When the Teacher Has More Bits: Self-Teacher Latent Distillation for Learned Image Compression

Abdellah El Mennaoui[1,2,3], Ghalia Hemrit[3], Joseph Meehan[3]
, and Jean-Luc Dugelay[1]

[1] EURECOM, Biot, France
[2] Sorbonne University, Paris, France
[3] Huawei Technologies, Valbonne, France

Abstract. Learned image compression (LIC) operates under a rate–distortion (RD) trade-off, where representation quality is constrained by the target bitrate. We revisit knowledge distillation in LIC through the lens of *bitrate asymmetry* and introduce a *self-teacher* distillation framework, where a high-rate instance of a codec supervises multiple lower-rate encoders of identical architecture. Because allocating more bits naturally leads to richer latent representations, the high-rate model provides informative supervision across rate levels. Direct latent matching, however, is problematic under tight rate budgets. We therefore propose *variance-normalized latent distillation* (VNLD), a rate-aware alignment strategy that scales channel-wise supervision by the teacher's variance, selectively transferring stable, informative structure while suppressing components that cannot be reliably reproduced at lower rates. Across different distillation objectives, networks, and bitrate levels, self-teacher distillation, particularly with VNLD, improves RD performance and yields consistent BD-rate gains over RD-only training. Our method remains compatible with fixed-decoder deployments, such as those targeted by JPEG AI standards.

Keywords: Learned Image Compression, Knowledge Distillation, Latent Representation, JPEG AI

1 Introduction

Learned image compression (LIC) has reshaped image coding by optimizing neural encoders and decoders under a rate–distortion (RD) objective [3, 4, 28]. Modern architectures incorporating hierarchical hyperpriors [4], autoregressive models [28], transformers [23, 25], and semantic priors [26] now surpass traditional codecs such as JPEG [30], JPEG 2000 [34], and even VVC [20].

Despite these architectural advances, training paradigms remain largely unchanged: each bitrate operating point is optimized independently using the same RD objective, without exploiting relationships across rate levels. This observation suggests a potential for knowledge transfer across bitrate levels.

Knowledge distillation (KD) [16] has been widely explored in recognition tasks and its application to learned image compression remains limited. In classical KD [16], teacher superiority typically arises from larger model capacity or architectural complexity. In contrast, in learned compression, models trained at higher Lagrangian multipliers λ are stronger simply because they are allowed to allocate more bits. Their latent representations are therefore richer not because of additional parameters, but because the rate constraint is relaxed. As a result, teacher superiority in LIC arises from rate allocation rather than architectural capacity.

Motivated by this observation, we explore whether a codec can effectively learn from itself across bitrates. In this work, we introduce a **self-teacher distillation** paradigm for learned image compression in which a high-rate model supervises several student models trained at lower-rate settings, enabling one-to-many cross-rate supervision. Instead of viewing the teacher as a larger network, we reinterpret it as a less rate-constrained instance of the same architecture. This motivates latent-space distillation, since unlike image-domain perceptual losses that compare reconstructions to the ground truth, the latent representation contains the rate-constrained compression features where bit allocation is formed.

To our knowledge, prior work has not *systematically* studied latent-space distillation *across bitrate operating points* in LIC, where teacher superiority is driven primarily by *rate allowance* rather than model capacity.

Transferring knowledge across bitrate levels is however not straightforward. Unlike conventional KD settings, a low-rate student cannot reproduce a distortion oriented teacher latent representation without violating its RD budget. Direct feature matching therefore imposes unrealistic constraints on the student and can lead to unstable optimization. This limitation complicates the application of distillation in learned compression, as low-rate students cannot directly reproduce distortion-oriented teacher representations without violating their rate budget.

To address this challenge, we propose **Variance-Normalized Latent Distillation (VNLD)**, a rate-aware latent distillation mechanism designed specifically for learned image compression. VNLD scales channel-wise supervision according to the variability of teacher features. Channels with high teacher variance often correspond to fine residual information. Since such components are difficult for a lower-rate student to reproduce under a tighter rate constraint, VNLD attenuates their contribution to the distillation loss. In contrast, stable channels encoding more global structure receive stronger supervision. This produces a rate-compatible supervision signal: rather than forcing the student to imitate the full high-rate teacher representation, VNLD guides it toward the components that can be more reliably transferred at its target bitrate.

We evaluate the self-teacher distillation framework across multiple distillation objectives, including standard latent MSE alignment, contrastive representation distillation (CRD) [36], and the proposed VNLD approach, and across different compression architectures: HPCM [24], BMS [5] and SegPIC [26]. These experi-

ments allow us to conclude that the benefits of self-teacher supervision are not architecture-specific. Across datasets, bitrate levels, and model architectures, self-teacher distillation consistently improves RD performance compared to RD-only training. Among the evaluated distillation objectives, VNLD achieves the most consistent improvements over MSE and CRD, leading to measurable BD-rate reductions on standard benchmark datasets.

A key advantage is that the proposed method operates at the encoder side, naturally supporting deployment scenarios in which the decoder and entropy model remain fixed. Such approach is particularly relevant for emerging compression standards such as JPEG AI [1].

In summary, this work makes the following contributions:

- We introduce a **self-teacher latent distillation framework** where a high-rate instance of the same codec supervises multiple lower-rate encoders sharing the same architecture.
- We propose **variance-normalized latent distillation (VNLD)**, a rate-aware latent alignment strategy that stabilizes cross-rate supervision by modulating the contribution of each latent channel according to its variance.
- We validate the approach across different architectures, datasets, and training settings, including deployment scenarios with frozen decoders.

The remainder of this paper is organized as follows. Section 2 reviews background and related work on distillation in learned image compression. Section 3 introduces the proposed self-teacher framework and the VNLD formulation. Sections 4 and 5 present experimental results under standard learned compression training settings. Section 6 evaluates the method under JPEG AI deployment constraints, and Section 7 concludes the paper.

2 Background and Related Work

Learned image compression models are typically optimized under the classical rate–distortion (RD) objective

$$\mathcal{L}_{RD} = R + \lambda D = -\log_2 p_{\hat{y}}(\hat{y}) + \lambda \|x - \hat{x}\|_2^2, \tag{1}$$

where R denotes the expected bitrate, λ the Lagrangian multiplier, D the reconstruction distortion, $p_{\hat{y}}$ the entropy model of the quantized latent representation $\hat{y}$, and x and $\hat{x}$ denote the input and reconstructed images, respectively.

Different operating points are obtained by varying λ following the variational compression framework of Ballé et al. [3, 4]. Increasing λ relaxes the rate constraint and allows the encoder to allocate more bits to the latent space, resulting in richer latent features at higher bitrates. Although the RD formulation governs how individual models are optimized, it treats each bitrate operating point independently and does not exploit potential relationships between models trained at different rates.

Knowledge distillation (KD) provides a general mechanism for transferring information between models and has been widely used to improve efficiency and representation learning. Initially introduced to compress large networks into smaller ones [16], KD has since expanded to feature-based supervision [32], relational and contrastive representation transfer [36], and self-distillation strategies where supervision emerges within the same model architecture [37]. These developments highlight that aligning internal representations can improve learning beyond simple output imitation.

Recent work has begun exploring KD within learned image compression. For instance, Chen *et al.* [10] propose a modular distillation framework that transfers knowledge from a complex teacher model to a lightweight student in order to reduce computational complexity while preserving rate–distortion performance . Similarly, other studies distill feature representations or entropy-related signals to build smaller and faster codecs [15]. While effective, these approaches largely follow the classical KD paradigm in which teacher superiority arises from architectural capacity or model scale.

In practice, learned image compression presents a different regime. Codecs trained at higher Lagrangian multipliers often perform better simply because they are allowed to allocate more bits, suggesting that teacher superiority may stem from **rate allowance rather than model capacity**. Existing KD approaches do not explicitly account for this property, nor do they study how knowledge can be transferred across bitrate levels within the same architecture.

Practical deployment scenarios also impose additional constraints on LIC systems. Recent standardization efforts, such as JPEG AI [1], aim to define interoperable neural image coding pipelines. In such frameworks, the decoding process is typically fixed in order to ensure compatibility across different encoders and implementations. Consequently, improvements to compression performance must often be achieved without modifying the decoder or entropy model, motivating approaches that focus on encoder-side optimization [11, 33].

3 Method

3.1 Self-Teacher Framework

Building on the RD formulation in Eq. (1), we introduce a self-teacher distillation framework that exploits bitrate asymmetry across operating points.

Instead of using a larger network as teacher, we employ a higher-rate instance of the same codec. A single teacher encoder trained at a high rate produces latent representations y_t that supervise multiple lower-rate student encoders producing y_s. The teacher is frozen during student training and distillation is applied to the *pre-quantization* latent representations. Fig. 1 illustrates the overall training framework.

The student (Encoder_s , Decoder_s) is optimized using the RD objective in Eq. (1) augmented with a latent-space distillation term:

$$\mathcal{L} = R + \lambda D + \beta \mathcal{L}_{KD}. \tag{2}$$

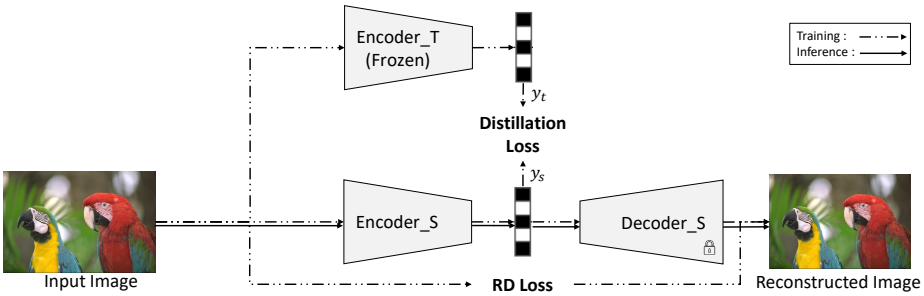

Fig. 1: Self-teacher distillation framework. A high-rate teacher encoder is frozen and supervises lower-rate students through a latent-space distillation loss.

We first consider standard latent alignment strategies based on MSE and contrastive representation distillation (CRD).

MSE Latent Alignment. Let $y_t, y_s \in \mathbb{R}^{C\times H'\times W'}$ denote teacher and student pre-quantization latents. A direct alignment baseline is mean-squared error:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{CH'W'}\|y_s - y_t\|_2^2. \tag{3}$$

Contrastive Representation Distillation (CRD). We also consider a contrastive objective by Tian et al. [36]. Global descriptors are obtained via spatial average pooling, $\mathbf{v}_t = \text{GAP}(y_t)$ and $\mathbf{v}_s = \text{GAP}(y_s)$, followed by projection heads. The student representation is encouraged to remain closer to the teacher than to negative samples using the InfoNCE loss [29]:

$$\mathcal{L}_{\text{CRD}} = -\log \frac{\exp(\mathbf{v}_s^T\mathbf{v}_t/\tau)}{\exp(\mathbf{v}_s^T\mathbf{v}_t/\tau) + \sum_{k=1}^{K}\exp(\mathbf{v}_s^T\mathbf{v}_t^{(k)}/\tau)}. \tag{4}$$

Here τ denotes a temperature parameter controlling the sharpness of the similarity distribution, K is the number of negative samples, and $\mathbf{v}_t^{(k)}$ represents negative teacher descriptors drawn from other samples in the batch.

3.2 Variance-Normalized Latent Distillation (VNLD)

Standard latent alignment implicitly assumes that all teacher channels are equally transferable. However, in LIC, channel variance is often correlated with bitrate allocation: high-variance channels typically encode fine residual details [3, 4, 28] that a lower-rate student cannot reproduce without exceeding its rate budget. As a result, simple MSE-based alignment may impose unrealistic constraints on low-rate students, while CRD avoids direct feature matching through relational supervision. Nevertheless, neither objective explicitly models the fact that teacher and student operate under different bitrate constraints. Consequently,

both treat all latent channels equally, regardless of whether they can realistically be reproduced by the student.

To address these challenges, we introduce Variance-Normalized Latent Distillation (VNLD), a rate-aware supervision mechanism that modulates channel-wise alignment according to the variance of teacher representations.

Let $y_t, y_s \in \mathbb{R}^{C\times H'\times W'}$ denote teacher and student latents before quantization. For each channel c, we compute the teacher standard deviation:

$$\sigma_t^c = \mathrm{Std}(y_t^c), \tag{5}$$

and stabilize small values with

$$\tilde{\sigma}_t^c = \max(\sigma_t^c, \epsilon).$$

The VNLD loss is defined as

$$\mathcal{L}_{VNLD} = \frac{1}{CH'W'} \sum_{c,i,j} \frac{(y_{s,ij}^c - y_{t,ij}^c)^2}{(\tilde{\sigma}_t^c)^2}. \tag{6}$$

Dividing the squared residual by the teacher variance corresponds to variance-weighted residual minimization. Such formulations arise naturally in heteroscedastic regression and uncertainty-aware learning, where observations with larger variance receive lower confidence weights [22]. VNLD can also be interpreted as minimizing a Mahalanobis-type distance in latent space, where channel-wise variance defines a confidence-weighted alignment that adapts supervision strength across latent dimensions [6, 27].

VNLD therefore implements a variance-weighted alignment, where teacher variance acts as an inverse-confidence weight. Unlike standard feature normalization, the variance is computed from the high-rate teacher and reflects rate-dependent channel variability, making the weighting inherently bitrate-aware.

VNLD does not suppress high-variance channels. Instead, it prevents them from dominating the loss, allowing the student to allocate bits according to its own rate constraint: channels with large spatial variability, typically associated with fine residual information at high bitrates, are down-weighted during optimization. Stable structural channels exert stronger supervision.

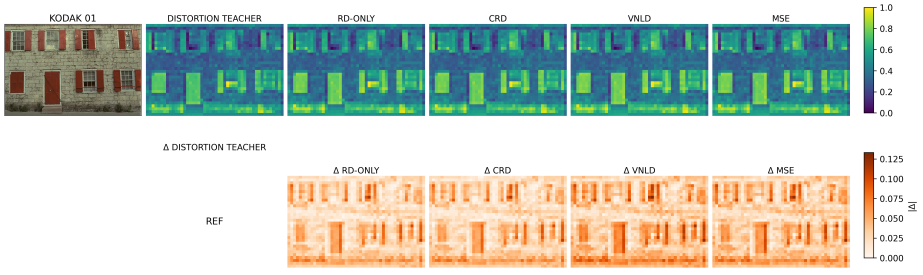

Fig. 2: Visualization of the highest-variance teacher latent channel on Kodak 01 and the corresponding student channels obtained with RD-only, MSE, CRD, and VNLD. The bottom row shows the absolute deviation from the teacher, $\Delta = |y_s - y_t|$.

This behavior is illustrated in Fig. 2, which visualizes the highest-variance teacher latent channel on Kodak 01 and the corresponding student channels obtained with RD-only training, MSE distillation, CRD, and VNLD. The VNLD representation reduces the dominance of unstable high-variance components, supporting the proposed inverse-variance weighting.

This normalization guides the student toward a rate-compatible approximation of the teacher representation, effectively projecting teacher features onto a subspace consistent with the student's bitrate budget.

VNLD requires one additional frozen-teacher encoder forward pass only during training and introduces no learnable parameters or additional inference-time computation, latency, or decoding complexity.

4 Experiments

We first evaluate the proposed self-teacher distillation framework in a standard LIC setting where both encoder and decoder are optimized jointly. This setup allows us to study the general behavior of the method without any deployment constraints.

4.1 Teacher–Student Construction

For each backbone, we construct a self-teacher distillation setup across multiple RD operating points. A single high-rate model trained with a large Lagrangian multiplier ($\lambda = 2$) serves as the teacher. The teacher network is trained independently and kept frozen while supervising student models trained at lower-rate.

The overall objective follows the classical rate–distortion formulation augmented with a distillation term, as defined in Eq. (2), where λ controls the rate–distortion trade-off and β balances the contribution of the distillation loss. The term $\mathcal{L}_{KD}$ corresponds to one of three alternatives: latent MSE alignment, contrastive representation distillation (CRD), or the proposed variance-normalized latent distillation (VNLD). Unless otherwise stated, we set $\beta = 10^{-3}$, which yields stable training behavior.

We analyzed the sensitivity to the distillation weight β, which balances the rate–distortion objective and the distillation term. Very small values ($\beta < 5 \times 10^{-4}$) make the distillation signal negligible, resulting in performance close to RD-only training. Conversely, large values ($\beta > 5 \times 10^{-3}$) over-emphasize latent alignment and degrade reconstruction quality, as the model sacrifices rate control to match teacher features. A value of $\beta = 10^{-3}$ provides a stable trade-off and is used in all experiments.

We additionally evaluated alternative cross-rate supervision configurations using the BMS network [5], introduced in more detail in Sec. 4.2. These include hierarchical supervision between adjacent rate levels, where each student at λ_i is supervised by the next higher-rate model λ_{i+1}, and a multi-target variant where the highest-rate model available in the pretrained operating points supervises all lower-rate students, i.e., $\lambda_6 \rightarrow [\lambda_1, \lambda_5]$. Both strategies produced only marginal

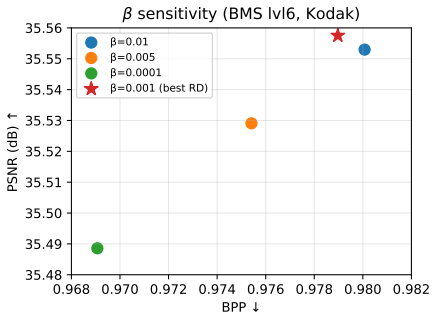

Fig. 3: β sensitivity on BMS lvl6/Kodak.

improvements and were consistently weaker than our final configuration, where a single substantially higher-rate teacher ($\lambda_T = 2$) supervises the lower-rate students. Table 1 reports this teacher-rate ablation on the lowest-rate BMS student and shows that nearby-rate teachers provide weaker supervision, while $\lambda_T = 2$ achieves the lowest student RD objective.

Table 1: Teacher-rate ablation on BMS [5] at $\lambda_S = 0.0018$.

λ_T	PSNR ↑	BPP ↓	RD loss ↓
0.0035	26.633	0.1153	0.3692
0.0130	26.641	0.1154	0.3691
0.0483	26.631	0.1154	0.3695
2	**26.880**	0.1213	**0.3613**

4.2 Experimental Setup

We evaluate two learned image compression architectures: HPCM [24], a recent high-capacity codec with hierarchical priors and modern transform networks, and BMS [5], a classical hyperprior-based baseline. Both architectures follow the standard neural transform coding pipeline, consisting of an analysis transform, quantization, entropy coding, and a synthesis transform. All components are optimized jointly during training.

All models are implemented in PyTorch and optimized using Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Training uses a batch size of 32 with 24 data-loading workers. Images are randomly cropped to 256×256 patches and augmented with random horizontal and vertical flips.

HPCM is trained on the Flickr dataset [31], while BMS is trained on COCO-Stuff [9]. Models are trained for 50 epochs with an initial learning rate of 10^{-5}, reduced on plateau (patience 8) down to 10^{-6}. The operating points follow $\lambda \in \{0.0018, 0.0035, 0.0067, 0.0130, 0.025, 0.0483\}$.

5 Results

We evaluate the proposed self-teacher distillation framework in a standard learned image compression setting where both encoder and decoder are optimized. For each network (HPCM [24], BMS-factorized codec [5]) we compare three distillation objectives: standard latent MSE alignment, contrastive representation distillation (CRD), and the proposed variance-normalized latent distillation (VNLD).

Evaluation is performed on three widely used compression benchmarks: Kodak [14], Tecnick [2], and the JPEG AI test set [19]. Performance is measured using BD-rate [7] with respect to PSNR, MS-SSIM, and LPIPSS [38]. Unless otherwise stated, results are reported relative to the Pretrained (RD-only) baseline of the same backbone in order to isolate the effect of the distillation objectives.

5.1 Quantitative Results

Table 3 reports BD-rate results for HPCM. Self-teacher distillation consistently improves RD performance over RD-only training. While both MSE and CRD provide improvements, VNLD yields the most consistent bitrate reductions across datasets and metrics.

Table 2: BD-rate reduction (%) compared with the Pretrained (RD-only) HPCM baseline.

Table 3: BD-rate (%) of self-teacher distillation variants with HPCM, reported relative to the pretrained RD-only HPCM baseline on Kodak, Tecnick, and JPEG AI. MSE and CRD denote latent MSE alignment and contrastive representation distillation, respectively, while VNLD denotes the proposed variance-normalized latent distillation.

	Kodak			Tecnick			JPEG AI		
	PSNR	MS-SSIM	LPIPS	PSNR	MS-SSIM	LPIPS	PSNR	MS-SSIM	LPIPS
HPCM (Pre)	0	0	0	0	0	0	0	0	0
+ MSE	-1.01	-0.74	-0.82	-0.91	-0.43	-0.57	-1.18	-1.22	-1.02
+ CRD	-0.32	-0.89	-1.05	-0.55	-0.71	-0.91	-1.19	-0.87	-1.32
+ VNLD	**-1.05**	**-0.94**	**-1.52**	**-0.65**	**-1.44**	**-1.80**	**-1.65**	**-1.24**	**-1.48**

Fig. 4 illustrates the corresponding RD curves on Kodak. Across all operating points, VNLD consistently shifts the RD curve toward lower bitrates, confirming that variance-aware supervision provides a more effective transfer signal than standard latent alignment.

To isolate the contribution of variance normalization, we compare standard latent MSE alignment with the full VNLD formulation. Table 4 shows that the proposed VNLD consistently improves BD-rate on average over all evaluation datasets. This confirms that the gains stem specifically from modulating channel-wise supervision rather than simply adding an auxiliary distillation term.

To further evaluate the generality of the proposed method beyond HPCM, we additionally test the self-teacher distillation framework on the BMS-factorized

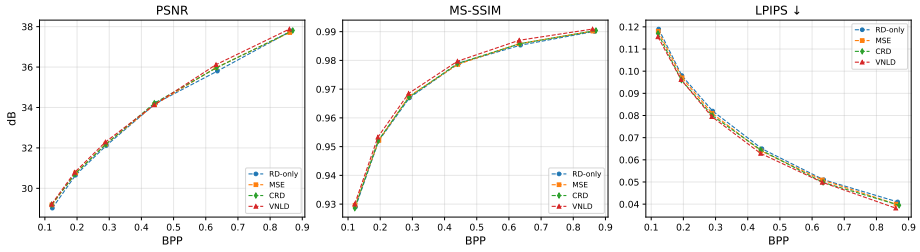

Fig. 4: Rate–distortion curves on Kodak for HPCM. Pretrained (RD-only) denotes the baseline, while MSE, CRD, and VNLD correspond to self-teacher variants.

Table 4: Effect of variance normalization in VNLD. Results are BD-rate (%) relative to the RD-only HPCM baseline.

	HPCM		
	PSNR	MS-SSIM	LPIPS
VNLD w/o variance	-1.03	-0.79	-0.70
VNLD (proposed)	**-1.12**	**-1.20**	**-1.60**

architecture [5], a classical hyperprior-based LIC model. As shown in Table 5, VNLD also achieves the best BD-rate reduction on BMS. Even for this lower complexity architecture, self-teacher distillation in particular with VNLD provides consistent bitrate reductions relative to Pretrained (RD-only) training.

Table 5: BD-rate (%) on the Kodak dataset for BMS relative to the Pretrained (RD-only) baseline. Negative values indicate bitrate savings.

	PSNR	MS-SSIM
+ MSE	-1.5	0.4
+ CRD	-1.6	-0.2
+ VNLD	**-1.7**	**-0.5**

Although the absolute BD-rate gains (1–2%) are numerically modest, such improvements are meaningful in competitive compression benchmarks where progress is typically incremental [17]. These gains are also obtained without increasing model complexity or modifying the decoder at inference time, as the teacher is trained only once and reused to supervise lower-rate models.

Across all tested backbones, self-teacher supervision improves RD performance relative to Pretrained (RD-only) training. This result confirms that the proposed distillation strategy is not tied to a particular model design. VNLD further provides the most consistent gains among the evaluated objectives, highlighting the importance of rate-aware variance normalization for cross-rate latent supervision.

5.2 Visual Results

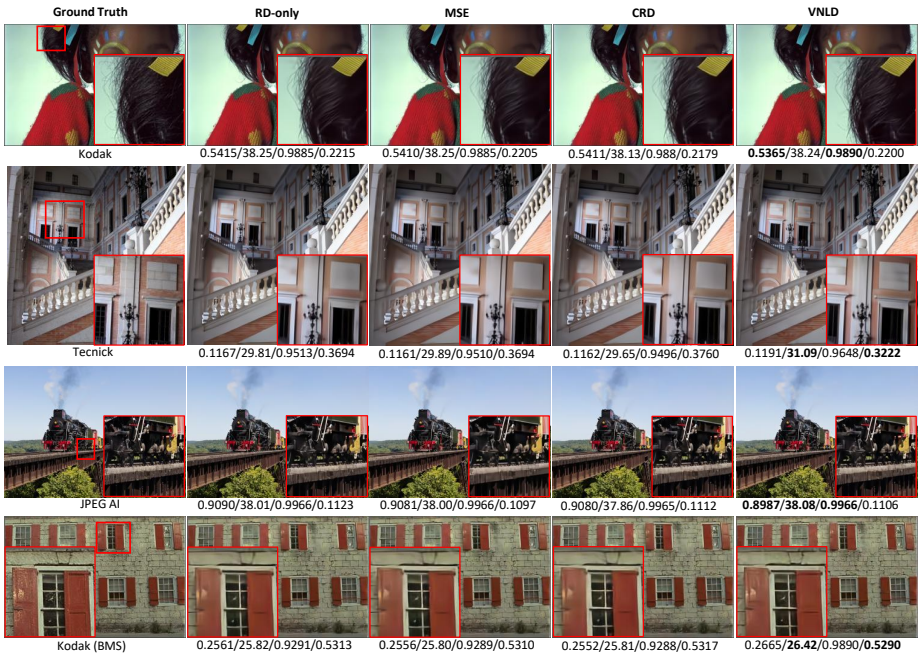

Fig. 5: Visual comparison of reconstructions. The first rows (1 to 3) correspond to HPCM, while the last row shows BMS on Kodak. Metrics: bpp / PSNR / MS-SSIM / LPIPS.

Fig. 5 presents qualitative comparisons between pretrained (RD-only) training, MSE distillation, CRD, and the proposed VNLD using the HPCM model across three different datasets. The last row additionally shows results for the BMS model on the Kodak dataset. For each image, we report *bpp / PSNR / MS-SSIM / LPIPS*.

RD-only model tends to produce over-smoothed reconstructions with loss of fine details. MSE and CRD partially improve structural consistency but remain limited in preserving high-frequency textures under low bitrate constraints. In contrast, VNLD yields sharper reconstructions with better texture preservation and clearer edges, resulting in improved perceptual quality.

6 JPEG AI Deployment Study

The emerging JPEG AI standard [18] aims to support learned image compression systems under practical deployment constraints, where encoder updates

may be possible while the decoder remains fixed in deployed devices. Such constraints motivate training strategies that improve compression performance without modifying the decoding pipeline.

To evaluate the proposed method under this deployment paradigm, we adopt **SegPIC** [26], a recent semantic-aware learned image compression architecture. SegPIC integrates Region-Adaptive Transform (RAT) and Scale Affine Layers (SAL) to exploit segmentation priors, together with attention mechanisms and autoregressive entropy modeling.

We select this model because it represents a recent generation of LIC architectures and enables us to evaluate the proposed approach across a wider range of modern compression models rather than limiting the study to earlier designs.

6.1 Deployment Evaluation with Frozen Decoder

To reflect realistic JPEG AI deployment constraints, the decoder and entropy model are kept fixed, and only the encoder is optimized for both teacher and student models. Table 6 reports BD-rate results for SegPIC under the frozen-decoder setup.

Table 6: BD-rate (%) of self-teacher distillation with SegPIC under frozen-decoder JPEG AI constraints, reported relative to the RD-only SegPIC baseline. The decoder and entropy model are kept fixed, and only the encoder is optimized.

	Kodak			CLIC Pro			JPEG AI		
	PSNR	MS-SSIM	LPIPS	PSNR	MS-SSIM	LPIPS	PSNR	MS-SSIM	LPIPS
SegPIC (RD only)	0	0	0	0	0	0	0	0	0
+ MSE	-0.32	-0.58	-0.61	-0.80	-0.30	-0.40	-1.09	**-2.82**	-3.41
+ CRD	+0.41	+0.22	+0.84	+4.90	+3.10	+2.70	**-1.86**	-2.51	-3.77
+ VNLD	**-0.47**	**-2.14**	**-0.74**	**-1.60**	**-1.20**	**-1.90**	-1.14	-2.63	**-4.25**

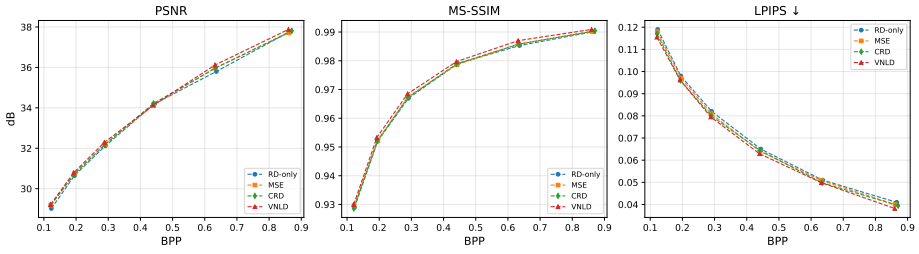

Fig. 6: Rate–distortion curves on Kodak for SegPIC with frozen decoder. RD-only denotes the baseline, while MSE, CRD, and VNLD correspond to self-teacher variants.

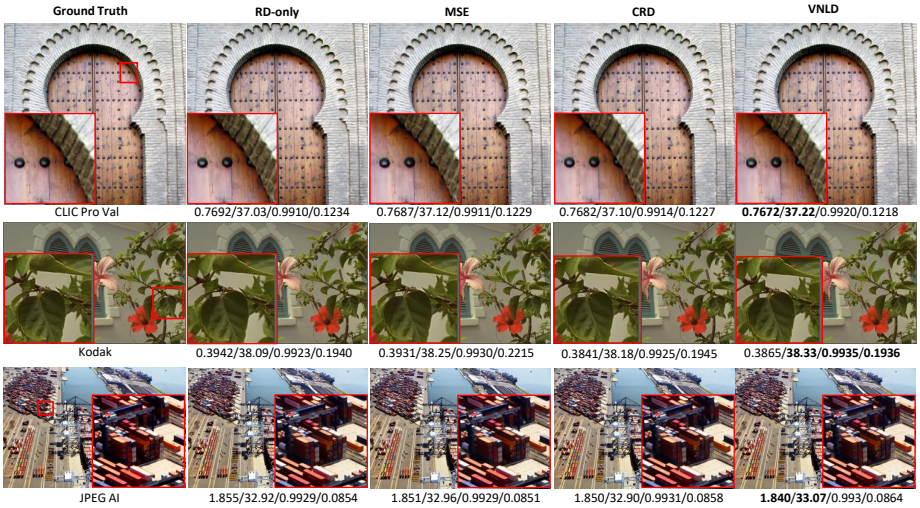

Fig. 7: Visual comparison of reconstructed images using the SegPIC model. For each example, methods are ordered from left to right as: Ground Truth, RD-only training, MSE distillation, CRD, and the proposed VNLD. Metrics are reported as bpp / PSNR / MS-SSIM / LPIPS.

Even in this constrained setting, distillation improves compression performance, with VNLD consistently achieving the strongest gains. Fig. 6 shows the corresponding rate–distortion curves on Kodak.

Fig. 7 presents qualitative comparisons between Pretrained (RD-only) training, MSE distillation, CRD, and the proposed VNLD of SegPIC model. For each image, we report *bpp / PSNR / MS-SSIM / LPIPS*. Visually, VNLD tends to better preserve fine structures and textures while reducing reconstruction artifacts compared to the other distillation strategies, which is consistent with the improvements observed in the quantitative evaluation.

6.2 Smartphone Image Specialization under JPEG AI Constraints

To further evaluate the behavior of rate-aware distillation in realistic deployment scenarios, we consider domain-specific specialization reflecting common smartphone image categories: *Selfie* [21], *Food101* [8], and *Landscape* [35]. To do that, we revisit our previous mobile-oriented LIC works [12, 13] by replacing standard encoder-side RD fine-tuning with the proposed self-distillation setup under fixed-decoder JPEG AI constraints.

We start from a pretrained SegPIC [26] model (Pre) trained on a large general-purpose dataset. This model serves as the baseline representing a generic compression model without domain specialization. For each image category (images are divided into 80% training, 10% validation, and 10% testing subsets to ensure consistent evaluation across datasets), we then fine-tune the encoder on

Table 7: BD-rate (%) of VNLD with SegPIC for in-domain specialization under fixed-decoder constraints, reported relative to the RD-only fine-tuned SegPIC model and the pretrained SegPIC baseline. Negative values indicate bitrate savings.

Dataset	PSNR		MS-SSIM	
	VNLD/RD-only	VNLD/Pre	VNLD/RD-only	VNLD/Pre
Selfie	-4.21	**-6.77**	-4.56	**-7.45**
Food101	-2.29	-6.16	-3.15	-7.17
Landscape	-2.61	-3.94	-3.24	-4.64
Kodak	+0.5	+0.71	+0.29	+0.56

the corresponding image set using either (i) the standard rate–distortion objective (RD-only) or (ii) the proposed VNLD distillation objective, while keeping the decoder and entropy model frozen to respect JPEG AI deployment constraints.

Table 7 reports BD-rate results relative to both the RD-only fine-tuned model and the pretrained baseline. VNLD consistently improves compression performance compared to both models across all categories.

On the Selfie dataset, VNLD achieves a **-6.77%** BD-rate gain in PSNR and **-7.45%** in MS-SSIM compared to the pretrained model, and also improves over the RD-only fine-tuned model. These results indicate that cross-rate supervision remains effective even under strong domain specialization.

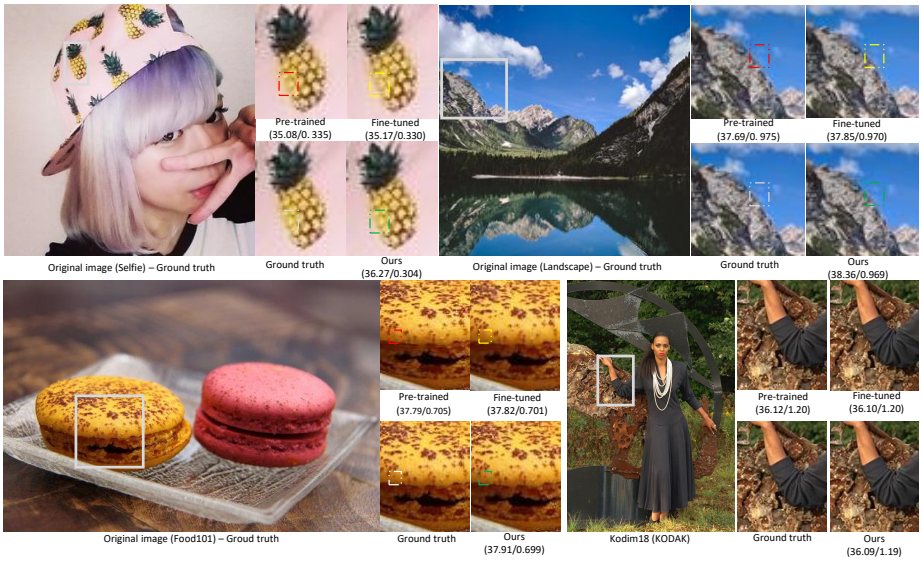

Fig. 8: Qualitative comparison of reconstructions on Selfie [21], Food101 [8], Landscape [35], and Kodak [14]. The leftmost images correspond to the uncompressed ground-truth images. Metrics are PSNR↑/bpp↓. VNLD preserves finer structures and textures compared to standard fine-tuning and the pretrained baseline.

We present visual comparisons in Figure 8 using SegPIC [26]. The reconstructed images highlight the differences between three models: (i) the pretrained SegPIC baseline trained on general data, (ii) the standard RD-only and (iii) the proposed VNLD model, both fine-tuned models specialized for the category dataset.

Overall, these results indicate that rate-aware VNLD-based latent distillation enables effective adaptation to domain-specific image distributions under JPEG AI deployment constraints. At the same time, the method remains robust across diverse image statistics, as demonstrated by consistent behavior on standard benchmark datasets such as Kodak, with only a marginal BD-rate increase, indicating limited performance degradation.

7 Conclusion

This work reframes knowledge distillation in learned image compression as latent-space supervision driven by bitrate asymmetry rather than model capacity. By treating a high-rate instance of a codec as a self-teacher for lower-rate students of identical architecture, it shows that the key role of supervision is not to increase capacity, but to shape how a fixed-capacity latent space allocates its limited bits.

Variance-normalized latent distillation (VNLD) operationalizes this idea by reweighting channel-wise alignment according to the teacher's latent variance, suppressing high-variance components that are infeasible at low rates and emphasizing stable, transferable structure. Experiments across multiple backbones, datasets, and distillation objectives demonstrate that this rate-aware supervision consistently improves RD performance over RD-only training and standard latent matching, including in encoder-only settings with frozen decoders and entropy models.

Evaluation results indicate that self-teacher distillation is a robust and general paradigm for LIC, complementary to architectural advances and compatible with deployment constraints such as JPEG AI–style decoding pipelines. More broadly, the findings suggest that future compression methods should treat distillation not merely as capacity transfer, but as a principled tool for rate-aware representation shaping across operating points.

References

1. Ascenso, J., Alshina, E., Ebrahimi, T.: The jpeg ai standard: Providing efficient human and machine visual data consumption. In: IEEE MultiMedia. vol. 30, pp. 9–17 (2023)
2. Asuni, N., Giachetti, A.: Testimages: A large-scale archive for testing visual devices and image processing algorithms. arXiv preprint arXiv:1406.6987 (2014)
3. Ballé, J., Laparra, V., Simoncelli, E.: End-to-end optimized image compression. In: ICLR (2017)
4. Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: Variational image compression with a scale hyperprior. In: ICLR (2018)

5. Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: Variational image compression with a scale hyperprior (2018)
6. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
7. Bjontegaard, G.: Calculation of average psnr differences between rd-curves. ITU SG16 Doc. VCEG-M33 (2001)
8. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: Proc. Eur. Conf. Comput. Vis. (ECCV). pp. 446–461 (2014)
9. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1209–1218 (2018)
10. Chen, Y., Lyu, Z., He, B., Cao, N., Chen, G., Lu, G., Zhang, W.: Knowledge distillation for learned image compression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
11. El Mennaoui, A., Dugelay, J.L., Hemrit, G.: Noise-aware encoder training and specialized denoiser for efficient jpeg ai compression and denoising. Journal of Physics: Conference Series **3128**(1), 012021 (2025). https://doi.org/10.1088/1742-6596/3128/1/012021
12. El Mennaoui, A., Hemrit, G., Dugelay, J.L.: Category-dependent learned image compression for smartphone photography with standard-compliant decoders. In: Proceedings of the 2025 IEEE International Conference on Image Processing (ICIP) (2025). https://doi.org/10.1109/ICIP55913.2025.11084461
13. El Mennaoui, A., Hemrit, G., Dugelay, J.L.: Optimized image compression for mobile photography. In: Proceedings of the 2025 Data Compression Conference (DCC). p. 364 (2025). https://doi.org/10.1109/DCC62719.2025.00052
14. Franzen, R.: Kodak lossless true color image suite. source: http://r0k.us/graphics/kodak **4**(2), 9 (1999)
15. Fu, H., Liang, F., Liang, J., Wang, Y., Fang, Z., Zhang, G., Han, J.: Fast and high-performance learned image compression with improved checkerboard context model, deformable residual module, and knowledge distillation. IEEE Transactions on Image Processing **33**, 4702–4715 (2024)
16. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop (2015)
17. Hu, Y., Yang, W., Ma, Z., Liu, J.: Learning end-to-end lossy image compression: A benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(8), 4194–4211 (2021)
18. Joint Photographic Experts Group (JPEG): JPEG AI—learning-based image coding standardization. https://jpeg.org/jpegai/ (2021)
19. Joint Photographic Experts Group (JPEG): JPEG AI Dataset. Online. Available: https://jpeg.org/jpegai/dataset.html (2022), accessed: April 2024
20. Joint Video Experts Team (JVET): Versatile video coding. Online. Available: https://jvet.hhi.fraunhofer.de/ (April 2021), accessed: June 2024
21. Kalayeh, M.M., Seifu, M., LaLanne, W., Shah, M.: How to take a good selfie? In: Proc. ACM Multimedia Conf. pp. 923–926 (2015)
22. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
23. Li, H., Li, S., Dai, W., Li, C., Zou, J., Xiong, H.: Frequency-aware transformer for learned image compression. In: The Twelfth International Conference on Learning Representations (ICLR) (2024)

24. Li, Y., Zhang, H., Li, L., Liu, D.: Learned image compression with hierarchical progressive context modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18834–18843 (2025)
25. Liu, J., Sun, H., Katto, J.: Learned image compression with mixed transformer-cnn architectures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14388–14397 (June 2023)
26. Liu, Y., Yang, W., Bai, H., Wei, Y., Zhao, Y.: Region-adaptive transform with segmentation prior for image compression. In: Proc. Eur. Conf. Comput. Vis. (ECCV). pp. 181–197 (2024)
27. Mahalanobis, P.C.: On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India **2**(1), 49–55 (1936)
28. Minnen, D., Ballé, J., Toderici, G.D.: Joint autoregressive and hierarchical priors for learned image compression. In: Proc. Adv. Neural Inf. Process. Syst. (NeurIPS). vol. 31, pp. 10771–10780 (2018)
29. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
30. Pennebaker, W.B., Mitchell, J.L.: JPEG Still Image Data Compression Standard. Springer-Verlag, New York, NY, USA (1992)
31. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV). pp. 2641–2649 (2015)
32. Romero, A., Ballas, N., Kahou, S., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: ICLR (2015)
33. Schäfer, M., Pientka, S., Pfaff, J., Schwarz, H., Marpe, D., Wiegand, T.: Rate-distortion optimized encoding for deep image compression. IEEE Open Journal of Circuits and Systems **2**, 633–647 (2021). https://doi.org/10.1109/OJCAS.2021.3124995
34. Taubman, D.S., Marcellin, M.W., Rabbani, M.: JPEG2000 Image Compression Fundamentals, Standards and Practice. Kluwer Academic, Boston, MA, USA (2002)
35. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
36. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: ICLR (2020)
37. Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K.: Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In: ICCV (2019)
38. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)