Advanced Image Comparison Metric for Discerning Subtle Distinctions in Visual Quality

Tamoghna Ghosh, Susanta Bhattacharjee, Neelanjan Bhattacharyya, Vicky Bisen, Mukesh Arora, Vishal Sinha, Kunjal Parikh
Intel Corporation

Abstract
Traditional image quality evaluation methods often fail to detect subtle differences crucial for high-stakes applications. Until now, discerning these minute variations has largely been a manual endeavour that is expensive, time-consuming, and lacks repeatability, while existing algorithms like PSNR, SSIM, LPIPS, and FID lack sensitivity. This paper introduces three advancements: effective use of existing metrics for subtle comparisons, a new no-reference metric for detecting fine artifacts, and a novel methodology for detailed image comparison. Using image super-resolution (SR) as a case study, our method captures critical nuances missed by previous analyses.

Author Keywords
Image Quality, Display Perception, Subtle Image Difference, Edge quality.

1. Introduction
In an era where technological advancements are rapidly transforming the visual landscape, the proliferation of sophisticated image generation algorithms has become a hallmark of progress. From the image super-resolution techniques enhancing the visual experience in our television screens and video games to the sharpening algorithms and other AI image enhances that bring our mobile phone displays to life, the quest for impeccable image quality is relentless. The market is continually flooded with new AI-based solutions, each purporting to surpass its predecessors in delivering superior visual experiences. Amidst this constant influx of innovation, consumers and professionals are facing a daunting challenge: determining which algorithm truly meets their specific needs. The subjective nature of visual perception further complicates this task, underscoring the necessity for an objective metric that can reliably evaluate and compare the nuanced performance of these cutting-edge image enhancement tools.

Previous Work: There already exist many robust tools for broad image comparisons. For instance, metrics like the Peak Signal-to-Noise Ratio (PSNR) [1] have been widely used for their simplicity and effectiveness in capturing overall image degradation. Similarly, the Structural Similarity Index (SSIM) [2] has provided a more perceptually relevant measure by considering changes in structural information, luminance, and contrast. Learned Perceptual Image Patch Similarity (LPIPS) [3] measures the similarity between activations of two image patches in a predefined network, aligning closely with human perception. Similarly, Fréchet Inception Distance (FID) [4] is a metric that compares the distribution of generated images to real images by analysing the mean and covariance statistics of both sets.

Despite their utility, these metrics share a common shortfall: they are not fine-tuned to detect *subtle differences*, particularly in small regions within an image. This oversight can lead to significant gaps in quality assessment, especially in cases where precision is paramount.

Moreover, these traditional metrics are not robust to certain simple image transformations like few *pixels shift*. Take, for instance, text displayed on a screen: when rendered natively at high resolution, the text appears crisp and clear. However, if the same text is first rendered at a lower resolution and then upscaled using a high-quality super-resolution model, a slight pixel shift may occur. This shift is not important distortion for human eye and hence a human will rate such generated image as very good compared to original. However, if we apply SSIM or LPIPS on these we noticed that we get very poor scores. Figure 1 shows one example to demonstrate this.

All these metrics discussed above need a reference image. We also need *no-reference* metrices for scenarios where the original image is unavailable for comparison. For instance, image sharpening – we may not have a best benchmark saying that this is the best sharpened image. In literature there are few popular no-reference metrices like: BRISQUE [5], which evaluates image quality using spatial features alone, providing a single holistic score for the entire image. Similarly, The Perception based Image Quality Evaluator (PIQE) [6] is an unsupervised, opinion-unaware metric that measures image quality with arbitrary distortion by estimating block-wise distortion and local variance.

The current no-reference metrics are unreliable for detecting subtle differences, especially edge sharpness on text as shown in section 4 Table3. Sharpened texts are easier to read. But, if an algo improves text sharpness from original, existing metric will treat that as drawback as the output is different from input. Accurate identification of these issues is crucial for developing advanced image quality algorithms. This gap presents an opportunity to create new edge quality metric. See Section 4 for details.

***Our approach*:** We seek to address these critical gaps by building upon the foundations laid by previous metrics. We introduce a novel methodology that enhances sensitivity to subtle differences and improves robustness against image transformations. Meaningful information for the human visual system is often concentrated in specific areas of an image rather than being uniformly distributed. For instance, pixels surrounding text or icons in a graphics-generated image hold more significance compared to background pixels, which are frequently just solid colours. Additionally, human vision primarily focuses on the central two degrees of the visual field. Current full-frame image comparison metrics average the differences in these informative areas with a larger number of mostly unchanged pixels that lack significant information. This averaging process dilutes the overall differences in images, which a human observer would easily detect and rate differently. Additionally, we have developed techniques to ensure that our metric remains consistent and reliable, even when images are rotated or shifted. This level of precision opens new possibilities for evaluating visual quality

ISSN 0097-996X/25/5603-1535-$1.00 © 2025

of applications where prolonged exposure to images is common, such as in detailed design work, thorough image analysis, or extensive reading.

In this paper, we will explore the limitations of existing image comparison metrics, discuss the importance of detecting subtle differences, and present our innovative solution. We introduce a novel *edge quality metric* for recognizing the critical role of edge quality in the perception of image sharpness and detail, our metric is adept at evaluating this aspect with precision.

By the end of this journey, we aim to demonstrate how our approach not only bridges the gap left by previous methods but also paves the way for a new standard in image quality assessment.

2. Capturing Subtle differences

We understand that it is crucial for any good metric to capture the subtle differences to have high correlation with the user perception. In the usual UX studies, the user is given couple of images (side by side / one after another) and is asked to rate the image quality. While the user compares/assess the complete image, his/her rating is highly dependent on the bad regions of the images. For example, if there is an image with ~95% regions in high quality and just a small fraction of ~5% with bad quality, the user's perception and the rating will be highly biased on that small fraction. But the objective metrics doesn't behave in this fashion rather they rate the image quality overall. In the above example, objective metric might give a high score of 0.95 and hence it will fail to correlate with the human perception. We drafted a methodology to solve this problem. We created a UI that gives user the flexibility to choose a small region in the full image and then provide rating against this area of interests (AOIs). The standard metrics like SSIM, LPIPS, PSNR, FID is run on these small image AOIs. Users rate selected AOIs from 1 to 5 based on their observations. User rating scale for Subjective study mention in below (Table 1). We used super resolution (SR) as a case study for this and compared outputs from two different SR algorithms for the subjective study.

Table 1. User rating scale for Subjective study

Rating	Description
1	Significantly different and easily distinguishable
2	Clear differences, requires some attention to identify.
3	Minor differences that are not immediately obvious.
4	Very similar with little differences (hard to detect).
5	Almost Identical: extremely difficult to tell apart.

After collecting subjective scores from several audiences and corresponding metric scores, we used Pearson's correlation to assess the strength of their relationship across various image categories (text, game, nature). The result for correlation analysis is mentioned below (Table 2).

However, there is still one more scenario where this correlation may not hold good. Suppose there is a shift of few pixels, but image structure is retained as is, the user's rating might not change or change slightly, but the objective metrics score would change significantly. Hence, it is important to nullify this impact in the objective metrics. In our designed metric, we take care of this limitation and is explained in section 3.

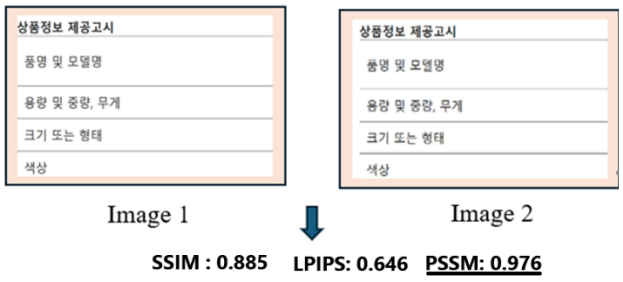
Table 2. Pearson Correlation Analysis for several users with and without AOI approach – Green (strong), Blue (Moderate) and Red (weak & anti) correlation. Statistical error with and without AOI approach was 0.1% Evaluation on two SR algorithm comparison.

Metric	Avg. Pearson correlation with AOI	Avg. Pearson correlation without AOI
SSIM	0.69	- 0.10
PSNR	0.60	- 0.09
LPIPS	0.41	- 0.01

3. Fixing Alignment Issue

Misalignment in images refers to the situation where corresponding pixels in two images do not perfectly overlap due to shifts, rotations, or other transformations. While evaluating any subjective metric, misalignment in the images can lead to inaccurate feature matching and distorted measurements. This issue is particularly problematic for Visual Quality (VQ) metrics such as PSNR, SSIM, and LPIPS. Misaligned images result in misleading VQ scores, compromising the reliability of the analysis. For example, consider image 1 and image 2, which are visually similar. However, image 2 is misaligned by a few pixels to the right. When we compute any of these standard VQ metrics, the scores can be significantly affected by this slight misalignment, as shown in Figure 1.

Figure 1. Effect of Translation on VQ Scores



To address this issue, our approach involves calculating VQ scores for various shifts of a reference grid over a target image and selecting the shift that maximizes the VQ score. Thus, we can find the required translation of image to undo the effect of shift as much as we can. This method ensures optimal alignment, thereby enhancing the accuracy of the VQ metrics. With these two improvements to SSIM we call our metric ***Precision Structural Similarity Metric (PSSM)***. Below is a simple image alignment algorithm we developed to achieve this:

Steps for AOI Alignment

1. Extract Reference AOI: Extract a AOI of size (G) from the reference image at coordinates (x_0, y_0).

2. Iterate and Extract Shifted AOIs: Shift the reference AOI over the target image in all possible directions within the range of ([-M, M]) pixels. Total number of valid pixel shifts (N):

$$N = (\min(M, H - G - x_0) - \max(-M, -x_0) + 1) \times (\min(M, H - G - y_0) - \max(-M, -y_0) + 1) \quad (1)$$

M: Maximum pixels shift in both x and y directions,
G: Size of the AOI,
H, W: Height, Width of the target image
(x_0, y_0): Starting coordinates of the reference AOI

3. Calculate and Compare VQ Metrics: Calculate VQ metrics between the reference AOI and each shifted AOI. Identify the best alignment by choosing the shift with the highest VQ score, indicating the closest match to the reference AOI.
By ensuring optimal alignment, the method enhances the reliability of VQ scores and makes them translation invariant.

4. Edge Quality Metric

To address limitations of existing no-reference metrics, we developed an edge quality metric that is pixel shift agnostic and detects subtle differences reliably as described in Figure 2. The EQ metric processes a single image by applying a Sobel filter to calculate edge strength. It sorts the edge strength in descending order and applies median filter to consider edge strength values above median. The remaining edge strength is normalized by dividing 255. Finally, it calculates the mean to generate a score between 0 and 1. This approach identifies subtle edge sharpness differences across various image types of specially text and icons.

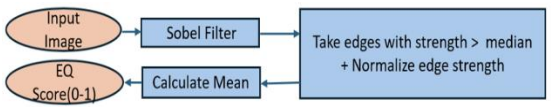


Figure 2. Edge Quality Metric Computation Algorithm

Below are few examples out of many that we tried out for EQ metric evaluation.

Table 3. All metrics scores are normalized to 100 – more implies better quality image

SI	Image	EQ	BRISQUE	PIQE
1		55.2	45.2	17.5
2		42.3	37.6	23.1
3		34.1	64.6	17.0

Table 3 shows that for the image (1), edges are very sharp, and the EQ score reflects this, while PIQE rate shows very poor-quality image. For the image (2), edges lack sharpness but still decent, but very low PIQE score suggests very poor quality. Image (3) clearly lacks edge sharpness compared to image 1 & image 2, detected by low EQ score, though very high BRISQUE rate it as very good quality image, contradicting human perception. These examples indicate that EQ is more reliable for assessing edge sharpness.

In the following section we will describe our framework for automatically identifying potential image crops or area of interests (AOIs) where subtle differences can potentially exist and then compare those AOIs to come with a score describing image quality.

Note that, EQ can be also used as a reference metric by computing the difference of EQ of reference image with generated image. Lower EQ absolute difference means better edge quality match.

5. Image Comparison Framework

To automate the manual AOI selection process, we developed a comprehensive method for comparing two images through a series of steps involving AOI processing, extraction, alignment, and VQ metric evaluation. The aim is to ensure that only the most informative regions of the image are analysed, thereby improving the accuracy and reliability of the comparison.

Figure 4 outlines this methodology in detail. Initially, Image1 is divided into tiny square AOIs. These AOIs are then filtered based on their standard deviation of pixel intensity, which serves as an indicator of information content. AOIs with higher variation in pixel values are retained, as they contain more useful information, while smooth regions like the sky, walls, or blurry backgrounds are removed. Figure 3 shows two examples of automatic AOI selection using this technique.

Auto AOI Selection: Calculate the pixel standard deviation (std) for each colour channel (Red, Green, Blue) within a AOI (AOI size = 25 to 100). The overall normalized std for the AOI denoted by *std_normalized* is computed as follows:

$$\text{std\_normalized} = \text{Max}_{C=R,\,G,\,B}\,(\text{std}(C)/128) \quad (2)$$

(C) denote a single channel of image, normalize it by dividing by the maximum std (128), which represents the variance of a uniform distribution from 0 to 255. AOIs with a normalized std less than a threshold α are skipped. After AOI selection, corresponding AOIs are extracted from Image2 and aligned with those from Image 1 by shifting them by "x" pixels in all possible directions, calculating the VQ Metric for each combination. After calculating the VQ Metric for each combination in all directions, we select the maximum VQ Metric score for each AOI pair. From these scores, we identify the worst first quartile results and compute their mean (Figure 5) to obtain the final VQ Metric, which is used for the overall comparison of the two images.

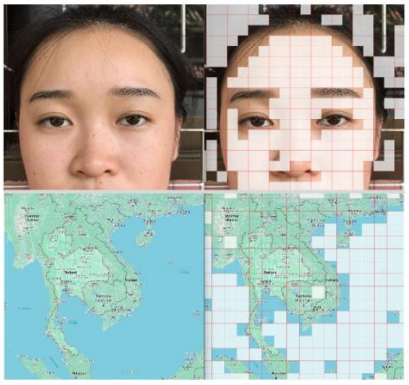


Figure 3. AOI selection is based on equation (2), the white cells in right images are removed and remaining are AOIs.

Here threshold α is chosen based on extensive experiments correlating subjective scores with selected AOI scores, ensuring that skipped AOIs contain minimal meaningful information, such as sky or white backgrounds.

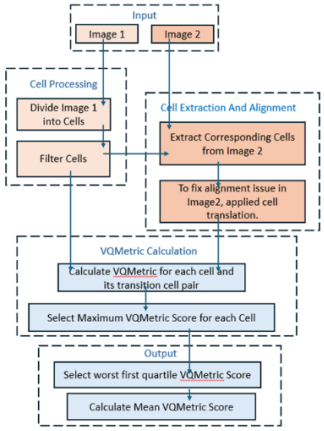


Figure 4. VQ Metric Evaluation Framework

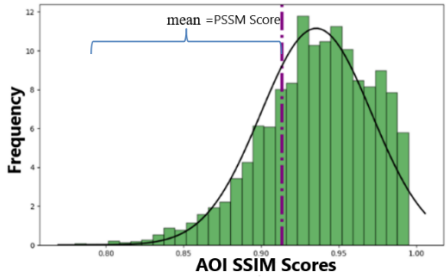


Figure 5. Score Distribution over all selected and aligned AOIs (for $\alpha = 0.25$). The scores less than first quartile are averaged to give the image level PSSM score.

AOI ratings from users (discussed in Table 2) were converted to image level ratings by taking an average of averages of AOIs ratings per user per image. Thereafter, as shown in Table 4 below, the correlation of the *PSSM* with these average image level user ratings are shown. We found AOIs of size 25x25 selected with threshold $\alpha = 0.25$ is best correlating as shown in Table 4 below. Also, the correlation score of 0.65 ($\alpha = 0.25$) is very close to the best correlation score from manual crops (Table 2).

For EQ metric also we found similar corelation with user ratings when evaluated on these selected AOIs. However, sensitivity of EQ metric to the AOI size selection is much lower than PSSM. So even with larger AOI size we get very similar correlation for EQ as with smaller AOI size.

Table 4. Shows the correlation of PSSM scores computed by our framework and user subjective ratings for various values of standard deviation and AOI size (height and width) in pixels.

AOI Size	Standard Deviation Threshold (α)	Pearson Correlation of PSSM with User Ratings
25	0.25	0.65
25	0.15	0.59
50	0.25	0.61
50	0.15	0.56
100	0.25	0.60
100	0.15	0.52

6. Results and Conclusion

PSSM and EQ can be used together to rank various candidate algorithms for super-resolution or image sharpening or any other visual quality enhancement algorithm. The candidate algorithms can be first sorted by best PSSM and them among them candidates with close PSSM values, the once with best EQ can be marked as the best candidate. Also, we can use absolute EQ difference in case we have a reference image to do EQ difference-based ranking. Table 5 shows that the candidates 1-3 have very close PSSM of 0.92 and hence we can't choose which one is best, only based on PSSM. Here, EQ helps us to choose the best – candidate 3 having lowest EQ difference with the native high resolution reference image. Algorithm 4 is a clear winner with highest PSSM and best EQ. 4 is a much powerful and complex algorithm. But, if we must choose among candidates 1-3, then 3 is the best candidate.

Table 5. EQ, PSSM scores for Candidate SR Algorithms

Candidate SR Algorithm	PSSM	EQ Difference
1	0.917	0.063
2	0.924	0.018
3	0.925	**0.003**
4	**0.961**	**0.002**

In conclusion, our research addresses the critical gaps in existing image comparison metrics by introducing three key advancements. First, we introduce metric **PSSM**. Second, we present a new no-reference metric (**EQ**) specifically designed to edge quality of images.

Third, we introduce a comprehensive methodology (Figure 4) that automates completely **PSSM** computation. These two metrics **PSSM** and **EQ** complement each other and facilitates the comparison of fine details of image quality without any manual intervention.

7. References

1. Fardo FA, Conforto VH, de Oliveira FC, Rodrigues PS. A formal evaluation of PSNR as quality measurement parameter for image segmentation algorithms. Centro Universitário da FEI, Sao Paulo, Brazil.
2. Ndajah P, Kikuchi H, Yukawa M, Watanabe H, Muramatsu S. SSIM image quality metric for denoised images. Department of Electrical and Electronics Engineering, Niigata University, Japan; 2010 Sep 3.
3. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. UC Berkeley; OpenAI; Adobe Research.
4. Yu Y, Zhang W, Deng Y. Frechet Inception Distance (FID) for evaluating GANs.
5. Mittal A, Moorthy AK, Bovik AC. No-reference image quality assessment in the spatial domain.
6. JeelanBasha S, Saranya M, AmruthaVarshini P, Sahithi N, Sravani P. Image quality assessment based on NIQE, PIQE, GLCM, and LBP using SVM. Geethanjali Institute of Science and Technology, India