DOI: 10.1111/cgf.15051

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

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

Enhancing image quality prediction with self-supervised visual masking

U. Çoğalan M. Bemana HP. Seidel and K. Myszkowski

Max-Planck-Institut für Informatik, Germany

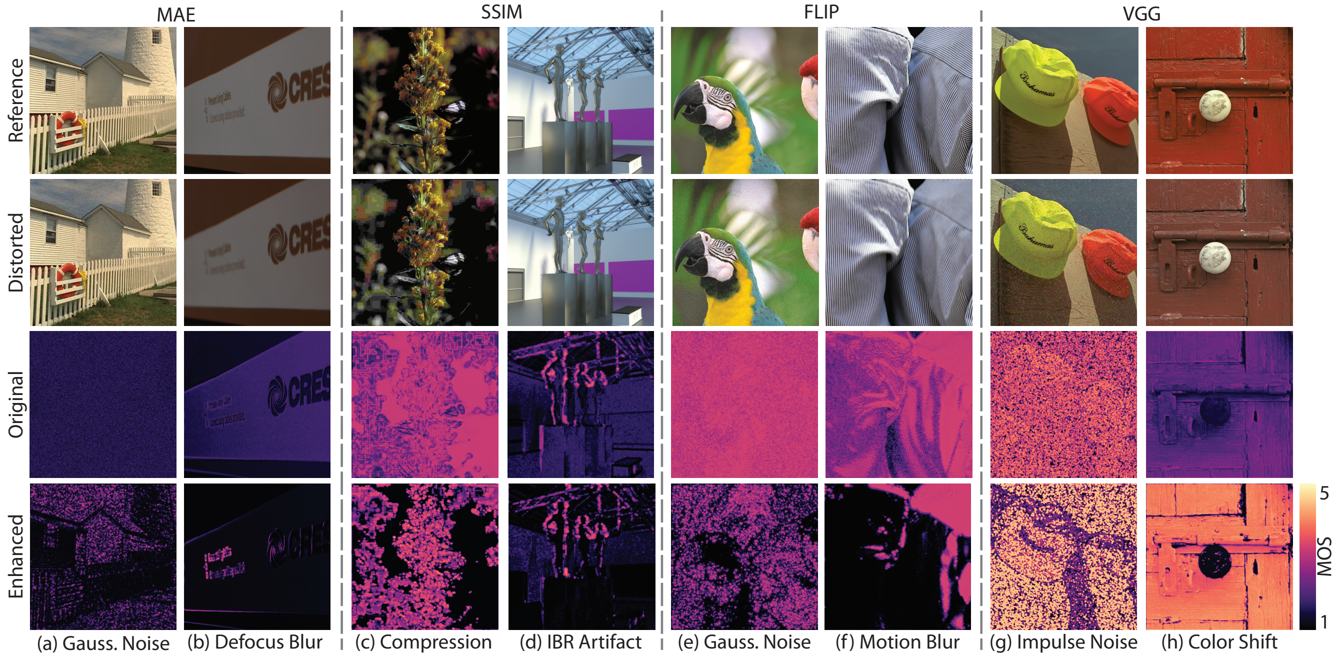

Figure 1: *We introduce self-supervised visual masking that enhances image quality prediction for existing quality metrics such as MAE, SSIM, FLIP, and VGG. Our work is inspired by the well-known characteristic of the Human Visual System (HVS), visual masking, which results in locally varying sensitivity to image artifact visibility that reduces with increasing contrast magnitude of the original image pattern. We found that the learned masking clearly outperforms its traditional hand-crafted versions and better adapts to specific distortion patterns. In the first two rows, we show the reference and distorted images, while the third and fourth rows show the error maps as predicted by the original metrics and their enhanced versions using our masking approach. As can be seen, our mask-enhanced metrics better predict the local distortion visibility by the human observer. For a more intuitive comparison, we scale each error map to fit within the mean opinion scores (MOS) range (please refer to Sec. 4.2 for more details). In this color scale, darker indicates less visible distortion.*

Abstract

Full-reference image quality metrics (FR-IQMs) aim to measure the visual differences between a pair of reference and distorted images, with the goal of accurately predicting human judgments. However, existing FR-IQMs, including traditional ones like PSNR and SSIM and even perceptual ones such as HDR-VDP, LPIPS, and DISTS, still fall short in capturing the complexities and nuances of human perception. In this work, rather than devising a novel IQM model, we seek to improve upon the perceptual quality of existing FR-IQM methods. We achieve this by considering visual masking, an important characteristic of the human visual system that changes its sensitivity to distortions as a function of local image content. Specifically, for a given FR-IQM metric, we propose to predict a visual masking model that modulates reference and distorted images in a way that penalizes the visual errors based on their visibility. Since the ground truth visual masks are difficult to obtain, we demonstrate how they can be derived in a self-supervised manner solely based on mean opinion scores (MOS) collected from an FR-IQM dataset. Our approach results in enhanced FR-IQM metrics that are more in line with human prediction both visually and quantitatively.

1. Introduction

Full-Reference Image Quality Metrics (FR-IQMs), which take as an input a pair of reference and distorted images, play a crucial role in a wide range of applications in digital image processing, such as image compression and transmission, as well as in evaluating the rendered content in computer graphics and vision. They are commonly used as a cost function in optimizing restoration tasks like denoising, deblurring, and super-resolution [DMWS21]. Consequently, developing FR-IQMs that accurately reflect the visual quality of images in accordance with the characteristics of the human visual system (HVS) is critical. The most commonly used FR-IQMs for evaluating image quality are the mean square error (MSE) or mean absolute error (MAE). While these per-pixel metrics are easy to compute, they assess image quality regardless of spatial content, leading to false positive predictions. This can be seen in Fig. 1a, where Gaussian noise is less noticeable in textured regions, while MAE predicts uniformly distributed error. Similarly, a depth-of-field blur is primarily visible on high-contrast fonts Fig. 1b, while MAE predicts the blur visibility also in smooth gradient regions. Other classic metrics like SSIM [WBSS04], while accounting for spatial content, often result in false positive predictions (the JPEG artifact and image-based rendering (IBR) artifact in Fig. 1c-d, respectively). A recent hand-crafted metric FLIP [ANA*20] is specifically designed to predict the visual differences in time-sequential image-pair flipping, which can make it too sensitive for side-by-side image evaluation, e.g., noise is less visible in high-contrast texture (Fig. 1e) or motion blur is not equally visible across different parts of an image (Fig. 1f). Recognizing that hand-crafted image features may not adequately capture the HVS complexity, modern metrics [ZIE*18] strive to assess the perceptual dissimilarity between images by comparing deep features extracted from classification networks [SZ15]. These metrics appear to better account for the HVS characteristics; however, they are designed to generate a single value per image pair and cannot provide correct visible error localization, as can be seen in the impulse noise example (Fig. 1g). Moreover, the features learned through training the classification networks tend to be less sensitive to global distortions, such as moderate color and brightness changes (Fig. 1h) that have less impact on reliable classification.

The objective of this work is not to develop a new perceptual FR-IQM; instead, we are interested in improving the quality prediction of existing metrics to align more closely with human judgment (Fig. 2). We also aim to enhance the accuracy of error map predictions by considering multiple factors such as image content, distortion levels, and distortion types. By detecting both the presence and evaluating the magnitude of visible distortion in each pixel, we aim to ensure that the metric predictions more accurately reflect the probability of a human observer detecting differences between a pair of images. In this regard, there have been several efforts toward incorporating the perceptual aspects of human vision, specifically visual masking [LF80, Fol94, WG84], into FR-IQM methods [Lub95, Dal93, MKRH11, MDC*21]. In simple words, visual masking refers to the phenomenon in which certain components of an image (in our application, distortions) may be less visible to the viewer due to the presence of other visual elements in the same image. Visual masking can affect image quality perception, making some image distortions less visible to the viewer [FSPG97, ZDL02]. However, existing visual masking models are typically hand-crafted and struggle to generalize effectively across various distortion types. Although learning a visual masking model appears to be a natural solution, the lack of reliable ground truth data for visual masking makes direct supervision impractical. In this work, we propose a self-supervised approach to predict visual masking using a dataset of images featuring a variety of distortions of different magnitudes whose quality has been evaluated in the mean opinion scores (MOS) experiment with human subjects [LHS19].

In summary, our work offers the following contributions:

- We propose a lightweight CNN that generates a mask for a given reference and distorted input pair. The predicted mask acts as a per-pixel weight and, when multiplied with the inputs, greatly improves the performance of the existing FR-IQMs. The incorporation of our learned mask into any FR-IQM is seamless and demands minimal computational resources. While the CNN is trained specifically for each metric, it learns a generic masking model capable of identifying various types of distortions.
- We demonstrate that our masking model can be generalized to deep features and used as a per-layer feature map weight.
- Our solution significantly enhances the accuracy of quality prediction for FR-IQMs across various test datasets. Furthermore, it produces per-pixel error maps that visually align more closely with human perception compared to the original FR-IQMs.
- We show the potential application of our approach as a loss function for training image denoising and motion deblurring.

2. Previous work

FR-IQMs can be categorized into classical metrics, which perform the computation directly in the image space, and learning-based metrics, which leverage deep feature models to assess image quality.

Classic metrics Basic FR-IQMs, such as MSE, RMSE, and MAE, compute the per-pixel difference to quantify image distortion. While these metrics are straightforward to calculate, their consistency with human vision is typically low. Such perceptual consistency can be improved by considering relative instead of absolute error, as in PSNR and the symmetric mean absolute percentage error (SMAPE) [VRM*18]. To account for the spatial aspects of the HVS, alternative metrics such as SSIM [WBSS04] are introduced, which consider image patches and measure local differences in luminance, contrast, and structural information. SSIM is further extended to multi-scale MS-SSIM [WSB03] and complex wavelet CW-SSIM [SWG*09] versions that capture both global and local structural information. FSIM [ZZMZ11] decomposes the image into multiple subbands using Gabor filters and compares subband responses between the reference and distorted images. By assuming that natural images have a specific distribution of pixel values, models based on information theory [SB05, SB06] measure the mutual information between images by comparing their joint histograms and taking into account the statistical dependencies between neighboring pixels. Classical metrics can offer either a single overall quality score or a visibility map indicating the distortion intensity. Watson-DCT [Wat93], VDM [Lub95], VDP [Dal93], HDR-VDP [MKRH11], and fovVideoVDP [MDC*21] measure either the visibility of distortions or perceived distortions magnitude, or both

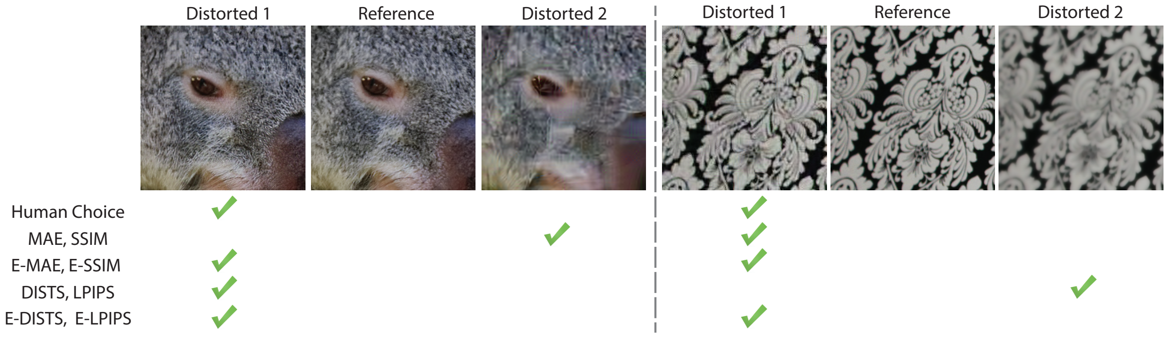

Figure 2: *Agreement of metric predictions with human judgments. We consider the classic (MAE and SSIM) and learning-based (LPIPS and DISTS) metrics, and we compare their prediction to their enhanced versions (E-MAE, E-SSIM, E-DISTS, and E-LPIPS) using our approach. On the left, we see a situation where MAE and SSIM favor JPEG-like artifacts over slightly resampled textures. On the right, we encounter a scenario where LPIPS and DISTS prefer blur over a subtle color shift. Our extended metric versions are better aligned with human choice. The images have been extracted from the PIPAL dataset [JHH*20].*

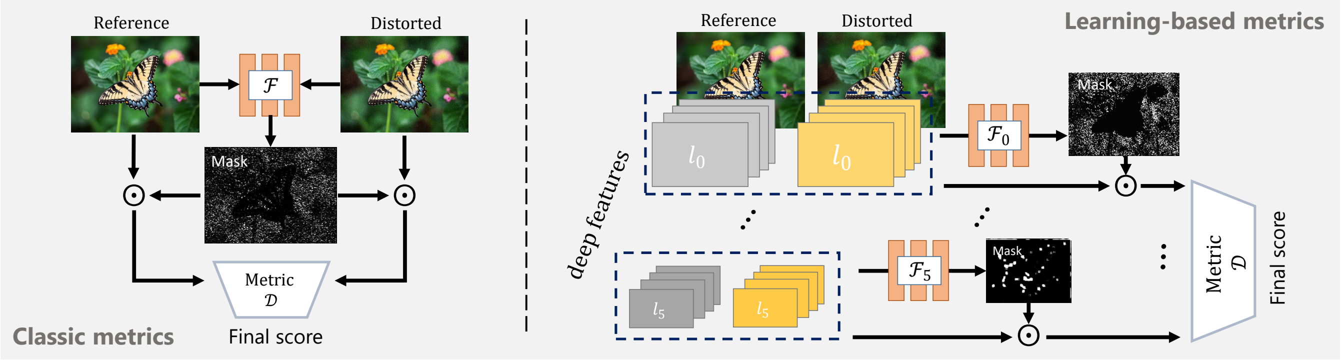

Figure 3: *Our proposed visual masking for enhancing classic metrics such as MAE and SSIM (left) and learning-based metrics such as DISTS or LPIPS (right). For classic metrics, the input to our mask predictor network $\mathcal{F}$ are sRGB images, while for learning-based metrics, the inputs are the VGG features extracted from the images. We learn the visual masks in a self-supervised fashion by minimizing the difference between the metric final score and human scores collected from an FR-IQM dataset.*

by considering various visual aspects such as luminance adaptation, contrast sensitivity, and visual masking. A more recent metric, FLIP [ANA*20], emphasizes color differences, and it is sensitive to even subtle distortions by emulating flipping between the compared image pair.

Deep learning-based metrics In recent years, research in FR-IQM has been placing greater emphasis on perceptual comparisons in deep feature space rather than image space to enhance the alignment with human judgments. Prashnani et al. [PCMS18] are among the first to utilize deep feature models learned from human-labeled data to predict perceptual errors. Zhang et al. [ZIE*18] demonstrate that internal image representations from classification networks can be used for image comparison. They propose the Perceptual Image Patch Similarity (LPIPS) index, which quantifies image similarity by measuring the ℓ_2 distances between pre-trained VGG features. To further improve the correlation with human judgments, they learn per-channel weights for selected VGG features using their collected perceptual similarity dataset. Recognizing that simple ℓ_p-norm measures fail to consider the statistical dependency of errors across different locations, Ding et al. [DMWS20] introduce the DISTS, which aims to measure the texture and structure similarity between feature pairs by comparing their global mean, variance, and correlations in the form of SSIM. Building upon this work, A-DISTS [DLZ*21] extended the approach to incorporate local structure and texture comparisons. Czolbe et al. [CKCI20] incorporate their extended Watson-DCT model [Wat93] as a measure of VGG feature distance. Moving away from deterministic point-wise feature comparisons, DeepWSD [LCZ*22] compares the overall distributions of features using the Wasserstein distance, a statistical measure for comparing two distributions. Nevertheless, the majority of the proposed IQMs metrics are targeted toward producing a single quality score and are not primarily designed to generate per-pixel error maps. In this regard, Wolski et al. [WGY*18] employ a custom CNN model trained in a fully supervised way using coarse user marking data to predict an error visibility map that highlights the regions where distortions are more likely to be noticeable.
Recently, deep learning-based no-reference metrics (NR-IQM) such as KonCept512 [HLSS20], HYPERIQA [SYZ*20], MUSIQ [KWW*21] and MANIQA [YWS*22] have been proposed. While NR-IQM methods often report impressive performance, their practical applicability remains limited. FR-IQM metrics are still predominant in CG applications, as the reference images are typically readily available.

In this work, we extend the classic and deep learning-based full-reference metrics by introducing a learnable component trained on perceptual MOS data in a self-supervised way. By implicitly analyzing local image content, our model derives per-pixel maps that mimic visual masking, effectively modeling the visual significance of distortions.

3. Self-supervised visual masking

This section elaborates on our methodology for perceptually calibrating the existing FR-IQMs. Given a reference and distorted pair (X and Y) $\in R^{H\times W\times C}$, we first learn a visual mask, $M \in R^{H\times W\times 1}$, which has the same spatial dimensions as the inputs. For classical metrics (Fig. 3-left), the input X and Y are sRGB images ($C=3$), while for learning-based metrics such as LPIPS, DISTS, or DeepWSD, the input X and Y are the VGG features extracted from the images and C is the number of channels in a given VGG layer (Fig. 3-right). The predicted mask is then element-wise multiplied with X and Y before being fed into an FR-IQM, $\mathcal{D}$. Note that, for learning-based metrics, a direct modulation of the input sRGB images by a mask M would distort their content and consequently reduce the VGG performance as it is originally trained on complete, non-masked images. Our solution with VGG feature modulation draws inspiration from classic FR-IQMs [Lub95, Dal93, MKRH11, MDC*21], where the response from handcrafted filter banks is transduced using a fixed, perception-motivated masking model [LF80, Fol94, WG84]. In our approach, the response from pre-trained VGG filters is modulated with a learned per-pixel mask M, where perception modeling is learned from the MOS data. We estimate the mask M by utilizing a lightweight CNN denoted as $\mathcal{F}$, which takes both X and Y as input. Mathematically, this can be expressed as:

$$M = \mathcal{F}(X,Y) \tag{1}$$

It is important to note that the network $\mathcal{F}$ is trained specifically for a metric $\mathcal{D}$. In the case of metrics such as LPIPS, DISTS, and DeepWSD, we follow their specific architecture and compute a mask for each layer using a separate $\mathcal{F}$, and the same mask is applied for all channels in a given layer (Fig. 3-right). The original spatial pooling is preserved for each metric, such as L1 distance in LPIPS, structural similarity in DISTS, or Wasserstein distance in DeepWSD. Since we cannot directly supervise the output of the mask generator network, we adopt a self-supervised approach to train it using an IQM dataset with a single quality score. The network's parameters are optimized by minimizing the ℓ_2 difference between the metric output value and human scores. Our loss is formulated as follows:

$$Loss = \|\mathcal{G}(\mathcal{D}(M\odot X, M\odot Y)) - q\|_2^2 \tag{2}$$

Here, $q \in [0,1]$ represents the normalized mean opinion score when comparing the images X and Y. As the metric response can vary in an arbitrary range, following a similar approach in [ZIE*18], a small network $\mathcal{G}$ is jointly trained to map the metric response to the human ratings.

3.1. Training and network details

For training, we use the KADID dataset [LHS19], which comprises 81 natural images that have been distorted using 25 types of traditional distortions, each at five different levels, making roughly 10k training pairs. Note that we train our mask generator network $\mathcal{F}$ for all the distortion categories together rather than for one specific category. We find that a lightweight CNN with three convolutional layers, each consisting of 64 channels, suffices for successful training. ReLU activation is applied after each layer, while we use Sigmoid activation for the final layer to keep the mask values in the range between 0 and 1. The computation overhead of our mask generator network is very negligible, and it takes only 12 ms to compute the mask on a 1080Ti GPU with an input resolution of $768\times 512\times 3$. Our mapping network $\mathcal{G}$ consists of two 32-channel fully connected (FC) ReLU layers, followed by a 1-channel FC layer with Sigmoid activation. The batch size for training is set to 4. We employ the Adam optimizer [KB15] with an initial learning rate of 10^{-4} and a weight decay of 10^{-6}.

Table 1: *Performance comparison of existing FR-IQMs and their enhanced versions using our approach (specified by the prefix E) on three standard IQM datasets. The prefix R denotes the original metric retraining on the KADID dataset, while the prefix S refers to employing a visual saliency mask instead of our mask. At the bottom, we include the corresponding results for NR-IQMs. Higher values of SRCC, PLCC, and KRCC indicate better quality prediction. The first and second best metrics for each dataset are indicated in bold and underlined, respectively. Additionally, the version with superior correlation is highlighted in dark gray for each metric.*

	CSIQ			**TID**			**PIPAL**		
Metric	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
FSIM	0.900	0.913	0.740	0.847	0.789	0.611	0.651	0.617	0.441
VIF	0.826	0.841	0.642	0.820	0.813	0.616	0.584	0.538	0.378
HDR-VDP-2	0.761	0.886	0.704	0.715	0.753	0.571	0.514	0.503	0.354
PieAPP	0.827	0.840	0.653	0.832	0.849	0.652	**0.729**	**0.709**	**0.521**
MAE	0.819	0.801	0.599	0.639	0.627	0.409	0.458	0.443	0.304
S-MAE	0.656	0.697	0.493	0.498	0.496	0.347	0.369	0.365	0.248
E-MAE	0.871	0.917	0.738	0.857	0.863	0.673	0.597	0.606	0.429
PSNR	0.851	0.837	0.645	0.726	0.714	0.540	0.468	0.456	0.314
E-PSNR	0.901	0.910	0.728	0.855	0.844	0.656	0.637	0.629	0.446
SSIM	0.848	0.863	0.665	0.697	0.663	0.479	0.550	0.534	0.373
E-SSIM	0.869	0.910	0.732	0.842	0.868	0.677	0.671	0.656	0.469
MS-SSIM	0.826	0.841	0.642	0.820	0.813	0.616	0.584	0.538	0.379
E-MS-SSIM	0.862	0.895	0.709	0.806	0.825	0.621	0.642	0.634	0.453
FLIP	0.731	0.724	0.527	0.591	0.537	0.413	0.498	0.442	0.306
E-FLIP	0.871	0.902	0.715	0.859	0.858	0.666	0.621	0.612	0.434
FovVideoVDP	0.795	0.821	0.632	0.742	0.727	0.544	0.565	0.509	0.358
E-FovVideoVDP	0.841	0.882	0.685	0.830	0.816	0.623	0.662	0.626	0.449
VGG	0.938	0.952	0.804	0.853	0.820	0.639	0.643	0.610	0.432
E-VGG	0.914	0.938	0.776	0.895	0.889	0.710	0.695	0.675	0.485
LPIPS	0.944	0.929	0.769	0.803	0.756	0.568	0.640	0.598	0.424
R-LPIPS	0.931	0.917	0.756	0.898	0.886	0.697	0.670	0.640	0.447
E-LPIPS	0.922	0.933	0.771	0.884	0.876	0.689	0.705	0.678	0.490
DISTS	0.947	0.947	0.796	0.839	0.811	0.619	0.645	0.626	0.445
E-DISTS	0.938	0.925	0.754	**0.903**	**0.915**	**0.725**	0.725	0.697	0.507
Watson-VGG	0.944	0.940	0.785	0.808	0.763	0.573	0.627	0.606	0.429
E-Watson-VGG	0.917	0.936	0.776	0.886	0.895	0.716	0.697	0.678	0.488
DeepWSD	0.949	**0.961**	0.821	0.879	0.861	0.674	0.593	0.584	0.409
R-DeepWSD	**0.955**	**0.961**	**0.823**	0.895	0.88	0.695	0.654	0.633	0.449
E-DeepWSD	0.937	0.937	0.775	0.905	0.892	0.710	0.704	0.672	0.485
HYPERIQA	0.769	0.757	0.573	0.679	0.662	0.489	0.325	0.363	0.250
MANIQA	0.874	0.827	0.642	0.784	0.760	0.572	0.404	0.407	0.276

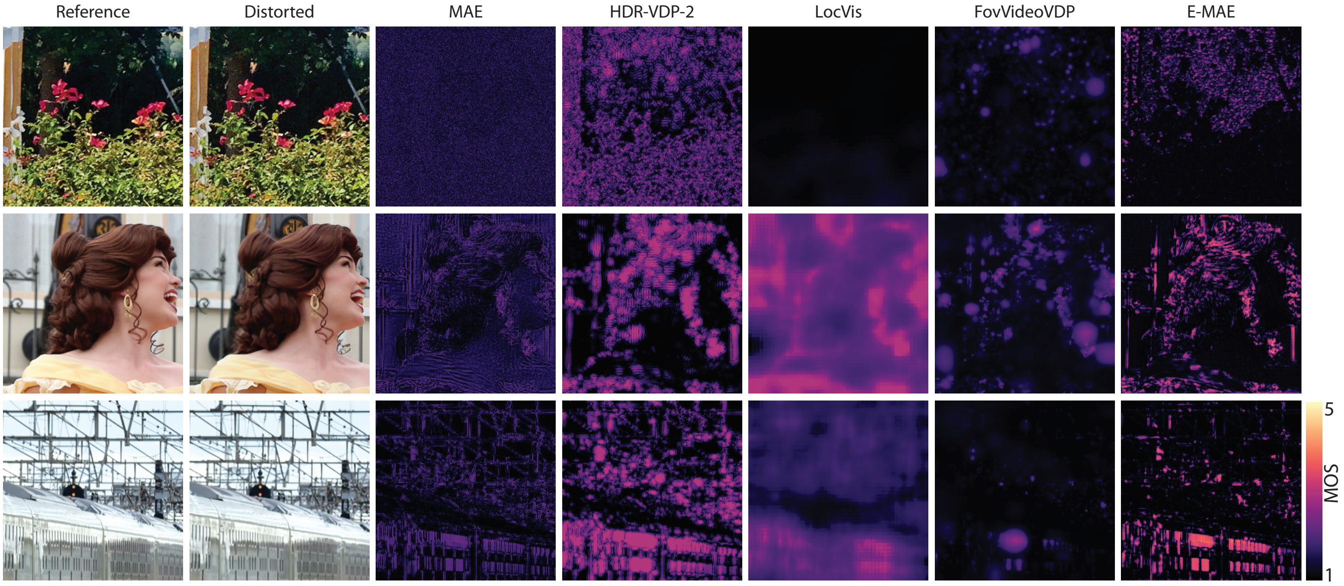

Figure 4: *Visual comparisons of distortion visibility maps for Gaussian noise (upper row) and superresolution artifacts (middle and bottom rows). The distortion examples are taken from the PIPAL dataset. The first two columns present the reference and distorted images, followed by the respective metric predictions: MAE, HDR-VDP-2 [MKRH11], LocVis [WGY*18], FovVideoVDP [MDC*21], and our E-MAE. Here, we additionally visualize the MAE map to better understand the characteristics of each distortion. As can be seen, the existing metrics tend to either underestimate or overestimate the distortion visibility. Note that LocVis and E-MAE have not seen distorted images with superresolution artifacts in their training.*

4. Results

In this section, we first present our experimental setup, which we use for our method evaluation and ablations of different training strategies.

4.1. Experimental setup

We employ our visual masking approach to enhance some of the classical metrics (MAE, PSNR, SSIM, MS-SSIM, FLIP, and fovVideoVDP) and recent learning-based methods (VGG, LPIPS, DISTS, Watson-VGG, and DeepWSD). Note for MS-SSIM, we used the same $\mathcal{F}$ across all scales, while the inputs are images at different scales. Moreover, the metric called VGG is computed by simply taking the ℓ_1 difference between VGG features for the same layers as originally chosen for LPIPS and DISTS. Deploying our masking model to PieAPP or any other metrics that create new CNN architectures from scratch is not practical as there is no intermediate component to which we can apply our masking model. Thus, our main focus remains on mainstream metrics that use features extracted from pre-trained networks for quality prediction. We assess the performance of our proposed approach on three well-established IQM datasets: CSIQ [LC10], TID2013 [PJI*15], and PIPAL [JHH*20]. The first two datasets mainly consist of synthetic distortions, ranging from 1k to 3k images. On the other hand, PIPAL is the most comprehensive IQM dataset due to its diverse and complex distortions, consisting of 23k images. Each reference image in this dataset was subjected to 116 distortions, including 19 GAN-type distortions. For evaluation, following [DMWS20], we resize the smaller side resolution of input images to 224 while maintaining the aspect ratio. Note that rescaling is only performed on the test datasets to match the image resolution in which the MOS data were collected. Our approach does not require rescaled inputs, and all visual figures in the paper are processed in their original resolution. For each dataset, three metrics are used for evaluation: Spearman's rank correlation coefficient (SRCC), Pearson linear correlation coefficient (PLCC), and the Kendall rank correlation coefficient (KRCC). The PLCC measures the accuracy of the predictions, while the SRCC indicates the monotonicity of the predictions, and the KRCC measures the ordinal association. The PLCC measures linear correlation, requiring both variables (metric output and MOS) to be on the same scale, hence, we mapped the metric scores to the MOS values using a four-parameter logistic function, consistent with established IQM methods [DMWS20, LCZ*22]. We do not use $\mathcal{G}$ for PLCC remapping; otherwise, we need to train a specific $\mathcal{G}$ for each metric on a given test set. Importantly, SRCC and KRCC scores do not require additional remapping, thus directly reflecting the correlation between metric output and MOS data.

4.2. Evaluations

In this section, we present the outcome of the quantitative (agreement with the MOS data) and qualitative (the quality of error maps) evaluation of our method. We also analyze the mask content and relate it with perceptual models of contrast and blur perception. Finally, we analyze the error map prediction of different distortion levels, and we consider the potential use of our enhanced E-MAE metric as a loss in denoising and deblurring image restoration tasks.

Quality prediction The experimental results are presented in Tbl. 1, where with the prefix E, we denote our proposed extension for each specific IQM. Our extension of traditional metrics, such as

MAE, PSNR, SSIM, FLIP, and fovVideoVDP, consistently improves their performance for all datasets. This is remarkable as those metrics are commonly used, and our simple extension can make their distortion prediction closer to the human observer. Interestingly, the enhanced E-MAE and E-PSNR outperform recent learning-based VGG, LPIPS, and DISTS in the TID dataset while showing a comparable performance for the PIPAL dataset. Notable improvements are also observed in both datasets for the recent learning-based metrics (E-VGG, E-LPIPS, E-DIST, Watson-VGG, and E-DeepWSD), positioning them at a level comparable to other state-of-the-art IQMs, such as PieAPP [PCMS18]. The only exception is the case of the small-scale CSIQ dataset, where the original learning-based metrics achieve high correlations with the MOS data and leave little space for further improvements. Please see our supplementary for a more detailed analysis.

We also consider retraining LPIPS per-channel weights using the KADID dataset (denoted as R-LPIPS in Tbl. 1), which improves correlation for TID and PIPAL datasets with respect to the original LPIPS. Compared to our E-LPIPS, such retraining is more prone to overfitting; it performs marginally better for the TID dataset, which has more distortion similarities with KADID, while it is significantly worse for the larger and more diverse PIPAL dataset. Similarly, training layer-specific weights for DeepWSD (R-DeepWSD) improves correlation, however, our E-DeepWSD achieves better performance. Moreover, channel/layer-wise weighing can not be reasonably applied to image-based metrics (MAE, SSIM, FLIP).

We evaluate the performance of recent NR-IQM methods MANIQA [YWS*22] and HYPERIQA [SYZ*20] that are trained on the KADID dataset. As it can be seen in Tbl. 1, the NR-IQM methods show significantly lower correlations with the MOS data, particularly for the PIPAL dataset, which indicates that FR-IQM methods can better generalize to unseen distortion types.

Visual saliency methods incorporate semantic information, however, they are not trained to discriminate between dominant distortions and salient features (e.g., faces). This seems to be a limiting factor in the direct saliency use in our image quality evaluation framework. To validate this observation, we employed a predicted saliency map from an off-the-shelf saliency network [Jia18] as a mask to the MAE metric that we denote as S-MAE in Tbl. 1. While in this simple attempt, we observe significantly lower correlations with the MOS data, we believe that our approach can be complemented by saliency, so that effectively distortion predictions are narrowed to image regions that are likely to be visually attended.

Error map prediction In Fig. 1, we show the error maps predicted by various existing IQMs and their enhanced versions for a set of images featuring different types of distortions. As the output of each metric can be in an unbounded range and vary across different metrics and their improved versions with our approach, for a more intuitive and fair comparison, instead of simply normalizing them within the range from zero to one using a Sigmoid function [ANA*20], we visualize the output of each metric after being scaled to the MOS range using a pre-trained scaling network $\mathcal{G}$. Specifically, we utilize KADID dataset and train a separate $\mathcal{G}$ for each metric to transform their raw response into values that align with human ratings (MOS). Note that for the enhanced version of each metric, the network $\mathcal{G}$ is already provided from the training step. In general, this scaling process is akin to mapping the metric scores to the MOS values using a four-parameter function when computing the correlation. As can be seen in Fig. 1, the enhanced error maps using our approach better align with the human perception of the distortion. A notable example is the case of Gaussian noise, where a metric like MAE predicts uniformly distributed error, and our masking approach effectively redistributes the error in terms of both their magnitude and local visibility. We provide more examples in our supplementary materials. Additionally, Fig. 4 showcases three examples where our E-MAE metric achieves better localized error maps compared to well-established visibility metrics such as HDR-VDP-2 [MKRH11], LocVis [WGY*18], and FovVideoVDP [MDC*21].

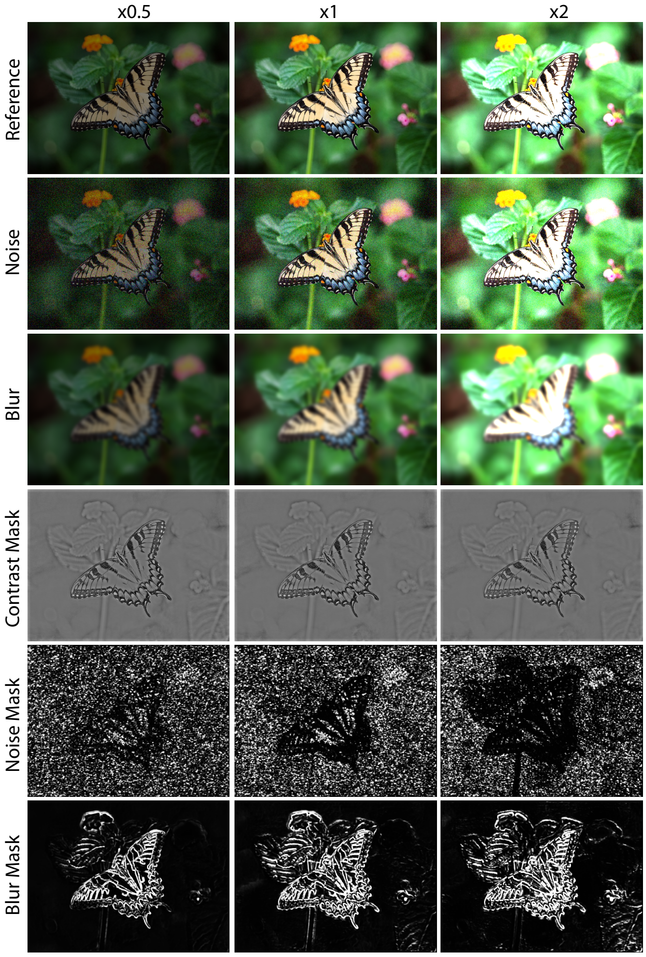

Figure 5: *Comparison of our E-MAE metric masks for the noise (fifth row) and blur (sixth row) distortions as a function of different image contrast ($\times0.5$, $\times1$, and $\times2$). In the fourth row, we also show a map with the human sensitivity to local contrast changes as predicted by a traditional model of visual contrast masking [TAKW*19, Eq.4]. In all cases, darker means more masking (less sensitive to distortion).*

Mask visualization It is also intriguing to see the learned mask, i.e., the output of the network $\mathcal{F}$, and to compare it with a traditional visual contrast masking model, such as the one used in JPEG2000 compression [ZDL02]. To this end, Fig. 5 presents our masks generated for noise and blur distortions. We consider the same distortion level and three levels of image contrast enhancement (×0.5, ×1, and ×2). In the case of noise distortion, our learned masks predict stronger visual masking in the high-contrast butterfly and better noise visibility in the out-of-focus smooth background. Increasing image contrast (×2) leads to even stronger visual masking in the butterfly area and the plant behind it. Reducing image contrast (×0.5) leads to the inverse effect. Such behavior is compatible with the visual contrast masking model [ZDL02, TAKW*19], where due to self-contrast masking, the higher the contrast of the original signal (e.g., on edges), the stronger the distortion should be to make it visible. Along a similar line, due to neighborhood masking, the higher the contrast texture, the stronger the visual masking as well. In the case of blur distortion, our learned mask predicts its strong visibility on high-contrast edges. The stronger the image contrast (×2), the blur visibility improves. Assigning a higher weight by our mask to high contrast regions agrees with perceptual models of blur detection and discrimination [WA11, SBG15].

Note that we derive each mask taking as an input both the reference and distorted images; the mask can resolve even per-pixel distortions, as in the case of noise (Fig. 5), and accordingly informs the E-MAE metric on the perceptual importance of such distortions. What is also remarkable is that the HVS might impose contradictory requirements on hand-crafted visual models that become specific for a given distortion. This is well illustrated in Fig. 5, where noise can be better masked by strong contrast patterns [ZDL02, TAKW*19] while blur is actually better revealed by strong contrast patterns [WA11]. Our learned E-MAE mask recognizes the distortion context and reacts as expected by penalizing less noise distortion in high-contrast and textured regions while penalizing more blur distortion at high-contrast edges. Interestingly, such local, seemingly contradictory behavior has been learned solemnly by providing multiple pairs of reference and distortion images along with the corresponding quality MOS rating, which is just a single number. No annotation on specific distortion types has been required in our training. Fig. 11 shows further examples that our learned masking is also informed about contrast masking by texture [FSPG97] and the contrast sensitivity function (CSF) [Dal93, Bar99, WAK*20].

Masks vs. metrics analysis Masks typically vary with distortion type, as demonstrated in Fig. 5 for noise and blur. In Fig. 6, we further illustrate the predicted mask across various metrics, including MAE, PSNR, SSIM, FLIP, and VGG for a given pair of reference and distorted images with Gaussian noise. As can be seen, metrics with similar characteristics, such as MAE, PSNR, and FLIP, tend to learn similar maps. For a more perceptually-informed metric like SSIM that partially models visual masking, our predicted mask adjusts its sensitivity by assigning lower weight to regions where SSIM exaggerates the error (e.g., in the grass area) and identity weight when accurately predicting the error magnitude (e.g., the body of lighthouse). When it comes to VGG, the mask learned for the early layer resembles the MAE mask since the initial convolutional layers tend to learn basic image features like edges and textures, whereas for the deeper layers, as the VGG learns more abstract features, the interpretation of the masks become less obvious. In the supplementary, we provide an additional example with blur distortion.

Employing the enhanced metric as a loss In this part, we investigate the benefit of the enhanced IQMs in optimizing image restoration algorithms. To this end, we employ MAE and MAE+$\lambda\cdot$ E-MAE as loss functions for training image denoising and motion deblurring using the state-of-the-art image restoration method, Restormer [ZAK*22]. For the denoising task, we select the images in the BSD400 dataset [MFTM01] as our training set and introduce synthetic noise to these images by applying additive white Gaussian noise with a randomly chosen standard deviation ranging between 0 and 50. We performed each training with the same number of iterations in an identical setup (e.g., learning rate). Then, we evaluate the trained models on five benchmark datasets, consistent with the ones used in [ZAK*22]. We conduct our evaluation for various noise levels and report the results in Tbl. 2. We can observe that training just with the MAE loss leads to higher PSNRs, in particular for higher noise levels, but at the same time, image blur and contrast loss can be observed (refer to Fig. 7). More perceptually inclined quality metrics penalize for such visual quality reduction, e.g., LPIPS is sensitive to excessive blur [ZIE*18]. Combining with an E-MAE loss component clearly improves such metrics' scores consistently across various noise levels as well as the visual quality. For the motion deblurring task, we employed the GoPro dataset [NHKML17] for the training and evaluation. The combination of MAE and E-MAE enhances the deblurring results across different quality metrics (Tbl. 3) and leads to a sharper appearance (Fig. 8). In both tasks, we empirically found that $\lambda = 1$ gives the best performance. We also observe that relying exclusively on the E-MAE loss component leads to worse results, which is expected, as indicated in [DMWS21]. In the supplementary, we provide more comparisons for other loss combinations, such as MAE+VGG and MAE+E-VGG that lead to similar conclusions.

4.3. Ablations

We perform a set of ablations to investigate the impact of reduced training data in terms of distortion levels, reference image number, and distortion type diversity on the E-MAE metric prediction accuracy.

Distortion levels The first experiment analyzes the importance of incorporating various distortion levels into our training set. In this regard, we train our network for the E-MAE metric using only one distortion level per category, and the results are reported in Fig. 9. Interestingly, for all the datasets (except PIPAL), an inverse U-shape trend emerged across five different distortion levels, where we observe the lowest correlation when training with the minimum and maximum distortion levels (levels 1 and 5). Conversely, a moderate amount of distortion (level 3) appears to be sufficiently representative for each distortion category and achieved a comparable correlation to training with all five levels. This behavior can be anticipated because, at the lowest and highest distortion levels, the distortions are either barely visible or strongly visible, leading to the consistent selection of mostly extreme rating scores. Consequently, when the network is exclusively exposed to images with one such extreme distortion and rating levels, it fails to learn to differentiate between them. On the other hand, at moderate distortion levels where distortions are partially visible or invisible, the network has a better opportunity to learn masks that behave differently for varying spatial locations.

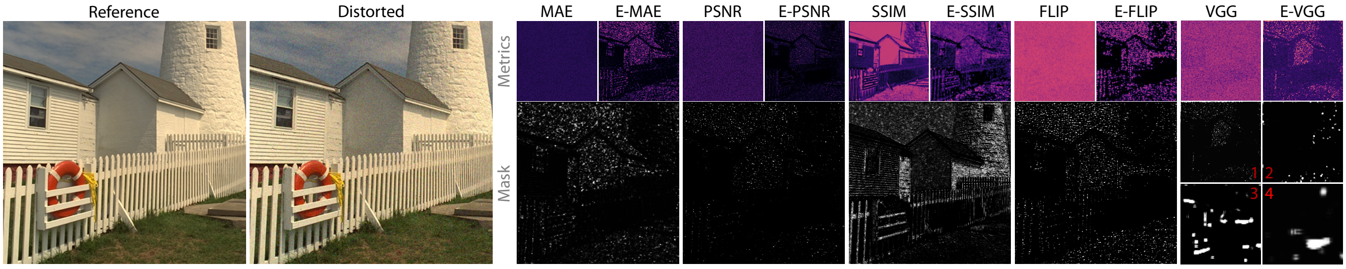

Figure 6: *Visualisation of predicted mask across different metrics for a given pair of reference and distorted images with Gaussian noise from the TID dataset. Note that the SSIM values have been remapped to 1-SSIM, where lower values indicate less visible errors. In the case of the PSNR, we show the 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.*

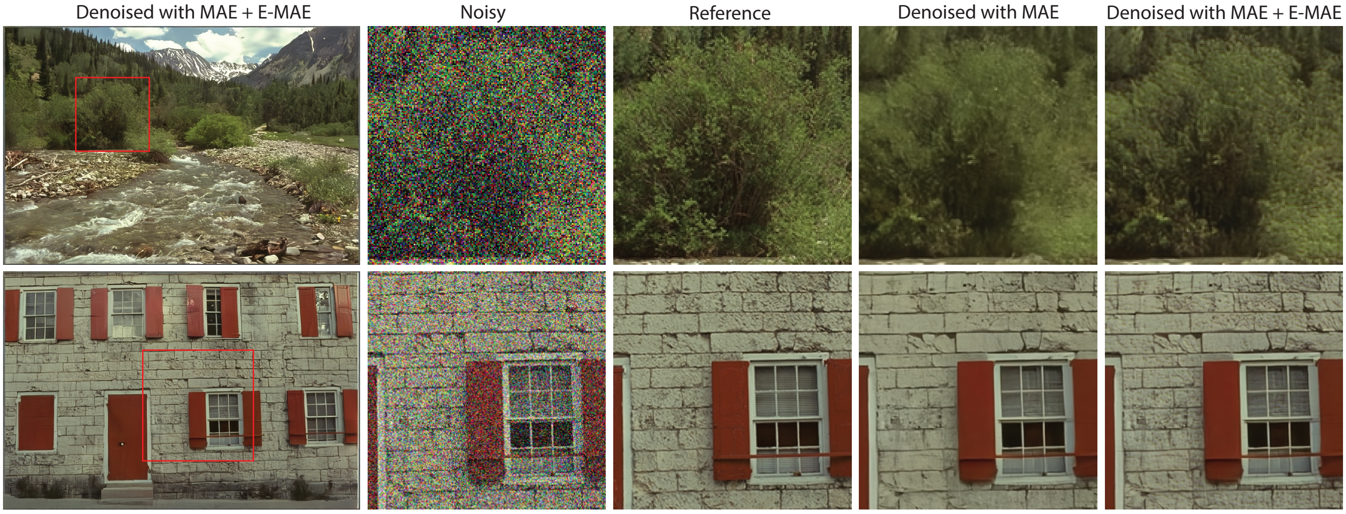

Figure 7: *Visual results in the image denoising task when employing MAE and MAE+E-MAE as loss functions. Considering that denoiser networks typically reduce noise through smoothing, our objective was to investigate whether the use of the E-MAE loss component could encourage the network to retain or hallucinate details, even if they do not precisely match the reference but their discrepancy from the ground truth is possibly not perceivable. As can be seen, the denoised images with the MAE+E-MAE loss yield sharper content and higher contrast.*

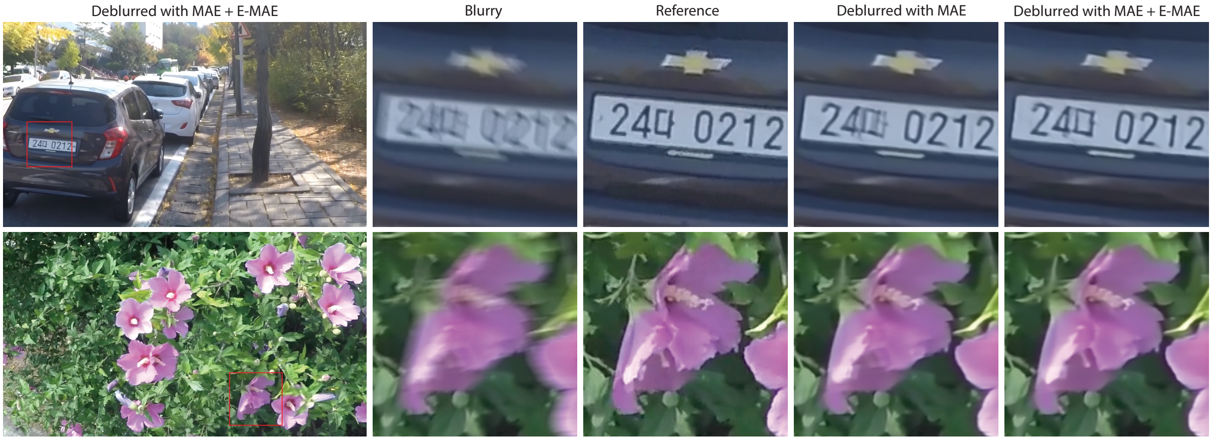

Figure 8: *Visual results for the motion deblurring task when employing MAE and MAE + E-MAE as loss functions.*

Dataset size Although we employ a large-scale KADID dataset in our training (25 distortion types × five distortion levels), the number of reference images is limited to 81. This ablation aims to investigate the training performance by even further reducing the number of reference images. To this end, we perform multiple runs of E-MAE metric training using randomly selected subsets of 20, 40, and 60 reference images. Fig. 10 presents the SRCC correlations averaged over multiple runs. The correlation differences between 40, 60, and the full set of 81 reference images are minor. In the case of 20 reference images, the performance is slightly lower and the

Table 2: *Evaluation of a blind Gaussian denoising task when employing MAE and the equal combination of MAE and E-MAE 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	34.36	**0.94**	0.058	0.0343	**31.94**	0.90	0.092	0.0849	**28.82**	**0.84**	0.163	3.187	**28.02**	**0.81**	0.182	4.258
MAE + E-MAE	**34.37**	**0.94**	**0.055**	**0.0145**	31.92	**0.91**	**0.087**	**0.0263**	28.71	**0.84**	**0.152**	**0.790**	27.88	**0.81**	**0.167**	**1.035**

Table 3: *Evaluation of a motion deblurring task when employing MAE and the equal combination of MAE and E-MAE as loss functions. We show the performance of the trained models on synthetic blur created using the GoPro dataset [NHKML17].*

Metric	PSNR↑	SSIM↑	LPIPS↓	E-MAE↓
MAE	31.70	0.92	0.1030	0.0192
MAE + E-MAE	**31.78**	**0.93**	**0.1018**	**0.0184**

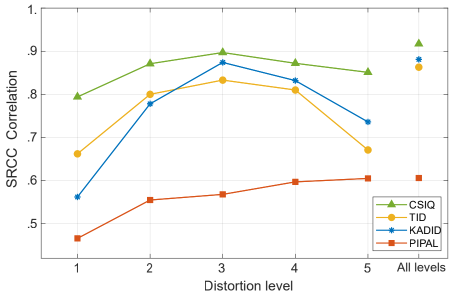

Figure 9: *Evaluation of E-MAE training performance using only selected distortion levels for each distortion category. We measure the SRCC correlation with the MOS data, and as a reference, we also include the results of complete training with all distortion levels.*

variance higher, which indicates that 20 scenes might not be enough to capture image content variability.

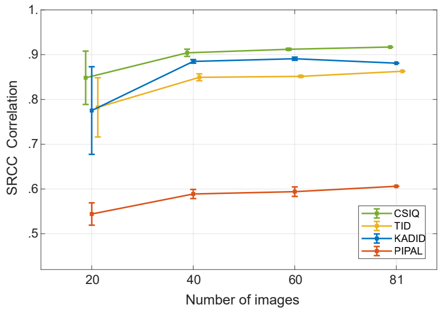

Figure 10: *Evaluation of E-MAE training performance using different numbers of the reference images (scenes). Multiple training runs have been performed for 20, 40, and 60 randomly selected scenes from the full set of 81 reference images. The data points represent the respective SRCC correlation averages over such runs, while the vertical bars depict the standard deviation.*

Distortion diversity We investigate the impact of separate E-MAE training on specific distortion subsets such as noise, blur, combined noise, and blur, as well as the complete KADID dataset. At the test time, we evaluate trained this way E-MAE versions on noise and blur subsets of the TID dataset, as well as its complete version. The results, presented in Tbl. 4, reveal that training solely on the noise category unsurprisingly improves the SRCC correlation within that category; however, it also enhances the overall correlation for the TID dataset with respect to the original MAE. Conversely, training exclusively on blur does not improve the performance within the blur category itself, as the blur distortion already exhibits a strong correlation (0.934) for the MAE metric, making any improvement marginal. On the other hand, we noticed that training with all categories combined significantly improves the correlation in both the noise and blur categories compared to training with only noise or blur categories, which can suggest that exposing the network to a wider range of distortion types enables better generalization.

Table 4: *The SRCC correlation with the MOS data for the E-MAE metric trained with specific distortion categories (noise, blur, noise&blur) and the entire (all) KADID dataset, as indicated in the brackets. The TID dataset is used for testing, where the "Category" columns indicate whether only the noise and blur subsets are considered or the entire dataset.*

Metric **Category:**	noise	blur	all
MAE	0.601	0.934	0.545
E-MAE (noise)	0.847	0.927	0.674
E-MAE (blur)	0.732	0.926	0.655
E-MAE (noise & blur)	0.841	0.936	0.726
E-MAE (all)	**0.906**	**0.955**	**0.857**

In our supplementary material, we additionally show the impact of each of the three factors on the predicted error maps.

5. Limitations and future work

The actual visual contrast masking is the function of the viewing condition and the display size [Cha13], which is often considered in the perceptual quality metrics [Dal93, MKRH11, MDC*21, ANA*20] but otherwise mostly ignored. However, the effectiveness of our visual masking model is limited to the experimental setup where human scores are obtained in the KADID dataset.

As we have demonstrated, incorporating our masking model into traditional metrics is straightforward, but it might be a difficult task for certain network architectures, such as PieAPP [PCMS18].

As shown in Fig. 11, in the context of the CSF reproduction, our metric might not be well calibrated for near contrast threshold stimuli, whose visibility is also affected by the viewing distance and display conditions. Wolski et al. [WGY*18] developed the LocVis dataset with local maps of distortion detection probability

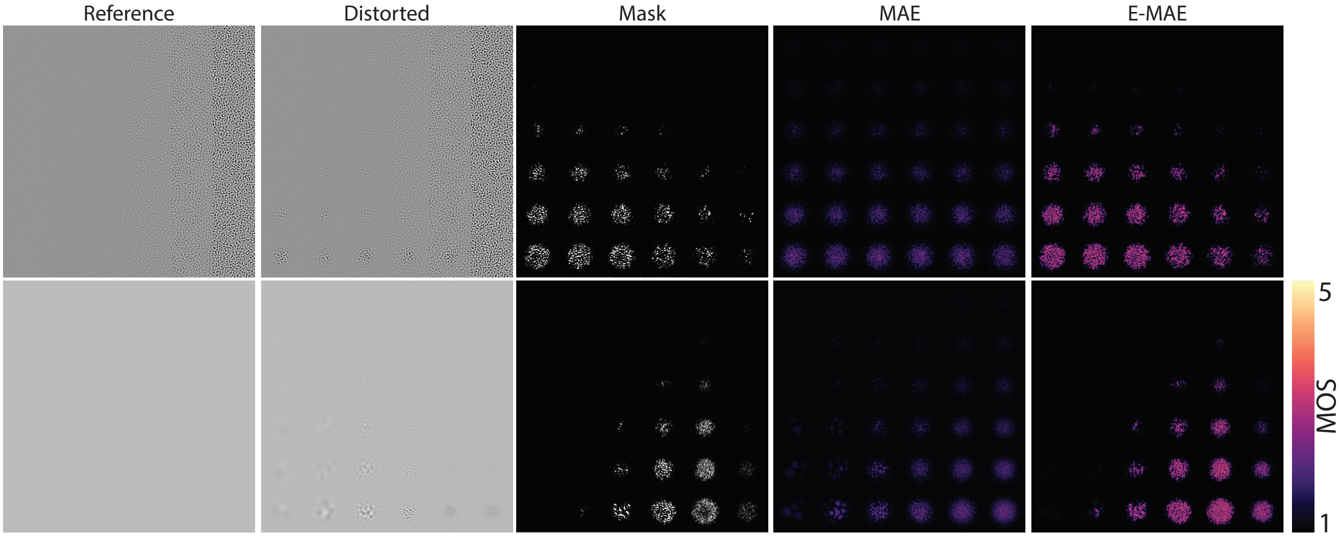

Figure 11: *Error map prediction for the MAE and E-MAE metrics along with learned weighting masks for two perception patterns from [ČHM*13]. These patterns were specifically designed to investigate various perceptual phenomena, including contrast sensitivity and contrast masking. In the first row, the background consists of a high-frequency pattern with increasing contrast toward the right and a stimulus pattern with decreasing contrast from bottom to top (which becomes more apparent when zoomed in). In this scenario, contrast masking is more pronounced with increasing background contrast that, in turn, reduces the stimulus visibility, and E-MAE correctly predicts this effect. The second row presents another example, showing a set of patterns where their spatial frequencies increase toward the right while their contrast decreases toward the top. In this case, the learned masking roughly follows an inverse U-shape characteristic, akin to the contrast sensitivity function (CSF) [Dal93, Bar99, WAK*20]. Our masking well approximates the sensitivity drop for high frequencies but tends to suppress the visibility of low-frequency patterns excessively. In spite of this drawback, we still find it quite remarkable that the CSF shape becomes apparent in our learned mask without any explicit training with calibrated near-threshold CSF data.*

that emphasize near-threshold distortions. Unfortunately, the LocVis dataset is not reliable for supra-threshold distortion in terms of their magnitude estimation, while we readily learn from the MOS data. We relegate as future work, combining such not-compatible distortion visibility and magnitude estimation data into a consistent training dataset for enhancing our masking model.

6. Conclusion

In this paper, we present a new approach towards reducing the notorious gap between the existing quality metric prediction and the actual distortion visibility by the human observer. We achieve this by self-supervised training of a metric-specific network using the existing distortion datasets labeled with mean opinion score (MOS). We show that although overall image quality is rated with a single MOS value in the training data, by securing sufficient diversity of such training, as detailed in our ablation study, the network can leverage global MOS into a meaningful per-pixel mask. The mask, through different weighting of local distortion visibility, which also adapts to specific distortion types, helps a given metric to aggregate such local information into the comprehensive MOS value, as imposed by the training data. The mask can be learned directly in the image space for traditional metrics or in the feature space for recent learning-based metrics. In either case, it is trivial to incorporate into most of the existing metrics. Remarkably, our approach improves the performance of commonly used metrics, such as MAE, PSNR, SSIM, and FLIP on all datasets we tested. The prediction accuracy of recent learning-based metrics is typically substantially enhanced.

Acknowledgements

Open Access funding enabled and organized by Projekt DEAL.

References

[ANA*20] Andersson P., Nilsson J., Akenine-Möller T., Oskarsson M., Åström K., Fairchild M. D.: FLIP: A Difference Evaluator for Alternating Images. *Proc. ACM Comput. Graph. Interact. Tech. 3*, 2 (2020), 15:1–15:23. 2, 3, 6, 9

[Bar99] Barten P. G.: *Contrast sensitivity of the human eye and its effects on image quality*. SPIE – The International Society for Optical Engineering, 1999. 7, 10

[Cha13] Chandler D. M.: Seven challenges in image quality assessment: past, present, and future research. *International Scholarly Research Notices* (2013). 9

[ČHM*13] Čadík M., Herzog R., Mantiuk R., Mantiuk R., Myszkowski K., Seidel H.-P.: Learning to predict localized distortions in rendered images. In *Computer Graphics Forum (Proc. Eurographics)* (2013), vol. 32, pp. 401–410. 10

[CKCI20] Czolbe S., Krause O., Cox I., Igel C.: A loss function for generative neural networks based on watson's perceptual model. In *Proc. NIPS* (2020). 3

[Dal93] Daly S.: The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity. In *Digital Image and Human Vision* (1993), pp. 179–206. 2, 4, 7, 9, 10

[DLZ*21] Ding K., Liu Y., Zou X., Wang S., Ma K.: Locally adaptive structure and texture similarity for image quality assessment. In *Proceedings of the 29th ACM International Conference on Multimedia* (2021), ACM. 3

[DMWS20] Ding K., Ma K., Wang S., Simoncelli E. P.: Image quality assessment: Unifying structure and texture similarity. *CoRR abs/2004.07728* (2020). 3, 5

[DMWS21] DING K., MA K., WANG S., SIMONCELLI E. P.: Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision 129* (2021), 1258–1281. 2, 7

[Fol94] FOLEY J.: Human luminance pattern-vision mechanisms: masking experiments require a new model. *J. Opt. Soc. Am. A 11*, 6 (1994), 1710–19. 2, 4

[FSPG97] FERWERDA J. A., SHIRLEY P., PATTANAIK S. N., GREENBERG D. P.: A model of visual masking for computer graphics. In *Proc. ACM SIGGRAPH* (1997), pp. 143–152. 2, 7

[HLSS20] HOSU V., LIN H., SZIRANYI T., SAUPE D.: Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing 29* (2020), 4041–4056. 3

[JHH*20] JINJIN G., HAOMING C., HAOYU C., XIAOXING Y., REN J. S., CHAO D.: PIPAL: A large-scale image quality assessment dataset for perceptual image restoration. In *Proc. ECCV* (2020), pp. 633–651. 3, 5

[Jia18] JIA S.: EML-NET: an expandable multi-layer network for saliency prediction. *CoRR abs/1805.01047* (2018). 6

[KB15] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. In *Proc. ICLR* (2015). 4

[KWW*21] KE J., WANG Q., WANG Y., MILANFAR P., YANG F.: Musiq: Multi-scale image quality transformer. In *Proc. ICCV* (2021), pp. 5128–5137. 3

[LC10] LARSON E. C., CHANDLER D. M.: Most apparent distortion: full-reference image quality assessment and the role of strategy. *J. Electronic Imaging 19* (2010), 011006. 5

[LCZ*22] LIAO X., CHEN B., ZHU H., WANG S., ZHOU M., KWONG S.: DeepWSD: Projecting degradations in perceptual space to wasserstein distance in deep feature space. In *Proceedings of the 30th ACM International Conference on Multimedia* (2022), ACM. 3, 5

[LF80] LEGGE G. E., FOLEY J. M.: Contrast masking in human vision. *J. Opt. Soc. Am. 70*, 12 (1980), 1458–1471. 2, 4

[LHS19] LIN H., HOSU V., SAUPE D.: KADID-10k: A large-scale artificially distorted IQA database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)* (2019), pp. 1–3. 2, 4

[Lub95] LUBIN J.: A visual discrimination model for imaging system design and development. In *Vision Models for Target Detection and Recognition* (1995), Peli E., (Ed.), World Scientific, pp. 245–283. 2, 4

[MDC*21] MANTIUK R. K., DENES G., CHAPIRO A., KAPLANYAN A., RUFO G., BACHY R., LIAN T., PATNEY A.: FovVideoVDP: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (Proc. SIGGRAPH) 40*, 4 (2021), 1–19. 2, 4, 5, 6, 9

[MFTM01] MARTIN D., FOWLKES C., TAL D., MALIK J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision* (2001), vol. 2, pp. 416–423. 7

[MKRH11] MANTIUK R., KIM K. J., REMPEL A. G., HEIDRICH W.: HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on Graphics (Proc. SIGGRAPH) 30*, 4 (2011), 1–14. 2, 4, 5, 6, 9

[NHKML17] NAH S., HYUN KIM T., MU LEE K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proc. CVPR* (2017), pp. 3883–3891. 7, 9

[PCMS18] PRASHNANI E., CAI H., MOSTOFI Y., SEN P.: Pieapp: Perceptual image-error assessment through pairwise preference. In *Proc. CVPR* (2018), pp. 1808–1817. 3, 6, 9

[PJI*15] PONOMARENKO N., JIN L., IEREMEIEV O., LUKIN V., EGIAZARIAN K., ASTOLA J., VOZEL B., CHEHDI K., CARLI M., BATTISTI F., JAY KUO C.-C.: Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication 30* (2015), 57–77. 5

[SB05] SHEIKH H. R., BOVIK A. C.: Information embedded in images: Joint estimation of information and distortion. In *Proceedings of the IEEE International Conference on Image Processing* (2005), vol. 2, pp. II–702. 2

[SB06] SHEIKH H., BOVIK A.: Image information and visual quality. *IEEE Transactions on Image Processing 15*, 2 (2006), 430–444. 2

[SBG15] SEBASTIAN S., BURGE J., GEISLER W. S.: Defocus blur discrimination in natural images with natural optics. *Journal of Vision 15*, 5 (2015), 16–16. 7

[SWG*09] SAMPAT M. P., WANG Z., GUPTA S., BOVIK A. C., MARKEY M. K.: Complex wavelet structural similarity: A new image similarity index. *IEEE Transactions on Image Processing 18*, 11 (2009), 2385–2401. 2

[SYZ*20] SU S., YAN Q., ZHU Y., ZHANG C., GE X., SUN J., ZHANG Y.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proc. CVPR* (2020), pp. 3664–3673. 3, 6

[SZ15] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR* (2015). 2

[TAKW*19] TURSUN C., ARABADZHIYSKA-KOLEVA E., WERNIKOWSKI M., MANTIUK R., SEIDEL H.-P., MYSZKOWSKI K., DIDYK P.: Luminance-contrast-aware foveated rendering. *ACM Transactions on Graphics (Proc. SIGGRAPH) 38*, 4 (2019). 6, 7

[VRM*18] VOGELS T., ROUSSELLE F., MCWILLIAMS B., RÖTHLIN G., HARVILL A., ADLER D., MEYER M., NOVÁK J.: Denoising with kernel prediction and asymmetric loss functions. *ACM Transactions on Graphics (Proc. SIGGRAPH) 37*, 4 (2018), 1–15. 2

[WA11] WATSON A. B., AHUMADA A. J.: Blur clarified: A review and synthesis of blur discrimination. *Journal of Vision 11*, 5 (2011), 10–10. 7

[WAK*20] WUERGER S., ASHRAF M., KIM M., MARTINOVIC J., PÉREZ-ORTIZ M., MANTIUK R. K.: Spatio-chromatic contrast sensitivity under mesopic and photopic light levels. *Journal of Vision 20*, 4 (2020), 23–23. 7, 10

[Wat93] WATSON A.: Visually optimal DCT quantization matrices for individual images. In *Proc. DCC '93: Data Compression Conference* (1993), pp. 178–187. 2, 3

[WBSS04] WANG Z., BOVIK A., SHEIKH H., SIMONCELLI E.: Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing 13*, 4 (2004), 600–612. 2

[WG84] WILSON H. R., GELB D. J.: Modified line-element theory for spatial-frequency and width discrimination. *J. Opt. Soc. Am. A 1*, 1 (1984), 124–131. 2, 4

[WGY*18] WOLSKI K., GIUNCHI D., YE N., DIDYK P., MYSZKOWSKI K., MANTIUK R., SEIDEL H.-P., STEED A., MANTIUK R. K.: Dataset and metrics for predicting local visible differences. *ACM Transactions on Graphics 37*, 5 (2018), 1–14. 3, 5, 6, 9

[WSB03] WANG Z., SIMONCELLI E., BOVIK A.: Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003* (2003), vol. 2, pp. 1398–1402 Vol.2. 2

[YWS*22] YANG S., WU T., SHI S., LAO S., GONG Y., CAO M., WANG J., YANG Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment, 2022. 3, 6

[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. 7, 9

[ZDL02] ZENG W., DALY S., LEI S.: An overview of the visual optimization tools in JPEG 2000. *Signal Processing: Image Communication 17*, 1 (2002), 85–104. 2, 6, 7

[ZIE*18] Zhang R., Isola P., Efros A. A., Shechtman E., Wang O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. CVPR* (2018). 2, 3, 4, 7

[ZZMZ11] Zhang L., Zhang L., Mou X., Zhang D.: Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing 20*, 8 (2011), 2378–2386. 2