



Compressed Feature Quality Assessment: Dataset and Baselines

Changsheng Gao
changsheng.gao@ntu.edu.sg
Nanyang Technological University
Singapore, Singapore

Guosheng Lin
gslin@ntu.edu.sg
Nanyang Technological University
Singapore, Singapore

Wei Zhou
ZhouW26@cardiff.ac.uk
Cardiff University
Cardiff, United Kingdom

Weisi Lin[†]
wslin@ntu.edu.sg
Nanyang Technological University
Singapore, Singapore

Abstract

The widespread deployment of large models in resource-constrained environments has underscored the need for efficient transmission of intermediate feature representations. In this context, feature coding, which compresses features into compact bitstreams, becomes a critical component for scenarios involving feature transmission, storage, and reuse. However, this compression process inevitably introduces semantic degradation that is difficult to quantify with traditional metrics. To address this, we formalize the research problem of Compressed Feature Quality Assessment (CFQA), aiming to evaluate the semantic fidelity of compressed features. To advance CFQA research, we propose the first benchmark dataset, comprising 300 original features and 12000 compressed features derived from three vision tasks and four feature codecs. Task-specific performance degradation is provided as true semantic distortion for evaluating CFQA metrics. We systematically assess three widely used metrics – MSE, cosine similarity, and Centered Kernel Alignment (CKA) – in terms of their ability to capture semantic degradation. Our findings demonstrate the representativeness of the proposed dataset while underscoring the need for more sophisticated metrics capable of measuring semantic distortion in compressed features. This work advances the field by establishing a foundational benchmark and providing a critical resource for the community to explore CFQA. To foster further research, we release the dataset and all associated source code at <https://github.com/chansongao/Compressed-Feature-Quality-Assessment>.

CCS Concepts

• **Information systems** → **Multimedia databases**; • **Computing methodologies** → **Image compression**; • **Human-centered computing** → *Ubiquitous and mobile computing design and evaluation methods*.

Keywords

Compressed Feature Quality Assessment (CFQA), Coding for Machines, Feature Coding

ACM Reference Format:

Changsheng Gao, Wei Zhou, Guosheng Lin, and Weisi Lin[†]. 2025. Compressed Feature Quality Assessment: Dataset and Baselines. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, October 27–31, 2025, Dublin, Ireland. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3746027.3758309>

1 Introduction

Rapid deployment of large foundation models (e.g., DINOv2 [34], LLaMA3 [14]) in distributed and resource-constrained environments has created a growing need to transmit intermediate features rather than raw signals [22]. Feature coding, which compresses these intermediate representations, plays a vital role in enabling scalable, privacy-preserving, and efficient systems. Unlike traditional image or video coding that prioritizes perceptual quality, feature coding targets the preservation of task-relevant semantics embedded in feature representations.

However, feature coding inevitably introduces semantic degradation: a loss of semantic information that may compromise the performance of large models. This degradation is fundamentally different from pixel-level distortions and often cannot be captured by conventional distortion metrics such as MSE or SSIM. Although task accuracy is a more reliable measurement of semantic distortion, it is impractical for practical deployments: downstream tasks may be inaccessible, costly to run, or unavailable in the feature coding process. These limitations highlight an urgent and underexplored problem: **Compressed Feature Quality Assessment (CFQA)** – How can we estimate the semantic distortion of compressed features without relying on downstream inference?

Solving CFQA presents several challenges. First, there is no public dataset that provides compressed features with corresponding task performance across multiple tasks and codecs. Second, existing similarity metrics lack validation for high-dimensional features and often fail to generalize across various tasks. Third, the field lacks a unified evaluation protocol to assess whether a given metric reliably measures semantic distortion.

To bridge this gap, we take the first step towards a systematic study of CFQA. In contrast to previous works that rely on task supervision or task-specific codecs, we treat CFQA as a standalone problem and aim to benchmark its core components: dataset, metrics, and evaluation protocols. In summary, we make the following contributions:

- **Benchmark Dataset:** We construct the first comprehensive CFQA dataset, consisting of 300 original features and 12000



This work is licensed under a Creative Commons Attribution 4.0 International License. *MM '25, Dublin, Ireland*

© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2035-2/2025/10
<https://doi.org/10.1145/3746027.3758309>

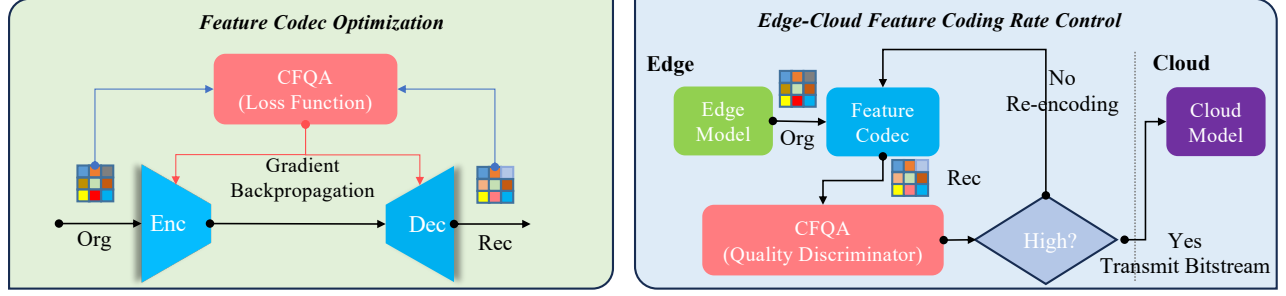


Figure 1: Exemplar application scenarios of compressed feature quality assessment.

compressed features from three vision tasks (image classification (Cls), semantic segmentation (Seg), and depth estimation (Dpt)) and four representative codecs (handcrafted and learning-based). The dataset enables quantitative analysis of semantic distortion across tasks, bitrates, and coding methods.

- **Ground-truth Semantic Distortion:** For each compressed feature, we provide task-specific semantic distortion labels computed by comparing task head outputs using original and compressed features. These labels serve as the ground truth for training and evaluating semantic distortion metrics.
- **Baseline Metric Evaluation:** We evaluate three representative metrics (MSE, cosine similarity, and Centered Kernel Alignment (CKA)) and analyze their sensitivity to feature coding and their ability to measure task-specific semantic distortion. Our analysis reveals their varying strengths and limitations, providing insight into their applicability to semantic distortion measurement.

By establishing a standardized benchmark and evaluation framework, this work bridges the gap between low-level compression and high-level semantic utility. We hope to catalyze the development of lightweight, generalizable, and task-agnostic CFQA metrics that advance the real-world adoption of feature coding. Our dataset and code are publicly available to foster reproducibility and future research.

2 Related Work

2.1 Feature Coding

Feature coding has received growing attention [1, 2, 6–8, 10, 17, 18, 23, 27, 30, 31, 39, 40, 43, 44] in edge-cloud collaborative intelligence scenarios. Recently, large model feature coding has attracted increasing interest [20–22]. These works extend feature coding to scenarios where features are transmitted, stored, and reused.

2.2 Semantic Distortion Measurement

Semantic distortion measurement methods are classified into three main categories: signal fidelity metrics, semantics fidelity metrics, and task-based metrics. Signal fidelity metrics [5, 10, 27, 33, 37] focus on measuring the distortion between original and reconstructed features using traditional metrics like MSE. This approach is directly borrowed from image compression techniques, where the signal similarity is regarded as the compression distortion.

Semantics fidelity metrics [2, 3, 7, 9, 11, 13, 16, 25, 26] assess the preservation of semantic information, where distortion is measured based on the performance drop in specific machine vision tasks. These methods provide a more task-relevant measure of distortion, linking reconstruction quality directly to task performance.

Task-based metrics [1, 19, 24, 29, 31, 36, 38, 40–43] directly measure semantic distortion using the task performance. These methods are specialized for a particular task and lack generalizability.

3 Problem Formulation and Applications

3.1 Problem Formulation

Given a pretrained model $\mathcal{M}$, we denote its extracted intermediate feature from an input $x \in \mathcal{X}$ as $\mathbf{f} = \mathcal{M}(x) \in \mathbb{R}^d$, where d is the feature dimension. A feature codec $\mathcal{C}$ encodes $\mathbf{f}$ into a compact bitstream and decodes it back to $\hat{\mathbf{f}} = \mathcal{C}^{-1}(\mathcal{C}(\mathbf{f}))$. Our objective is to evaluate how much *semantic information* is preserved in $\hat{\mathbf{f}}$ with respect to $\mathbf{f}$, *without accessing downstream task ground-truth or executing inference*.

We define this evaluation task as Compressed Feature Quality Assessment: Given a pair of original and compressed features $(\mathbf{f}, \hat{\mathbf{f}})$, estimate the semantic quality $q \in \mathbb{R}$ of the compressed feature, such that q strongly correlates with its performance on downstream tasks.

Assume a downstream task $\mathcal{T}$ with a head network $h_{\mathcal{T}}$, producing an output $y = h_{\mathcal{T}}(\cdot)$. Let the task-specific performance metric be denoted as $\mathcal{A}_{\mathcal{T}}(\cdot)$. When using compressed features, the task performance becomes: $s = \mathcal{A}_{\mathcal{T}}(h_{\mathcal{T}}(\hat{\mathbf{f}}))$. This value $s \in \mathbb{R}$ reflects the *ground-truth semantic utility* of the compressed feature. However, computing s requires task execution and labels, which may be unavailable or expensive. The goal of CFQA is to estimate a score $q = Q(\mathbf{f}, \hat{\mathbf{f}})$, where $Q(\cdot)$ is a *task-agnostic quality metric*, such that $q \approx s$ in terms of correlation across samples.

3.2 Application Scenarios

We illustrate two typical application scenarios of CFQA in Figure 1. In the first scenario (left of Figure 1), CFQA can be directly integrated into the codec training process as a supervisory signal to align compression objectives with downstream task performance. Recent studies [17, 36] show that task-aware feature coding benefits significantly from semantic-aware supervision. CFQA offers such guidance without requiring end-to-end task labels. As shown on the left of Figure 1, both the original and reconstructed features

Task	Source	Num. of Org. Feat.	Pre-processing	Feature Codecs	Num. of Comp. Feat.	Feature Shape	GT Distortion
Cls	ImageNet	100	Resize	HM, VTM	4000	257×1536	Rank
Seg	VOC 2012	100	Flip and Crop	Multi-task Hyperprior	4000	2×1370×1536	mIoU difference
Dpt	NYUv2	100	Flip	Task-specific Hyperprior	4000	2×4×1611×1536	RMSE difference

Table 1: Abstract information of the proposed dataset for CFQA. (Refer to Sec. 4 for more details.)

are input to the CFQA module, which estimates the semantic distortion. The distortion score is then used to generate gradients for optimizing the encoder and decoder. With a well-designed CFQA metric, the resulting codec learns to preserve semantic information even under low-bitrate constraints.

In the second scenario (right of Figure 1), CFQA is essential in edge-cloud collaborative systems, where features are extracted at the edge and transmitted to the cloud for inference. Since the downstream model resides on the cloud side, semantic distortion cannot be directly measured at the edge. Here, CFQA is used as a proxy to estimate task-relevant semantic distortion. As illustrated on the right of Figure 1, the compressed feature is first evaluated by the CFQA module. If it is judged to be of high quality, the bitstream is transmitted to the cloud. Otherwise, the edge device re-encodes the feature at a higher bitrate to better preserve semantic information. This strategy ensures the reliability of transmitted features while reducing unnecessary bandwidth consumption.

Although our work focuses on semantic distortion due to compression, we emphasize that the value of CFQA extends beyond codec benchmarking: it serves as a crucial component in a variety of systems where features are extracted, compressed, transmitted, cached, or reused.

4 Dataset Construction

4.1 Overview

Our proposed dataset is designed to support the analysis of *semantic distortion* introduced by lossy compression. It includes three tasks, 300 original features, and 12000 compressed features from 4 feature codecs. The features cover diverse image processing methods and multiple feature extraction strategies. The overall information of the dataset is presented in Table 1.

4.2 Model and Task Selection

Since current feature coding research primarily focuses on visual signals, we initiate the study of CFQA with visual features. We adopt DINOv2 [34] as our backbone feature extractor due to its strong generalization capability and widespread adoption in general-purpose vision tasks. To cover a broad range of semantic distortion, we select three widely studied vision tasks: image classification (Cls), semantic segmentation (Seg), and depth estimation (Dpt). These tasks span coarse-to-fine semantic understanding: Cls focuses on image-level category prediction, Seg introduces spatial semantics with class-wise alignment at the pixel level, and Dpt requires detailed geometry prediction.

This multi-task approach is critical for evaluating the generalizability and sensitivity of CFQA metrics, ensuring they remain robust across different semantic requirements.

4.3 Source Data Collection

To ensure semantic diversity, we select 100 representative samples for each task. For Cls, we sample 100 ImageNet [12] images from 100 distinct categories, each correctly predicted by the DINOv2 classifier. For Seg, we sample 100 images from the Pascal VOC 2012 [15] validation set, covering all 20 semantic categories. For Dpt, we sample 100 images from the NYUv2 [32] dataset, spanning all 16 scenes. This sampling strategy balances feature diversity with manageable dataset scale, enabling rigorous yet tractable evaluation.

4.4 Original Feature Collection

For each image, we extract task-specific intermediate features from DINOv2’s designated split points. These split points are chosen based on their semantic richness and alignment with common practice in split computing. For Cls, we resize the image to 224×224 and extract features from the 40th Vision Transformer (ViT) block, which produces features in the shape of 257×1536 (256 patch tokens and 1 class token). For Seg, we flip the original image horizontally and extract features from the same 40th ViT block, resulting in $2 \times 1370 \times 1536$ features. For Dpt, we collect multi-scale features from the 10th, 20th, 30th, and 40th ViT blocks. The original and its horizontally flipped images generate a stacked tensor of shape $2 \times 4 \times 1611 \times 1536$.

The inclusion of diverse image pre-processing techniques and varying split points emulates real-world input variability, enabling rigorous evaluation of CFQA metric generalizability.

4.5 Compressed Feature Collection

To simulate different types and strengths of semantic distortion, we compress original features through four codecs. All original features are flattened into 2D arrays before encoding.

Handcrafted Codecs. We select two handcrafted codecs: **HM Intra coding** (configured with `encoder_intra_main_rext.cfg`) and **VTM Intra coding** (configured with `encoder_intra_vtm.cfg`). Before encoding, we first uniformly quantize the original feature values to $[0, 1023]$. For both codecs, we use the YUV 4:0:0 (monochrome) format for feature coding and set the quantization parameters to $\in \{2, 4, 6, 8, \dots, 16, 18, 20\}$ to simulate various bitrates (measured by Bits Per Feature Point, **BPFP**) and distortion levels.

Learning-Based Codecs. We adopt the Hyperprior model [4] as it represents a milestone in learning-based feature coding. Since most learning-based feature coding methods build upon this architecture, it serves as an ideal testbed for studying the semantic distortion characteristics of learning-based methods.

To investigate how optimization strategies affect compressed feature quality, we implement two distinct variants. **Multi-Task Hyperprior Codec:** The codec is trained on features extracted from

the three tasks. **Task-Specific Hyperprior Codec:** The codec is trained exclusively on features extracted from a single task. All codecs are optimized following the training protocols established in [22]. The rate-distortion trade-off parameters (λ) can be found in our released models.

4.6 Semantic Distortion Collection

To establish a rigorous benchmark for evaluating CFQA metrics, we require ground-truth measurements of semantic distortion. We define the true semantic distortion as the performance degradation in downstream tasks when using compressed features $\hat{\mathbf{f}}$ versus original features $\mathbf{f}$. For **Cls**, we measure the deviation in prediction confidence by computing the *rank* in the softmax function generated from $\hat{\mathbf{f}}$. The original feature $\mathbf{f}$ achieves perfect ranking (rank=1), with higher ranks indicating more severe semantic degradation. For **Seg**, we compute the mIoU difference between the segmentation masks predicted from $\hat{\mathbf{f}}$ and $\mathbf{f}$. For **Dpt**, we compute the RMSE difference between the depth maps predicted from $\hat{\mathbf{f}}$ and $\mathbf{f}$.

These task-specific scores provide quantitative measures of semantic distortion, serving as the foundation for assessing CFQA metric performance.

5 Experiments and Analysis

5.1 Baseline CFQA Metrics

To cover diverse types of distortions and similarity relationships between features, we select three complementary metrics: MSE, cosine similarity, and CKA [28]. These metrics collectively span from local element-wise to global structural comparisons.

- **MSE:** MSE measures feature distortion at the element level. It is widely used in signal processing fields.
- **Cosine similarity:** Cosine similarity measures the angular difference between feature vectors, normalized by their magnitudes. This metric captures directional alignment in the feature space, which is often more robust to scale distortions and is considered to better reflect semantic similarity in high-dimensional representations [35].
- **CKA:** CKA measures similarity between two features using normalized HSIC (Hilbert-Schmidt Independence Criterion). It captures higher-order statistical relationships, making it particularly suitable for comparing architectural differences and global feature patterns.

5.2 Evaluation Protocol

We define an evaluation protocol to assess how well a CFQA metric predicts the semantic distortion of compressed features. PLCC and SROCC are used in our experiments.

- **PLCC (Linearity):** PLCC quantifies how well the predicted quality scores approximate the actual semantic distortion in a linear sense.
- **SROCC (Monotonicity):** SROCC assesses the consistency in the ranking between predicted and true distortions.

Specifically, for each original feature $\mathbf{f}_i$, we compute these two metrics on its 10 predicted quality scores and 10 true semantic distortions. Each metric is evaluated on all three tasks and across all codecs.

5.3 Rate-Accuracy Performance Analysis

Table 2 presents the rate-accuracy performance of the handcrafted codecs. For all three tasks, both codecs exhibit a broad range of bitrates and corresponding accuracy levels, spanning from nearly lossless reconstruction to significant performance degradation. This wide variation highlights their effectiveness in simulating traditional coding distortions. However, the Dpt task exhibits less variation in both bitrate and performance. This difference stems from the use of multi-scale features, which include lower-level layers with higher redundancy and reduced semantic abstraction. These characteristics lead to smaller fluctuations in both bitrate and distortion.

Table 3 shows the rate-accuracy performance of two learning-based codecs. Unlike the handcrafted codecs, the learning-based codecs show more diverse performance patterns across tasks. The multi-task Hyperprior codec achieves better performance than the task-specific Hyperprior codec for Cls, while the task-specific Hyperprior codec performs better in Dpt. These results indicate that incorporating fine-grained task features improves the codec optimization of coarse-grained features, whereas fine-grained features benefit more from training on a single task-specific distribution.

These four codecs comprehensively simulate a wide range of compression distortions, making them not only highly suitable but also essential for supporting research in compressed feature quality assessment.

5.4 CFQA Performance Analysis

5.4.1 Average Performance Analysis. Table 4 presents the average CFQA performance of the three baseline metrics. Overall, the three baseline metrics show higher PLCC and SROCC values for the handcrafted codecs compared to the learning-based codecs. This indicates that the semantic distortions introduced by handcrafted codecs are more stable. The larger fluctuations in semantic distortion observed for learning-based codecs are likely due to the inherent variability in the training process, as these models are data-driven and task-specific. Additionally, since handcrafted codecs are block-based, they tend to exhibit more consistent and predictable distortion patterns.

Among the three baseline metrics, cosine similarity consistently demonstrates a higher degree of linearity and monotonicity in relation to ground-truth semantic distortion. This is due to the metric's ability to capture angular relationships between feature vectors, making it particularly effective for high-dimensional features such as those produced by ViTs.

For handcrafted codecs, the three baseline metrics show better performance in capturing HM-generated distortions compared to VTM. This is likely because VTM employs more complex coding tools, resulting in more intricate and potentially less predictable distortion patterns. In contrast, for learning-based codecs, significant performance variation is observed across the three baseline metrics. This variability is expected, as learning-based encoders produce a wider range of distortion types, making it more challenging for simple metrics to achieve linear fitting.

Among the three tasks, Seg presents the most complex and difficult-to-fit distortions. For most codecs, the baseline metrics show poorer fitting for Seg compared to Cls and Dpt. This reflects

HM						VTM					
Cls		Seg		Dpt		Cls		Seg		Dpt	
BPFP	Acc.	BPFP	mIoU	BPFP	RMSE	BPFP	Acc.	BPFP	mIoU	BPFP	RMSE
32	100	32	83.39	32	0.37	32	100	32	83.39	32	0.37
0.006	0.00	0.002	4.59	0.19	1.73	0.006	1.00	0.004	22.56	0.23	1.59
0.008	0.00	0.004	11.39	0.37	1.40	0.01	1.00	0.01	40.18	0.36	1.36
0.012	2.00	0.009	32.22	0.64	1.27	0.02	10.00	0.02	47.59	0.52	1.25
0.02	6.00	0.02	47.79	1.01	1.00	0.03	13.00	0.04	55.05	0.75	1.23
0.04	18.00	0.05	59.63	1.41	0.81	0.06	26.00	0.09	65.24	1.02	1.06
0.08	25.00	0.11	67.87	1.82	0.63	0.11	44.00	0.17	72.34	1.26	0.92
0.15	55.00	0.21	74.54	2.16	0.55	0.23	81.00	0.31	76.81	1.52	0.83
0.33	84.00	0.41	78.40	2.55	0.47	0.49	92.00	0.56	79.79	1.85	0.73
0.67	94.00	0.73	80.49	2.91	0.43	0.86	97.00	0.91	81.33	2.23	0.59
1.12	97.00	1.13	81.86	3.28	0.40	1.26	98.00	1.28	82.04	2.67	0.52

Table 2: Rate-accuracy performance of handcrafted codecs, with the first row showing original features’ performance.

Multi-Task Hyperprior						Task-Specific Hyperprior					
Cls		Seg		Dpt		Cls		Seg		Dpt	
BPFP	Acc.	BPFP	mIoU	BPFP	RMSE	BPFP	Acc.	BPFP	mIoU	BPFP	RMSE
32	100	32	83.39	32	0.37	32	100	32	83.39	32	0.37
0.01	0.00	0.0001	2.65	0.0003	1.68	0.34	16.00	0.05	50.67	0.12	2.09
0.12	18.00	0.10	58.74	0.10	1.79	0.43	35.00	0.08	60.69	0.18	1.66
0.58	44.00	0.46	76.70	0.52	1.55	0.56	55.00	0.14	65.45	0.30	1.22
0.71	52.00	0.57	77.80	0.64	1.24	0.65	61.00	0.21	71.89	0.42	0.88
1.01	59.00	0.85	79.30	0.91	0.99	1.15	70.00	0.29	74.76	0.57	0.82
1.19	67.00	1.01	80.06	1.07	1.45	1.39	72.00	0.42	78.10	0.70	0.73
1.34	77.00	1.14	80.50	1.20	0.77	1.81	78.00	0.65	79.09	1.06	0.60
1.45	81.00	1.23	80.66	1.29	0.92	1.96	81.00	0.94	79.59	1.30	0.49
1.75	94.00	1.45	81.22	1.52	0.76	2.15	84.00	1.41	81.02	1.39	0.44
2.18	96.00	1.77	81.73	1.84	0.42	2.36	88.00	1.65	82.02	1.48	0.43

Table 3: Rate-accuracy performance of learning-based codecs, with the first row showing original features’ performance.

		MSE		Cosine Similarity		CKA	
Codec	Task	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
HM	Cls	0.6641	0.9113	-0.7368	-0.9116	-0.5788	-0.9089
	Seg	-0.6210	-0.6729	0.6866	0.6729	0.6133	0.6726
	Dpt	0.8601	0.9036	-0.8799	-0.9036	-0.8829	-0.9027
VTM	Cls	0.6220	0.8939	-0.7089	-0.8939	-0.5499	-0.8932
	Seg	-0.4784	-0.5213	0.5131	0.5213	0.4885	0.5213
	Dpt	0.8344	0.8750	-0.8572	-0.8747	-0.8478	-0.8754
Hyperprior (Multi-Task)	Cls	-0.0281	0.4707	-0.8907	-0.7165	-0.3277	-0.6593
	Seg	0.0595	-0.0251	0.3496	0.0604	-0.0059	0.0847
	Dpt	0.6466	0.6675	-0.4752	-0.6528	-0.5876	-0.6004
Hyperprior (Task-Specific)	Cls	-0.0220	0.1303	-0.5486	-0.6084	-0.2622	-0.3365
	Seg	0.3259	0.3235	-0.0552	-0.1669	-0.3138	-0.2187
	Dpt	-0.0787	0.4915	-0.8427	-0.8258	-0.5483	-0.5638

Table 4: Average CFQA performance of the three baseline metrics on Cls, Seg, and Dpt tasks.

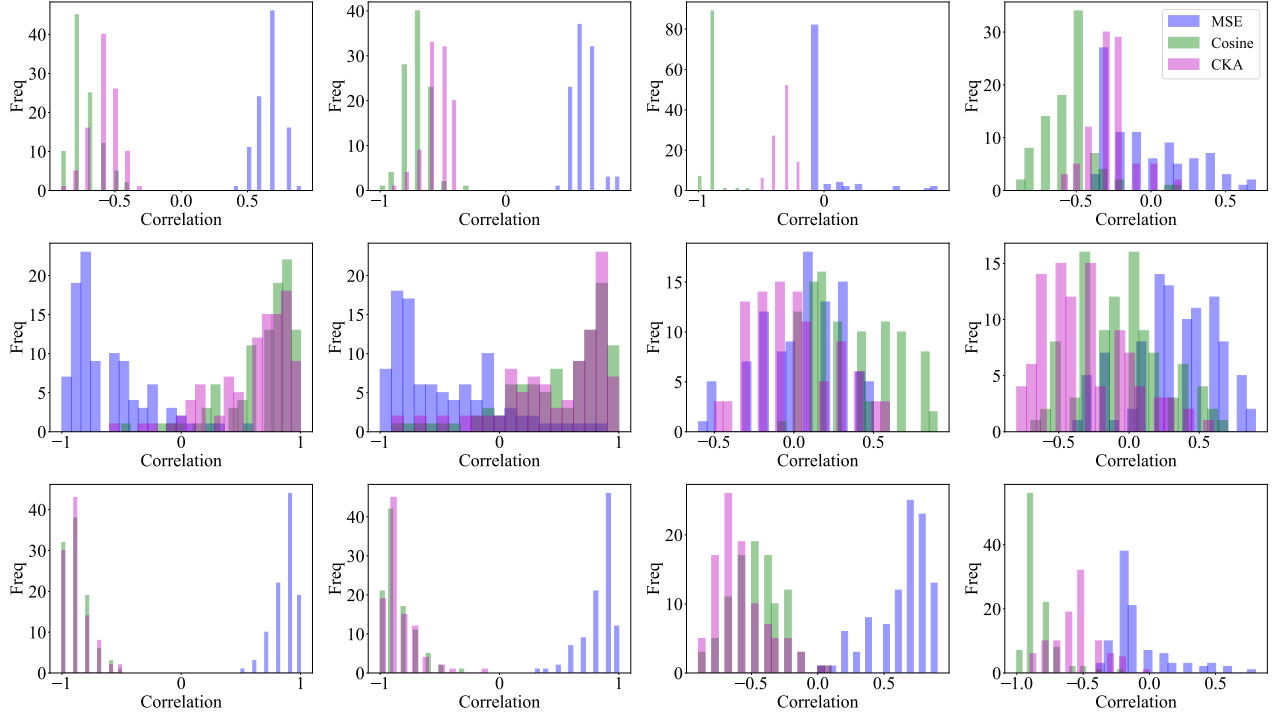


Figure 2: PLCC distribution visualization of MSE, cosine similarity, and CKA. The first, second, and third rows correspond to the Cls, Seg, and Dpt tasks, respectively. The first, second, third, and fourth columns correspond to the HM, VTM, multi-task-trained Hyperprior, and task-specific-trained Hyperprior codes.

the higher semantic complexity in Seg, which is more challenging for these metrics to capture.

5.4.2 Distribution Analysis. Figure 2 visualizes the distribution of PLCC. To provide a clearer view, all PLCC values are rounded to the nearest tenth before computing the frequency histograms.

Overall, the PLCC distributions for handcrafted codecs are more concentrated compared to those of learning-based codecs. This aligns with the fact that handcrafted codecs introduce more consistent and predictable semantic distortion patterns.

Among the three baseline metrics, cosine similarity exhibits the most concentrated PLCC distributions. This observation aligns with its superior average PLCC values, as reported in Table 4, indicating that cosine similarity provides a more stable semantic distortion measurement across varying conditions.

Across all three metrics, Seg shows more dispersed PLCC distributions compared to Cls and Dpt. In some cases, such as with CKA, both positive and negative correlations are observed within the same metric. This highlights the high complexity of distortion patterns in segmentation features. The broader distribution further confirms that existing baseline metrics struggle to model such complex semantic degradation accurately.

5.5 Discussion

Our evaluation reveals three key limitations of current CFQA metrics. First, signal-based metrics like MSE show limited task sensitivity and often correlate poorly with actual semantic degradation.

Second, while cosine similarity and CKA capture structural relationships, they demonstrate instability when handling nonlinear distortions from learned codecs. Most importantly, none of these metrics achieves consistent performance across all tasks and compression methods, indicating that conventional similarity measures alone cannot provide universal CFQA solutions. These findings underscore the necessity for developing more adaptive, task-aware quality estimators – a direction we plan to explore in future work.

6 Conclusion

This paper introduces the concept of Compressed Feature Quality Assessment, a crucial research area for evaluating the semantic distortion of compressed features in systems where features are transmitted, stored, and reused. We present the first benchmark dataset for CFQA, which lays the groundwork for further research in this field. We assess the widely used metrics in CFQA and provide insights for interested researchers. Moving forward, we will focus on developing adaptive CFQA metrics capable of generalizing across diverse tasks and coding strategies.

Acknowledgments

This work was supported by the Ministry of Education of Singapore under Grant T2EP20123-0006. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

- [1] Saeed Ranjbar Alvar and Ivan V. Bajić. 2019. Multi-Task Learning with Compressible Features for Collaborative Intelligence. In *ICIP*. 1705–1709. doi:10.1109/ICIP.2019.8803110
- [2] Saeed Ranjbar Alvar and Ivan V. Bajić. 2020. Bit Allocation for Multi-Task Collaborative Intelligence. In *ICASSP*. 4342–4346. doi:10.1109/ICASSP40776.2020.9054770
- [3] Saeed Ranjbar Alvar and Ivan V. Bajić. 2021. Pareto-Optimal Bit Allocation for Collaborative Intelligence. *IEEE Transactions on Image Processing* 30 (2021), 3348–3361. doi:10.1109/TIP.2021.3060875
- [4] Johannes Ballé, David C. Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. 2018. Variational image compression with a scale hyperprior. *ArXiv abs/1802.01436* (2018).
- [5] Yangang Cai, Peiyin Xing, and Xuesong Gao. 2022. High Efficient 3D Convolution Feature Compression. *IEEE Transactions on Circuits and Systems for Video Technology* (2022), 1–1. doi:10.1109/TCSVT.2022.3200698
- [6] Qiaoxi Chen, Changsheng Gao, and Dong Liu. 2024. End-to-End Learned Scalable Multilayer Feature Compression for Machine Vision Tasks. In *ICIP*. 1781–1787. doi:10.1109/ICIP51287.2024.10647798
- [7] Zhuo Chen, Ling-Yu Duan, Shiqi Wang, Weisi Lin, and Alex C. Kot. 2020. Data Representation in Hybrid Coding Framework for Feature Maps Compression. In *ICIP*. 3094–3098. doi:10.1109/ICIP40778.2020.9190843
- [8] Zhuo Chen, Kui Fan, Shiqi Wang, Lingyu Duan, Weisi Lin, and Alex Chichung Kot. 2020. Toward Intelligent Sensing: Intermediate Deep Feature Compression. *IEEE Transactions on Image Processing* 29 (2020), 2230–2243. doi:10.1109/TIP.2019.2941660
- [9] Zhuo Chen, Kui Fan, Shiqi Wang, Ling-Yu Duan, Weisi Lin, and Alex Kot. 2019. Lossy Intermediate Deep Learning Feature Compression and Evaluation. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*. Association for Computing Machinery, New York, NY, USA, 2414–2422. doi:10.1145/3343031.3350849
- [10] Hyomin Choi and Ivan V. Bajić. 2018. Deep Feature Compression for Collaborative Object Detection. In *ICIP*. 3743–3747. doi:10.1109/ICIP.2018.8451100
- [11] Hyomin Choi and Ivan V. Bajić. 2021. Latent-Space Scalability for Multi-Task Collaborative Intelligence. In *ICIP*. 3562–3566. doi:10.1109/ICIP42928.2021.9506712
- [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*. 248–255. doi:10.1109/CVPR.2009.5206848
- [13] Lingyu Duan, Jiaying Liu, Wenhan Yang, Tiejun Huang, and Wen Gao. 2020. Video Coding for Machines: A Paradigm of Collaborative Compression and Intelligent Analytics. *IEEE Transactions on Image Processing* 29 (2020), 8680–8695. doi:10.1109/TIP.2020.3016485
- [14] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88 (2010), 303–338.
- [16] Ruoyu Feng, Xin Jin, Zongyu Guo, Runsen Feng, Yixin Gao, Tianyu He, Zhizheng Zhang, Simeng Sun, and Zhibo Chen. 2022. Image coding for machines with omniscient feature learning. In *ECCV*. Springer, 510–528.
- [17] Changsheng Gao, Yiheng Jiang, Li Li, Dong Liu, and Feng Wu. 2024. DMOFC: Discrimination Metric-Optimized Feature Compression. In *PCS*. 1–5. doi:10.1109/PCS60826.2024.10566361
- [18] Changsheng Gao, Yiheng Jiang, Siqi Wu, Yifan Ma, Li Li, and Dong Liu. 2025. IMOFC: Identity-Level Metric Optimized Feature Compression for Identification Tasks. *IEEE Transactions on Circuits and Systems for Video Technology* 35, 2 (2025), 1855–1869. doi:10.1109/TCSVT.2024.3467124
- [19] Changsheng Gao, Zhuoyuan Li, Li Li, Dong Liu, and Feng Wu. 2024. Rethinking the Joint Optimization in Video Coding for Machines: A Case Study. In *DCC*. 556–556.
- [20] Changsheng Gao, Shan Liu, Feng Wu, and Weisi Lin. 2025. Cross-architecture universal feature coding via distribution alignment. *arXiv preprint arXiv:2506.12737* (2025).
- [21] Changsheng Gao, Zijie Liu, Li Li, Dong Liu, Xiaoyan Sun, and Weisi Lin. 2025. DT-UFC: Universal Large Model Feature Coding via Peak-to-Balanced Distribution Transformation. *arXiv preprint arXiv:2506.16495* (2025).
- [22] Changsheng Gao, Yifan Ma, Qiaoxi Chen, Yenan Xu, Dong Liu, and Weisi Lin. 2024. Feature Coding in the Era of Large Models: Dataset, Test Conditions, and Benchmark. *arXiv preprint arXiv:2412.04307* (2024).
- [23] Sha Guo, Zhuo Chen, Yang Zhao, Ning Zhang, Xiaotong Li, and Lingyu Duan. 2023. Toward Scalable Image Feature Compression: A Content-Adaptive and Diffusion-Based Approach. In *Proceedings of the 31st ACM International Conference on Multimedia (Ottawa ON, Canada) (MM '23)*. Association for Computing Machinery, New York, NY, USA, 1431–1442. doi:10.1145/3581783.3611851
- [24] Robert Henzel, Kiran Misra, and Tianying Ji. 2022. Efficient Feature Compression for the Object Tracking Task. In *ICIP*. 3505–3509. doi:10.1109/ICIP46576.2022.9897802
- [25] Yuzhang Hu, Sifeng Xia, Wenhan Yang, and Jiaying Liu. 2020. Sensitivity-Aware Bit Allocation for Intermediate Deep Feature Compression. In *VCIP*. 475–478. doi:10.1109/VCIP49819.2020.9301807
- [26] Ademola Ikusan and Rui Dai. 2021. Rate-Distortion Optimized Hierarchical Deep Feature Compression. In *ICME*. 1–6. doi:10.1109/ICME51207.2021.9428228
- [27] Yeongwoong Kim, Hyewon Jeong, Janghyun Yu, Younhee Kim, Jooyoung Lee, Se Yoon Jeong, and Hui Yong Kim. 2023. End-to-End Learnable Multi-Scale Feature Compression for VCM. *IEEE Transactions on Circuits and Systems for Video Technology* (2023), 1–1. doi:10.1109/TCSVT.2023.3302858
- [28] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*. PMLR, 3519–3529.
- [29] Shibao Li, Chenxu Ma, Yunwu Zhang, Longfei Li, Chengzhi Wang, Xuerong Cui, and Jianhang Liu. 2023. Attention-Based Variable-Size Feature Compression Module for Edge Inference. *The Journal of Supercomputing* (2023).
- [30] Yifan Ma, Changsheng Gao, Qiaoxi Chen, Li Li, Dong Liu, and Xiaoyan Sun. 2024. Feature Compression With 3D Sparse Convolution. In *VCIP*. 1–5.
- [31] Kiran Misra, Tianying Ji, Andrew Segall, and Frank Bossen. 2022. Video Feature Compression for Machine Tasks. In *ICME*. 1–6. doi:10.1109/ICME52920.2022.9859894
- [32] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*.
- [33] Benben Niu, Xiaoran Cao, Ziwei Wei, and Yun He. 2021. Entropy Optimized Deep Feature Compression. *IEEE Signal Processing Letters* 28 (2021), 324–328. doi:10.1109/LSP.2021.3052097
- [34] Maxime Quab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).
- [35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 8748–8763.
- [36] Saurabh Singh, Sami Abu-El-Haija, Nick Johnston, Johannes Ballé, Abhinav Shrivastava, and George Toderici. 2020. End-to-End Learning of Compressible Features. In *ICIP*. 3349–3353. doi:10.1109/ICIP40778.2020.9190860
- [37] Shurun Wang, Shiqi Wang, Wenhan Yang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. 2022. Towards Analysis-Friendly Face Representation With Scalable Feature and Texture Compression. *IEEE Transactions on Multimedia* 24 (2022), 3169–3181. doi:10.1109/TMM.2021.3094300
- [38] Zixi Wang, Fan Li, Yunfei Zhang, and Yuan Zhang. 2023. Low-Rate Feature Compression for Collaborative Intelligence: Reducing Redundancy in Spatial and Statistical Levels. *IEEE Transactions on Multimedia* (2023), 1–16. doi:10.1109/TMM.2023.3303716
- [39] WG2. April 2023. Call for Proposals on Feature Compression for Video Coding for Machines. ISO/IEC JTC 1/SC 29/WG 2, N282 (April 2023).
- [40] Ning Yan, Changsheng Gao, Dong Liu, Houqiang Li, Li Li, and Feng Wu. 2021. SSSIC: Semantics-to-Signal Scalable Image Coding With Learned Structural Representations. *IEEE Transactions on Image Processing* 30 (2021), 8939–8954. doi:10.1109/TIP.2021.3121131
- [41] Wenhan Yang, Haofeng Huang, Yueyu Hu, Ling-Yu Duan, and Jiaying Liu. 2024. Video Coding for Machines: Compact Visual Representation Compression for Intelligent Collaborative Analytics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024), 1–18. doi:10.1109/TPAMI.2024.3367293
- [42] Zhongzheng Yuan, Samyak Rawlekar, Siddharth Garg, Elza Erkip, and Yao Wang. 2022. Feature Compression for Rate Constrained Object Detection on the Edge. In *2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval*. 1–6. doi:10.1109/MIPR54900.2022.00008
- [43] Zhicong Zhang, Mengyang Wang, Mengyao Ma, Jiahui Li, and Xiaopeng Fan. 2021. MSFC: Deep Feature Compression in Multi-Task Network. In *ICME*. 1–6. doi:10.1109/ICME51207.2021.9428258
- [44] Lingyu Zhu, Binzhe Li, Riyu Lu, Peilin Chen, Qi Mao, Zhao Wang, Wenhan Yang, and Shiqi Wang. 2024. Learned Image Compression for Both Humans and Machines via Dynamic Adaptation. In *ICIP*. IEEE, 1788–1794.