RankDVQA: Deep VQA based on Ranking-inspired Hybrid Training Supplementary Material

Chen Feng, Duolikun Danier, Fan Zhang and David Bull
Visual Information Laboratory, University of Bristol, Bristol, UK, BS1 5DD
{chen.feng, duolikun.danier, fan.zhang, dave.bull}@bristol.ac.uk

A. Training Database Generation in Stage 1

To generate the training content, we used 230 HD source sequences from the BVI-DVC [16] and the CVPR 2022 CLIC video compression challenge [7] database. These sequences (and their re-sampled versions, see Section 3.1 in the main paper) were compressed using four different video codecs at four quantization levels to create diverse distortion types. The codec versions and configurations are summarized in Table 1.

B. Comprehensive Evaluation Results

In the main paper, we evaluated the performance of the proposed metrics (FR-RankDVQA and NR-RankDVQA) and 18 (10 for FR and 8 for NR) competing metrics using the Spearman Rank Order Correlation Coefficient (SROCC) and an F-test. Here, we additionally test the performance of these methods with the Pearson Linear Correlation Coefficient (PLCC), the Outlier Ratio (OR) and the Root Mean Squared Error (RMSE). Furthermore, we benchmark our approaches against three additional deep learning-based metrics (WaDIQA [1], VSFA [13], CONVIQT [18] (unsupervised learning) and RAPIQUE [24]). The full SROCC, PLCC, RMSE and OR results are summarized in Table 2, 3, 4 and 5 respectively. It can be observed that the proposed FR-RankDVQA and NR-RankDVQA show the best overall performance compared to other quality metrics in each track (FR and NR).

C. Full Complexity Figures

Table 2 (Complexity column) also presents the complexity figures of proposed metrics and other benchmark metrics. It is noted tht the average time required by FR-RankDVQA to process one HD frame (1920 $\times$ 1080) is 3.88 times that of VMAF, similar to other under-performing deep FR-VQA methods, such as DeepVQA [11] (4.05$\times$) and C3DVQA [27] (3.37$\times$). The runtime was estimated using a computer with a 2.4GHz Intel CPU and an NVIDIA P100 graphic card.

D. Detailed Quantitative Evaluation on Single Sources

As we discussed in the main paper, it is important for any VQA method to accurately differentiate the quality between distorted versions of the same source or from different sources. Due to limited space, we only summarized the overall SROCC values for single sources, but did not provide full quantitative results for each database. Here we report the single-source SROCC values for all tested methods, following the calculation procedures in [9]. Specifically, the SROCC value is first calculated based on all the distorted sequences for each source content, and these SROCC values are averaged among all sources for each database and reported in Table 6.

E. Additional Ablation Study Results

Effectiveness of the residual patch. In the main paper, we only presented key ablation study results where the primary contributions of the proposed methods including PQANet, STANet and the ranking-inspired losses are evaluated. To further verify the contribution of the residual patch in Stage 1 (see Section 3.1 of main paper), which was not included in the [9], we removed the residual patch calculation step from the proposed PQANet, and concatenated features $[\mathcal{F}_D\ \mathcal{F}_{\text{diff}}\ \mathcal{F}_R]$ after the Feature Extraction (FE) Module, where the $\mathcal{F}_{\text{diff}}$ is the difference feature maps between two normalized features $\mathcal{F}_D$ and $\mathcal{F}_R$.

$$\mathcal{F}_{\text{diff}} = \| \mathrm{f}_{norm}(\mathcal{F}_D) - \mathrm{f}_{norm}(\mathcal{F}_R) \|_1 . \tag{1}$$

This variant, denoted as V5 (Res), is similar to that in [9], and its performance is reported in Table 7 alongside other ablation study results. By comparing this variant to FR-RankDVQA, we can observe higher SROCC values for the latter on all tested video databases. This confirms the effectiveness of using residual frames in Stage 1.

Codec	Version	Configuration parameters
H.264/AVC	x264-r3094-bfc87b7	Preset veryslow configuration [6]. QP=[32,37,42,47]
HEVC HM	16.20	Random access configuration [22]. QP=[32,37,42,47]
AOM AV1	1.0.0-5ec3e8c	The same configuration as in [28]. QP=[32,43,55,63]
VVC VTM	7.0	Random access configuration [2]. QP=[32,37,42,47]

Table 1. The employed video codecs and configurations.

SROCC↑ (F-test)	NFLX	NFLX-P	BVI-HD	BVI-CCHD	BVI-CCHDDO	MCL-V	SHVC	VQEGHD3	Overall	Complexity
Full Reference VQA Methods										
PSNR	0.6218 (-1)	0.6596 (-1)	0.6143 (-1)	0.6166 (-1)	0.7497 (-1)	0.4640 (-1)	0.7380 (-1)	0.7518 (-1)	0.6520	0.35×
SSIM [25]	0.5638 (-1)	0.6054 (-1)	0.5992 (-1)	0.7194 (-1)	0.8026 (-1)	0.4018 (-1)	0.5446 (-1)	0.7361 (-1)	0.6216	0.38×
MS-SSIM [26]	0.7136 (-1)	0.7394 (-1)	0.7652 (-1)	0.7534 (-1)	0.8321 (0)	0.6306 (-1)	0.8007 (0)	0.8457 (-1)	0.7601	0.47×
WaDIQA [1]	0.5713 (-1)	0.6593 (-1)	0.6646 (-1)	0.6516 (-1)	0.7041 (-1)	0.6072 (-1)	0.6731 (-1)	0.6910 (-1)	0.6528	1.29×
DeepQA [10]	0.7298 (-1)	0.6995 (-1)	0.7106 (-1)	0.6202 (-1)	0.6705 (-1)	0.6832 (-1)	0.7176 (-1)	0.7881 (-1)	0.7024	1.76×
LPIPS [29]	0.6793(-1)	0.7859 (-1)	0.6670 (-1)	0.6838 (-1)	0.7678 (-1)	0.6579 (-1)	0.6360 (-1)	0.8075 (0)	0.7107	2.03×
DeepVQA [11]	0.7352 (-1)	0.7609 (-1)	0.7330 (-1)	0.6924 (-1)	0.8120 (0)	0.6142 (-1)	0.8041 (0)	0.7805 (-1)	0.7540	4.05×
C3DVQA [27]	0.7730 (-1)	0.7714 (-1)	0.7393 (-1)	0.7203 (-1)	0.8137 (0)	0.7126 (-1)	0.8194 (0)	0.7329 (-1)	0.7641	3.37×
DISTS [5]	0.7787 (-1)	0.9325 (0)	0.7030 (-1)	0.6303 (-1)	0.7442 (-1)	0.7792 (-1)	0.7813 (0)	0.8254 (0)	0.7718	2.28×
ST-GREED [17]	0.7470 (-1)	0.7445 (-1)	0.7769 (-1)	0.7738 (-1)	0.8259 (0)	0.7226 (-1)	0.7946 (0)	0.8079 (-1)	0.7842	0.83×
VMAF 0.6.1 [15]	0.9254 (0)	0.9104 (0)	0.7962 (-1)	0.8723 (0)	0.8783 (0)	0.7766 (-1)	0.9114 (0)	0.8442 (-1)	0.8644	1×
FR-RankDVQA	0.9393	0.9184	0.8659	0.8991	0.9037	0.8391	0.9142	0.8979	**0.8972**	3.88×
No reference VQA methods										
VIIDEO [20]	0.4550 (-1)	0.5527 (-1)	0.1297 (-1)	0.1308 (-1)	0.2523 (-1)	0.0406 (-1)	0.2033 (-1)	0.1881 (-1)	0.2440	5.61×
TLVQM [12]	0.4652 (-1)	0.4720 (-1)	0.3124 (-1)	0.1622 (-1)	0.3420 (-1)	0.2758 (-1)	0.4983 (0)	0.5382 (-1)	0.3469	1.68×
VSFA [13]	0.6221 (-1)	0.6993 (-1)	0.2118 (-1)	0.3480 (-1)	0.3876 (-1)	0.2978 (-1)	0.3031 (-1)	0.5350 (-1)	0.4256	11.2×
BRISQUE [19]	0.7828 (0)	0.7861 (0)	0.2033 (-1)	0.3738 (-1)	0.3746 (-1)	0.3154 (-1)	0.3601 (-1)	0.5467 (-1)	0.4716	0.06×
NIQE [21]	0.7959 (0)	0.8269 (0)	0.1932 (-1)	0.4247 (-1)	0.5225 (-1)	0.3985 (-1)	0.6210 (0)	0.5291 (-1)	0.5390	0.12×
MDTVSFA [14]	0.5137 (-1)	0.6024 (-1)	0.3725 (-1)	0.4068 (-1)	0.5547 (-1)	0.5712 (0)	0.6165 (0)	0.6422 (0)	0.5311	6.98×
CONVIQT [18]	0.6989 (-1)	0.7962 (0)	0.3489 (-1)	0.3706 (-1)	0.5381(-1)	0.6323 (0)	0.4983 (0)	0.6217 (0)	0.5631	0.98×
RAPIQUE [24]	0.6178 (-1)	0.6962 (-1)	0.5690 (-1)	0.6024 (-1)	0.5951 (-1)	0.5392 (-1)	0.5477 (0)	0.7240 (0)	0.6114	0.34×
VIDEVAL [23]	0.7899 (0)	0.7261 (0)	0.5884 (-1)	0.6974 (0)	0.7620 (0)	0.4836 (-1)	0.6428 (0)	0.5326 (-1)	0.6529	2.54×
GSTVQA [4]	0.8109 (0)	0.7858 (0)	0.4132 (-1)	0.7447 (0)	0.7665 (0)	0.7385 (0)	0.6710 (0)	0.7011 (0)	0.7040	2.89×
NR-RankDVQA	0.8346	0.7944	0.7326	0.7628	0.7994	0.7631	0.7118	0.8346	**0.7791**	3.76×

Table 2. The SROCC performance and relative complexity of the proposed methods and other benchmark approaches on eight HD test databases. The values in each cell x(y) correspond to the SROCC value (x) and F-test result (y) at 95% confidence interval. y=1 suggests that the metric is superior to FR-RankDVQA in the full reference track or NF-RankDVQA in the no reference track (y=-1 if the opposite is true), while y=0 indicates that there is no significant difference between them. The figures in red and blue indicate the highest and second highest SROCC values respectively in each column. The complexity column presents the run time of the proposed metrics and the competing metrics.

F. Additional Subjective Examples

In the main paper, we only presented two visual examples. Here, in Figure 1, we have included additional subjective examples that further demonstrate the accurate quality prediction performance of the proposed method.

G. Limitations of the proposed method

Despite the excellent performance achieved by the proposed approach on eight HD test databases, our focus in this work is solely on full HD compressed video content. The performance of this method on higher resolution (4K, 8K), higher frame rate (120fps, 240fps) or different video for-

PLCC↑	NFLX	NFLX-P	BVI-HD	BVI-CCHD	BVI-CCHDDO	MCL-V	SHVC	VQEGHD3	Overall
Full Reference VQA Methods									
PSNR	0.6436	0.7065	0.6238	0.6355	0.7488	0.4707	0.7370	0.7459	0.6640
SSIM [25]	0.5584	0.6368	0.5969	0.6752	0.7659	0.4033	0.4879	0.7066	0.6039
MS-SSIM [26]	0.6879	0.7534	0.6840	0.7155	0.7285	0.6424	0.7236	0.7736	0.7136
WaDIQA [1]	0.7683	0.7842	0.7427	0.6488	0.7127	0.5954	0.5053	0.6972	0.6818
DeepQA [10]	0.5943	0.6599	0.6540	0.6882	0.7557	0.5011	0.7350	0.7160	0.6630
LPIPS [29]	0.6348	0.7921	0.6541	0.6875	0.7582	0.6460	0.6071	0.7644	0.6930
DeepVQA [11]	0.8092	0.7005	0.7489	0.6754	0.7491	0.6097	0.8109	0.7640	0.7335
C3DVQA [27]	0.7633	0.7959	0.7830	0.7649	0.8462	0.7274	0.8398	0.7557	0.7845
DISTS [5]	0.7643	0.9195	0.6595	0.7038	0.7683	0.7718	0.8159	0.8020	0.7756
ST-GREED [17]	0.7239	0.8186	0.7960	0.7562	0.8193	0.6943	0.7814	0.7975	0.7734
VMAF 0.6.1 [15]	0.9333	0.9356	0.7921	0.8476	0.8840	0.7781	0.9379	0.8670	0.8720
FR-RankDVQA	0.9295	0.9421	0.8817	0.8910	0.8728	0.8339	0.9065	0.9282	0.8982
No reference VQA methods									
VIIDEO [20]	0.5608	0.6550	0.1695	0.2846	0.4272	0.0251	0.4745	0.1864	0.3479
TLVQM [12]	0.3898	0.3656	0.3357	0.3597	0.5380	0.4049	0.4215	0.3009	0.3895
VSFA [13]	0.5421	0.6401	0.2371	0.2612	0.3171	0.3417	0.4067	0.4942	0.4050
BRISQUE [19]	0.7872	0.8551	0.2941	0.3135	0.4811	0.3212	0.4757	0.5788	0.5133
NIQE [21]	0.8161	0.8890	0.2087	0.5293	0.5789	0.3955	0.5461	0.5366	0.5625
MDTVSFA [14]	0.5847	0.5935	0.3526	0.5337	0.5368	0.5075	0.5751	0.6469	0.5413
CONVIQT [18]	0.7447	0.8442	0.3589	0.5102	0.5880	0.6136	0.5993	0.6430	0.6141
RAPIQUE [24]	0.6952	0.7065	0.5957	0.6207	0.6153	0.6008	0.5573	0.7082	0.6375
VIDEVAL [23]	0.7568	0.7447	0.6338	0.6706	0.7424	0.5294	0.5979	0.5362	0.6515
GSTVQA [4]	0.7896	0.7332	0.5569	0.6921	0.7131	0.7504	0.6319	0.6930	0.6950
NR-RankDVQA	0.8269	0.8371	0.7083	0.7175	0.7859	0.7850	0.6936	0.7504	0.7631

Table 3. The PLCC performance of the proposed methods and other benchmark approaches on eight HD test databases. The figures in red and blue indicate the highest and second highest PLCC values respectively in each column.

mats (HDR or UGC content) has not been not fully investigated in this paper. As future work, it would be interesting to see how the proposed method generalizes to these content types. We are also aware of the relatively high complexity associated with this model, which leads to both slower inference speed and higher GPU memory usage. This is mainly due to the large size of the neural network involved. Various techniques such as model compression [3] and knowledge distillation [8], can be employed to reduce the model complexity and the number of network parameters. This will also be our future work.

OR↓	NFLX	NFLX-P	BVI-HD	BVI-CCHD	BVI-CCHDDO	MCL-V	SHVC	VQEGHD3	Overall
Full Reference VQA Methods									
PSNR	-	-	0.5677	0.5556	0.4000	0.6458	-	0.5278	0.5394
SSIM [25]	-	-	0.5573	0.6389	0.6889	0.6875	-	0.8611	0.6867
MS-SSIM [26]	-	-	0.7292	0.6389	0.6889	0.7292	-	0.8611	0.7294
WaDIQA [1]	-	-	0.7635	0.6389	0.6889	0.5312	-	0.8611	0.6967
DeepQA [10]	-	-	0.7329	0.6365	0.6256	0.7250	-	0.7844	0.7009
LPIPS [29]	-	-	0.7292	0.6389	0.6889	0.7292	-	0.8611	0.7294
DeepVQA [11]	-	-	0.4583	0.4167	0.3889	0.7083	-	0.8611	0.5667
C3DVQA [27]	-	-	0.3958	0.2222	0.3222	0.5938	-	0.4028	0.3874
DISTS [5]	-	-	0.7292	0.6389	0.6889	0.7292	-	0.8611	0.7294
ST-GREED [17]	-	-	0.4375	0.4444	0.4444	0.7292	-	0.6667	0.5444
VMAF 0.6.1 [15]	-	-	0.3958	0.2870	0.2556	0.4479	-	0.3472	0.3467
FR-RankDVQA	-	-	0.3750	0.2222	0.2333	0.4062	-	0.2778	0.3029
No reference VQA methods									
VIIDEO [20]	-	-	0.7188	0.6204	0.6222	0.7292	-	0.8611	0.7103
TLVQM [12]	-	-	0.6979	0.6463	0.6556	0.6146	-	0.8194	0.6868
VSFA [13]	-	-	0.8096	0.6019	0.6333	0.7083	-	0.7778	0.7062
BRISQUE [19]	-	-	0.6771	0.6389	0.5667	0.6875	-	0.7222	0.6585
NIQE [21]	-	-	0.7188	0.5833	0.5556	0.6146	-	0.6111	0.6167
MDTVSFA [14]	-	-	0.7889	0.6111	0.6000	0.6883	-	0.7778	0.6932
CONVIQT [18]	-	-	0.6406	0.5556	0.5111	0.5833	-	0.5417	0.5665
RAPIQUE [24]	-	-	0.7563	0.5926	0.5778	0.6875	-	0.7639	0.6756
VIDEVAL [23]	-	-	0.6375	0.5764	0.5611	0.6755	-	0.7275	0.6356
GSTVQA [4]	-	-	0.6106	0.5796	0.5463	0.6317	-	0.7537	0.6244
NR-RankDVQA	-	-	0.5208	0.5370	0.5111	0.5312	-	0.6250	0.5450

Table 4. The outlier ratio (OR) performance of the proposed methods and other benchmark approaches on eight HD test databases. The figures in red and blue indicate the lowest and second lowest OR values respectively in each column. On NFLX, NFLX-P and SHVC databases, OR values cannot be calculated as the standard error for each subjective score is not available.

RMSE↓	NFLX	NFLX-P	BVI-HD	BVI-CCHD	BVI-CCHDDO	MCL-V	SHVC	VQEGHD3
Full Reference VQA Methods								
PSNR	0.8711	18.1675	13.4933	12.4575	13.2023	1.9722	1.8148	0.6743
SSIM [25]	0.9442	19.7931	13.8163	16.9298	20.4117	2.0436	2.6850	1.0116
MS-SSIM [26]	1.1382	25.6703	17.3173	16.9304	20.4118	2.2414	2.6850	0.6713
WaDIQA [1]	0.7286	15.9273	11.5064	16.9820	20.9321	2.2414	2.3170	1.0234
DeepQA [10]	1.1055	24.8074	16.6020	16.2461	19.5864	2.2213	2.6670	0.9887
LPIPS [29]	1.1382	25.6703	17.3173	16.9298	20.4115	2.2414	2.6850	0.6696
DeepVQA [11]	0.6687	18.3192	11.3789	16.8554	19.0823	2.2030	1.7713	0.6682
C3DVQA [27]	0.7353	15.5421	10.9109	13.5351	13.4093	2.0251	1.5713	0.6713
DISTS [5]	1.1382	11.6703	17.3173	16.9298	20.4117	2.2414	1.7850	0.5816
ST-GREED [17]	0.8478	15.6804	10.2283	13.7927	16.9766	2.0709	2.4554	0.5852
VMAF 0.6.1 [15]	0.4086	9.0598	10.4955	7.9969	9.1395	1.3996	0.9316	0.5057
FR-RankDVQA	0.4140	8.8504	8.5652	7.7287	9.1992	1.3063	0.9348	0.4895
No reference VQA methods								
VIIDEO [20]	0.9424	19.3972	17.1016	16.0882	18.5468	2.2414	2.3639	1.0116
TLVQM [12]	1.0482	23.8930	16.2426	16.0748	17.4833	2.0519	2.4348	0.9691
VSFA [13]	1.0071	23.0343	17.0335	17.3363	19.3548	2.1744	2.5706	1.6880
BRISQUE [19]	0.7020	13.3095	16.5372	16.2730	17.6206	2.1287	2.3618	0.8295
NIQE [21]	0.6578	11.7546	16.9631	14.1962	16.3204	2.0614	2.6850	0.8536
MDTVSFA [14]	0.9587	21.9299	16.0909	14.0597	18.4092	2.1440	2.5257	0.9422
CONVIQT [18]	0.7441	12.6833	16.0597	13.2822	16.2269	1.7407	2.6170	0.9873
RAPIQUE [24]	0.9370	21.7140	14.9286	16.7457	17.6228	2.1307	2.5265	0.8116
VIDEVAL [23]	0.7219	21.5357	14.8032	16.6525	16.1987	2.0668	2.3504	0.9164
GSTVQA [4]	0.7031	20.9120	15.6097	13.1453	17.0783	1.7349	2.4956	0.8942
NR-RankDVQA	0.6205	12.9074	13.6189	13.0835	16.0654	1.7385	2.2541	0.8052

Table 5. The root mean squared error (RMSE) performance of the proposed methods and other benchmark approaches on eight HD test databases. The figures in red and blue indicate the lowest and second lowest OR values respectively in each column.

SROCC↑ (SS)	NFLX	NFLX-P	BVI-HD	BVI-CCHD	BVI-CCHDDO	MCL-V	SHVC	VQEGHD3	Overall (SS)
Full Reference VQA Methods									
PSNR	0.9787	0.9866	0.8893	0.9384	0.9542	0.9722	0.9286	0.9325	0.9476
SSIM [25]	0.9798	0.9802	0.8893	0.9407	0.9407	0.9742	0.9286	0.9272	0.9451
MS-SSIM [26]	0.9852	0.9866	0.8893	0.9407	0.9327	0.9861	0.9286	0.9325	0.9477
WaDIQA [1]	0.9150	0.9403	0.8791	0.8858	0.9360	0.8627	0.8765	0.8514	0.8934
DeepQA [10]	0.9064	0.8974	0.8464	0.8906	0.9088	0.8944	0.8594	0.8799	0.8854
LPIPS [29]	0.9852	0.9866	0.8734	0.9291	0.9569	0.9325	0.9345	0.9325	0.9413
DeepVQA [11]	0.9205	0.9237	0.8904	0.8902	0.9372	0.8994	0.8891	0.8950	0.9060
C3DVQA [27]	0.9754	0.9866	0.8782	0.9291	0.9475	0.9881	0.9167	0.9325	0.9443
DISTS [5]	0.9852	0.9866	0.8680	0.9345	0.9609	0.9762	0.9167	0.9299	0.9447
ST-GREED [17]	0.9852	0.9866	0.8875	0.9338	0.9340	0.9802	0.9286	0.9325	0.9460
VMAF 0.6.1 [15]	0.9852	0.9866	0.8893	0.9431	0.9367	0.9683	0.9226	0.9325	0.9455
FR-RankDVQA	0.9896	0.9966	0.9492	0.9837	0.9788	0.9861	0.9786	0.9884	**0.9814**
No reference VQA methods									
VIIDEO [20]	0.6877	0.6662	0.2656	0.3260	0.3145	0.0317	0.2024	0.4245	0.3648
TLVQM [12]	0.8405	0.8437	0.8104	0.8102	0.8572	0.8094	0.8051	0.815	0.8239
VSFA [13]	0.5117	0.5953	0.3844	0.8047	0.6714	0.8036	0.2679	0.7976	0.6046
BRISQUE [19]	0.9825	0.9847	0.6792	0.8627	0.7980	0.7560	0.5774	0.9417	0.8236
NIQE [21]	0.9342	0.9743	0.7668	0.8114	0.9044	0.7202	0.8750	0.9510	0.8672
MDTVSFA [14]	0.9785	0.9847	0.8875	0.8786	0.9340	0.6607	0.8333	0.9404	0.8872
CONVIQT [18]	0.7611	0.9494	0.7488	0.7562	0.8168	0.9037	0.8571	0.8596	0.8322
RAPIQUE [24]	0.9820	0.9847	0.8262	0.8606	0.8963	0.6687	0.9048	0.9378	0.8877
VIDEVAL [23]	0.8524	0.8664	0.8345	0.8527	0.9114	0.8154	0.9022	0.8619	0.8621
GSTVQA [4]	0.9101	0.9105	0.8776	0.8795	0.9470	0.8608	0.9145	0.9113	0.9014
NR-RankDVQA	0.9825	0.9850	0.8897	0.8616	0.9179	0.9048	0.9286	0.9425	**0.9266**

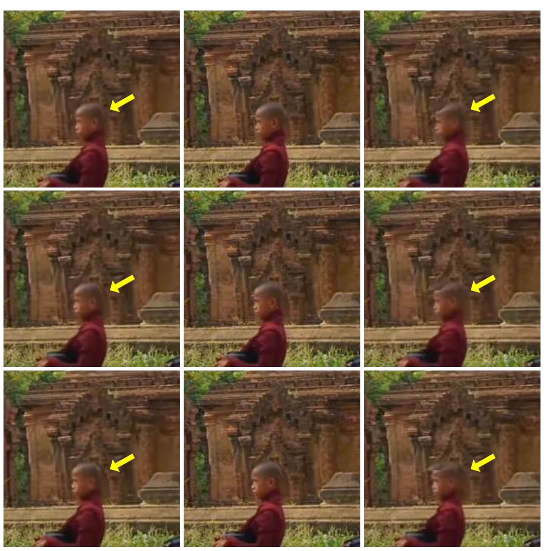
Table 6. The single source based SROCC value for the proposed methods and other benchmark approaches on eight HD test databases. The figures in red indicate the highest SROCC values in each column.

SROCC↑	NFLX	NFLX-P	BVI-HD	BVI-CCHD	BVI-CCHDDO	MCL-V	SHVC	VQEGHD3	Overall
Ablation Study Results									
V1 ($\ell 1$)	0.8793	0.8816	0.7583	0.7792	0.8523	0.7678	0.8238	0.8501	0.8190
V2 ($\ell 2$)	0.8812	0.8883	0.7612	0.7794	0.8568	0.7696	0.8234	0.8507	0.8263
V3 (C3D)	0.9034	0.8964	0.8233	0.8763	0.8961	0.8054	0.8692	0.8465	0.8653
V4 (S1)	0.9201	0.8983	0.8361	0.8825	0.8987	0.8231	0.8966	0.8544	0.8762
V5 (Res)	0.9002	0.8803	0.8203	0.8963	0.8982	0.8354	0.8312	0.8765	0.8673
FR-RankDVQA	0.9393	0.9184	0.8659	0.8991	0.9037	0.8391	0.9142	0.8979	**0.8972**

Table 7. Comprehensive ablation study results in terms of SROCC on eight HD test databases.

Distorted A		Reference		Distorted B
56.58	√	Human		51.10
64.71		VMAF	√	66.19
67.49	√	RankDVQA		60.08

Distorted A		Reference A	Reference B		Distorted B
48.43		Human	Human	√	72.65
74.4415	√	VMAF	VMAF		74.42
61.77		RankDVQA	RankDVQA	√	76.78



Distorted A		Reference		Distorted B
77.57	√	Human		61.24
84.83		VMAF	√	87.62
79.69	√	RankDVQA		70.55

Distorted A		Reference A	Reference B		Distorted B
50.00		Human	Human	√	82.50
69.56	√	VMAF	VMAF		60.37
68.41		RankDVQA	RankDVQA	√	74.98

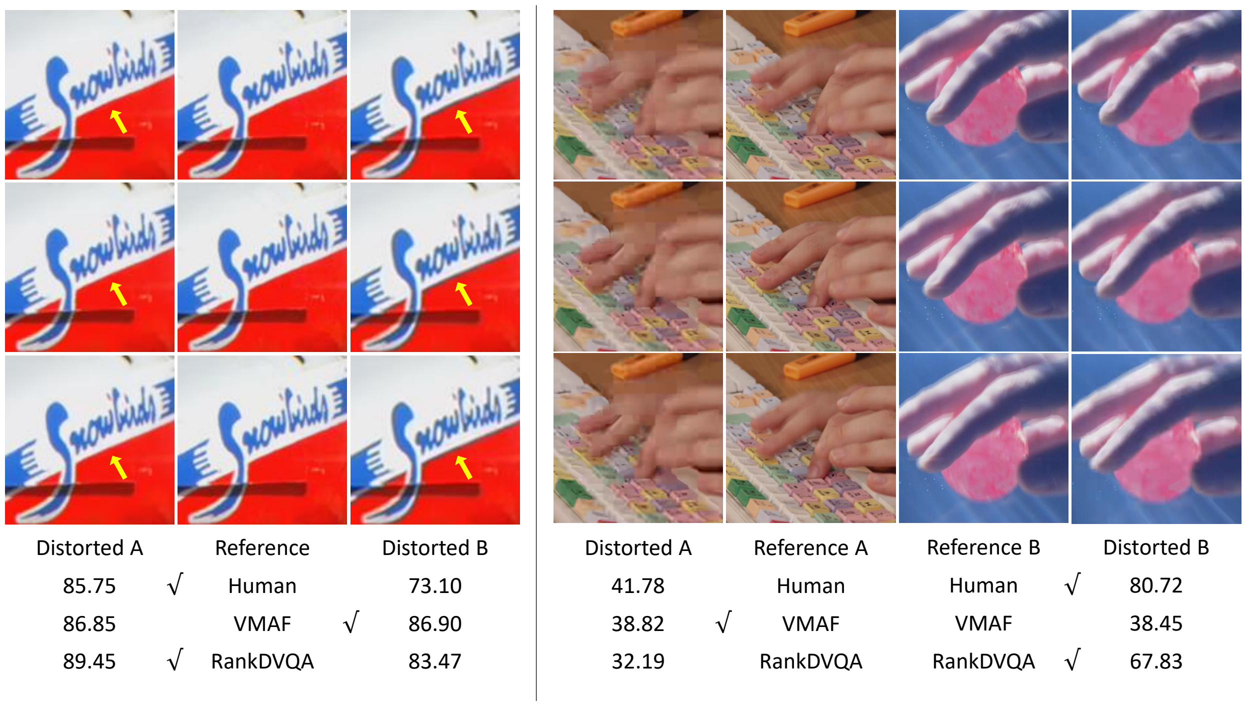


Figure 1. Additional visual examples demonstrating the superiority of the proposed FR quality metric. From top to bottom, the three images in each column correspond to three consecutive frames in the same sequence. The left frames are from single-source, and the right frames are from dual-source. FR-RankDVQA correctly predicts a higher quality index for the distorted video with a higher subjective quality score.

References

[1] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, 2018. 1, 2, 3, 4, 5, 6

[2] F. Bossen, J. Boyce, X. Li, V. Seregin, and K. Sühring. JVET common test conditions and software reference configurations for SDR video. In *the JVET Meetings, JVET-M1001*. ITU-T and ISO/IEC, 2019. 2

[3] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006. 3

[4] Baoliang Chen, Lingyu Zhu, Guo Li, Fangbo Lu, Hongfei Fan, and Shiqi Wang. Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):1903–1916, 2022. 2, 3, 4, 5, 6

[5] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020. 2, 3, 4, 5, 6

[6] Advanced Encoding Guide. x264 general settings. *https://silentaperture.gitlab.io/mdbook-guide/encoding/x264.html/*, 2011. 2

[7] Advanced Encoding Guide. the fifth workshop and challenge on learned image compression (video track). *http://compression.cc/*, 2022. 1

[8] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 3

[9] Qiqi Hou, Abhijay Ghildyal, and Feng Liu. A perceptual quality metric for video frame interpolation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 234–253. Springer, 2022. 1

[10] Jongyoo Kim and Sanghoon Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1969–1977, 2017. 2, 3, 4, 5, 6

[11] Woojae Kim, Jongyoo Kim, Sewoong Ahn, Jinwoo Kim, and Sanghoon Lee. Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 219–234, 2018. 1, 2, 3, 4, 5, 6

[12] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019. 2, 3, 4, 5, 6

[13] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2351–2359, 2019. 1, 2, 3, 4, 5, 6

[14] Dingquan Li, Tingting Jiang, and Ming Jiang. Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision*, 129:1238–1257, 2021. 2, 3, 4, 5, 6

[15] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 2016. 2, 3, 4, 5, 6

[16] Di Ma, Fan Zhang, and David Bull. BVI-DVC: a training database for deep video compression. *IEEE Transactions on Multimedia*, 2021. 1

[17] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. St-greed: Space-time generalized entropic differences for frame rate dependent video quality prediction. *IEEE Transactions on Image Processing*, 30:7446–7457, 2021. 2, 3, 4, 5, 6

[18] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Conviqt: Contrastive video quality estimator, 2022. 1, 2, 3, 4, 5, 6

[19] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 2, 3, 4, 5, 6

[20] Anish Mittal, Michele A Saad, and Alan C Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, 25(1):289–300, 2015. 2, 3, 4, 5, 6

[21] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 2, 3, 4, 5, 6

[22] K. Sharman and K. Suehring. Common test conditions for hm video coding experiments. In *the JCT-VC meeting, JCTVC-AF1100*. ITU-T, ISO/IEC, 2018. 2

[23] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. UGC-VQA: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing*, 30:4449–4464, 2021. 2, 3, 4, 5, 6

[24] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Rapique: Rapid and accurate video quality prediction of user generated content. *IEEE Open Journal of Signal Processing*, 2:425–440, 2021. 1, 2, 3, 4, 5, 6

[25] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 2, 3, 4, 5, 6

[26] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multi-scale structural similarity for image quality assessment. In *Proc. Asilomar Conference on Signals, Systems and Computers*, volume 2, page 1398. IEEE, 2003. 2, 3, 4, 5, 6

[27] Munan Xu, Junming Chen, Haiqiang Wang, Shan Liu, Ge Li, and Zhiqiang Bai. C3DVQA: Full-reference video quality assessment with 3d convolutional neural network. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4447–4451, 2020. 1, 2, 3, 4, 5, 6

[28] Fan Zhang, Angeliki V Katsenou, Mariana Afonso, Goce Dimitrov, and David R Bull. Comparing VVC, HEVC and AV1 using objective and subjective assessments. *arXiv preprint arXiv:2003.10282*, 2020. 2

[29] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3, 4, 5, 6