An Empirical Analysis of Hyperprior Side-Information in Direct Latent-Space Image Classification

Anonymous Authors[1]

Abstract

Executing computer vision tasks directly within the compressed latent space of variational autoencoders (VAEs) offers significant computational advantages by bypassing the decompression bottleneck. In this paper, we investigate the semantic utility of hierarchical hyperpriors—traditionally used for spatial entropy estimation—as a side-information gating mechanism for direct latent-space classification. Utilizing a balanced 100,000-image subset of the AGAR microbial dataset, we demonstrate that a baseline Latent-ResNet operating strictly on primary latents achieves a mean Top-1 accuracy of 96.32%, closely trailing a pixel-space EfficientNet-B0 (97.13%). Contrary to theoretical intuition, our proposed Fusion-Gated Hyperprior architecture yields a slight performance degradation (95.57%) alongside increased total system latency. This empirical ablation study suggests that at the specific compression fidelity of Quality Level 3, primary latent representations are semantically saturated for structural classification tasks, rendering hyperprior variance data redundant and mildly noisy. These findings provide bounded system-design parameters for deploying latency-optimized inference pipelines on pre-compressed data arrays.

1. Introduction

Deep convolutional image compression has systematically outperformed traditional linear-transform codecs (e.g., JPEG, BPG) in both rate-distortion efficiency and the preservation of perceptual semantics (Ballé et al., 2017). This is particularly advantageous in data-intensive domains such as high-throughput computational biology, where vast repositories of macro-imagery must be stored under strict constraints. However, standard downstream analysis mandates a "decode-then-classify" workflow. Fully decompressing a VAE-encoded image back into pixel space for standard ResNet or EfficientNet processing introduces a severe computational bottleneck, often requiring hundreds of GFLOPS per image (Choi et al., 2019).

A compelling operational paradigm involves direct latent-space classification, which circumvents the decoder entirely. Direct latent-space evaluation builds upon foundational work demonstrating the viability of feature extraction without decoding (Torfason et al., 2018). Subsequent frameworks have expanded this domain through joint optimization for machine and human perception (Codevilla et al., 2021; Wang et al., 2022), knowledge transfer techniques (Tu et al., 2023), and dynamic feature adaptation (Deng & Karam, 2023). These works establish that the core feature maps generated by compression encoders maintain sufficient spatial and semantic integrity to facilitate accurate classification, drastically reducing inference latency (Liu et al., 2021).

Modern VAE compression architectures, such as the model introduced by Minnen et al. (2018), decompose images into two distinct tensors: a primary latent representation ($\hat{y}$) and a hyperprior representation ($\hat{z}$). The hyperprior captures spatial dependencies to model the entropy of the primary latents, essentially acting as an uncertainty or variance map. Given that high variance in compression often correlates with complex object structures (e.g., microbial colonies against a uniform agar background), we hypothesized that hyperprior data could serve as a powerful spatial attention mask to augment latent classification.

In this work, we conduct a rigorous empirical analysis to determine whether fusing hyperprior side-information improves classification accuracy in the compressed domain. We present an ablation study comparing a single-tensor Latent-ResNet against a dual-tensor Fusion-Gated architecture. Our results provide valuable insight into the information dynamics of learned compression manifolds and offer pragmatic guidance for deploying search-on-compressed-data systems.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Methodology

2.1. Dataset Adaptation and Constraints

We utilize the Annotated Germs for Automated Recognition (AGAR) database (Majchrowska et al., 2021). Originally formatted for object detection, we adapted the dataset for multi-class classification by extracting bounding box crops of individual colonies from the "countable" subset. These encompass five species: *B. subtilis*, *C. albicans*, *E. coli*, *P. aeruginosa*, and *S. aureus*. Crops were padded via zero-value letterboxing to standard 224×224 dimensions to preserve structural morphology. We sampled a balanced dataset of 100,000 images, utilizing a 70/15/15 distribution for training, validation, and testing. To rigorously prevent train/test leakage, the dataset was constructed utilizing a strict plate-disjoint method; the data split was partitioned exclusively at the plate level, ensuring that colonies extracted from the same physical agar plate do not appear across different evaluation splits.

2.2. Representation Extraction Pipeline

We implemented the hierarchical VAE architecture of Minnen et al. (2018), utilizing a pre-trained backbone optimized for Mean Squared Error at Quality Level 3. The objective of this study is to evaluate the inherent semantic density of the existing latents; therefore, the weights of the compression encoder were strictly frozen. A forward pass of the input image x yields the primary quantized latents $\hat{y} \in \mathbb{R}^{C \times H/16 \times W/16}$. A secondary pass through the frozen hyper-autoencoder yields the scale parameters $\hat{\sigma}$ derived from the hyper-latents $\hat{z}$.

2.3. Architectural Configurations

To isolate the utility of the hyperprior, we evaluated three classification paradigms against a standard pixel-space EfficientNet-B0 baseline:

1. **Latent-MLP:** A rudimentary baseline that flattens $\hat{y}$ into a 1D vector processed by a Multi-Layer Perceptron. This tests whether the latent data is linearly separable.

2. **Latent-ResNet (Figure 1):** A spatially-aware baseline. The $\hat{y}$ tensor is ingested by a custom head comprising an initial convolution followed by progressive residual blocks with spatial downsampling.

3. **Fusion-Gated Hyperprior (Figure 2):** Our proposed mechanism for integrating side-information. The hyperprior scale parameters ($\hat{\sigma}$) and the primary latents ($\hat{y}$) are concatenated and passed through a trainable 1×1 convolutional gate. The fused tensor is defined as:

$$f_{fused} = \sigma(W_\sigma * \hat{\sigma}) \odot (W_y * \hat{y}) \tag{1}$$

This operation forces the network to weight the primary features based on the spatial entropy localized by the hyperprior before routing the tensor into the ResNet head.

All architectures were trained for 10 epochs using Adam optimization with fixed hyperparameters to isolate architectural impact.

3. Experimental Results

3.1. Ablation of Classification Capabilities

The empirical data is organized to first establish functional bounds (Table 1) and subsequently isolate the impact of the hyperprior (Table 2). First, the collapse of the Simple MLP ($54.36 \pm 0.00\%$) demonstrates that while semantic data is preserved post-quantization, it is highly non-linear. The representation mandates convolutional architectures to successfully aggregate localized biological features.

Second, our core ablation over five independent trials reveals that the Latent-ResNet resolved this structural mismatch, achieving $96.32 \pm 0.18\%$ accuracy. This establishes a narrow performance degradation relative to the uncompressed pixel-space baseline (97.13%), confirming that direct latent classification is highly viable for complex morphologies.

Finally, contrary to our hypothesis regarding spatial attention, the Fusion-Gated architecture underperformed the single-tensor baseline, yielding a mean accuracy of $95.57 \pm 0.31\%$. The integration of hyperprior side-information induced a measurable drop in predictive capability, indicating that the scale parameters $\hat{\sigma}$ do not inherently possess complementary discriminative utility for this specific task.

3.2. Inference Speed Evaluation

Operating within the latent space incurs a temporal penalty during the initial encoding pass. Consequently, the Total Latency for all compressed-domain models approaches or slightly exceeds the pixel-space EfficientNet. Notably, the Total Latency of the Fusion-Gated model is higher due to the secondary pass required to extract $\hat{\sigma}$ (1.96 ms vs 1.68 ms), while the isolated Head Latency remains equivalent (0.17 ms). If deployed in an environment where images are ingested and natively archived as compressed tensors, the decoding step is eliminated, enabling the system to evaluate pre-compressed archives at nearly $6,000$ images per second.

4. Discussion

4.1. The Semantic Saturation Hypothesis

The primary analytical finding of this work is the failure of the hyperprior gating mechanism to enhance model accu-

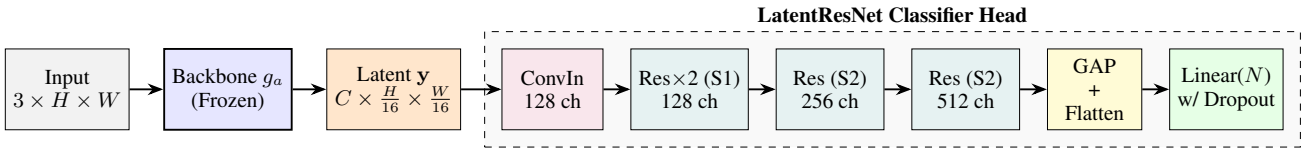

Figure 1. **Latent-ResNet Architecture.** Spatial latents $\mathbf{y}$ are maintained in a 2D tensor and processed via residual blocks with progressive spatial downsampling.

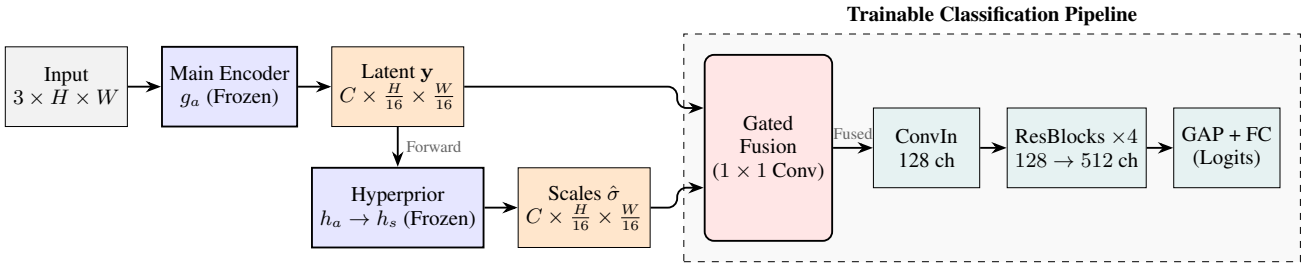

Figure 2. **Fusion-Gated Hyperprior Architecture.** The core latents $\mathbf{y}$ and hyperprior scale parameters $\hat{\sigma}$ are routed through a trainable Gated Fusion module to construct an attention-weighted tensor.

racy. We theorize this is due to a phenomenon of *semantic saturation* at high compression bitrates. At Quality Level 3 (typically corresponding to moderate bitrates ensuring high perceptual fidelity), the primary latent representation $\hat{y}$ is sufficiently expansive to capture the fine-grained textures and structural edges of the microbial colonies without significant information loss.

Because the primary manifold is saturated with requisite data, the introduction of the hyperprior variance map $\hat{\sigma}$ acts as a redundant input stream. Instead of providing useful spatial guidance, the network is forced to learn weights that actively suppress the hyperprior data to avoid overfitting. The reduction in accuracy reflects the noise introduced by this unnecessary complexity.

4.2. Systems Engineering Implications

This negative result provides a highly constructive directive for systems engineers constructing latent-space inference pipelines. It demonstrates that at moderate to high storage fidelities, architectural complexity can be minimized. Systems do not need to expend computational budgets extracting and fusing hyper-latents to maintain high performance. By routing only the primary latents $\hat{y}$ directly into a lightweight ResNet head, pipelines can simultaneously maximize accuracy and minimize overall inference latency and memory bandwidth.

4.3. Limitations and Future Scope

The conclusions drawn regarding hyperprior redundancy are currently constrained strictly to models operating at Quality Level 3. A critical avenue for future research is an exhaustive rate-distortion ablation. As compression ratios increase (Quality Levels 1 and 2), the primary tensor $\hat{y}$ will undergo severe quantization, inevitably losing core textural details. Under those extreme conditions, the semantic saturation hypothesis predicts that the structural outline provided by the hyperprior will transition from a redundant signal to a critical compensatory feature, potentially validating the Fusion-Gated approach at the lower bounds of the bit-rate curve.

5. Conclusion

This study evaluated the efficacy of integrating hyperprior side-information to augment direct latent-space image classification. Using a compressed adaptation of the AGAR microbial dataset, our empirical analysis revealed that a baseline Latent-ResNet operating exclusively on primary latents achieves near-parity with uncompressed models (96.32%). Integrating hyperprior variance data as a spatial attention gate provided no discriminative advantage, yielding slightly lower accuracy (95.57%) and increased overall computational overhead. We conclude that at the tested compression fidelity, primary latent representations are semantically saturated, rendering complex multi-tensor fusion unnecessary. This finding streamlines the design of latency-optimized, compressed-domain analysis systems, dictating a preference for simple, single-tensor extraction architectures.

Impact Statement

This research contributes to the optimization of machine learning infrastructure by interrogating the efficiency of compressed-domain image classification. By demonstrating that complex, multi-tensor hyperprior fusion mechanisms are redundant at high compression fidelities, this work al-

Table 1. Preliminary Baselines: Reference models establishing upper bound (pixel-space) and linear separability lower bounds (primary latent). Latency indicates mean processing time.

MODEL	REPRESENTATION	ACCURACY (%)	TOTAL LATENCY (MS)	HEAD LATENCY (MS)
EFFICIENTNET-B0	PIXEL	97.13 ± 0.00	1.69 ± 0.02	1.69 ± 0.02
SIMPLE MLP	PRIMARY LATENT ($\hat{y}$)	54.36 ± 0.00	1.72 ± 0.02	0.02 ± 0.00

Table 2. Core Findings ($n = 5$ Trials): Ablation of hyperprior side-information. Total Latency indicates total pipeline execution, while Head Latency isolates the duration of the classifier execution, excluding frozen encoding phases.

MODEL	REPRESENTATION	ACCURACY (%)	TOTAL LATENCY (MS)	HEAD LATENCY (MS)
LATENT-RESNET	PRIMARY LATENT ($\hat{y}$)	$\mathbf{96.32} \pm 0.18$	$\mathbf{1.68} \pm 0.01$	0.17 ± 0.01
FUSION-GATED	FUSED ($\hat{y} + \hat{\sigma}$)	95.57 ± 0.31	1.96 ± 0.02	0.17 ± 0.01

lows system architects to design leaner, more computationally efficient inference pipelines. Executing analytical tasks directly on single-tensor pre-compressed data—without the need for full decompression or redundant feature extraction—presents opportunities to significantly reduce computational overhead, lower energy expenditure, and minimize memory bandwidth limits in massive-scale laboratory screening. The methodologies discussed are foundational data-processing techniques and do not interface directly with sensitive demographic data or autonomous individual decision-making systems. Consequently, we anticipate the primary societal impact to be a reduction in the environmental footprint and operational costs associated with high-throughput computer vision tasks.

References

Ballé, J., Laparra, V., and Simoncelli, E. P. End-to-end optimized image compression, 2017. URL https://arxiv.org/abs/1611.01704.

Choi, Y., El-Khamy, M., and Lee, J. Variable rate deep image compression with a conditional autoencoder, 2019. URL https://arxiv.org/abs/1909.04802.

Codevilla, F., Simard, J. G., Goroshin, R., and Pal, C. Learned image compression for machine perception, 2021. URL https://arxiv.org/abs/2111.02249.

Deng, Y. and Karam, L. J. Dnn-compressed domain visual recognition with feature adaptation, 2023. URL https://arxiv.org/abs/2305.08000.

Liu, J., Sun, H., and Katto, J. Learning in compressed domain for faster machine vision tasks. pp. 01–05, 12 2021. doi: 10.1109/VCIP53242.2021.9675369.

Majchrowska, S., Pawłowski, J., Guła, G., Bonus, T., Hanas, A., Loch, A., Pawlak, A., Roszkowiak, J., Golan, T., and Drulis-Kawa, Z. Agar a microbial colony dataset for deep learning detection, 2021.

Minnen, D., Ballé, J., and Toderici, G. Joint autoregressive and hierarchical priors for learned image compression, 2018. URL https://arxiv.org/abs/1809.02736.

Torfason, R., Mentzer, F., Agustsson, E., Tschannen, M., Timofte, R., and Van Gool, L. Towards Image Understanding from Deep Compression without Decoding. *arXiv e-prints*, art. arXiv:1803.06131, March 2018. doi: 10.48550/arXiv.1803.06131.

Tu, H., Li, L., Zhou, W., and Li, H. Learning in compressed domain via knowledge transfer, 2023. URL https://openreview.net/forum?id=AcyZ0Q5p6G8.

Wang, Z., Qin, M., and Chen, Y.-K. Learning from the CNN-based Compressed Domain. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 405. IEEE, January 2022. doi: 10.1109/WACV51458.2022.00405.