MILO: A Lightweight Perceptual Quality Metric for Image and Latent-Space Optimization

UĞUR ÇOĞALAN, Max Planck Institute for Informatics, Germany
MOJTABA BEMANA, Max Planck Institute for Informatics, Germany
KAROL MYSZKOWSKI, Max Planck Institute for Informatics, Germany
HANS-PETER SEIDEL, Max Planck Institute for Informatics, Germany
COLIN GROTH, Max Planck Institute for Informatics, Germany

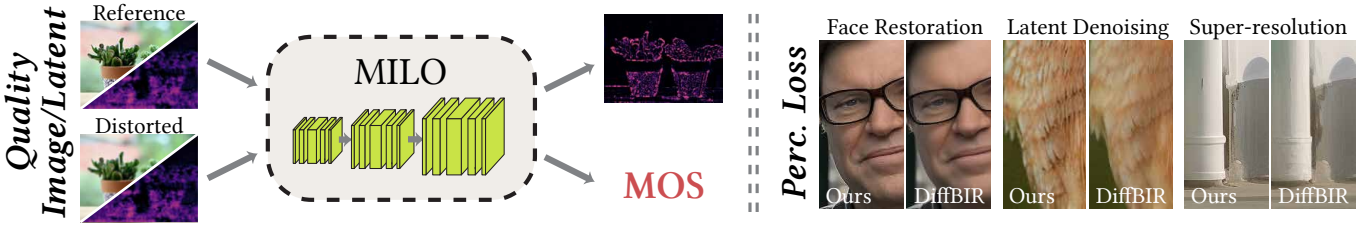

Fig. 1. We present *MILO*, our perceptual metric for full-reference image quality assessment, **solely trained on learned perceptual data**. MILO processes either RGB or VAE-decoded latent input pairs and outputs a perceptual quality score in the MOS range along with a spatial visibility map. When applied as a perceptual loss function, MILO allows through a curriculum learning strategy for spatially guided and content-aware refinement in image restoration tasks and diffusion-based optimization.

We present MILO (Metric for Image- and Latent-space Optimization), a lightweight, multiscale, perceptual metric for full-reference image quality assessment (FR-IQA). MILO is trained using pseudo-MOS (Mean Opinion Score) supervision, in which reproducible distortions are applied to diverse images and scored via an ensemble of recent quality metrics that account for visual masking effects. This approach enables accurate learning without requiring large-scale human-labeled datasets. Despite its compact architecture, MILO outperforms existing metrics across standard FR-IQA benchmarks and offers fast inference suitable for real-time applications. Beyond quality prediction, we demonstrate the utility of MILO as a perceptual loss in both image and latent domains. In particular, we show that spatial masking modeled by MILO, when applied to latent representations from a VAE encoder within Stable Diffusion, enables efficient and perceptually aligned optimization. By combining spatial masking with a curriculum learning strategy, we first process perceptually less relevant regions before progressively shifting the optimization to more visually distorted areas. This strategy leads to significantly improved performance in tasks like denoising, super-resolution, and face restoration, while also reducing computational overhead. MILO thus functions as both a state-of-the-art image quality metric and as a practical tool for perceptual optimization in generative pipelines.

CCS Concepts: • **Computing methodologies** → **Image processing**; *Image representations*; Supervised learning.

Authors' addresses: Uğur Çoğalan, Max Planck Institute for Informatics, Saarbrücken, Germany, ugurcogalan@gmail.com; Mojtaba Bemana, Max Planck Institute for Informatics, Saarbrücken, Germany, mbemana@mpi-inf.mpg.de; Karol Myszkowski, Max Planck Institute for Informatics, Saarbrücken, Germany, karol@mpi-inf.mpg.de; Hans-Peter Seidel, Max Planck Institute for Informatics, Saarbrücken, Germany, hpseidel@mpi-inf.mpg.de; Colin Groth, Max Planck Institute for Informatics, Saarbrücken, Germany, c.groth@nyu.edu.

ACM 0730-0301/2025/12-ART
https://doi.org/10.1145/3763340

Additional Key Words and Phrases: Full-Reference Image Quality Assessment, FR-IQA, Perceptual Metric, Perceptual Loss, Latent Masking

ACM Reference Format:
Uğur Çoğalan, Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel, and Colin Groth. 2025. MILO: A Lightweight Perceptual Quality Metric for Image and Latent-Space Optimization. *ACM Trans. Graph.* 44, 6 (December 2025), 11 pages. https://doi.org/10.1145/3763340

1 INTRODUCTION

For centuries, people have captured real and imagined scenes through photography and painting. Today, visual expression increasingly relies on computer-generated imagery, especially images synthesized by deep neural networks. As demand for high-quality generation grows, so does the need for reliable metrics to assess visual quality as perceived by the human observer and guide optimization in image synthesis and restoration pipelines.

Traditional metrics like Mean Absolute Error (MAE) and Peak Signal-to-Noise Ratio (PSNR) rely on per-pixel comparisons and often fail to align with human perception. Perceptual metrics such as LPIPS [Zhang et al. 2018] and DISTS [Ding et al. 2020] improve on this by leveraging deep features from pre-trained networks like VGG. These methods are trained on human-labeled perceptual data at the patch level, which is costly to collect and may still fall short of capturing perceived quality in the context of full images.

Due to the labor-intensive nature of repeated and controlled human judgment collection, perceptual datasets are limited in scope. For example, the widely used KADID-10k dataset [Lin et al. 2019] contains only 81 reference images, each with 25 distortions at five distortion levels, amounting to over 10,000 image pairs that need annotation. This scarcity is particularly problematic given that perceived quality is strongly dependent on image content and local

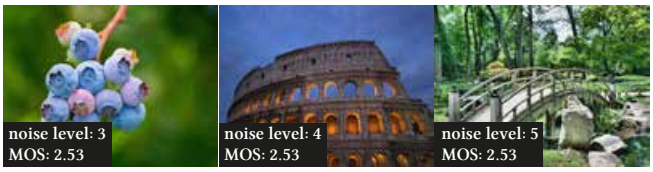

Fig. 2. Three distortion images of the KADID10k dataset that apply significantly different amounts of noise to the image. Yet, the subjective quality (MOS) is rated the same by human observers.

context. As illustrated in Fig. 2, even large variations in distortion strength can result in identical subjective ratings due to masking or content-based tolerance. However, the underlying challenge extends beyond data scarcity: it lies in the absence of models that capture variations in perceptual sensitivity across spatial structure and frequency content. Effective IQA methods must account for these factors to predict human judgments reliably and support optimization tasks that require fine-grained, spatially aware perceptual guidance.

We propose MILO, a compact and efficient multiscale FR-IQA metric that addresses these limitations. MILO is trained using pseudo-MOS labels generated by an ensemble of visual masking-aware metrics applied to synthetically distorted images through a forward model. This strategy enables large-scale supervision without the need for manual annotations. The resulting model predicts both global perceptual scores and spatial visibility maps, which allow for the localization of the distortion visibility based on the underlying content. MILO achieves state-of-the-art performance on common IQA benchmarks while being significantly faster than existing learning-based metrics due to its lightweight structure. Beyond quality prediction, we show that MILO is effective as a perceptual loss in restoration and generative tasks. In particular, our method can operate directly in the latent space of diffusion models. Most contemporary generative pipelines perform image synthesis through iterative refinement of the latent. Applying perceptual guidance in this domain is advantageous, as it avoids repeated decoding into image space and allows perceptual supervision to act on high-level semantic representations. Our method enables spatially resolved loss weighting in latent space, improving the quality and efficiency of diffusion-based optimization. The code for our metric, trained models, and further information are available at https://milo.mpi-inf.mpg.de/. Our contributions summarize as:

- a fast perceptual metric (MILO) that achieves state-of-the-art accuracy in full-reference image quality assessment.
- a pseudo-supervision strategy based on reproducible distortions and a metric ensemble to synthesize large-scale, annotated training data.
- extension of our metric to operate directly in the latent space of diffusion models for efficient, perceptual optimization.
- an effective perceptual loss in restoration tasks through curriculum learning and spatial masking.

2 RELATED WORK

This section explores former research and discusses advancements in image quality assessment and data augmentation techniques.

2.1 Image Quality Assessment

Image quality assessment (IQA) is primarily divided into full-reference (FR) and no-reference (NR) approaches. In this paper, we mainly focus on FR metrics. Traditional FR-IQA methods, such as MAE, PSNR, and SSIM [Wang et al. 2004], rely on direct comparisons between reference and distorted images to evaluate quality. FSIM [Zhang et al. 2011b] incorporates phase congruency and gradient magnitude to better model human perception. FLIP [Andersson et al. 2020] is a task-specific metric designed to assess perceptual differences in rendered graphics with high spatial precision. However, these traditional metrics often fall short of representing visual quality as perceived by humans because they do not account for image content, e.g., high-frequency vs. low-frequency background, or perceptual phenomena such as masking. In recent years, learning-based models [Ding et al. 2021, 2020; Liao et al. 2022; Prashnani et al. 2018; Zhang et al. 2018] have been developed to more accurately predict perceptual quality, leveraging deep image features to better align with human perception. Notably, Zhang et al. [2018] demonstrated that internal image representations from classification networks could be used to assess image similarity. Their Perceptual Image Patch Similarity (LPIPS) index computes image similarity by measuring ℓ_2 distances between features extracted from pre-trained VGG networks. This idea builds on earlier perceptual feature-based losses, such as those used in style transfer and super-resolution networks [Johnson et al. 2016], where VGG features provided a proxy for perceptual content. Ding et al. [2020] introduced the DISTS metric, which evaluates texture and structure similarity by comparing the global statistics—mean, variance, and correlations—of feature pairs, similar to SSIM. Shifting away from deterministic feature comparisons, DeepWSD [Liao et al. 2022] employed the Wasserstein distance, a statistical method for comparing the distributions of features, offering a more holistic approach to IQA. TOPIQ [Chen et al. 2024a] extends this idea further by leveraging token-based representations from transformers, enabling better generalization across image types. More recently, Çoğalan et al. [2024] proposed enhancing FR-IQA by learning visual masking models that modulate the metric input in a self-supervised manner using MOS data. Their masking network improves both traditional and learning-based metrics by integrating perceptual phenomena such as masking—an aspect also explicitly modeled in perceptually grounded methods like HDR-VDP-2 [Mantiuk et al. 2011] and FovVideoVDP [Mantiuk et al. 2021], which incorporate display parameters and foveation but remain limited in general applicability due to their complexity and reliance on calibrated inputs. Despite these advances, many of these methods are trained on relatively small datasets, such as KADID [Lin et al. 2019] and PIPAL [Jinjin et al. 2020], limiting their ability to generalize to larger, more varied image collections. For example, WaDIQaM [Bosse et al. 2017] demonstrated promising performance by combining deep feature regression with patch-wise quality estimation, but it still depends on limited per-image labeled data. However, the reliance on manually annotated datasets —which require extensive human labor— restricts both the size and diversity of training samples. In this paper, we propose a learnable FR-IQA metric that addresses these limitations by leveraging an augmented

dataset with automatically assigned perceptual scores and a lightweight CNN architecture capable of generalizing across diverse image content. Compared to existing perceptually optimized metrics, our method aims to generalize across both natural and synthetic distortions.

2.2 Data Augmentation

Human-labeled datasets are essential for training perceptually aligned IQA models, as demonstrated by PieApp, WaDIQaM, and the work of Çoğalan et al. [2024]. However, creating such datasets is costly and time-consuming, and their limited size restricts generalization across diverse content and distortions. This makes conventional augmentation strategies—common in classification or segmentation—largely inapplicable. Patch sampling [Ahmed et al. 2022; Bosse et al. 2017] and image mixing [Shi et al. 2024] have been used to increase data extent. DSMix blends two distorted images with a binary mask, assuming a linear perceptual relationship between them—an untested and likely unreliable premise for human quality judgments. Similarly, random geometric or photometric transformations [Chen et al. 2024b] aim to improve metric robustness but do not introduce new perceptual content.

Recent work explores the use of multimodal large language models (MLLMs) for text-driven IQA augmentation [Wu et al. 2024; You et al. 2023]. Wu et al. [2024] evaluated various prompting techniques and psychophysical procedures to be used for IQA, finding that current MLLMs fail to reliably model fine-grained perceptual differences, such as subtle color shifts or quality distinctions between similar images. Diffusion models are also promising, capable of generating diverse and semantically consistent images by learning probabilistic distributions [Ho et al. 2020; Saharia et al. 2022]. They have been applied successfully in domains like medical imaging and remote sensing [Ozbey et al. 2023]. Methods like DatasetDM [Wu et al. 2023] enable perception-aware augmentation with corresponding annotations. However, these models still struggle with generating images suitable for reliable quality assessment.

In contrast, we propose a distortion-aware augmentation strategy based on reproducible forward models, generating perceptually diverse reference-distortion pairs. Scores from an ensemble of existing metrics provide pseudo-supervision, eliminating the need for human annotations while ensuring perceptual validity. This strategy allows for massive training of our learning-based FR metric.

3 METRIC FOR IMAGE- AND LATENT-SPACE OPTIMIZATION (MILO)

We propose MILO, a perceptual quality metric that predicts both a global quality score and a spatial visibility map for a given reference-distorted image pair. However, as discussed in Sec. 2.2 existing perceptual datasets are limited in diversity and scale, restricting generalization—particularly for content-sensitive distortions.

To address this limitation, we introduce a two-stage framework. First, we generate a synthetic training set by applying parameterized distortions (e.g., Gaussian blur, additive noise) to diverse images (see Sec. 3.1). Corresponding perceptual scores are assigned via an ensemble of masking-aware full-reference metrics [Çoğalan et al. 2024], yielding dense and content-aware supervision without manual annotations. Second, we train a lightweight multi-scale CNN that predicts both MOS and a pixel-wise visibility map. The architecture (see Fig. 3) combines residual blocks and multiscale features to model content-dependent distortion visibility efficiently (see Sec. 3.2). Our metric generalizes beyond its pseudo-supervision, producing state-of-the-art performance on standard IQA benchmarks.

Our metric is further extended to operate directly on VAE encoded latents to avoid costly image-space decoding (see Sec. 3.3) and employ curriculum learning to schedule loss weighting from low-visibility to high-visibility regions for perceptually aligned optimization (see Sec. 3.4).

3.1 Training Data Augmentation

To augment existing IQA datasets with reliable perceptual annotations, we develop a pseudo-MOS prediction module based on an ensemble of masking-aware FR-IQAs. The goal is to assign perceptual scores to newly generated distorted images without relying on extensive studies with human participants. We leverage the forward distortion model provided by the KADID10k dataset [Lin et al. 2019] to synthetically reproduce traditional distortion types (e.g., Gaussian blur, additive noise, JPEG compression) at controlled levels on arbitrary reference images. This procedure ensures that each generated image follows a known and interpretable degradation process, instead of handling learning-incurred distortions or domain-specific artifacts. The reference images for the data augmentation are randomly chosen from the ImageNet dataset [Deng et al. 2009].

For each distorted image, we compute the outputs of masking-aware versions E-VGG, E-LPIPS, E-DISTS, and E-DeepWSD as proposed in [Çoğalan et al. 2024], due to their strong performance (see Tbl. 1). We aggregate their scores using equal weights to avoid overfitting due to limited human-labeled data. We refer to this combined metric as *Ensemble*, which leverages the complementary strengths of its components and achieves higher predictive accuracy than any single metric alone. This aligns with findings from Lou et al. [2024], where combining diverse IQA metrics led to more stable and reliable predictions.

The resulting pseudo-MOS scores are used to train our multiscale quality metric and are critical to reaching state-of-the-art performance without requiring large-scale human annotations.

3.2 Multiscale CNN Model

We define our perceptual quality model as a multiscale convolutional neural network F that estimates a pixel-wise perceptual mask $M \in [0,1]^{H\times W}$ and a global quality score $S_{\text{raw}} \in \mathbb{R}$ for a given distorted image $Y \in \mathbb{R}^{H\times W\times 3}$ with respect to a reference image $X \in \mathbb{R}^{H\times W\times 3}$. The overall function of the network can be expressed as $(_{\text{raw}}, M) = F(X, Y)$. Unlike many recent deep FR-IQA approaches that rely on large pretrained backbones or increase network capacity, we follow a multiscale architecture to improve performance and interpretability without adding computational burden [Nah et al. 2017; Ranjan and Black 2017]. This design is inspired by prior work, which demonstrated that perceptual error prediction can be improved using spatial weighting informed by human visual masking [Çoğalan

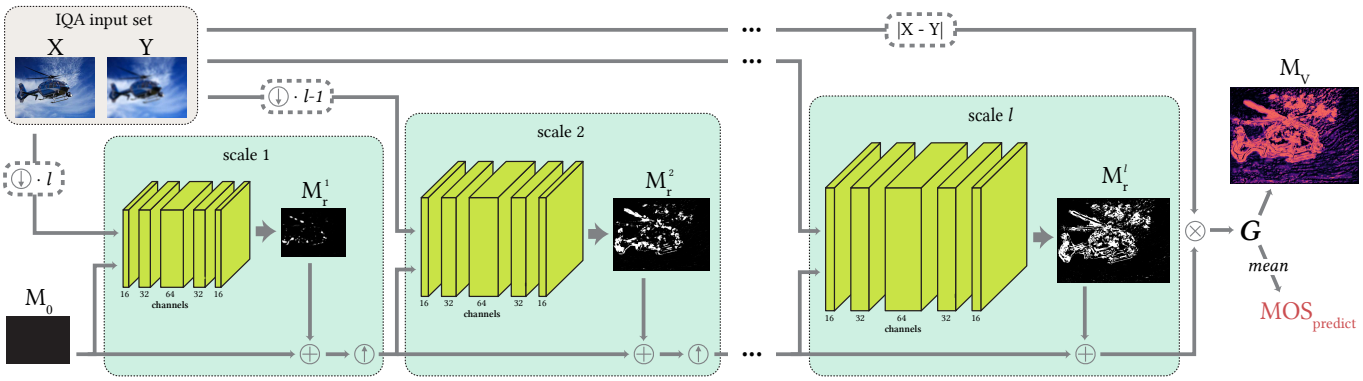

Fig. 3. Overview of the MILO pipeline. The method takes a pair of input images (or their VAE-encoded latents)—reference X and distorted Y—and processes them at multiple scales. At each scale, a shared CNN module predicts a residual visibility mask M_r, which is recursively refined across scales. The final visibility map M_V highlights perceptually salient distortion regions. A complementary global quality score in the MOS scale reflects predicted human judgment of perceptual quality.

et al. 2024; Mantiuk et al. 2011, 2021]. We found a multiscale architecture to be a good choice for this task, as the different scales correspond to different spatial frequency bands. High-frequency distortions, such as impulse noise, may be better captured at finer scales, while low-frequency distortions, such as color changes, correspond more to coarser scales. However, in comparison with classical HVS models that are based on narrowband masking effects measured in controlled settings, natural images can exhibit complex correlations across scales. A multiscale architecture benefits from propagating coarse-level priors to prevent inconsistent visibility estimates between scales.

We apply a multi-level image pyramid using bicubic downscaling to both reference and distorted images:

$$\{X^l, Y^l\}_{l=1}^{L}, \quad \text{where } X^l, Y^l \in \mathbb{R}^{\frac{H}{2^{l-1}} \times \frac{W}{2^{l-1}} \times 3} \tag{1}$$

Here, $l = 1$ denotes the coarsest scale and $l = L$ the original resolution. Each scale level is processed by a shared convolutional subnetwork ϕ_l that finds a residual error mask M_r:

$$M_r^l = \phi_l(X^l, Y^l, M^{l-1}) \tag{2}$$

with $M^0 = 0$. To obtain the input mask for the next scale, the error mask of $l-1$ is combined with the output of ϕ_l and upsampled using bilinear interpolation (Up):

$$M^l = \mathrm{Up}\left(\phi_l(X^l, Y^l, M^{l-1}) + M^{l-1}\right) \tag{3}$$

The residual mask is fed to all scales rather than computed independently per scale because it is not frequency-selective in the classical sense. Instead, it encodes spatially localized, content-dependent visibility as a joint function of local contrast, structure, and texture. Feeding this contextual information forward allows finer scales to leverage the coarse-scale visibility prediction without relearning it.

The CNN at each scale consists of five convolutional layers with channel sizes 16–32–64–32–16. All layers use ReLU activations, except the final layer, which uses a sigmoid to constrain the mask values in $[0, 1]$. The mask of the finest resolution M^L is used to modulate the local error of the input images:

$$\tilde{E} = M^L \odot E, \quad \text{where } E = |X - Y| \tag{4}$$

where $\odot$ denotes element-wise multiplication. The final global error S_{raw} is aggregated over the finest scale:

$$S_{\text{raw}} = \frac{1}{HW} \sum_{i,j} \tilde{E}_{i,j} \tag{5}$$

To ensure comparability with the value range of MOS labels, S_{raw} and $\tilde{E}$ are normalized by a lightweight scalar mapper $G : \mathbb{R} \to \mathbb{R}$, similar to previous work [Çoğalan et al. 2024; Zhang et al. 2018]. The aggregated error S_{raw} is transformed into the predicted MOS score MOS_{predict}, while the final visibility map is obtained by $M_V = G(\tilde{E})$. The small network G is trained jointly with F and learns a mapping from metric scale to perceptual scale.

Compared with the single-scale version of this model (finest scale), incorporating multiple resolutions leads to significantly better results, especially when trained on augmented data. MILO is trained exclusively on pseudo-MOS scores generated by *Ensemble* to guide the learning. Therefore, the training inputs are only the reference and distorted image and the pseudo-MOS score. For further training details, please refer to the supplementary material.

3.3 Latent Space Masking

Aside from RGB inputs, our perceptual masks can equally be derived from images decoded into latent space. Latents are essential for modern generative models like diffusion, which operate on latent representations for efficiency and guidance. Stable diffusion models [Rombach et al. 2022], for instance, perform generation in a compact latent space encoded via a variational autoencoder (VAE). Direct latent-space operation reduces memory and compute overhead by avoiding repeated VAE decoding. Our perceptual masking further enhances this process by preserving spatial structure, enabling targeted restoration in semantically relevant regions. Processing the latent images requires no further modification of the pipeline.

3.4 Curriculum Learning with Perceptual Masking

Image restoration models are typically trained using global quality scores or uniform loss functions, even though different regions of

an image vary significantly in perceptual importance. Our perceptual mask addresses this limitation by providing spatially resolved information about distortion visibility. This enables a curriculum learning strategy [Bengio et al. 2009] in which the network first learns to restore less perceptually critical areas before focusing on regions where distortions are more visible. This gradual increase in task difficulty aligns better with how learning systems adapt and helps the network prioritize regions that matter most for perceived image quality.

We define curriculum learning over the course of training by introducing a scheduling parameter $\alpha \in [0, 1]$ that gradually shifts the loss emphasis from masked (easier, lower distortion visibility) regions to unmasked (difficult, perceptually salient) regions. As before, we define the reference and predicted images as X and Y, respectively, and the perceptual mask by $M \equiv M^L$, which is normalized to $[0, 1]$. The pixel-wise absolute difference is defined by $D = |X - Y|$. Using α, we define a curriculum-weighted L1 loss:

$$\mathcal{L}_{\text{curr}} = [(1 - \alpha)(1 - M) \cdot D + \alpha M \cdot D]_{\text{mean}} \tag{6}$$

The parameter α is scheduled as a function of the current training epoch e and the total number of epochs N. We evaluate two scheduling strategies. The first one is a linear scheduling strategy:

$$\alpha = \frac{e}{N} \tag{7}$$

The second strategy modulates the emphasis by a cosine function:

$$\alpha = \frac{1}{2}\left(1 - \cos\left(\pi \cdot \frac{e}{N}\right)\right) \tag{8}$$

While the linear schedule increases uniformly, the cosine variant introduces a slow start and accelerates mid-training, providing a smoother transition between easy and hard regions. With this flexibility, we aim to enable fine control over the difficulty progression, which we show to be beneficial for stable and perceptually aligned restoration performance in Sec. 5.

4 RESULTS

To validate the performance of our metric, we present quantitative (Sec. 4.1–4.2) and qualitative results (Sec. 4.3).

4.1 Predictions of Image Quality

Table 1 presents the correlations between predicted MOS values and human-labeled ground truth scores across standard FR-IQA datasets. We compare our metric against both traditional and learning-based approaches. Please see the supplementary for more information about the experimental setup. Traditional metrics such as MAE, SSIM, HDR-VDP-2, and ColorVideoVDP show consistently lower performance in predicting perceptual image quality. These methods, while computationally efficient, rely on handcrafted assumptions and may not account for spatial context or complex perceptual phenomena. In contrast, our approach delivers substantially higher correlation with human ratings across all benchmarks. Compared to learning-based metrics, like LPIPS or DeepWSD, our method achieves superior prediction accuracy. Despite relying on a significantly smaller network, MILO outperforms other methods by combining targeted architecture design with semantically meaningful data and an effective augmentation strategy (see Sec. 3.1).

Table 1. Comparison of our MILO metric against existing FR-IQA methods on three standard datasets. MILO_I and MILO_L denote the image-based and latent-based versions of our method, respectively. Numbers in parentheses indicate the number of reference images used for training (# training pairs = # reference images * 5 levels * 25 distortion types); * denotes training with only one random distortion at five distortion levels per reference (# training pairs = # reference images * 5 levels)—please see Sec. 6 for more details. Higher values indicate better quality prediction.

	CSIQ			TID			PIPAL		
Metric	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
PSNR	0.851	0.837	0.645	0.726	0.714	0.540	0.468	0.456	0.314
SSIM	0.848	0.863	0.665	0.697	0.663	0.479	0.550	0.534	0.373
FSIM	0.900	0.913	0.740	0.847	0.789	0.611	0.651	0.617	0.441
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
VGG	0.938	0.952	0.804	0.853	0.820	0.639	0.643	0.610	0.432
LPIPS	0.944	0.929	0.769	0.803	0.756	0.568	0.640	0.598	0.424
DISTS	0.947	0.947	0.796	0.839	0.811	0.619	0.645	0.626	0.445
DeepWSD	0.949	0.961	0.821	0.879	0.861	0.674	0.593	0.584	0.409
FovVideoVDP	0.795	0.821	0.632	0.742	0.727	0.544	0.565	0.509	0.358
ColorVideoVDP	0.885	0.895	0.728	0.864	0.853	0.672	0.625	0.587	0.418
FLIP	0.731	0.724	0.527	0.591	0.537	0.413	0.498	0.442	0.306
WaDIQaM	0.909	0.923	0.827	0.857	0.854	0.655	0.665	0.675	0.484
TOPIQ	0.954	0.961	0.827	0.906	0.908	0.728	0.710	0.693	0.503
E-VGG	0.943	0.942	0.784	0.909	0.892	0.714	0.676	0.650	0.464
E-LPIPS	0.945	0.940	0.783	0.900	0.886	0.703	0.707	0.675	0.489
E-DISTS	0.930	0.916	0.743	0.910	0.899	0.717	0.729	0.701	0.509
E-DeepWSD	0.932	0.934	0.769	0.903	0.892	0.710	0.665	0.632	0.450
Ensemble	0.955	0.952	0.805	**0.921**	**0.911**	**0.738**	0.734	0.704	0.513
MILO_I (1k)	0.954	0.961	0.823	0.887	0.869	0.680	0.730	**0.711**	**0.519**
MILO_I (10k)	0.965	0.967	0.834	0.885	0.866	0.677	0.731	0.708	0.516
MILO_I (20k)	**0.967**	**0.968**	**0.838**	0.885	0.866	0.678	0.731	0.705	0.513
MILO_I (50k)	0.966	0.967	0.835	0.888	0.871	0.683	**0.736**	**0.711**	**0.519**
MILO_I (100k*)	0.962	0.962	0.824	0.885	0.860	0.673	0.728	0.701	0.509
MILO_I (1M*)	0.961	0.959	0.818	0.897	0.876	0.692	0.713	0.690	0.499
MILO_L (100k*)	0.963	0.958	0.817	0.877	0.856	0.668	0.714	0.678	0.489
MILO_L (500k*)	0.960	0.957	0.814	0.878	0.860	0.671	0.715	0.684	0.495
MILO_L (1M*)	0.963	0.960	0.821	0.881	0.862	0.675	0.711	0.683	0.493

We additionally evaluate our proposed *Ensemble*, a predictor based on refined error metrics from recent work [Çoğalan et al. 2024]. *Ensemble* was used as the basis of our pseudo-MOS generation. The results demonstrate how *Ensemble* can achieve strong performance by combining the strengths of its components. Remarkably, although MILO is trained on the *Ensemble* outputs, it generalizes better to unseen data, such as the extensive PIPAL benchmark, where most distortions arise from undertrained neural networks that produce non-deterministic and unpredictable artifacts. This indicates that MILO's multiscale network can capture perceptual structure beyond the representational scope of its training labels.

4.2 Runtime Performance

We further compare the learning-based metrics' inference time, averaging 1000 runs (NVIDIA Quadro RTX 8000 GPU). Images have a resolution of 512×384, and we consider the pure metric processing time (no I/O). As shown in Tbl. 2, our method achieves significantly

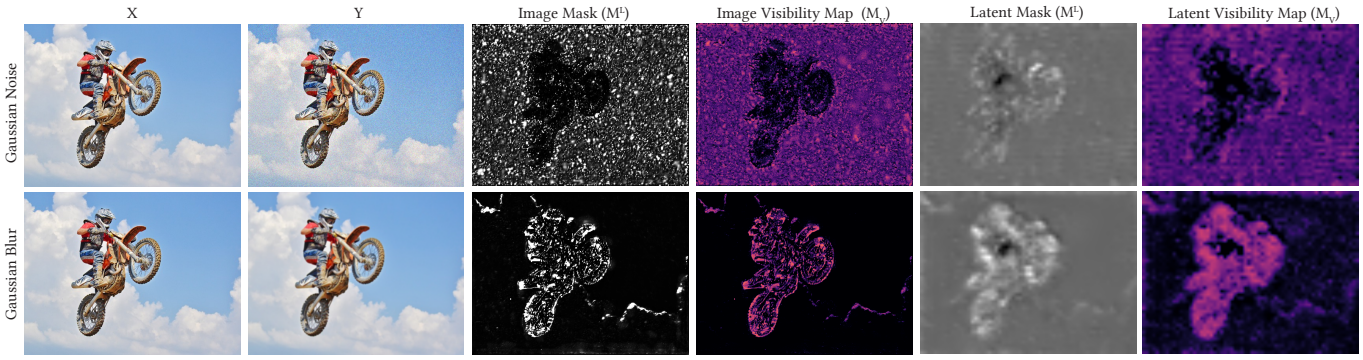

Fig. 4. Visualization of residual masks and visibility maps in both image and latent space for two distortion types: Gaussian noise (top row) and blur (bottom row). In the Image Masks (M^L, Eq. 4), darker regions indicate stronger visual masking. For Gaussian noise, masking increases with contrast, resulting in reduced noise visibility in the high-contrast motorcycle region. Conversely, the bright motorcycle region in the blur example reflects weak masking, as blur is more noticeable along high-contrast edges. The Image Visibility Maps (M_V, Fig. 3) capture the resulting visibility of these distortions. The Latent Masks and Latent Visibility Maps exhibit spatial and magnitude correlations with their image-space counterparts, though their patterns are less visually interpretable. Their smoother, more blurred appearance results from the compact image representation produced by the VAE encoder of Lin et al. [2024], which required 8× upsampling for visualization in this figure.

Table 2. Average inference time per image for learning-based FR metrics.

Method	PieApp	DeepWSD	LPIPS	DISTS	
Time (ms)	72.75	24.53	18.97	21.03	
Method	TOPIQ	WaDIQaM	Ensemble	MILO_I	MILO_L
Time (ms)	357.69	172.06	105.40	**3.94**	**2.16**

faster inference speed, with an average processing time of only 3.94 ms per image. Notably, this remarkable speed advantage is achieved without compromising accuracy (see Sec. 4.1). We also observe that the latent version of our metric (MILO_L) is faster than its image-space counterpart (MILO_I), due to the reduced dimensionality of the compressed latent representation.

4.3 Quality of the Visibility Map

Fig. 4 compares the visibility maps and residual masks produced by MILO_I and MILO_L to evaluate the interpretability and perceptual relevance of the learned error weighting. We consider the same image subjected to two different distortions: Gaussian noise (top row) and blur (bottom row). As expected, noise visibility is more strongly masked in the high-contrast motorcycle region than in the low-contrast sky. Conversely, blur is more perceptible along the high-contrast edges of the motorcycle pattern.

These results suggest that the network has implicitly learned the link between local error visibility and perceptual masking, without requiring explicit supervision. The learned mask acts as a visibility estimator that reflects both the type and spatial structure of the distortion in context with the image content. Notably, the network learns to generate per-pixel visibility maps using only per-image pseudo-MOS supervision. While a similar insight was reported in [Çoğalan et al. 2024], our approach replaces human MOS annotations with synthetic pseudo-MOS data. Leveraging large-scale training on this data, our lightweight multi-scale metric achieves significantly improved image quality prediction accuracy.

5 APPLICATIONS

Beyond image quality prediction, our metric can be applied directly as a perceptual loss in applications that aim for image restoration or optimization. Unlike conventional loss functions (e.g., L1 or SSIM loss), which treat all image regions equally, our approach provides spatially varying guidance by the perceptual mask M^L. Through guidance, we allow restoration networks to prioritize learning in perceptually important regions. This spatial weighting facilitates curriculum learning that adapts the training objective over time, guiding the network to first solve perceptually easier subproblems before addressing more challenging regions. In the following, we demonstrate the practical applicability of our metric to improve on many image restoration and diffusion optimization tasks. For extensive visual results, please also see the provided HTML.

5.1 Loss for Image Restoration

We evaluate the applicability of our metric as a perceptual loss in image restoration tasks. To this end, we integrate it as the training objective in the common Restormer framework [Zamir et al. 2022] and test for restoration performance in image denoising and motion deblurring.

5.1.1 Image Denoising. We train on BSD400 [Martin et al. 2001a] with additive Gaussian noise ($\sigma \in [0, 50]$) and evaluate on five standard benchmarks from Restormer [Zamir et al. 2022]. Fig. 5 shows that our method yields sharper, more detailed outputs than the baseline. Tbl. 3 compares seven training variants: ℓ_1 loss baseline, two full-reference setups using all 25 distortions, two with increased reference diversity but only one distortion per image at five distortion levels, and two curriculum learning variants (1M* model with linear and cosine scheduling). Our loss consistently improves perceptual quality. Notably, using more reference images—even with fewer distortions per image—improves generalization. While this benefit was not reflected in the classic metric evaluation scores like PLCC or SRCC (Tbl. 1), it becomes evident in the downstream restoration

Table 3. Denoising results at different noise levels (σ) measured by SSIM, LPIPS and our *Ensemble* metrics. The baseline uses MAE loss. Numbers in parentheses show the number of training references; * indicates one random distortion at five distortion levels per reference (instead of the full set of 25 distortion types)—please see Sec. 6 for more details. Curriculum learning (curr) uses the *MILO (1M*)* model.

Method	$\sigma = 25$			$\sigma = 50$		
	SSIM↑	LPIPS↓	Ens.↓	SSIM↑	LPIPS↓	Ens.↓
MAE	0.9046	0.0966	0.2582	0.8342	0.1686	0.3615
MILO (10k)	0.9038	0.0887	0.2472	0.8331	0.1525	0.3458
MILO (50k)	**0.9050**	0.0837	0.2493	0.8337	0.1432	0.3487
MILO (100k*)	0.9040	0.0834	0.2451	0.8333	0.1451	0.3427
MILO (1M*)	0.9038	0.0766	0.2385	0.8332	0.1340	0.3373
MILO (curr_{lin})	0.9037	0.0717	0.2325	0.8330	0.1282	0.3324
MILO (curr_{cos})	0.9028	**0.0691**	**0.2323**	**0.8309**	**0.1256**	**0.3313**

performance. Curriculum learning provides further gains, with the cosine scheduling being in slight favour.

5.1.2 Motion Deblurring. For motion deblurring, we use the GoPro dataset [Nah et al. 2017] for both training and evaluation. We apply only the best-performing version of our metric from the previous experiments—trained with 1M different images at five distortion levels and cosine curriculum learning—as it showed clear advantages over other configurations. Visual comparisons are provided in Fig. 6. Supervision with our metric leads to significantly better performance compared to ℓ_1 loss results (see Tbl. 4).

5.2 Latent Space Optimization

Latent-space optimization has become a powerful strategy for guiding generative models, especially in the context of diffusion-based image restoration. DiffBIR [Lin et al. 2024] is a recent state-of-the-art blind image restoration framework. It decouples the restoration process into two stages. First, a task-specific restoration module removes degradations in the image. Second, a generative diffusion model, IRControlNet [Lin et al. 2024], reconstructs perceptual details in the latent space using stable diffusion.

We extend DiffBIR by integrating MILO_L (1M* version) into the framework's second stage. During inference, MILO operates as a latent-space guidance mechanism for the optimization process. Unlike the original region-adaptive MSE guidance of DiffBIR, which requires decoding the latent representation to RGB space at each step, our method applies masking directly in latent space. This avoids repeated VAE decoding and enables faster, memory-efficient optimization. MILO computes a spatial distortion mask (M^L) based on perceptual differences between the latent representations of the distorted image and the cleaned image of the restoration module of the first stage. This mask is then used to modulate the loss function with a curriculum learning scheme (Eq. 8, Sec. 3.4), progressively focusing training on perceptually salient regions. Our method requires no retraining of the diffusion module and integrates seamlessly into the existing inference process of DiffBIR.

We apply our approach to multiple blind restoration tasks: denoising, super-resolution, and face restoration. Visual comparisons are provided in Fig. 7 and in the supplementary material. For quantitative analysis, we employ the no-reference metrics TOPIQ [Chen et al. 2024a], CLIP-IQA [Wang et al. 2023], and MUSIQ [Ke et al. 2021] since no reference image is available during inference. Results in Tbl. 5 are listed as an average score for all datasets. For more detailed analysis of different noise levels and individual datasets, please refer to the supplementary material.

Table 4. Quantitative results for the motion deblurring task. Our method employs a curriculum learning scheme and is compared to the baseline MAE loss.

Model	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**	**Ensemble ↓**
MAE	29.1441	0.8862	0.1436	0.5600
MILO	**29.6523**	**0.8965**	**0.1293**	**0.5462**

5.2.1 Blind Denoising. For the blind diffusion-based image denoising, we evaluate our method on five standard datasets: Set12 [Zhang et al. 2017], BSD68 [Martin et al. 2001b], Urban100 [Huang et al. 2015], Kodak24 [Franzen 1999], and McMaster [Zhang et al. 2011a]. Here, we show the results for additive white Gaussian noise with $\sigma = 50$. Visually, MILO preserves sharper structures and generates more detailed textures. In contrast, the outputs from the MSE-based guidance tend to appear oversmoothed and lose high-frequency detail. Quantitative results further support these observations. Our method consistently outperforms the baseline in all metrics, supporting the qualitative observations of improved quality.

5.2.2 Super-Resolution. We evaluate blind super-resolution using the three real-world datasets from prior work: the real image dataset of DiffBIR [Lin et al. 2024], the RealSR dataset [Cai et al. 2019], and the DIV2K-val dataset [Agustsson and Timofte 2017]. The qualitative results show that our method recovers finer details and better contrast. Edges appear sharper and textures more defined, whereas the original MSE-based guidance produces slightly blurrier outputs, partially lacking local detail. The quantitative evaluations confirm the visual observations. Our method outperforms the baseline across all datasets and enables more perceptually accurate restoration without retraining of the diffusion pipeline or runtime compromises.

5.2.3 Face Restoration. We evaluate face restoration performance using the LFW [Huang et al. 2008] and Wider [Zhou et al. 2022] datasets. Visually, a clear difference can be seen. Our method preserves facial structure more accurately, enhances fine-scale details such as wrinkles and facial hair, and avoids the over-smoothing observed with MSE-based loss. Surprisingly, quantitative results in Tbl. 5 do not support these findings. Although we could not observe visually objectionable failure cases, the no-reference metrics get worse, which is a known problem in the domain [Hu et al. 2025; Voznesensky et al. 2022].

6 ABLATIONS

In addition to ablations discussed before, such as the scheduling scheme of curriculum learning or the image domain of the metric (RGB vs. latent), we report the results of two further experiments.

Influence of Dataset Size. To evaluate the influence of dataset size, we generated synthetic training data from arbitrary images using

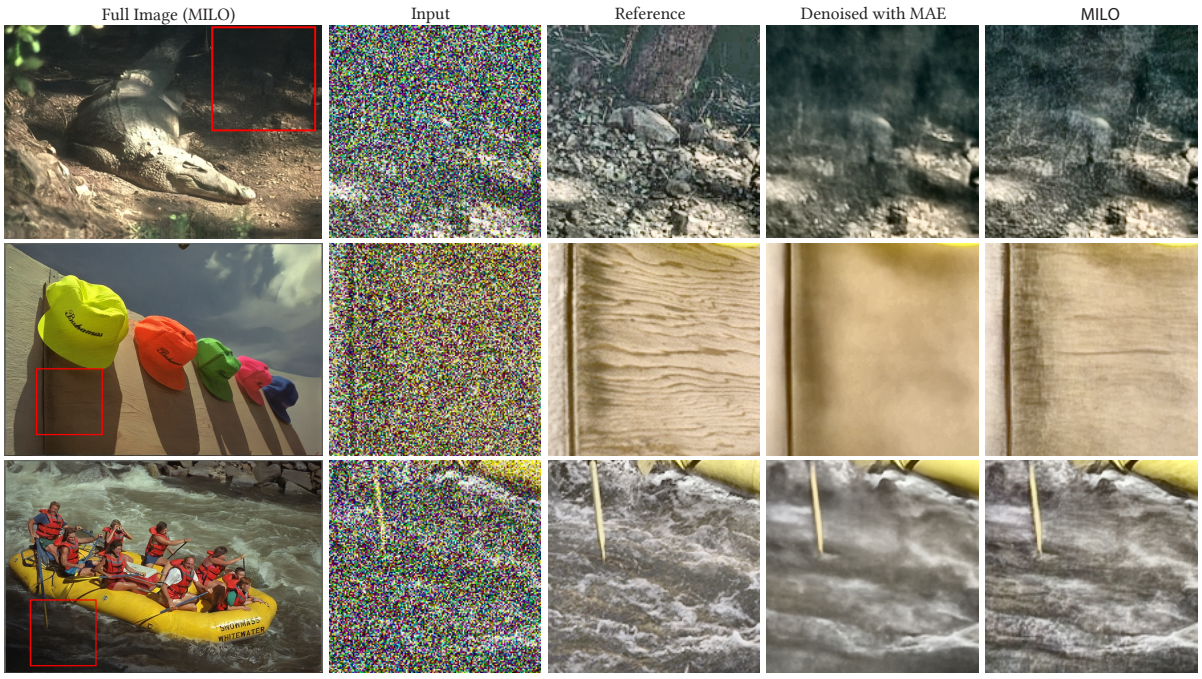

Fig. 5. Visual comparison for the denoising task using Restormer. Our metric-based supervision preserves sharper details and reduces oversmoothing compared to training with standard ℓ_1 loss. The denoising with Restormer gives more importance to the darker regions, and as a result, we can get sharper details under low-light conditions. Here, we apply local tonemapping [Farbman et al. 2008] to the cropped insets for better visualisation.

Fig. 6. Visual comparison for the deblurring task using Restormer.

Table 5. Quantitative results for the three diffusion-based blind optimization tasks evaluated by established no-reference metrics. The baseline is the standard DiffBIR implementation. Here, we use the 1M* model of MILO with curriculum learning.

Method	Denoising			Super-Resolution			Face Restoration		
	TOPIQ↑	CLIP-IQA↑	MUSIQ↑	TOPIQ↑	CLIP-IQA↑	MUSIQ↑	TOPIQ↑	CLIP-IQA↑	MUSIQ↑
DiffBIR	0.6651	0.6895	68.3026	0.6685	0.7238	68.9008	**0.7584**	**0.8015**	**75.8715**
MILO	**0.6731**	**0.7097**	**69.0815**	**0.6997**	**0.7529**	**70.1442**	0.6933	0.7516	75.8026

our distortion forward model and ensemble labeling strategy. As shown in Tbl. 1, increasing the number of reference images also improves the performance of the metric. For static metric correlation analysis, performance saturates after about 10,000 reference images. However, this saturation is not observed when MILO is used as a perceptual loss in downstream tasks such as denoising. Interestingly, in this setting, performance continues to improve with more reference images, indicating that data diversity might be meaningful beyond the saturation point of IQA correlations.

All Distortions vs. Randomized Samples. We further examine the necessity to apply all distortions to each reference image. We compare full distortion coverage (25 distortions per reference) against having one randomly sampled distortion per reference and more references instead (see Tbl. 3). Results indicate an increase in restoration performance in the randomized scheme, while the metric correlations are comparable (see Tbl. 1). These results suggest that the diversity of reference images is more significant than full distortion training to achieve good restoration performance. Still, we found

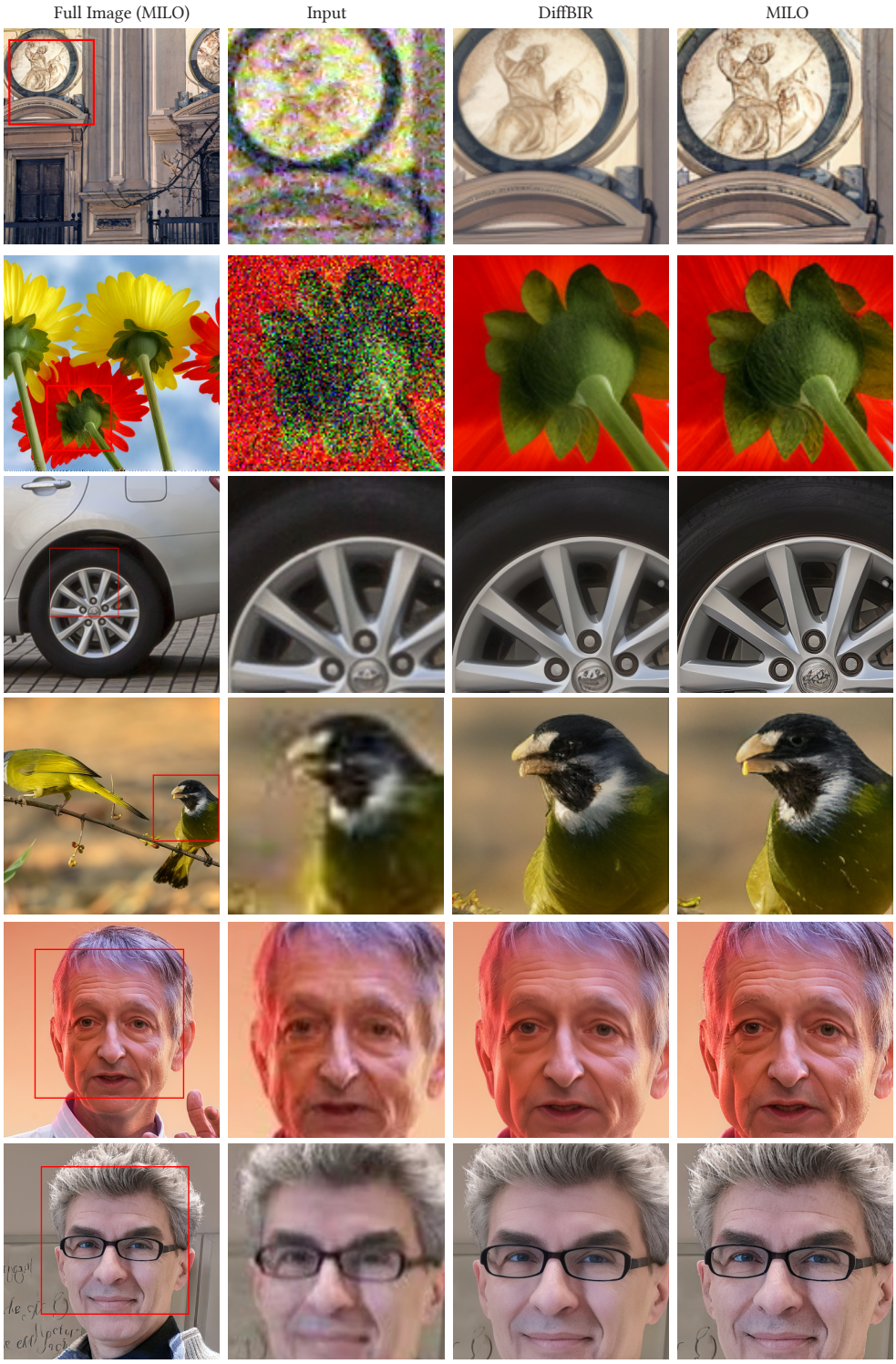

Fig. 7. Visual comparison of using MILO for perceptually guided optimization in comparison to the original loss of DiffBIR [Lin et al. 2024]. Comparisons are given for diffusion-based blind denoising (first two rows), blind super-resolution (two middle rows), and face restoration (last two rows).

that it is important to apply all distortion levels for each distortion type, so that the network can learn non-linearity in distortion perception versus its magnitude (cf. Fig. 1 in the supplementary).

7 DISCUSSION

MILO presents a new, perceptually accurate image quality metric that leverages pseudo-MOS labeling, curriculum learning, and a multiscale CNN architecture. In the following, we would like to discuss and give further intuition about the core behaviour and functionality of MILO.

Efficiency of the Pipeline. The pipeline was shown to be effective despite its simplicity. We attribute this effectiveness to the fact that the CNN learns core behaviors of the HVS directly from data. For that, it does not require explicit psychophysical calibration. The multiscale residual masks effectively emulate spatial-frequency dependencies in a multi-dimensional manner. In comparison with psychophysic-based metrics, MILO (like many other learnable metrics) is not conditioned by distance to display or display characteristics. In the wild, this information is often not available. Learning from diverse image statistics, therefore, makes the method more robust and adaptable.

Without sharing the mask between scales the network fails to converge, given the results of additional experiments. The guidance from the intermediate masks, is needed to follow the intended masking cues, and to make informed decisions during the processing stages.

Furthermore, MILO is designed to predict relative perceptual differences rather than absolute detection thresholds, which are more beneficial for restoration guidance.

Relation of the Design to the Human Visual System. The overall pipeline design is inspired by and aligns with established properties of the human visual system (HVS), though it remains simpler than full psychophysical models. The pyramid structure follows the HVS principle of coarse-to-fine analysis. MILO further reflects spatially selective integration by modulating local error before pooling. This follows human behavior, not integrating error uniformly, but instead being attracted to perceptually salient regions. When collecting MOS data, which we replicate in our pseudo-MOS data, strong distortions are likely to attract attention, modulated by their perceptibility due to visual masking. Near-threshold distortions are more likely to be perceived in salient image content that is more likely to be actively attended. Unlike fixed hand-crafted models, where edges are always strong maskers, MILO's learned masking is distortion-dependent: edges mask noise effectively but do not mask blur, where distortions become even more visible (see Fig. 4). This adaptive behavior better matches how the HVS responds under different distortion types.

Edge-enhanced reconstruction. The reconstructed images of restoration, particularly from denoising and deblurring, show a behaviour similar to edge sharpening and higher-contrast. We hypothesize that in our pseudo-MOS data, sharp edges and higher-contrast textures are promoted, leading to higher MOS values. This behavior could be learned from data involving blur distortion or compression artifacts, where sharper regions are associated with better-perceived quality. Note that MILO does not support naive contrast enhancement—for example, the contrast associated with noise or blockiness, which in pseudo-MOS data corresponds to low image quality, is not enhanced.

8 LIMITATIONS AND FUTURE WORK

While MILO delivers strong performance in both image quality prediction and optimization tasks, it has limitations. The dataset used for training the *Ensemble* components contains human scores collected in uncontrolled, online settings. As a result, the quality of the augmented training data used for our visual masking model is inherently limited by the noise and variability in the individual viewing conditions, such as distance to the display and ambient room lighting. The MILO visibility map is not calibrated using any threshold-based visibility data, unlike some existing metrics that attempt this calibration [Mantiuk et al. 2011, 2021]. However, accurately predicting per-pixel image differences as perceived by human observers remains an open challenge for any existing metric. Instead, in this paper, we focus on relative spatial differences in distortion visibility across varying image content, which provide meaningful information for both network training and test-time guidance in diffusion-based generative tasks. In such generative scenarios, MILO cannot correct for failures of the underlying generative model; if the diffusion process produces semantically incorrect outputs, the metric has limited ability to guide recovery.

As future work, we plan to use generative models to augment data when explicit forward models are unavailable, such as for complex distortions in PIPAL. Lastly, we aim to extend MILO to a no-reference formulation by integrating content-aware perceptual priors.

9 CONCLUSION

We presented MILO, a fast and accurate multiscale metric for full-reference image quality assessment, trained by pseudo-MOS supervision without human annotations. Beyond prediction, MILO serves as an effective perceptual loss in both image and latent space. Applied at inference time to diffusion-based restoration, it enables efficient and perceptually guided optimization, improving quality in tasks like denoising, super-resolution, and face restoration.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the Saarland/Intel Joint Program on the Future of Graphics and Media.

REFERENCES

Eirikur Agustsson and Radu Timofte. 2017. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1122–1131. https://doi.org/10.1109/CVPRW.2017.150

Nisar Ahmed, HM Shahzad Asif, Abdul Rauf Bhatti, and Atif Khan. 2022. Deep ensembling for perceptual image quality assessment. *Soft Computing* 26, 16 (2022), 7601–7622.

Pontus Andersson, Jim Nilsson, Tomas Akenine-Möller, Magnus Oskarsson, Kalle Åström, and Mark D Fairchild. 2020. FLIP: A Difference Evaluator for Alternating Images. *Proc. ACM Comput. Graph. Interact. Tech.* 3, 2 (2020), 15–1.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *International Conference on Machine Learning*. 41–48.

Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. 2017. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing* 27, 1 (2017), 206–219.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. 2019. Toward Real-World Single Image Super-Resolution: A New Benchmark and A New Model. *CoRR* abs/1904.00523 (2019). arXiv:1904.00523 http://arxiv.org/abs/1904.00523

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2024a. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing* (2024).

Zewen Chen, Haina Qin, Juan Wang, Chunfeng Yuan, Bing Li, Weiming Hu, and Liang Wang. 2024b. PromptIQA: Boosting the Performance and Generalization for No-Reference Image Quality Assessment via Prompts. *arXiv:2403.04993* (2024).

Uğur Çoğalan, Mojtaba Bemana, Hans-Peter Seidel, and Karol Myszkowski. 2024. Enhancing image quality prediction with self-supervised visual masking. In *Computer Graphics Forum*, Vol. 43. Wiley Online Library, e15051.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Keyan Ding, Yi Liu, Xueyi Zou, Shiqi Wang, and Kede Ma. 2021. Locally Adaptive Structure and Texture Similarity for Image Quality Assessment. In *Proceedings of the 29th ACM International Conference on Multimedia*. ACM. https://doi.org/10.1145/3474085.3475419

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. 2020. Image Quality Assessment: Unifying Structure and Texture Similarity. *CoRR* abs/2004.07728 (2020). https://arxiv.org/abs/2004.07728

Zeev Farbman, Raanan Fattal, Dani Lischinski, and Richard Szeliski. 2008. Edge-preserving decompositions for multi-scale tone and detail manipulation. *ACM Trans. Graph.* 27, 3 (Aug. 2008), 1–10. https://doi.org/10.1145/1360612.1360666

Rich Franzen. 1999. Kodak Lossless True Color Image Suite. http://r0k.us/graphics/kodak/. Online; accessed 20 May 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33. 6840–6851.

Peng Hu, Chunming He, Lei Xu, Jingduo Tian, Sina Farsiu, Yulun Zhang, Pei Liu, and Xiu Li. 2025. IQPFR: An Image Quality Prior for Blind Face Restoration and Beyond. *arXiv preprint arXiv:2503.09294* (2025).

Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*. Erik Learned-Miller and Andras Ferencz and Frédéric Jurie, Marseille, France. https://inria.hal.science/inria-00321923

Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. 2015. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5197–5206.

Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S. Ren, and Dong Chao. 2020. PIPAL: A Large-Scale Image Quality Assessment Dataset for Perceptual Image Restoration. In *Proc. ECCV*. 633–651.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*. Springer, 694–711.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5148–5157.

Xingran Liao, Baoliang Chen, Hanwei Zhu, Shiqi Wang, Mingliang Zhou, and Sam Kwong. 2022. DeepWSD: Projecting Degradations in Perceptual Space to Wasserstein Distance in Deep Feature Space. In *Proceedings of the 30th ACM International Conference on Multimedia*. ACM. https://doi.org/10.1145/3503161.3548193

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. 2019. KADID-10k: A Large-scale Artificially Distorted IQA Database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. 1–3. https://doi.org/10.1109/QoMEX.2019.8743252

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. 2024. DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior. arXiv:2308.15070 [cs.CV]

Jianxun Lou, Xinbo Wu, Padraig Corcoran, Paul L. Rosin, and Hantao Liu. 2024. TranSalNet+: Distortion-aware saliency prediction. *Neurocomputing* 600 (2024), 128155. https://doi.org/10.1016/j.neucom.2024.128155

Rafał Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. 2011. 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.

Rafał K Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. 2021. FovVideoVDP: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 40, 4 (2021), 1–19.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. 2001a. 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*, Vol. 2. 416–423.

David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. 2001b. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, Vol. 2. IEEE, 416–423.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proc. CVPR*. 3883–3891.

Mert Ozbey, Radu Timofte, Bahadir Gunturk, and Shai Avidan. 2023. Restoring Fine Details in Low-Resolution Images with Diffusion Models. *arXiv preprint arXiv:2301.00483* (2023).

Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. 2018. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proc. CVPR*. 1808–1817.

Anurag Ranjan and Michael J. Black. 2017. Optical Flow Estimation Using a Spatial Pyramid Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2720–2729. https://doi.org/10.1109/CVPR.2017.291

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

Chitwan Saharia, William Chan, Jonathan Ho, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022. Image Super-Resolution via Iterative Refinement. *arXiv preprint arXiv:2204.06125* (2022).

Jinsong Shi, Pan Gao, Xiaojiang Peng, and Jie Qin. 2024. DSMix: Distortion-Induced Sensitivity Map Based Pre-training for No-Reference Image Quality Assessment. arXiv:2407.03886 [cs.CV] https://arxiv.org/abs/2407.03886

Aleksander S Voznesensky, Aleksandr M Sinitca, Evgeniy D Shalugin, Sergei A Antonov, and Dmitrii I Kaplun. 2022. No-Reference Metrics for Images Quality Estimation in a Face Recognition Task. In *International Conference on Actual Problems of Applied Mathematics and Computer Science*. Springer, 462–474.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 2555–2563.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

Tianhe Wu, Kede Ma, Jie Liang, Yujiu Yang, and Lei Zhang. 2024. A comprehensive study of multimodal large language models for image quality assessment. *arXiv preprint arXiv:2403.10854* (2024).

Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. 2023. Datasetdm: Synthesizing data with perception annotations using diffusion models. *Advances in Neural Information Processing Systems* 36 (2023), 54683–54695.

Zhiyuan You, Zheyuan Li, Jinjin Gu, Zhenfei Yin, Tianfan Xue, and Chao Dong. 2023. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. *arXiv preprint arXiv:2312.08962* (2023).

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proc. CVPR*. 5728–5739.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. 2017. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing* 26, 7 (2017), 3142–3155.

Lei Zhang, Xiaolin Wu, Antoni Buades, and Xin Li. 2011a. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *Journal of Electronic imaging* 20, 2 (2011), 023016–023016.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. 2011b. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing* 20, 8 (2011), 2378–2386. https://doi.org/10.1109/TIP.2011.2109730

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proc. CVPR*.

Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. 2022. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems* 35 (2022), 30599–30611.