A Psychovisual Analysis on Deep CNN Features for Perceptual Metrics and A Novel Psychovisual Loss

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

Abstract

The efficacy of Deep Convolutional Neural Network (CNN) features as perceptual quality features has been demonstrated by researchers. Nevertheless, any thorough analysis in the context of human visual perception on 'why deep CNN features perform well as perceptual features?', 'Which layers are better?', 'Which feature maps are better?' and most importantly, 'Why are some better?' has not been studied. In this paper, we address these issues and provide an analysis for deep CNN features in terms of Human Visual System (HVS) characteristics. We characterize the frequency tuning of feature maps in a trained deep CNN (e.g., VGG-16) by applying grating stimuli of different spatial frequencies as input. We observe that feature maps behave as spatial frequency-selective filters whose characteristics vary with depth. We analyze the frequency sensitivity of deep features in relation to the human contrast sensitivity function and design a novel Visual Frequency Sensitivity Score (VFSS) to explain and quantify how good different deep features are as perceptual quality features. Based on our analysis, we propose a weighting mechanism to discriminate between feature maps on the basis of their perceptual properties and use this weighting to improve the VGG perceptual loss. The results will demonstrate that the proposed psychovisual loss improves the perception-distortion trade-off when used for CNN based image restoration.

1. Introduction

Image quality assessment is divided into two major subfields: (i) No-reference Image quality assessment (NR-IQA) [1]; and (ii) Full-reference Image quality assessment (FR-IQA) [2]. NR-IQA refers to the problem of quantifying the quality of an image without any reference for comparison. FR-IQA, on the other hand, is the problem of assessing the quality of an image relative to a given reference image. In this paper, we focus on FR-IQA. The PSNR (Peak Signal to Noise Ratio) and SSIM (Structural Similarity Index) [3], being fair metrics of distortion between two images, are not a satisfactory metrics to measure perceptive visual quality. The difference and trade-off between distortion and quality has been well explained in [4]. The human visual perception process is fairly complex and still not understood completely. The problem of FR-IQA, on an extremely low level, could be expressed as how differently two images activate different regions in our visual cortex. Considering the highly non-linear nature of CNNs and their arguable similarity to the human visual cortex, it is only natural to assume that deep features perform very well as FR perceptual metrics.

The perceptual loss proposed by Johnson et al. [5] was one of the first to demonstrate how effective the deep CNN features can be as perceptual metrics, especially when used in loss functions for image restoration. After this proposal, the perceptual loss began to be popularly adopted in many image restoration problems such as super-resolution, style transfer, denoising etc. [6],[7],[8]. However even now, little is understood about the efficacy of the deep feature manifold as a perceptual metric, especially in the context of human visual perception. There are important questions such as 'which feature maps are better perceptual features?', 'which layers are better?', 'which and most importantly, 'why some are better than others?'. Zhang et al. [9] and Blau et al. [4] further demonstrated how effective the deep features really are, but a deeper analysis addressing the raised issues is still missing. This paper addresses these important questions in the context of human visual perception.

The spatial frequency theory [10],[11] in neuro-science states that the neurons in the human visual cortex are tuned to specific spatial frequencies or bands of frequencies. This is commonly explained by the contrast sensitivity function of HVS which has different sensitivities to different spatial frequencies of stimuli. An experimental technique, that has been used by neuro-scientists for decades, is used to verify the spatial frequency theory and to find the contrast sensitivity function of HVS [12]. The technique is to stimulate the human visual cortex by showing observers different gratings of different characteristics such as spatial frequency,
eccentricity and orientation. The responses to the stimuli are recorded in the form of functional magnetic resonance imaging (fMRI) or electroencephalography (EEG) potentials. This technique has been instrumental to understand the functional characteristics of the human brain and to verify proposed theories in neuro-science [13],[14]. An interesting question arises: Considering the similarities between CNNs and the visual cortex, (i) could the spatial frequency theory be applied to neural networks as well? (ii) are different feature maps tuned to different spatial frequencies? and (iii) if yes, what implications can this have?

In order to explore these issues, we stimulate the VGG-16 [15] neural network with concentric sinusoidal gratings of different spatial frequencies, and record the responses of feature maps in the form of mean activations of feature maps versus the spatial frequency. We establish a hypothesis that feature maps that are more sensitive to spatial frequencies that the visual cortex is most sensitive to, are better perceptual quality features. We empirically verify our hypothesis by proposing a metric called Visual Frequency Sensitivity Score (VFSS) and utilizing it on various experiments.

To the best of our knowledge, this is the first analysis of its type. Considering the highly non-linear nature of neural networks, it is unlikely that a tractable mathematical analysis of frequency response such as DFT and DCT could be applied. The grating based methods might be a good alternative to quantify the frequency responses of neural network feature maps.

The main contributions of our work are summarized as follows:

1. Taking inspiration from neuro-scientific experiments, we devise a novel technique to quantify the spatial frequency tuning characteristics of deep CNN features.

2. We present the first analysis and quantification to determine which feature maps are better perceptual quality features in the context of frequency sensitivity in human visual perception.

3. We propose a Visual Feature Sensitivity Score (VFSS), that can be used in feature selection and improvement of techniques that utilize deep features as full-reference perceptual metrics.

4. We apply our analysis to propose a weighted improvement of the VGG perceptual loss which we call the psychovisual loss. The proposed loss results in a better perception-distortion trade-off compared to the perceptual loss when used for image restoration.

2. A Psychovisual Approach

2.1. Spatial Frequency Theory

The spatial frequency theory theorizes that the visual cortex operates on a code of spatial frequency, and specific regions or neurons in the visual cortex are tuned to specific spatial frequencies [11]. The theory is contrary to earlier theories of Hubel and Wisel which used straight edges and lines to characterize perception in the visual cortex [16]. The theory is supported by psycho-visual experiments which employ the Fourier theory. According to the Fourier theory, a light distribution in a retinal image can be expressed as a linear combination of basis harmonic components. Grating stimuli have therefore been used in psycho-visual experiments to study the characteristics of the visual cortex.
Experiments such as the ones described in [11], [14] and [17] show that the neurons in various regions of the visual cortex behave as spatial frequency filters. This behavior gives a great insight into 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.

2.2. Methods

Our experimental method is inspired by the grating stimulus experiments used by neuro-scientists to study characteristics of the visual cortex. We generate concentric sinusoidal gratings of a fixed contrast and varying spatial frequencies (cycles per degree), use them to stimulate a trained VGG-16 [15] network and record the responses of the feature maps in the form of mean activation versus spatial frequency. Fig. 1 illustrates the overall scheme of measuring the spatial frequency responses of feature maps in various convolution layers of the trained VGG-16 network. The reason we are using a concentric pattern is to eliminate the factor of orientation sensitivity from our analysis. Some concentric grating stimulus patterns are shown as input to the trained VGG-16 network in Fig. 1.

Some prominent response patterns are observed for feature maps in various convolution layers. Fig.2 shows some representative spatial frequency responses of the feature maps in the trained VGG-16 network for concentric stimuli of various spatial frequencies.

Figure 2: Spatial frequency responses of the feature maps in several internal convolution layers of the trained VGG-16 network. The x-axis in each subfigure indicates the spatial frequency in cycles per degree, and each y-axis is the mean activation of a feature map in a convolution layer.

3. Deep Features as Perceptual Metrics

3.1. Motivation

The main motivation behind using deep CNN features as perceptual metrics is that: (i) instead of a distance measure between two images being a good FR metric, computing distance after non-linear transformation of images into a high dimensional manifold, might result in a better perceptual quality measure. The high dimensional manifold in this case is the manifold of CNN features. The general form for the perceptual loss [5] is given by Eq. (1)

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

Where $\Phi^k_m$ 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 [22], image restoration [7] and style transfer [8] problems.

However, little is known about the internal characteristics of deep features as perceptual features. In the next sections, we will demonstrate a method to discriminate between different feature maps in a layer on the basis of their efficacy as perceptual quality features. Prior knowledge about the efficacy of different feature maps will allow us to propose a weighted extension of the perceptual loss which proves to be very effective.

3.2. Visual Frequency Sensitivity

In this section, we will use the results of the grating experiment to introduce the concept of visual frequency sensitivity. Let us consider the FR-IQA problem. Noise or
distortions result in alterations in the frequency spectrum of an image. When important visual frequency components are altered, the resulting distortions are more perceptible to the observers. The more sensitive an FR-IQA metric is to these perceptual distortions, the better it performs an objective quality measure. The frequencies where the HVS is highly sensitive are termed as Sensitive Visual Frequencies (SVFs). The CSF is a measure of the visual system’s sensitivity over different spatial frequencies.

Considering the presented analysis on the spatial frequency selective behavior of deep feature maps. Our hypothesis is that the feature maps that are more sensitive to SVFs, can be better perceptual features, which can more appropriately be used for perceptual metrics. Consider Fig.3. Suppose some reference image, and a distorted version of the reference are separately input to a CNN. The distortion will effect important visual frequencies in the image spectrum, and cause the CNN feature maps to activate differently for both images. Hypothetically, a feature map that is more sensitive to sensitive visual frequencies should serve better as an indicator of perceptual differences. In Fig.3, it can be seen that different feature maps may have different sensitivity to different spatial frequencies.

We model this behavior by proposing a metric we call the Visual Frequency Sensitivity Score (VFSS). The VFSS of a feature map is defined as

\[ VFSS(k, m) = \sum_f CSF(f) \frac{\partial a_m^k}{\partial f} \]  

where ‘k’ is the index for the convolution layer, ’m’ is the feature map index in each convolution layer, ‘CSF’ is the contrast sensitivity function (CSF), ‘a’ is the mean activation of the feature map and ’f’ is the spatial frequency in cycles per degree. The VFSS quantifies the average sensitivity of a feature map to weighted by the CSF over different spatial frequencies. The feature maps having higher VFSS values should serve more importantly as perceptual features according to our hypothesis, because they can be more sensitive to visually perceivable distortions in input images. Therefore, the VFSS can be used:

- To determine which feature maps are better perceptual quality features;
- To calculate a layer wise Mean Visual Frequency Sensitivity Score (MVFSS) to determine which layers have more feature maps that are better perceptual quality features;
- To acquire prior knowledge about potential efficacy of deep feature maps as perceptual quality features that can benefit applications such as FR-IQA, Image restoration, deep image compression and accelerating SR and IQA CNN’s.

### 3.3. A Novel Psychovisual Loss for Image Restoration

As hypothetically, the VFSS is a measure of how good a feature map as is a perceptual quality feature, the feature maps weighted by the VFSS should improve perceptual properties of the VGG perceptual loss [5]. Therefore, we propose a psychovisual loss by extending Eq. (1) to Eq. (3):

\[ l_{ps} = \frac{1}{MWH} \sum_{m=1}^{M} w_m^k \| \Phi_k^m(I_{out}) - \Phi_k^m(I_{GT}) \|_2^2 \]  

where

\[ w_m^k = \frac{VFSS(k, m)}{\sum_{t=1}^{M} VFSS(k, t)} \]  

It should be noted that \( w_m^k \) is defined as a normalized VFSS. The psychovisual loss in Eq. (3) indicates the distance between a ground-truth image and a restored one in the feature space projected by the trained network (e.g., VGG-19) where the projections are weighted by the normalized VFSS in Eq. (4). The experimental results will demonstrate that the psychovisual loss in Eq. (3) helps achieve a better perception-distortion trade-off compared to the perceptual loss when used for image restoration.

### 4. Experiments

#### 4.1. Overview

In order to verify the effectiveness and applicability of the proposed approach based on our hypothesis, we perform extensive experiments. For this, we first verify our hypothesis by performing objective quality assessment (OQA) experiments based on our proposed methods and comparing them with the results of subjective quality assessments (SQA). This experiment will indicate how well the Objective metrics are aligned with human judgment of quality. Secondly, we perform an image restoration experiment (super-resolution) with our proposed psychovisual loss to demonstrate that we achieve a better perception-distortion trade-off.

The first experiment includes two OQA experiments in comparison with the results of human SQA. For this, we use the image quality dataset, called LIVE dataset, with subjective quality scores from the LIVE Lab [2]. A more detailed description about of the dataset can be found in [2]. The Gaussian blur distorted images in the LIVE dataset have been used in our experiments. The first OQA experiment is to verify our hypothesis by comparing the performances of Eq. (1) and Eq. (3) with respect to the results of SQA.

In order to further reinforce the validity of our hypothesis, we design another OQA experiment. We take two
smaller subsets from the set of feature maps in layers of a pretrained VGG-16. The first subset consists of 15% of the feature maps with the highest VFSS scores in a layer. The second subset consists of 15% of the feature maps with the lowest non-zero VFSS scores in a layer. Incorporating the subsets into Eq.(3), for the first subset with the Top-15% VFSS feature maps, we use

\[ w^k_{m} = \begin{cases} 
1, & \text{if } m \in \text{Top} - 15\%. \\
0, & \text{otherwise.}
\end{cases} \quad (5) \]

For the second subset with the Bottom-15% VFSS feature maps, we use Eq.(3) with

\[ w^k_{m} = \begin{cases} 
1, & \text{if } m \in \text{Bottom} - 15\%. \\
0, & \text{otherwise.}
\end{cases} \quad (6) \]

If our hypothesis is correct, Eq.(3) should serve as a much better objective metric with weights in Eq. (5) compared to weights in Eq. (6).

For an image restoration experiment, we compare our proposed psychovisual loss with the VGG perceptual loss [5] in the context of super-resolution performance. We have used the VDSR [23] network for super-resolution (SR) of x4 upscaling with the DIV2K dataset [24] for our experiments. We demonstrate that training with the psychovisual loss results in a better perception-distortion trade-off compared to training with the classical VGG perceptual loss [5].

### 4.2. Results and Discussions

#### 4.2.1 Objective Quality Experiments

Tables 1 and 2 compare the Root-Mean-Square-Error (RMSE), LCC (Linear Correlation Coefficient) and SROCC (Spearman Rank Order Correlation Coefficient) performances for the Eq. (1) and Eq. (3) on the Gaussian blur distortion dataset, for different layers of the trained VGG-16 network with the Subjective Quality Scores in the LIVE dataset. The log of the error scores in Eq. (1) and Eq. (3) were fitted with subjective scores using a third degree polynomial. The RMSE, LCC and SROCC were calculated after fitting as explained in [2]. As observed in Tables 1 and 2, the psychovisual loss \( l_{pv} \) performs better as an OQA metric compared to the classical VGG perceptual loss \( l_p \) for all layers under investigation. This result demonstrates the efficacy of \( l_{pv} \) as an FR Objective Quality Metric and shows that it is more aligned with human perceptual quality judgment. It also reinforces our hypothesis regarding the importance of visual frequency sensitivity in the performance of feature maps as perceptual features.

Whether we can use the VFSS to explain the OQA performance of a convolution layer against other layers is a different problem compared to the cases of the performance comparison among different feature maps within the same layer, which are a much simpler problem. All the feature maps in a layer can be thought of having the same complexity level because they reside in a same transformed manifold. However, two feature maps in different layers are difficult to be compared directly as perceptual features because they reside in different transformed manifolds that are constructed through different degrees of non-linear transforms. That is, when we compare different layers, due to the non-linear nature of neural networks, a variety of complex factors comes into play. So, it is difficult to understand the internal behaviors of neural networks.

In order to study whether the VFSS can be used to explain the efficacy of different layers in comparison to one another, we propose the Mean Visual Frequency Sensitivity Score of different layers (MVFSS) to measure the mean visual frequency sensitivity of each layer which can then be used to determine which layers have effective feature maps for perceptual metrics. The MVFSS of the \( k\)-th layer is cal-

| Layer   | Loss | RMSE  | LCC  | SROCC |
|---------|------|-------|------|-------|
| ReLU1_1 (64) | \( l_p \) | 11.4039 | 0.6989 | 0.7144 |
|         | \( l_{pv} \) | 11.3369 | 0.7032 | 0.7164 |
| ReLU2_2 (128) | \( l_p \) | 10.3392 | 0.7612 | 0.7766 |
|         | \( l_{pv} \) | 9.7808 | 0.7897 | 0.8112 |
| ReLU3_3 (256) | \( l_p \) | 9.4553 | 0.8052 | 0.8288 |
|         | \( l_{pv} \) | 9.3989 | 0.8078 | 0.8305 |
| ReLU4_4 (512) | \( l_p \) | 9.1521 | 0.8189 | 0.8404 |
|         | \( l_{pv} \) | 10.0056 | 0.7786 | 0.7979 |
| ReLU5_5 (512) | \( l_{pv} \) | 9.9556 | 0.7811 | 0.8005 |

| Layer   | Loss | RMSE  | LCC  | SROCC |
|---------|------|-------|------|-------|
| ReLU1_1 (64) | Bottom-15% | 11.3535 | 0.7021 | 0.7118 |
|         | Top-15% | 11.3381 | 0.7031 | 0.7179 |
| ReLU2_2 (128) | Bottom-15% | 11.3508 | 0.7023 | 0.7159 |
|         | Top-15% | 9.9606 | 0.7809 | 0.8010 |
| ReLU3_3 (256) | Bottom-15% | 10.046 | 0.7765 | 0.7951 |
|         | Top-15% | 9.7041 | 0.7935 | 0.8195 |
| ReLU4_4 (512) | Bottom-15% | 10.3312 | 0.7617 | 0.7852 |
|         | Top-15% | 9.1932 | 0.8170 | 0.8451 |
| ReLU5_5 (512) | Bottom-15% | 10.5636 | 0.7490 | 0.7541 |
|         | Top-15% | 9.6460 | 0.7692 | 0.8188 |

Table 1: Our proposed Psychovisual loss is more correlated with human subjective assessment of perceptual quality compared to the perceptual loss.

Table 2: Our VFSS quantification of how good different deep features are as perceptual features is correct as the Top 15% highest VFSS scored features maps perform much better compared to the 15% lowest VFSS scored feature maps.
Table 3: The \( MVFSS_{Top-15\%} \) of each layer under investigation.

| Layer | \( MVFSS_{Top-15\%} \) |
|-------|-------------------------|
| ReLU1 | 0.33                    |
| ReLU2 | 28.53                   |
| ReLU3 | 198.65                  |
| ReLU4 | 270.24                  |
| ReLU5 | 175.23                  |

Table 4: MVFSS measures for five different layers of the trained VGG-16 network

| Layer | \( MVFSS \) |
|-------|-------------|
| ReLU1 | 0.068       |
| ReLU2 | 8.8097      |
| ReLU3 | 64.5575     |
| ReLU4 | 35.6725     |
| ReLU5 | 3.8894      |

calculated as

\[
MVFSS(k) = \frac{1}{M} \sum_m VFFS(k,m) \quad (7)
\]

where \( M \) is the total number of feature maps in the \( k\)-th layer.

Table 4 shows the MVFSS values measured for five different layers of the trained VGG-16 network. It is evident that the ReLU3 layer has the highest MVFSS value followed by the ReLU4 layer. It is noted that these two layers have significantly higher MVFSS values than the other three layers. Considering our analysis, these two layers should deliver features that perform very well as perceptual metrics. The results in Table 1 show that the ReLU3 and ReLU4 layers show the best OQA performances compared to the other three layers. The ReLU1 layer in Table 4 has the lowest MVFSS value and it also shows the worst OQA performance shown in Table 1. Furthermore, the MVFSS drops significantly from the ReLU4 layer to the ReLU5 layer and such performance drop can also be observed in Table 1. Although the change in MVFSS performance is not completely analyzed layer-to-layer, a definite correlation between the MVFSS and OQA performances over different layers is evident. This experimental evidence supports our hypothesis that feature maps, that are more sensitive to spatial frequencies which the human visual system is most sensitive to, are better for perceptual metrics.

Table 2 shows results of the OQA experiment on the Gaussian Blur distorted images from the LIVE subjective quality data-set. It can be seen that the Top-15% VFSS scored feature maps perform much better than the Bottom-15% VFSS scored feature maps for all layers under investigation. This results shows that that the Top-15% feature maps deliver features whose discrimination between images is more correlated with human quality judgment as opposed to the Bottom-15%. These results further reinforce our hypothesis and demonstrate the effectiveness of our approach for the selection of deep features in applications that utilize them as perceptual metrics.

To investigate which layers have a better concentration of high VFSS scored feature maps and whether this concentration can be correlated to the performance of layers as perceptual metrics, we calculate the Mean VFSS scores for the Top-15% VFSS scored feature maps in each layer. For example, in the layer ReLU3 which has 256 feature maps, the \( MVFSS_{Top-15\%} \) will the average VFSS for the 39 highest VFSS scored feature maps. Table 3 shows that the performance of the Top-15% feature maps in different layers in Table 2 is very well correlated with the \( MVFSS_{Top-15\%} \) of the layers. The ReLU3 has the highest \( MVFSS_{Top-15\%} \) score in Table 3 and it can be seen to have the best performing Top-15% feature maps in Table 2. In terms of the SROCC and LCC, the ordering of performance of each layer in Table 2 is correlated exactly with the \( MVFSS_{Top-15\%} \) ordering of the layers. This is exactly in accordance with our hypothesis that layers with feature maps sensitive to important visual frequencies have better perceptual features.

The previous OQA experiments verify that the visual frequency sensitivity of feature maps is an important characteristic that determines the efficacy of deep features as perceptual quality features. To our best knowledge, this is the first analysis of deep CNN features as perceptual features in the context of human visual perception. Results with additional networks and distortion types are provided in the supplementary material.

4.2.2 Image Restoration (Super-Resolution) Experiments

We have used the VDSR [23] network and the DIV2K dataset [24] for a x4 Super-Resolution (SR) problem in our experiments. We select the ReLU4 layer of the VGG-16 for the perceptual loss as it is commonly used. In general for SR, as explained in [5], the perceptual loss [5] is used in combination with pixel-wise losses such as \( l_1 \) or \( l_2 \) loss. Eq. (8) gives the general expression for the perceptual loss.

\[
L_p = \alpha \cdot l_1 + (1 - \alpha) \cdot l_p \quad (8)
\]

Using the VFSS-based weighting in Eq. (4), we extend Eq. (1) to our proposed psychovisual loss in Eq. (9):

\[
L_{pv} = \beta \cdot l_1 + c \cdot (1 - \beta) \cdot l_{pv} \quad (9)
\]

The coefficient ‘c’ has been introduced so that the proposed psychovisual loss has a comparable magnitude to per-
Figure 4: The psychovisual loss (Eq. 9) improves the perception-distortion trade-off compared to the perceptual loss (Eq. 8) for x4 SR with the VDSR on the DIV2K dataset.

The psychovisual loss i.e \( l_p \approx c \cdot l_{pv} \). This is helpful for implementation and comparison and does not effect the performance. We have empirically chosen \( c = 509 \).

It should be reminded that the traditional metrics such as PSNR and SSIM are known to be not highly correlated with perceptual quality and the distortion measured by PSNR and SSIM for a degraded image indicates the net deviation from its reference image. Many different images of different perceptual qualities may have a same distortion for a given reference. This is why objective distortion measures such as PSNR and SSIM do not necessarily account for perceptual quality. To quantify no-reference perceptual quality, a combination of the metrics such as NIQE[25] and NRQM[26] has been used as a perceptual indicator (PI) in recent works [27]. The work of Blau et.al [4] demonstrates that distortion and perceptual quality are in a trade-off relation and this trade-off is the correct measure for quantifying the efficacy of image restoration algorithms as explained in [27]. Fig. 4 demonstrates the our proposed psychovisual loss in Eq.(9) provides a better perception-distortion trade-off compared to the perceptual loss in Eq.(8) when parameters are varied.

Fig.5 shows the x4 SR reconstructions by two VDSR networks: one was trained by the loss function in Eq. (8) and the other by the loss function in Eq. (9). Consider Fig.5-(b) and Fig.5-(d) in which the restored images have roughly the same distortion (SSIM) with the ground truth in Fig.5-(a) but the psychovisual loss improves the perceptual quality (lower PI). Furthermore, even at significantly lower distortion, the perceptual loss based restored image in Fig.5-(c) fails to achieve as good perceptual quality as the psychovisual loss based restored image in Fig.5-(b). Similar results can be deduced from Fig.5-(h) to -(k). This clearly demonstrates that the proposed psychovisual loss in Eq. (9) has the ability to deliver a better perception-distortion trade-off compared to the perceptual loss in Eq.(8).

4.2.3 Potential Applications

Our proposed approach is the first method to quantify how good a deep feature map is as a perceptual feature. It can be used to select layers and feature maps of Deep CNNs for use in perceptual metrics/losses and other applications. The perceptual loss is extensively used for many applications and we have clearly demonstrated that our proposed psychovisual loss delivers a better perception-distortion trade-off and can be used for image restoration, style transfer and FR-IQA applications. Furthermore, deep CNN features are of great importance to the problem of deep image compression and prior knowledge about the potential efficacy of feature maps may help in better quantization techniques to improve the quality of compressed images. Prior knowledge of perceptual efficacy of deep features may also aid in feature compression or using a small number of feature maps for accelerating CNN performance.

5. Conclusions

We characterize the spatial frequency selectivity of deep CNN feature maps using an experimental technique inspired by neuro-science. We propose and verify a hypothesis that the feature maps that show more sensitivity to spatial frequencies which the human visual system is most sensitive to, perform better as perceptual features for perceptual metrics. We quantify this spatial frequency sensitivity of a feature map using a novel metric, called the Visual Frequency Sensitivity Score (VFSS). We demonstrate that the VFSS metric can be used to quantify which feature maps can extract better perceptual features, and which layers have better feature maps. This paper is the first to quantify and explain how good different feature maps are as perceptual quality features, resulting in a novel psychovisual loss that can effectively be used for CNN-based image restoration. In comparison to the popular and widely used VGG perceptual loss, the proposed psychovisual loss applied for a super-resolution problem results in a better perception-distortion trade-off.

References

[1] K. B. V Kamble, “No-reference image quality assessment algorithms: A survey,” Optik - International Journal for Light and Electron Optics, vol. abs/1501.00092, 2015.

[2] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, “A statistical evaluation of recent full reference image quality
Figure 5: x4 SR experiment results on the VDSR show that the psychovisual loss \((L_{pv})\) in Eq. (9) delivers a better perception-distortion trade-off compared to the perceptual loss \((L_p)\) in Eq. (8).
assessment algorithms,” *IEEE Transactions on Image Processing*, vol. 15, pp. 3440–3451, 2006. 1, 4, 5, 10

[3] 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*, vol. 13, no. 4, pp. 600–612, 2004. 1

[4] Y. Blau and T. Michaeli, “The perception-distortion tradeoff,” *IEEE CVPR*, 2018. 1, 7

[5] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” 2016. 1, 3, 4, 5, 7

[6] 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)*, pp. 105–114, 2017. 1

[7] 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

[8] 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)*, pp. 2414–2423, 2016. 1, 3

[9] 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

[10] J. J. Kulikowski, S. Marvecelja, 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*, vol. 43, pp. 187–198, 1982. 1

[11] L. Maffei and A. Fiorentini, “The visual cortex as a spatial frequency analyser.” *Vision research*, vol. 13 7, pp. 1255–67, 1973. 1, 2

[12] *Emergent Techniques for Assessment of Visual Performance*. Committee on Vision, National Research Council, 1985. 1

[13] M. G. B. W. K. L. S. E. K. R. M. Everson, A. K. Prashanth, “Representation of spatial frequency and orientation in the visual cortex,” *Proceedings of the National Academy of Sciences*, vol. 95 14, 1998. 2

[14] 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*, vol. 20 22, pp. 8504–14, 2000. 2

[15] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014. 2, 3

[16] D. H. Hubel and T. N. Wiesel, “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex,” *The Journal of Physiology*, vol. 160 1, pp. 106–154. 2

[17] J. J. Kulikowski, S. Marvecelja, 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*, vol. 43, pp. 187–198, 1982. 2

[18] S. K. O. I. Kato D, Baba M, “Effects of generalized pooling on binocular disparity selectivity of neurons in the early visual cortex.” *Philosophical Transactions of the Royal Society B: Biological Sciences.*, vol. 371 1697, 2016.

[19] L. G. T. Russell L. De Valois, Duane G. Albrecht, “Spatial frequency selectivity of cells in macaque visual cortex,” *Vision Research*.

[20] M. van Wyk, H. Wässle, and W. R. Taylor, “Receptive field properties of on- and off-ganglion cells in the mouse retina.” *Visual neuroscience*, vol. 26 3, pp. 297–308, 2009.

[21] M. A. W. Bhuvanesh Awasthi, Jason Friedman, “Faster, stronger, lateralized: Low spatial frequency information supports face processing,” *Neuropsychologia*, vol. 49 13, pp. 3583–3590, 2011.

[22] 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*, vol. 27, pp. 206–219, 2017. 3

[23] 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)*, pp. 1646–1654, 2016. 5, 6

[24] E. Agustsson and R. Timofte, “Ntire 2017 challenge on single image super-resolution: Dataset and study,” 2017 *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1122–1131, 2017. 5, 6
A. Supplementary Material

B. Implementation Details

In this section, we present details behind the generation of sinusoidal gratings of different spatial frequencies.

The contrast sensitivity function is expressed on the domain of spatial frequency in cycles per degree (cyc/deg). The cycles per degree express the number of sine cycles captured by the observer per unit degree of observation. Obviously, the distance of viewing and dimensions of the screen play an important part in this measurement.

We essentially generate gratings in the computer simulation in cycles per pixel. Let the display screen being used in the experiment have a height ‘h’ inches and resolution ‘r’ pixels per inch. The optimal viewing distance in psychovisual experiments should satisfy a function called the PVD [7]. The PVD is a function that expresses the optimal ratio of viewing distance to the height of the display screen. The optimal viewing distance ‘d’ for the screen with height ‘h’ can be calculated using the PVD.

The transformation between cycles/degree and cycles/pixel is

\[
\frac{\text{cycles}}{\text{pixel}} = \frac{\text{cycles}}{\text{degree}} \times \frac{\text{degrees}}{\text{pixel}}
\]

Where

\[
\frac{\text{pixels}}{\text{degree}} = \frac{180}{\pi \times d \times r}
\]

Therefore,

\[
\frac{\text{cycles}}{\text{pixel}} = \frac{\text{cycles}}{\text{degree}} \times \frac{\pi \times d \times r}{180}
\]

We have tested with a number of different display systems of SD, 2K and 4K resolutions. Considering that the PVD takes the viewing angle into account, the changes in the resultant spatial frequencies of the gratings are small and insignificant. Therefore, it can be concluded that the choice of display system has a negligible effect on the experiment.

C. Additional Experiments and Results

C.1. Additional Networks and Distortions

So far, considering its wide use, we have only investigated the VGG-16 feature maps and demonstrated the efficacy of our approach. In this section, we will extend our approach to additional networks such as;

- AlexNet [31]
- ResNet-18 [29]
- SqueezeNet [30]
and a variety of distortions such as:

- Gaussian Blur
- Fast-Fading distortions.
- White Noise.
- Multiple Distortions (camera image acquisition process where images are first blurred due to narrow depth of field or other defocus and then corrupted by white Gaussian noise to simulate sensor noise.)

The fast-fading and white noise corrupted images and subjective scores were obtained from the LIVE Image Quality Assessment Database [2]. The multiple distortion images and subjective scores were obtained from the LIVE Multiply Distorted Image Quality Database [32].

We demonstrate that our quantification for the efficacy of feature maps as perceptual quality features is correct and our proposed perceptual loss delivers superior performance and is much more correlated with human subjective assessment of quality compared to the traditional perceptual loss. Considering the results in Table. - (1) to - (6), it can be observed that our proposed psychovisual loss ($l_{pv}$) is a superior metric compared to the perceptual loss ($l_p$) for a wide variety of networks, their layers and different types of distortions. The superiority is verified by its better correlation with human subjective assessment of perceptual quality and deliverance of better perception-distortion trade-off for the x4 SR experiment we have performed, as shown in Fig. 4 of the main text. Considering the widespread use of the perceptual loss in image restoration, we believe that our proposed psychovisual loss can serve as a much better alternative.

As supplements to Table. 2 in the main text, Tables. - (7) and -(8) demonstrate that our proposed quantification about how good different feature maps are as perceptual quality features is correct for other networks and distortions as well. It can clearly be seen that the 15% of the feature maps with the highest VFSS scores are a much better set of perceptual quality features compared to the 15% of the feature maps with the lowest non-zero VFSS scores.
Table 5: Our proposed Psychovisual loss is more correlated with human subjective assessment of perceptual quality compared to the perceptual loss.

| SqueezeNet- Multiple Distortions | Metric | RMSE  | LCC  | SROCC |
|---------------------------------|--------|-------|------|--------|
| fire2-ReLU_exp3x3               | $l_p$  | 15.2720 | 0.6132 | 0.5589 |
|                                 | $l_{pv}$ | 14.2447 | 0.6761 | 0.6110 |
| fire4-ReLU_exp1x1               | $l_p$  | 14.5749 | 0.6570 | 0.6051 |
|                                 | $l_{pv}$ | 14.1847 | 0.6795 | 0.6292 |
| fire6-ReLU_exp3x3               | $l_p$  | 15.5951 | 0.5910 | 0.5482 |
|                                 | $l_{pv}$ | 14.0247 | 0.6873 | 0.6607 |

Table 6: Our proposed Psychovisual loss is more correlated with human subjective assessment of perceptual quality compared to the perceptual loss.

| SqueezeNet - Gaussian Blur       | Metric | RMSE  | LCC  | SROCC |
|---------------------------------|--------|-------|------|--------|
| fire2-ReLU_exp3x3               | $l_p$  | 11.0894 | 0.7185 | 0.7375 |
|                                 | $l_{pv}$ | 9.4110  | 0.8072 | 0.8323 |
| fire4-ReLU_exp1x1               | $l_p$  | 10.2650 | 0.7652 | 0.7867 |
|                                 | $l_{pv}$ | 8.9756  | 0.8265 | 0.8525 |
| fire8-ReLU_exp1x1               | $l_p$  | 10.8534 | 0.7326 | 0.7531 |
|                                 | $l_{pv}$ | 10.4643 | 0.7545 | 0.7683 |

Table 7: Our proposed Psychovisual loss is more correlated with human subjective assessment of perceptual quality compared to the perceptual loss.

| AlexNet - Gaussian Blur         | Metric | RMSE  | LCC  | SROCC |
|---------------------------------|--------|-------|------|--------|
| ReLU_2                          | $l_p$  | 9.5549 | 0.8006 | 0.8229 |
|                                 | $l_{pv}$ | 9.2554 | 0.8143 | 0.8611 |
| ReLU_3                          | $l_p$  | 6.1683 | 0.9221 | 0.9327 |
|                                 | $l_{pv}$ | 4.4803  | 0.9597 | 0.9588 |
| ReLU_5                          | $l_p$  | 8.4921 | 0.8464 | 0.8615 |
|                                 | $l_{pv}$ | 6.8048  | 0.9044 | 0.9078 |

Table 8: Our proposed Psychovisual loss is more correlated with human subjective assessment of perceptual quality compared to the perceptual loss.

| ResNet18 - Gaussian Blur        | Metric | RMSE  | LCC  | SROCC |
|---------------------------------|--------|-------|------|--------|
| Res4a_Relu                      | $l_p$  | 5.6106 | 0.9360 | 0.9349 |
|                                 | $l_{pv}$ | 4.7937 | 0.9537 | 0.9509 |
| Res4b_Relu                      | $l_p$  | 5.9010 | 0.9290 | 0.9284 |
|                                 | $l_{pv}$ | 5.3772  | 0.9414 | 0.9370 |
| Res5a_Relu                      | $l_p$  | 6.2196 | 0.9208 | 0.9242 |
|                                 | $l_{pv}$ | 5.6946  | 0.9340 | 0.9321 |

Table 9: Our proposed Psychovisual loss is more correlated with human subjective assessment of perceptual quality compared to the perceptual loss.

| AlexNet- Fast Fading Noise      | Metric | RMSE  | LCC  | SROCC |
|---------------------------------|--------|-------|------|--------|
| ReLU_3                          | $l_p$  | 4.2471 | 0.9670 | 0.9631 |
|                                 | $l_{pv}$ | 3.6872 | 0.9753 | 0.9710 |
| ReLU_4                          | $l_p$  | 4.5008 | 0.9629 | 0.9585 |
|                                 | $l_{pv}$ | 3.9098  | 0.9721 | 0.9667 |
| ReLU_5                          | $l_p$  | 5.5131 | 0.9438 | 0.9388 |
|                                 | $l_{pv}$ | 4.9294  | 0.9553 | 0.9511 |

Table 10: Our proposed Psychovisual loss is more correlated with human subjective assessment of perceptual quality compared to the perceptual loss.

| ResNet18- White Noise           | Metric | RMSE  | LCC  | SROCC |
|---------------------------------|--------|-------|------|--------|
| Res4a_Relu                      | $l_p$  | 3.8925 | 0.9707 | 0.9703 |
|                                 | $l_{pv}$ | 3.1727 | 0.9806 | 0.9758 |
| Res5a_Relu                      | $l_p$  | 5.6544 | 0.9370 | 0.9504 |
|                                 | $l_{pv}$ | 4.4505  | 0.9615 | 0.9688 |
| Res5b_Relu                      | $l_p$  | 5.6111 | 0.9380 | 0.9444 |
|                                 | $l_{pv}$ | 4.8910  | 0.9533 | 0.9578 |
Table 11: Our VFSS quantification of how good different deep features are as perceptual features is correct as the Top 15% highest VFSS scored features maps perform much better compared to the 15% lowest VFSS scored feature maps.

| SqueezeNet - Gaussian Blur | Layer       | Metric | RMSE   | LCC    | SROCC  |
|---------------------------|-------------|--------|--------|--------|--------|
|                           | fire2-ReLU_3x3 | Bottom-15% | 10.9308 | 0.7280 | 0.7469 |
|                           |             | Top-15%   | 9.0796  | 0.8220 | 0.8501 |
|                           | fire4-ReLU_1x1 | Bottom-15% | 9.5820  | 0.7993 | 0.8262 |
|                           |             | Top-15%   | 8.2463  | 0.8559 | 0.8809 |
|                           | fire6-ReLU_3x3 | Bottom-15% | 11.1143 | 0.7170 | 0.7347 |
|                           |             | Top-15%   | 7.8518  | 0.8703 | 0.8810 |

Table 12: Our VFSS quantification of how good different deep features are as perceptual features is correct as the Top 15% highest VFSS scored features maps perform much better compared to the 15% lowest VFSS scored feature maps.

| AlexNet - Fast Fading | Layer | Metric | RMSE   | LCC    | SROCC  |
|-----------------------|-------|--------|--------|--------|--------|
|                        | ReLU_1 | Bottom-15% | 8.4147 | 0.8634 | 0.8573 |
|                        |       | Top-15%   | 7.3503 | 0.8977 | 0.9011 |
|                        | ReLU_3 | Bottom-15% | 4.9245 | 0.9554 | 0.9538 |
|                        |       | Top-15%   | 3.6900 | 0.9752 | 0.9702 |
|                        | ReLU_4 | Bottom-15% | 5.0769 | 0.9526 | 0.9471 |
|                        |       | Top-15%   | 3.9109 | 0.9721 | 0.9672 |

Figure 7: VFSS distributions in different layers of the VGG-16 show that only a small proportion of feature maps in a layer are most effective as perceptual features.
Figure 8: Supplement single Image x4 SR results for the VDSR to demonstrate the better perception-distortion trade-off using our proposed psychovisual loss $L_{pv}$ compared to the perceptual loss $L_p$. 