Variance-Normalized Latent Distillation (VNLD) for Domain-Specific Learned Image Compression under JPEG AI Constraints

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

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

Abstract. We introduce *variance-normalized latent distillation (VNLD)*, a new loss function for learned image compression. VNLD transfers rich latent representations from a teacher to a student through channel-wise variance-normalized alignment, which amplifies informative channels and stabilizes training, while keeping the decoder frozen to comply with JPEG AI decoding requirements. The teacher is obtained by fine-tuning SegPIC, a state-of-the-art compression model, from its best pretrained checkpoint on one image category with high distortion weight (not rate–distortion (RD) constrained), and the student is initialized from the same checkpoint and optimized with the combined RD and VNLD losses. Experiments on the three dominant domains of smartphone photography (selfies, food, landscapes) show that VNLD achieves up to **6.77% BD-rate savings** over the pretrained baseline and outperforms standard fine-tuning, while preserving performance on general benchmarks (Kodak, JPEG AI) and remaining robust on unseen datasets within the same category. These results position VNLD as a candidate loss for future JPEG AI-compliant learned image compression.

Keywords: Learned Image Compression · Latent Space Distillation · JPEG AI

1 Introduction

The rapid proliferation of smartphones equipped with high-resolution cameras has led to an unprecedented surge in the number of images captured and shared daily. Efficient image compression has therefore become essential to reduce storage requirements and transmission bandwidth, especially for mobile devices with limited computational resources. Among user-generated content, a few categories such as selfies, food, and landscapes dominate social media platforms [26], making them particularly important targets for optimized compression. Preserving perceptual quality in these categories is critical to enhance user experience while reducing the overall storage and communication burden.

Traditional codecs such as JPEG [1], JPEG 2000 [4], VVC [20], and BPG [5] have been widely deployed but are reaching their performance limits. In contrast, learned image compression (LIC) has recently emerged as a promising alternative, leveraging deep neural networks to achieve superior rate–distortion performance [3, 6]. SegPIC [15] exemplifies this new generation of models by integrating semantic priors into the compression process, providing strong efficiency while remaining compatible with the JPEG AI decoding pipeline [23]. We adopt SegPIC as the backbone for our experiments due to its state-of-the-art compression performance and its ability to incorporate semantic priors, making it particularly suitable for our category-specific training and distillation framework.

Despite these advances, most LIC models are trained on general-purpose datasets such as COCO [24] or ImageNet [27], which do not fully capture the statistical regularities of smartphone images. This domain gap limits compression efficiency for user-centric applications. In our previous work [25, 22], we addressed this issue by fine-tuning SegPIC on category-specific datasets (selfies, food, landscapes), while keeping the decoder fixed to comply with JPEG AI standardization [16]. While this fine-tuning improved compression efficiency, it relied only on the rate–distortion loss, ignoring the richer semantic information encoded in the latent space.

In this paper, we introduce *variance-normalized latent distillation (VNLD)*, a novel latent-space distillation loss to overcome this limitation. The proposed approach adopts a teacher–student design: the teacher is obtained by fine-tuning the best pretrained SegPIC with a high distortion weight (high-λ) on the target category (e.g., Selfie), ensuring that its latent space retains rich semantic and structural information, while the student is initialized from the pretrained SegPIC checkpoints and trained with our loss function Eq. (8).

Knowledge transfer occurs directly in the latent space, guided by a channel-wise-variance-normalized alignment loss that emphasizes informative channels and stabilizes training. This design allows the student to inherit domain-specific representations without losing the robustness of the pretrained backbone. Importantly, the decoder remains frozen, in line with the JPEG AI vision where the decoder is standardized and cannot be retrained.

We instantiate our framework on three representative smartphone images domains (selfies, food and landscapes) as case studies. Experiments on these datasets, along with cross-category evaluations [29] and general benchmarks such as Kodak [11] dataset and the JPEG AI test set [14] show that our method achieves up to **6.77% BD-rate savings** in terms of the PSNR over the pretrained model and consistently outperforms standard fine-tuning on the specialized domain, all while preserving JPEG AI compatibility with a standardized decoder.

Contributions. ((i) We propose variance-normalized latent distillation **(VNLD)**, a channel-wise latent alignment loss that transfers knowledge from a high-λ teacher (*not RD constrained*) to a pretrained student, while keeping the decoder frozen to comply with JPEG AI constraints. (ii) We show that VNLD

consistently improves RD trade-offs over standard fine-tuning across three domains (selfies, food, landscapes) and generalizes to unseen data, achieving up to **6.77% BD-rate savings** compared to the pretrained baseline.

2 Background and Related Work

Learning-based image compression (LIC) has shown remarkable progress in recent years, surpassing traditional codecs by jointly optimizing the rate–distortion trade-off with deep neural networks. Ballé *et al.* [3] introduced a variational autoencoder (VAE) framework with an entropy bottleneck, while Minnen *et al.* [6] proposed joint autoregressive and hierarchical priors for improved entropy modeling. More recently, Liu *et al.* [15] presented SegPIC, a segmentation-prior-guided framework that leverages class-agnostic masks through modules such as the Region-Adaptive Transform (RAT) and Scale Affine Layer (SAL), achieving state-of-the-art compression performance.

Most existing LIC models are trained on large-scale, general-purpose datasets such as COCO [24] and ImageNet [27]. While such broad training ensures generalization across diverse content, it often fails to capture the statistical regularities of smartphone images, particularly selfies, food, and landscapes, which dominate user-captured photos [26]. This limitation motivates the development of domain-specific compression strategies that can exploit category-specific redundancies to improve efficiency while preserving perceptual quality.

Any domain-specific solution must also comply with standardization requirements. The JPEG AI initiative [16] is currently defining a new generation of learning-based codecs that achieve state-of-the-art compression while ensuring interoperability across devices and applications. A key constraint of JPEG AI is the use of a fixed decoder, which requires encoder optimizations to remain fully compatible with the standardized decoding pipeline. As a result, evaluation on the JPEG AI test set has become essential to ensure fair and standard-compliant comparisons.

Knowledge distillation (KD). Introduced by Hinton *et al.* [7], KD was originally proposed to transfer knowledge from a teacher to a student using softened outputs. Later works such as FitNets [8] and cross-modal distillation [9] extended this idea to intermediate features, demonstrating that latent representations can act as effective supervisory signals.

The latent space in learned image compression is not only an intermediate feature but the central representation that drives entropy modeling and image reconstruction. Within the JPEG AI framework [16], its role is even more critical, since a single encoder produces a latent representation that is consumed by multiple standardized decoders. Beyond compression, these decoders can support downstream tasks such as classification, detection, segmentation, and image restoration [2]. This highlights the importance of learning latent representations that are both informative and well-structured.

However, despite this central role, knowledge distillation remains relatively underexplored in the context of learned image compression. In particular, lever-

aging KD directly in the latent space is a natural yet underutilized direction, as it enables the transfer of structured, category-specific information between models. Latent-space distillation therefore provides a principled mechanism to guide the encoder toward richer representations while preserving compatibility with a fixed decoder. Since the alignment is performed at the encoder level, the decoder can remain unchanged, ensuring full compliance with the JPEG AI framework while enabling domain-adaptive optimization.

3 Variance-Normalized Latent Distillation (VNLD) Loss

3.1 Loss Formulation

While knowledge distillation (KD) is typically applied at the output level [7], prior works have shown that intermediate features can serve as reliable supervisory signals [8, 9]. We extend this principle to the latent space, which in learned image compression constitutes the central bottleneck for entropy modeling and reconstruction.

We introduce *variance-normalized latent distillation* (VNLD), a distillation strategy designed to transfer domain-specific representations between learned image compression models while preserving decoder compatibility. VNLD operates directly in the latent space at the bottleneck *pre-quantization latent* y (i.e., the output of the analysis transform), before entropy modeling and synthesis.

To ensure bitstream compatibility with the standardized JPEG AI decoder, the synthesis transform (decoder) and the probability model are kept frozen, while only the analysis transform (encoder) is updated. Let $y_t, y_s \in \mathbb{R}^{C\times H'\times W'}$ denote the latent representations produced by the teacher and student encoders, respectively. For each channel, we compute the spatial standard deviation of the teacher features, σ_t, stabilized with a threshold $\epsilon = 0.02$:

$$\sigma_t = \max\big(\mathrm{Std}(y_t),\, \epsilon\big). \tag{1}$$

The distillation term aligns the student with the teacher while adaptively weighting each channel by its variability:

$$\mathcal{L}_{\text{distill}} = \frac{1}{|\Omega|} \sum_{i\in\Omega} \left(\frac{y_s^i - y_t^i}{\sigma_t^i} \right)^2, \tag{2}$$

where Ω denotes the spatial domain. This formulation departs from raw feature matching by normalizing alignment with the teacher's variability: high-variance (often noisy) channels are downweighted, while low-variance, stable channels are emphasized, which guides the student to inherit the most informative latent-space dimensions. The clamping mechanism improves stability.

3.2 Theoretical Justification

We provide theoretical insights into why latent-space distillation is effective even when teacher and student share the same SegPIC architecture.

Rate–Distortion Perspective An encoder trained at Lagrangian multiplier λ minimizes

$$\min_{p(z|x)} \; I(X;Z) + \lambda\,\mathbb{E}[d(X,\hat{X})], \tag{3}$$

where $I(X;Z)$ is the mutual information between source X and latent vector Z. Along the rate–distortion curve,

$$\frac{\mathrm{d}R}{\mathrm{d}D} = -\lambda. \tag{4}$$

Thus, a larger λ yields lower distortion but higher rate, implying

$$I(X;Z_{\lambda_j}) > I(X;Z_{\lambda_i}), \quad \lambda_j > \lambda_i. \tag{5}$$

High-λ encoders preserve richer latent representations, making them natural teachers for guiding students.

Information Bottleneck Perspective In the information bottleneck (IB) view, the objective for a secondary task Y is

$$\max \; I(Z;Y) - \beta\, I(X;Z). \tag{6}$$

A latent representation that retains more of X (larger $I(X;Z)$) can also achieve higher $I(Z;Y)$. A high-λ teacher therefore acts as an upper bound on task-relevant information, and distilling its latent features helps the student approach this bound under a tighter bitrate. This view is consistent with prior knowledge distillation research. While early methods focused on matching output distributions [7], later work showed that aligning intermediate feature representations yields stronger supervision [8, 9]. Distillation in the latent space therefore provides a principled way to transfer structured representations from teacher to student.

Interpretation Together, these perspectives explain why VNLD is effective in learned image compression: the teacher retains high-fidelity latents, while the student inherits them to improve its rate–distortion trade-off without altering the decoder. Unlike conventional KD, VNLD operates directly in the latent space, which in LIC is the central bottleneck that governs entropy modeling and reconstruction.

4 SegPIC backbone

The SegPIC model, proposed by Liu *et al.* [15], is a learned image compression framework that leverages segmentation priors to guide the compression process. It introduces two key modules, the Region-Adaptive Transform (RAT) and the Scale Affine Layer (SAL), which enable region-specific feature transformations

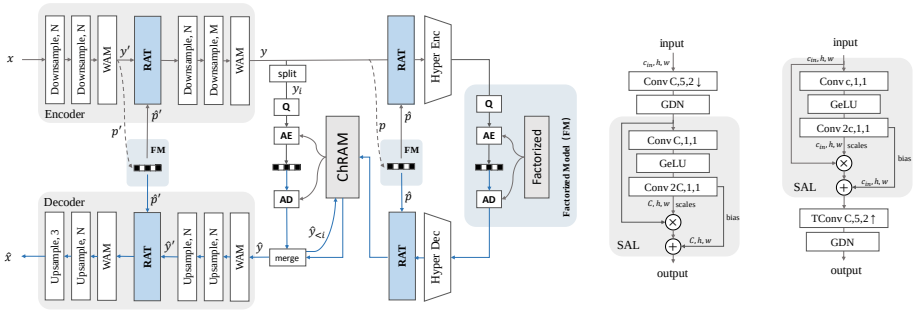

Fig. 1. SegPIC framework from [15].

and enhance semantic representation, making the model particularly suitable for smartphone photography scenarios.

As shown in Figure 1, the framework integrates several components including RAT, SAL, the Window Attention Module (WAM), Channel-wise Auto-Regressive Model (ChARM), Factorized Model (FM), and Generalized Divisive Normalization (GDN). The diagram highlights how RAT and SAL support region-specific compression and feature enhancement, while downsampling and upsampling blocks enable compact and expressive latent-space representations.

5 Training Setup

We conducted experiments on three representative datasets that dominate smartphone photography: the Selfie dataset [18], Food101 [19], and a subset from the Landscape dataset YFCC100M [21]. For all three domains, we used 10,000 images per category with an 80/10/10 split for training, validation, and testing, following the same protocol as in our previous work [25]. This ensured a consistent experimental setup across datasets and enabled fair comparisons of domain-specific specialization.

In all cases, the student was initialized from a generic pretrained SegPIC checkpoint. The teacher, in contrast, was obtained from the best-performing pretrained SegPIC model in terms of reconstruction quality ($\lambda = 0.0483$) and further fine-tuned on the target dataset with a high distortion weight ($\lambda = 2$). Unlike a standard encoder trained for compression, this high-λ teacher is designed to preserve as much semantic and structural information as possible, producing latents that are close to a ground-truth representation of the image. Its role is therefore to act as a supervisory reference in the latent space rather than as a competitive compressor. Both teacher and student shared the same architecture and employed a fixed decoder, ensuring full bitstream compatibility with JPEG AI [23].

It is important to highlight that the teacher operates at a high-rate point (47.3 dB at 2.7 bpp on Selfie), producing latents rich in semantic and struc-

tural information. VNLD aligns the student to these latents while the student is trained under standard rate–distortion constraints.

We adopt the standard rate loss $\mathcal{L}_{\text{rate}}$ and distortion loss $\mathcal{L}_{\text{distortion}}$, where

$$\mathcal{L}_{\text{rate}} = -\log_2 p_{\hat{y}}(\hat{y}), \qquad \mathcal{L}_{\text{distortion}} = \|x - \hat{x}\|_2^2, \tag{7}$$

with $\hat{y}$ the quantized latent representation, $p_{\hat{y}}$ its estimated probability mass function, x the original image, and $\hat{x}$ its reconstruction.

The overall training objective is then defined as

$$\mathcal{L} = \mathcal{L}_{\text{rate}} + \lambda\, \mathcal{L}_{\text{distortion}} + \beta\, \mathcal{L}_{\text{distill}}, \tag{8}$$

with $\lambda \in \{0.0018, 0.0035, 0.0067, 0.0130, 0.025, 0.0483\}$ controlling the RD trade-off. After preliminary experiments, we fixed $\beta = 0.001$, which provided stable convergence and consistent improvements.

All models were trained for 150 epochs with batch size 16, 24 workers, and Adam optimizer (initial learning rate 1×10^{-5}, reduced on plateau with patience of 8 epochs and minimum learning rate of 1×10^{-6}). No data augmentation was applied. To ensure comparability, we repeated this setup for all rate–distortion trade-offs using λ values. The decoder remained frozen in all experiments to comply with JPEG AI requirements [23], ensuring that only the encoder adapts while the bitstream remains compatible with a standardized decoder.

Preliminary experiments showed that $\beta > 0.005$ destabilized training, while $\beta < 0.0005$ made the distillation term negligible. We therefore fixed $\beta = 0.001$, which provided a stable and effective balance with the rate–distortion objective.

6 Results

In this section we compare three different configurations: (i) the pretrained SegPIC baseline [15], (ii) the standard fine-tuned model optimized with the $R + \lambda D$ loss [25, 22], and (iii) our proposed variance-normalized latent distillation (VNLD).

6.1 BD-Rate Results on Category-Specific and Benchmark Datasetss

Figure 2 presents the PSNR vs. BPP rate–distortion curves for the three models across smartphone representative categories: selfies, food, and landscapes. The results confirm that VNLD consistently outperforms standard fine-tuning and surpasses the pretrained baseline in all three domains.

Tables 1 and 2 report BD-rate [28] results in terms of PSNR and MS-SSIM. The first table evaluates performance on category-specific images, whereas the second evaluates performance on the standard benchmark datasets, Kodak and the JPEG AI test set.

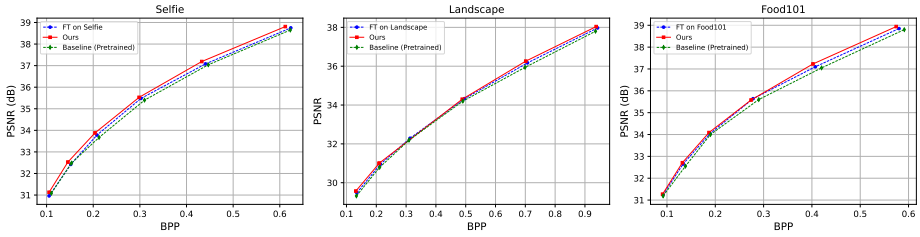

Fig. 2. Rate–distortion performance (PSNR vs. BPP) across three smartphone photography categories: Selfie, Food, and Landscape. VNLD consistently outperforms both the pretrained SegPIC [15] and standard fine-tuning.

Table 1. BD-Rate (%) results for in-domain training (Selfie, Food101, Landscape). Cross-generalization is reported only for the Selfie-trained model, evaluated on Flickr30k Selfie. Negative values indicate bitrate savings. (FT: fine tuned using the RD only loss)

Dataset	**PSNR**		**MS-SSIM**	
	VNLD/FT [25]	**VNLD/Pretrained**	**VNLD/FT [25]**	**VNLD/Pretrained**
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%
Flickr30k Selfie	-2.71%	-6.11%	-2.01%	-7.09%

Table 2. BD-Rate (%) of VNLD compared to the pretrained SegPIC baseline on general benchmark datasets (Kodak and JPEG AI) after specialization on each category.

Training category	**Kodak**		**JPEG AI**	
	PSNR	**MS-SSIM**	**PSNR**	**MS-SSIM**
Selfie	+0.71%	+0.56%	+1.21%	+0.81%
Food101	+1.03%	+0.3%	+0.98%	+1.13%
Landscape	+0.9%	-0.21%	+0.83%	+1.19%

These results highlight that latent-space distillation consistently outperforms standard fine-tuning, with a **-6.77%** BD-rate in terms of PSNR and **-7.45%** BD-rate saving in terms of MS-SSIM over the pretrained model on the selfie dataset. On Kodak and JPEG AI, performance remains comparable, indicating no loss of generalization.

6.2 Qualitative Results

We present visual comparisons in Figure 3. The reconstructed images highlight the differences between three models.

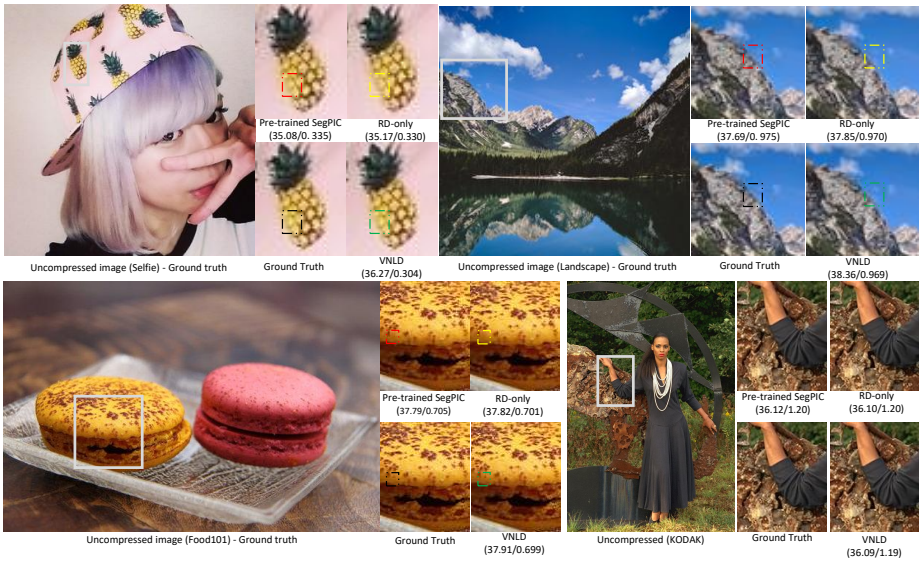

Fig. 3. Qualitative comparison of reconstructions on Selfie, Food101, Landscape, and Kodak. Metrics are PSNR↑/bpp↓. VNLD preserves finer structures and textures compared to standard fine-tuning and the pretrained baseline. All images are from publicly available datasets (Selfie [18], Food101 [19], and Kodak [11]).

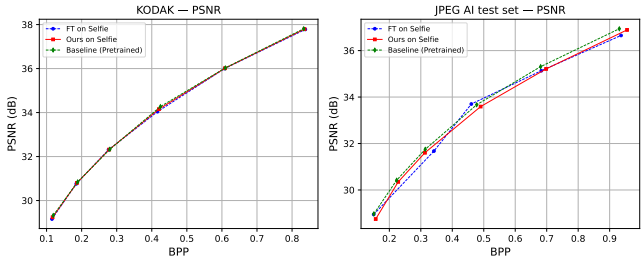

Fig. 4. Rate–distortion curves (PSNR vs. BPP) on the Kodak dataset and the JPEG AI test set. Across both benchmarks, the pretrained baseline [15], standard fine-tuning [22, 25], and our proposed VNLD approach yield nearly identical performance.

Visual inspection confirms the numerical gains: the VNLD model reconstructs sharp edges and subtle textures more faithfully, especially at lower bitrates. Compared to standard fine-tuning, our loss yields clearer structures with fewer artifacts, demonstrating that latent-space alignment with a specialized teacher provides more robust guidance than optimizing the rate–distortion trade-off alone.

The rate–distortion curves on the Kodak dataset and the JPEG AI test set (Fig. 4) show that all models achieve nearly identical performance on these

general benchmarks. On the Kodak dataset, the reconstructed image quality remains comparable across models, with only a small loss of +0.71% BD-rate (PSNR) and +0.56% (MS-SSIM), as illustrated in Fig. 3. This indicates that specialization on a specific category such as selfies does not significantly reduce reconstruction quality on general benchmark datasets.

7 Conclusion and Future Work

We introduced **variance-normalized latent distillation (VNLD)**, a new loss function for domain-specific learned image compression, instantiated on three representative categories of smartphone photography: selfies, food, and landscapes. Unlike conventional fine-tuning with the $R + \lambda D$ loss, our method aligns a pretrained SegPIC student to a teacher obtained by fine-tuning the best pretrained SegPIC on the target dataset with a high distortion weight using VNLD loss, thereby emphasizing informative channels while keeping the decoder fixed to comply with JPEG AI standards. Across all three domains, VNLD consistently outperforms both the pretrained baseline and standard fine-tuning, achieving up to **6.77% BD-rate savings** in-domain, while preserving performance on general datasets such as Kodak and the JPEG AI test set and maintaining cross-category generalization.

Looking ahead, we plan to train VNLD directly on large-scale general datasets and explore its integration into different LIC architectures. By demonstrating both domain-specific gains and benchmark-level robustness, we position VNLD not only as an effective specialization strategy but also as a candidate benchmark loss for future learned image compression research under JPEG AI.

References

1. Pennebaker, W.B., Mitchell, J.L.: *JPEG Still Image Data Compression Standard*. Springer-Verlag, New York, NY (1992)
2. El Mennaoui, A., Dugelay, J.-L., Hemrit, G.: Noise-aware encoder training and specialized denoiser for efficient JPEG AI compression and denoising. In: *Proc. London Imaging Meeting (LIM)* (2025)
3. Ballé, J., Laparra, V., Simoncelli, E.P.: End-to-end optimized image compression. In: *Proc. Int. Conf. Learn. Represent. (ICLR)* (2017)
4. Taubman, D.S., Marcellin, M.W.: JPEG2000 image compression fundamentals, standards and practice. Kluwer Academic, Boston, MA, USA (2002)
5. Bellard, F.: BPG image format. Online: `https://bellard.org/bpg/` (2014), accessed June 2025
6. Minnen, D., Ballé, J., Toderici, G.D.: Joint autoregressive and hierarchical priors for learned image compression. In: *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, pp. 10771–10780 (2018)
7. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: *NIPS Deep Learning and Representation Learning Workshop* (2015)
8. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: FitNets: hints for thin deep nets. In: *Proc. Int. Conf. Learn. Represent. (ICLR)* (2015)

9. Gupta, S., Hoffman, J., Malik, J.: Cross-modal distillation for supervision transfer. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2827–2836 (2016)
10. Mentzer, F., Agustsson, E., Tschannen, M., Timofte, R., Van Gool, L.: Conditional probability models for deep image compression. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4394–4402 (2018)
11. Franzen, R.: Kodak lossless true color image suite. Available online: `http://r0k.us/graphics/kodak` (1999)
12. Rippel, O., Bourdev, L.: Real-time adaptive image compression. In: *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 2922–2930 (2017)
13. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: *Proc. Asilomar Conf. Signals, Systems and Computers*, pp. 1398–1402 (2003)
14. Joint Photographic Experts Group (JPEG): JPEG AI dataset. Available online: `https://jpeg.org/jpegai/dataset.html`
15. 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)
16. Joint Photographic Experts Group (JPEG): JPEG AI—learning-based image coding standardization. Available online: `https://jpeg.org/jpegai/` (2021)
17. Nahai, N.: *Webs of Influence: The Psychology of Online Persuasion*. Pearson Education (2012)
18. Kalayeh, M.M., Seifu, M., LaLanne, W., Shah, M.: How to take a good selfie? In: *Proc. ACM Multimedia Conf.*, pp. 923–926 (2015)
19. 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)
20. Joint Video Experts Team (JVET): Versatile Video Coding (VVC). Online: `https://jvet.hhi.fraunhofer.de/` (2021), accessed June 2024
21. 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. *Commun. ACM* **59**(2), 64–73 (2016)
22. El Mennaoui, A., Hemrit, G., Dugelay, J.-L.: Category-dependent learned image compression for smartphone photography with standard-compliant decoders. In: *Proc. IEEE Int. Conf. Image Process. (ICIP)* (2025)
23. Ascenso, J., Alshina, E., Ebrahimi, T.: The JPEG AI standard: providing efficient human and machine visual data consumption. *IEEE MultiMedia* **30**(1), 9–17 (2023)
24. Caesar, H., Uijlings, J., Ferrari, V.: COCO-stuff: Thing and stuff classes in context. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1209–1218 (2018)
25. El Mennaoui, A., Hemrit, G., Dugelay, J.-L.: Optimized image compression for mobile photography. In: *Proc. Data Compression Conf. (DCC)*, p. 364 (2025)
26. Massip, E., Hidayati, S.C., Cheng, W.-H., Hua, K.-L.: Exploiting category-specific information for image popularity prediction in social media. In: *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, pp. 45–46 (2018)
27. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 248–255 (2009)
28. Bjontegaard, G.: Calculation of average PSNR differences between RD-curves. ITU-T SG16 Doc. VCEG-M33 (2001)

29. 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)