FRIEREN: A Lightweight System for Face Resizing Image Detail Quality Evaluation via Robust Estimation of Image Naturalness

Yuan-Kang Lee
Communication Engineering
National Taiwan University
Taipei, Taiwan
r12942062@ntu.edu.tw

Kuan-Lin Chen
Communication Engineering
National Taiwan University
Taipei, Taiwan
r13942067@ntu.edu.tw

Jian-Jiun Ding
Communication Engineering
National Taiwan University
Taipei, Taiwan
jjding@ntu.edu.tw

***Abstract*—In real-time video conferencing systems, webcams often apply image resizing methods, such as nearest-neighbor, bilinear, bicubic, and Lanczos interpolation, to highlight facial regions and enhance user experience. However, interpolations introduce significant high-frequency artifacts that distort perceived image quality. Our work discovered that existing state-of-the-art image quality assessments (IQA) greatly overestimate the sharpness in images resized by nearest-neighbor interpolation, failing to extract useful information in the image's high-frequency components. To address this, we propose FRIEREN (Face Resizing Image Detail Quality Evaluation via Robust Estimation of Image Naturalness), a novel IQA for detail quality evaluation that integrates measures of image naturalness, including motion noise, spatial noise, and HVS-based sharpness. Designed features are fed to Kolmogorov-Arnold Networks (KANs) for quality prediction. Experimental results show that FRIEREN effectively and accurately evaluates the face image detail quality scaled by different interpolations with low computational complexity, making it suitable for quality-aware vision systems.**

***Index Terms*—Image quality assessment, interpolation, noise estimation, sharpness measure, Kolmogorov-Arnold networks**

Fig. 1: Enlarged face images interpolated using the nearest-neighbor, bilinear, bicubic, and Lanczos methods, respectively. It is recommended to zoom in on-screen to observe the different visual effects of different interpolations. From human perception, the detail qualities of interpolated images should be Lanczos $>$ bicubic $>$ bilinear $>$ nearest-neighbor.

I. Introduction

Image quality assessment (IQA) has become essential for the development of multimedia communication nowadays, especially those involving real-time visual interaction. In applications like video conferencing, webcams often use face detection algorithms to automatically center and enlarge facial regions. This enhancement is typically achieved through interpolation techniques such as nearest-neighbor, bilinear, bicubic, and Lanczos methods. Applying different interpolations can cause varying effects on their visual quality [1] [2] [3]. IQA methods aim to analyze the clarity of face images, which is correlated with the precision of the recognition system [4] and the experience of video conferencing. Yet, we discovered that current IQA approaches hugely overestimate the facial image sharpness enlarged using the nearest-neighbor interpolation.

Fig. 1 shows the images of the mannequin face enlarged by different interpolations, and it can be observed that the face image enlarged by the nearest-neighbor method is heavily affected by severe block artifacts. Existing advanced IQAs mistakenly interpret noise-related high-frequency components as textures, underscoring a critical limitation in their ability to distinguish useful high-frequency information in images. To overcome this drawback, we propose a face resizing image detail quality evaluation via robust estimation of image naturalness, named FRIEREN, which is based on two main observations of the human visual system (HVS):

1) Edge-related high-frequency elements are the most important factors in how humans evaluate image sharpness.
2) For the HVS perception, the ranking of detail quality levels in facial images after applying different interpolations is as follows: Lanczos, bicubic, bilinear, and nearest-neighbor method (from the best to the worst).

Three realistic mannequins are used in part of our experiments, shown in Fig. 2. To prevent any impact from post-processing on facial details, all images are captured in RAW format using a Sony IMX383-AAQK image sensor. To the best of our knowledge, our work is the first one that addresses interpolation effects on facial image quality. Fig.

3. shows the overall algorithm flowchart of FRIEREN. The main contributions of our work are summarized as follows:

- A novel no-reference image quality assessment method for face resizing images, FRIEREN, is proposed. By incorporating motion noise, spatial noise, and sharpness measurements, facial detail quality can be evaluated in alignment with human perception.
- The detail quality degradation induced by interpolation can be quantified using our proposed motion and spatial noises and HVS-based sharpness calculation effectively.
- FRIEREN processes images with a size 0f 1920x1080 in 0.1017 seconds, showing its potential for real-time image quality monitoring systems.

II. Proposed Method

A. Quality Score Evaluation using KANs

Predicting image quality requires a regression model that maps features to quality scores to enhance alignment with human visual judgment. We employ Kolmogorov-Arnold Networks (KANs) [5] for the regression in FRIEREN. KANs, based on the Kolmogorov-Arnold theorem, differ from MLPs by avoiding linear outputs. It enables KANs to effectively capture non-linear relationships, making them a way better regression model for simulating human perception. We use the Adam (Adaptive moment estimation) optimizer [6] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to update the model parameters for optimal quality prediction performance. To train the quality prediction model, 64% of the images were randomly selected for training, with 16% used for validation and the remaining 20% for testing. The proposed image features in FRIEREN are introduced in the following sections.

B. Motion Noise Estimation

High-level temporal noise disrupts visual understanding of multimedia content. It cannot be completely eliminated due to the sensor limitations and environmental conditions. Hence, determining the effect of temporal noise is essential to quantify the facial details that a camera can reproduce in images. To adapt FRIEREN into a no-reference method, motion noise is introduced to calculate the temporal noise influence using only one single image frame. Frame modification is applied for motion noise estimation. Denoted *I* as the original image frame. By discarding the last row and column of pixels from *I*, the new image frame *I'* with slightly different content is created. Then, image frames *I* and *I'* are both enlarged to the same dimensions, simulating motion variations typically seen in a video sequence. Let *D* be denoted as the absolute pixel-wise difference between the two frames. The motion noise of the image frame, σ_m, is calculated as in (1)

$$\sigma_m = \sqrt{\frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} (D_{ij} - \mu_D)^2} \tag{1}$$

where μ_D represents the average of the frame difference and *M* and *N* are the height and width of *I*, respectively.

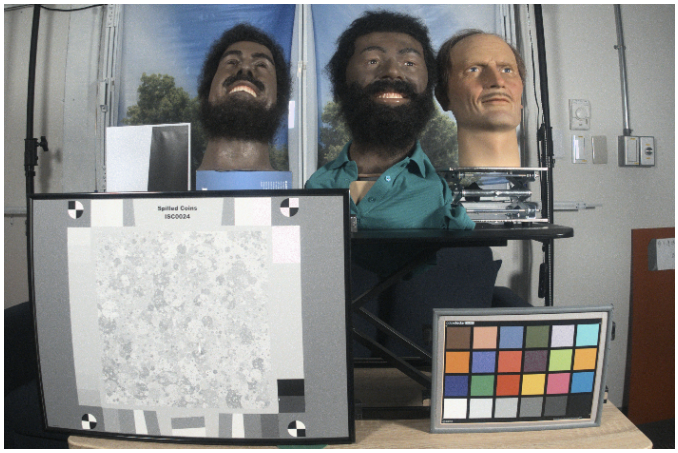

Fig. 2: The three mannequins used in our experiments.

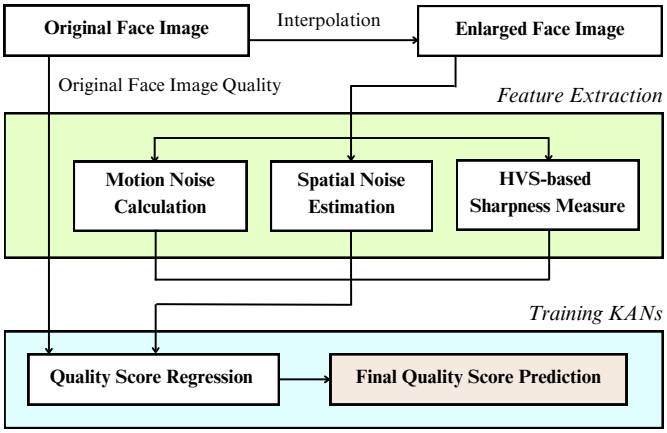

Fig. 3: The overall algorithm framework of FRIEREN.

C. Spatial Noise Estimation

The method we used in FRIEREN is an improved version of the noise estimation model from an adaptive Wiener Filter-based image denoising method [7] [8].The 2D Discrete Wavelet Transform (DWT) decomposes an image into LL, LH, HL, and HH sub-bands, capturing varying detail levels. High-frequency information, including noise and edges, is contained in the LH, HL, and HH sub-bands. Effective noise estimation depends on the ability to distinguish noise from these high-frequency components. By removing edges in the LH, HL, and HH layers, the noisy components can be identified. Denote LL_i, LH_i, HL_i, and HH_i as frequency layers obtained at decomposition level *i* in the 2D DWT process. Let *EM* be the edge detection result applied to the LL_i layer, and *IEM* be the inverted edge mask. The noise energy map *NEM* is expressed as in (2) and (3)

$$IEM = 1 - EM \tag{2}$$

$$NEM = \sqrt{a \cdot LH_i^2 + b \cdot HL_i^2 + c \cdot HH_i^2} \cdot IEM \tag{3}$$

where *a*, *b*, and *c* are weighting parameters for regulating the influence of LH_i, HL_i, and HH_i coefficients, respectively.

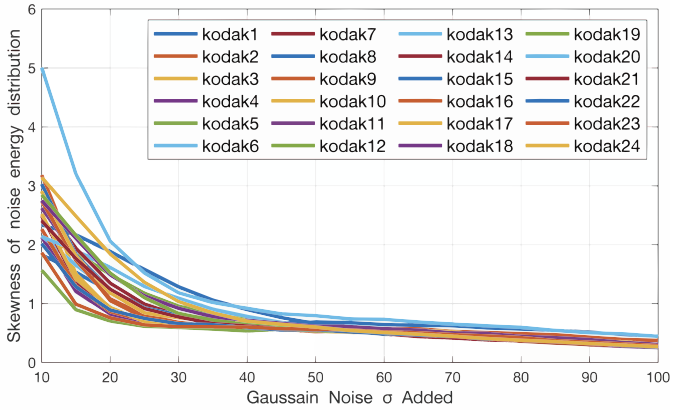

Fig. 4: The relationship between the added σ_{noise} and the skewness of N_D in the Kodak24 dataset.

The noise energy distribution N_D is obtained by removing all zero elements from the *IEM-NEM* multiplication. Noise dominates the edges when an image suffers from severe noise, causing underestimation in the noise influence evaluation. Our proposed approach addresses this problem by recognizing that varying noise energy distributions require different quantiles to identify the most representative energy data for estimation. To validate our hypothesis, we added Gaussian noise with standard deviations ranging from 10 to 100 (in intervals of 5) to all images in the Kodak24 dataset [9] and computed their skewness of N_D, illustrated in Fig. 4. This observation shows that the skewness of N_D remains below 1 consistently when images suffer from high-level noise, and a concise linear relationship exists between the skewness of N_D and the noise standard deviation. By using this mathematical characteristic, we identify the specific quantile that captures data representing the overall noise energy distribution of an image, which can be used as its spatial-domain noise level. Let γ_n denoted as the skewness of an image's noise energy distribution. The image's estimated spatial noise σ_s can be calculated in (4) and (5)

$$Q_n = \alpha_1 \cdot \gamma_n + \alpha_2 \tag{4}$$

$$\sigma_s = Q_n \text{ th quantile of } N_D \tag{5}$$

where Q_n represents the specific quantile for estimation under different noise distribution N_D, and α_1 and α_2 are constant coefficients. Because edge detection performs well on images that suffer from low-level noise ($\sigma_{noise} < 30$), Q_n is set as 45 when γ_n exceeds 1. In real-world scenarios such as video conferencing, image enlargement is commonly applied when a face appears in a frame to make the subject more visually prominent. Since faces before the enlargement are relatively small in the scene, they inherently contain a constrained amount of visual details. Higher decomposition levels can reveal subtle features that might be missed at lower levels [10], improving the noise estimation performance within small face images. We use the decomposition level of 3 in our method. The edge mask *EM* is obtained by applying the Sobel operator on the LL_3 layer.

D. HVS-based Sharpness Measure

Sharpness determines how well-defined the textures and edges of objects appear in images. Inspired by the previous task of spatial noise estimation, our proposed image sharpness measure integrates the spatial-domain and transform-domain methods. We employ edge detection on the image's LL_1 sub-band to effectively capture the edge-related high-frequency components that contain the critical sharpness information. Three levels of DWT high-frequency sub-bands are incorporated into the sharpness calculation. Let the edge mask *EMask* be the result of the Sobel operator applied to the LL_1 sub-band. A fourfold dilation is implemented on the edge mask to encompass all edge-related elements. The edge map *Emap* at the decomposition level *i* can be expressed as in (6)

$$EMap_i = \sqrt{LH_i^2 + HL_i^2 + HH_i^2} \cdot EMask \tag{6}$$

Since the size of the DWT layers is downsampled by a factor of 2 after each decomposition, *EMask* should be resized to match the dimensions of $EMap_2$ and $EMap_3$. The final edge map $EMap_f$ is expressed as in (7)

$$EMap_f = EMap_1 + EMap_2 + EMap_3 \tag{7}$$

where both $EMap_1$ and $EMap_2$ must also be resized to have the same sizes as $EMap_3$. The proposed image sharpness measure, *ISM*, is expressed as the mean value of the final energy map $EMap_f$. *ISM* is then used to assess the clarity of a face image in our model. The method, which combines spatial and transform domain information, minimizes the impact of noise as much as possible while emphasizing the importance of edge features when calculating image clarity.

III. Experiments

A. Experimental Setups

We utilize two face image datasets in our experiments: our collected mannequin face images and the MS1MV2 dataset [11]. Our dataset validates quality degradation due to interpolation, while MS1MV2 demonstrates FRIEREN's effectiveness compared to other IQA methods. In our dataset, gamma correction is applied to generate additional face images, which are enlarged using nearest-neighbor, bilinear, bicubic, and Lanczos interpolation at scales from 2× to 5× (step 0.5). For MS1MV2, 10,000 randomly selected images are upscaled to 2× using all four methods.

A challenge emerges in this task: currently, there are no mean opinion scores (MOS) evaluated by the human subjective judgment for face images enlarged using different interpolations. Since adopting full-reference image quality assessments (FRIQAs) to generate credible pseudo-MOS has proven highly effective [12], we adopt PSNR as our target metric. Suppose that an original face image is scaled up by a factor of β using each of the four interpolations. To produce the PSNR value, we first resize the original face image to $1/\beta\%$ of its size. Next, we enlarge its size by a factor of β, restoring it to the original dimensions. Finally, the PSNR value is calculated using the original image and the interpolated image.

To target MOS values for each interpolation, we scale the MOSs in proportion to the relative PSNR values. That is, the relationship between the MOS values mirrors the ratio of their corresponding PSNRs, which can be expressed in (8):

$$\begin{aligned} &\mathrm{MOS}_{\mathrm{Nearest}} : \mathrm{MOS}_{\mathrm{Bilinear}} : \mathrm{MOS}_{\mathrm{Bicubic}} : \mathrm{MOS}_{\mathrm{Lanczos}} \\ &= \mathrm{PSNR}_{\mathrm{Nearest}} : \mathrm{PSNR}_{\mathrm{Bilinear}} : \mathrm{PSNR}_{\mathrm{Bicubic}} : \mathrm{PSNR}_{\mathrm{Lanczos}} \end{aligned} \quad (8)$$

We use CLIB-FIQA [13], the most advanced face image quality assessment method, to evaluate face images enlarged using Lanczos interpolation. These scores serve as the reference MOS values. After that, all the corresponding target MOS values of each image both in our dataset and in the MS1MV2 dataset are calculated for further regression.

B. Noise Analysis

We hypothesize that the negative influence caused by the nearest-neighbor interpolation can be calculated by the estimations of motion and spatial noise. In Eq. (3), the weighting parameters *a*, *b*, and *c* are set to 0.5, 0.5, and 1, respectively. In Eq. (4), α_1 and α_2 are set to -40 and 85, respectively. Temporal noise is calculated using all previous frames in our dataset. For the MS1MV2 dataset, frame modifications outlined in Section II are applied for motion noise estimation.

The mean estimations of σ_m and σ_s for each enlarged face image in our dataset and the MS1MV2 dataset are listed in Table I and Table II, respectively. The highest results are shown in **bold**. The results support our theory regarding the side effects of the nearest interpolation on face image quality.

TABLE I: Average Temporal And Spatial Noise Estimations For Each Type Of Enlarged Image In Our Mannequin Dataset

	Nearest	**Bilinear**	**Bicubic**	**Lanczos**
Temporal Noise	**0.6178**	0.4880	0.5150	0.5141
Spatial Noise	**18.4668**	14.7780	17.6856	17.7233

TABLE II: Average Motion And Spatial Noise Estimations For Each Type Of Enlarged Image In The MS1MV2 Dataset

	Nearest	**Bilinear**	**Bicubic**	**Lanczos**
Motion Noise	**4.3453**	3.2759	3.5767	3.5807
Spatial Noise	**25.8520**	24.1216	25.1551	25.0978

C. Performance of Quality Score Prediction

Table III and Table IV demonstrate the average PSNR values for each enlarged face image in our dataset and the MS1MV2 dataset, respectively. The highest values are shown in **bold**.

TABLE III: Average PSNR Results For Each Type Of Enlarged Face Images In Our Mannequin Dataset

	Nearest	**Bilinear**	**Bicubic**	**Lanczos**
Mean PSNR	28.6731	29.1761	29.3089	**29.3140**

TABLE IV: Average PSNR Results For Each Type Of Enlarged Face Images In The MS1MV2 Dataset

	Nearest	**Bilinear**	**Bicubic**	**Lanczos**
Mean PSNR	29.9370	31.7721	33.5590	**33.5791**

TABLE V: Performance Comparison of Quality Prediction on Testing Face Images in the MS1MV2 Dataset.

	MGVG	CDV	CLIB	FRIEREN
PLCC	-0.6411	-0.6860	0.7816	**0.8954**
SROCC	-0.6905	-0.6434	0.7327	**0.8723**

In addition to the three proposed image features, CLIB-FIQA scores on the original unenlarged images are also used in the regression. Quality prediction performance is assessed using Pearson linear correlation coefficient (PLCC) and Spearman rank order correlation coefficient (SROCC). To ensure FRIEREN's robustness, train-validation-test operations are repeated randomly 10 times, with median PLCC and SROCC values reported as final results. The prediction in the MS1MV2 dataset, assessed with FRIEREN and other existing state-of-the-art no-reference IQA methods such as MGVG [14], CDV [15], and CLIB-FIQA, are illustrated in Table V. The average predicted quality scores for all methods are illustrated in Table VI. The highest values are shown in **bold**. It is worth noting that CLIB-FIQA is specifically designed for the quality evaluation of face images and trained on the MS1MV2 dataset. FRIEREN's quality prediction best reflects the trend of target MOS values across enlarged face images.

IV. Conclusion

While existing state-of-the-art IQA methods perform well under conventional distortions, our findings reveal that they fall short in evaluating the quality of face images resized via interpolation. FRIEREN is specifically designed to address this limitation. Beyond algorithmic robustness, FRIEREN is characterized by its computational efficiency, achieving real-time performance even in unoptimized implementations. This makes it highly suitable for deployment in embedded systems and real-time video pipelines where processing latency and resource constraints are critical.

TABLE VI: Average Quality Scores of Different IQA Methods In The MS1MV2 Dataset. HVS-preferred Quality Ranking: Lanczos $>$ Bicubic $>$ Bilinear $>$ Nearest-neighbor.

	MGVG	CDV	CLIB	FRIEREN
Nearest	**56.2486**	**32.4979**	**0.7551**	0.6176
Bilinear	28.9318	16.5068	0.7361	0.6615
Bicubic	33.4192	19.1522	0.7443	0.6621
Lanczos	31.9301	18.2963	0.7441	**0.6629**

REFERENCES

[1] M. Moran, M. Faria, G. Giraldi, L. Bastos, and A. Conci, "Do radiographic assessments of periodontal bone loss improve with deep learning methods for enhanced image resolution?," Sensors, vol. 21, no.6, p. 2013, 2021.

[2] N. Z. F. N. Azam, H. Yazid, and S. A. Rahim, "Performance analysis on interpolation-based methods for fingerprint images," in 2022 IEEE 10th Conference on Systems, Process & Control (ICSPC), Dec. 2022, pp. 135–140.

[3] S. Gupta, V. Sandeep, P. R. Rege, V. J. Vijayalakshmi, K. Parashar, and T. Vijayaraj, "Benchmarking different types of interpolation methods for image super-resolution," in 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Jun. 2024, pp. 1–6.

[4] K. Li, H. Chen, F. Huang, S. Ling, and Z. You, "Sharpness and brightness quality assessment of face images for recognition," Scientific Programming, vol. 2021, no. 1, p. 4606828, 2021.

[5] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, and M. Tegmark, "KAN: Kolmogorov-Arnold networks," arXiv preprint arXiv:2404.19756, 2024.

[6] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in Proc. 3rd Int. Conf. Learn. Representations (ICLR 2015), May 2015, vol. 1.

[7] Y.-K. Lee and J.-J. Ding, "Efficient and Accurate DWT-based Image Noise Estimation Using Edge, Skewness, and Statistical Information," Proceedings of the 2024 8th International Conference on Graphics and Signal Processing (ICGSP), pp. 24–29, Jun. 2024

[8] Y.-K. Lee and J. - J. Ding, "Efficient Color Image Denoising using DWT-based Noise Estimation and Adaptive Wiener Filter," 2024 8th International Conference on Imaging, Signal Processing and Communications (ICISPC), pp. 47–51, Jul. 2024

[9] Kodak Lossless True Color Image Suite [Online]. Available: https://r0k.us/graphics/kodak/

[10] Y. Tao, T. Scully, A. G. Perera, A. Lambert, and J. Chahl, "A low redundancy wavelet entropy edge detection algorithm," Journal of Imaging, vol. 7, no. 9, p. 188, 2021.

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4690–4699.

[12] J. Wu, J. Ma, F. Liang, W. Dong, G. Shi, and W. Lin, "End-to-end blind image quality prediction with cascaded deep neural network," IEEE Trans. Image Process., vol. 29, pp. 7414–7426, 2020.

[13] F. Z. Ou, C. Li, S. Wang, and S. Kwong, "CLIB-FIQA: Face image quality assessment with confidence calibration," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2024, pp. 1694–1704

[14] Y. Zhan and R. Zhang, "No-reference image sharpness assessment based on maximum gradient and variability of gradients," IEEE Trans. Multimedia, vol. 20, no. 7, pp. 1796–1808, Jul. 2017.

[15] C. Shi, Y. Lin, and X. Cao, "No reference image sharpness assessment based on global color difference variation," Chinese J. Electron., vol. 33, no. 1, pp. 293–302, 2024.