC [16]	-67.61%	9.81	-52.12%	3.22	-46.79%	2.53	6.89(9.7%)
	Channel Selection [17]	-37.94%	5.66	-12.80%	0.73	-17.79%	0.91	5.43(7.7%)
	Adapt-ICMH [18]	-89.97%	16.60	-60.37%	3.59	-58.84%	3.29	1.66(2.3%)
	GSFA-ICM (Ours)	**-93.45%**	**16.63**	**-64.63%**	**3.76**	**-66.13%**	**3.55**	**1.26(1.8%)**
Lu2022 [19]	Base Codec	0	0	0	0	0	0	7.51(100%)
	TransTIC [16]	-58.32%	9.96	-46.30%	2.77	-48.47%	2.70	1.61(21.4%)
	ICMH-Net [20]	-18.75%	3.36	-9.07%	0.63	-10.77%	0.65	3.98(53.0%)
	Channel Selection [17]	-37.17%	6.28	6.84%	-0.55	16.51%	-0.94	0.91(12.1%)
	Adapt-ICMH [18]	**-88.57%**	**16.90**	-55.14%	3.55	-54.43%	3.21	0.28(3.7%)
	GSFA-ICM (Ours)	-86.34%	14.92	**-56.96%**	**4.01**	**-58.69%**	**3.28**	**0.22(2.9%)**
Cheng2020 [21]	Base Codec	0	0	0	0	0	0	26.60(100%)
	ICMH-Net [20]	-47.46%	10.4	-8.81%	0.53	-12.18%	0.74	4.43(16.6%)
	Channel Selection [17]	-41.58%	8.77	-11.66%	0.73	5.22%	0.22	1.34(4.8%)
	Adapt-ICMH [18]	-87.56%	20.27	-49.34%	3.13	-59.90%	3.48	0.41(1.5%)
	GSFA-ICM (Ours)	**-89.93%**	**20.96**	**-54.93%**	**3.30**	**-62.13%**	**3.56**	**0.31(1.2%)**
mbt2018 [22]	Base Codec	0	0	0	0	0	0	7.03(100%)
	ICMH-Net [20]	-15.99%	3.55	-6.02%	0.39	-4.97%	0.31	3.98(56.6%)
	Channel Selection [17]	-41.30%	9.90	-23.07%	1.52	-15.09%	1.05	0.91(12.9%)
	Adapt-ICMH [18]	-82.00%	18.71	-56.17%	3.84	-52.65%	3.17	0.28(4.1%)
	GSFA-ICM (Ours)	**-85.50%**	**19.13**	**-59.65%**	**3.93**	**-54.74%**	**3.27**	**0.22(3.1%)**

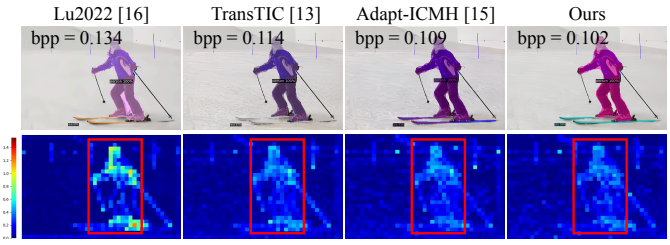


Fig. 5. Qualitative results of our GSFA-ICM with other SOTA methods. For different subgraphs, the first row is the visualization results of our GSFA-ICM compared with other methods for instance segmentation on a single image from the COCO2017 dataset, and the second row is the bit allocation map for latent $\hat{y}$ of each method. The values of bit allocation maps denote the average negative log likelihood of each element in $\hat{y}$ across all channels.

B. Experimental Results

To provide a comprehensive evaluation, the BD-rate [25] metric is reported, which represents the percentage of bitrate change while achieving the same task performance compared to the anchor. Furthermore, BD-mAP results are adopted to assess the average performance gains at equivalent bitrates.

1) Rate-task Performance: The rate-task performance of our GSFA-ICM and the SOTA methods is shown in Fig. 4. We observe that our GSFA-ICM maintains a competitive rate-task performance compared to existing methods in image classification, object detection, and instance segmentation. More results are shown in the supplementary materials.

We compute the BD-rate, accuracy improvements, and the number of trainable parameters for different methods across multiple codecs, as summarized in Table I. These metrics provide a comprehensive assessment of each method in terms of both efficiency and effectiveness, highlighting the lightweight nature and practical advantages of our GSFA-ICM framework. Our GSFA-ICM achieves superior performance to other SOTA methods with fewer trainable parameters (equivalent to about 3% of the base codec parameters), which demonstrates the parameter efficiency of our GSFA-ICM.

TABLE II
COMPARISON OF COMPUTATIONAL METRICS FOR DIFFERENT METHODS.

Methods	GFLOPs	kMACs/pixel	Params(M)
TransTIC [16]	33.117	505.329	1.61
Adapt-ICMH [18]	22.044	336.358	0.28
GSFA-ICM (Ours)	**21.273**	**324.602**	**0.22**

TABLE III
ABLATION STUDY FOR DIFFERENT CONFIGURATIONS ON THE INSTANCE SEGMENTATION TASK. (✓ = ENABLED, ✗ = DISABLED).

Models	HSF	HSA	FFT	Wavelet	GFA	BD-rate↓	Params (M)
Model 1	✗	✓	✓	✓	✓	12.58%	0.21
Model 2	✓	✗	✓	✓	✓	8.07%	0.22
Model 3	✓	✓	✗	✓	✓	20.73%	0.18
Model 4	✓	✓	✓	✗	✓	11.36%	0.19
Model 5	✓	✓	✓	✓	✗	14.74%	0.20

2) Qualitative Results: Fig. 5 presents the visualization results and the corresponding bit allocation maps produced by the compared methods for instance segmentation on a single image from the COCO2017 dataset. At a lower bitrate, our GSFA-ICM achieves superior visualisation results. To reduce the bitrate while maintaining effective task performance, our GSFA-ICM allocates more bits to the foreground than to the background, and more bits to the edges of the object, thereby reducing the bit allocation for the details inside the object.

C. Complexity Comparison

The complexity of the competing methods built on the base codec Lu2022 [19] is evaluated on the COCO2017 dataset

in terms of giga floating-point operations (GFLOPs), kilo-multiply-accumulate-operations per pixel (kMACs/pixel), and trainable parameters, as summarized in Table II. Our GSFA-ICM achieves the lowest complexity compared to Adapt-ICMH [18] and TransTIC [16]. More complexity comparison results are shown in the supplementary materials.

D. Ablation Study

As shown in Table III, we conduct ablation studies of our GSFA-ICM for instance segmentation on the COCO2017 dataset using the base codec Lu2022 [19]. Notably, our GSFA-ICM serves as the anchor and uses 0.22M trainable parameters.

1) Effectiveness of the Spatial-Context Adapter: Model 1 excludes the HSF and retains only the HSA, while Model 2 removes the HSA and preserves the HSF. Compared to the GSFA-ICM, Model 1 exhibits a substantial BD-rate increase of 12.58% with a marginal parameter reduction of 0.01M. Similarly, Model 2 results in an 8.07% increase in BD-rate without affecting the number of parameters. The results demonstrate that both the HSF and HSA individually contribute to performance improvements, and their combination produces the best results with a negligible increase in the number of parameters.

2) Effectiveness of the Frequency-Spectrum Adapter: Model 3 and Model 4 utilize only the Wavelet-based or FFT-based adaptation, respectively. Both Model 3 and Model 4 exhibit performance degradation compared to the GSFA-ICM. This comparison demonstrates that the two transforms capture complementary spectral features. Although the integration of both adaptations in the GSFA-ICM results in a slight increase in parameters, this overhead is considered tolerable given the significant performance gains achieved.

3) Effectiveness of the Gated Fusion Adapter: To assess the gated fusion strategy, Model 5 replaces the Gated Fusion Adapter with a simple element-wise addition operation. Model 5 suffers a severe performance drop, indicated by a 14.74% increase in BD-rate. While the simple addition in Model 5 offers a trivial reduction in parameters, these results confirm that the GFA is crucial for dynamically balancing spatial and frequency information.

IV. Conclusion

In this paper, we propose a new PEFT-based ICM framework, GSFA-ICM. Specifically, we introduce the Gated Spatial-Frequency Adapter (GSFA) to eliminate redundancy in the spatial and frequency domains through PEFT, enabling rapid adaptation to machine vision tasks. The proposed GSFA is plug-and-play and compatible with existing NIC models. Extensive experiments demonstrate that our GSFA-ICM consistently outperforms existing SOTA methods on multiple machine vision tasks, while requiring fewer trainable parameters.

References

[1] Jinming Liu, Heming Sun, and Jiro Katto, "Learned image compression with mixed transformer-cnn architectures," in *CVPR*, 2023, pp. 14388–14397.

[2] Chuqin Zhou, Guo Lu, Jiangchuan Li, Xiangyu Chen, Zhengxue Cheng, Li Song, and Wenjun Zhang, "Controllable distortion-perception tradeoff through latent diffusion for neural image compression," in *AAAI*, 2025, vol. 39, pp. 10725–10733.

[3] Siqi Wu, Yinda Chen, Dong Liu, and Zhihai He, "Conditional latent coding with learnable synthesized reference for deep image compression," in *AAAI*, 2025, vol. 39, pp. 12863–12871.

[4] Yuanchao Bai, Xu Yang, Xianming Liu, Junjun Jiang, Yaowei Wang, Xiangyang Ji, and Wen Gao, "Towards end-to-end image compression and analysis with transformers," in *AAAI*, 2022, vol. 36, pp. 104–112.

[5] Yuxiao Sun, Yao Zhao, Meiqin Liu, Chao Yao, and Weisi Lin, "Embedding compression distortion in video coding for machines," in *ICME*, 2025, pp. 1–6.

[6] Stefano Della Fiore, Alessandro Gnutti, Marco Dalai, et al., "TFIC: End-to-end text-focused image compression for coding for machines," *arXiv preprint arXiv:2503.19495*, 2025.

[7] Ruoyu Feng, Xin Jin, Zongyu Guo, Runsen Feng, Yixin Gao, Tianyu He, Zhizheng Zhang, Simeng Sun, and Zhibo Chen, "Image coding for machines with omnipotent feature learning," in *ECCV*, 2022, pp. 510–528.

[8] Jiancheng Zhao, Xiang Ji, Zhuoxiao Li, Zunian Wan, Weihang Ran, Mingze Ma, Muyao Niu, Yifan Zhan, Cheng-Ching Tseng, and Yinqiang Zheng, "All-in-one transferring image compression from human perception to multi-machine perception," *arXiv preprint arXiv:2504.12997*, 2025.

[9] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al., "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 220–235, 2023.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian 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.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask R-CNN," in *ICCV*, 2017, pp. 2961–2969.

[13] K Senthamilselvan and Lakshmi Dhevi, "Wireless transmission based image quality analysis using uni-level haar wavelet transform," *Circuits and Systems*, vol. 7, no. 8, pp. 1816–1821, 2016.

[14] Johannes Ballé, Valero Laparra, and Eero P Simoncelli, "End-to-end optimized image compression," *arXiv preprint arXiv:1611.01704*, 2016.

[15] Han Li, Shaohui Li, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong, "Frequency-aware transformer for learned image compression," in *ICLR*, 2024.

[16] Yi-Hsin Chen, Ying-Chieh Weng, Chia-Hao Kao, Cheng Chien, Wei-Chen Chiu, and Wen-Hsiao Peng, "TransTIC: Transferring transformer-based image compression from human perception to machine perception," in *ICCV*, 2023, pp. 23297–23307.

[17] Jinming Liu, Heming Sun, and Jiro Katto, "Improving multiple machine vision tasks in the compressed domain," in *ICPR*, 2022, pp. 331–337.

[18] Han Li, Shaohui Li, Shuangrui Ding, Wenrui Dai, Maida Cao, Chenglin Li, Junni Zou, and Hongkai Xiong, "Image compression for machine and human vision with spatial-frequency adaptation," in *ECCV*, 2024, pp. 382–399.

[19] Ming Lu, Peiyao Guo, Huiqing Shi, Chuntong Cao, and Zhan Ma, "Transformer-based image compression," *arXiv preprint arXiv:2111.06707*, 2021.

[20] Lei Liu, Zhihao Hu, Zhenghao Chen, and Dong Xu, "ICMH-Net: Neural image compression towards both machine vision and human vision," in *ACM MM*, 2023, pp. 8047–8056.

[21] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *CVPR*, 2020, pp. 7939–7948.

[22] David Minnen, Johannes Ballé, and George D Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *in NeurIPS*, vol. 31, 2018.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al., "Microsoft COCO: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[25] Gisle Bjontegaard, "Calculation of average PSNR differences between RD-curves," *ITU-T SG16, Doc. VCEG-M33*, 2001.