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

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

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 — Supplementary Material —. *ACM Trans. Graph.* 44, 6 (December 2025), 5 pages. https://doi.org/10.1145/3763340

1 TRAINING DETAILS OF MILO

The training of our model was conducted using the AdamW [Loshchilov and Hutter 2019] optimizer with default parameters β_1 and β_2, and a weight decay of 10^{-6}. We used a batch size of 16 and employed a MultiStepLR scheduler with a step size of 1000 iterations. The proposed metric is implemented using the PyTorch framework [Paszke et al. 2019]. In our implementation we set $L = 4$. Training was performed on an NVIDIA Quadro RTX 8000 GPU for approximately one day, when trained with 10,000 reference images along with all corresponding distorted versions. The loss function applied was the L2 loss between the ground truth MOS and the predicted MOS, formally defined as:

$$\mathcal{L} = \left\| \mathrm{MOS}_{\mathrm{pred}} - \mathrm{MOS}_{\mathrm{gt}} \right\|_2^2$$

Model checkpoints were saved every 100,000 iterations, and training was terminated when the maximum correlation on the Kadid training data split was reached. To create the augmented perceptual dataset, we used the ImageNet dataset [Deng et al. 2009]. Each image is cropped to the largest possible square region and then resized to a resolution of 256×256.

For training the latent space masking, the perceptual dataset is encoded by the VAE of stable diffusion models [Rombach et al. 2022]. This encoding process preserves essential perceptual and semantic information when transforming the RGB images into a lower-dimensional latent representation. For the latent masking, the same training setup and model architecture are used as for image-space masking.

In terms of training time, $\mathrm{MILO_I}$ requires approximately 2.5 hours for 100,000 iterations, whereas $\mathrm{MILO_L}$ completes the same number of iterations in just 26 minutes.

2 FR-IQA EXPERIMENTATION

2.1 Experimental Setup

We employ a broad set of full-reference image quality assessment (FR-IQA) metrics to predict perceptual scores, which are then correlated with ground-truth human opinion scores across multiple datasets. The evaluated methods include traditional metrics such as MAE, PSNR, and SSIM [Wang et al. 2004], FSIM [Zhang et al. 2011], FLIP [Andersson et al. 2020], and HDR-VDP-2 [Mantiuk et al. 2011], as well as learning-based approaches like VGG [Johnson et al. 2016], LPIPS [Zhang et al. 2018], DISTS [Ding et al. 2020], DeepWSD [Liao et al. 2022], TOPIQ [Chen et al. 2024], PieAPP [Prashnani et al. 2018], and WaDIQaM [Bosse et al. 2017]. We additionally include the enhanced versions of selected metrics using the visual masking model of Çoğalan et al. [2024]. The performance of all metrics is assessed on four well-established IQM datasets: CSIQ [Larson and Chandler 2010], TID2013 [Ponomarenko et al. 2015], KADID-10k [Lin et al. 2019], and PIPAL [Jinjin et al. 2020]. CSIQ and TID2013 primarily feature synthetic distortions, with dataset sizes ranging from 1k to 3k images. The KADID-10k dataset consists of 81 pristine images and applies 25 traditional distortions, like Gaussian blur or JPEG compression, at five distinct levels. In contrast, PIPAL is the most comprehensive dataset, containing 23k images with a diverse set of distortions. Each reference image in PIPAL features 116 distortions, including 19 generated by GAN-based methods.

For training the learnable metrics, we split the more comprehensive KADID dataset randomly by reference image, using an 80-20 ratio for training and testing. The training is performed separately and tested on the disjoint test set as well as on the other three datasets, respectively. For the data augmentation, we synthesize arbitrary distortion counterparts using the forward model of KADID for random clean images out of the ImageNet dataset [Deng et al. 2009]. The corresponding MOS scores are predicted using the metric ensemble of the masking-aware E-metrics of Çoğalan et al. [2024] (see Sec. 3.1 of the main paper).

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

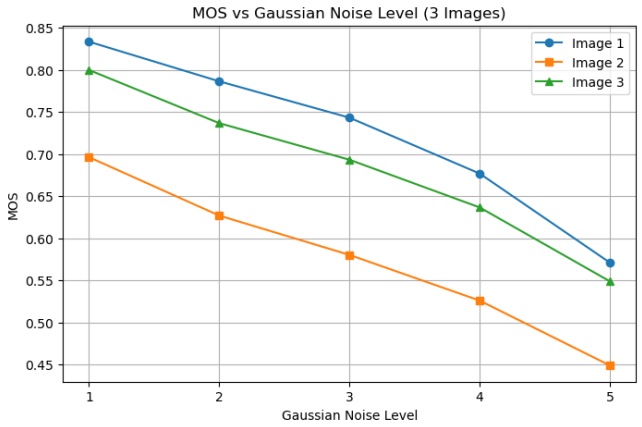

Fig. 1. Visualization of predictions of three representative images under different levels of noise. Higher levels represent a stronger distortion intensity.

For each dataset, three evaluation metrics are used: Spearman's rank correlation coefficient (SRCC), Pearson's linear correlation coefficient (PLCC), and Kendall's rank correlation coefficient (KRCC). PLCC quantifies the accuracy of predictions, SRCC assesses their monotonicity, and KRCC measures ordinal association. Since PLCC captures linear correlation, it requires both the metric output and MOS to be on the same scale. To ensure comparability, metric scores are mapped to MOS values using a four-parameter logistic function, following standard IQM practices [Ding et al. 2020; Liao et al. 2022].

2.2 Distortion Levels

In the paper, we employ the forward distortion model of KADID10k [Lin et al. 2019]. In the work of Lin et al. [2019], the authors define five levels per distortion type, e.g., setting five gradularly increasing kernel sizes for gaussian blur. These predefined intensities (i.e. levels) are also used for the training of MILO to account for varying severity of image artifacts. Given that these different levels are used, it can be interesting to evaluate how the metric interprets human quality judgements under different levels of the same distortion. Such evaluation is given in Fig. 1 for three representative images. While the noise level is linearly increasing, the MOS relationship is non-linear, indicating that a more complex relationship was learned by the metric. Also, we can observe slightly different curves per image, suggesting that the MOS is content dependent. At the same time, the overall trend is always decreasing MOS with higher level, which follows our natural understanding of distortion behaviour.

3 DETAILED APPLICATION RESULTS

This section provides extended quantitative results for the application of our proposed perceptual metric as a loss function in image restoration tasks. Here, we report results across four key domains: image denoising, blind diffusion-based denoising, single-image super-resolution, and face restoration. In each case, we report performance on standard benchmarks and compare against state-of-the-art. The main paper demonstrated that integrating our metric as supervision leads to consistent improvements in perceptual image quality.

Image Denoising. We use the BSD400 dataset [Martin et al. 2001] for training and evaluate on five standard benchmarks as in the Restormer study [Zamir et al. 2022]. Synthetic white Gaussian noise with standard deviations $\sigma = 15$, 25, and 50 is added during training. Table 1 reports results across the three noise levels. Our metric-guided model achieves sharper reconstructions and enhanced detail preservation, outperforming the ℓ_1, particularly under high-noise conditions. Only for the simple, not perceptually accurate method PSNR, this difference is not highlighted. When applying the LPIPS loss, the results exhibit checkerboard-like artifacts, which explains the relatively low quantitative performance.

Blind Diffusion-based Denoising. To test the flexibility of our loss in blind denoising, we apply it in a stable diffusion-based pipeline. The denoising network is trained without access to the noise level, and the model is supervised using our perceptual metric. Table 2 shows results for the same $\sigma = 15$, 25, and 50 settings. Despite the lack of explicit noise information, our loss improves restoration quality robustly across all no-reference metrics.

Super-Resolution. We evaluate our loss on three super-resolution benchmarks: Real [Lin et al. 2024], RealSR [Cai et al. 2019], and DIV2K-val [Agustsson and Timofte 2017]. The task involves 4× upscaling, and we supervise the DiffBIR architecture [Lin et al. 2024] using our perceptual loss. Table 3 summarizes the results. Across all datasets, our model produces more natural details compared to the MAE-based standard loss, also reflected in improved scores from metrics such as CLIP-IQA and MUSIQ.

Face Restoration. We test our method on two face datasets—LFW [Huang et al. 2008] and Wider [Zhou et al. 2022]—for blind face restoration. As noted in the main paper, existing no-reference metrics often penalize perceptually improved restorations in this setting. Tables 4 report results separately for LFW and Wider. Models trained with our loss produce visually superior results (please see the provided HTML file), particularly in restoring detailed facial features. However, most no-reference metrics assign lower scores to these outputs, reflecting a known misalignment between NR-IQA scores and perceptual quality in face restoration tasks [Hu et al. 2025; Voznesensky et al. 2022].

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

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.

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. 2024. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing* (2024).

Table 1. Quantitative results for the denoising task at different noise levels (σ) given by multiple metrics and our ensemble. The baseline is standard MAE loss, which is also the method that is applied in the original Restormer framework. Numbers in parentheses indicate the number of reference images used for training (each paired with 25 distortion types at 5 levels). Each reference image is paired with all 25 distortions unless marked with *, where only one random distortion per reference is used. The curriculum learning variants are based on the *MILO (1M*)* model.

Method	$\sigma = 15$				$\sigma = 25$				$\sigma = 50$			
	PSNR↑	SSIM↑	LPIPS↓	Ens.↓	PSNR↑	SSIM↑	LPIPS↓	Ens.↓	PSNR↑	SSIM↑	LPIPS↓	Ens.↓
MAE	34.42	0.9391	0.0598	0.2007	31.98	0.9046	0.0966	0.2582	28.86	0.8342	0.1686	0.3615
LPIPS	23.15	0.5506	0.0343	0.1942	19.64	0.4061	0.0596	0.2765	15.70	0.2478	0.1245	0.3963
MILO (10k)	34.28	0.9383	0.0553	0.1940	31.84	0.9038	0.0887	0.2472	28.69	0.8331	0.1525	0.3458
MILO (50k)	34.40	0.9392	0.0532	0.1962	31.96	0.9050	0.0837	0.2493	28.81	0.8337	0.1432	0.3487
MILO (100k*)	34.22	0.9384	0.0526	0.1944	31.78	0.9040	0.0834	0.2451	28.65	0.8333	0.1451	0.3427
MILO (1M*)	34.26	0.9383	0.0484	0.1897	31.82	0.9038	0.0766	0.2385	28.66	0.8332	0.1340	0.3373
MILO (curriculum *linear*)	34.26	0.9382	0.0454	0.1859	31.82	0.9037	0.0717	0.2325	28.70	0.8330	0.1282	0.3324
MILO (curriculum *cosine*)	34.22	0.9378	0.0436	0.1860	31.76	0.9028	0.0691	0.2323	28.62	0.8309	0.1256	0.3313

Table 2. Comparison of diffusion-based blind denoising using the original DiffBIR framework and our extended version with latent masking from MILO. Results are reported for different noise levels (σ) using the established no-reference metrics TOPIQ, CLIP-IQA, and MUSIQ.

Method	$\sigma = 15$			$\sigma = 25$			$\sigma = 50$		
	TOPIQ↑	CLIP-IQA↑	MUSIQ↑	TOPIQ↑	CLIP-IQA↑	MUSIQ↑	TOPIQ↑	CLIP-IQA↑	MUSIQ↑
DiffBIR	0.6955	0.7303	69.8351	0.6844	0.7160	69.3131	0.6651	0.6895	68.3026
MILO	0.6972	0.7433	70.0311	0.6874	0.7307	69.6133	0.6731	0.7097	69.0815

Table 3. Comparison of diffusion-based blind super-resolution results using the original DiffBIR framework and our MILO-based latent masking extension. Results are reported on three datasets and evaluated by the common no-reference metrics TOPIQ, CLIP-IQA, and MUSIQ.

Method	**Real Dataset**			**RealSR Dataset**			**DIV2K-val**		
	TopIQ ↑	CLIP-IQA ↑	MUSIQ ↑	TopIQ ↑	CLIP-IQA ↑	MUSIQ ↑	TopIQ ↑	CLIP-IQA ↑	MUSIQ ↑
DiffBIR	0.7072	0.7706	72.9941	0.6211	0.6609	64.2512	0.6773	0.7398	69.4571
MILO	0.7062	0.7769	73.5812	0.6817	0.7217	65.9904	0.7112	0.7600	70.8609

Table 4. Comparison of diffusion-based blind face restoration results using the original DiffBIR framework and our MILO-based latent masking extension. Results are reported on two datasets and evaluated by the common no-reference metrics TOPIQ, CLIP-IQA, and MUSIQ.

Method	**LFW Dataset**			**Wider Dataset**		
	TopIQ ↑	CLIP-IQA ↑	MUSIQ ↑	TopIQ ↑	CLIP-IQA ↑	MUSIQ ↑
DiffBIR	0.7640	0.7948	76.4206	0.7528	0.8083	75.3224
MILO	0.6993	0.7420	76.0020	0.6873	0.7613	75.6032

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, 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

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

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.

Eric C. Larson and Damon M. Chandler. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *J. Electronic Imaging* 19 (2010), 011006.

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]

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=Bkg6RiCqY7

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.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. 2001. 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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library.

Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication* 30 (2015), 57–77.

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

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.

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.

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

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.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. 2011. 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.

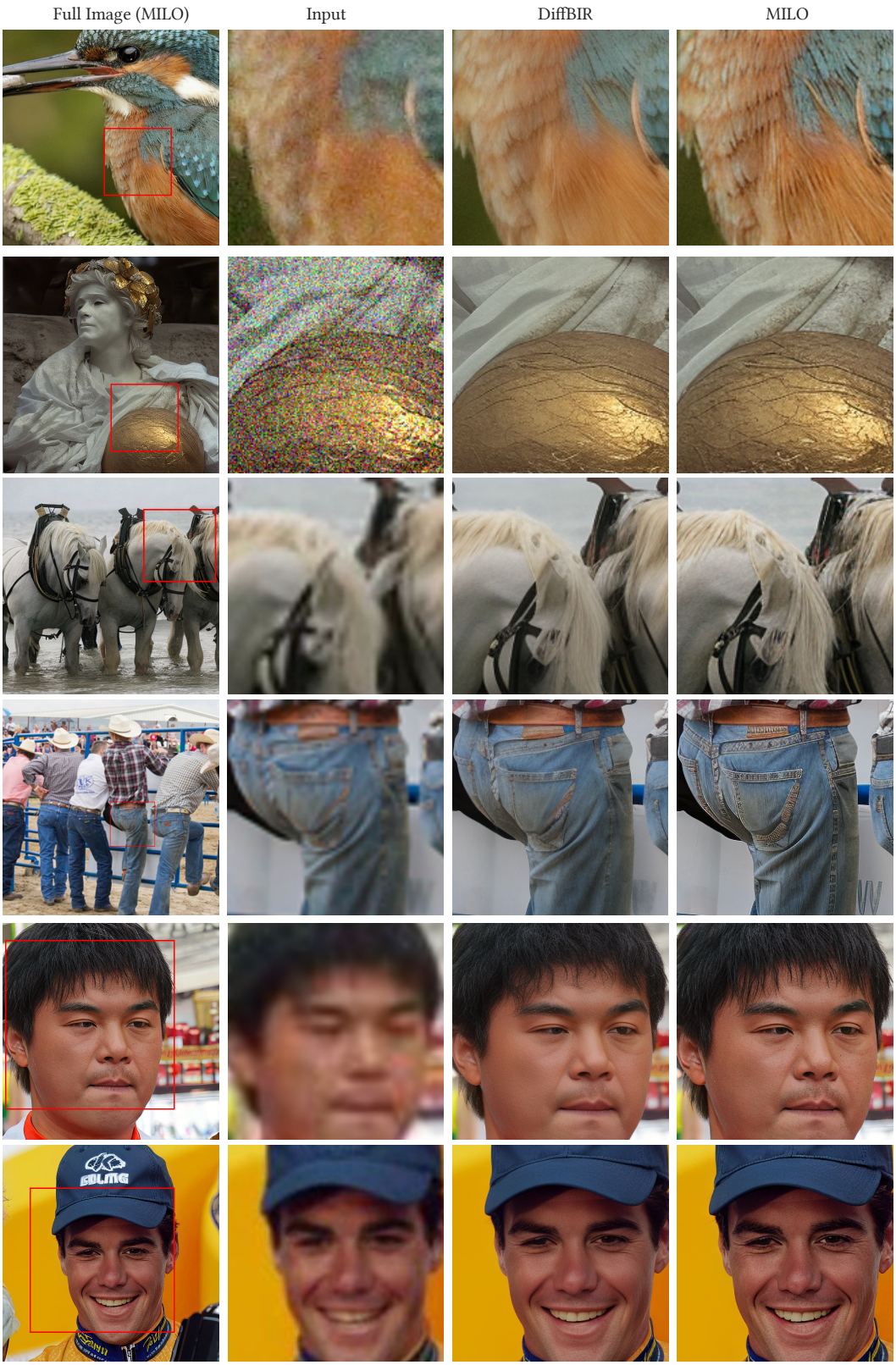

Fig. 2. Visual comparison of using MILO for perceptually guided optimization in comparison to the original loss of DiffBIR [Lin et al. 2024]. Comparisons are given for diffusion-based blind denoising (first two rows), blind super-resolution (two middle rows), and face restoration (last two rows).