Diversity Preserving Coresets for Image Quality Assessment

Supplementary Material

A. More on Related Works

In this section, we highlight the key challenges associated with applying existing coreset selection methods to the task of Image Quality Assessment (IQA). More broadly, adapting existing coreset selection strategies to IQA tasks is non-trivial due to the regression-based nature of IQA, the complexity of its latent representations which often lack well-defined local clusters, and the inherent incoherence between semantic content, distortions, and the perceived quality scores of images.

Uncertainty-based approaches [6] struggle in this setting, as typical uncertainty estimation techniques rely on discrete class probabilities, which are not directly applicable to continuous output domains like IQA. Similarly, gradient matching methods [16, 20] depend on precise gradient estimates from the model. In IQA, such gradients can be unreliable or noisy due to the intricate interplay between semantic content and image distortions. Approaches based on loss or decision boundaries - such as loss-based methods [25] and decision boundary-driven strategies [7] typically depend on identifying misclassifications, which are not well-defined when working with continuous quality predictions. Standard loss functions in IQA, like Mean Squared Error (MSE) and Mean Absolute Error (MAE), capture the discrepancy between predicted and true quality scores, but these differences are not easily translated into selection criteria for coresets.

Bi-level optimization techniques [3, 11, 27] often require access to ground-truth labels and involve computationally expensive nested optimization loops, which further limits their utility in IQA tasks, especially when labeled data is scarce. Dataset Quantization (DQ) [33] combines coreset selection with patch-level quantization to create a compact, representative dataset. It selects key samples and stores only image patches, which are later reconstructed using generative models like Masked Auto-Encoders (MAEs). Techniques like DQ that utilize saved image patches and regenerate images through Masked Auto-Encoders have shown strong performance for classification tasks, where preserving semantic content is the primary concern. However, such methods introduce reconstruction artifacts that can degrade perceived image quality, making them unsuitable for IQA tasks.

Despite progress in general coreset techniques, these limitations highlight the need for tailored methods that can address the intricate requirements of IQA and improve the effectiveness of coreset selection for this domain.

B. IQA Datasets

We evaluated the effectiveness of our proposed coreset selection strategy, Q-Diverse, on seven widely used image quality assessment (IQA) benchmarks, each offering distinct characteristics in terms of content and distortion types.

- **KADID-10K** [19]: Composed of 10,125 images generated by applying 25 types of synthetic distortions to 81 reference images, this dataset captures a broad spectrum of degradations such as noise, blur, and compression artifacts. It serves as a comprehensive benchmark for studying coreset selection under controlled distortion scenarios.
- **TID2013** [22]: This dataset includes 3,000 images created from 24 pristine images, each distorted with 25 synthetic distortion types across five severity levels. It presents a wide range of perceptual quality impairments including additive noise and color degradations, making it suitable for evaluating robustness to artificial noise patterns.
- **KonIQ-10K** [14]: Featuring 10,073 images with real-world distortions, this dataset reflects authentic visual imperfections resulting from diverse capture conditions. The naturalistic quality variations make it a valuable resource for evaluating performance in practical IQA applications.
- **SPAQ** [8]: This dataset comprises 11,125 images captured using multiple smartphone devices, showcasing quality variations due to mobile-specific challenges such as sensor noise, exposure issues, and motion blur. It emphasizes authentic distortions prevalent in mobile photography.

- **AGIQA-3K** [18]: Containing distortions generated through deep generative models, AGIQA-3K includes 2,982 images with novel and complex artifacts. This dataset offers a unique testbed for IQA methods intended to handle content generated by AI models and other emerging technologies.
- **PIPAL** [10]: The PIPAL dataset is a large-scale image quality assessment (IQA) benchmark designed for perceptual image restoration tasks. It contains distorted images, including outputs from traditional and GAN-based super-resolution methods, with corresponding subjective quality scores.
- **FLIVE** [30]: This dataset contains around 40,000 real-world, sourced from open platforms, the collection spans diverse content, sizes, and natural distortions. It also includes 120,000 image patches at varied scales and aspect ratios for fine-grained perceptual quality analysis.

As detailed in Table 1, we report the Mean Opinion Score (MOS) range for each dataset, along with the number of images selected for various dataset fractions (from 1% to 95%). This structured evaluation allows for a consistent analysis of coreset performance across a wide range of dataset sizes. Together, these benchmarks provide a diverse and comprehensive platform for validating coreset selection methods across synthetic, authentic, and AI-generated distortion domains.

Table 1. Dataset statistics and coreset sizes for different dataset fractions. Each dataset varies in distortion type and MOS range. The coreset sizes (1% to 95%) correspond to the number of selected images at each fraction. 'Full' represent full training set.

Dataset	Distortion Category	#images	Coreset Size = #images corresponding to dataset fraction								Full
			1.0%	5.0%	10.0%	30.0%	50.0%	70.0%	90.0%	95.0%	
KADID-10K	Synthetic	10125	70	354	708	2125	3543	4960	6377	6731	7086
TID2013	Synthetic	3000	20	104	209	629	1049	1469	1889	1994	2099
KonIQ-10k	Authentic	10073	70	352	705	2115	3525	4935	6345	6698	7051
SPAQ	Authentic	11125	77	389	778	2336	3893	5450	7008	7397	7787
AGIQA-3K	AI Generated	2982	20	104	208	625	1043	1460	1877	1981	2086
PIPAL	GAN Based	23200	208	1044	2088	6264	10440	14616	18792	19836	20880
FLIVE	Authentic	39810	278	1393	2786	8359	13933	19506	25079	26472	27866

C. Baseline Methods

We benchmarked the performance of our proposed method, Q-Diverse, against five widely recognized coreset selection techniques: Herding [4], k-Center Greedy [23], Contextual Diversity (CD) [1], and Moderate Coreset [26].

- **Herding** [4]: This method incrementally selects representative samples to approximate the target data distribution in the feature space. It minimizes the difference between the empirical mean of the coreset and that of the full dataset, without requiring model training. The approach is inspired by maximum entropy principles.
- **k-Center Greedy** [23]: This algorithm solves a variant of the facility location problem, where the objective is to select a subset from dataset such that $CS \subset D$ of k points so the maximum distance between any point in $D \setminus CS$ and its nearest neighbor in CS is minimized. Due to the NP-hard nature of this problem, a greedy approximation is employed. The method provides theoretical guarantees by relating the average loss over the selected subset to that over the full dataset, and is commonly used in active learning for CNNs.
- **Contextual Diversity (CD)** [1]: Designed to improve active learning in convolutional neural networks, CD prioritizes selection based on contextual diversity rather than purely visual or predictive uncertainty. It captures spatial variations and co-occurrence patterns in local regions, making it more effective in selecting informative and contextually diverse samples.
- **Moderate Coreset** [26]: Unlike fixed-range coreset methods, which rely on predefined score thresholds that may not generalize across varying distributions, this technique uses the median score as a stable selection criterion. Although more adaptable to handle noise, it applies a global median threshold without considering class-specific variations, which may limit its effectiveness when class densities differ significantly.
- **PGCS** [21]: PGCS is a coreset selection method for IQA that partitions the latent space of multitask-trained encoder features and performs adaptive sampling within each partition. The partitions ensure coverage of distortions, semantics, and perceptual quality, while adaptive sampling balances representativeness.

D. Time Complexity Analysis

In this section, we present a detailed time complexity analysis of the proposed Q-Diverse algorithm. Let the input dataset $\mathcal{X}$ contain N images, each encoded into d-dimensional embeddings via the content and quality encoders. Let b denote the batch size, $K = \lceil N/b \rceil$ be the number of batches, l be the number of Nyström landmarks, r the number of retained eigenvectors (based on energy threshold ϵ), and M the total coreset size. We analyze each stage of the algorithm below.

D.1. Content-Quality-Aware Embedding Extraction

For each batch $\mathcal{B}_i$, we compute embeddings using f_c and f_q, followed by pairwise Euclidean distances in both embedding spaces. Using optimized vector operations (e.g., matrix broadcasting), the pairwise distance matrices $\mathbf{Dist}^{(C)}$ and $\mathbf{Dist}^{(Q)}$ are computed in $O(b^2)$ time instead of the naive $O(b^2 d)$.

$$T_{\text{embed}}^{(i)} = O(bd + b^2) \tag{1}$$

Over all K batches, the embedding extraction cost is:

$$T_{\text{embed}} = O(Nd + Nb) \tag{2}$$

D.2. Spectral Space Construction via Nyström Approximation

For each batch, we construct a similarity matrix from the distance matrix using the RBF kernel and compute the graph Laplacian. The Nyström method is then used to approximate the spectral embedding.

$$\begin{aligned} T_{\text{spectral}}^{(i)} &= O(b^2) \quad \text{(kernel and Laplacian)} \\ &\quad + O(b \cdot l + l^3 + b \cdot l \cdot r) \end{aligned} \tag{3}$$

Summing over all K batches yields:

$$T_{\text{spectral}} = O(Nb + Nlr + Kl^3) \tag{4}$$

D.3. Diversity-Based Sampling in Spectral Space

Given spectral embeddings $\mathbf{Z} \in \mathbb{R}^{b \times r}$, we perform pairwise similarity computation and iterative selection using a geometric diversity criterion. Pairwise distances and the RBF kernel matrix take $O(b^2 r + b^2)$ time. The diversity sampling step proceeds for M_i rounds per batch, where $M_i \approx M/K$. Each round requires updating scores for $O(b)$ points via matrix-vector operations involving the kernel submatrix inverse, which we conservatively upper bound by $O(b \cdot M_i^2)$ per batch.

$$T_{\text{sampling}}^{(i)} = O(b^2 r + b \cdot M_i^2) \tag{5}$$

Aggregating over all batches, the total sampling cost becomes:

$$T_{\text{sampling}} = O\left(\frac{N^2 r}{K} + \frac{NM^2}{K^2}\right) \tag{6}$$

D.4. Total Time Complexity

Combining all components, the total time complexity of Q-Diverse is:

$$\boxed{T_{\text{total}} = O\left(Nd + Nb + Nlr + Kl^3 + \frac{N^2 r}{K} + \frac{NM^2}{K^2}\right)} \tag{7}$$

The use of optimized Euclidean operations and the Nyström method ensures the algorithm scales efficiently with large datasets, while batch processing prevents memory overload and allows parallelization across batches.

E. Implementation Details

In this section, we provide the detailed implementation settings in our experiments. We begin by introducing hyperparameters for our proposed Q-Diverse method, followed by the baseline configurations used for comparison. Subsequently, we describe the implementation details of two IQA architectures, MANIQA [29] and MUSIQ [15]. These details ensure the reproducibility of our results and enable fair performance comparisons across different methods.

E.1. Q-Diverse

For all experiments, we used a fixed regularization constant $\varepsilon = 1 \times 10^{-6}$ to ensure numerical stability during the inversion of kernel submatrices. The energy threshold ϵ that controls the spectral embedding dimensionality was consistently set to 0.95 in all datasets. The quality–content trade-off parameter γ was set to 0.5 for all datasets except TID2013, where it was increased to 0.7. The number of Nyström landmarks (l) and the mini-batch size (b) were selected based on the size of the data set. Specifically, we used $(l = 600, b = 1000)$ for KADID-10K and KonIQ-10K, $(l = 600, b = 1500)$ for SPAQ. Q-Diverse is capable of operating in non-batch mode for small-scale datasets where batch processing may be unnecessary. These values were empirically determined to balance computational efficiency with the representation quality of the selected coreset. Q-Diverse leverages content-aware image embeddings using ResNet-50 [12] and quality-aware image embeddings from ARNIQA [2].

E.2. Baseline Implementations

The baseline coreset selection methods [1, 4, 23, 26] were not originally designed to operate on spectral embeddings constructed from distinct content- and quality-aware representations. Therefore, to ensure a fair comparison, we employed the LIQE [32] image encoder instead of conventional deep features (e.g., ResNet-50 [12] or VGG [24]) for all baselines, as it inherently captures both scene semantics and distortion characteristics within its learned features. Details regarding the implementation and adaptations of each baseline are summarized below:

- **Herding** [4] and **k-Center Greedy** [23]: Both methods were implemented using the DeepCore repository[1]. The original feature extractors were replaced with the LIQE encoder, and the resulting embeddings were used as input for coreset selection.
- **Contextual Diversity (CD)** [1]: We utilized the official implementation available[2]. For application in IQA, we discretized the continuous Mean Opinion Scores (MOS) into a fixed number of bins. The quality logits produced by the LIQE encoder were then used to construct a co-occurrence matrix among the discretized quality levels, following the original CD formulation.
- **Moderate Coreset** [26]: The publicly available official repository[3] was employed. As the method requires categorical labels, we transformed the continuous MOS scores into discrete bins, treating each bin as a separate class label. This enabled the method to be applied in the IQA setting while preserving its original framework.
- **PGCS** [21]: We employed the publicly available official repository[4]. We kept the hyperparameter values unchanged for consistency. .

E.3. IQA Architecture: MANIQA

We trained and evaluated MANIQA on coresets selected from all seven datasets described in Section B, using the official implementation[5] with default settings. Input images were resized to 224×224, with random horizontal flips applied (probability 0.7). The model architecture includes two stages with Transposed Attention Blocks and Scale Swin Transformer Blocks, using embedding dimensions $D_1 = 768$, $D_2 = 384$, MLP size $D_m = 768$, number of heads $H = 4$, window size 4, and scaling factor $\alpha = 0.80$. Training used the ADAM optimizer with batch size 8, learning rate 1×10^{-5}, weight decay 1×10^{-5}, cosine annealing schedule for 25 epochs using Mean Square Error Loss (MSE) loss. To ensure fair comparisons across dataset fractions and methods, test sets remained fixed, and SRCC and PLCC were reported.

[1] https://github.com/PatrickZH/DeepCore
[2] https://github.com/sharat29ag/CDAL
[3] https://github.com/tmllab/2023_ICLR_Moderate-DS.git
[4] https://github.com/Arpita2012/PGCS
[5] [https://github.com/IIGROUP/MANIQA

E.4. IQA Architecture: MUSIQ

We employed the official MUSIQ implementation[6] with default settings for all experiments. MUSIQ utilizes a Transformer-based architecture with a hidden dimension of 384, 6 attention heads, and an MLP with a hidden size of 1152. Dropout is set to 0.1 for attention, feed-forward, and embedding layers. Layer normalization with $\varepsilon = 1 \times 10^{-12}$ is used for numerical stability. A grid size of 10 is adopted for spatial embeddings. Training was conducted using the SGD optimizer with a batch size of 2 and a cosine annealing learning rate scheduler ($T_{\text{max}} = 3 \times 10^4$, $\eta_{\text{min}} = 0$). The learning rate and number of training epochs for MUSIQ were customized for each dataset. For KonIQ-10K [14], KADID-10K [19], PIPAL [10], FLIVE [30] and SPAQ [8], we used a learning rate of 1×10^{-4}, with 30 training epochs. For AGIQA-3K, the model was trained with a learning rate of 1×10^{-5} for 10 epochs.

F. Extended Results

This section presents extended experimental results to further validate the effectiveness of the proposed Q-Diverse approach. We begin by reporting the quantitative performance of Q-Diverse on the MUSIQ architecture across different dataset fractions, demonstrating its ability to maintain high performance with significantly reduced training data. We then provide t-SNE visualizations to analyze the spatial distribution and diversity of the selected coresets. Finally, we evaluate the quality range coverage of the selected coresets across multiple IQA benchmarks, confirming that Q-Diverse preserves the full perceptual spectrum of the original datasets. Together, these results highlight the diversity and representativeness of Q-Diverse coreset selection.

F.1. Results on MUSIQ

Table 2 presents the performance of different coreset selection strategies when training MUSIQ [15] across various dataset fractions. Q-Diverse consistently achieves superior results over the baselines, especially at lower fractions, highlighting its capacity to retain essential quality-relevant diversity with minimal data. Performance with only 30%–70% of the dataset remains on par with full data training, indicating that Q-Diverse effectively discards redundant or low-utility samples. These findings reinforce the effectiveness of informed coreset selection in enhancing model generalization by constructing compact yet representative training subsets.

F.2. Coresets as Calibration Sets

Coresets are typically evaluated based on their training performance when used with deep neural networks. In this section, we additionally explore the effectiveness of the Q-Diverse-selected coreset in the context of model compression, specifically neural network pruning. Pruning strategies often target neurons or blocks based on certain importance criteria. For this preliminary analysis, we adopt the RDPrune framework [28], which emphasizes the use of a calibration set to generate rate-distortion curves. Instead of using random samples or noise-based selections, we utilize our Q-Diverse coreset as the calibration set to prune the BaseCNNIQA architecture[7]. We follow the original implementation provided by the authors of RDPrune[8] along with adaptation for IQA task, keeping all other settings unchanged. Table 3 presents a performance comparison of different coreset selection methods when used as the calibration set for pruning. We apply RDPrune with a target sparsity of 20% on the BaseCNNIQA network. The results demonstrate how coreset selection influences the effectiveness of pruning.

F.3. Performance on Image Classification

While Q-Diverse is primarily designed for IQA, we extend its application to a standard image classification dataset CIFAR-10 [17] in order assess its generalization capability. Table 4 reports classification accuracy using coresets selected by different methods. For this evaluation, Q-Diverse was configured with 500 landmark count and a batch size of 1500. Although Q-Diverse is not specifically optimized for classification, it achieves performance comparable to other coreset selection methods. ZCore [9] is also a label-free method; however, it relies on a strong assumption that feature elements follow a triangular distribution. In contrast, Q-Diverse imposes no such distributional assumptions, making it more broadly applicable. For this experiment, the hyperparameter γ was set to 0.7 to emphasize content-specific features over quality embeddings. These results suggest that IQA, as inspired by prior work such as [5], support efficient neural network training as well coreset selection.

[6] https://github.com/anse3832/MUSIQ
[7] https://github.com/zwx8981/DBCNN-PyTorch
[8] https://github.com/Akimoto-Cris/RD_PRUNE

Table 2. Performance comparison of coreset selection methods for MUSIQ [15], trained on dataset fractions chosen by each method and evaluated on the corresponding test sets. **Bold** values indicate the best performance for each dataset.

Dataset Fraction	Methods	KADID-10K		TID2013		KonIQ-10k		SPAQ		AGIQA-3K		PIPAL		FLIVE	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
5%	Herding	0.3457	0.3602	0.1617	0.1435	0.4430	0.4351	0.6967	0.6860	0.0203	0.1513	0.3434	0.3806	0.2453	0.2799
	K-center	0.4092	0.4207	0.1667	0.1808	0.4624	0.4514	0.7125	0.7007	0.0806	0.2374	0.3463	0.3864	0.2412	0.2784
	CD	0.1511	0.1713	0.2097	0.1754	0.4655	0.4414	0.7344	0.7142	0.0631	0.1907	0.3420	0.3859	0.2341	0.2797
	Moderate	0.3576	0.3686	0.2101	0.1985	0.5954	0.5803	0.7271	**0.7189**	0.0398	0.2003	0.3473	0.3988	**0.2484**	0.2794
	Q-Diverse	**0.4369**	**0.4503**	**0.2119**	**0.2046**	**0.6101**	**0.6158**	**0.7437**	0.7124	**0.0840**	**0.2398**	**0.3479**	**0.4184**	0.2430	**0.2842**
10%	Herding	0.3874	0.3929	0.2135	0.2459	0.3604	0.3383	0.7536	0.7223	0.0646	0.1176	0.3924	0.4493	0.2643	0.3064
	K-center	0.4425	0.4643	0.2379	0.2588	0.5573	0.5183	0.7410	0.7345	0.1650	0.2894	0.3987	0.4432	0.2639	0.3022
	CD	0.2081	0.2303	0.2403	0.2732	0.4606	0.4342	0.7412	0.6807	0.1738	0.2893	0.3871	0.4311	0.2635	0.3107
	Moderate	0.4719	0.4854	0.2436	0.2786	0.5851	0.5700	0.7685	0.7649	0.1542	0.2463	0.3979	0.4438	0.2665	0.3150
	Q-Diverse	**0.5113**	**0.5309**	**0.3430**	**0.3630**	**0.6633**	**0.6640**	**0.7985**	**0.7920**	**0.1961**	**0.3253**	**0.4044**	**0.4555**	**0.2731**	**0.3166**
30%	Herding	0.3098	0.3172	0.2279	0.2430	0.6161	0.6101	0.8168	0.7983	0.1692	0.2967	0.4812	0.5195	0.3284	0.3648
	K-center	0.4600	0.4856	0.2769	0.3098	0.5719	0.5502	0.8016	0.7720	0.2915	0.4274	0.4836	0.5237	0.3385	0.3745
	CD	0.2399	0.2653	0.2843	0.3381	0.6262	0.6095	0.8216	0.7931	0.1197	0.2658	0.4850	0.5264	0.3408	0.3792
	Moderate	0.4249	0.4393	0.3478	0.3765	0.6349	0.6187	0.7819	0.7783	0.3794	0.4521	**0.4933**	0.5304	0.3419	0.3845
	Q-Diverse	**0.5225**	**0.5361**	**0.4024**	**0.4273**	**0.6589**	**0.6202**	**0.8429**	**0.8443**	**0.4501**	**0.4601**	0.4901	**0.5440**	**0.3479**	**0.3857**
50%	Herding	0.5431	**0.5711**	0.2776	0.3138	0.6239	0.6064	0.8319	0.8204	0.4722	0.4542	0.5326	0.5865	0.3772	0.4336
	K-center	0.4001	0.4371	0.3485	0.3913	0.6293	0.6275	0.8179	0.8206	0.3535	0.4599	0.5463	0.5876	0.3826	0.4414
	CD	0.4922	0.5045	0.2540	0.2074	0.6950	0.6766	0.8295	0.7888	0.4492	0.4828	0.5348	0.5819	0.3834	0.4437
	Moderate	0.4721	0.4848	0.3924	0.4050	0.6173	0.5882	0.8124	0.8151	0.4646	**0.5289**	0.5438	0.5904	0.3878	0.4537
	Q-Diverse	**0.5444**	0.5561	**0.4710**	**0.5164**	**0.7185**	**0.7192**	**0.8483**	**0.8373**	**0.4827**	0.4907	**0.5498**	**0.5951**	**0.4077**	**0.4625**
70%	Herding	0.5002	0.5104	0.4697	0.5196	0.6785	0.6751	0.8143	0.8145	0.5783	0.5885	0.5818	0.6165	0.4475	0.5404
	K-center	0.4455	0.4725	0.4628	0.5275	0.7361	0.7358	0.8206	0.8106	0.5704	0.5774	0.5793	0.6154	0.4438	0.5523
	CD	0.4432	0.4478	0.4466	0.4972	0.6426	0.5984	0.8297	0.8300	0.5026	0.5114	0.5805	0.6158	0.4494	0.5545
	Moderate	0.4968	0.5059	0.4117	0.4404	0.6003	0.5826	0.7888	0.7840	0.4992	0.5451	0.5850	0.6256	0.4506	0.5541
	Q-Diverse	**0.5352**	**0.5384**	**0.4898**	**0.5449**	**0.7392**	**0.7380**	**0.8366**	**0.8345**	**0.5855**	**0.5891**	**0.5855**	**0.6263**	**0.4579**	**0.5700**
90%	Herding	0.5112	0.5225	0.4148	0.4984	0.6723	0.6450	0.8285	0.8292	0.5731	0.5604	0.6247	0.6358	0.5309	0.6263
	K-center	0.4951	0.5227	0.4477	0.5239	0.6709	0.6418	0.8149	0.8179	0.5733	0.5934	0.6252	0.6358	0.5355	0.6248
	CD	0.4530	0.4556	0.4417	0.5360	0.6635	0.6526	0.8109	0.8134	0.4036	0.4617	0.6258	0.6359	0.5388	0.6201
	Moderate	0.4642	0.4809	0.4752	0.5489	0.6528	0.6165	0.8162	0.8136	0.4501	0.5224	0.6348	0.6428	0.5303	**0.6384**
	Q-Diverse	**0.5476**	**0.5399**	**0.5118**	**0.5557**	**0.7183**	**0.6975**	**0.8379**	**0.8379**	**0.5865**	**0.5981**	**0.6382**	**0.6468**	**0.5454**	0.6356
95%	Herding	0.4759	0.4901	0.4717	0.5459	0.7116	0.6981	0.7967	0.7850	0.5227	0.5499	0.6518	0.6857	0.5563	0.6416
	K-center	0.4574	0.4745	0.4994	0.5520	0.7043	0.6881	0.7602	0.7549	0.4431	0.4961	0.6531	0.6847	0.5548	0.6411
	CD	0.4796	0.4990	0.4820	0.5527	0.6995	0.6863	0.8018	0.7910	0.4822	0.5176	0.6544	0.6863	0.5620	0.6412
	Moderate	0.4989	0.5169	0.4910	0.5528	0.6648	0.6397	0.8217	0.8237	0.5734	0.5985	0.6516	0.6811	0.5585	0.6527
	Q-Diverse	**0.5272**	**0.5255**	**0.5028**	**0.5632**	**0.7329**	**0.7281**	**0.8290**	**0.8321**	**0.5970**	**0.6081**	**0.6549**	**0.6984**	**0.5655**	**0.6580**
Full		0.5248	0.5265	0.4671	0.5461	0.7471	0.7531	0.8262	0.8299	0.5923	0.5994	0.6543	0.6968	0.5628	0.6591

F.4. Q-Diverse: Quality Range Coverage

Figure 1 presents the quality range coverage achieved by Q-Diverse selected coresets, each comprising 10% of the entire train dataset, on five standard IQA benchmarks. The plots compare the distribution of Mean Opinion Scores (MOS) between the full dataset and the corresponding coreset, illustrating that Q-Diverse consistently preserves the quality distribution, ensuring broad coverage across the perceptual quality spectrum.

F.5. Visualization

To understand the spatial distribution of selected samples, we employed t-SNE visualization [13] of each dataset's LIQE [32] embeddings along with coreset selections at varying dataset fractions (from 1% to 50%). In Figure 2 to Figure 6, the coreset points were highlighted in red, where darker shades indicated regions with higher selection density. These visualizations revealed how different methods behaved as the selection budget increased. In particular, Q-Diverse consistently exhibited superior spatial coverage, selecting points that spanned the entire embedding space while maintaining minimal redundancy. This widespread and diverse selection highlighted Q-Diverse's ability to capture both content-related and quality-related variations effectively.

Table 3. Performance comparison while using coreset as "Calibration set" for neural network pruning. Results are reported with calibration set (10% of dataset size) - selected using different coreset methods. RDPrune [28] method used for pruning a BaseCNNIQA network with target sparsity of 20%.

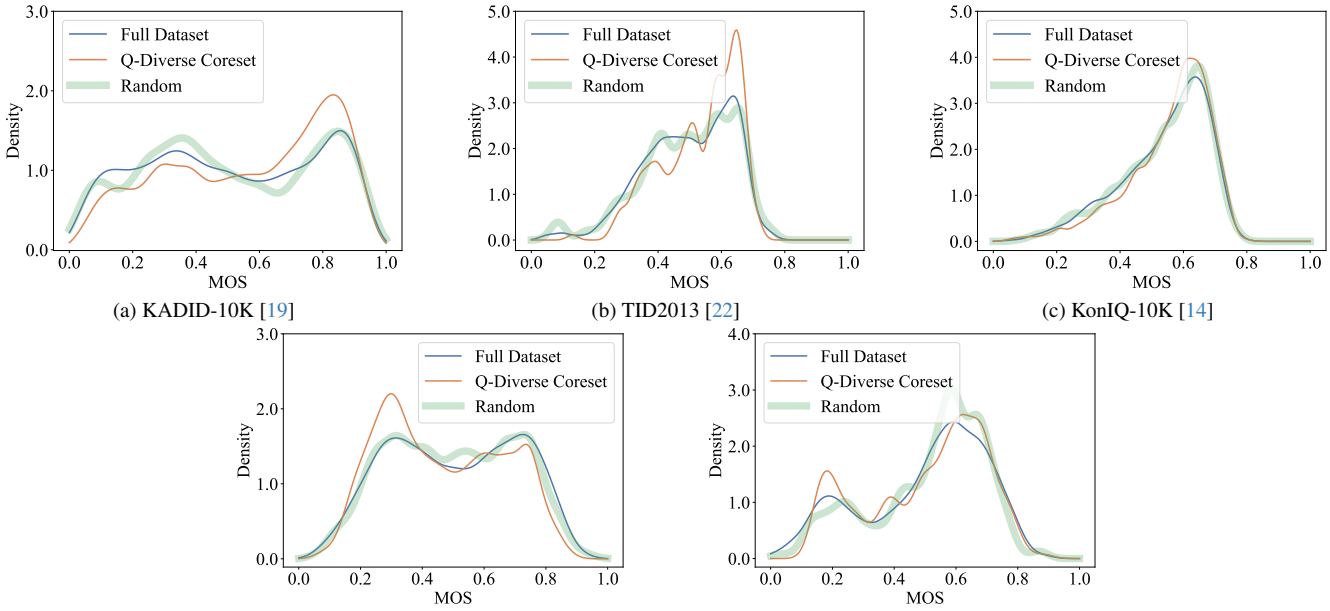
Calibration Set (% dataset fraction)	Coreset Methods	FLOPs	SRCC	PLCC
AGIQA-3K (10%)	Random	0.8132	0.7767	0.8473
	Herding	0.7220	0.7738	0.8430
	K-center Greedy	0.7615	0.7749	0.8463
	CD	0.7307	0.7759	0.8479
	Moderate	0.6358	0.7777	0.8513
	PGCS	0.6280	0.7782	0.8512
	Q-Diverse	**0.6103**	**0.7795**	**0.8540**
KADID-10K (5%)	Random	0.7696	0.9454	0.9489
	Herding	0.8186	0.9488	0.9511
	K-center Greedy	0.6617	0.9486	0.9519
	CD	0.7807	0.9473	0.9507
	Moderate	0.8148	0.9419	0.9466
	PGCS	0.6647	0.9465	0.9516
	Q-Diverse	**0.6526**	**0.9490**	**0.9519**
TID2013 (10%)	Random	0.6973	0.8976	0.9162
	Herding	0.6509	0.8986	0.9139
	K-center Greedy	0.7540	0.8933	0.9132
	CD	0.7439	0.8713	0.9040
	Moderate	0.6958	0.8870	0.9094
	PGCS	0.6612	0.8956	0.9097
	Q-Diverse	**0.6500**	**0.9085**	**0.9179**
KonIQ-10K(5%)	Random	0.7504	0.7931	0.8334
	Herding	0.7608	0.7800	0.8185
	K-center Greedy	0.7853	0.7906	0.8296
	CD	0.7601	0.7974	0.8307
	Moderate	0.7545	0.7990	0.8330
	PGCS	0.7566	0.7992	0.8337
	Q-Diverse	**0.7451**	**0.7995**	**0.8339**

Table 4. Performance comparison of Q-Diverse with other coreset selection methods on CIFAR-10 [17] for ResNet-18 [12]. Results marked with * and # are obtained from DeepCore [31] and ZCore [9], respectively, to ensure consistency with standard benchmarks.

Dataset Fraction	Herding*	K-center*	CD*	Moderate#	Z-core#	Q-Diverse
10%	63.5 ± 3.4	75.8 ± 2.4	58.5 ± 2.0	83.61 ± 0.09	84.18 ± 0.21	**84.93 ± 0.23**
30%	80.1 ± 2.2	90.9 ± 0.4	90.8±0.5	89.71±0.14	90.97± 0.44	**91.02±0.28**

G. Limitations

A notable limitation of Q-Diverse lies in its reliance on the quality of the underlying feature encoders. Although we adopt separate encoders for content and quality to better capture their distinct and complementary roles, the overall effectiveness of the method can be influenced by the representational strength of these pretrained networks. This sensitivity may become more pronounced in cross-domain settings, where possible misalignment between encoder outputs and perceptual ground truth could impact coreset fidelity. Additionally, while the geometric diversity sampling component is inherently sequential, its practical impact is mitigated in batch-wise implementations, making it less of a bottleneck in most settings.



(a) KADID-10K [19] (b) TID2013 [22] (c) KonIQ-10K [14]

(d) SPAQ [8] (e) AGIQA-3K [18]

Figure 1. Quality range coverage plots for selected coresets (Dataset Fraction = 10%) by Q-Diverse versus the full dataset. Each plot illustrates how well the Q-Diverse coreset maintains the distribution of MOS scores.

H. Future Work

Moving forward, we aim to explore self-supervised strategies to learn more robust feature embeddings tailored to coreset selection in IQA. Moreover, the principles behind Q-Diverse, joint modeling of quality and content can be extended beyond. Potential applications include dataset curation, personalized photo selection, and IQA continual learning setups where perceptual relevance and diversity are both critical.

References

[1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *ECCV*, pages 137–153. Springer, 2020. 2, 4

[2] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. ARNIQA: Learning distortion manifold for image quality assessment. In *IEEE WACV*, pages 3592–3602, 2024. 4

[3] Zalán Borsos, Mojmir Mutny, and Andreas Krause. Coresets via bilevel optimization for continual learning and streaming. *NeurIPS*, 33:14879–14890, 2020. 1

[4] Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. In *UAI*, pages 109–116, 2010. 2, 4

[5] Zhuo Chen, Weisi Lin, Shiqi Wang, Long Xu, and Leida Li. Image quality assessment guided deep neural networks training. *arXiv preprint arXiv:1708.03880*, 2017. 5

[6] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *ICLR*, 2020. 1

[7] Melanie Ducoffe and Frederic Precioso. Adversarial active learning for deep networks: A margin based approach. *arXiv preprint arXiv:1802.09841*, 1802. 1

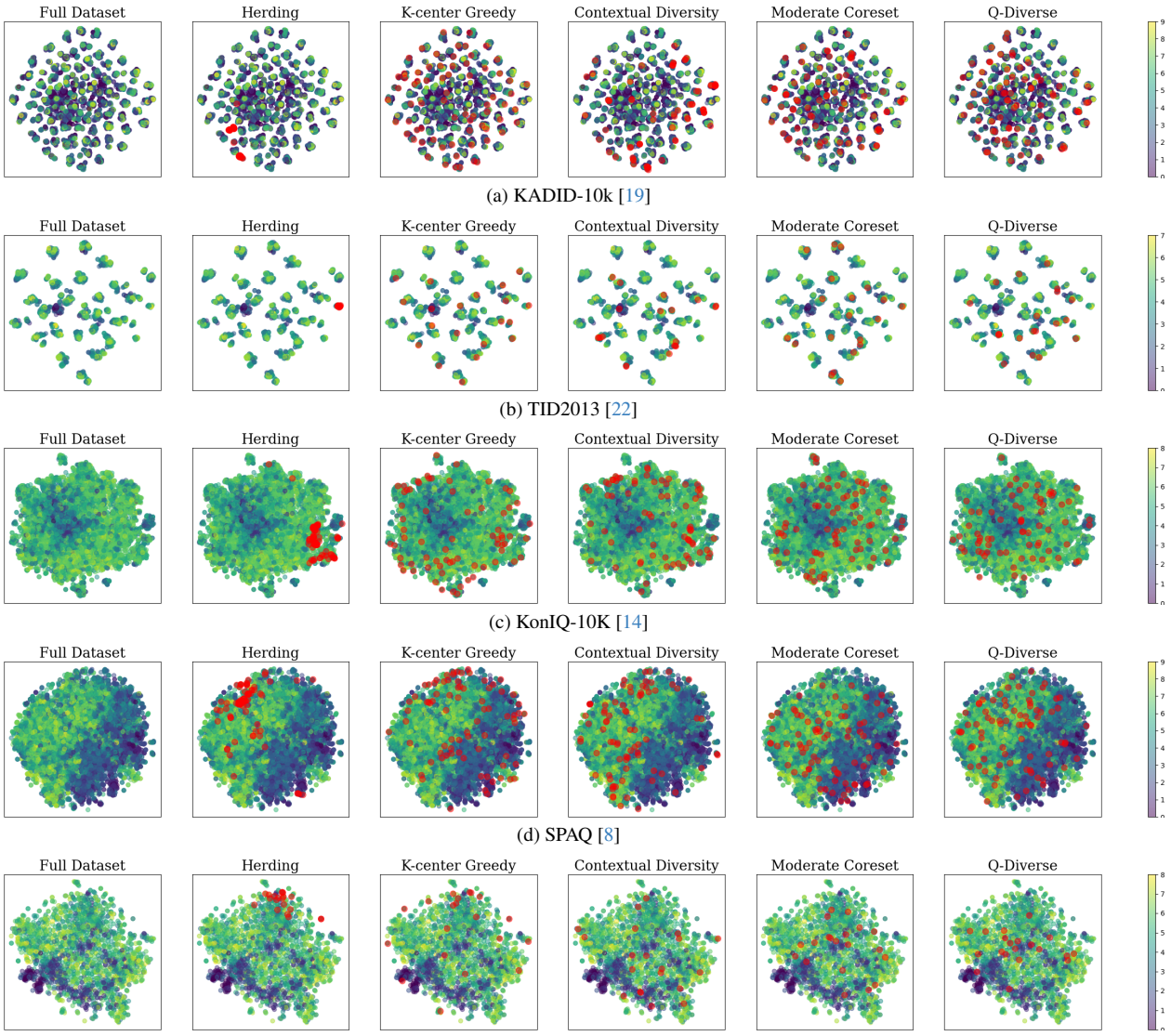
[8] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *IEEE CVPR*, pages 3677–3686, 2020. 1, 5, 8, 9, 10, 11, 12, 13

[9] Brent A Griffin, Jacob Marks, and Jason J Corso. Zero-shot coreset selection: Efficient pruning for unlabeled data. *arXiv preprint arXiv:2411.15349*, 2024. 5, 7

[10] Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy S Ren, and Chao Dong. PIPAL: A large-scale image quality assessment dataset for perceptual image restoration. In *ECCV*, pages 633–651, 2020. 2, 5

[11] Jie Hao, Kaiyi Ji, and Mingrui Liu. Bilevel coreset selection in continual learning: A new formulation and algorithm. *NeurIPS*, 36, 2024. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016. 4, 7



(a) KADID-10k [19]

(b) TID2013 [22]

(c) KonIQ-10K [14]

(d) SPAQ [8]

(e) AGIQA-3K [18]

Figure 2. t-SNE visualizations of full datasets and corresponding coreset selections (Dataset Fraction = 1%). Points selected in the coreset are highlighted in red. Regions with deeper red saturation denote higher density of selected samples, indicating spatial concentration in the embedding space. The colorbar represents the distribution of MOS across the dataset.

[13] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *NeurIPS*, 15, 2002. 6

[14] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE TIP*, 29:4041–4056, 2020. 1, 5, 8, 9, 10, 11, 12, 13

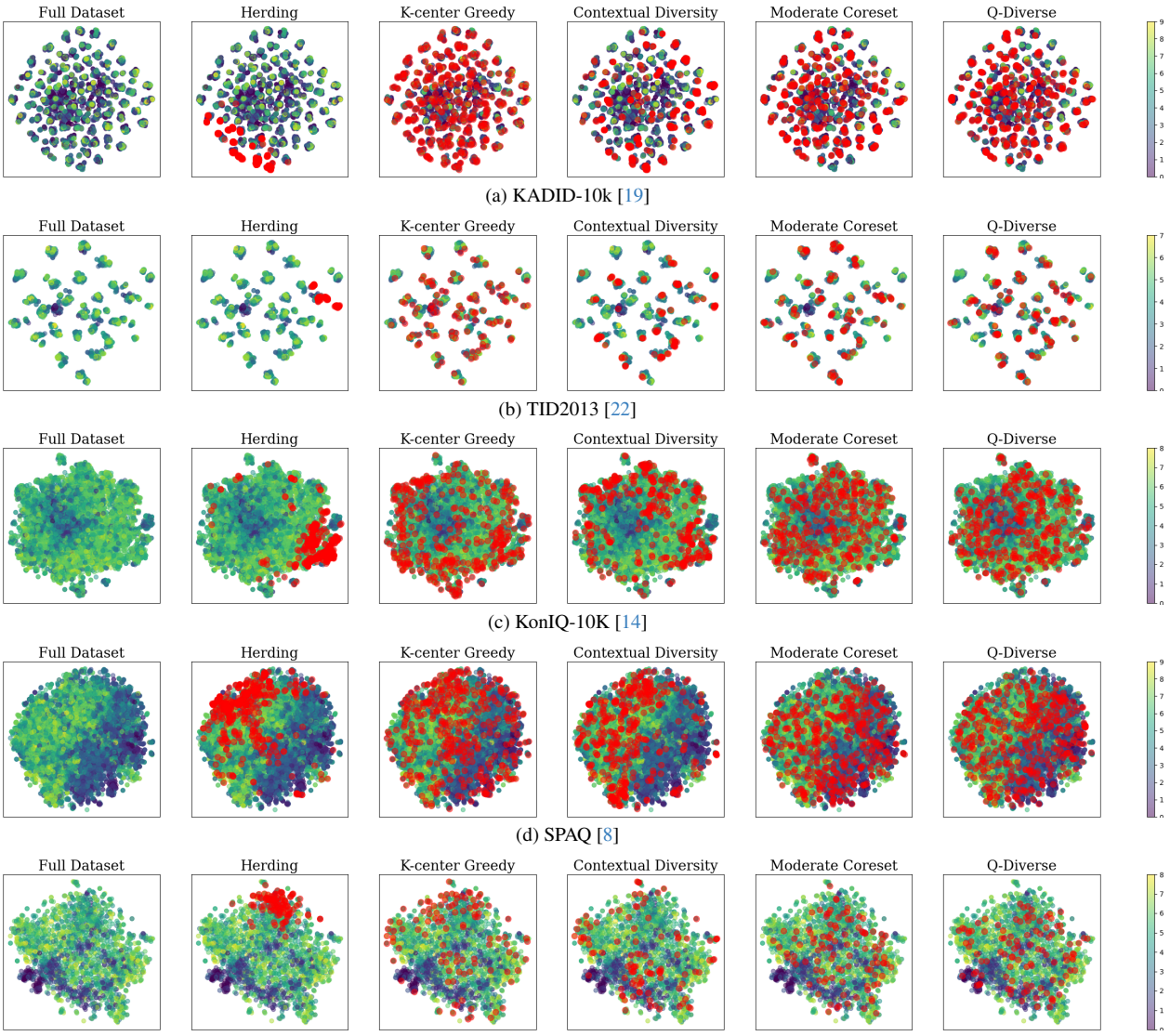
[15] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *IEEE CVPR*, pages 5148–5157, 2021. 4, 5, 6

[16] Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *ICML*, pages 5464–5474, 2021. 1

[17] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 5, 7

[18] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Agiqa-3k: An open database for ai-generated image quality assessment. *IEEE TCSVT*, 2023. 2, 8, 9, 10, 11, 12, 13

[19] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *IEEE QoMEX*, pages 1–3, 2019. 1, 5, 8, 9, 10, 11, 12, 13



(a) KADID-10k [19]

(b) TID2013 [22]

(c) KonIQ-10K [14]

(d) SPAQ [8]

(e) AGIQA-3K [18]

Figure 3. t-SNE visualizations of full datasets and corresponding coreset selections (Dataset Fraction = 5%). Points selected in the coreset are highlighted in red. Regions with deeper red saturation denote higher density of selected samples, indicating spatial concentration in the embedding space. The colorbar represents the distribution of MOS across the dataset.

[20] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *ICML*, pages 6950–6960. PMLR, 2020. 1

[21] Arpita Nema, Hanwei Zhu, and Weisi Lin. Holistic coreset selection for data efficient image quality assessment. In *IEEE ICIP*, pages 2432–2437, 2025. 2, 4

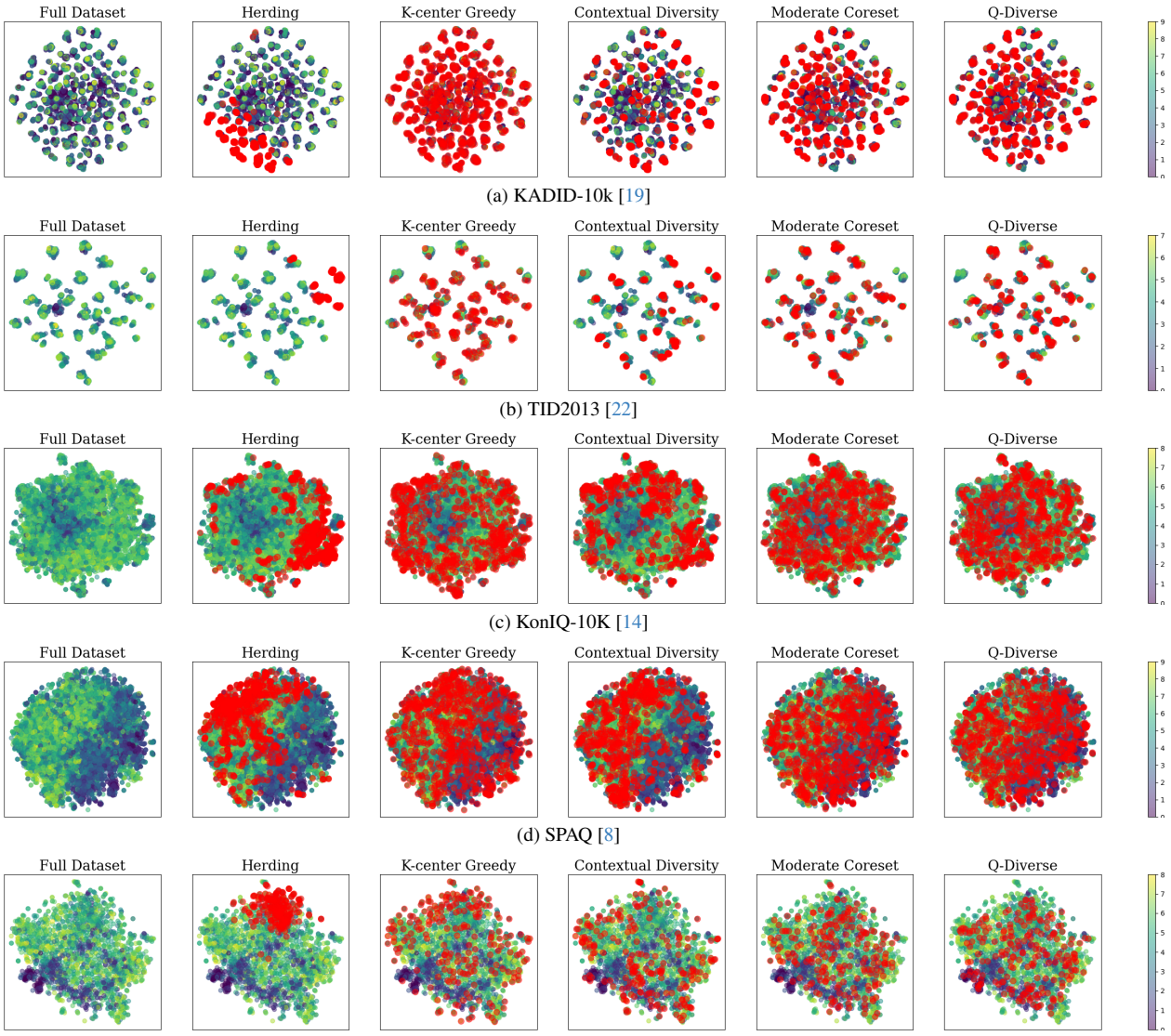
[22] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database TID 2013: Peculiarities, results and perspectives. *SPIC*, 30:57–77, 2015. 1, 8, 9, 10, 11, 12, 13

[23] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018. 2, 4

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4

[25] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. In *ICLR*, 2018. 1

[26] Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *ICLR*, 2022. 2, 4



(a) KADID-10k [19]

(b) TID2013 [22]

(c) KonIQ-10K [14]

(d) SPAQ [8]

(e) AGIQA-3K [18]

Figure 4. t-SNE visualizations of full datasets and corresponding coreset selections (Dataset Fraction = 10%). Points selected in the coreset are highlighted in red. Regions with deeper red saturation denote higher density of selected samples, indicating spatial concentration in the embedding space. The colorbar represents the distribution of MOS across the dataset.

[27] Xiaobo Xia, Jiale Liu, Shaokun Zhang, Qingyun Wu, Hongxin Wei, and Tongliang Liu. Refined coreset selection: Towards minimal coreset size under model performance constraints. In *ICML*, 2024. 1

[28] Kaixin Xu, Zhe Wang, Xue Geng, Min Wu, Xiaoli Li, and Weisi Lin. Efficient joint optimization of layer-adaptive weight pruning in deep neural networks. In *IEEE ICCV*, pages 17447–17457, 2023. 5, 7

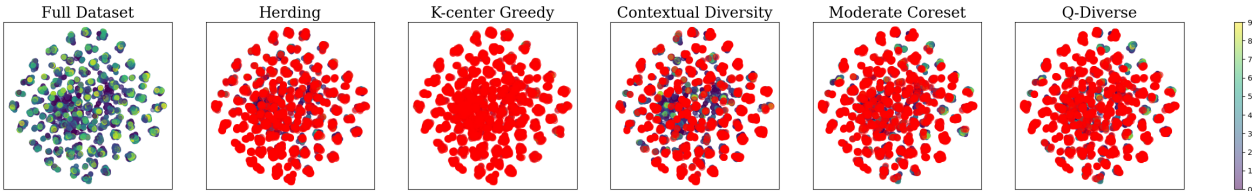
[29] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *IEEE CVPR*, pages 1191–1200, 2022. 4

[30] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *IEEE CVPR*, pages 3575–3585, 2020. 2, 5

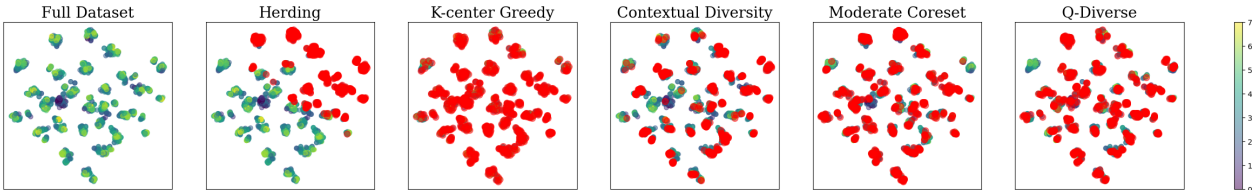
[31] Patrick ZH. Deepcore: A comprehensive library for coreset selection in deep learning, 2023. Accessed: 2024-11-08. 7

[32] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE CVPR*, pages 14071–14081, 2023. 4, 6

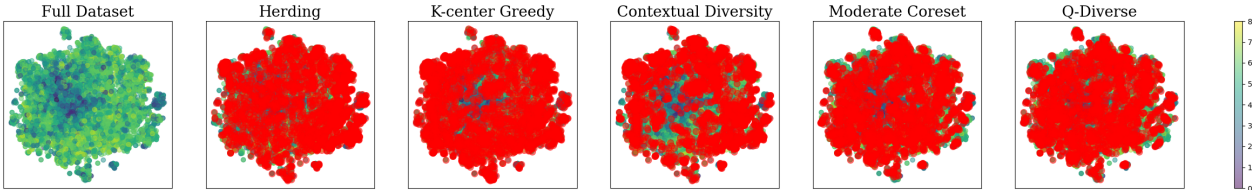
[33] D. Zhou, K. Wang, J. Gu, X. Peng, D. Lian, Y. Zhang, Y. You, and J. Feng. Dataset quantization. In *IEEE CVPR*, pages 17205–17216, 2023. 1



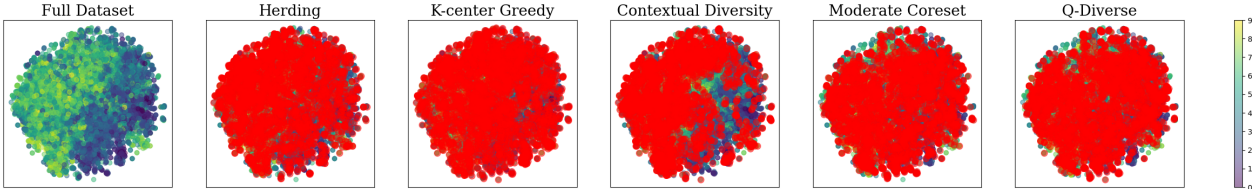
(a) KADID-10k [19]



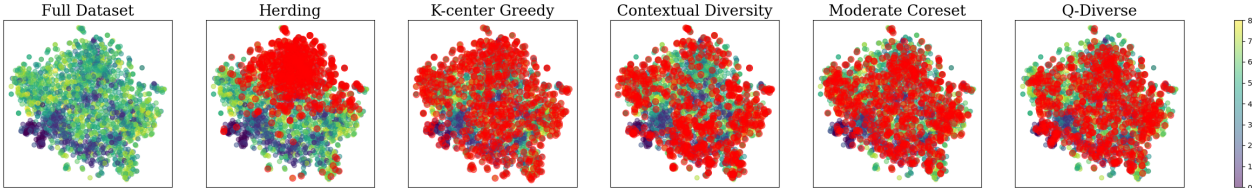
(b) TID2013 [22]



(c) KonIQ-10K [14]

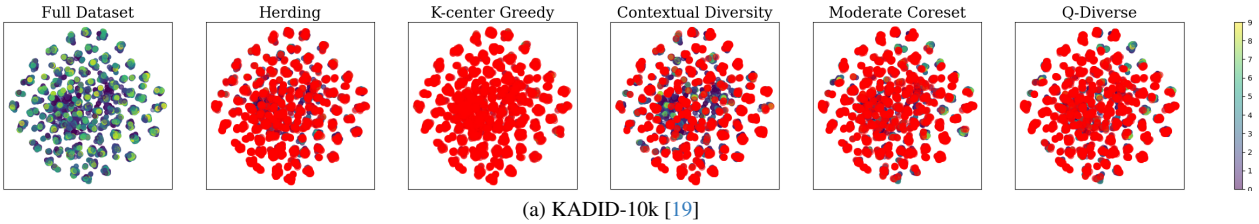


(d) SPAQ [8]

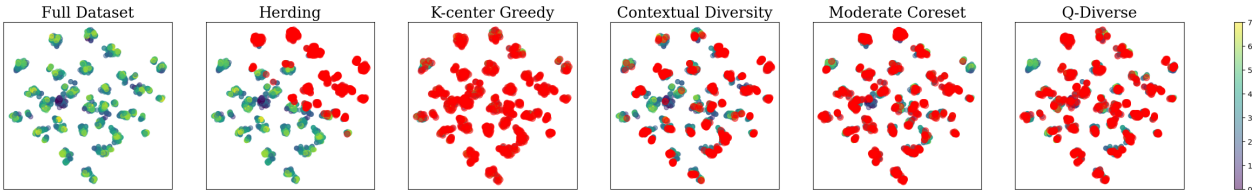


(e) AGIQA-3K [18]

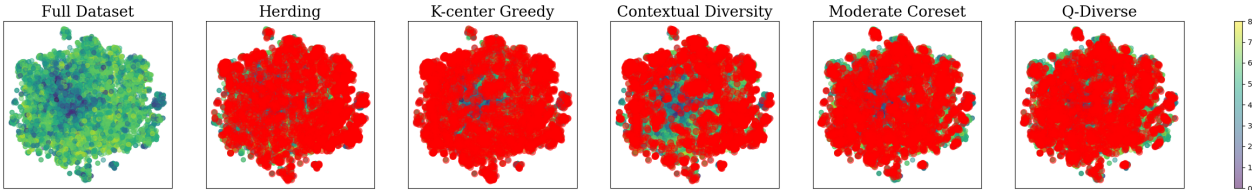
Figure 5. t-SNE visualizations of full datasets and corresponding coreset selections (Dataset Fraction = 30%). Points selected in the coreset are highlighted in red. Regions with deeper red saturation denote higher density of selected samples, indicating spatial concentration in the embedding space. The colorbar represents the distribution of MOS across the dataset.



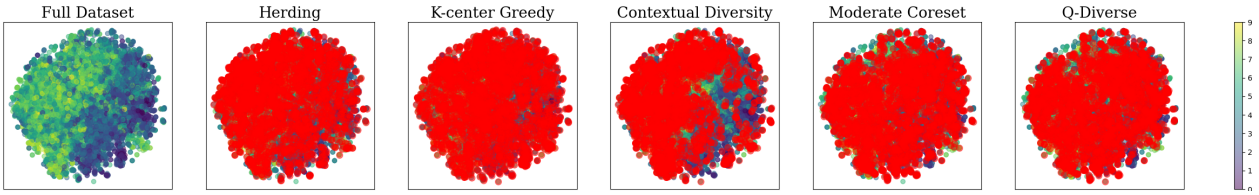
(a) KADID-10k [19]



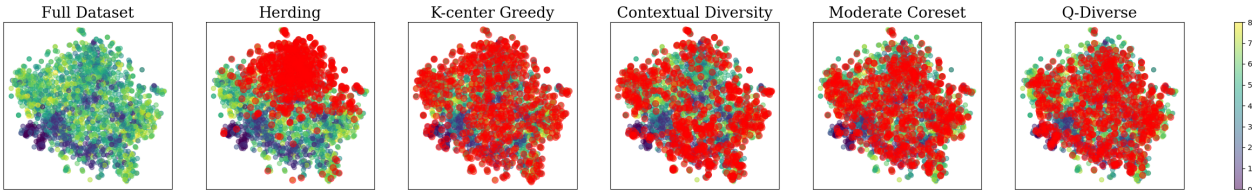
(b) TID2013 [22]



(c) KonIQ-10K [14]



(d) SPAQ [8]



(e) AGIQA-3K [18]

Figure 6. t-SNE visualizations of full datasets and corresponding coreset selections (Dataset Fraction = 50%). Points selected in the coreset are highlighted in red. Regions with deeper red saturation denote higher density of selected samples, indicating spatial concentration in the embedding space. The colorbar represents the distribution of MOS across the dataset.