H-008

Subjective Evaluation of Super-Resolution Image Reconstructed by Trainable Regularization

Supatta Viriyavisuthisakul† Parinya Sanguansat‡ Toshihiko Yamasaki§

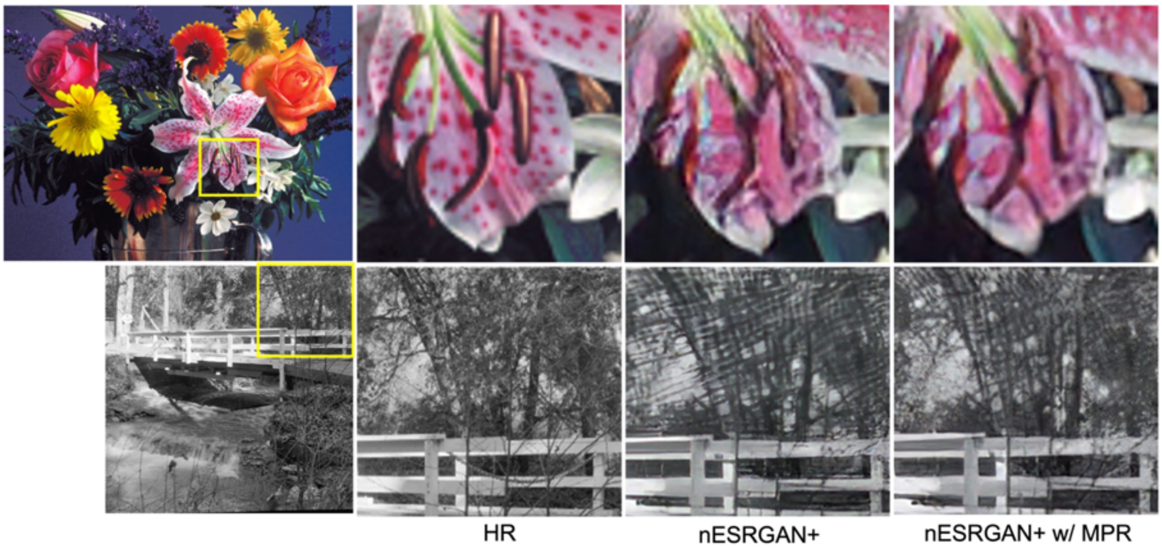

Figure 1 Visual comparison between nESRGAN+ (baseline) and nESRGAN+ with multiple parametric regularization.

1. Introduction

Super-resolution image reconstruction techniques aim to enhance the spatial resolution of low-resolution images, improving their visual quality and enabling finer details to be discerned. These techniques have gained significant attention in various fields, including medical imaging, surveillance, and digital photography. As the demand for high-resolution imagery continues to grow, accurately assessing the quality of super-resolution reconstructed images becomes important [1].

Objective image quality assessment (IQA) metrics, such as peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [2], have traditionally been employed to measure the perceptual quality of reconstructed images. However, the correlation between objective metrics and human perceptual judgments is not always consistent, particularly in the case of super-resolution image reconstruction [1]. Consequently, subjective evaluation that involves human assessments is proposed to show the perceived quality of super-resolution reconstructed images. Typically, Mean Opinion Score (MOS) testing is applied for subjective evaluation methods. It allows direct assessment of visual quality by collecting human perceptual judgments. MOS testing involves presenting observers with pairs or sets of images and asking them to rate the quality based on their subjective perception [3].

† Japan Advanced Institute of Information Technology, Japan
‡ Panyapiwat Institute of Management, Thailand
§ The University of Tokyo, Japan

However, MOS testing can be time-consuming and resource-intensive, requiring a significant number of human participants to achieve reliable results.

In recent years, researchers have explored alternative subjective evaluation approaches that can supplement or replace MOS testing [4]. These methods offer valuable insights into the relationship between objective metrics and human perception, potentially providing efficient and cost-effective means of evaluating super-resolution image reconstruction.

This research aims to investigate the subjective evaluation and traditional IQA of super-resolution image. By examining the correlation between objective metrics and subjective assessments, we aim to gain a deeper understanding of the strengths and limitations of current IQA metrics in capturing human perception. We extend the image evaluation to the subjective domain, comparing the baseline and the images generated by using regularization techniques. The scores from human perception are compared with image quality assessment (IQA) matrices to analyze the agreement between subjective and objective evaluations.

2. Quality evaluation for super-resolution images

Several studies have highlighted the limitations of IQA metrics in evaluating the visual quality of super-resolution images [5], [6]. Research by Jiang et al. [7] demonstrated that while objective metrics achieved high scores for super-resolution images, human observers perceived them to be of lower quality due to the presence of artifacts and loss of fine details. This inconsistency between objective metrics and human perception requires the incorporation of subjective evaluation methods to provide a more comprehensive assessment of super-resolution image quality. Therefore, subjective evaluation methods, such as Mean Opinion Score (MOS) testing, have been extensively employed to gather human perceptual judgments. These methods involve presenting pairs or sets of super-resolution images to human understanding and collecting their subjective ratings based on perceived quality. MOS testing provides a direct measure of human perception and serves as a valuable benchmark for evaluating the effectiveness of super-resolution techniques.

3. Methodology

For evaluating the quality of super-resolution (SR) images, subjective evaluation can be applied in providing human perceptual judgments. Normally, MOS has been widely used to assess the quality of SR images by asking observers to rate an image for its quality on a particular score. This method pro-vides a broad and detailed assessment of the SR image quality. However, this approach can be time-consuming and impractical when dealing with multiple SR methods and many images.

To address this challenge, we conduct a subjective evaluation methodology that enables instantaneous decision-making by comparing the results from different SR methods with a comprehensive set of perceptual criteria. We aim to capture the overall perceived quality of SR images without the need for pixel-level analysis providing a quicker and more efficient evaluation process. It can collect immediate subjective judgments from observers. This subjective activity is conducted by 30 people on test set from Set14.

The three images are shown for each question that are ground-truth, output from nESRGAN+ as a baseline [6], and output from extended the loss function of nESRGAN+ with multiple parametric regularization [8]. A set of images consists of image patches that contain significantly complicated details, as shown in Figure 1. Each sample is selected by varying the texture.

4. Experimental results

In this study, we conduct a binary scoring comparison between the generated images of the baseline and nESRGAN+ with MPR. To evaluate the reconstructed images, PSNR, SSIM and human observers scores visual results are utilized on benchmark datasets on Set14, as demonstrated in Table 1.

PSNR measured the average pixelwise difference between target and distorted images. It can be used to control quality of digital signal transmission. However, it conflicts with human perceptual as it only considers pixel intensities and ignores the image structure. Therefore, SSIM was proposed to measure the similarity of the image structure between the ground truth and reconstructed local patches. A higher PSNR and SSIM score indicate better quality of images.

Our subjective evaluation is instant decision making by human evaluators. A small group of 30 people evaluates 14 sample sets from Set14. For each sample, they choose the image that is most similar to the provided ground truth. The results are presented in Table 1, revealing that nESRGAN+ with MPR can produce higher quality output compared to the original nESRGAN+. While the objective metrics, such as PSNR and SSIM, show minimal differences between the two approaches, the subjective evaluations using the same sample demonstrate a substantial gap.

Table 1 The average IQA scores comparison of PSNR (dB) and SSIM between SISR methods and nESRGAN+ with MPR on Set14.

Dataset	Evaluation matrices	nESRGAN+	nESRGAN+ w/ MPR
Set14	PSNR	24.81	**24.91**
	SSIM	0.782	**0.786**
	Human	158/420	**262/420**

This result suggests that MPR can help the generator to effectively reconstruct the fine details that align with human perceptual judgments.

5. Conclusions

In this study, we conducted a subjective and objective image quality evaluation for image super resolution comparison between the baseline and nESRGAN+ with MPR. PSNR and SSIM are used in objective evaluation. In this paper, we proposed to apply instant decision approach based on human perception instead of using MOS score. This method can offer more practical and time-efficient alternative to MOS evaluation. It enables researchers and practitioners to quickly assess and compare the perceived quality of SR images without the need for pixel-level analysis. The instant decision making provides an overall impression of the image quality and facilitates a more efficient evaluation process. The results from both domains show that MPR can provide better detail and quality than the baseline.

References

[1] S. M. A. Bashir, Y. Wang, M. Khan, and Y. Niu, "A comprehensive review of deep learningbased single image super-resolution," *PeerJ Comput Sci*, vol. 7, pp. 1–56, 2021.

[2] 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

[3] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives," *Multimed Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[4] Afnan, F. Ullah, Yaseen, J. Lee, S. Jamil, and O. J. Kwon, "Subjective Assessment of Objective Image Quality Metrics Range Guaranteeing Visually Lossless Compression," *Sensors 2023, Vol. 23, Page 1297*, vol. 23, no. 3, p. 1297.

[5] X. Wang *et al.*, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," in *European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 1–23.

[6] N. C. Rakotonirina and A. Rasoanaivo, "ESRGAN+ : Further Improving Enhanced Super-Resolution Generative Adversarial Network," *IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 3637–3641, 2020.

[7] Jiang, Q., Liu, Z., Gu, K., Shao, F., Zhang, X., Liu, H., & Lin, W, "Single Image Super-Resolution Quality Assessment: A Real-World Dataset, Subjective Studies, and an Objective Metric," *IEEE Transactions on Image Processing*, vol. 31, pp. 2279–2294, 2022.

[8] S. Viriyavisuthisakul, N. Kaothanthong, P. Sanguansat, M. Le Nguyen, and C. Haruechaiyasak, "Parametric regularization loss in super-resolution reconstruction," *Machine Vision and Applications*, vol. 33:71, no. 5, pp. 1–21, 2022.