Supplementary for: "Training Task-Specific Image Reconstruction Loss"

Anonymous WACV submission

Paper ID 784

This document includes additional details that could not be included in the main paper due to the lack of space. This comprises: *a)* manifold assumption validation and visual comparison with SR-GAN discriminator as compared to our multi-scale discriminators; *b)* quantitative results in terms of average PSNR and LPIPS for all application across 3 benchmark datasets *c)* qualitative results for the JPEG artefact removal application; *d)* ablation study on the number of seed images and the number of discriminators of the Multi-Scale Discriminative Feature (MDF) loss; *e)* hyper-parameter tuning for the VGG and LPIPS feature-wise loss functions; and *f)* performance of loss functions as quality predictors. Finally, we provide an HTML report which comprises all the results.

1. Image manifold assumption

The main objective of GANs [3] in image restoration is to learn a discriminator model that differentiates between image manifolds [6, 11, 9, 8, 2]. This is based on the hypothesis that input samples (e.g. noisy images) and their corresponding ground truth samples lie on two different manifolds. The generator model thereby learns a mapping function from one manifold to another, resulting in photo-realistic images closer to the natural image manifold [5, 2].

However, in this paper, we propose that learning the natural image manifold, which is often the task attributed to the discriminator, is less important than being able to detect errors introduced by the generator. Moreover, learning the natural image manifold requires the GAN to be trained with thousands of natural and fake images, making the training process computationally intensive. Here, we show that our task-specific discriminators, trained on a single image, can be used as feature extractors for the loss function because they learn the generator errors rather than the natural image manifold.

To validate this claim, a multi-scale discriminator trained on *a single image* for the task of JPEG artefact removal is employed as feature extractor. We randomly sample 100 natural images from the ILSVRC validation dataset [10]. From these images we generate *a)* JPEG compressed images using a compression quality between 7 and 10, *b)* blurry image samples by downsampling and upsampling the images by a factor of 4 using bi-linear filter and *c)* scrambled images by randomly permuting the pixels on each level of the Laplacian pyramid. Such permutations distort the second-order statistic, but preserve the composition of the spatial spectrum. The JPEG trained discriminator is used to extract the latent feature space of each set of images. The feature space for each image is the average across the channels and the resulting feature vector is reduced to a dimensionality of 3 using t-SNE for visualization. Fig. 1 shows the plot of the features from each set of images. The visualization shows that the discriminator does not learn the natural image manifold and cannot discriminate between natural and randomly permuted images. It also cannot discriminate between blurred and original images, but performs well in detecting JPEG artifacts regardless of image content.

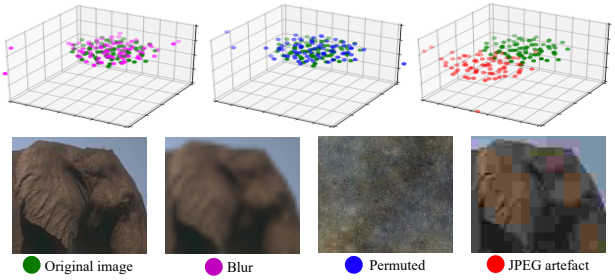

Figure 1: Manifold assumption validation: The figure shows the 3D t-SNE plots of the latent feature vectors extracted from diverse sets of images using multi-scale discriminators trained for the JPEG artefact removal task. Our JPEG-tuned discriminator cannot differentiate between the original and permuted images (middle plot), yet is a very effective feature-extractor for a loss function for JPEG task.

1.1. Image manifold comparison

In this section, we repeat the experiment conducted above, instead this time for a fully trained SR-GAN [5] discriminator. This further bolsters our claim that the task-specific discriminators of our MDF loss function learn to detect the generator distortions instead of the entire natural

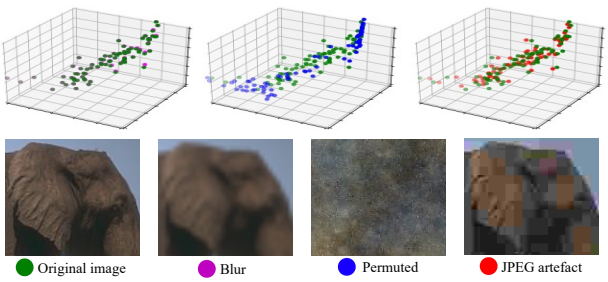

Figure 2: Manifold assumption validation: The figure shows the 3D t-SNE plots of the latent feature vectors extracted from diverse sets of images using an SR-GAN discriminator trained on DIV2K dataset [1]. The SR-GAN discriminator cannot differentiate between the original and jpeg images (right plot), thereby cannot be used as an effective feature extractor to detect and remove distortions.

image manifold. This thereby allows our MDF loss function, trained on a single image, to be used to effective feature extractors between the generated and the reference image.

We chose the same sample of 100 natural images from the ILSVRC validation dataset [10]. From these images we generated a) JPEG compressed images using a compression quality between 7 and 10, b) blurry image samples by downsampling and upsampling the images by a factor of 4 using bi-linear filter and c) scrambled images by randomly permuting the pixels on each level of the Laplacian pyramid. Such permutations distort the second-order statistic, but preserve the composition of the spatial spectrum. A trained SR-GAN discriminator is used to extract the latent feature space of each set of images. The feature space for each image is chosen after the Global Average Pooling (GAP) layer of the network. We used t-SNE to reduce the dimensionality of the feature vector to 3 for visualization. Fig. 2 shows the plot of the features from each set of images. The visualization shows that the discriminator of SR-GAN learns the natural image manifold (unlike our multi-scale discriminator) and can discriminate between natural and randomly permuted images. However, it cannot discriminate between the JPEG compressed and original images, making it an inferior feature extractor to detect and remove distortions.

2. Quantitative results

The quantitative results for all four applications are shown as distributions in Fig. 3 for real world mobile phone captured images from DPED dataset [4]. The differences in means (magenta dots in Fig. 3) are small but statistically significant for most comparisons (one-tailed t-test with H_1 show that the quality score is higher for our method, red $*$ symbols are shown if the difference is significant at $\alpha = 0.05$). The means, however, are not the best indicator of performance of different losses. This is because the differences in loss functions are mostly visible in smooth or flat parts of the images, which occupy only small percentage of all pixels but have a substantial impact on the perceived image quality (as demonstrated in Sec. 4.3 of the main paper). The advantage of our loss is better visible for the worst-case results, shown in Fig. 3 as the lower 5^{th} percentile of values (black asterisks). In majority of the comparison, MDF loss produces fewer images with low quality values, especially in terms of LPIPS. We also report the quantitative results in terms of average PSNR and LPIPS in Table. 2.

3. JPEG artefact removal results

In this section, we provide qualitative results showing comparison between three sample reconstructed images from the BSD Test Set using our (MDF) loss with various other loss functions for the task of JPEG artefact removal application. The test images are compressed with a quality factor of 10 and a more challenging factor of 7. Fig. 5 shows the results for the compression quality factor 7. The performance of the various loss functions seems to be comparable for the quality factor of 10, however, our model substantially provides artefact removal, especially in the uniform areas of the image for a much challenging codec quality of 7. The same was also observed in the subjective experiment conducted (see Sec. 4.3 of the main paper). Additional qualitative results can be seen in the HTML report attached to the Supplementary Material.

4. Ablation study

4.1. Scales of Discriminators

Since our MDF loss function comprises a series of discriminators trained on a single image at various scales, we need to select the optimal number of scales (the hyper-parameter K in Equation 2 of the main paper) to achieve the best performance. We perform an ablation study on training the EDSR model [7] using only the coarsest scale discriminator and subsequently adding finer scales. We observe a significant increase in quality of the images generated with the increase in the number of discriminators. As shown in Table 1, our loss performs the best when all 8 scales are employed.

Number of seed images Next we investigate the impact of increasing the number of seed images while training the MDF loss function. The plot in Fig. 4 shows that the performance of EDSR increases by only 0.03 dB when trained on 4 images and then it saturates. We did not observe any improvement in visual quality. Because the increase in performance in negligible when adding more seed images, we used a single image for training in our results.

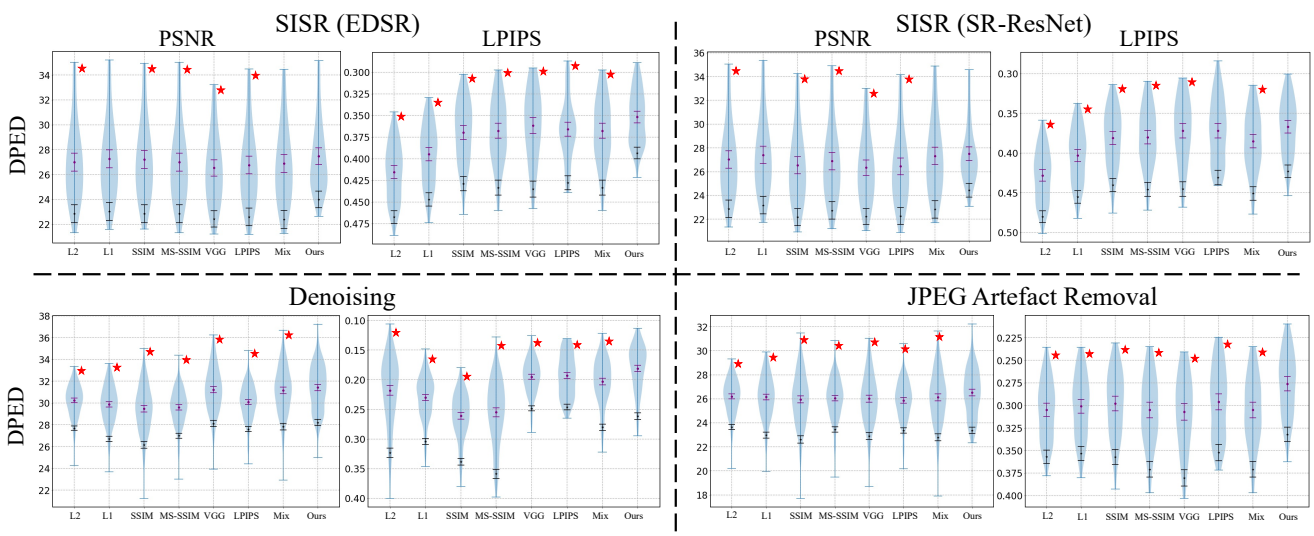

Figure 3: Additional violin plots illustrating the distribution of the PSNR [dB] ↑ and LPIPS ↓ values for DPED dataset [4] for all applications. Note that the y-axis is reversed for LPIPS so that the quality improves towards the top of each plot. The error bars show the 95% confidence intervals for the mean (magenta) and the 5^{th} percentile (black). The latter CIs were computed by bootstrapping. The red asterisks indicate that one-tailed t-test on the means gives statistically significant difference at $\alpha = 0.05$. It is worth noting that our loss produced fewer images with low quality values.

Table 1: Ablation study on training the SISR model (EDSR) using different scales of our loss. The scale number represents the number of scales included in the MDF loss. The inference results are reported for the BSD dataset.

Scales	1	2	3	5	7	8
PSNR ↑	22.55	23.89	24.43	24.89	25.27	**25.70**
LPIPS ↓	0.392	0.357	0.354	0.311	0.305	**0.286**

5. Hyper-parameter tuning for VGG and LPIPS

In Fig. 6 we show the qualitative results for the trade-off between the MSE and LPIPS/VGG network components in the joint loss function. For fair comparison, we conducted a hyper-parameter search over the scalar λ controlling the weight of the feature-wise loss function. We searched over the values in $\{\lambda : \lambda = 10^k, k = -3, .., 3\}$. The greater λ parameter is, the more LPIPS/VGG components contribution is. In our experiments across all image restoration applications, we found the best results are produced when $\lambda = 1$ for VGG and $\lambda = 0.1$ for LPIPS loss. Additional qualitative results are provided in the HTML report.

6. Image quality metrics and loss functions

To further investigate the performance of loss functions as quality predictors, we generated a set of images that were distorted by blur, noise, added sinusoidal grating, contrast and brightness changes. The distortions were generated so

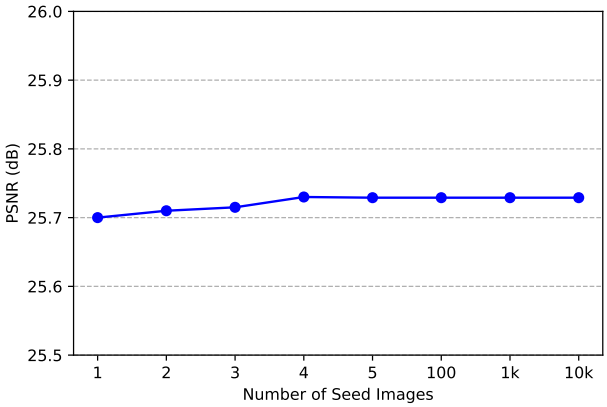

Figure 4: Performance of EDSR model with the increasing the number of seed images used for training the MDF loss function. Note that PSNR increases only by 0.03 dB and saturates for larger number of images. The inference results are reported for the BSD dataset.

that they degraded the image in equal steps of PSNR. Fig. 7 presents an example of images with introduced distortions at three PSNR levels. The experiment shows a failure case of PSNR, predicting the same quality even though the distortions due to contrast and brightness are much less objectionable than the others to a human observer.

In Fig. 8, we show the loss values computed for the increasing amount of distortions of different types for differ-

Table 2: Comparison of our proposed Multi-Scale Discriminative Feature (MDF) loss function with other losses on 3 public benchmark datasets for four tested applications. Results show PSNR [dB] ↑ / LPIPS ↓. The numbers in red indicate the best performance and the ones in blue the second best.

Dataset	L_2	L_1	SSIM	MS-SSIM	VGG	LPIPS	MS-SSIM + L_1	Ours
Single Image Super-Resolution (EDSR [7])								
DIV2K	28.70 / 0.342	29.22 / 0.315	29.21 / 0.293	28.70 / 0.342	28.10 / 0.278	28.34 / 0.283	28.87 / 0.283	**29.51 / 0.276**
DPED	26.99 / 0.415	27.26 / 0.394	27.22 / 0.369	27.00 / 0.367	26.54 / 0.361	26.76 / 0.366	26.88 / 0.368	**27.48 / 0.351**
BSD	25.28 / 0.320	25.66 / 0.304	25.52 / 0.309	24.70 / 0.301	24.44 / 0.298	24.49 / 0.296	25.08 / 0.306	**25.70 / 0.286**
Single Image Super-Resolution (SR-ResNet [5])								
DIV2K	27.57 / 0.343	27.76 / 0.321	27.05 / 0.325	27.20 / 0.320	26.83 / 0.301	27.00 / 0.307	27.49 / 0.313	**27.95 / 0.295**
DPED	27.03 / 0.428	27.41 / 0.403	26.54 / 0.381	26.89 / 0.380	26.34 / 0.372	26.45 / 0.372	27.32 / 0.385	**27.50 / 0.367**
BSD	24.56 / 0.337	24.68 / 0.328	24.07 / 0.370	24.18 / 0.364	23.19 / 0.315	23.42 / 0.310	24.48 / 0.336	**25.07 / 0.293**
Image Denoising [12]								
DIV2K	29.75 / 0.233	29.55 / 0.236	29.47 / 0.275	29.62 / 0.263	30.80 / 0.215	29.61 / 0.215	30.05 / 0.225	**31.25 / 0.192**
DPED	30.24 / 0.218	29.87 / 0.230	29.48 / 0.261	29.60 / 0.255	31.23 / 0.195	30.09 / 0.191	31.15 / 0.203	**31.36 / 0.181**
BSD	29.92 / 0.240	29.71 / 0.248	29.39 / 0.285	29.55 / 0.262	30.40 / 0.203	29.81 / 0.203	30.39 / 0.214	**30.42 / 0.192**
JPEG Artefact Removal [12]								
DIV2K	26.50 / 0.303	26.71 / 0.295	26.32 / 0.295	26.37 / 0.301	26.48 / 0.315	26.27 / 0.281	26.50 / 0.299	**26.77 / 0.261**
DPED	26.20 / 0.305	26.15 / 0.301	25.95 / 0.298	26.05 / 0.305	26.01 / 0.307	25.87 / 0.296	26.12 / 0.305	**26.53 / 0.276**
BSD	25.64 / 0.316	25.71 / 0.310	25.43 / 0.309	25.49 / 0.313	25.54 / 0.308	25.39 / 0.308	25.52 / 0.312	**25.75 / 0.293**

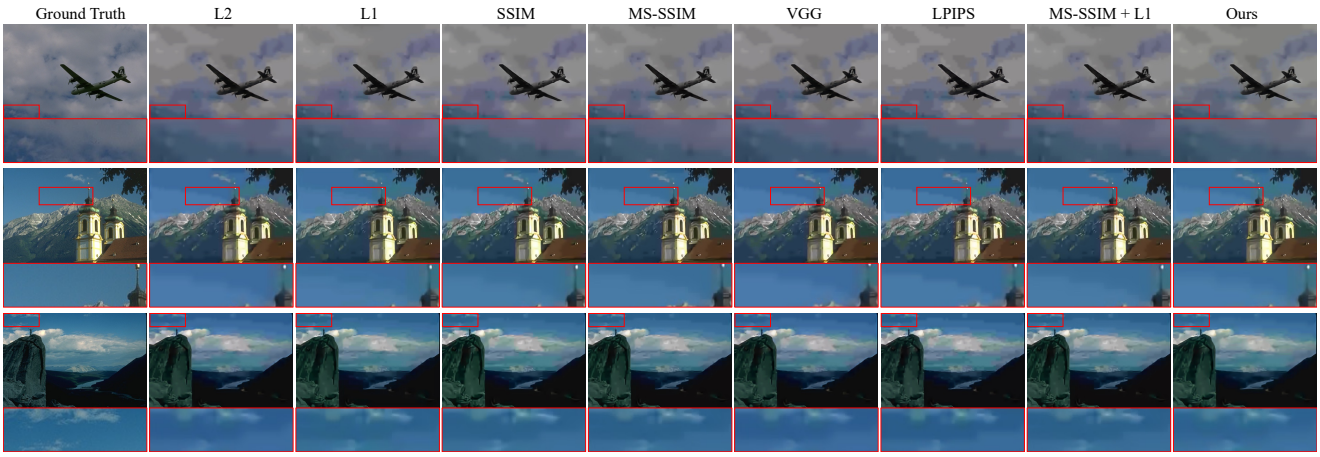

Figure 5: Results for JPEG artefact removal (compression quality = 7) using DnCNN model [12] trained using different losses. Our loss improves artefact reduction, especially in the uniform areas of an image. Qualitative results in terms of PSNR and LPIPS are reported in Table 2. Best viewed when zoomed.

ent loss functions. Despite the same PSNR value, the distortions due to noise, blur and added sinusoidal wave are much more noticeable than those due to contrast and brightness change (refer to Fig. 7). The loss functions derived from quality metrics (SSIM, MS-SSIM) and also feature-wise losses (VGG, LPIPS) penalize more the distortions that result in higher degradation of quality. In contrast, MDF losses penalize the most the distortions that are relevant for a given task: blur in case of SISR (MDF SR), blur and noise in case of denoising, and contrast followed by the mixture of all distortions in case of JPEG artifact removal. This is another example demonstrating that an effective loss (MDF) function does not need to predict image quality.

7. HTML report

In the Supplementary Material, we provide a comprehensive HTML report, showing the results for each loss function across different image reconstruction applications for various datasets. We further provide results for the ablation study and the hyper-parameter selection. The HTML report, including all the inference images, are attached with the supplementary material. Please visit the URL `HTML_Report_Paper_ID_784.html` inside the folder named "Report_784".

Due to size limitations for the supplementary material, we include the first 30 images from each test set. Images are stored as JPEGs with a quality of 90 to ensure that coding

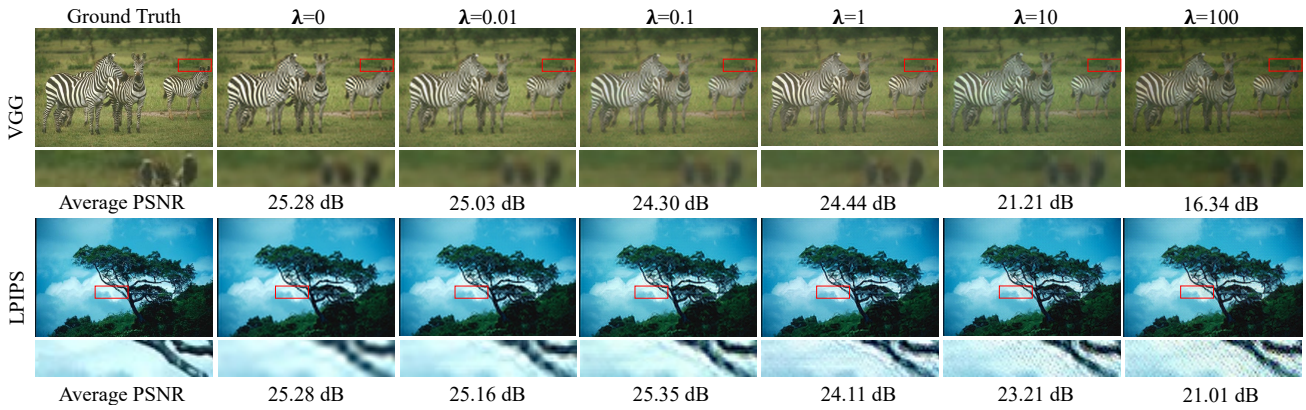

Figure 6: Comparison of the single-image super resolution (SISR) results (EDSR) when trained using a weighted sum of VGG/LPIPS and MSE feature-wise losses: MSE + λ VGG/LPIPS. The average PSNR is reported for the entire test set.

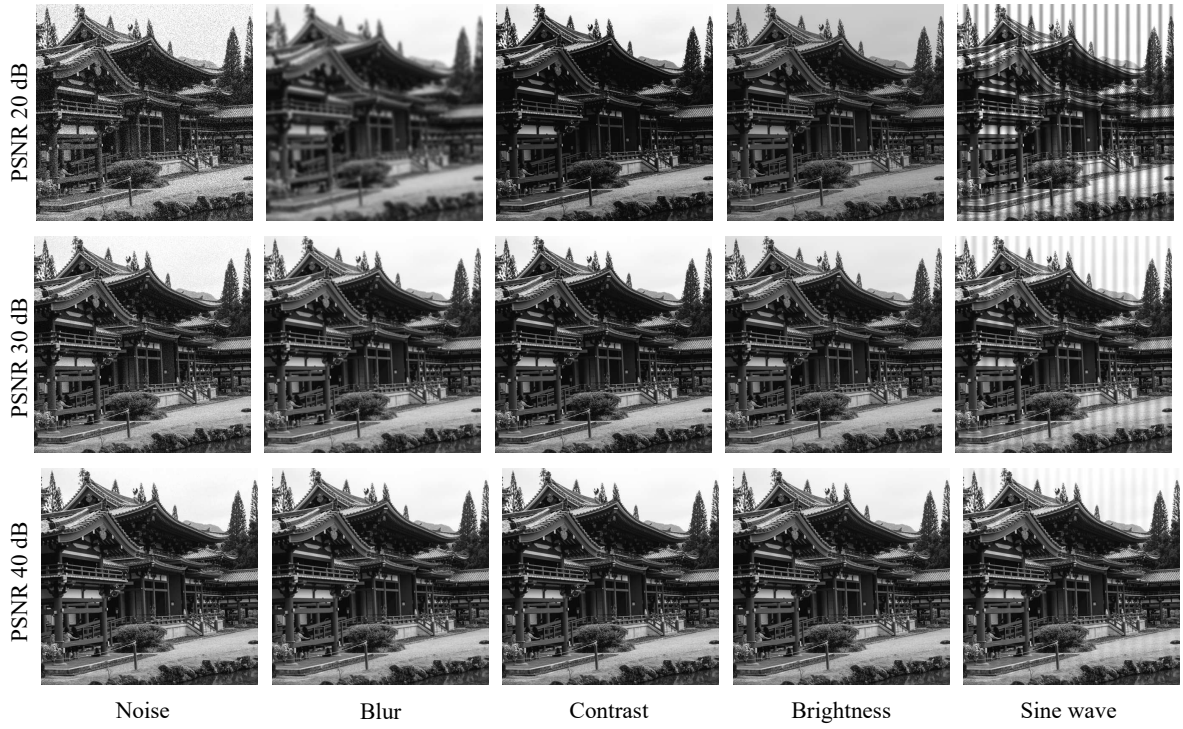

Figure 7: Examples of images used to test the sensitivity of loss functions to different types of distortions. We introduced artifacts so that the each distortion results in the same PSNR level (across each row). Here we provide examples of images at 20 dB, 30 dB and 40 dB. Note that the perceived quality differs between the columns despite the same PSNR level.

distortions do not distort the results. Upon acceptance, the code and the complete set of inference outputs will be made public for the research community.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 2

[2] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. *arXiv preprint arXiv:1506.05751*, 2015. 1

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and

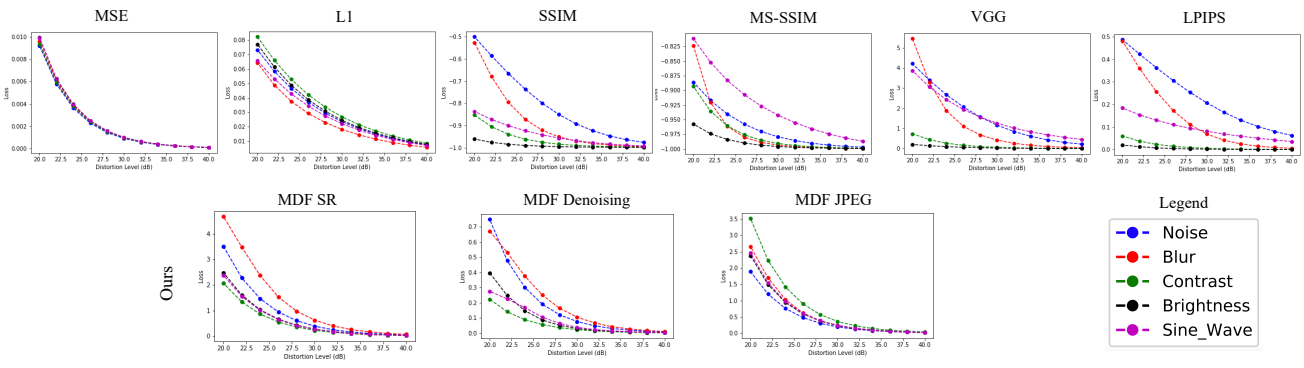

Figure 8: Loss values for the increasing amount of distortions of different types. The distortion levels have been generated to result in equal PSNR values, shown on the x-axis. Despite the same PSNR value, the distortions due to noise, blur and added sinusoidal wave are much more noticeable than those due to contrast and brightness change (refer to Fig. 7). The MDF loss accurately predicts the perceived magnitude of task specific distortions for which it is trained.

Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 1

[4] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3277–3285, 2017. 2, 3

[5] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1, 4

[6] Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2479–2486, 2016. 1

[7] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 2, 4

[8] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 1

[9] Aamir Mustafa and Rafal K Mantiuk. Transformation consistency regularization-a semi-supervised paradigm for image-to-image translation. *arXiv preprint arXiv:2007.07867*, 2020. 1

[10] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1, 2

[11] Raymond Yeh, Chen Chen, Teck Yian Lim, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with perceptual and contextual losses. *arXiv preprint arXiv:1607.07539*, 2(3), 2016. 1

[12] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 4