Rethinking Knowledge Transfer in Image Quality Assessment: A Perceptual Preference Structure Alignment Perspective

Supplementary Material

1. Comparison with State-of-the-Art Methods

1.1. Cross-Domain Transfer (PreSTA-S)

We compare PreSTA-S with state-of-the-art IQA and UDA-based IQA methods under the synthetic-to-authentic transfer setting (KADID-10k $\rightarrow$ LIVEC, KonIQ-10k, BID). Results are summarized in Tab. 1. PreSTA-S achieves the best average performance across all datasets, outperforming both general IQA models (e.g., RankIQA [9], DBCNN [15], HyperIQA [12], MUSIQ [6], VCRNet [11], KGANet [18], CLIPIQA+ [13]) and the recent LMM-based Q-Align [14] in terms of both SRCC and PLCC.

Compared with UDA-based IQA methods that align the feature distribution $P(X)$ (DANN [3], UCDA [2], RankDA [1], StyleAM [10], FreqAlign [8], DGQA [7]), PreSTA-S further improves the average SRCC from 0.716 to 0.784 and the average PLCC from 0.710 to 0.794. This supports our key premise: when perceptual preference structures differ across domains, simply aligning $P(X)$ alone is insufficient. By selecting preference-consistent source samples, PreSTA-S reduces the discrepancy in $P(Y|X)$, leading to more robust and data-efficient transfer.

1.2. Targeted Joint Transfer (PreSTA-J)

We further compare PreSTA-J with state-of-the-art methods under the targeted joint transfer setting on LIVEC and BID, as summarized in Tab. 2. PreSTA-J consistently outperforms models trained solely on each target dataset (e.g., MUSIQ [6], DBCNN [15], HyperIQA [12], TreS [4]), demonstrating the effectiveness of leveraging auxiliary data through preference-guided selection.

Compared with joint-training approaches that aggregate multiple large-scale datasets (UNIQUE [16], LIQE [17]), PreSTA-J achieves the best overall performance while using substantially less data. Specifically, PreSTA-J relies on only a few thousand selected auxiliary samples per target, whereas joint training uses six full datasets [17]. This result indicates that aligning perceptual preference structures is more critical than simply increasing data scale.

2. Practical Estimation of Target PPR

In practical transfer scenarios, the target perceptual preference representation is typically unavailable from the full target dataset and must instead be estimated from a limited number of labeled samples. We therefore study this setting from two perspectives. We first analyze the statistical stability of PPR under limited-label estimation, and then evaluate whether PreSTA remains effective when the target PPR is estimated from only a small labeled subset.

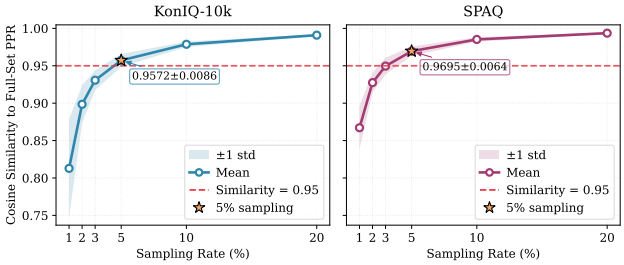

Figure 1. Cosine similarity between subset-estimated and full-set estimated PPR as a function of the sampling rate (100 runs). Shaded regions denote ± 1 std. Both KonIQ-10k and SPAQ achieve similarity above 0.95 at 5% sampling, indicating high stability of PPR estimation under limited target labels.

2.1. Statistical Stability of PPR under Limited Labels

To validate the robustness of PPR estimation, we analyze how closely a subset-estimated PPR approximates the PPR estimated from the full dataset. Specifically, for each target dataset, we randomly sample a fraction of labeled examples, compute the corresponding PPR, and measure its cosine similarity to the full-set estimated PPR. This process is repeated 100 times for each sampling rate, and we report the mean and standard deviation across runs.

As shown in Fig. 1, experiments on KonIQ-10k and SPAQ demonstrate that PPR remains highly stable even under very low sampling rates. With only 5%labeled samples, the mean cosine similarity already exceeds 0.95 (KonIQ-10k: 0.9572 ± 0.0086; SPAQ: 0.9695 ± 0.0064), while the standard deviation remains below 0.01. These results indicate that PPR captures a robust dataset-level preference structure rather than subset-specific noise, and can therefore be reliably estimated with very limited annotation cost.

2.2. Transfer Performance with Few-Label PPR Estimation

We next evaluate whether a subset-estimated target PPR is sufficient for practical transfer. In this experiment, for each target dataset, we randomly sample 5% labeled data to estimate the target PPR r_t, and use the remaining 95% samples for testing. This random split is repeated 10 times, and we report the mean and standard deviation across runs. We also include the full-data baseline and the PreSTA-S based

Table 1. Cross-domain synthetic-to-authentic transfer (KADID-10k→LIVEC, KonIQ-10k, BID).

Methods	LIVEC		KonIQ-10k		BID		Average	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
RankIQA [9]	0.491	0.495	0.603	0.551	0.510	0.367	0.535	0.471
DBCNN [15]	0.572	0.589	0.639	0.618	0.620	0.609	0.613	0.606
HyperIQA [12]	0.490	0.487	0.545	0.556	0.379	0.282	0.472	0.442
MUSIQ [6]	0.517	0.524	0.554	0.573	0.575	0.600	0.549	0.566
VCRNet [11]	0.561	0.548	0.517	0.525	0.542	0.545	0.540	0.540
KGANet [18]	0.575	-	0.528	-	-	-	-	-
CLIPIQA+ [13]	0.512	0.543	0.511	0.515	0.474	0.442	0.499	0.500
Q-Align [14]	0.702	0.744	0.668	0.665	-	-	-	-
DANN [3]	0.499	0.484	0.638	0.636	0.586	0.510	0.574	0.543
UCDA [2]	0.382	0.358	0.496	0.501	0.348	0.391	0.408	0.417
RankDA [1]	0.451	0.455	0.638	0.623	0.535	0.582	0.542	0.553
StyleAM [10]	0.584	0.561	0.700	0.673	0.637	0.567	0.640	0.600
FreqAlign [8]	0.618	0.588	0.748	0.721	0.674	0.708	0.680	0.673
DGQA [7]	0.696	0.690	0.681	0.687	0.770	0.753	0.716	0.710
PreSTA-S	**0.744**	**0.756**	**0.774**	**0.781**	**0.833**	**0.846**	**0.784**	**0.794**

Table 2. Targeted joint transfer on LIVEC and BID.

Method	**LIVEC**		**BID**	
	SRCC	PLCC	SRCC	PLCC
MUSIQ [6]	0.785	0.828	0.744	0.774
DBCNN [15]	0.835	0.883	0.864	0.883
HyperIQA [12]	0.855	0.878	0.848	0.868
TreS [4]	0.846	0.877	0.853	0.871
UNIQUE [16]	0.854	0.884	0.852	0.875
LIQE [17]	0.904	0.910	0.875	**0.900**
PreSTA-J	**0.905**	**0.919**	**0.885**	0.898

on full-set estimated PPR as references.

The results are summarized in Tab. 3. PreSTA-S remains highly effective even when the target PPR is estimated from only 5% labeled data, achieving clear improvements over the baseline on both KonIQ-10k and SPAQ. Moreover, its performance is very close to that obtained with full-set estimation, with only marginal gaps of 0.013 and 0.011 SRCC, respectively. These results indicate that the effectiveness of PreSTA does not rely on full target-set statistics, and that few-label PPR estimation is sufficient for practical deployment.

3. More Ablation Experiments

3.1. Comparison with Additional Baselines

We further compare PreSTA-S with two additional baselines: Random Sampling and KMM Reweighting [5]. Random Sampling selects a 20% source subset uniformly at random, which isolates the effectiveness of our preference-guided selection strategy from the mere reduction in training data size. KMM Reweighting operates on the full source dataset to align the marginal feature distributions between domains via importance weighting.

Table 3. Transfer performance (SRCC) under few-label target-PPR estimation for KADID-10k → KonIQ-10k and SPAQ. For PreSTA-S (5% Est.), the results are reported as mean ± std over 10 runs.

Method	KonIQ-10k	SPAQ	Data %
Baseline	0.682	0.805	100%
PreSTA-S (5% Est.)	0.761 ± 0.006	0.861 ± 0.003	20%
PreSTA-S (Full-set Est.)	**0.774**	**0.872**	20%

As shown in Tab. 4, under the same 20% data budget, PreSTA-S substantially outperforms Random Sampling by 0.103 SRCC on KonIQ-10k and 0.055 SRCC on SPAQ. Moreover, even with only 20% of the source data, PreSTA-S surpasses KMM Reweighting trained on the full source set. These results confirm that the gains of PreSTA arise from explicitly aligning perceptual preference structure, rather than from data reduction or conventional marginal distribution alignment.

3.2. Effect of Backbone Architecture

To verify whether the effectiveness of PreSTA depends on a specific backbone architecture, we compare Swin-B and ResNet-50 under the same cross-domain transfer setting (KADID-10k → KonIQ-10k, SPAQ). For each backbone, we report the full-data baseline and the corresponding PreSTA-S result using 20% selected source data.

Table 4. Comparison with additional baselines (SRCC) under KADID-10k → KonIQ-10k and SPAQ. Random Sampling results are reported as mean ± std over 10 runs.

Method	KonIQ-10k	SPAQ	Data %
Baseline	0.682	0.805	100%
Random Sampling	0.671 ± 0.008	0.817 ± 0.009	20%
KMM Reweighting	0.658	0.838	100%
PreSTA-S	**0.774**	**0.872**	20%

Table 5. Backbone ablation (SRCC) for PreSTA-S under KADID-10k → KonIQ-10k and SPAQ.

Backbone	Method	KonIQ-10k	SPAQ	Data %
ResNet-50	Baseline	0.625	0.758	100%
	PreSTA-S	0.755	0.853	20%
Swin-B	Baseline	0.682	0.805	100%
	PreSTA-S	**0.774**	**0.872**	20%

As shown in Tab. 5, PreSTA-S consistently improves over the baseline for both architectures. With Swin-B, PreSTA-S substantially improves the SRCC from 0.682 to 0.774 on KonIQ-10k and from 0.805 to 0.872 on SPAQ. Similar gains are observed with ResNet-50. These results indicate that the effectiveness of PreSTA is not tied to a specific backbone architecture. Since Swin-B achieves stronger overall performance, we adopt it as the default backbone.

3.3. Effect of Feature Layer

We further study the effect of feature selection for computing PPR. Using Swin-B as the backbone, we compare two settings: (1) using only the final-stage feature (Stage 4), and (2) using features aggregated from all four stages. In both cases, features are globally pooled and used to compute the PPR, and PreSTA-S selects 20% of the source data under the same protocol.

As shown in Tab. 6, using hierarchical features leads to substantially better performance on KonIQ-10k and comparable performance on SPAQ. This suggests that aggregating hierarchical features provides a more comprehensive characterization of dataset-level perceptual preference by jointly capturing low-level and high-level cues. Based on this observation, we adopt hierarchical features as the default in our method.

4. Runtime Analysis

We analyze the computational overhead of PreSTA under the synthetic-to-authentic transfer setting (KADID-10k → KonIQ-10k and SPAQ), with all results averaged over the two target datasets. PreSTA introduces an offline sample selection stage that takes approximately 43 minutes. However, since only 20% of the source data is retained, the subsequent training time is reduced by approximately 144 minutes. This results in a net saving of approximately 99 minutes overall. Despite the additional selection step, PreSTA reduces the total end-to-end runtime, improving both transfer performance and computational efficiency in practice.

Table 6. Layer-choice ablation (SRCC) for PPR computation under KADID-10k → KonIQ-10k and SPAQ using Swin-B.

Backbone	Feature Strategy	KonIQ-10k	SPAQ
Swin-B	Final only (Stage 4)	0.721	**0.856**
	All stages (Stage 1–4)	**0.755**	0.853

References

[1] Baoliang Chen, Haoliang Li, Hongfei Fan, and Shiqi Wang. No-reference screen content image quality assessment with unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 30:5463–5476, 2021. 1, 2

[2] Pengfei Chen, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Unsupervised curriculum domain adaptation for no-reference video quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5178–5187, 2021. 1, 2

[3] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 1, 2

[4] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1220–1230, 2022. 1, 2

[5] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19, 2006. 2

[6] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 1, 2

[7] Aobo Li, Jinjian Wu, Yongxu Liu, and Leida Li. Bridging the synthetic-to-authentic gap: Distortion-guided unsupervised domain adaptation for blind image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28422–28431, 2024. 1, 2

[8] Xin Li, Yiting Lu, and Zhibo Chen. Freqalign: Excavating perception-oriented transferability for blind image quality assessment from a frequency perspective. *IEEE Transactions on Multimedia*, 2023. 1, 2

[9] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1040–1049, 2017. 1, 2

[10] Yiting Lu, Xin Li, Jianzhao Liu, and Zhibo Chen. Styleam: Perception-oriented unsupervised domain adaption for no-reference image quality assessment. *IEEE Transactions on Multimedia*, 27:2043–2058, 2024. 1, 2

[11] Zhaoqing Pan, Feng Yuan, Jianjun Lei, Yuming Fang, Xiao Shao, and Sam Kwong. Vcrnet: Visual compensation restoration network for no-reference image quality assessment. *IEEE Transactions on Image Processing*, 31:1613–1627, 2022. 1, 2

[12] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3667–3676, 2020. 1, 2

[13] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 1, 2

[14] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, pages 54015–54029, 2024. 1, 2

[15] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018. 1, 2

[16] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021. 1, 2

[17] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14071–14081, 2023. 1, 2

[18] Tianwei Zhou, Songbai Tan, Baoquan Zhao, and Guanghui Yue. Multitask deep neural network with knowledge-guided attention for blind image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1, 2