EUROGRAPHICS 2024 / A. Bermano and E. Kalogerakis
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 43 (2024), Number 2

Supplementary Materials of "Enhancing image quality prediction with self-supervised visual masking"

1. Analysis of poorer performance for the CSIQ dataset

We further investigate the correlation of each metric across six distortion categories for the CSIQ dataset. We found that our approach slightly improves or maintains high correlations for the majority of the distortion categories; the only exception is the *global contrast decrements* category, where we see a significant decrease in the correlation across all metrics, resulting in an overall negative impact on the correlation. We attribute this behavior to the fact that the global contrast change results in strong brightness differences where our masking model apparently can not generalize to this specific unseen distortion category. Fig. 1 illustrates two sample images from this category where our predicted mask for the E-MAE metric exhibits less sensitivity to changes in brightness, particularly noticeable in the sky regions.

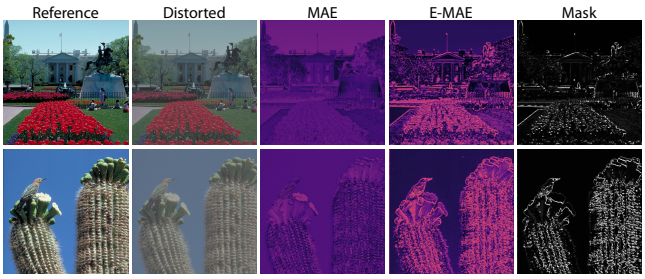

Figure 1: *Visualizations of error maps for MAE and its enhanced version (E-MAE) alongside the predicted masks for two sample examples within the "global contrast decrements" category from the CSIQ dataset.*

2. Additional results for analyzing the visual mask

In addition to Fig. 6 in the main paper, in Fig. 2, we further inspect the visual masks predicted by our approach across multiple metrics, using an example of motion blur distortions from the PIPAL dataset. As can be seen, the presence of blur is not uniform across the entire image; it becomes particularly noticeable when the direction of the motion blur is different from the pattern of the shirt (the right and upper parts). Here, we observe similar characteristics of predicted masks for MAE, PSNR, FLIP, and the first layer of VGG metrics, as in Fig. 6. For SSIM, which already includes a divisive contrast component akin to visual masking modeling, our predicted mask assigns identity weights to regions where SSIM accurately predicts errors and lowers weights in areas where SSIM exaggerates the error.

3. Mask visualization for the ablation experiments

In this section, we aim to investigate how the quantitative measures in ablation experiments are reflected in the predicted error maps. To this end, we show the E-MAE error map within various experimental setups (detailed in Sec. 4.3 of the main paper) for a Gaussian noise distortion example from the TID dataset. Fig. 3 shows the error maps when the metric is trained with only one distortion level per category. We observe that our enhanced error maps have less visual similarity compared to training across all five levels when it is trained using the lowest and highest distortion levels, while it has the highest similarity when trained with distortion level 3. This observation is aligned with our correlation measurement in Fig. 9. Additionally, Fig. 4 shows the error maps when E-MAE is trained with a subset of images in the training set. Training with 20 reference images appears insufficient in generating accurate visual masking, which is aligned with our findings in Fig. 10, where a reduction in correlation is observed with just 20 images. Conversely, training with 40 or 60 images closely approximates the results of training with the entire dataset, similarly reflected in the error maps. Lastly, in Fig. 5, we present the maps obtained through training with different subsets of distortion categories from the training set. Here, we observe that training exclusively with noise and blur can not produce precise masking, and including more categories is necessary to produce more localized masking. This is consistent with the correlation measures reported in Tbl. 4 in the main paper.

Employing the enhanced VGG metric as a loss Following the experiments in optimizing image restoration algorithms in main paper, we trained the state-of-the-art image restoration method, Restormer [ZAK*22] for the image-denoising with MAE + VGG and MAE + E-VGG in an identical conditions as stated in the main paper. The results are reported in Tbl. 1. The trained method with VGG shows a better LPIPS score as expected; however, we found denoising with E-VGG looks visually better, particularly in smooth low contrast regions (Fig. 6).

References

[ZAK*22] ZAMIR S. W., ARORA A., KHAN S., HAYAT M., KHAN F. S., YANG M.-H.: Restormer: Efficient transformer for high-resolution image restoration. In *Proc. CVPR* (2022), pp. 5728–5739. 1, 2

Table 1: *Evaluation of a blind Gaussian denoising task when employing VGG and the equal combination of VGG and E-VGG as loss functions. We show the performance of the trained models on synthetic Gaussian noise created with four distinct noise levels (σ) averaged across five benchmark datasets, consistent with the ones used in [ZAK*22].*

	σ = 15				σ = 25				σ = 50				σ = 60			
Loss	PSNR↑	SSIM↑	LPIPS↓	E-MAE↓	PSNR↑	SSIM↑	LPIPS↓	E-MAE↓	PSNR↑	SSIM↑	LPIPS↓	E-MAE↓	PSNR↑	SSIM↑	LPIPS↓	E-MAE↓
MAE + VGG	34.16	0.936	**0.033**	0.0356	31.68	0.900	**0.056**	0.0882	28.51	0.826	**0.111**	0.3279	27.66	0.797	0.132	0.4451
MAE + E-VGG	**34.34**	**0.939**	0.049	**0.0340**	**31.91**	**0.905**	0.078	**0.0837**	**28.78**	**0.835**	0.139	**0.3131**	**27.98**	**0.811**	0.155	**0.4216**

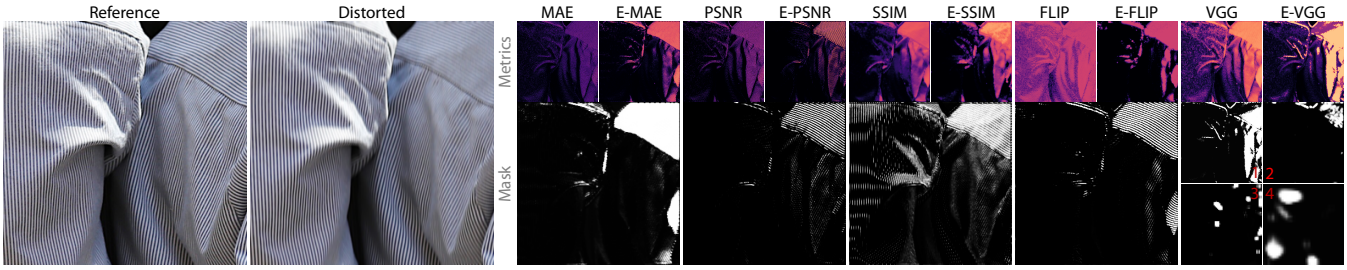

Figure 2: *Visualisation of predicted mask across different metrics for a given pair of reference and distorted images with motion blur from the PIPAL dataset. The SSIM values have been remapped to 1-SSIM, where lower values indicate less visible errors. In the case of the PSNR, we show the error map for the measured MSE. For the VGG metric, we visualize the predicted mask for all layers, while the error map is shown only for the first layer.*

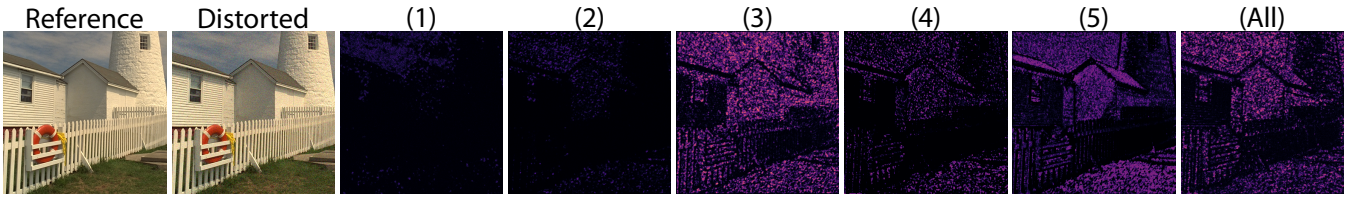

Figure 3: *Visualization of E-MAE error maps when it is trained with different levels of distortion.*

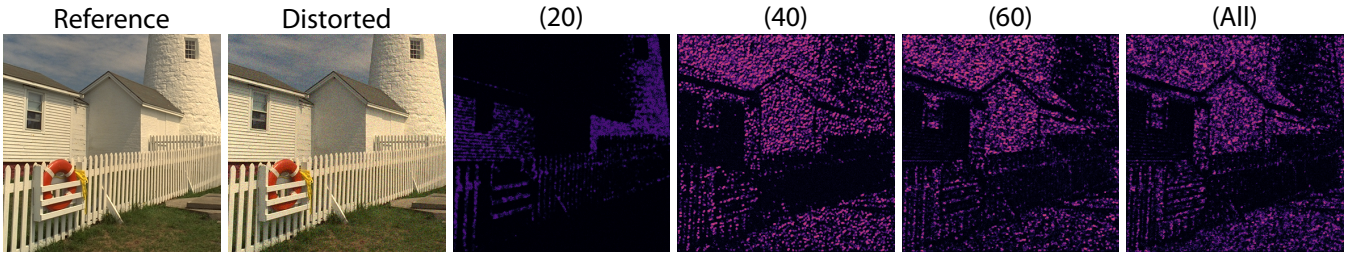

Figure 4: *Visualization of E-MAE error maps when it is trained with a different number of training images.*

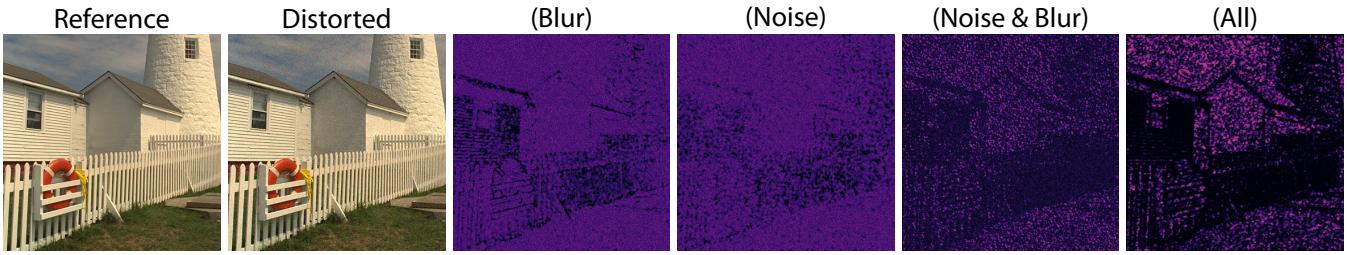

Figure 5: *Visualization of E-MAE error maps when it is trained with different distortion categories.*

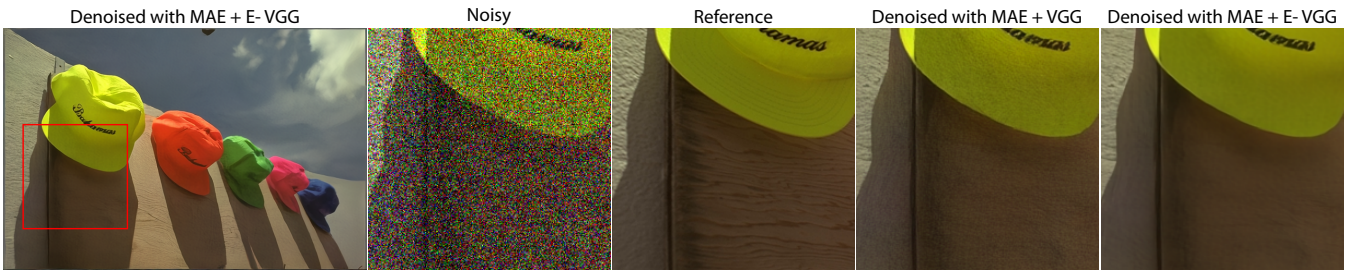

Figure 6: *Visual results in the image denoising task when employing MAE+VGG and MAE+E-VGG as loss functions. Denoising with MAE+VGG typically remains the noise in the dark region. On the other hand, MAE+E-VGG removes the noise successfully that matches better with human perception.*