

# No Reference Quality Assessment Metric for Multi-spectral and Multi-Modal Image Fusion using Sparse Approximate Variational Autoencoder

Milind S. Patil<sup>1</sup>, Pradip B. Mane<sup>2</sup>

Submitted: 22/12/2023 Revised: 28/01/2024 Accepted: 08/02/2024

**Abstract:** Unlike natural image quality assessment approaches, satellite stereo images have various quality criteria in different application contexts, making it difficult to develop an appropriate objective evaluation model. The area of perceptual quality evaluation has evolved significantly and continues to expand. In the low-level computer vision field, no reference image quality assessment (NRIQA) is critical. Deep neural networks are gaining popularity for NRIQA applications. Existing deep learning-based systems are generally supervised and depend on an unrealistically huge number of labelled training data. Model-based techniques are unsupervised and flexible, but they depend on handmade priors. The majority of extant No reference image quality assessment (NR-IQA) models were designed for synthetically distorted images, however they perform badly on in-the-wild images, which are frequently used in a variety of practical applications. Blind Image Quality Evaluation Metric for Multi-spectral and Multi-modal Image Fusion Techniques is developed in this research. This No reference quality measure is examined and compared to numerous well-known cutting-edge methods and mean opinion score. The proposed quality evaluation regression models successfully predict quality score. When compared to the MOS score, archives score with 96% similarity. The suggested approach has a Pearson correlation value of 0.96 and a Spearman's rank correlation coefficient of 0.83. To use the abundant self-supervisory information and decrease the model's uncertainty, we impose self-consistency between the outputs of our quality assessment model for each image and its sparse code book. Our results demonstrate that our suggested technique outperforms other methods on Fused image datasets with distorted images.

**Keywords:** Deep Neural Network (DNN), Sparse Approximate Variational Autoencoder, Quality Assessment Regression Model

## 1. Introduction

In recent years, remote sensing satellite technology has been extensively applied in a variety of industries, including land resources, marine resources, agriculture, forestry, water conservation, seismic monitoring, and the environment, resulting in huge economic and social advantages. Unlike natural image quality assessment approaches, satellite stereo images have various quality criteria in different application contexts, making it difficult to develop an appropriate objective evaluation model. To be able to accurately and reliably predict the perceived image quality without having access to the reference image is an essential capability for a number of computer vision applications, as well as for the social media and streaming media industries. Image Quality Assessment (IQA) is concerned with the difficulty of measuring and forecasting human judgements of image quality. IQA was developed by experts in the field. Determining the quality of damaged images without having any knowledge of pristine reference images or the kind of aberrations that are present is the focus of the No-Reference

(NR) or blind International Quality Assurance (IQA) technique. In order to generate robust and trustworthy quality forecasts that align well with subjective evaluations, NR-IQA models are tailored to meet the requirements of the design. Full reference (FR) [1]-[3], reduced reference (RR) [4]-[6], and no reference (NR)/blind [7]-[9] are the three categories that are used to classify international quality assurance (IQA) systems. These categories are determined by the availability of fused images. It is possible to employ FR-IQA methods in order to directly compare the distorted image to its fused image when fused images are available. For the purpose of calculating the visual quality score, just a piece of the fused image has to be calculated when using RR-IQA techniques. When it comes to practical applications, fused images are often unavailable; thus, NR-IQA measurements should be used instead. NR-IQA procedures do not always perform as well as FR measures. This is due to the fact that the visual quality score is assessed only via the use of distorted images, without the utilisation of reference images. The BIQA technique, on the other hand, may be used for a wider range of applications, including but not limited to image and video retargeting, streaming media, and computer graphics. The requirements for computation are often rather low since there is no need to manage references. The consequence of this is that a growing number of scholars are working on creating NR-

<sup>1</sup> Research Scholar, All India Shri Shivaji Memorial Society's Institute of Information Technology, Pune, India. patilmilind78@gmail.com

<sup>1</sup> Assistant Professor, Vishwakarma Institute of Information Technology, Pune, India

<sup>2</sup> Principal, All India Shri Shivaji Memorial Society's Institute of Information Technology, Pune, India. pbmane6829@gmail.com

\* Corresponding Author Email: patilmilind78@gmail.com

IQA methods.

No Reference Quality Assessment (NRQA) is difficult in satellite imaging for various reasons:

- a. **Absence of Ground Truth:** Unlike other image domains, it is sometimes hard to get a high-quality, ground-truth image of the scene taken by the satellite. This makes direct comparisons for determining quality impractical. Imagine attempting to analyse the quality of a satellite image of a distant forest without ever having been there!
- b. **Subjectivity and Context Dependence:** Perception of image quality is extremely subjective and application- and user-specific. What defines a "good" quality image for land cover categorization may vary dramatically from what is required for urban change detection. NRQA approaches fail to capture these distinctions without considering user intent.
- c. **Diverse Image.** Satellite images have a wide variety of features, including spectral resolution, spatial resolution, sensor-specific artefacts, and atmospheric influences. Designing a single NRQA measure that successfully accounts for all of these variances is a substantial problem.
- d. **Limited Training Data:** Developing good NRQA models requires a substantial quantity of labelled data, which may be difficult and costly to get in the satellite images domain. This scarcity reduces the accuracy and generalizability of these models.
- e. **Computational Complexity:** Some NRQA techniques, such as those based on deep learning, are computationally costly, rendering them unsuitable for real-time applications or resource-constrained contexts.

Despite these obstacles, researchers are working to create improved NRQA approaches for satellite images. These techniques examine a variety of tactics, including:

- i. **Leveraging past knowledge** entails incorporating prior knowledge about common image qualities and degradation kinds.
- ii. **Using image statistics:** Analysing statistical aspects of the image to determine quality.
- iii. **Exploring task-specific information:** Adapting the NRQA approach to the unique application or user requirements.
- iv. **Learning from Limited Data:** Developing effective approaches for training NRQA models on smaller datasets.
- v. **While flawless NRQA in satellite imagery remains a research goal,** these developments pave the path

for more robust and dependable quality assessment in this increasingly important sector.

In regard to fact, the majority of the currently available NR-IQA algorithms disregard the intrinsic uncertainty that is present in datasets. Due to the fact that NR-IQA datasets are produced in a subjective manner, there is observational noise that introduces a level of corruption into the target values. Thus, there does not exist an accurate mapping  $y \sim F(x)$  between label  $y$  and data  $x$ .

Revised mapping form can be interpreted as,  $y = F(x) + noise(x)$ .

When all of these factors are taken into account, it is necessary to do a statistical modelling of the observational noise ( $x$ ). In order to do this, we include uncertainty learning into the NR-IQA model and propose the Sparse approximate variational autoencoder (SA-VAE) regression model. Attempting to quantify the noise that is already present in a model or dataset is the primary emphasis of the study of uncertainty. The epistemic uncertainty and the heteroscedastic aleatoric uncertainty are the two main uncertainties that should be of concern to you. The epistemic uncertainty is caused by the noise that is present in the model parameters or the model outputs. Aleatoric uncertainty is inherently present within the dataset itself. A significant number of people make an effort to include uncertainty into models in the hope of achieving better results. [10] and [11]. In order to examine uncertainty, some people are considering the possibility of building a generic learning paradigm. In their work [12, 13], Geng and colleagues make an effort to represent an instance by using a particular distribution rather than one or more labels. The authors Pate et al. [21] use a risk level technique to evaluate the amount of uncertainty. In recent times, there has been a significant amount of focus placed on uncertainty analysis on neural networks [14-18]. Increasing the robustness and effectiveness of models is accomplished by the use of uncertainty analysis in a variety of tasks, including semantic segmentation [17], face recognition [18], and item identification [19]-[20]. The primary objective of VAEs is to acquire the knowledge necessary to learn a mapping between high-dimensional observations and a representation space with fewer dimensions. This mapping should be such that the original observations may be approximated using the representation with reduced dimensions.

Sparse coding-based variational autoencoder (SA-VAE) is an acronym for sparse approximation variational autoencoder, which is the name of the model that has been proposed. The SA-VAE has a number of advantages over the work that came before it. To begin, it is possible to teach it from beginning to finish, and it does not have any problems with posterior or codebook collapse. In the second place, it makes it possible for us to generate a mapping of

the input image and quality index that is more precise. In conclusion, the latent sparse codes make it possible for us to do regression value via the use of image patch grouping.

The following is a summary of the contributions that this work has made:

1. The strategy that has been suggested IF distortions or artefacts should be analysed, and parametric and non-parametric characteristics should be constructed depending on the identified distortion.
2. Our company offers a deep learning approach that covers the whole process, from beginning to finish, for quality assessment regression models that predict quality score.
3. We have proposed a variational autoencoder that is based on a Sparse Approximated Code for the purpose of quality prediction, which will then be followed by a Quality Assessment Regression Model. During the training process, we make advantage of the rich self-supervisory information by using self-consistency between the output for each image and its sparse codebook. This allows us to reduce the network's sensitivity.
4. Extensive testing on fused image datasets (for distortions) demonstrates that our proposed technique yields positive results across a wide range of datasets.

## 2. Literature Survey

Existing IQA algorithms that are based on deep learning are mostly dependent on subjective human evaluations (MOS/DMOS), and they portray the quality prediction issue as a job that involves either regression or classification. As a consequence of this, the models are unable to make direct use of the relative arrangements of the images. The limited availability of large labelled IQA datasets is one of the challenges that must be overcome in order to construct CNN-based IQA models. The process of annotating IQA datasets is one that is both time-consuming and expensive. A self-supervised collaborative autoencoder is developed by Z. Zhou and colleagues [21] in order to represent the information regarding the content and the distortion in a separate manner. Subsequently, a self-adaptive weighting-based quality predictor is developed in order to achieve a balance between the individual representations of the content and the distortions through the process of image quality prediction. Attention processes are often used in the activities that are associated with computer vision [22–24]. An end-to-end saliency-guided architecture that incorporates spatial and transposed attention was introduced by Yang et al. [25] in the context of NR-IQA. According to X. Ma et al. [26], For the purpose of marking the distorted image, a large number of FR-IQA measures were used as an

alternative to subjective quality annotation. Because there is no clear image, a deep neural network is trained to make predictions about numerous FR-IQA scores without any information being provided. The final quality predicting score is obtained by combining the predictions of a number of different FR-IQA scores. This is accomplished via a self-supervised FR-IQA score aggregator that is based on an adversarial auto-encoder score. Using deconstructed large-kernel convolutions, L. Yu et al. [27] presented a lightweight attention technique that extracts multiscale features. Additionally, they presented a one-of-a-kind feature improvement module that predicts No-Reference Image Quality Assessment. Both of these approaches are described in the article. A hierarchical no-reference Stereoscopic Image Quality Assessment approach was given by J. Si et al. [18]. This method takes into consideration binocular interaction and binocular fusion, and it also incorporates automatic weight learning. Y. Zhu and colleagues [29] proposed a model that includes a self-supervised feature learning approach that is needed in order to relieve the shortage of data and learn complete feature representations. Additionally, the model includes a self-attention-based feature fusion module in order to include a self-attention mechanism. Through the use of stacked autoencoders, J. Yang et al. [30] presented a blind quality evaluation measure that was both effective and novel. This measure was based on graphical and textual regions. The characteristics that have been created from these two domains, in addition to their subjective evaluations, are input into two regressors for the purpose of training. It is only possible for each regressor to deliver a single projected score. In conclusion, a weighted model is used in order to provide the ultimate perceptual quality score of a test SCI. This score is derived from the two expected values provided. The No-reference image quality model is developed by K. Ding and colleagues [31], who also include a set tolerance for texture resampling. Using a convolutional neural network, we construct an injective and differentiable function that is capable of converting images into multi-scale over complete representations. Full-reference image quality assessment measures are provided by Varga, D. et al. [32]. These measures characterise the global changes that occur in an image as a result of Grunwald-Letnikov derivatives and the local changes that occur as a result of image gradients. Additionally, visual saliency is used for the purpose of weighing changes in images and highlighting areas of the image that are significant to the human visual system. K. Lamichhane and colleagues [33] proposed an objective quality metric that involves the use of deep neural networks. The human visual system is taken into account by the measure, which does so by calculating the saliency map and natural scene statistical features of the image that is being examined. The convolutional layers and the regression units are the two components that make up the neural network. A two-stream CNN-based no-reference LF

image quality assessment (LF-IQA) technique is presented by S. Alamgeer and colleagues under the reference number 34. Rich distortion-related spatial and angular binocular characteristics are extracted using the two-stream CNN in order to achieve the purpose of evaluating the quality of the LF contents.

### 3. Methodology

The Encoder layer is responsible for compressing the input image into a representation of the same latent space. The input fused image is converted into a compressed representation that has a decreased dimension using this process. The size of the code, also known as the bottleneck, is the most critical hyperparameter to consider when customising the autoencoder. The amount of data that has to be compressed is determined by it. In addition to that, it may be used as a regularisation term. When making adjustments to autoencoders, it is important to bear in mind that the number of layers is an important factor. The complexity of the model rises with increasing depth, whereas with decreasing depth, the analysis process is sped up. It is often a challenging endeavour to choose nodes for each layer of their structure. The number of nodes in the autoencoder reduces with each consecutive layer because the input to each layer becomes less as it progresses through the levels. In order to manage sparse autoencoders, the number of nodes that are present in each hidden layer is used. Because it is challenging to build a neural network with a variable number of nodes in its hidden layers, sparse autoencoders operate by penalising the activation of certain neurons in the hidden levels. This is done in order to accommodate the difficulty of the construction process. In other words, it suggests that the loss function is subjected to a penalty that is precisely proportional to the number of neurons that are activated.

For the purpose of regularising the neural network, the sparsity function prevents extra neurons from being involved in the process. Regularizers may be broken down into two categories:

1. Increasing the size of the model is possible by the use of the L1 Loss approach, which is a general regularizer to utilise.
2. In contrast to the L1 Loss methodology, the KL-divergence method analyses activations across several samples simultaneously, as opposed to just adding up all of the activations. It is our intention to restrict the average activity of each neuron throughout this collection.

Variational autoencoders, often known as VAEs, are models that use autoencoders that are more conventional in order to address a specific problem. An autoencoder is taught to learn to represent the input solely in a compressed form that

is referred to as the latent space or the bottleneck. This is accomplished via the training process. On the other hand, the latent space that is generated as a result of training is not necessarily continuous, and as a consequence, it may be challenging to interpolate. Variational autoencoders are concerned with this specific problem and express their latent characteristics as a probability distribution, resulting in a continuous latent space that can be sampled and interpolated easily. An strategy that is well-known for learning sparse codes, which involves iterating the recursive equation that is shown below until it converges. The l1-norm regularised least square problem of sparse codebook is addressed by this research via the use of an iterative sparse approximation that incorporates a one-of-a-kind pre-conditioner system.

Sparse approximation may be used to generate image quality indices, which has various benefits over conventional techniques. This methodology would generate Sparsed code  $x \in \mathbb{R}^n$  from its noisy measurements. General sparse approximation can be estimated as,

$$\begin{aligned} b &= Ax + n \\ &\in \mathbb{R}^M \end{aligned} \quad (1)$$

Where  $A \in \mathbb{R}^{M \times N}$  and  $n \in \mathbb{R}^M$  is the environmental noise.

The traditional Least Squares (LS) approach necessitates a large number of observations.

$M \geq N$  and  $A$  has full rank  $N$  to recover  $x = (A^T A)^{-1} A^T b$ . Current Compressed Sensing (CS) techniques could recreate  $x$  from a much smaller number of observations  $M \leq N$ . As long as the input image appears sparse, the aforementioned basis pursuits speech denoising dilemma can be solved:

$$\min_{x \in \mathbb{R}^N} \frac{1}{2} \|Ax - b\|^2 + \tau \|x\|_1 \quad (2)$$

Where  $\tau > 0$  is a specified normalization coefficient,  $\|x\| = \sqrt{\sum_{i=1}^N x_i^2}$  and  $\|x\|_1 = \sqrt{\sum_{i=1}^N |x_i|}$  denote the  $l_2$  and the  $l_1$  norms of  $x$ , respectively,

Including the constraint clearly defines the solution key space in (6) the pseudo recovery, proposed methodology would use a simple optimization strategy to get them out.

$$\hat{x} = [x^+; x^-] \in \mathbb{R}^{2N} \geq 0 \text{ and } \hat{A} = [A, -A] \in \mathbb{R}^{M \times 2N} \quad (3)$$

where,  $x_i^+ = \max(x_i, 0)$  and  $x_i^- = \max(-x_i, 0)$ , the  $Ax = \hat{A}\hat{x}$  and,  $\|x\|_1 = \|\hat{x}\|_1$  and hence Eq.) can be solved with respect to  $\hat{A}$  and  $\hat{x} \geq 0$ .

As a result, proposed method just need to acknowledge the version of Eq. shown below for  $x \geq 0$

$$\min_{x \in \mathbb{R}^N} \frac{1}{2} \|Ax - b\|^2 + \tau e^\tau x \quad (4)$$

$$stx \geq 0$$

Since in Eq.(4) an optimization problem is convex with nothing but linear constraints that fulfills Slater's condition, then this can discover optimized solution by addressing its Karush-Kuhn-Tucker (KKT) system:

$$A^T Ax - s - A^T b + \tau e = 0 \quad (5a)$$

$$X \cdot Se = 0 \quad (5b)$$

$$(x, s) \geq 0 \quad (5c)$$

Where

$$X = \text{Diag}(x) \text{ and } S = \text{Diag}(s)$$

The above equation indicates diagonal matrices consist of primal coefficient  $x$  and dual coefficient value  $x$  and dual coefficient  $s$ , respectively, and 0 and  $e$  indicate an entirely zero or all one array whose size should be apparent from reference, respectively. In the multi-dimensional manner, the modules cooperatively increase the interaction among different regions of images globally and locally. The Inverse Sparse Approximation (ISA) addresses a transformed Karush-Kuhn-Tucker method by merely substituting Eq. (5b) for Eq. (5b) in the basic Karush-Kuhn-Tucker framework.

$$Xse = \sigma \mu e \quad (6)$$

Where  $\mu = x^T s / N$  goes to 0, Whenever it converges, it returns to zero and  $\sigma \in [0,1]$  is a centeredness element. A  $\sigma$  closer to 1 will prompt search results further towards the interior  $(x, s) > 0$ . Moving from a specific point  $(x, s)$ , the novel Karush-Kuhn-Tucker system's orientation could be calculated as

$$A^T A \Delta x - \Delta s = r_d \quad (7a)$$

$$S \Delta x + X \Delta s = r_e \quad (7b)$$

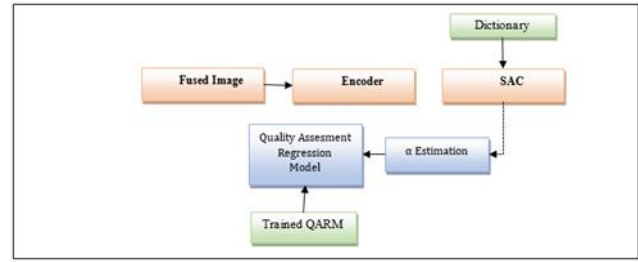
Where  $r_d$  indicated the stationary residual and complementary slackness residual  $r_e$  can be expressed by

$$r_d = s - \nabla h(x) \quad (8a)$$

$$r_e = \sigma \mu e - XSe \quad (8b)$$

Here,  $\nabla h(x) = A^T Ax - A^T b + \tau e$  is the gradient of the objective function.

$$h(x): \frac{1}{2} \|Ax - b\|^2 + \tau e^T x \quad (9)$$



**Fig.1.** Proposed SAC based NR-IQ Estimation

In Algorithm 1, the proposed method represents Sparse Approximated Code (SAC) with predictor-corrector steps, which employs the SAC framework. It can be widely regarded as one of the most effective of the different sparse approximation. To ensure quicker convergence, the proposed SAC uses different initializations, which have simplified but appropriate coefficients, a new preconditioner, and adaptive tolerance. Although Eq. (5a) must be fulfilled at all times, the proposed SAC allows quite versatile  $x, s$  that violate Eq. (5a) during initial setup and subsequent iterations, that only need Eq. (5a) to be fulfilled at convergence.

Encoded feature vector is used as a input of the SAC network:  $X = [x_1, x_2, \dots, x_k]$  where  $k$  is the number of the encoded coefficient per computation. In order to accelerate the convergence of the framework, the Min-Max Normalization is suggested, through which the variables for each data dimension are sequentially reshaped and normalised to  $[0, 1]$  range:

$x = \left[ \frac{x - \min}{\max - \min} \right]$  Min is the minimum at each column, and max is the maximum at each column. Then data by restructuring 1 by 1024 deep attributes into 32 by 2 matrix form before convolution is generated. The output of the  $j$ th feature map on the  $i$ th unit of the 1 convolution layer is:

$$x_i^{l,j} = \sigma [b_j + \sum_{a=1}^m w_a^j x_{i+a-1}^{l-1,j}] \quad (10)$$

$b_j$  is the bias term for  $j$ th feature map,  $m$  is the kernel size,  $j$  a  $w$  is the weight of  $j$ th feature map and  $a$ th filter index and  $\sigma$  is the activation function.

### 3.1. Training of VAE:

Initially, satellite-fused images are obtained. The whole dataset is preprocessed and cleaned adequately for the quality estimate job. The data is then divided into three sets: training, validation, and test. The training set trains the VAE, the validation set monitors training progress and adjusts hyperparameters, and the test set evaluates the final model's performance. The VAE's encoder converts the input data to a lower-dimensional latent space representation. The encoder is made up of many neural network layers that extract characteristics from the data. The proposed model included a probabilistic aspect by expressing the latent space as a distribution rather than a single point. This is

commonly accomplished by employing a normal distribution with the Sparse Approximated Code anticipated by the encoder. Reconstruction loss is usually assessed using a mean squared error or similar distance metric. KL divergence loss pushes the latent space distribution to resemble a standard normal distribution. This helps to keep the model consistent and prevents overfitting. A VAE's performance may be influenced by a variety of hyperparameters, including the learning rate, the number of encoder and decoder layers, and the latent space dimensions.

Algorithm: 1 Sparse Approximated Code (SAC) Framework

**Inputs:**  $\epsilon$  : Choose  $(x^0, s^0) > 0$  stop accuracy  $\epsilon$  (e.g.  $1e-6$ ),

Total Epochs is  $k_{max}$ .

for  $k = 1, 2, \dots, k_{max}$  do

    Perform Prediction Step : set  $\sigma \leftarrow 0.001$

$(x^k, s^k, \alpha_p, \alpha_d) = UPDATE(x^{k-1}, s^{k-1}, \sigma)$

    if  $\mu_k \leq \epsilon h(x^k)$  and  $\|r_d^k\|$  then

        e Break

**Outputs:**  $x^k$

**Function:**  $UPDATE(x^{k-1}, s^{k-1}, \sigma)$

    Compute  $\Delta x, \Delta s$  with  $\sigma, x^{k-1}, s^{k-1}$

    Compute  $\alpha_p, \alpha_d$  with  $x^{k-1}, s^{k-1}, \Delta x, \Delta s$

Update  $(x^k, s^k) \leftarrow (x^{k-1} + \alpha_p \Delta x, s^{k-1} + \alpha_d \Delta s)$

Return  $(x^k, s^k, \alpha_p, \alpha_d)$

Overall, Sparse Approximated Code (SAC) Framework is a promising strategy for assessing image quality since it is flexible, adaptable, and has the potential for quick evaluation.

- Captures perceptual aspects: Considers how effectively the image may be portrayed in accordance with human perception.
- Handles various distortions: Adaptable to a variety of distortions, including compression artefacts, noise, and blur.

- Potentially reduced complexity: Depending on the vocabulary and metrics used, this may be computationally efficient.

## 4. Experiment & Results

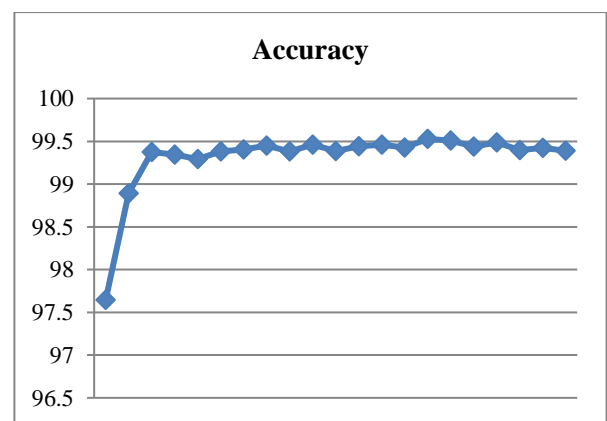
We carried out the subjective experiment for subjective quality score. All the pansharpened images were displayed on a 55 inch MI TV. We adapted double stimulus continuous quality scale test methodology in the experiment, in which the reference and test images are simultaneously presented (displayed side-by-side) on the screen. Based on the properties of fused images, grading criteria in our subjective study depend on two factors: distortions or artifacts.

### 4.1. Time Complexity:

The cumulative amount of time required by the proposed system across all convolution layers

$$O(\sum_{l=1}^d n_{l-1} \cdot s_1^2 \cdot n_1 \cdot m_1^2) \quad (11)$$

The convolutional layer index is set to 1, and the total number of convolutional layers is denoted by the letter d. The total number of filters that are used in the lth layer is denoted by the letter n, which is also referred to as the number of input channels in the lth layer. Despite the fact that the filter has a spatial size of s, the feature map that is produced has a spatial size of m. However, the time cost associated with completely connected layers and pooling layers is always between 5 and 10 percent of the total computational time, which is not included in the composition that was described before. An example of the training performance of the recommended model is provided below in the form of a figure. At first, the performance of the model improves with each epoch, but later it remains same.



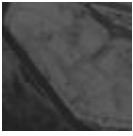
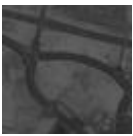


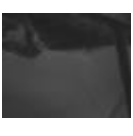

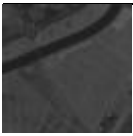

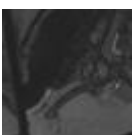
**Fig. 2.** Training Accuracy of proposed model vs epochs

Unlike regular autoencoders, which are simply concerned with reconstruction, VAEs additionally aim for a latent space distribution that is near to the standard normal. This creates complexity, which may have varying effects on training accuracy. Depending on the setting, "training



accuracy" is refer to a weighted mix of reconstruction and KL divergence. For the purpose of illustrating the influence that many measurements have on selecting the most optimal fusion results, we provide an example.

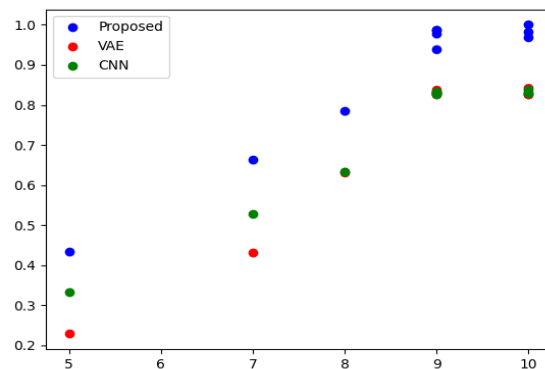
**Table 1.** Compassion of Proposed method NR-IQA

Fused Image	MOS	Proposed IR-IQA	Self-Adaptive Weighting based VAE [21]	CNN [34]
	9	0.987858	0.83023	0.829046
	9	0.987321	0.826752	0.828684
	10	1.00	0.826977	0.82854
	10	0.96918	0.826258	0.831533
	9	0.939202	0.837491	0.833493
	9	0.977617	0.830724	0.826063
	8	0.785954	0.632175	0.632891
	5	0.433124	0.22857	0.333841
	7	0.662485	0.432456	0.528965

In order to conduct an objective evaluation of the performance of our strategy, we compare it to the two ways that we selected. Taking into consideration the findings, one might reach the following conclusions:

1) Current comparable measures have relatively poor assessment performances when compared to MOS metrics. This is due to the fact that spectrum distortion is not reflected into conventional metrics, which results in an inaccurate portrayal of the distortions.

2) When compared to three training-based M1 measures, our method performs better than any of them combined. As a result of the fact that these measurements are very dependent on the characteristics and training samples that were used in their training, our method is more robust than the training-based metrics that are currently in use. The strategy that we choose is the most efficient. The possible reason for this may be that the methods of feature extraction that are used in these measurements are not suitable for the circumstances that we are now facing. All things considered, our approach takes into consideration the effects of spectrum distortion, spatial distortion, and affects on information indexes, which ultimately leads to a distribution that is congruent with subjective results. Developing a universal image quality matrix is difficult owing to the subjective and application-specific nature of quality perception. However, VAEs may be excellent tools for providing significant insights and features for developing successful image quality evaluation systems.

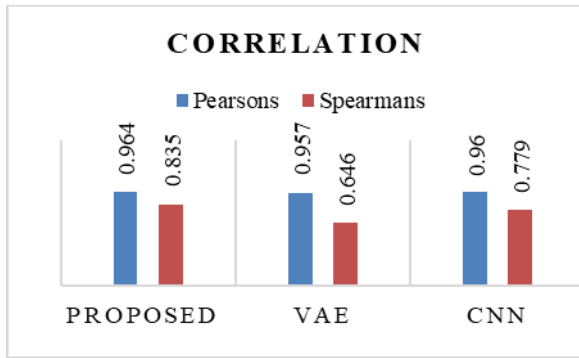


**Fig. 3.** Correlation of Proposed Method with Existing

Figure 3 shows correlation graph of proposed method with existing Method. Proposed method have better close to MOS score than VAE and CNN

**Table 2.** MoS Correlation Score

	Proposed	VAE	CNN
Pearsons	0.964	0.957	0.96
Spearmans	0.835	0.646	0.779



**Fig. 4.** Correlation of MoS Score

It is a statistical correlation coefficient that examines the linear connection between Mos and Predicted Score. The Pearson coefficient of correlation is a common example of such a coefficient. For all intents and purposes, it is only a normalised measurement of covariance, with the result always falling somewhere between 0. The product of the covariances of two variables and the standard deviations of those variables is what it is called. In terms of Pearson Correlation, the suggested method performs better than both the VAE and CNN-based Quality ratings. The correlation coefficient of Spearman's rank is an example of a statistic. Pearson's correlation analyses linear relationships, while Spearman's correlation investigates monotonic connections, regardless of whether or not they are linear. The Quality Score that has been suggested correlates linearly with the Quality Score that is based on MoS.

## 5. Conclusion

The results of this research provide a method for evaluating the quality of fused images that does not need references. For the purpose of evaluating distortions, we use a variational autoencoder that is based on a Sparse Approximated Code, which is then followed by a Quality Assessment Regression Model. Through the extraction of quality-aware characteristics from both the spatial domain and the spatial-temporal domain, the proposed model is able to evaluate levels of distortion. In order for the proposed model to be able to completely leverage the distortions and texturing that are available in the current database, we train the spatial feature extractor from the very beginning to the very finishing. In spite of the fact that our approach is successful, there is always room for improvement. Some of the distortion is not as irritating as others. As a consequence of this, it is essential for future development to recognise certain types of distortion, such as appealing unreal texture, unpleasing unreal texture, and so on.

## References

[1] H. R. Sheikh, M. F. Sabir and A. C. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms," in *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp.

3440-3451, Nov. 2006, doi: 10.1109/TIP.2006.881959.

- [2] S. Bosse, D. Maniry, K. -R. Müller, T. Wiegand and W. Samek, "Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment," in *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206-219, Jan. 2018, doi: 10.1109/TIP.2017.2760518.
- [3] W. Sun, Q. Liao, J. -H. Xue and F. Zhou, "SPSIM: A Superpixel-Based Similarity Index for Full-Reference Image Quality Assessment," in *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4232-4244, Sept. 2018, doi: 10.1109/TIP.2018.2837341.
- [4] A. Rehman and Z. Wang, "Reduced-Reference Image Quality Assessment by Structural Similarity Estimation," in *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3378-3389, Aug. 2012, doi: 10.1109/TIP.2012.2197011.
- [5] Z. Wang and A. C. Bovik, "Reduced- and No-Reference Image Quality Assessment," in *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 29-40, Nov. 2011, doi: 10.1109/MSP.2011.942471.
- [6] J. Wu, W. Lin, G. Shi and A. Liu, "Reduced-Reference Image Quality Assessment With Visual Information Fidelity," in *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1700-1705, Nov. 2013, doi: 10.1109/TMM.2013.2266093.
- [7] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," in *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695-4708, Dec. 2012, doi: 10.1109/TIP.2012.2214050.
- [8] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen and J. Häkkinen, "CID2013: A Database for Evaluating No-Reference Image Quality Assessment Algorithms," in *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 390-402, Jan. 2015, doi: 10.1109/TIP.2014.2378061.
- [9] Z.M. Parvez Sazzad, Y. Kawayoke, Y. Horita, No reference image quality assessment for JPEG2000 based on spatial features, *Signal Processing: Image Communication*, Volume 23, Issue 4, 2008, Pages 257-268, ISSN 0923-5965, <https://doi.org/10.1016/j.image.2008.03.005>.
- [10] A. Der Kiureghian and O. Ditlevsen, "Aleatory or epistemic? does it matter?" *Structural safety*, vol. 31, no. 2, pp. 105-112, 2009.
- [11] M. H. Faber, "On the treatment of uncertainties and probabilities in engineering decision analysis," 2005.
- [12] X. Geng, "Label distribution learning," *IEEE Trans.*



- Knowl. Data Eng., vol. 28, no. 7, pp. 1734–1748, 2016.
- [13] X. Geng and L. Luo, “Multilabel ranking with inconsistent rankers,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2014, pp. 3742–3747.
- [14] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, “Weight uncertainty in neural network,” pp. 1613–1622, 2015.
- [15] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” pp. 1050–1059, 2016.
- [16] A. Kendall and Y. Gal, “What uncertainties do we need in bayesian deep learning for computer vision?” vol. 30, 2017.
- [17] A. Kendall, V. Badrinarayanan, and R. Cipolla, “Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding,” arXiv preprint arXiv:1511.02680, 2015.
- [18] J. Chang, Z. Lan, C. Cheng, and Y. Wei, “Data uncertainty learning in face recognition,” pp. 5710–5719, 2020.
- [19] F. Kraus and K. Dietmayer, “Uncertainty estimation in one-stage object detection,” in IEEE Trans. Intell. Transp. Syst. Conf. (ITSC). IEEE, 2019, pp. 53–60.
- [20] T. Yu, D. Li, Y. Yang, T. M. Hospedales, and T. Xiang, “Robust person re-identification by modelling feature uncertainty,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 552–561.
- [21] Z. Zhou, F. Zhou and G. Qiu, "Blind Image Quality Assessment based on Separate Representations and Adaptive Interaction of Content and Distortion," in IEEE Transactions on Circuits and Systems for Video Technology, doi: 10.1109/TCSVT.2023.3299328.
- [22] Mingdeng Cao, Yanbo Fan, Yong Zhang, Jue Wang, and Yujiu Yang. Vdtr: Video deblurring with transformer. arXiv preprint arXiv:2204.08023, 2022.
- [23] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In Proc. of ICCV, 2017.
- [24] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang.
- [25] Restormer: Efficient transformer for high-resolution image restoration. arXiv preprint arXiv:2111.09881, 2021
- [26] Sheng Yang, Qiuping Jiang, Weisi Lin, and Yongtao Wang.: An end-to-end saliency-guided deep neural network for no-reference image quality assessment. In Proc. of ACM MM, 2019.
- [27] X. Ma, S. Zhang, C. Liu and D. Yu, "Bridge the gap between full-reference and no-reference: A totally full-reference induced blind image quality assessment via deep neural networks," in China Communications, vol. 20, no. 6, pp. 215-228, June 2023, doi: 10.23919/JCC.2023.00.023.
- [28] L. Yu, J. Li, F. Pakdaman, M. Ling and M. Gabbouj, "MAMIQA: No-Reference Image Quality Assessment Based on Multiscale Attention Mechanism With Natural Scene Statistics," in IEEE Signal Processing Letters, vol. 30, pp. 588-592, 2023, doi: 10.1109/LSP.2023.3276645.
- [29] J. Si, B. Huang, H. Yang, W. Lin and Z. Pan, "A no-Reference Stereoscopic Image Quality Assessment Network Based on Binocular Interaction and Fusion Mechanisms," in IEEE Transactions on Image Processing, vol. 31, pp. 3066-3080, 2022, doi: 10.1109/TIP.2022.3164537.
- [30] Y. Zhu, Y. Li, W. Sun, X. Min, G. Zhai and X. Yang, "Blind Image Quality Assessment Via Cross-View Consistency," in IEEE Transactions on Multimedia, doi: 10.1109/TMM.2022.3224319.
- [31] J. Yang et al., "No Reference Quality Assessment for Screen Content Images Using Stacked Autoencoders in Pictorial and Textual Regions," in IEEE Transactions on Cybernetics, vol. 52, no. 5, pp. 2798-2810, May 2022, doi: 10.1109/TCYB.2020.3024627.
- [32] K. Ding, K. Ma, S. Wang and E. P. Simoncelli, "Image Quality Assessment: Unifying Structure and Texture Similarity," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 5, pp. 2567-2581, 1 May 2022, doi: 10.1109/TPAMI.2020.3045810.
- [33] Varga, D. “Full-Reference Image Quality Assessment Based on Grünwald–Letnikov Derivative, Image Gradients, and Visual Saliency.” Electronics 2022, 11, 559. <https://doi.org/10.3390/electronics11040559>
- [34] Kamal Lamichhane, Marco Carli, Federica Battisti, “A CNN-based no reference image quality metric exploiting content saliency”, Signal Processing: Image Communication, Volume 111, 2023, 116899, ISSN 0923-5965, <https://doi.org/10.1016/j.image.2022.116899>.
- [35] Alamgeer, S., Farias, M.C. “A two-stream cnn based visual quality assessment method for light field images.” Multimed Tools Appl 82, 5743–5762 (2023). <https://doi.org/10.1007/s11042-022-13436-4>