RankIQA: Learning from Rankings for No-reference Image Quality Assessment Supplementary Material

Xialei Liu
Computer Vision Center
Barcelona, Spain
xialei@cvc.uab.es

Joost van de Weijer
Computer Vision Center
Barcelona, Spain
joost@cvc.uab.es

Andrew D. Bagdanov
MICC, University of Florence
Florence, Italy
andrew.bagdanov@unifi.it

1. Introduction

In this supplementary material we supply more comprehensive details about additional tables and graphs related to the main experiments, results on additional IQA datasets, and ranked dataset generation.

2. Additional experimental results

Here we provide insight into the ability of our RankIQA approach to learn to discriminate distortions, results illustrating the convergence properties of our fast Siamese backpropagation technique, and experimental results on additional IQA datasets.

Learning from Rankings and IQA discrimination. We trained our network on the Places2 [12] dataset until convergence, but performed *no fine-tuning* on IQA scores. We then plot as histograms the output of our Siamese network on images from the Waterloo [4] dataset distorted with four different distortions as shown in Fig. 1. In the plot, we divide the observations according to the true distortion level (indicated by the color of the histogram). The model discriminates different levels of distortions on Waterloo, even though the acquisition process and the scenes of the two datasets are totally different.

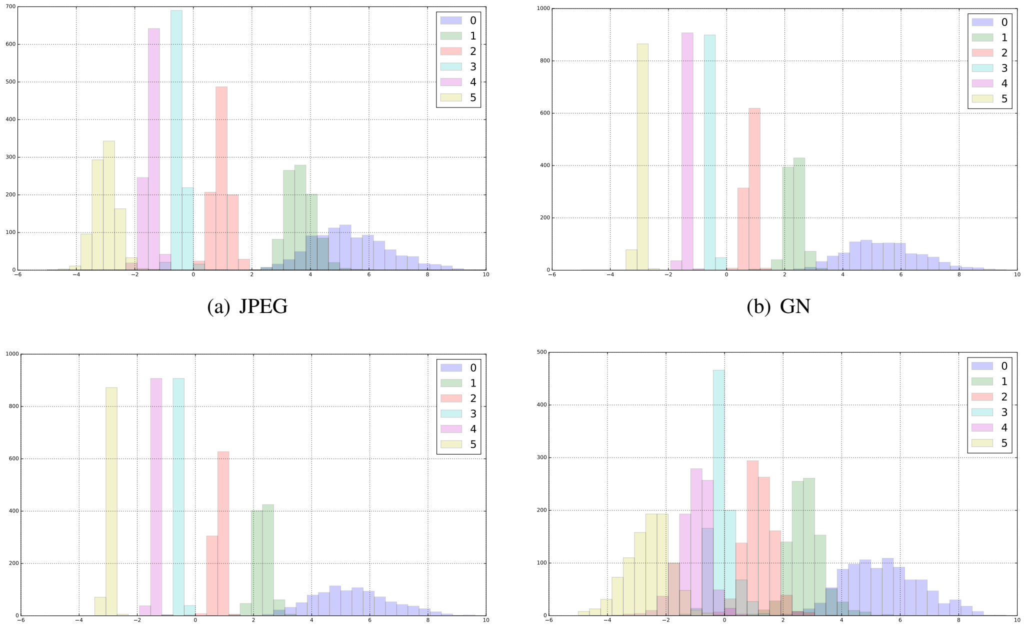
Efficient Siamese backpropagation. We compare our method to both standard random pair sampling, and a hard-negative mining method similar to [10][1]. The comparison of convergence rate on JP2K, JPEG, GB and GN is shown in Fig. 2. For all four distortions, the efficient Siamese backpropagation not only converges much faster, but also converges to a considerably lower loss depending on how difficult is the specific distortion. It is notable that for the easiest distortion GN, our method converges fast and obtains slightly better performance than hard-negative mining method in the end, however for other three relatively difficult distortions, our method achieves a significant improvement. In addition, we train four distortions jointly on entire LIVE [9] dataset. The comparison of convergence rate for three methods is shown in Fig. 3. The same conclusion as individual distortion can be drawn.

Baseline performance analysis. The performance evaluation (LCC) on the entire TID2013 [7] database is shown in Table 1. Our RankIQA method achieves superior results on almost all individual distortions even without ever using the TID2013 dataset. However, performance decreases for ALL distortions, which is because of nonlinear relationship between predicted and ground truth scores that is impossible to capture without fine-tuning on TID2013. After fine-tuning on the TID2013 database (RankIQA+FT), we considerably improve the ALL score, and improve the baseline by 13% on LCC. However, in the fine-tuning process to optimize the ALL score the network balances the various distortions, and this results in a decrease in performance for several individual distortions.

Evaluation on CSIQ [3]. We compare the performance of our method using the VGG-16 network with state-of-the-art methods. Four distortion types from CSIQ are used in this experiments (the distortions shared with the LIVE dataset: JPEG, JP2K, GB and GN). We randomly split the reference images and distorted versions from LIVE into 80% training samples and 20% testing, and compute the average LCC and SROCC scores on the testing set after training to convergence. This process is repeated ten times and the results are averaged. These results are shown in Table 2. Note that our RankIQA approach achieves comparable results compared to other start-of-the-art methods even without having access to the CSIQ dataset. Especially for SROCC, our method is superior than others except HOSA. After fine-tuning on CSIQ, we obtain about 1% higher on LCC and about 2% higher on SROCC than the state-of-the-art methods.

Evaluation on MLIVE [2]. In this experiment, We compare the performance of our method using the VGG-16 network with state-of-the-art methods. MLIVE is randomly split into 80% training samples and 20% testing, and the

[1] We experimented with several hard-negative mining methods and found this to obtain the best results.



(a) JPEG

(b) GN

(c) GB

(d) JP2K

Figure 1. Siamese network output for JPEG, GN, GB, and JP2K distortions at 6 different levels. These graphs illustrate that the Siamese network successfully manages to separate the different distortion levels, even without fine-tuning on IQA scores. (Corresponding to Figure 2 in main submission.)

Method	#01	#02	#03	#04	#05	#06	#07	#08	#09	#10	#11	#12	#13
Baseline	0.547	0.452	0.715	0.162	0.704	0.535	0.503	0.763	0.683	0.827	0.817	0.598	0.666
RankIQA	**0.883**	**0.852**	**0.906**	**0.725**	**0.919**	**0.839**	**0.904**	0.810	**0.897**	**0.955**	**0.937**	0.671	0.426
RankIQA+FT	0.652	0.588	0.796	0.326	0.780	0.703	0.776	**0.811**	0.819	0.894	0.894	**0.755**	**0.798**
Method	#14	#15	#16	#17	#18	#19	#20	#21	#22	#23	#24	ALL	
Baseline	0.170	0.322	0.046	0.500	0.370	0.466	0.375	0.660	0.614	0.793	0.773	0.663	
RankIQA	**0.484**	**0.639**	**0.369**	**0.665**	0.591	**0.833**	0.622	**0.875**	**0.806**	**0.891**	0.750	0.566	
RankIQA+FT	0.472	0.626	0.260	0.628	**0.629**	0.593	**0.661**	0.798	0.782	0.834	**0.874**	**0.799**	

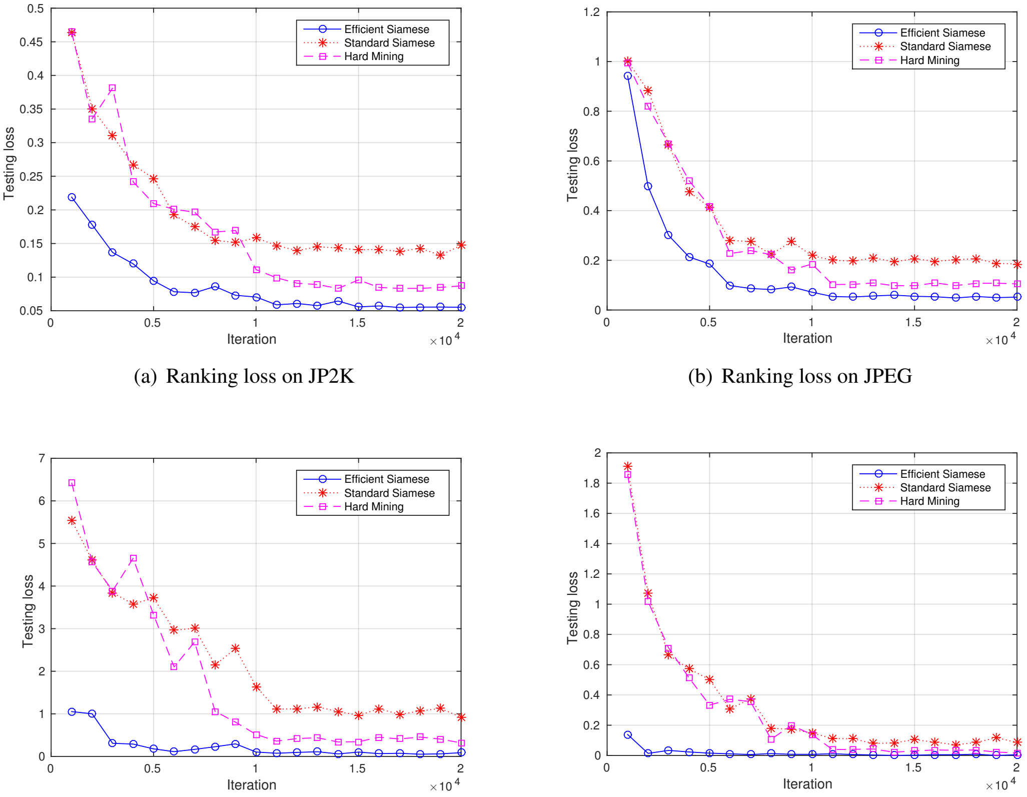
Table 1. Performance evaluation (LCC) on the entire TID2013 database. The Baseline approach fine-tunes an ImageNet-trained network on TID2013 data. Our RankIQA approach fine-tunes an ImageNet-trained network using only ranked images, and RankIQA+FT is our learning-from-ranking approach further fine-tuned on TID2013 data. (Corresponding to Table 2 in main submission.)

Method	LCC	SROCC
DIIVINE [6]	0.898	0.876
BRISQUE [5]	0.928	0.910
BLIINDS-II [8]	0.932	0.914
HOSA [11]	0.948	0.930
RankIQA	0.911	0.918
RankIQA+FT	**0.960**	**0.947**

Table 2. Average LCC and SROCC on CSIQ

average LCC and SROCC scores are computed on the testing set after training to convergence. This process is repeated ten times and the results are averaged. These results are shown in Table 3. MLIVE is more challenging than LIVE and CSIQ because multiple distortions are applied into each reference image. Similar conclusions as for LIVE and CSIQ can be drawn: RankIQA captures the factors that vary between different distortions, and achieves results similar to other methods. After fine-tuning, we obtain about 1% higher on LCC and about 2% higher on SROCC than the state-of-the-art.

Ablation Analysis To further probe the relation between performance of the metric and the number of images trained in ranked datasets and distortion levels, we did two experiments on the train-test split of TID2013. In the first experiment we fixed the number of reference images for generating distortions, and vary the number of distortion levels from 2 to 6 (we used 5 in the paper). The resulting LCC is in Table 4, from which observe that accuracy increases with more distortion levels until 5. In the second experiment we fixed the number of distortion levels, and vary the number



(a) Ranking loss on JP2K

(b) Ranking loss on JPEG

(c) Ranking loss on GB

(d) Ranking loss on GN

Figure 2. Convergence properties of our approach. Convergence of ranking loss on JP2K, JPEG, GB and GN distortions for our approach versus standard Siamese and hard-negative mining. (Corresponding to Figure 3 in main submission.)

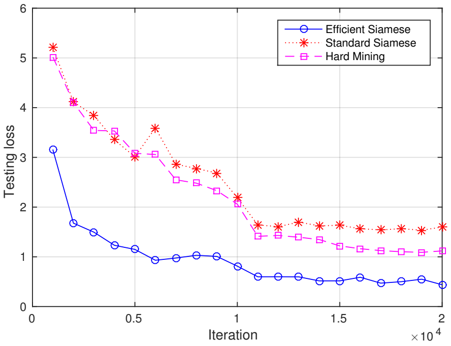


Figure 3. The networks are trained jointly on entire LIVE dataset including JPEG, JP2K, GB and GN distortions. Convergence of ranking loss for our approach versus standard Siamese and hard-negative mining.

Method	LCC	SROCC
DIIVINE [6]	0.894	0.874
BRISQUE [5]	0.921	0.897
BLIINDS-II [8]	0.903	0.887
HOSA [11]	0.926	0.902
RankIQA	0.898	0.893
RankIQA+FT	**0.936**	**0.921**

Table 3. Average LCC and SROCC on MLIVE

Distortion levels	2	3	4	5	6
LCC	0.73	0.75	0.82	0.85	0.84
Percent of dataset	20%	40%	60%	80%	100%
LCC	0.82	0.83	0.84	0.85	0.85

Table 4. The relation between LCC and distortion levels and percent of dataset.

of reference images from 20% to 100% of all reference images from Waterloo (we used 80% in the paper). The LCC increases gradually with percent of dataset. From these two experiments we see that distortion levels have more impact on accuracy than number of images. To conclude, more data still has the potential to increase performance, but levels of distortion are crucial for the final accuracy.

3. Synthetic ranked dataset generation

We synthetically generate a range of distortions from reference images. The types of distortions generated depend on the target test dataset.

To test on the LIVE [9] and CSIQ [3] datasets. We generate four types of distortions which are widely used and shared by the two datasets: Gaussian Blur (GB), Gaussian Noise (GN), JPEG Compression (JPEG), and JPEG2000 Compression (JP2K). Following [4] we generate all distortions at five levels, and the details are:

- **Gaussian Blur**: 2D circularly symmetric Gaussian blur kernels with standard deviations of [1.2, 2.5, 6.5, 15.2, 33.2] are used to distort the original images.
- **Gaussian Noise**: Gaussian noise is added to the original images, where variances are set to [0.001, 0.006, 0.022, 0.088, 1.000] for the five distortion levels, respectively.
- **JPEG Compression**: The quality factor that determines the DCT quantization matrix is set to be [43, 12, 7, 4, 0] for the five levels, respectively.
- **JPEG2000 Compression**: The compression ratio is set to be [52, 150, 343, 600, 1200] for the five levels, respectively.

To test on the MLIVE [2] dataset. MLIVE includes two datasets of images distorted by two types of distortions. The first is distorted by Gaussian Blur followed by JPEG (GB+JPEG). The second is distorted by Gaussian Blur followed by Gaussian Noise (GB+GN). The whole database consists of 450 distorted images. The distortion details are:

- **GB+JPEG**: 2D circularly symmetric Gaussian blur kernels with standard deviations of [1.2, 2.5, 6.5] are used to distort the reference images. Then the quality factor that determines the DCT quantization matrix is set to be [43, 12, 7] for the 3 levels, respectively. For each reference image, there are 9 distorted versions generated.
- **GB+GN**: 2D circularly symmetric Gaussian blur kernels with standard deviations of [1.2, 2.5, 6.5] are used to distort the reference images. Gaussian noise is then added to the images, with variances set to [0.001, 0.006, 0.022] for the 3 distortion levels, respectively. For each reference image, there are 9 distorted versions generated.

To test on TID2013 [7]. The TID2013 dataset consists of 25 reference images with 3000 distorted images from 24 different distortion types at 5 degradation levels. Mean Opinion Scores are in the range [0, 9]. Distortion types include a range of noise, compression, and transmission artifacts. We generate 17 out of the 24 distortions for training our networks. For the distortions which we could not generate, we apply fine-tuning from the network trained from the other ones. The generations details are as follows (distortions in **bold** are synthetically generated, while those in normal typeface we do not generate):

- **#01 additive white Gaussian noise**: The local variance of the Gaussian noise added in RGB color space is set to be [0.001, 0.005, 0.01, 0.05].
- **#02 additive noise in color components**: The local variance of the Gaussian noise added in the YCbCr color space is set to be [0.0140, 0.0198, 0.0343, 0.0524].
- #03 additive Gaussian spatially correlated noise: there was insufficient detail in the original TID2013 paper [7] about how spatially correlated noise was generated and added to reference images.
- #04 masked noise: there was insufficient detail in the original TID2013 paper [7] about how masks were generated.
- **#05 high frequency noise**: The local variance of the Gaussian noise added in the Fourier domain is set to be [0.001, 0.005, 0.01, 0.05] after which it is multiplied by a high-pass filter.
- **#06 impulse noise**: The local variance of "salt & pepper" noise added in RGB color space is set to be [0.005, 0.01, 0.05, 0.1].
- **#07 quantization noise**: The quantization step is set to be [27, 39, 55, 76].
- **#08 Gaussian blur**: 2D circularly symmetric Gaussian blur kernels are applied with standard deviations set to be [1.2, 2.5, 6.5, 15.2].
- **#09 image denoising**: The local variance of the Gaussian noise added in RGB color space is [0.001, 0.005, 0.01, 0.05]. Followed by the same denoising process as in [1].
- **#10 JPEG compression**: The quality factor that determines the DCT quantization matrix is set to be [43, 12, 7, 4].
- **#11 JPEG2000 compression**: The compression ratio is set to be [52, 150, 343, 600].
- #12 JPEG transmission errors: the precise details of how JPEG transmission errors were introduced was not clear and we were unable to reproduce this distortion type.

- #13 JPEG2000 transmission errors: the precise details of how JPEG2000 transmission errors were introduced was not clear and we were unable to reproduce this distortion type.

- **#14 non eccentricity pattern noise**: Patches of size 15x15 are randomly moved to nearby regions [7]. The number of patches is set to [30, 70, 150, 300].

- **#15 local blockwise distortion of different intensity**: Image patches of 32x32 are replaced by single color value (color block) [7]. The number of color blocks we distort is set to be [2, 4, 8, 16].

- **#16 mean shift**: Mean value shifting generated in both directions is set to be: [-60,-45,-30,-15] and [15, 30, 45, 60].

- **#17 contrast change**: Contrast change generated in both directions is set to be: [0.85, 0.7, 0.55, 0.4] and [1.2, 1.4, 1.6, 1.8].

- **#18 change of color saturation**: The control factor as in TID2013 paper [7] is set to be: [0.4, 0, -0.4, -0.8].

- **#19 multiplicative Gaussian noise**: The local variance of the Gaussian noise added is set to be [0.05, 0.09,0.13, 0.2].

- #20 comfort noise: the authors of [7] used a proprietary encoder unavailable to us.

- #21 lossy compression of noisy images: the authors of [7] used a proprietary encoder unavailable to us.

- **#22 image color quantization with dither**: The quantization step is set to be: [64, 32, 16, 8].

- **#23 chromatic aberrations**: The mutual shifting of in R and B channels is set to be [2, 6, 10, 14] and [1, 3, 5, 7], respectively.

- #24 sparse sampling and reconstruction: the authors of [7] used a proprietary encoder unavailable to us.

4. Shallow network details

In this section we give additional details of the Shallow network we use in our network comparison experimental analysis. The shallow network has four convolutional layers and one fully connected layer as shown in Table 5.

Layer	Output size	Description
Input layer	3x227x227	RGB input images
Layer 1	32x57x57	conv1 (3x3), relu1, pool1 (4x4)
Layer 2	32x14x14	conv2 (3x3), relu2, pool2 (4x4)
Layer 3	32x12x12	conv3 (3x3), relu3
Layer 4	32x1x1	conv4 (12x12)
Layer 5	1	fc1

Table 5. Details of the Shallow network used for evaluation.

5. Discussion

In this work we address the problem of scarcity in IQA data and proposed a method which learns from synthetically generated ranked image datasets. Our fast backpropagation method for Siamese networks is general and can be applied to a wide class of loss functions. In this supplementary material we gave specific details of how distortions are generated, and also showed that our efficient Siamese backpropagation technique converges faster and to a lower objective than standard Siamese network training and hard-negative mining. Finally, we demonstrated that that the results of our approach generalize to a broad range and number of distortions, as indicated by our state-of-the-art results on CSIQ, MLIVE, and TID2013.

References

[1] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007. 4

[2] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik. Objective quality assessment of multiply distorted images. In *Signals, Systems and Computers (ASILOMAR), 2012 Conference Record of the Forty Sixth Asilomar Conference on*, pages 1693–1697. IEEE, 2012. 1, 4

[3] E. C. Larson and D. M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006–011006, 2010. 1, 4

[4] K. Ma, Q. Wu, Z. Wang, Z. Duanmu, H. Yong, H. Li, and L. Zhang. Group MAD competition – a new methodology to compare objective image quality models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 4

[5] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *Image Processing, IEEE Transactions on*, 21(12):4695–4708, 2012. 2, 3

[6] A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, 2011. 2, 3

[7] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, et al. Color image database tid2013: Peculiarities and preliminary results. In *Visual Information Processing (EUVIP), 2013 4th European Workshop on*, pages 106–111. IEEE, 2013. 1, 4, 5

[8] M. A. Saad, A. C. Bovik, and C. Charrier. Blind image quality assessment: A natural scene statistics approach in

the dct domain. *Image Processing, IEEE Transactions on*, 21(8):3339–3352, 2012. 2, 3

[9] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. Live image quality assessment database. http://live.ece.utexas.edu/research/quality. 1, 4

[10] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 118–126, 2015. 1

[11] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing*, 25(9):4444–4457, 2016. 2, 3

[12] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. 1