Image Comparative Index (ICI): A Pixel-Wise Image Similarity Metric for Computational Super-Resolution (SR) Microscopy

Shiraz S Kaderuppan
Faculty of SAgE
Newcastle University
Newcastle upon Tyne, NE7 1RU, U.K.
S.S.O.Kaderuppan@newcastle.ac.uk

Anurag Sharma
Faculty of SAgE
Newcastle University
Newcastle upon Tyne, NE7 1RU, U.K.
Anurag.Sharma@newcastle.ac.uk

Muhammad Ramadan Saifuddin
Faculty of SAgE
Newcastle University
Newcastle upon Tyne, NE7 1RU, U.K.
Ramadan.Saifuddin@newcastle.ac.uk

Wai Leong Eugene Wong
Engineering Cluster
Singapore Institute of Technology
Singapore 138683, Singapore
eugene.wong@singaporetech.edu.sg

Wai Lok Woo
Computer and Information Sciences
Northumbria University
Newcastle-upon-Tyne, NE1 8ST, U.K.
wailok.woo@northumbria.ac.uk

TENCON 2024 - 2024 IEEE Region 10 Conference (TENCON) | 979-8-3503-5082-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/TENCON61640.2024.10902742

***Abstract*—We compare the outputs of several widely-utilized image quality metrics (such as the PSNR, SSIM, MS-SSIM, $FSIM_C$ & IMMSE) in the context of AI-mediated image super-resolution (SR) microscopy derived from a GAN. Although these metrics are often employed in the image analytics space for assaying image quality, the findings of our study indicate that such metrics may not be quite suitable for determining the quality of a GAN-generated image in the context of image SR microscopy, namely in the preservation of fine details and structures which are crucial aspects of high-resolution images (as determined visually). For instance, in some of the assayed images, we observed that these metrics returned a relatively favorable score, while on closer visual inspection, the generated image was observed to be prone to reconstruction artifacts, bearing little similarity to the Expected (ground truth) image. In this respect, we have sought to develop a custom image quality metric capable of assessing image similarity on a pixel-wise scale irrespective of the bit-depth of the image. Our proposed metric [termed the *image comparative index* (ICI)] has proven to be a viable determinant of similarity between 2 images, closely corroborating with a pixel-wise map of the differences between the assayed images, thereby allowing for more detailed analysis and identification of specific areas of improvement. We postulate that these results represent an important consideration for researchers seeking to utilize deep convolutional neural networks (DCNNs) for image SR (particularly when assessing the similarity between a DCNN-generated and the ground truth images during model training) with potential extrapolations of the proposed ICI metric in image classification & object detection use-cases as well.**

Keywords—deep learning, image quality metrics, object detection, image classification, computer vision, computational super-resolution microscopy, PSNR, SSIM, IMSE, GANs

I. Introduction

Deep learning, in particular deep convolutional neural networks (DCNNs), has played a fundamental role in spurring recent innovations in AI-mediated image analysis, including (amongst others) object detection, segmentation, recognition and classification [1]. This has resulted in a recent ongoing drive globally to develop more intricate and advanced DCNN architectures for computer vision applications, including generative adversarial networks (GANs) [2] (such as BigGAN [3], SpectralGAN [4], etc), object detection & image classification algorithms (such as YOLO v9 [5]), and even image super-resolution frameworks (such as BSRGAN [6] & A-Net [7]). In this light, there would emerge an imperative need to *quantify* the performance of each of these DCNNs in their respective allocated tasks, so as to facilitate the selection of the optimal DCNN algorithm for a designated role.

Common metrics often utilized for quantifying the DCNN performance (particularly in the field of image analytics) include accuracy (as defined by the confusion matrix), precision, recall, specificity, F1 score, as well as the area under the curve (AUC) for the Receiver Operator Characteristic (ROC) [8]. These are often supplemented with *traditional* image quality metrics, which are also employed for determining factors such as the presence of image noise [e.g. PSNR [9] & image MSE (IMMSE) / image signature based mean square error (ISMSE) [10])] and the similarity between 2 images (SSIM [11], MS-SSIM [12] & $FSIM_C$ [13]) – 2 aspects which are particularly essential when evaluating the performance of a DCNN-generated image in the context of image classification and super-resolution (SR). Despite the widespread ubiquity of these metrics in assessing the general performance of a DCNN for image analysis, we have encountered significant issues when attempting to use these quantifiers to assay the performance of a DCNN-mediated SR image, the details of which are depicted in the **Results & Analysis** section of this study. In this respect, we have developed an image quality metric (IQM) capable of pixel-wise discrimination between image pairs irrespective of their bit-depth, which we termed the *image comparative index* (ICI) to be subsequently assessed alongside the afore-mentioned metrics. At this juncture, it would be prudent to mention one such popular metric often utilized for assessing similarity between data points in a plot & as a loss function for DCNN training [i.e. the *mean absolute error* (MAE)], although it is incapable of handling higher dimensional data input (such as RGB colour images or video streams). In contrast, our proposed ICI metric proves to be capable of identifying structural homologs between RGB image pairs (based on weakly discernible features), a finding which is not seemingly replicable via other structural similarity indices (such as SSIM or MS-SSIM) while also being resilient to the bit-depth of the

acquired image (by allowing a 24-bit RGB image to be compared against a 30-bit/42-bit/48-bit RGB RAW file), thus further predicating a novelty of the ICI metric in this respect. The results supporting this deduction are presented in the subsequent **Results & Analysis** section of the present study.

II. Methodology

The presently-proposed ICI metric may be mathematically expressed as follows:

$$\text{ICI} = \frac{\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\left|\frac{A_{ij}}{2^{q}-1}-\frac{C_{ij}}{2^{r}-1}\right|\right)_{\mathbf{R}}+\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\left|\frac{A_{ij}}{2^{q}-1}-\frac{C_{ij}}{2^{r}-1}\right|\right)_{\mathbf{G}}+\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\left|\frac{A_{ij}}{2^{q}-1}-\frac{C_{ij}}{2^{r}-1}\right|\right)_{\mathbf{B}}}{3mn} \quad (1)$$

where $\boldsymbol{A_{ij}}$ & $\boldsymbol{C_{ij}}$ represents the pixel in the ith row & j^{th} column of the reference image (***A***) and the test image (***C***) for which the ICI score is being computed; **R**, **G** & **B** represent the Red, Green & Blue channels of a 3-channel colour image; *m* & *n* represent the rows & columns (dimensions) respectively of the images ***A*** or ***C*** (both ***A*** and ***C*** having identical dimensions of *m* x *n*); *q* & *r* represent the bit-depth of a *single* channel in the reference image (***A***) and test image (***C***) respectively (e.g. *q* = 8 for a 24-bit RGB image).

From (1), we may deduce that the *lower* the ICI score, the *better* the image match / similarity, with the ICI scores in the present context having a range of [0, 1]. Hence, for all of the figures and metrics shown in the following **Results & Analysis** section, we assess the ICI scores in this respect, comparing their values against those of the widely-utilized image quality metrics (such as PSNR, IMMSE, SSIM & MS-SSIM). Imperatively, from (1), we may also note that the ICI metric facilitates the pixel-wise comparison between 2 images having different bit-depths (e.g. a 24-bit RGB image against a 30-bit/42-bit RGB RAW image such as is acquired by most DSLRs), through the assimilation of a normalization factor.

All of the image quality metrics were computed in MATLAB R2022b (© 1984-2022, The MathWorks, Inc.), with 4 of them (i.e. PSNR, SSIM, IMMSE, MS-SSIM) derived using their native MATLAB implementations. FSIM_C was determined using a custom MATLAB function published by [13]. These 5 image quality metrics were chosen as they are widely-employed in the industry for evaluating similarities and discrepancies between image pairs (other metrics such as F1, AUC & ROC are also relatively popular, although in the interest of time and scope of the present study, we have chosen to compare our proposed ICI metric against PSNR, SSIM, IMMSE, MS-SSIM & FSIM_C). The images (**Source** & **Expected**) as well as the U-Net & O-Net models used in Figs. 1 & 2 were obtained from [14]. The BSRGAN model (BSRGAN x2) used for the results shown in Table 1 was obtained from [9].

III. Results & Analysis

For the initial tests of this study, we assimilated the BSRGAN architecture (described in [9]) to super-resolve a set of images. The corresponding image quality metrics (such as PSNR, SSIM, MS-SSIM, FSIM_C & IMMSE) coupled with our own self-devised metric (i.e. the ICI) scores were computed for each pair of **Generated-Expected** images, using MATLAB R2022b (© 1984-2022, The MathWorks, Inc.). The findings gleaned from this aspect are illustrated in the following Table:

A. BSRGAN

TABLE I Comparison of Source & BSRGAN-Generated Images using Popular Image Quality Metrics

	Images		Metrics					
	Source	Generated	PSNR	IMMSE	SSIM	MS-SSIM	FSIM_C	ICI
❶			28.9243	83.3003	0.9809	R – 0.8898 G – 0.9531 B – 0.8219	0.9299	0.010883
❷			29.875	66.924	0.9608	R – 0.9504 G – 0.9667 B – 0.8762	0.9134	0.007504
❸			29.4908	73.1135	0.9865	R – 0.9560 G – 0.9641 B – 0.8954	0.9517	0.008208

Table 1. A table illustrating the **Source** and BSRGAN-generated image (scaling factor: 2x). The corresponding image metrics were computed using the **Source** image as the reference (& the **Generated** image as the test) image. A *lower* value for the IMMSE & ICI (& a *higher* value for the PSNR, SSIM, MS-SSIM and FSIM_C) scores correspond to *better* results.

From Table 1, it may be observed that the ICI metric has been shown to closely corroborate visual evidence when assaying the similarity of images from a pixel-wise perspective. Interestingly, it appears that the ICI scores agree with the PSNR and IMMSE quantifiers in determining that the images in Row 2 exhibit the highest quality & closest similarity (between the **Source** & **Generated** images) while those in Row 1 are of the lowest quality & the least similar – a finding which seemingly disagrees with that of the SSIM & FSIM_C indices for these image pairs (both SSIM & FSIM_C postulate that the images with the highest similarity are in Row 3, while those with the lowest similarity scores are in Row 2). Through a visual comparison, we determine that the images in Row 2 do indeed show a closer similarity (between the **Source** & **Generated** images) as compared to those from Rows 1 & 3 (the regions demarcated by the ellipses shown in Table 1 clearly exhibit these differences). However, we also note that there are several regions which appear dark in the figures in Row 2 – it is likely that the SSIM & FSIM_C (coupled with MS-SSIM) metrics penalize the scoring of the image similarities for regions where features are not so evident (as indicated by these darkened regions).

B. U-Net

In the second segment of our testing, we seek to utilize a SR GAN model based on the U-Net architecture [15]. Model training & validation was performed in Python 3.8.5, and as previously, the performance of the model was assayed by means of the same image quality metrics. Fig. 1 & Table 2 below depict our findings obtained in this respect:

FIGURE I U-NET-GENERATED IMAGES FOR FEATURE-WISE COMPARISON AGAINST THE TARGET IMAGE

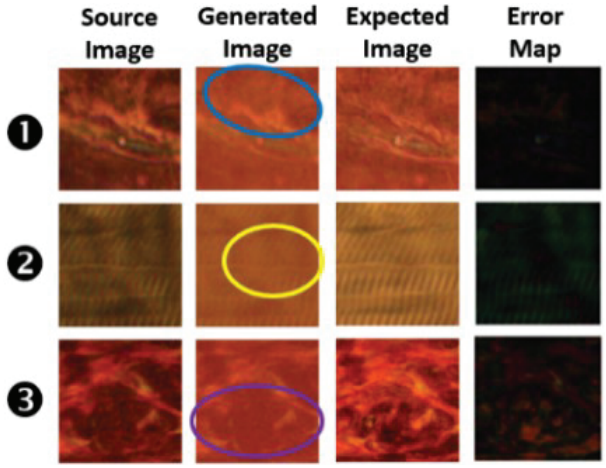

Figure 1. A figure illustrating the U-Net **Generated** images from the respective **Source** images, coupled with their respective **Expected** (ground truth) **Image** and **Error Image**. The yellow ellipse (Column 2) highlights the seemingly discreet striations in the circled region – a feature which is noted to be clearly apparent in the **Expected** Image.

TABLE II COMPARISON OF THE U-NET-GENERATED VS EXPECTED IMAGES USING VARIOUS IMAGE QUALITY METRICS

Metric / Image	PSNR	IMMSE	SSIM	MS-SSIM	$FSIM_C$	ICI
Row 1	27.2286	123.0902	0.9842	R – 0.7821 G – 0.8382 B – 0.8461	0.9164	0.017121
Row 2	23.2917	304.7261	0.9641	R – 0.6583 G – 0.8099 B – 0.8080	0.9002	0.006532
Row 3	21.4657	463.9968	0.9436	R – 0.6190 G – 0.7536 B – 0.8537	0.8525	0.039653

Table 2. A table illustrating the computed metrics (PSNR, IMSE, SSIM, MS-SSIM, $FSIM_C$ and ICI) scores for the images depicted in Fig. 1 previously. Here, the reference image is the **Expected** (i.e. ground truth) image while the U-Net **Generated** image is the test image (which is being compared against the reference).

From Fig. 1 & Table 2, we observe (upon close visual inspection) that the images in Column 2 depict the greatest similarity between the **Generated** & **Expected** images, while those in Column 3 depict the least similarity (the regions of interest being highlighted within the ellipses shown in Fig. 1). Again, this agrees with the ICI scores for these images, while the SSIM, MS-SSIM & $FSIM_C$ scores seem to suggest otherwise; these metrics indicate that the highest scoring image pairs are from Column 1 (not Column 2), although the images in Column 3 are also depicted as the worst scoring image pairs. The PSNR & IMMSE metrics too corroborate with the SSIM, MS-SSIM & $FSIM_C$ scores – an aspect which may lead one to concur that the best scoring image pairs are from Column 1 (although visual inspection suggests otherwise). A closer look at the **Error** maps (images) for the respective Columns reveals that a difference in image *intensities* between the **Generated** & **Expected** images might be the underlying cause of this anomaly, although evaluation of the images from a structural basis (i.e. by comparing the features present in both the **Generated** & **Expected** images) shows that the conclusions gleaned from these metrics (PSNR, SSIM, IMMSE, MS-SSIM & $FSIM_C$) might not accurately reflect the true state of these images, implying the superiority of the proposed ICI metric (over these other metrics) in the current context of image comparison.

C. O-Net

In the third section of our analyses, we implement a SR GAN model based on a relatively recently-published DCNN SR architecture, named O-Net [14]. We used the model supplied by the authors of this study [14], and assessed the performance of our proposed ICI metric against the afore-mentioned key quantifiers of image quality (i.e. PSNR, SSIM, IMSE, MS-SSIM & $FSIM_C$). The results from this analysis are presented in Fig. 2 & Table 3 as follows:

FIGURE II O-NET-GENERATED IMAGES FOR FEATURE-WISE COMPARISON AGAINST THE TARGET IMAGE

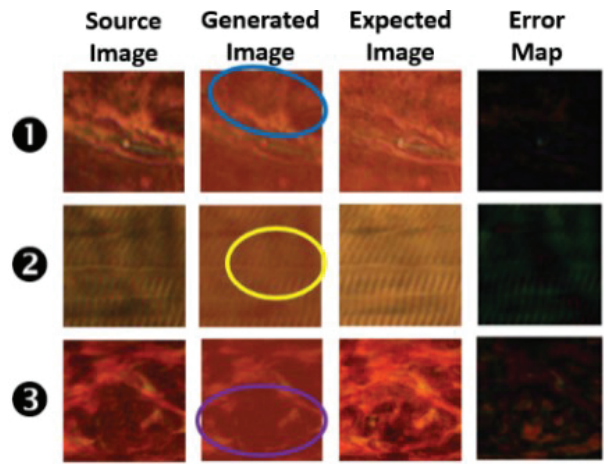

Figure 2. A figure illustrating the O-Net **Generated** images from the respective **Source** images, coupled with their respective **Expected** (ground truth) **Image** and **Error Image**. Notice the more obvious striations in the circled region demarcated by the yellow ellipse (Column 2) – a feature which is observed to be less apparent for the U-Net **Generated** image previously (Fig. 1, Column 2).

TABLE III COMPARISON OF THE O-NET-GENERATED VS EXPECTED IMAGES USING VARIOUS IMAGE QUALITY METRICS

Metric / Image	PSNR	IMMSE	SSIM	MS-SSIM	$FSIM_C$	ICI
Row 1	23.3513	300.575	0.9684	R – 0.7798 G – 0.8370 B – 0.8388	0.9165	0.002249
Row 2	19.4188	743.3572	0.9378	R – 0.6858 G – 0.8055 B – 0.8086	0.9014	0.000532
Row 3	20.1476	628.5186	0.9293	R – 0.6236 G – 0.7413 B – 0.8613	0.8416	0.018471

Table 3. A table illustrating the computed metrics (PSNR, IMMSE, SSIM, MS-SSIM, $FSIM_C$ and ICI) scores for the images depicted in Figure 2. Here, the reference image is the **Expected** (i.e. ground truth) image while the O-Net **Generated** image is the test image (which is being compared against the reference). reference).

Further image analysis using the images & values shown in Fig. 2 & Table 3 depict the best ICI-scoring image was from Column 2, while the worst ICI scores were obtained from

Column 3. This observation, however, doesn't correlate with the other assayed metrics [PSNR, IMMSE, SSIM & $FSIM_C$ indicate the image in Column 1 to attain the best scores, while the worst scores are obtained by the images in Column 2 (PSNR & IMMSE) and Column 3 (SSIM & $FSIM_C$)]. Again, close visual inspection of the **Generated-Expected** image pairs in Fig. 2 (based on the RoIs highlighted within the ellipses in this Figure) concur with the deductions drawn from the ICI scores, portraying the viability of the ICI metric as a quantitative determinant of image similarity across multiple channels. Moreover, comparing the ICI scores of the O-Net **Generated** image for Column 2 (Table 3) against its U-Net counterpart (Table 2) also suggests the O-Net **Generated** image bearing a closer similarity to the **Expected** (ground truth) image than its U-Net counterpart – a finding which closely agrees with a visual inspection of these images in Figs. 1 & 2 (notice the more obvious striations evident in the yellow ellipse-RoI of the O-Net **Generated** image in Fig. 2 as opposed to the U-Net **Generated** image in Fig. 1). Once again, this inference is *not* reflected when comparing the PSNR, IMMSE & SSIM scores for these images (although the $FSIM_C$ scores agree with the ICI values), highlighting the superiority of ICI as a quantitative SR image homology metric.

IV. Potential Limitations & Future Work

Although the proposed ICI metric has been depicted to be highly conducive for elucidating structural homology between 2 computationally SR images, it might *putatively* be prone to inherent noise present within the image – an aspect which has not been thoroughly evaluated in the current study to avoid the introduction of potential confounders in the present work. One possible way to overcome this would be to first pre-process the images using suitable image denoising algorithms (e.g. median filtering), although this might have the consequence of negating some of the fine SR structural details as well. To circumvent this, we may suggest that the user evaluate the error maps and other metrics (such as PSNR & IMMSE) together with the ICI scores to decide if a favourable score for any of these metrics is indeed reflective of the superior image quality, rather than the removal of pseudo-noise features (an aspect which was also surfaced in [14]).

At this juncture, it would also be prudent to highlight prior work done by other researchers in the field of MAE utilization for image analysis, namely [16] where the authors demarcated *image quality analysis/assessment* (IQA) into 2 main categories – (i) *subjective* QA, which (though being highly reliable) is expensive, inefficient and cannot be applied to real-time analytical systems, and (ii) *objective* QA (which may be further classified into full-reference (FR), reduced-reference (RR) and no-reference (NR) QA) [16]. According to [16], the current objective QA approaches (though being replicable and exhibiting significant potential for assimilation in quantitative image analysis) do not conform exactly to subjective QA (which is based on the *human visual system* (HVS)), leading the authors of [16] to propose an alternative approach in their paper where they converted the RGB images to their grayscale equivalents, prior to generating a gradient map ($\boldsymbol{I}_{grd}$) through convolution of the grayscale image with a Sobel operator and subsequently evaluating (& correcting) the impact of potential distortions on the MAE outputs. Nonetheless, [16] highlighted potential areas for improvement of their proposed approach (despite yielding relatively promising results in their paper), namely where the images are blurry, compressed, pixel-quantized, or experience changes in image contrast. Similarly, our proposed ICI index might function well when applied to the context of image super-resolution (as in the current study), but (unlike [16]) would also work well in the context of blurry, quantized, or contrast-variant image pairs. As such, a potential deployment of the proposed ICI metric would lie in its assimilation for cellular identification, forensic pathology, or the identification of genetic / chromosomal mutations (where pixel-wise determination of image discrepancies is essential), albeit necessitating further verification of the ICI's suitability in this context.

V. Conclusion

The presently-proposed image comparison measure (i.e. the image comparative index / ICI) represents a key determinant for evaluating the similarity between 2 images, especially in the field of computational SR. Here, it is demonstrated that ICI scores correspond more closely to a visual comparison of the 2 images when contrasted against other popular metrics (e.g. SSIM or MS-SSIM), implicating the potential use of the ICI as a viable quantitative determinant of image quality and similarity. In this regard, we may postulate that the ICI metric holds significant promise for comparative quantitative image analysis making it well-suited across industries where computational vision is prevalent (e.g. in diagnostics, quality inspection and surveying).

Data Availability

The images & code utilized for this study may be downloaded from https://github.com/shiraz14/ICI-metric.

References

[1] R. Goel, A. Sharma and R. Kapoor, "Object Recognition Using Deep Learning," *Journal of Computational and Theoretical Nanoscience*, vol. 16, pp. 4044-4052, 2019.

[2] J. J. Jeong, A. Tariq, T. Adejumo, H. Trivedi, J. W. Gichoya and I. Banerjee, "Systematic Review of Generative Adversarial Networks (GANs) for Medical Image Classification and Segmentation," *Journal of Digital Imaging*, vol. 35, no. 2, pp. 137-152, 2022.

[3] "Generating Images with BigGAN | TensorFlow Hub," [Online]. Available: https://www.tensorflow.org/hub/tutorials/biggan_generation_with_tf_hub. [Accessed 4 12 2023].

[4] S. Jung and M. Keuper, "Spectral Distribution Aware Image Generation," *arXiv*, p. 2012.03110, 2020.

[5] C-Y. Wang, I-H. Yeh and H-Y. M. Liao, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," *arXiv*, p. 2402.13616, 2024.

[6] K. Zhang, J. Liang, L. van Gool and R. Timofte, "Designing a Practical Degradation Model for Deep Blind Image Super-Resolution," in *International Conference on Computer Vision*, [Online], 2021.

[7] W. Ouyang, A. Aristov, M. Lelek, X. Hao and C. Zimmer, "Deep learning massively accelerates super-resolution localization microscopy," *Nature Biotechnology*, vol. 36, no. 5, pp. 460-468, 2018.

[8] D. Wu, "Applications of Deep Neural Networks in Image Classification," in *2nd International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)*, Nanjing, China, 2022.

[9] A. Tanchenko, "Visual-PSNR measure of image quality," *Journal of Visual Communication and Image Representation*, vol. 25, pp. 874-878, 2014.

[10] Z. Cui, Z. Gan, G. Tang, F. Liu and X. Zhu, "Image Signature Based Mean Square Error for Image Quality Assessment," *Chinese Journal of Electronics*, vol. 24, pp. 755-760, 2015.

[11] G. P. Renieblas, A. T. Nogués, A. M. González, N. Gómez-Leon and E. G. D. Castillo, "Structural similarity index family for image quality assessment in radiological images," *Journal of Medical Imaging*, vol. 4, p. 035501, 2017.

[12] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, vol. 2, pp. 1398-1402, 2003.

[13] L. Zhang, L. Zhang, X. Mou and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378-2386, 2011.

[14] S. S. Kaderuppan, E. W. L. Wong, A. Sharma and W. L. Woo, "O-Net: A Fast and Precise Deep-Learning Architecture for Computational Super-Resolved Phase-Modulated Optical Microscopy," *Microscopy and Microanalysis*, pp. 1-15, 2022.

[15] Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Computer Vision Group, Freiburg, 2015. [Online]. Available: https://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a/. [Accessed 30 Apr 2019].

[16] H. Sihan, and L. Sumei, "A Weighted Mean Absolute Error Metric for Image Quality Assessment," in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, [Online], 2020.