A HVS-inspired Attention Map to Improve CNN-based Perceptual Losses for Image Restoration

Taimoor Tariq, Juan Luis Gonzalez and Munchurl Kim
Korea Advanced Institute of Science and Technology (KAIST)
{taimoor.tariq, juanluisgb, mkimee}@kaist.ac.kr

Abstract

Deep Convolutional Neural Network (CNN) features have been demonstrated to be effective perceptual quality features. The perceptual loss, based on feature maps of pre-trained CNN’s has proven to be remarkably effective for CNN based perceptual image restoration problems. In this work, taking inspiration from the the Human Visual System (HVS) and our visual perception, we propose a spatial attention mechanism based on the dependency human contrast sensitivity on spatial frequency. We identify regions in input images, based on underlying spatial frequency where the visual system might be most sensitive to distortions. Based on this prior, we design an attention map that is applied to feature maps in the perceptual loss, helping it to identify regions that are of more perceptual importance. The results will demonstrate that the proposed technique improves the correlation of the perceptual loss with human subjective assessment of perceptual quality and also results in a loss which delivers a better perception-distortion trade-off compared to the widely used perceptual loss in CNN based image restoration problems.

1. Introduction

In this paper, we focus on Full-Reference Image quality metrics [21] for image restoration. The PSNR and SSIM [25], have been traditionally used as FR quality metrics between two images, but they have been shown to quantify distortion, not perceptual quality. The difference and trade-off between distortion and quality is well explained in [3]. The human visual perception process is fairly complex and still not understood completely. Recently, we have observed an increased use of deep CNN features as perceptual quality features for Image restoration and quality assessment.

The perceptual loss proposed by Johnson et al. [11] was one of the first to demonstrate how effective the deep CNN features can be as perceptual quality features, especially when used in loss functions for image restoration. The perceptual loss is now a benchmark and standard loss function popularly adopted in many image to image translation problems such as super-resolution, style transfer, denoising etc. [15],[24],[5]. The recent PIRM Super-Resolution Challenge Report [2] clearly iterates that the perceptual loss is the standard benchmark loss function for perceptual image restoration and is also used in combination with other losses such as the GAN loss. Zhang et al. [27] and Blau et al. [3] further demonstrate how effective the deep features really are as perceptual quality features. Considering the difficulty in quantifying perceptual quality and the importance of effective differentiable metrics, it is surprising that a significant amount of research as been subjected to novel CNN architectures for image restoration but the loss functions, which are of as much if not more importance, remain relatively un-addressed.

In this work, we exploit a known and well established property of the HVS called contrast sensitivity. Psychovisual experiments have revealed that the visual system’s ability to respond to contrast, varies significantly with spatial frequency of the stimulus [17]. Neuro-scientists have been able to quantify this variation using a function called the Contrast Sensitivity Function (CSF) [20]. Considering that it is an important representative factor of our visual perception, the CSF has been used extensively by the community to improve image processing algorithms and techniques [20].

We use a variation of the CSF filtering technique of images to extract maps which represent regions which the human visual system is most sensitive to in terms of contrast. Theoretically, due to our enhanced contrast sensitivity while perceiving these regions, distortions in these regions should be more perceptible and prominent to the human observer. The map is then applied to feature maps of pre-trained networks used in the perceptual loss to give more emphasis to important areas during restoration. The results will demonstrate that the technique improves the perceptual quality of restored images and delivers a better perception-distortion trade-off compared to the standard and popularly used perceptual loss [11].
2. A HVS inspired Attention

2.1. Contrast Sensitivity Function (CSF)

The spatial theory frequency aims to correlate the response of the visual cortex with spatial frequency of input stimulus [17]. The theory is contrary to earlier theories of Hubel and Wisel which used straight edges and lines to characterize perception in the visual cortex [7]. The theory is supported by psycho-visual experiments which employ sinusoidal gratings. According to the Fourier theory, a light distribution of an image on the retina can be expressed as a linear combination of basis harmonic components. Grating stimuli have therefore been used in psycho-visual experiments to represent basis harmonic components.

Experiments such as the ones described in [17], [9] and [14] show that the neurons in various regions of the visual cortex are tuned to specific spatial frequencies. This behavior gives an interpretation of the function of different regions of the cortex, supporting the claim that a major function of the primary visual cortex is to split images into harmonic components for further processing.

The net response/sensitivity of the HVS to different spatial frequencies is summarized in the CSF. Fig. 2(a) shows the human CSF derived through psychovisual experiments. The units of the CSF are in cycles/degree. This unit translates to the number of cycles per visual degree of the observer.

2.2. CSF based Map Generation

Let \( I_{GT} \) be the Y channel of the ground truth image with size \( M \times N \). The 2D DFT of the image will be

\[
F[u, v] = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I_{GT}[m, n]e^{-j2\pi \left( \frac{m}{M} u + \frac{n}{N} v \right)}
\]

(1)

Where \( u \) and \( v \) are the horizontal and vertical frequency indicators respectively. As iterated earlier, the CSF is defined on the spatial frequency scale of cycles per visual degree. The current frequency units of the DFT need to be appropriately converted into the required scale for effective application of the CSF. Let,

\[
f(u) = \frac{u - 1}{\Delta M}, f(v) = \frac{v - 1}{\Delta N}
\]

(2)

where \( \Delta = 0.25 \text{mm} \) is the dot pitch. The cycles/degree equivalent of the DFT frequency units with a viewing distance of ‘dis’ millimeters will be \( s(u, v) \) [1] where,

\[
s(u, v) = \frac{\pi}{180 \times \arcsin \left( \frac{1}{\sqrt{1 + dis^2}} \right)} \times \sqrt{f(u)^2 + f(v)^2}
\]

(3)

Now we have effectively expressed the DFT on the spatial frequency units of cycles/degree. To generate a map that represents the most important spatial information, we will approximate the input image using a narrow band of spatial frequencies around the peak of the CSF. Let,

\[
\hat{F}(u, v) = \begin{cases} 
F(u, v), & s_l \leq s(u,v) \leq s_h, \\
0, & \text{otherwise}
\end{cases}
\]

(4)

Now we take the inverse 2D DFT of the approximated representation to generate a map as shown in Eq. (5).

\[
T[m, n] = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \hat{F}[u, v]e^{j2\pi \left( \frac{m}{M} u + \frac{n}{N} v \right)}
\]

(5)

As a final step, we normalize the matrix ‘T’ so that its largest element is unity. \( \|T\|_M \) is the max norm of T in Eq.(6) (largest element of the matrix).

\[
\mu = \frac{1}{\|T\|_M} \times T
\]

(6)

The viewing distance for HDTV displays should be selected for a 30 degree viewing angle. Considering parameters for viewing distance and for effective perception of common distortions (checkerboard artifacts etc) introduced by the use of the perceptual loss, we selected the viewing distance to be 55cm. We empirically selected the range of the filter to be the frequencies corresponding to 45% of the peak of the CSF i.e. \( s_l = 2 \) cycles per degree and \( s_h = 23 \) cycles per degree.
3. Deep Features for Perceptual Metrics/Losses

3.1. Perceptual Loss

Deep CNN features are the result of non-linear transformations on images into a high dimensional manifold. The perceptual loss is based on the fact that distance between features on the transformed manifold might result in a better perceptual quality measure compared to distance between image pixels themselves. The high dimensional manifold in this case is the manifold of CNN features. The standard form of the perceptual loss, as per [11] is given in Eq.(7)

\[ l_p = \frac{1}{M \times W \times H} \sum_{m=1}^{M} \| \Phi^k_m(I_{out}) - \Phi^k_m(I_{GT}) \|^2 \]  

(7)

where '\( \Phi^k_m \)' is the 'm'\textsuperscript{th} feature map in the 'k'\textsuperscript{th} layer with 'M' number of feature maps with dimensions '\( H \times W \)'. This approach and its variants have proven to be remarkably effective as perceptual features in FR-IQA methods [4], image restoration [24] and style transfer [5] problems. The perceptual loss is a benchmark loss function in perceptual image restoration as demonstrated in [2].

3.2. Proposed Spatially Attentive Perceptual Loss

The CSF based attention map is a representation of regions which produce higher contrast sensitivity in perception. These regions will be the regions on which distortions will be most perceptible to human observers. The perceptual loss in the form of Eq. (7) does not spatially discriminate between information in feature maps on the basis of their perceptual importance. We propose extending the perceptual loss to a form in Eq. (8).

\[ l_{att} = \frac{1}{M \times W \times H} \sum_{m=1}^{M} \| \mu^k(I_{GT}) \odot (\Phi^k_m(I_{out}) - \Phi^k_m(I_{GT})) \|^2 \]  

(8)

where \( \mu^k(I_{GT}) \) is the CSF based attention map. The superscript 'k' represents that is has been appropriately re-sized to match the feature map dimensions in the k-th layer. The formulation in Eq. (8) will help the perceptual loss to give more importance in restoration of regions which are perceptually more important and have a visual susceptibility to perception of distortions which result in a loss of contrast, commonly associated with Super-Resolution and Blurring.

The results will demonstrate that our proposed extension provides a superior metric that is more correlated with human subjective assessment of perceptual quality and delivers a better perception-distortion trade-off when used for CNN based image restoration.

4. Experimental Setup

4.1. Overview

To demonstrate the superiority of our proposed extension over the classical and widely used perceptual loss, we will make use of two distinct experimental techniques. The first being objective quality assessment (OQA) tests and the second being an image restoration experiment.

OQA tests are a standard method to evaluate the effectiveness of perceptual quality metrics. The best judges of perceptual quality are human observers. OQA tests try to correlate the performance of objective metrics with human judgment of perceptual quality. The more correlated a metric is with human subjective assessment of quality, the better it can be inferred as a perceptual quality metric. For human subjective assessment of perceptual quality, we will use the LIVE subjective image quality database [22]. The data-set consists of human subjective scores in the form of Differential Mean Opinion Scores (DMOS) for a large number of images with a wide variety of distortions. The correlation of objective metric scores with human subjective scores is quantified using three metrics namely the Root Mean Square Error (RMSE), Linear Correlation Coefficient
(LCC) and the Spearman Rank Order Correlation Coefficient (SROCC) after curve fitting of the objective and subjective scores using a non-linear polynomial.

For comparing the objective metrics in Eq. -(7) and -(8), we will use feature maps from different layers of a wide variety of pre-trained networks such as the VGG-16 [23], ResNet-18 [6], SqueezeNet [8] and AlexNet [13] to reinforce the validity of our approach. We will also repeat our OQA test for a wide variety of distortions such as Gaussian Blur, JPEG and JPEG2000 Compression Noise and Multiple Distortions [10] (camera image acquisition process where images are rst blurred due to narrow depth of field or other de-focus and then corrupted by white Gaussian noise to simulate sensor noise). A diverse experimental setup with multiple networks, their layers and different types of distortions will reinforce the validity and effectiveness of our approach.

Our second experiment will be an x4 Image Super-Resolution experiment using the well known VDSR network trained on the standard DIV2K data-set using the loss functions in Eq. -(7) and -(8) used in combination with the $l_2$ loss. The two loss function that will be used to train the network will hence be:

$$L_p = \alpha.l_2 + (1-\alpha).l_p$$ (9)

and

$$L_{p\text{att}} = \alpha.l_2 + (1-\alpha).l_{p\text{att}}$$ (10)

To evaluate and compare the performance of both loss functions, we will compare the perception-distortion trade-off they both deliver. The Structural Similarity Index (SSIM) and the Peak Signal to Noise Ratio (PSNR) have traditionally, some what incorrectly been used as metrics to evaluate image restoration techniques. These metrics measure distortion between two images and are not sufficient to quantify the perceptual quality. Recent works such as [3] and [27] have pointed out that to quantify the efficacy of image restoration techniques, we need to analyze something known as the perception-distortion trade-off and perceptual quality. In recent works [2], The perceptual quality has been quantified using a metric known as perceptual index (PI) which is a combination of two No-Reference quality metrics called the Naturalness Image Quality Evaluator (NIQE) [19] and the NRQM [16]. The PI of an image $I_{in}$ is thus defined in Eq.(11)

$$PI(I_{in}) = \frac{1}{2}((10 - NRQM(I_{in})) + NIQE(I_{in}))$$ (11)

The lesser the PI, the better the perceptual quality of the image. The distortion (measured by the SSIM and PSNR) and the perceptual quality (PI) are in a trade-off relation, as explained in [3]. The effectiveness of one technique over another can be quantified by its ability to deliver a better trade-off i.e lesser PI at the same SSIM. We will sufficiently demonstrate that our proposed loss does exactly that in comparison to the perceptual loss.

4.2. Results and Discussions

4.2.1 Objective Quality Assessment Experiments

Table. 1 demonstrates the correlation of the perceptual loss (Eq. 7) and our proposed extension (Eq. 8) with human subjective assessment of perceptual quality. The experiment is repeated over multiple pre-trained networks, their layers and for various types of distortions. It can be observed that our proposed extension is much more correlated with human subjective assessment of perceptual quality and thus a superior metric. The superiority is quantified in higher LCC and SROCC scores. A graphical representation is shown in Fig. 3 where it can be seen that the objective metric in Fig. 3-(b) is more correlated with human subjective DMOS compared to the perceptual loss ($l_p$). This result is for loss with feature maps ‘Fire3_ReLU, expand1x1’ layer of the SqueezeNet on Gaussian Blur distorted images.

Figure 3: It can be seen that our proposed extension ($l_{p\text{att}}$) is more correlated with human subjective DMOS compared to the perceptual loss ($l_p$). This result is for loss with feature maps ’Fire3_ReLU, expand1x1’ layer of the SqueezeNet on Gaussian Blur distorted images.

4.2.2 x4 Super Resolution Experiments

For our x4 SR experiment, we trained a VDSR on the DIV2K data-set on both the loss functions in Eq.-(9) and -(10) under the exact same training conditions, using the ‘ReLU3.2’ layer of the VGG-16 for the loss functions. We performed three training and testing trials for each loss function by varying the control parameter $\alpha$. The higher the $\alpha$ in Eq.-(9) and -(10), the more weight is given to the pixel-wise $l_2$ loss compared to the $l_p$ or $l_{p\text{att}}$ loss causing an increase in the SSIM and a decrease in the perceptual quality (higher PI). This behavior is in accordance with the perception-distortion trade-off [3]. Fig. 4 shows that even...
Table 1: The attention module improves the correlation with human subjective scores of the modified perceptual loss ($l_{att}^{p}$) in Eq.(8) compared to the widely used standard perceptual loss ($l_p$) in Eq.(7) which leads to the conclusion that $l_{att}^{p}$ is a better objective perceptual quality metric. The results are repeated across multiple pre-trained networks, their layers and different types of common distortions as well as their combinations (Multiple Distortions).

| Layer               | Metric | RMSE  | LCC   | SROCC  |
|---------------------|--------|-------|-------|--------|
| Fire2-ReLU_expand3x3| $l_p$  | 15.2720 | 0.6132 | 0.5589 |
|                     | $l_{att}^{p}$ | 13.9830 | 0.6906 | 0.6084 |
| Fire4_ReLU_expand1x1| $l_p$  | 14.5749 | 0.6570 | 0.6051 |
|                     | $l_{att}^{p}$ | 12.9300 | 0.7434 | 0.6688 |
| Fire6_ReLU_expand3x3| $l_p$  | 15.5951 | 0.5910 | 0.5482 |
|                     | $l_{att}^{p}$ | 14.3458 | 0.6704 | 0.6135 |

Table 1.(b)

| Layer               | Metric | RMSE  | LCC   | SROCC  |
|---------------------|--------|-------|-------|--------|
| Fire2-ReLU_expand3x3| $l_p$  | 11.0894 | 0.7185 | 0.7375 |
|                     | $l_{att}^{p}$ | 10.1010 | 0.7737 | 0.8010 |
| Fire3_ReLU_expand1x1| $l_p$  | 10.5569 | 0.7494 | 0.7641 |
|                     | $l_{att}^{p}$ | 8.2405  | 0.8561 | 0.8712 |
| Fire4_ReLU_expand1x1| $l_p$  | 10.2650 | 0.7652 | 0.7867 |
|                     | $l_{att}^{p}$ | 8.5886  | 0.8425 | 0.8591 |

Table 1.(c)

| Layer               | Metric | RMSE  | LCC   | SROCC  |
|---------------------|--------|-------|-------|--------|
| ReLU2.2             | $l_p$  | 10.3392 | 0.7612 | 0.7766 |
|                     | $l_{att}^{p}$ | 8.5227  | 0.8451 | 0.8694 |
| ReLU3.2             | $l_p$  | 9.4553  | 0.8052 | 0.8288 |
|                     | $l_{att}^{p}$ | 8.7198  | 0.8372 | 0.8572 |
| ReLU4.2             | $l_p$  | 9.4181  | 0.8069 | 0.8301 |
|                     | $l_{att}^{p}$ | 8.7319  | 0.8367 | 0.8453 |

Table 1.(d)

| Layer               | Metric | RMSE  | LCC   | SROCC  |
|---------------------|--------|-------|-------|--------|
| ReLU1.1             | $l_p$  | 7.4194  | 0.8917 | 0.8836 |
|                     | $l_{att}^{p}$ | 5.4533  | 0.9431 | 0.9342 |
| ReLU2.2             | $l_p$  | 7.0756  | 0.8992 | 0.8669 |
|                     | $l_{att}^{p}$ | 5.7946  | 0.9355 | 0.9285 |
| ReLU3.2             | $l_p$  | 6.7050  | 0.9126 | 0.9024 |
|                     | $l_{att}^{p}$ | 6.1821  | 0.9262 | 0.9167 |

Table 1.(e)

| Layer               | Metric | RMSE  | LCC   | SROCC  |
|---------------------|--------|-------|-------|--------|
| ReLU                | $l_p$  | 8.8626  | 0.8313 | 0.8450 |
|                     | $l_{att}^{p}$ | 7.5581  | 0.8805 | 0.8914 |
| Conv                | $l_p$  | 9.3736  | 0.8089 | 0.8281 |
|                     | $l_{att}^{p}$ | 6.4939  | 0.9133 | 0.9216 |
| ReLU_2              | $l_p$  | 9.5549  | 0.8006 | 0.8229 |
|                     | $l_{att}^{p}$ | 0.8294  | 0.8327 | 0.8532 |

Table 1.(f)

if the $\alpha$ parameter is varied, our proposed attentive perceptual loss always delivers a significantly better perception-distortion trade-off compared to the widely used perceptual loss [11]. At roughly the same average SSIM on the test set, training with $L_{att}^{p}$ gives delivers images with much better perceptual quality. Furthermore, even at much lower distortions, the perceptual loss fails to achieve as good perceptual quality as our proposed extension delivers at much higher
Table 1.(g)

| Layer           | Metric | RMSE  | LCC     | SROCC  |
|-----------------|--------|-------|---------|--------|
| Res2a_ReLU (JPEG) |        |       |         |        |
|                 | \(l_p\) | 7.1066 | 0.8833  | 0.8651 |
|                 | \(l_{att}\) | 6.5831 | 0.9134  | 0.8820 |
| Res2a_ReLU (JPEG2000) | |       |         |        |
|                 | \(l_p\) | 6.4062 | 0.9205  | 0.9144 |
|                 | \(l_{att}\) | 5.6115 | 0.9396  | 0.9316 |
| Res2b_Branch2a (JPEG) | |       |         |        |
|                 | \(l_p\) | 6.8992 | 0.8947  | 0.8704 |
|                 | \(l_{att}\) | 6.5296 | 0.9149  | 0.8926 |

Table 1.(h)

| Layer | Metric | RMSE  | LCC     | SROCC  |
|-------|--------|-------|---------|--------|
| Conv_1 |        |       |         |        |
|       | \(l_p\) | 6.6372 | 0.9144  | 0.9076 |
|       | \(l_{att}\) | 5.6783 | 0.9381  | 0.9310 |
| ReLU_1 |        |       |         |        |
|       | \(l_p\) | 6.7111 | 0.9124  | 0.9056 |
|       | \(l_{att}\) | 5.7262 | 0.9370  | 0.9298 |
| Conv_2 |        |       |         |        |
|       | \(l_p\) | 5.0185 | 0.9520  | 0.9459 |
|       | \(l_{att}\) | 4.8431 | 0.9564  | 0.9573 |

Figure 4: The proposed attentive perceptual loss \(L_{att}\) significantly improves the perception-distortion trade-off compared to the widely used perceptual loss \(L_p\). Close examination of the crops reveals that the \(L_{att}\) significantly helps in suppressing checkerboard artifacts, which are a common nuisance while using the perceptual loss. For example, while comparing Fig.5-(c) and -(d), the artifacts are clearly visible in Fig.5-(d) whereas Fig.5-(c) has a much smoother all round texture. Similar observations can be made in Fig.5-(h) and -(i) if the blue ocean background is carefully observed.

In summary, taking inspiration from the HVS, we have proposed an extension to the popular and widely used perceptual loss for image restoration. The results demonstrate that our proposed extension is inherently a better quality metric and it significantly improves the performance compared to the perceptual loss [11]. A significant amount of research has been invested into novel CNN architectures for image restoration but the issue of loss functions is still relatively un-addressed. A few variants of the perceptual loss such as the contextual loss [18] have also been proposed which instead of an \(l_2\) distance between feature maps, calculates an approximated version of the KL-Divergence. Our proposed attention map is by definition applicable to any extension of the perceptual loss that uses deep CNN features as perceptual quality features.

4.2.3 Gaussian De-Blurring Experiment

We perform another image restoration experiment to demonstrate the effectiveness of our proposed spatially attentive extension to the perceptual loss over the classical and widely used perceptual loss [11]. The experiment targets restoration of heavily Gaussian blurred images using the well known DnCNN [26] architecture. The DIV2K training images were corrupted by large standard deviation gaussian blur (\(\sigma = 7\)) and the DnCNN architecture was trained for restoration. The network was separately trained using the perceptual loss \(L_p\) and the proposed spatially attentive perceptual loss \(L_{att}\).

Fig.6 shows that \(L_{att}\) significantly improves the perception distortion trade-off in comparison to the perceptual loss \(L_p\). At the same level of distortion, training with the \(L_{att}\) results in much better perceptual quality. Fig. 7 shows images to demonstrate the improvement in perceptual quality and the perception-distortion trade-off. Considering that the perceptual loss is a standard benchmark and widely used loss function for CNN based perceptual image restoration [2], the improvement is significant and has a wide variety of practical applications.
Figure 5: Single Image x4 SR on the VDSR [12] architecture. Our proposed attentive perceptual loss ($L_{p}^{att}$) in Eq.(10) delivers better perceptual quality (less PI) at same levels of distortion (SSIM) compared to the perceptual loss ($L_{p}$) in Eq.(9).

5. Conclusions

Taking inspiration from the Human Visual System (HVS), more specifically the spatial frequency dependence of contrast sensitivity in our visual perception, we propose an attention map to improve the widely used perceptual loss for image restoration. Through objective quality assessment experiments, we verify that our proposed attentive perceptual loss is a better perceptual quality metric compared to the perceptual loss, owing to a better correlation with human subjective assessment of quality. Through an x4 Super-Resolution and Gaussian De-blurring experiment, we also
The proposed attentive perceptual loss ($L^{att}$) significantly improves the perception-distortion trade-off compared to the widely used perceptual loss ($L_p$) in a ($\sigma = 7$) gaussian deblurring experiment with the DnCNN on the DIV2K dataset. The two different points in each curve are obtained by varying the values of $\alpha$ in Eq.-(9) and -(10) in the main paper and averaging the results over the DIV2K test set.

Figure 6: The proposed attentive perceptual loss ($L^{att}$) significantly improves the perception-distortion trade-off compared to the widely used perceptual loss ($L_p$) in a ($\sigma = 7$) gaussian deblurring experiment with the DnCNN on the DIV2K data-set. The two different points in each curve are obtained by varying the values of $\alpha$ in Eq.-(9) and -(10) in the main paper and averaging the results over the DIV2K test set.

demonstrate that our proposed extension has the ability to deliver a significantly better perception-distortion trade-off compared to the perceptual loss. Considering that the perceptual loss is the benchmark loss function in perceptual image restoration, our proposed extension can effectively be used to train novel image restoration CNNs for perceptually enhanced performance.

References

[1] M. B. Amor, F. Kammoun, and N. Masmoudi. Improved performance of quality metrics using saliency map and csf filter for standard coding h264/avc. Multimedia Tools and Applications, 77:19377–19397, 2017.

[2] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor. 2018 pirm challenge on perceptual image super-resolution. 2018.

[3] Y. Blau and T. Michaeli. The perception-distortion tradeoff. IEEE CVPR, 2018.

[4] S. Bosse, D. Maniry, K.-R. Muller, T. Wiegand, and W. Samek. Deep neural networks for no-reference and full-reference image quality assessment. IEEE Transactions on Image Processing, 27:206–219, 2017.

[5] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2414–2423, 2016.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

[7] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. The Journal of Physiology, 160:1:106–154.

[8] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer. SqueezeNet: Alexnet-level accuracy with 0.5mb model size. CoRR, abs/1602.07360, 2017.

[9] N. P. Issa, C. Trepel, and M. P. Stryker. Spatial frequency maps in cat visual cortex. The Journal of Neuroscience : the official journal of the Society for Neuroscience, 20:22:8504–14, 2000.

[10] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik. Objective quality assessment of multiply distorted images. 2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), pages 1693–1697, 2012.

[11] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016.

[12] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1646–1654, 2016.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. Commun. ACM, 60:84–90, 2012.

[14] J. J. Kulikowski, S. Marvcelja, and P. O. Bishop. Theory of spatial position and spatial frequency relations in the receptive fields of simple cells in the visual cortex. Biological Cybernetics, 43:187–198, 1982.

[15] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 105–114, 2017.

[16] C. Ma, C. Yang, X. Yang, and M. Yang. Learning a no-reference quality metric for single-image super-resolution. CoRR, abs/1612.05890, 2016.

[17] L. Maffei and A. Fiorentini. The visual cortex as a spatial frequency analyser. Vision research, 13:7:1255–67, 1973.
Figure 7: \((\sigma = 7)\) Gaussian Deblurring on the DnCNN [26] to demonstrate a better perception-distortion trade-off using our proposed \(L_{\text{att}}\) compared to the perceptual loss \(L_p\). At almost the same level of distortions (SSIM), training with \(L_{\text{att}}\) results in restored images with much better perceptual quality (lower PI) compared to training with the widely used perceptual loss \(L_p\).

[18] R. Mechrez, I. Talmi, F. Shama, and L. Zelnik-Manor. Maintaining natural image statistics with the context-
tual loss. 2018. 6

[19] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20:209–212, 2013. 4

[20] M. J. Nadenau, S. Winkler, D. Alleysson, and M. Kunt. Human vision models for perceptually optimized image processing - a review. 2000. 1

[21] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15:3440–3451, 2006. 1

[22] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15:3440–3451, 2006. 3

[23] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 4

[24] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang. Esrgan: Enhanced super-resolution generative adversarial networks. 2018. 1, 3

[25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1

[26] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26:3142–3155, 2017. 6, 9

[27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *IEEE CVPR*, 2018. 1, 4