KonIQ++: Boosting No-Reference Image Quality Assessment in the Wild by Jointly Predicting Image Quality and Defects — Supplementary Material

Shaolin Su[1]
shaolin_su@mail.nwpu.edu.cn
Vlad Hosu[2]
vlad.hosu@uni-konstanz.de
Hanhe Lin[3]
h.lin2@rgu.ac.uk
Yanning Zhang[1]
ynzhang@nwpu.edu.cn
Dietmar Saupe[2]
dietmar.saupe@uni-konstanz.de

[1] School of Computer Science and Engineering
Northwestern Polytechnical University
Xi'an, China

[2] Department of Computer and Information Science
University of Konstanz
Konstanz, Germany

[3] National Subsea Centre
Robert Gordon University
Aberdeen, UK

Here we give more details about the IQA database we created and analyse the performance of our proposed model. We show more examples of learnt features (heatmaps) for each prediction branch, and failure cases for the predictor.

1 Subjective Annotations for KonIQ++

1.1 User Study Interface

We first show the user interface for the subjective study in Figure 1. Participants were asked to identify if there is a degradation present, and only if they chose "Yes" then they were asked to identify all visible degradations. More than one degradation could be selected. Participants needed to provide a quality rating in either case.

1.2 Database properties

The distribution of distortion magnitudes for the four different types of annotated degradations is shown in Figure 2. All distortions except *blur* generally have small magnitudes, meaning only a small fraction of the participant selected them.

In Figure 3 we show the correlations between the distortion types. As an image can be labeled with multiple types of distortions, the total "distortion", representing the average of the magnitudes of the individual distortion types, negatively correlates the highest with quality with an SRCC of -0.92. The next most prevalent factor that negatively affects quality

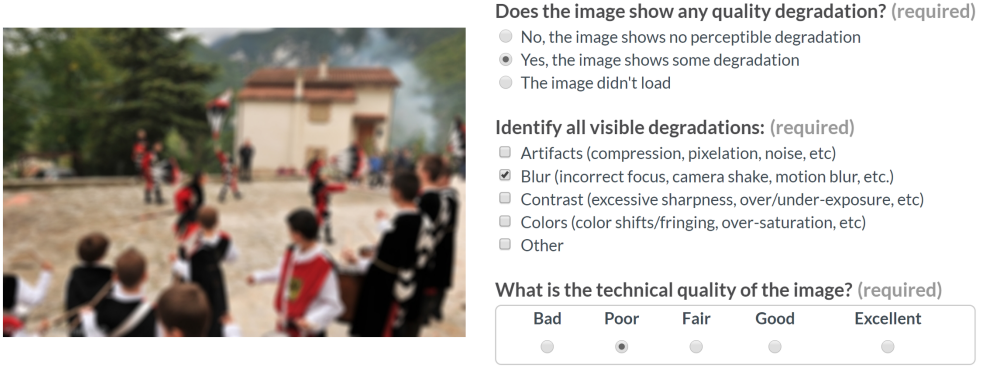


Figure 1: The user interface for the subjective study. *The query image was displayed on top of the questionaire in the actual study.*

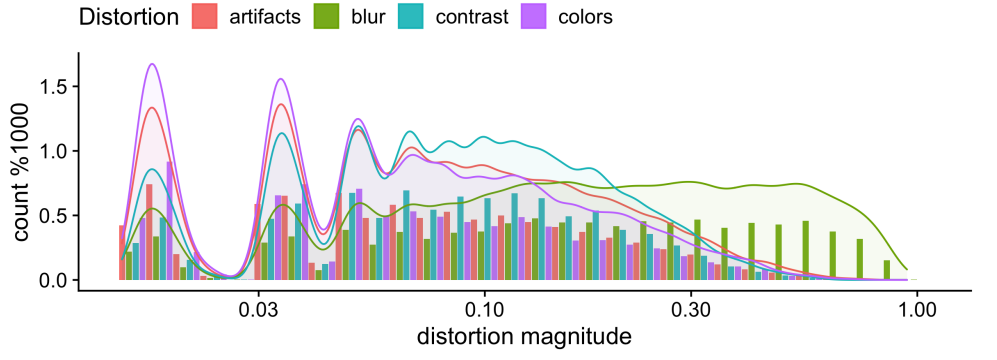


Figure 2: Distribution of distortion magnitudes for the KonIQ++ dataset. The kernel based density estimation is shown (as a line plot) as well as the histograms. The Y-axis represents the density value and the bin-counts divided by 1000. The distortion type "other" is not shown here as it was very rarely chosen. The distortion magnitude is shown using log-scale.

is the level of blur, having an SRCC of -0.82. It shows that the quality rating cannot be derived entirely from the degradation amount.

2 Interactions between image quality and distortion

In general, quality is inversely correlated with the magnitude of the distortions, see Figure 3. There are exceptions to this rule, where perceived image quality is not negatively affected by the distortion. In Figure 4, we show some example images taken from the KonIQ-10k dataset to illustrate this. We also show their quality score q and the distortion magnitude from subjective study d in Figure 4. The scores all range from 0 to 1, a higher value in q indicates a better quality, and a higher value in d indicates more distortions being detected.

In subfigure (a) – (c), the images are all of good perceptual quality, but some distortions are apparent. Concretely, distortions can be detected in some parts of subfigure (a) and (b), *i.e.* the over-exposed sky in subfigure (a) and the out-of-focus background in subfigure (b),

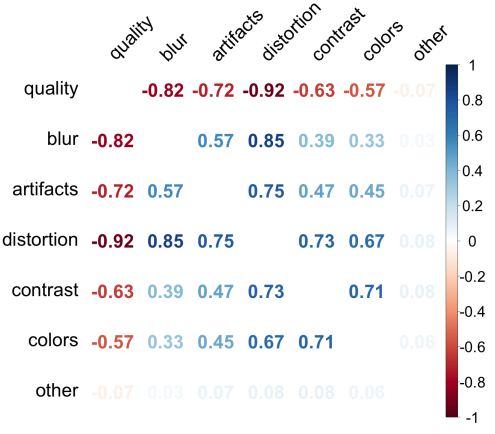


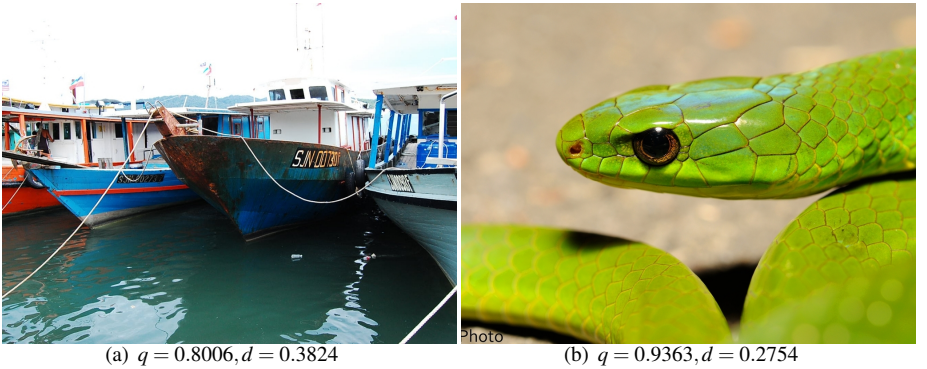
Figure 3: SRCC between distortion magnitudes and quality MOS. The field "distortion" represents the average magnitude of all distortions an image was labeled with.

while noise-like artifacts can be found in most parts of subfigure (c). Even though these distortions are obvious, they do not significantly affect image quality. In subfigure (d), we show an example image which is barely distorted, but the perceived quality is not better than that images from subfigure (a) – (c).

The above examples indicate that image quality and the presence of distortion interact in a much more complicated way during perception. Therefore, instead of simply extracting shared features and predicting image quality and distortion with multi-head regressors, as adopted in previous works [1], we propose a model which captures image quality and distortion features more precisely, and thus makes more accurate predictions on both image quality and distortion.

3 Additional visualizations for quality and distortion features

In this section, we show more distortion and quality feature heatmaps extracted from the proposed model. In Figures 5, 6, 7, the images contain several defects such as wrong contrast, blur, and other minor problems. The distortion prediction side network consistently responds to the presence of image defects, *i.e.* strong responses to incorrectly exposed regions in images from Figure 5, out of focus regions in images from Figure 6, and weak responses to images which are barely distorted from Figure 7. On the other hand, the quality prediction side network focuses on regions which are critical to perceived quality, *i.e.* large main areas of the scene and foreground objects. The results further demonstrate the two side networks are learning task specific features, allowing the model to make precise predictions on either distortions or image quality.



(a) $q = 0.8006, d = 0.3824$ (b) $q = 0.9363, d = 0.2754$

(c) $q = 0.8189, d = 0.3182$ (d) $q = 0.7492, d = 0.0952$

Figure 4: We show in some cases image quality is not necessarily dependent on distortions. In subfigure (a) – (c), the image exhibits high quality but also contains visible degradations. In subfigure (d), though barely distorted, the image quality is not better than that images from subfigure (a) – (c).

4 Ablation studies

In this section, we report the results of several ablation experiments on the model architecture. We first substitute the proposed FFRM module by simply concatenating input features and fusing them using 1×1 convolutions to observe the effectiveness of the FFRM module. The model with the simple feature fusion scheme is denoted as w/o FFRM. We then modified the model to predict image quality and defects in a single side network to evaluate if using separate side networks as in our originally proposed model is effective. The model is denoted as w/o separate. Next, we tested the model performances when different stages of backbone features are extracted and fused. We denote Stage=1, Stage=2, Stage=3 and Full models which receive the last 1, 2, 3 and all 4 stages of features extracted from the backbone network, respectively. All modified models are trained and tested on the proposed KonIQ++ dataset. The performances on IQA and defects prediction are shown in Table 1.

Based on Table 1, we made several observations. First, when substituting the proposed FFRM module with simple concatenation and convolution operations, model performances dropped w.r.t. both quality and defects predictions. This demonstrates that the feature refining operation is necessary in our model. Second, as an increasing number of stages of backbone features are fused and refined, model performances on image quality prediction

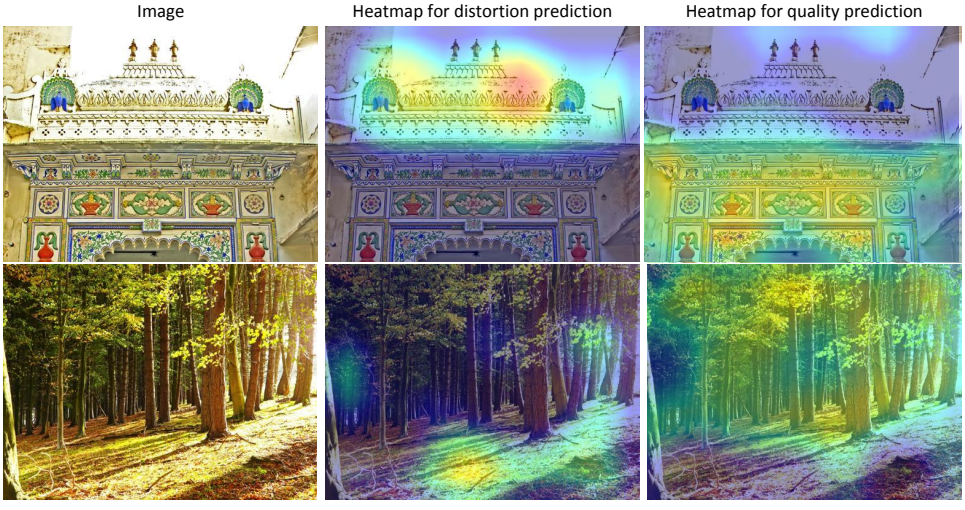


Figure 5: Additional results on heatmaps taken from distortion and quality prediction side networks of the model. The selected images contain visible contrast problem.

Model	Quality		Artifacts		Blur		Contrast		Colors	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
w/o FFRM	0.935	0.945	0.718	0.821	0.867	0.884	0.654	0.703	0.678	0.783
Stage=1	0.936	0.947	0.722	0.820	0.867	0.886	0.665	0.717	0.672	0.779
Stage=2	0.938	0.947	0.719	**0.823**	0.861	0.883	**0.673**	**0.723**	0.666	0.772
Stage=3	0.939	0.947	**0.726**	0.827	0.870	0.886	0.661	0.709	**0.683**	0.784
w/o separate	0.939	0.947	0.715	0.814	0.864	0.887	0.667	0.714	0.667	0.774
Full	**0.940**	**0.948**	0.723	**0.823**	**0.873**	**0.888**	0.672	0.718	0.677	**0.785**

Table 1: Ablation results on the proposed model with different modifications.

increase accordingly, indicating that multiple stages of image features are beneficial to IQA. However, the prediction accuracy for defects does not always follow the trend; this might be because image defects are more related with low-level feature representations, and combining features with deep semantic representations does not offer extra information. Last, when predicting image quality and defects in one side network, the performances are a bit lower than using two separate side networks. These results validate the effectiveness of using two separate side networks for learning task specific representations.

5 Failure cases and analysis

In this section, we show some failure cases for the model prediction. When testing the proposed model on the KonIQ-10k test subset, we select images whose deviation of the predicted quality score relative to the subjective MOS was large. We chose a threshold $\theta = 12$; the MOS scores range from 0 to 100. In total, there were 45 images with a prediction error larger than θ. In Figure 8 we show two examples of such images where the prediction was based on obviously incorrect parts of the image. The heatmaps were extracted from the quality prediction branch in order to interpret how model prediction fails, and shed some light on how the model could be further improved.

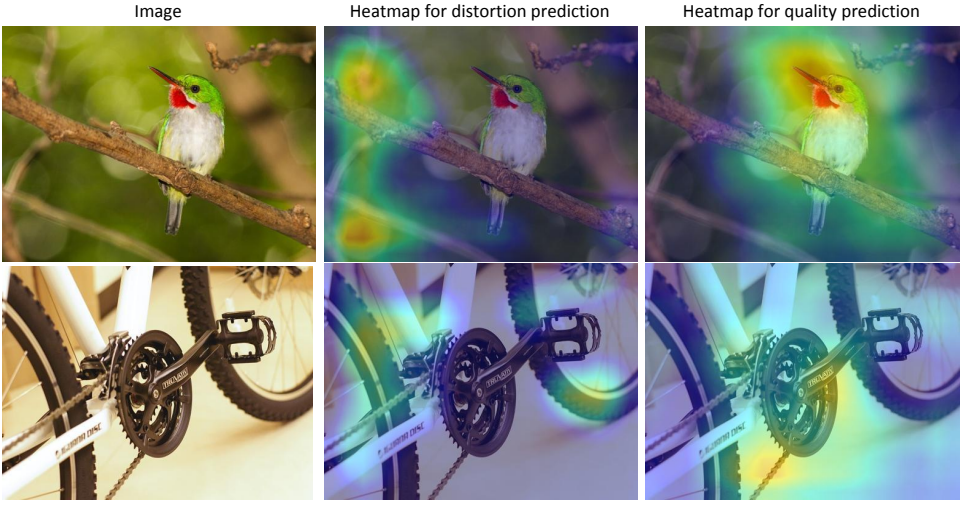


Figure 6: Additional results on heatmaps taken from distortion and quality prediction side networks of the model. The selected images are contaminated with blur.

We found that in the failure cases, the activation maps exclude regions which are critical for quality perception, thus leading to inaccurate predictions. Specifically, it is possible that the quality prediction side-network focuses on the background with more regular patterns and ignores highly variable regions such as those depicting people. These cases suggest that in order to improve model performance we could be imposing constraints on "quality critical" regions such as those depicting people or other salient areas. Collecting annotations for "quality critical" regions and adopting them for supervision could also be beneficial to improve model performance in future work. For instance, we could collect region-wise quality scores and distortion types indicating which part of the image was involved when making the judgement about the perceived quality.

References

[1] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3677–3686, 2020.

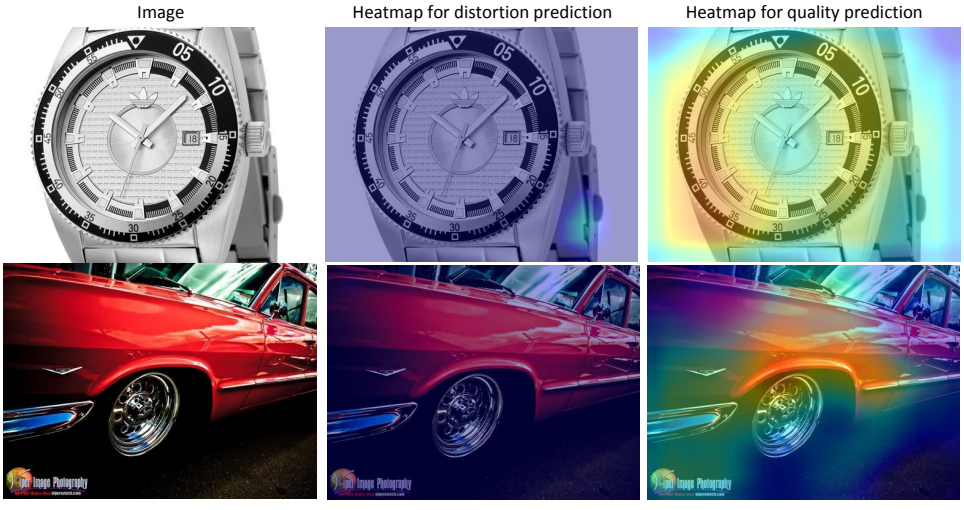


Figure 7: Additional results on heatmaps taken from distortion and quality prediction side networks of the model. The selected images are barely distorted.

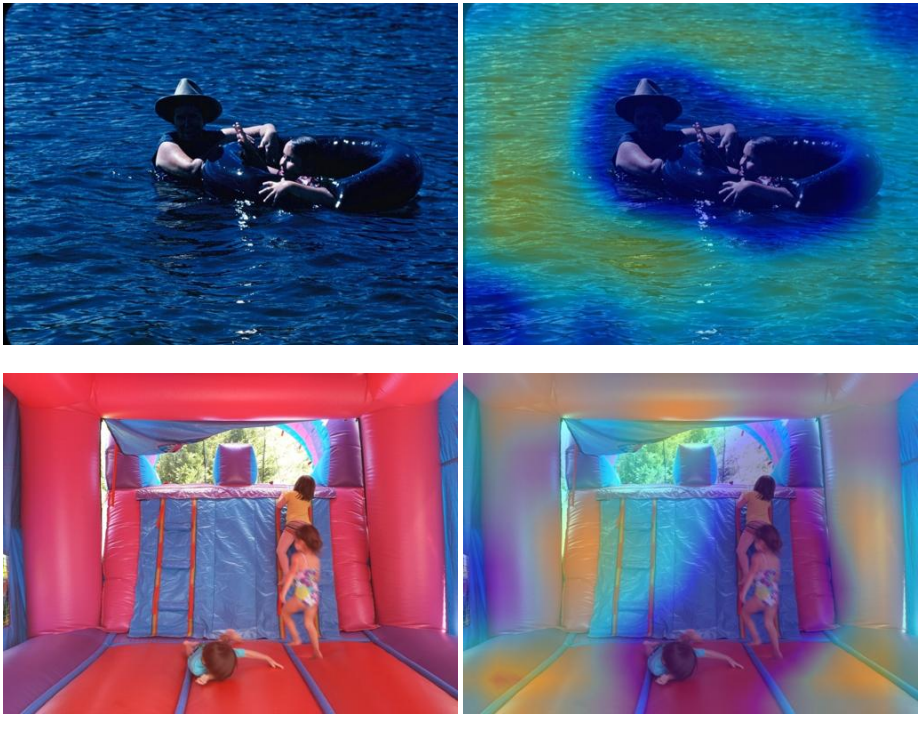


Figure 8: We show some failure images for model quality prediction. Their quality feature heatmaps are shown as well. For the image on the first row, predicted quality score and MOS are 62.31 and 49.52, respectively. For image on the second row, predicted quality score and MOS are 65.37 and 46.52.