Beyond Compression: Revisiting the Encoder for Multi-Task Learning in AI-Based Image Compression

Mohamed El-Amine Bellebna[1*], Sid-Ahmed Berrani[1,2], Mohamed Riadh Temmar[1], Jean-Luc Dugelay[3]

[1*]Ecole Nationale Polytechnique, 10 rue des Frères Oudek, 16200, Algiers, Algeria.
[2]National School of Artificial Intelligence, Sidi Abdellah Campus, 16201, Algiers, Algeria.
[3]EURECOM, 450 route des Chappes 06410, Biot, France.

*Corresponding author(s). E-mail(s): mohamed_el_amine.bellebna@g.enp.edu.dz;
Contributing authors: sidahmed.berrani@ensia.edu.dz;
mohamed_riadh.temmar@g.enp.edu.dz; jean-luc.dugelay@eurecom.fr;

Abstract

This paper proposes a multi-task framework for learning-based image compression in which multiple tasks share a common latent representation while preserving compatibility with a single frozen reconstruction decoder. Unlike existing approaches that retrain both encoder and decoder for each task configuration, the proposed method adapts only the encoder and task-specific heads, maintaining decoder standardization and interoperability. Built upon the HiFiC codec, the framework supports additional tasks such as image super-resolution and facial feature extraction from the compressed domain. An adaptive multi-task loss balances compression efficiency and task performance. Experiments at different bitrates demonstrate that heterogeneous tasks can be integrated within a shared latent space while preserving competitive rate-distortion performance. These results support the development of interoperable AI-based compression systems for both visual reconstruction and downstream inference, under a fixed, shared decoder.

Keywords: Learning-based image compression, multi-task learning, rate-distortion, auto-encoders.

1 Introduction

Traditional image compression standards such as JPEG and JPEG 2000 [1, 2] rely on fixed transforms to reduce spatial redundancy. The growing demand for higher compression efficiency and compatibility with automated analysis has motivated the development of learning-based image compression methods, where deep neural networks replace handcrafted components with end-to-end optimized architectures.

Modern learned codecs typically follow an autoencoder structure [3], in which an encoder maps the image to a compact latent representation and a decoder reconstructs it under a rate-distortion objective [4–6]. Representative models such as HiFiC [5], HFD [7], and SegPic [6] demonstrate the effectiveness of this paradigm.

Recent efforts, including JPEG AI[1] [8, 9], aim to standardize learning-based codecs whose latent representations can support both human

[1]https://jpeg.org/jpegai/

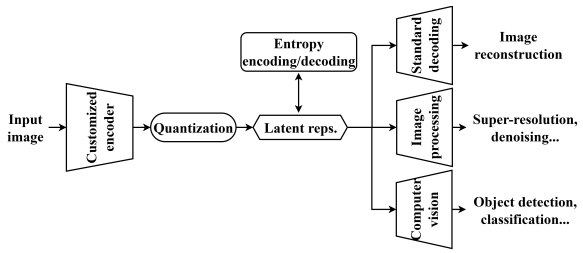

Fig. 1: The scope of JPEG AI core-engine [8].

reconstruction and downstream machine vision tasks (see Fig. 1). In parallel, multi-task learning approaches have explored sharing compressed representations across tasks [10, 11]. However, existing solutions typically retrain both encoder and decoder for each task configuration, producing tightly coupled codec pairs and limiting interoperability.

In this work, we propose a multi-task framework that preserves decoder standardization while enabling encoder customization. Multiple task-specific encoders generate latent representations compatible with a single frozen reconstruction decoder. This design ensures interoperability at the decoding stage while allowing encoder-side adaptation to different downstream objectives.

We instantiate the framework using HiFiC [5] as backbone, though the proposed design is codec-agnostic and applicable to other architectures. The decoder is kept fixed, while the encoder and task-specific heads are adapted to support tasks such as image super-resolution and facial feature extraction. An adaptive multi-task training strategy dynamically balances compression efficiency and task performance under a shared latent space constraint.

Experiments conducted at two target bitrates demonstrate that heterogeneous tasks can coexist within a single latent representation while maintaining competitive rate-distortion performance and controlled trade-offs across tasks.

This study addresses the following key research questions:

1. How can encoder customization be enabled and implemented for diverse applications while maintaining a single fixed reconstruction decoder?
2. How can adaptive multi-task learning mechanisms be designed to balance compression and task performances in a shared latent space, in alignment with the goals of next-generation AI-based codecs (such as JPEG AI)?

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed framework and training strategy. Section 4 reports experimental results, and Section 5 concludes the paper.

2 Related Works

Multi-task learning has been explored in compression and vision systems through shared latent representations. Early works employed autoencoder-based frameworks where a common encoder feeds multiple decoders to support heterogeneous objectives [12]. In learning-based image compression, hyperprior-based variational models such as HiFiC [5] established strong rate-distortion performance and serve as a backbone for many subsequent architectures [13]. These developments highlight the growing interest in exploiting compact latent spaces beyond pure reconstruction.

More recent approaches aim to jointly optimize compression for human perception and downstream machine tasks [14]. Human-machine codecs introduce multi-branch designs to enable segmentation, detection, or analysis directly from compressed features [15, 16]. In the video domain, VNVC [17] proposes a versatile neural coding framework that jointly optimizes video reconstruction and machine vision tasks through shared intermediate feature representations. However, these methods remain jointly trained, task-specific codec instances where encoder and decoder are optimized together for predefined task bundles. As a result, the produced latent representations are not designed to be interoperable across independently trained encoders under a fixed decoder constraint.

Beyond compression, multi-task learning has shown the effectiveness of shared representations and adaptive loss balancing [18]. In applied settings, compressed-domain representations have also been adapted for face recognition and collaborative intelligence scenarios [19, 20]. These works confirm the potential of shared latent spaces to support heterogeneous objectives while preserving competitive performance.

Despite progress in shared latent representations, current multi-task compression methods

remain task-dependent and do not enforce decoder standardization. JPEG AI explicitly pursues a unified learned bitstream for both human and machine consumption [8, 9]. This objective introduces new constraints on interoperability and standard compliance that are not fully addressed by existing architectures. In this context, our work enforces a frozen decoder constraint, enabling multiple encoder-side adaptations to produce latent representations compatible with a single standardized reconstruction decoder.

3 The Proposed Multi-Task Learning Approach

Our research investigates the feasibility of exploiting the latent representation of input images for multi-task learning within AI-based image compression. In the baseline configuration, limited to image compression and reconstruction, the encoder is trained to minimize a mono-task objective composed of a reconstruction loss and a regularization term enforcing a bitrate constraint. This loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \gamma \cdot r(y) \tag{1}$$

where $\mathcal{L}_{\text{rec}}$ denotes the reconstruction loss, $r(y)$ is the estimated bitrate of the latent representation y, and γ is a coefficient that balances the trade-off between reconstruction fidelity and compression efficiency.

In our multi-task setting, we extend the latent representation to serve additional computer vision and image processing tasks. This introduces new challenges, particularly the need to balance performance across multiple tasks. In particular, if task-specific losses are not properly balanced, dominant tasks can bias the shared representation, leading to suboptimal trade-offs for others.

To address these issues, we adopt a global loss formulation that combines the losses of all tasks using task-specific weights and an adaptive bitrate constraint. The multi-task global loss is:

$$\mathcal{L} = \left(\sum_{i=1}^{N} \lambda_{T_i} \cdot \mathcal{L}_{T_i} \right) + \gamma \cdot r(y) \tag{2}$$

where N is the number of tasks, and $\mathcal{L}_{T_i}$ and λ_{T_i} denote respectively the loss and a static weighting hyperparameter for task T_i. The term $r(y)$ remains the estimated bitrate of the latent representation, and γ is now treated as an adaptive function that changes according to the current bitrate relative to the target. In particular, γ is automatically reduced to a small value (e.g., 2^{-4}) when the target bitrate is reached or exceeded, thus dynamically relaxing the pressure on compression once bitrate requirements are met.

This formulation balances task-specific losses while preventing the bitrate constraint from unnecessarily degrading task performance. By combining static task weights with an adaptive bitrate control, the framework ensures stable optimization and effective multi-task learning under a fixed decoder architecture.

To address these challenges, we consider the following application framework and describe our methodology within this setting.

3.1 Application Framework

The considered application framework is illustrated in Fig. 2. It relies on three pipelines: (1) standard image reconstruction using the pretrained HiFiC decoder [5]; (2) image super-resolution, an image processing task; and (3) a computer vision task focused on facial feature extraction for face recognition task. The decoder, which is the non-trainable standard part of the network, is illustrated by the red box in Fig. 2, and the blue box on the left indicates the customizable encoder trained for the target application. The encoder, along with the other components (yellow boxes), constitutes the trainable elements.

A complexity analysis of the main components of this framework is provided in supplementary material (Section S3).

3.1.1 Image Reconstruction (IR)

Decoder. It is used to reconstruct the original input image from the latent representation generated by the encoder. In our schema the decoder is "frozen" and remains standard throughout the experiments. This approach guarantees interoperability regardless of the encoder's customizations or the targeted application constraints.

In our implementation, we have used the AI-based compression model HiFiC [5]. It is worth pointing out that our approach is not specific to

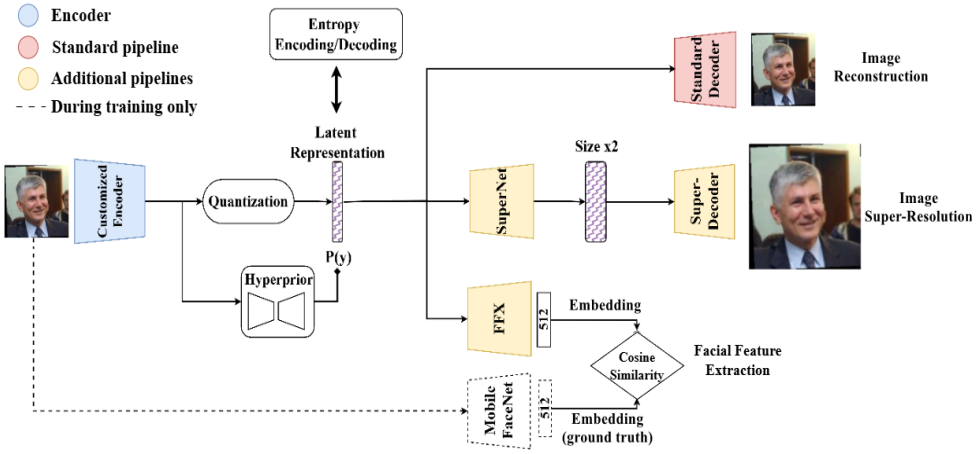

Fig. 2: Our proposed multi-task framework. a unique latent representation for compression and other tasks.

HiFiC; it can be applied to any other AI-based compression models, such as SegPiC [6].

3.1.2 Image Super-Resolution (ISR)

SuperNet. It is inspired by the EDSR model [21]. We have adapted it to function as a task-specific transcoder within our framework. In this role, SuperNet performs latent transcoding by operating directly on the shared latent representation and transforming it into a higher-resolution latent representation suitable for ISR. The dimensionality of the latent representation is doubled here. The used model, described in supplementary material (Fig. S1), consists of 16 residual blocks with 64 filters per convolution layer (3×3), offering a lightweight yet effective design.

SuperDecoder (SRD). It has the same architecture as the decoder used previously for IR. During training, its initial weights are inherited from the HiFiC decoders. This decoder takes the doubled-size latent representation from the SuperNet and generates a super-resolution image.

3.1.3 Facial Feature Extraction (FFX)

FFX. This module serves as a task-specific transcoder that transforms the shared latent representation into a facial feature vector compatible with the output space of MobileFaceNet [22]. It extracts identity-relevant features from the compressed domain without decoding the latent into a full image. As highlighted in yellow in Fig. 2, this intermediate transcoding step enables efficient FFX directly from the compressed domain. The architecture of the FFX module is detailed in supplementary material (Table S1).

Non-trainable MobileFaceNet (Ground Truth Generator). MobileFaceNet is a lightweight convolutional neural network designed for efficient face recognition on resource-constrained devices. It directly extracts robust facial feature embeddings from the input image. In our framework, a separate, non-trainable MobileFaceNet processes the original input image to generate ground truth embeddings. These embeddings are used to evaluate the quality of the facial feature vector produced by the trainable FFX module.

3.2 Training and Loss Normalization

The fundamental structure comprises an encoder E, which takes an original image X as input and produces a latent representation $E(X)$. This representation will serve as the foundation for the three tasks. The first pipeline involves the standard reconstruction, where $E(X)$ is injected into the decoder D resulting in a reconstruction $D(E(X))$. Here, the loss function is:

$$\mathcal{L}_{Deco} = 1 - SSIM(X, D(E(X))) \qquad (3)$$

In the second pathway, we inject the latent representation in SuperNet R yielding a new representation $R(E(X))$ that has the same number of channels as the first latent representation with a

double image resolution. This zoomed-out representation is injected in the decoder SRD resulting in a reconstruction $SRD(R(E(X)))$ with double the resolution of the original image. Resulting in the following loss function:

$$\mathcal{L}_{ISR} = 1 - SSIM(X, SRD(R(E(X)))) \quad (4)$$

Here also we have used the $SSIM$ measure to minimize the cost between the ISR image produced by our model and the ground truth. The third pipeline is dedicated to FFX towards face recognition. The latent representation $E(X)$ is introduced into a module F, which extracts essential facial features from the latent representation, producing a 3-channel image denoted as $F(E(X))$. This image is subsequently entered into the MobileFaceNet model M, generating an embedding $\hat{e}$ suitable for face recognition purposes. In order to train the module F, we inject the original image directly into MobileFaceNet resulting in a reference embedding e. Then, the target objective is to minimize the angle between the $\hat{e}$ and the reference embedding e, where the loss function is:

$$\mathcal{L}_{FaceFX} = 1 - \frac{e \cdot \hat{e}}{\|e\| \cdot \|\hat{e}\|} \quad (5)$$

Finally, the global loss function, including normalization coefficients, is given by:

$$\mathcal{L} = \lambda_{Deco} \cdot \mathcal{L}_{Deco} + \lambda_{ISR} \cdot \mathcal{L}_{ISR} + \lambda_{FaceFX} \cdot \mathcal{L}_{FaceFX} + \gamma \cdot r(y) \quad (6)$$

The bitrate term $r(y)$ in Equation (6) is computed as $r(y) = -\log P(y)$, where $P(y)$ is the probability model of the quantized latent y estimated by the hyperprior [4]. This follows HiFiC's entropy-based rate estimation [5], where entropy coding (e.g., arithmetic coding [23]) compresses the latent according to $P(y)$.

The training procedure optimizes the global loss function using multiple partial terms, as expressed in Equation (6). Normalization plays a crucial role in multi-task learning, as heterogeneous loss scales can result in unbalanced optimization. In particular, tasks with larger raw losses may dominate the training process, potentially suppressing others, a problem previously identified in [18].

To address this, we adopt a three-phase training strategy that combines static normalization coefficients with an adaptive bitrate-aware term. Each phase plays a distinct role in managing task integration and maintaining overall performance:

1. **Encoder adaptation phase:** The encoder is first adapted to the target dataset via transfer learning on the reconstruction task to establish a reconstruction baseline.
2. **Task incorporation phase:** Each target task is trained independently from the shared latent representation. During this phase, the encoder remains frozen to preserve the original compression pipeline, and only the task-specific pipeline is optimized. This setup allows incorporating new tasks (e.g., ISR or FFX) without altering the latent representation. The best loss value achieved for each task is recorded and later used to normalize its contribution during joint optimization.
3. **Joint-learning phase:** In this phase, the encoder is released for training and optimized jointly with multiple task pipelines. To preserve the integrity of the latent representation and ensure that the output remains decodable, the reconstruction task is always included in the loss function. Static normalization weights λ_{T_i} are computed as the inverse of the best loss values $\hat{\mathcal{L}}_{T_i}$ obtained during the task incorporation phase. Because training resumes from the pretrained encoder and task-specific weights corresponding to these $\hat{\mathcal{L}}_{T_i}$ values, each normalized loss starts at a value of 1 at $t = 0$:

$$\lambda_{T_i} \cdot \mathcal{L}_{T_i}(t = 0) = \frac{1}{\hat{\mathcal{L}}_{T_i}} \cdot \hat{\mathcal{L}}_{T_i} = 1. \quad (7)$$

This provides a balanced initialization for multi-task optimization and ensures equal early contributions across all tasks.

Additionally, the global loss function incorporates an adaptive coefficient γ, which dynamically controls the bitrate term $r(y)$ to ensure efficient compression during training. Here, N denotes the number of tasks used in the multi-task learning setup. For the main analysis, we consider the following adaptive function for γ:

$$\gamma = \begin{cases} N \cdot (1 + \hat{r}(y) - r_{\text{target}}), & \text{if } \hat{r}(y) > r_{\text{target}} \\ 2^{-4}, & \text{otherwise} \end{cases} \tag{8}$$

Here, $\hat{r}(y)$ refers to the actual bitrate.

4 Experiments

We have conducted an extensive validation of our methodology through a series of experiments to assess its effectiveness. Our models are trained using Adam optimizer with the following hyper-parameters: batch size, number of epochs, learning rate and weight decay initialized to $10, 80, 10^{-4}$ and 10^{-6} respectively. The learning rate is updated after each epoch using the following function: $lr_t = lr_{t-1} * 0.95^{epoch}$.

Tasks are evaluated independently and jointly to measure the impact of multi-task integration.

4.1 Datasets

Our experimental evaluation employs multiple datasets covering both facial and natural scenes to assess the proposed multi-task framework:

CelebA Dataset [24]: 162,770 training images, 19,867 validation images, and 19,962 test images with diverse facial identities and poses. CelebA is used for training and evaluation of IR, ISR, and FFX. Horizontal flip data augmentation is applied during training to increase diversity and improve generalizability.

Kodak Dataset [25]: 24 high-quality natural images, used exclusively for testing IR and ISR.

J16 Dataset [26]: 16 high-resolution natural images (1336×872 to 3680×2456). This dataset is also used solely for testing.

During training, CelebA images are resized to 256×256 and bicubically downsampled by a factor of 2 to produce 128 × 128 encoder inputs. The encoder operates exclusively at 128 × 128. For IR and FFX, the 128 × 128 images serve as ground truth, whereas for ISR the corresponding 256×256 images are used as supervision targets.

At test time, CelebA follows the same preprocessing. Kodak and J16 images retain their native resolution and are downsampled by a factor of 2 to generate low-resolution inputs. For IR, the low-resolution input serves as ground truth, whereas for ISR the original images serve as ground truth.

Sample images are shown in Supplementary Material (Fig. S2).

4.2 Evaluation Metrics

Reconstruction and super-resolution are evaluated using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). SSIM is used as the training loss to promote perceptual quality, while both metrics are reported at test time.

For FFX, identity preservation is assessed using Cosine Similarity (CosSim) between embeddings extracted from the decoded image and a non-trainable reference model. Higher CosSim values indicate better feature consistency.

Unless otherwise specified, all reported results correspond to the mean per-image metric over the entire test set (Tables 1-4).

4.3 Training using the Standard Latent Representation - Baseline

To assess the impact of compression level, we use two pretrained HiFiC models: **HiFiC Low** (trained with a target bitrate of 0.14 bpp) and **HiFiC High** (trained with a target bitrate of 0.45 bpp), as defined in [5]. These fixed encoder-decoder instances allow evaluation across distinct compression regimes.

We begin by establishing a baseline to assess how well the pretrained HiFiC encoder, originally trained for IR only, supports additional task modules. In this first integration phase, we fine-tune the encoder on the CelebA dataset for IR only, while keeping the decoder frozen. This preserves compatibility, ensuring that the resulting latent representation remains decodable by the standard, shared, fixed decoder.

After adapting the encoder for IR, we freeze it and train the additional tasks from scratch, specifically the ISR pipeline (comprising the SuperNet and SuperDecoder) and the FFX module. This simulates an early deployment scenario where new capabilities are added to the AI-based compression framework instance without retraining the core compression pipeline.

Table 1 reports performance on CelebA (validation and test), J16, and Kodak under HiFiC High and HiFiC Low. Under HiFiC High (0.468 bpp), the model achieves 31.56 dB PSNR and 0.941 SSIM on CelebA (Test) for IR, with a

Table 1: Baseline performance for IR, ISR, and FFX.

Quality	Dataset	IR		ISR		FFX	bpp
		PSNR	SSIM	PSNR	SSIM	CosSim	
HiFiC Low	CelebA (Val)	27.16	0.871	28.84	0.865	0.683	0.152
	CelebA (Test)	27.24	0.872	28.94	0.868	0.684	0.152
	J16	23.38	0.741	29.21	0.796	n.a.	0.175
	Kodak	23.53	0.719	29.55	0.787	n.a.	0.168
HiFiC High	CelebA (Val)	31.46	0.940	32.43	0.921	0.692	0.469
	CelebA (Test)	31.56	0.941	32.56	0.923	0.692	0.468
	J16	26.13	0.837	30.96	0.839	n.a.	0.490
	Kodak	26.12	0.821	31.11	0.827	n.a.	0.485

cosine similarity of 0.692 for FFX. Under HiFiC Low (0.152 bpp), performance decreases (e.g., 27.24 dB PSNR and 0.872 SSIM on CelebA Test), as expected under stronger compression.

These results reflect a latent representation optimized for IR only. Since the encoder is not updated for the additional tasks, this phase serves as a compatibility check. In Section 4.4, the encoder is jointly fine-tuned with the new tasks to learn task-relevant features under both HiFiC Low and High settings.

4.4 Considering Jointly Two Tasks

We now consider the proposed multi-task learning framework and, as a first step, address the case of jointly training two tasks. In all the following experiments, the reconstruction decoder remains frozen and shared across configurations, while the encoder and task-specific heads are updated. We first examine the joint training of two-task pairs, where each additional task (ISR or FFX) is trained together with standard IR. We then extend the framework to the three-task setting that combines IR, ISR (an image processing task), and FFX (a computer vision task) within a single model. In the following multi-task experiments, we use pretrained models from the standard latent representation as baselines, with performance metrics provided in Section 4.3. Additionally, we implement an adaptive loss function to control the compression task through the dynamic adjustment of the weighting factor γ, as defined in Equation (8). This adaptive strategy enables flexible bitrate adjustment based on task demands, ensuring effective task balancing without traditional loss normalization. This approach helps achieve a dynamic balance between tasks, further enhancing the overall performance of the multi-task learning framework.

4.4.1 Image Reconstruction+ISR

In this configuration, we jointly train IR and ISR tasks within our framework. Both tasks are structurally aligned and rely on similar visual fidelity objectives, primarily optimized using SSIM. As a result, the joint optimization leads to stable and sometimes improved performance without introducing task conflict in the shared latent representation.

Table 2 shows that under the **HiFiC High** setting (average 0.495 bpp), CelebA (Test) IR remains robust, maintaining a PSNR of 31.56 dB while slightly improving SSIM from 0.941 to 0.944. ISR benefits even more significantly, improving from 32.56 dB PSNR and 0.923 SSIM (baseline) to 33.00 dB PSNR and 0.929 SSIM. This improvement suggests that joint optimization enables the encoder to retain richer, task-relevant features, information likely discarded when optimizing solely for reconstruction, thereby enhancing ISR without compromising the primary compression objective.

On the **J16** and **Kodak** datasets, joint training also leads to near-equivalent or slightly enhanced performance. For example, the reconstruction PSNR on J16 drops marginally from 26.13 dB to 26.04 dB, but SSIM increases from 0.837 to 0.838. On Kodak, both reconstruction and ISR metrics are nearly identical to the baseline, confirming the stability of joint learning across unseen test domains.

Under the **HiFiC Low** regime (average 0.152 bpp), we observe a similar pattern. The reconstruction performance on CelebA (Test) drops slightly from 27.24 dB to 26.93 dB PSNR, with a minimal SSIM change (0.872 to 0.871). However, ISR remains comparable, and J16/Kodak show almost identical scores to their respective baselines.

Table 2: Joint training performance for IR and ISR.

Quality	Dataset	IR		ISR		bpp
		PSNR	SSIM	PSNR	SSIM	
HiFiC Low	CelebA (Val)	26.86	0.869	28.60	0.868	0.152
	CelebA (Test)	26.93	0.871	28.70	0.871	0.152
	J16	23.13	0.739	29.05	0.798	0.176
	Kodak	23.19	0.717	29.28	0.790	0.169
HiFiC High	CelebA (Val)	31.46	0.943	32.86	0.928	0.495
	CelebA (Test)	31.56	0.944	33.00	0.929	0.495
	J16	26.04	0.838	31.07	0.842	0.505
	Kodak	25.91	0.822	31.05	0.830	0.500

Importantly, the bitrates remain consistent with their respective compression targets: 0.152 bpp for HiFiC Low and 0.495 bpp for HiFiC High. The joint model meets the compression constraint while achieving better task-specific fidelity than the baseline that used a reconstruction-only latent. These results suggest that jointly training IR and ISR tasks is compatible with a shared encoder–decoder pipeline and can improve perceptual quality in the evaluated setting.

4.4.2 Image Reconstruction+FFX

When jointly optimizing IR and FFX tasks, we observe a trade-off between spatial fidelity and identity preservation. As shown in Table 3, the CelebA (Test) reconstruction performance under the **HiFiC High** setting (0.486 bpp) decreases slightly from 31.56 dB PSNR and 0.941 SSIM (baseline) to 30.55 dB PSNR and 0.933 SSIM. Meanwhile, the cosine similarity for FFX improves from 0.692 to 0.831, confirming that the latent representation aligns better with identity features.

Similar patterns hold for **HiFiC Low**, where the PSNR drops from 27.24 dB to 25.99 dB and SSIM from 0.872 to 0.853, while the cosine similarity improves markedly from 0.684 to 0.802. These results suggest that although some reconstruction quality is sacrificed, the encoder adapts to prioritize features beneficial for face recognition.

On the **J16** and **Kodak** datasets, where facial identity labels are unavailable, only reconstruction metrics are reported. The drops in PSNR are consistent with those observed on CelebA, but SSIM shows only a small change relative to the baseline, indicating that the general image structure is preserved.

The bitrate constraints are respected: the encoder-decoder pipeline maintains 0.152 bpp in HiFiC Low and 0.486 bpp in HiFiC High. These results show that the shared latent representation can be successfully adapted to serve both visual fidelity and identity preservation.

4.5 Considering Jointly Three Tasks

As additional tasks are incorporated into the framework, ISR and FFX alongside IR, the shared latent representation must support more diverse objectives. This naturally introduces trade-offs, as the encoder must now balance between preserving spatial fidelity, generating high-frequency details, and maintaining identity-aware features.

Despite these competing demands, the jointly trained three-task model achieves consistently high performances across all tasks and datasets (see Table 4)[2]. On the CelebA (Test) set, the reconstruction task maintains 30.28 dB PSNR and 0.933 SSIM, while ISR reaches 32.10 dB PSNR and 0.923 SSIM. The FFX task also benefits, achieving a cosine similarity of 0.811.

These results confirm that joint optimization enables the encoder to embed more task-relevant features without severely compromising individual performance. On J16, for instance, reconstruction drops slightly from 26.13 dB (baseline High) to 25.39 dB, and on Kodak from 26.12 dB to 25.32 dB. A similar trend is observed for SSIM, which shows only a marginal decrease relative to the baseline, whereas cosine similarity on CelebA improves, indicating better identity preservation.

In terms of bitrate, the three-task configuration maintains a reasonable 0.467 bpp under HiFiC High, only marginally lower than the two-task IR + ISR configuration (0.495 bpp) and closely aligned with the baseline (0.468 bpp). This

[2] Additional qualitative reconstructions and super-resolution examples on J16 at low and high bitrates are reported in the Supplementary Material (Figs. S3-S4).

Table 3: Joint training performance for IR and FFX.

Quality	Dataset	IR		FFX	bpp
		PSNR	SSIM	CosSim	
HiFiC Low	CelebA (Val)	25.93	0.852	0.800	0.152
	CelebA (Test)	25.99	0.853	0.802	0.152
	J16	22.74	0.721	n.a.	0.177
	Kodak	22.92	0.699	n.a.	0.170
HiFiC High	CelebA (Val)	30.48	0.932	0.829	0.486
	CelebA (Test)	30.55	0.933	0.831	0.486
	J16	25.59	0.824	n.a.	0.511
	Kodak	25.64	0.808	n.a.	0.504

Table 4: Multi-task framework performance across IR, ISR, and FFX.

Quality	Dataset	IR		ISR		FFX	bpp
		PSNR	SSIM	PSNR	SSIM	CosSim	
HiFiC Low	CelebA (Val)	26.09	0.856	28.30	0.864	0.792	0.159
	CelebA (Test)	26.15	0.857	28.40	0.866	0.794	0.159
	J16	22.73	0.727	28.92	0.795	0.727	0.175
	Kodak	22.85	0.705	29.31	0.787	0.705	0.209
HiFiC High	CelebA (Val)	30.21	0.931	31.98	0.921	0.809	0.467
	CelebA (Test)	30.28	0.933	32.10	0.923	0.811	0.467
	J16	25.39	0.822	30.86	0.840	n.a.	0.500
	Kodak	25.32	0.806	30.87	0.829	n.a.	0.490

indicates that the model can efficiently encode multi-task features while keeping compression overhead minimal, demonstrating the scalability and adaptability of the proposed multi-task framework.

5 Conclusion

This paper presents a multi-task learning framework for AI-based image compression that enables image reconstruction, image super-resolution, and facial feature extraction from a shared latent representation while preserving compatibility with a single frozen reconstruction decoder. By decoupling encoder adaptation from decoder standardization, the proposed approach addresses interoperability constraints aligned with JPEG AI objectives. Experimental results across multiple bitrate regimes demonstrate that heterogeneous tasks can coexist within a shared latent space while maintaining controlled rate-distortion trade-offs.

One important limitation of the proposed approach is that perceptually aligned tasks (like reconstruction and super-resolution) coexist more effectively than semantically divergent tasks (like facial feature extraction).

Future work will thus focus on three directions. First, we aim to characterize task compatibility in multi-task compression, in order to design grouping or training strategies that mitigate interference between heterogeneous objectives.

Second, we will study the scalability of the shared latent representation as the number and diversity of tasks increase, analyzing the trade-off between task diversity and performance.

Finally, we will investigate the role of transcoding as an intermediate transformation between the shared latent space and task-specific pipelines. Instead of processing the latent directly, each task could benefit from an intermediate transformation that aligns the latent representation with its objectives.

Declarations

Funding: Not applicable.

References

[1] Joint Photographic Experts Group (JPEG). JPEG 2000 Part 1: Core Coding System. ISO/IEC JTC 1/SC 29/WG 1; 2000. ISO/IEC 15444-1 — ITU-T T.800.

[2] Christopoulos C, Skodras A, Ebrahimi T. The JPEG2000 still image coding system: an overview. IEEE Transaction Consumer Electronics. 2000;46(4).

[3] Kingma DP, Welling M. Auto-Encoding Variational Bayes. In: Proc. of the Int. Conf. on Learning Representations, Canada; 2014. .

[4] Ballé J, Minnen D, Singh S, Hwang SJ, Johnston N. Variational image compression with a scale hyperprior. In: Proc. of the Int. Conf. on Learning Representations, Canada; 2018. .

[5] Mentzer F, Toderici G, Tschannen M, Agustsson E. High-Fidelity Generative Image Compression. In: Proc. of the Conf. on Neural Information Processing Systems, virtual; 2020. .

[6] Liu Y, Yang W, Bai H, Wei Y, Zhao Y. Region-adaptive transform with segmentation prior for image compression. In: Proc. of the European Conf. on Computer Vision, Italy; 2024. .

[7] Hoogeboom E, Agustsson E, Mentzer F, Versari L, Toderici G, Theis L. High-Fidelity Image Compression with Score-based Generative Models. CoRR. 2023;abs/2305.18231.

[8] Ascenso J, Alshina E, Ebrahimi T. The JPEG AI Standard: Providing Efficient Human and Machine Visual Data Consumption. IEEE Multimedia. 2023;30(1).

[9] Alshina E, Ascenso J, Ebrahimi T. JPEG AI: The First Int. Standard for Image Coding Based on an End-to-End Learning-Based Approach. IEEE Multimedia. 2024;31(4).

[10] Yang M, Yang F, Murn L, Blanch MG, Sock J, Wan S, et al. Task-Switchable Pre-Processor for Image Compression for Multiple Machine Vision Tasks. IEEE Transaction on Circuits and Systems for Video Technology. 2024;34(7).

[11] Cao J, Yao X, Zhang H, Jin J, Zhang Y, Ling BWK. Slimmable Multi-Task Image Compression for Human and Machine Vision. IEEE Access. 2023;11.

[12] Sebai D, Shah AU. Semantic-oriented learning-based image compression by Only-Train-Once quantized autoencoders. Signal, Image and Video Processing. 2023;17(1).

[13] Qiu L, Bai E, Wu Y, Cao Y. Learned image compression with dual frequency branch modulation. Signal, Image and Video Processing. 2025;19(17).

[14] Chamain LD, Racapé F, Bégaint J, Pushparaja A, Feltman S. End-to-end optimized image compression for multiple machine tasks. CoRR. 2021;abs/2103.04178.

[15] Li H, Li S, Ding S, Dai W, Cao M, Li C, et al. Image Compression for Machine and Human Vision with Spatial-Frequency Adaptation. In: Proc. of the European Conf. on Computer Vision, Italy; 2024. .

[16] Guo S, Sui L, Zhang CL, Chen Z, Yang W, Duan L. A Unified Image Compression Method for Human Perception and Multiple Vision Tasks. In: Proc. of the European Conf. on Computer Vision, Italy; 2024. .

[17] Sheng X, Li L, Liu D, Li H. VNVC: A Versatile Neural Video Coding Framework for Efficient Human-Machine Vision . IEEE Transactions on Pattern Analysis & Machine Intelligence. 2024;46(07).

[18] Kendall A, Gal Y, Cipolla R. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In: Proc. of the Conf. on Computer Vision and Pattern Recognition, USA; 2018. .

[19] Wang S, Wang S, Yang W, Zhang X, Wang S, Ma S, et al. Towards Analysis-Friendly Face Representation With Scalable Feature and Texture Compression. IEEE Transaction on Multimedia. 2022;24.

[20] Wang Z, Li F, Zhang Y, Zhang Y. Low-Rate Feature Compression for Collaborative Intelligence: Reducing Redundancy in Spatial and Statistical Levels. IEEE Transaction on Multimedia. 2024;26.

[21] Lim B, Son S, Kim H, Nah S, Lee KM. Enhanced Deep Residual Networks for Single Image Super-Resolution. In: Proc. of the Conf. on Computer Vision and Pattern Recognition Workshops, USA; 2017. .

[22] Chen S, Liu Y, Gao X, Han Z. MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices. In: Proc. of the Chinese Conf. of Biometric Recognition, China; 2018. .

[23] Marpe D, Schwarz H, Wiegand T. Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard. IEEE Transaction on Circuits and Systems for Video Technology. 2003;13(7).

[24] Liu Z, Luo P, Wang X, Tang X. Deep Learning Face Attributes in the Wild. In: Proc. of the Int. Conf. on Computer Vision, Chile; 2015. .

[25] Kodak.: Kodak PhotoCD Dataset. Accessed: January 10, 2026. Available from: http://r0k.us/graphics/kodak/.

[26] JPEG-AI.: Test Images: JPEG-AI MMSP Challenge. Accessed: January 10, 2026. https://jpegai.github.io/test_images/.