A HVS-inspired Attention to Improve Loss Metrics for CNN-based Perception-Oriented Super-Resolution

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 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 the underlying spatial frequency, which are not generally well reconstructed during Super-Resolution but are most important in terms of visual sensitivity. Based on this prior, we design a spatial attention map that is applied to feature maps in the perceptual loss and its variants, helping them to identify regions that are of more perceptual importance. The results demonstrate the our technique improves the ability of the perceptual loss and contextual loss to deliver more natural images in CNN based super-resolution.

1. Introduction

Inverse problems such as Single image super-resolution (SISR) and de-noising have been a subject of research for quite some time. Recently, Convolutional Neural Networks (CNNs) have been showed remarkable results for SISR. Traditionally, the emphasis of CNN based SISR has been distortion oriented, where the goal is to minimize the distortion using metrics such as the Peak Signal to Noise Ration (PSNR) and Structural Similarity Index (SSIM). Recently though, perception oriented super-resolution has gained wide interest, where the goal is deep learning techniques to get the best perceptual quality of restored images, maintaining naturalness and photo-realistic appearance. This paper addresses loss metrics for perception oriented super-resolution.

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 Full-Reference 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]. Zhang et al. [27] and Blau et al. [3] further demonstrate how effective the deep features really are as perceptual quality features. More recently, Mecherez et al. [18] proposed a variation of the perceptual loss called the contextual loss, which still employs deep CNN features as perceptual quality features but uses an approximation of the KL-divergence to quantify distance. The contextual loss has been demonstrated to be quite effective in maintaining natural image statistics during SISR. The recent PIRM Super-Resolution Challenge Report [2] clearly iterates that the perceptual loss and the contextual loss are the most widely used loss functions for CNN based perceptual image Super-Resolution.

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. The map is then applied to feature maps of pre-trained networks used in the perceptual loss and contextual loss to give more emphasis to spatially 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] and con-
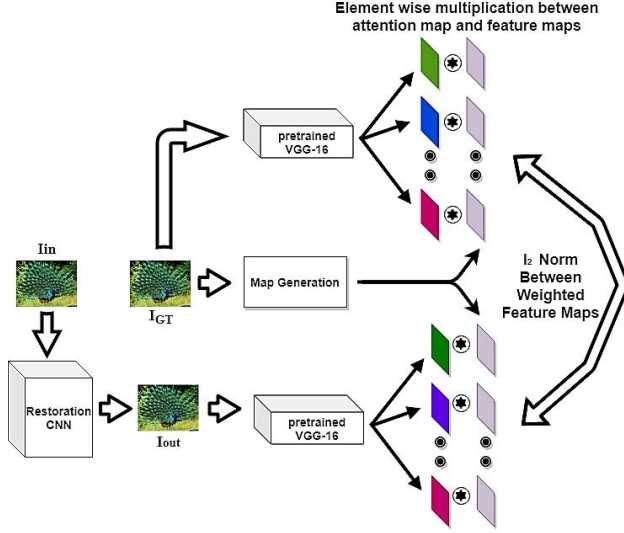
Figure 1: The distance between feature representations of reference and restored images is generally used as a perceptual metric in the perceptual loss and its variants. We propose a spatial attention map, applied to feature representations to enable them to focus on more important regions during the learning process.

textual loss [18].

2. A HVS inspired Attention

2.1. Motivation

In SISR, the low resolution image (I_{LR}) can be modeled as a low-pass filtered and down-sampled version of the high-resolution image (I_{HR}). In the CNN based inverse problem of estimating I_{HR} from I_{LR}, there can be an infinite number of of possible HRs that have the same low-frequency components (relatively easily reconstructed) but different high frequency components. Therefore, the goal for perception oriented SISR should put more emphasis on perceptually important frequency components.

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. Therefore, considering the SISR problem, we will design a CSF based spatial attention mechanism to help perceptual losses put more emphasis on reconstructing perceptually important spatial information, where there is minimal contrast masking in human perception.

2.2. CSF based Map Generation

Let I_{GT} be the Y channel of the ground truth image with size MxN. The 2D DFT of the image will be

$$F[u,v] = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I_{GT}[m,n].e^{-j2\pi(\frac{u}{M}m+\frac{v}{N}n)}$$
(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 $\Delta = 0.25$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).

$$T[m,n] = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} \hat{F}[u,v].e^{j2\pi(\frac{u}{M}m+\frac{v}{N}n)}$$
(5)

As a final step, we normalize the matix '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_u=23 cycles per degree.

(a) CSF

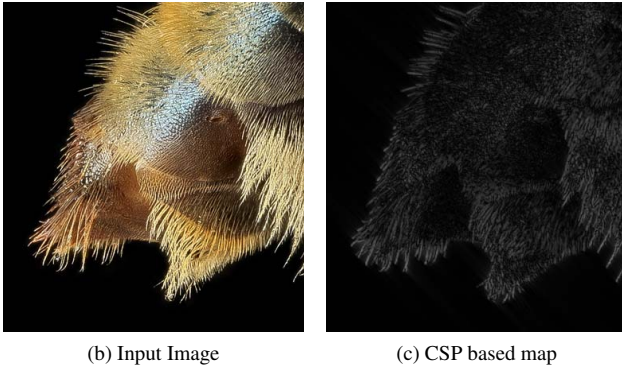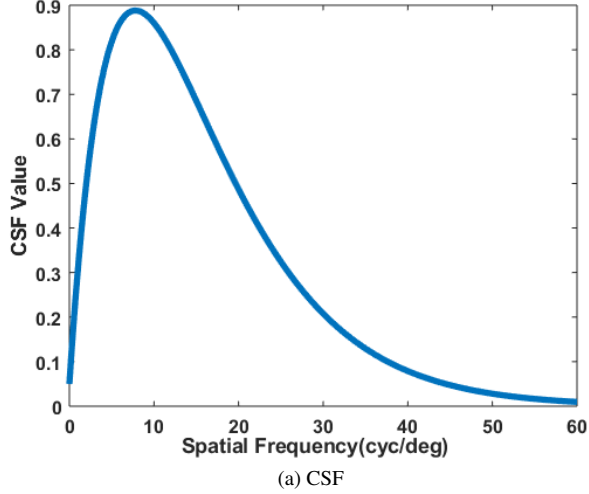(b) Input Image   (c) CSP based map

Figure 2: A Human Contrast Sensitivity Function (CSF), a sample input image and its corresponding attention map.

3. Loss Functions for CNN based Perception-Oriented SISR

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)

$$l_p = \frac{1}{M \times W \times H} \sum_{m=1}^{M} \| \Phi_m^k(I_{\text{out}}) - \Phi_m^k(I_{\text{GT}}) \|_2^2 \quad (7)$$

Where 'Φ_m^k' is the 'm^{th}' feature map in the 'k^{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], im-

age 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. Contextual Loss

The contextual loss [18] takes deep feature representations of images as input and aims to minimize an approximated version of the KL-Divergence. Let $x_i = \Phi_i^k(I_{\text{out}})$ and $y_i = \Phi_i^k(I_{\text{GT}})$ and the normalized pairwise distance $d_{ij} = dist(x_i, y_j)/(\min_k dist(x_i, y_j) + \epsilon)$. The pair-wise affinities are firstly defined as:

$$A_{ij} = \frac{\exp(1 - d_{ij}/h)}{\sum_l \exp(1 - d_{il}/h)} \quad (8)$$

The contextual loss between feature distribution's of the $k - th$ layer of the pre-trained CNN (eg. VGG-16) can be expressed as

$$l_{cx}(x^k, y^k) = -\log\left(\frac{1}{N} \sum_j \max_i A_{ij} \right) \quad (9)$$

As demonstrated in [18], the contextual loss helps maintain natural image statistics in images during SISR.

3.3. Proposed Spatially Attentive Loss Function's

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. (10).

$$l_p^{att} = \frac{1}{M \times W \times H} \sum_{m=1}^{M} \| \mu^k(I_{\text{GT}}) \odot (\Phi_m^k(I_{\text{out}}) - \Phi_m^k(I_{\text{GT}})) \|_2^2$$

$$(10)$$

where $u^k(I_{\text{GT}})$ 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. (10) 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.

For the contextual loss, similarly, instead of inputting the set of feature maps $\Phi^k(I_{\text{out}})$ and $\Phi^k(I_{\text{GT}})$, we will input the feature maps weighted by out attention map i.e $\mu^k(I_{\text{GT}}) \odot \Phi^k(I_{\text{out}})$ and $\mu^k(I_{\text{GT}}) \odot \Phi^k(I_{\text{GT}})$. The modified contextual loss will thus be:

$$l_{cx}^{att} = l_{cx}(\mu^k(I_{\text{GT}}) \odot x^k, \mu^k(I_{\text{GT}}) \odot y^k) \quad (11)$$

The results will demonstrate that our proposed extension improves both the perceptual and contextual loss, delivering a much better perception-distortion trade-off when used for CNN based SISR.

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.

4.2. Objective Quality Test

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 -(10), 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 distortions such as Gaussian Blur and Multiple Distortions [10] (camera image acquisition process where images are first blurred due to narrow depth of field or other de-focus and then corrupted by white Gaussian noise to simulate sensor noise).

4.3. Super-Resolution Experiments

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 using the perceptual and contextual loss separately in in combination with the l_2 loss. For the perceptual loss, we will train the network will network with;

$$L_p = \alpha.l_2 + (1 - \alpha).l_{\mathrm{p}} \tag{12}$$

and

$$L_p^{att} = \alpha.l_2 + (1 - \alpha).l_{\mathrm{p}}{}^{att} \tag{13}$$

and for the contextual loss, we will train the network with

$$L_{cx} = \alpha.l_2 + (1 - \alpha).l_{\mathrm{cx}} \tag{14}$$

and

$$L_{cx}^{att} = \alpha.l_2 + (1 - \alpha).l_{\mathrm{cx}}{}^{att} \tag{15}$$

4.3.1 Perception-Distortion Trade-off

To evaluate and compare the improved performance of our extensions in Eq.(13) and Eq.(15), we will compare the perception-distortion trade-off they deliver in comparison to the standards in Eq.(12) and Eq.(14) respectively. The Structural Similarity Index (SSIM) and the Peak Signal to Noise Ratio (PSNR) have traditionally, somewhat 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 perception-oriented 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 Evaluater (NIQE) [19] and the NRQM [16]. The PI of an image I_{in} is thus defined in Eq.(16)

$$PI(I_{in}) = \frac{1}{2}((10 - NRQM(I_{in})) + NIQE(I_{in})) \tag{16}$$

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 attention improves the perception-distortion trade-off for both the perceptual loss and contextual loss.
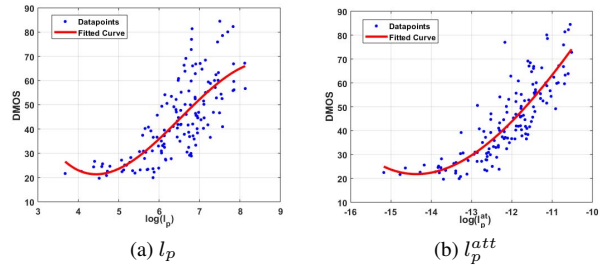

(a) l_p (b) l_p^{att}

Figure 3: It can be seen that our proposed extension (l_p^{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.

Table 1: The attention module improves the correlation with human subjective scores of the modified perceptual loss (l_p^{att}) in Eq.(10) compared to the widely used standard perceptual loss (l_p) in Eq.(7) which leads to the conclusion that l_p^{att} is a better objective perceptual quality metric. The results are repeated across multiple pre-trained networks and their layers.

(SqueezeNet) - (Multiple Distortions)				
Layer	**Metric**	**RMSE**	**LCC**	**SROCC**
Fire2-ReLU_expand3x3	l_p	15.2720	0.6132	0.5589
	l_p^{att}	13.9830	0.6906	0.6084
Fire4_ReLU_expand1x1	l_p	14.5749	0.6570	0.6051
	l_p^{att}	12.9300	0.7434	0.6688
Fire6_ReLU_expand3x3	l_p	15.5951	0.5910	0.5482
	l_p^{att}	14.3458	0.6704	0.6135

Table 1.(b)

(SqueezeNet) - (Gaussian Blur)				
Layer	**Metric**	**RMSE**	**LCC**	**SROCC**
Fire2-ReLU_expand3x3	l_p	11.0894	0.7185	0.7375
	l_p^{att}	10.1010	0.7737	0.8010
Fire3_ReLU_expand1x1	l_p	10.5569	0.7494	0.7641
	l_p^{att}	8.2405	0.8561	0.8712
Fire4_ReLU_expand1x1	l_p	10.2650	0.7652	0.7867
	l_p^{att}	8.5886	0.8425	0.8591

Table 1.(c)

(VGG-16) - (Gaussian Blur)				
Layer	**Metric**	**RMSE**	**LCC**	**SROCC**
ReLU2_2	l_p	10.3392	0.7612	0.7766
	l_p^{att}	8.5227	0.8451	0.8694
ReLU3_2	l_p	9.4553	0.8052	0.8288
	l_p^{att}	8.7198	0.8372	0.8572
ReLU4_2	l_p	9.4181	0.8069	0.8301
	l_p^{att}	8.7319	0.8367	0.8453

Table 1.(d)

(VGG-16) - (JPEG2000)				
Layer	**Metric**	**RMSE**	**LCC**	**SROCC**
ReLU1_1	l_p	7.4194	0.8917	0.8836
	l_p^{att}	5.4533	0.9431	0.9342
ReLU2_2	l_p	7.0756	0.8992	0.8669
	l_p^{att}	5.7946	0.9355	0.9285
ReLU3_2	l_p	6.7050	0.9126	0.9024
	l_p^{att}	6.1821	0.9262	0.9167

Table 1.(e)

(AlexNet) - (Gaussian Blur)				
Layer	**Metric**	**RMSE**	**LCC**	**SROCC**
ReLU_1	l_p	8.8626	0.8313	0.8450
	l_p^{att}	7.5581	0.8805	0.8914
Conv_2	l_p	9.3736	0.8089	0.8281
	l_p^{att}	6.4939	0.9133	0.9216
ReLU_2	l_p	9.5549	0.8006	0.8229
	l_p^{att}	0.8294	0.8327	0.8532

Table 1.(f)

(ResNet-18) - (Multiple Distortions)				
Layer	**Metric**	**RMSE**	**LCC**	**SROCC**
Res2a_ReLU	l_p	13.8896	0.6956	0.6522
	l_p^{att}	12.9627	0.7419	0.6717
Res3a_ReLU	l_p	11.1586	0.8166	0.7805
	l_p^{att}	9.4828	0.8714	0.8273
Res4a_ReLU	l_p	9.4211	0.8732	0.8448
	l_p^{att}	9.0407	0.8839	0.8518

4.4. Results and Discussions

4.4.1 Objective Quality Assessment Experiments

Table. 1 demonstrates the correlation of the perceptual loss (Eq. 7) and our proposed extension (Eq. 10) with human subjective assessment of perceptual quality. The experiment is repeated over multiple pre-trained networks and their layers. 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
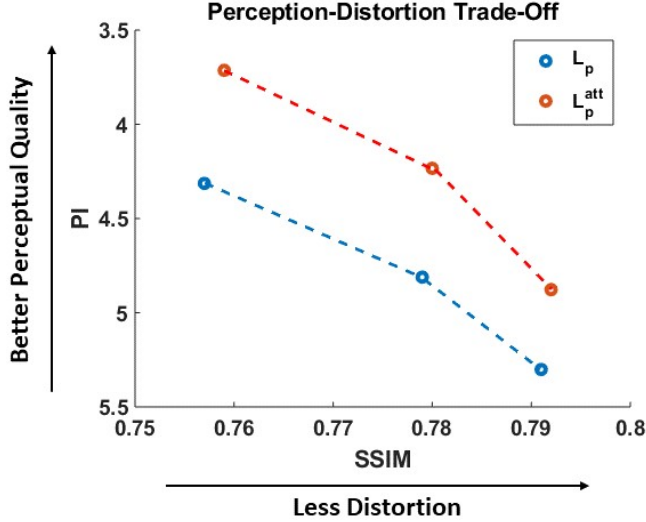
Figure 4: The proposed attentive perceptual loss (L_p^{att}) significantly improves the perception-distortion trade-off compared to the widely used perceptual loss (L_p) in an x4 SR experiment with the VDSR on the DIV2K data-set. The three different points in each curve are obtained by varying the values of α in Eq.-(12) and -(13).

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 one in Fig.3-(a). The next results will demonstrate that this effectiveness enables our proposed extension to improve perceptual image restoration compared to the traditional and widely used perceptual loss.

4.4.2 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.-(12) and -(13) 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 α. The higher the α in Eq.-(12) and -(13), the more weight is given to the pixel-wise l_2 loss compared to the l_p or l_p^{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 if the α 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_p^{att} gives delivers images with much better perceptual quality. Furthermore, even at much lower distor-

tions, the perceptual loss fails to achieve as good perceptual quality as our proposed extension delivers at much higher distortions, indicating a very significant improvement.

Fig. 4 shows some crops from the DIV2K test set images restored using the VDSR trained on Eq.-(12) and -(13). It can be seen that at the same level of distortions (SSIM), the attentive perceptual loss (L_p^{att}) improves the perceptual quality (lower PI) compared to the perceptual loss (L_p). Close examination of the crops also reveals that the L_p^{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.

Similarly, for an x4 SR experiment on the VDSR, Fig. 6 demonstrates the effectiveness of our attention in improving the ability of the contextual loss to deliver more natural images during CNN based SISR. Training the VDSR with Eq. (15) delivers a much better perception-distortion trade-off compared to training it with Eq. (14). This results clearly demonstrate the efficacy of our proposed attention. As the perceptual loss and contextual loss are the most widely used loss functions in perceptual Super-Resolution, as iterated in [2], the proposed technique has high practical impact in CNN based Super-Resolution.

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 mechanism to improve the widely used perceptual loss and contextual loss for perception-oriented SISR. 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 experiment, we also demonstrate that our proposed attention has the ability significantly improve the perceptual loss and contextual loss to deliver more natural images. Considering that the perceptual loss and contextual loss are the two most widely used loss functions in perceptual image restoration, our proposed extensions can effectively be used to train novel image restoration CNNs for perception oriented Super-Resolution.

6. Acknowledgement

(a) Ground Truth   (b) L_p^{att} (SSIM: 0.744, PI: 3.3497)   (c) L_p (SSIM: 0.744, PI: 3.9229)

(d) Ground Truth   (e) L_p^{att} (SSIM: 0.680, PI: 3.4226)   (f) L_p (SSIM: 0.680, PI: 3.7100)

(g) Ground Truth   (h) L_p^{att} (SSIM: 0.777, PI: 3.2619)   (i) L_p (SSIM: 0.778, PI: 4.0202)

Figure 5: Single Image x4 SR on the VDSR [12] architecture. Our proposed attentive perceptual loss (L_p^{att}) in Eq.(13) delivers better perceptual quality(less PI) at same levels of distortion(SSIM) compared to the perceptual loss (L_p) in Eq.(12).

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,
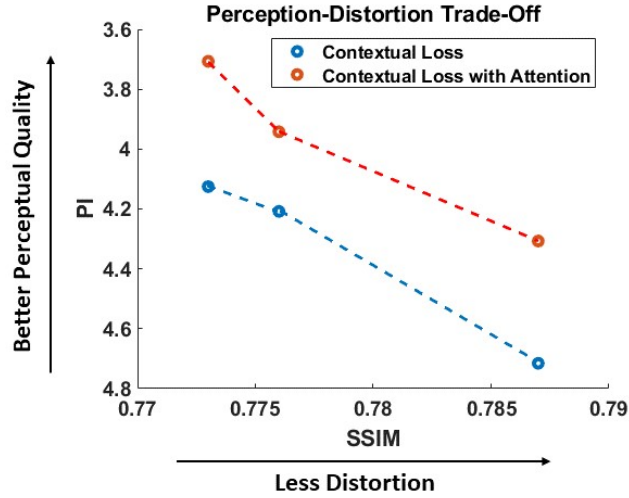
Figure 6: The proposed attentive contextual loss (L_{cx}^{att}) significantly improves the perception-distortion trade-off compared to the contextual loss (L_{cx}) in an x4 SR experiment with the VDSR on the DIV2K data-set. The three different points in each curve are obtained by varying the values of α in Eq.-(14) and -(15).

(a) Ground Truth   (b) L_{cx}^{att} (SSIM: 0.786, PI: 3.9865)   (c) L_{cx} (SSIM: 0.786, PI: 4.5587)

(d) Ground Truth   (e) L_{cx}^{att} (SSIM: 0.742, PI: 4.0813)   (f) L_{cx} (SSIM: 0.741, PI: 4.7239)

Figure 7: x4 SISR on the VDSR [12] architecture. Our proposed attentive contextual loss (L_{cx}^{att}) in Eq.(15) delivers better perceptual quality (less PI) at same levels of distortion (SSIM) compared to the contextual loss (L_{cx}) in Eq.(14).

2017. 2

[2] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor. 2018 pirm challenge on perceptual image super-resolution. 2018. 1, 3, 4, 6

[3] Y. Blau and T. Michaeli. The perception-distortion tradeoff. *IEEE CVPR*, 2018. 1, 4, 6

[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. 3

[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. 1, 3

[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. 4

[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 50x fewer parameters and ¡0.5mb model size. *CoRR*, abs/1602.07360, 2017. 4

[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. 4

[11] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016. 1, 3, 6

[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. 7, 8

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60:84–90, 2012. 4

[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. 1

[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. 4

[17] L. Maffei and A. Fiorentini. The visual cortex as a spatial frequency analyser. *Vision research*, 13 7:1255–67, 1973. 1

[18] R. Mechrez, I. Talmi, F. Shama, and L. Zelnik-Manor. Maintaining natural image statistics with the contextual loss. 2018. 1, 2, 3

[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.

[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. 4

[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.

[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.

[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