Gamma Knife: Is my Patient's MRI Scan too Blurry to be used for Gamma Knife Radiosurgery?

1st Alexandra Ferentinos
School of Data Science
University of Virginia
Charlottesville, Virginia
kzk8qq@virginia.edu

2nd Samantha Johnson
School of Data Science
University of Virginia
Charlottesville, Virginia
nnx3ue@virginia.edu

3rd Peter Landers
School of Data Science
University of Virginia
Charlottesville, Virginia
bjs6pj@virginia.edu

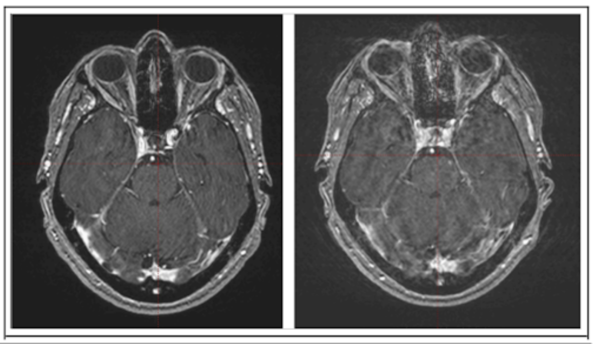

Figure 1: Typical MRI scans for patients who will undergo Gamma Knife radiosurgery. Left figure: A less-blurry MRI. Right figure: A more blurry MRI. The main research question is how to tell if either or both of these images are too blurry to use for GK SRS treatment planning.

Fig. 1. MRI Image Slices

***Abstract*—Gamma Knife Stereotactic Radiosurgery (GK SRS) is a technique where patients are treated with high doses of radiation to very small targets. Because of the small targets and high doses of radiation, accuracy and precision in targeting is critical to achieving the desired clinical result and avoiding unwanted complications. Ideally, the imaging taken on patients would be perfect. However in the real world the quality of the images can vary. One common problem is patient motion, voluntary or involuntary, which can cause artifacts in the resulting images (Figure 1). Currently, evaluation of whether an image is of sufficient quality to use is mainly a subjective determination by the treatment team, generally after an imaging session has closed. A tool which could make an objective determination if an image is too degraded to use would be a practically useful tool, as sending patients for repeat scans is costly and often results in unwanted delays to what are critical treatments.**

I. Introduction

Building on the critical need for precision in Gamma Knife Stereotactic Radiosurgery (GK SRS) described in the abstract, this project addresses the challenge of objectively determining MRI image quality suitable for treatment planning. The goal of this research is to develop a machine learning system that analyzes MRI DICOM files of human brain scans to quantitatively assess image quality, focusing specifically on blur detection that may result from patient movement or equipment factors. This system aims to serve as a decision support tool for clinicians, providing standardized metrics that can be used to determine if an image meets the quality threshold required for accurate GK SRS treatment planning. The research questions focus on two primary objectives: (1) identifying and exploring comprehensive metrics to quantify MRI sharpness that correlate with clinical usability standards, and (2) developing an accurate classification model to categorize MRI scans as acceptable or too blurry for GK SRS planning, despite limited availability of training data. This work serves as a proof-of-concept that could ultimately reduce treatment delays and eliminate unnecessary repeat scans, improving both operational efficiency and patient care in radiosurgery settings, but would require fine-tuning before implementation.

II. Literature Review

A. *Image Quality in SRS*

Gamma Knife Stereotactic Radiosurgery (GK SRS) requires high spatial precision due to the small target volumes and high radiation doses involved. Fatemi et al. [1] highlighted the importance of correcting geometric distortions induced by gradient nonlinearity using 3D distortion correction methods. They found that acceptable distortion should be within 1 mm in the axial plane and up to 2 mm along the z-direction at 10 cm from the isocenter.

B. *Quantitative MRI Quality Metrics*

Vecchiato et al. [2] demonstrated that objective metrics such as gradient entropy and white matter homogeneity correlate well with radiologist assessments when evaluating motion artifacts. These findings support the feasibility of building automated systems that assess image quality in a way that reflects clinical judgment.

C. *Deep Learning for Image Assessment*

Koch et al. [3] used Siamese neural networks as a method to generalize learning to new data distributions even with limited training data. Their findings are particularly relevant for this study due to the small dataset available for training.

D. Clinical Motivations

Trigeminal neuralgia (TN) requires precise diagnosis and treatment planning. Allam et al. [4] highlight that TN is frequently mis-diagnosed despite clear diagnostic criteria due to variable symptom presentation. Patient movement during MRI acquisition can obscure critical anatomical details essential for accurate diagnosis and treatment planning. This motion-induced blurring challenges clinicians interpreting images, potentially leading to suboptimal treatment decisions. An objective, automated tool for quantifying image blur before Gamma Knife stereotactic radiosurgery would address this clinical need, improving diagnostic accuracy, reducing treatment planning time, and enhancing patient outcomes by ensuring high-quality imaging for surgical planning.

III. Data Description

Due to issues with the Institution Review Board (IRB) approval, access to raw data was given on February 21, 2025 and at the end of the project, the team only had access to 60 MRI scans. Each MRI scan was labeled by the team's sponsor as 0 for sharp, 1 for average, or 2 for blurry. Each MRI scan was broken down into a series of DICOM files, each representing a 2D slice of the full 3D volume. These DICOMs were already processed by software from the machine manufacturer in the form of raw k-space data undergoing transforms and potential planar adjustments/shifts/rotations, as well as some level of blurriness corrections as a result of hardware quality. The images' metadata supported this assumption by showing signs of further processing. As a result, we did not seek to further correct the image quality but provide a reasonable assessment of the current state. Further, it is assumed the blurriness under question likely arises from either shifts in movement from the patient (breathing, heartbeat, general-bodily fluid flow), or is some Gaussian noise caused by the imaging machinery).

IV. Methodology

A. Processing and Augmentation

In terms of image quality assessment methods, we identified a handful that we were interested in testing. Some methods for calculating a blurriness score without using a reference image were variance of Laplacian, Sobel Gradient, Tenegrad, and Modified Laplacian. The variance of the Laplacian is fairly simple; measuring the quantity of edges in an image by looking at the spread of areas where pixel intensity changes rapidly. However, this method tends to be sensitive to noise [5]. Sobel Gradient is also fairly simple–computing how much pixel brightness changes both horizontally and vertically but can also be sensitive to noise [6]. Tenegrad is similar to Sobel Gradient but is more sensitive to small details, so it performs better on images with finer qualities [7].

Additionally, we identified two methods that require a reference image: Peak Signal-to-Noise Ratio and Structural Similarity Index Measure. These methods compare the pixels of the image of interest to a reference image to determine if the image of interest is blurrier or clearer than the reference image [8]. The main limitation of these methods is having a good reliable reference image.

Given the limited sample size of our data we employed a variety of data augmentation techniques to enhance our dataset. For each image, we performed a data augmentation task and then saved the augmented image, doubling the size of the initial dataset. Some of the data augmentation techniques used include performing random horizontal rotations, random vertical rotations, random changes of less than 0.01 to the amount of brightness/contrast in the image, and introducing Gaussian blur to some images. These data augmentation techniques substantially increased the size of our dataset and helped improve initial model results.

Early on, we began to notice that processing the MRI images was eating up a lot of memory on Ivy, our virtual machine. Therefore, we employed a few different data processing techniques to help reduce the amount of compute required to load our images and train our models. After loading the MRI images, we cropped each one to ensure that we were removing unnecessary blank space in the image and therefore only focusing on the area of importance, the brain. Additionally, we refined our data pipeline so that we loaded in the MRI images one time, performed all of our data processing and augmentation, and then saved the images to a tensor flow dataset on the virtual machine. This allowed us to save unnecessary process repetition and memory consumption.

B. Support Vector Machine

The team considered simpler models starting with a Support Vector Machine (SVM) as those generally perform well with less data, but tend to fall off performance-wise compared to Convolutional Neural Networks (CNN) on larger datasets. With this in mind, the team classified MRIs as a binary blurry or not blurry. The independent statistics used were the Mean Sobel Gradient (MSG), Tenengrad score (TS), Laplacian, and eventually a modified Laplacian. The MSG the magnitude pixel gradients in both x- and y-directions while the TS is basically the square of those gradients. The goal of the SVM turned into a simple, binary classification task–whether an image was blurry (0) or sharp (1). The testing set for the model would only include unaltered images, and augmented versions of the test images were not considered for training.

Each image was passed through a processing pipeline as previously described while having normalized values between 0 and 1 and was outputted in CSV format for model ingestion. Initially, modeling results were taken without augmentation, but as expected, the model would guess that the image was sharp every time with a roughly 90% accuracy. This spurred the need for some level of augmentation. Post-augmentation results without adding any artificial blur spurred similar results, so we tried adding Gaussian blur to some images to balance the sample set. Once we began working with a more balanced set, the SVM began to predict images as blurry.

C. Siamese Neural Network

Initially, our team anticipated building a CNN to solve our task, but given the small sample size and imbalance of our data, we pivoted and developed a Siamese Neural Network (SNN) instead. SNNs are generally effective with small datasets because they focus on learning relative similarity between embeddings and, therefore, do not require a large amount of labeled data for each class.

Specifically, our SNN focuses on learning to differentiate between blurry and sharp images based on their feature embeddings rather than pixel-wise differences. Since our SNN focus on embedding similarity, it is more equipped to handle class imbalance as opposed to other deep learning techniques.

The SNN we developed, outlined in Figure 2, contains two identical sub-networks that have the same architecture, parameters, and weights. The MRI we want to classify is run through one sub-network, and our reference image is run through the other. At the end of each sub-network, the input image is embedded. After the images are embedded, the absolute distance is computed between the embeddings of the input image and the embeddings of the reference image to determine if the input image is blurry or not.

Before running our images through the SNN, we had to generate image pairs. Essentially, this means that each input image needed to be paired up with the reference image and then the pair needed to be labeled, 1 for when the pairs had the same classification or 0 for when they had different classifications.

Our SNN model has three different blocks. Our first two block contain the same three layers: a 3D convolutional layer to capture features from the images, a 3D pooling layer to down-sample the volume and a batch normalization layer to normalize the feautres. Our third block contains a 3D convolutional layer to extract high level features, a 3D global average pooling layer to reduce the dimensions, and a fully connected dense layer to produce the final embedding. The head of our Siamese model takes the embeddings from the two branches and computes the distance between their vectors. Then the output layer takes the distance and outputs a similarity score between 0 and 1. Specifically, our model uses binary cross-entropy loss to predict if the input pair is similar or not.

One drawback to using SNNs, is that they are highly sensitive to the reference image used. If the reference image does not accurately represent the class then some input images will likely be mislabeled. Therefore, it is important to invest time picking the reference image that most accurately represents the class trying to be identified and updating it as needed. For our SNN, we chose a blurry reference image that did not appear very blurry to the non-medically trained human eye. Going forward, the reference image that is used in this model should continue to be refined to ensure that the reference image contains the desired level of blur for Gamma Knife surgery. Picking a reference image may appear like a time-intensive, daunting task, but it allows for our model to be extensible to other medical applications that may be able to tolerate a

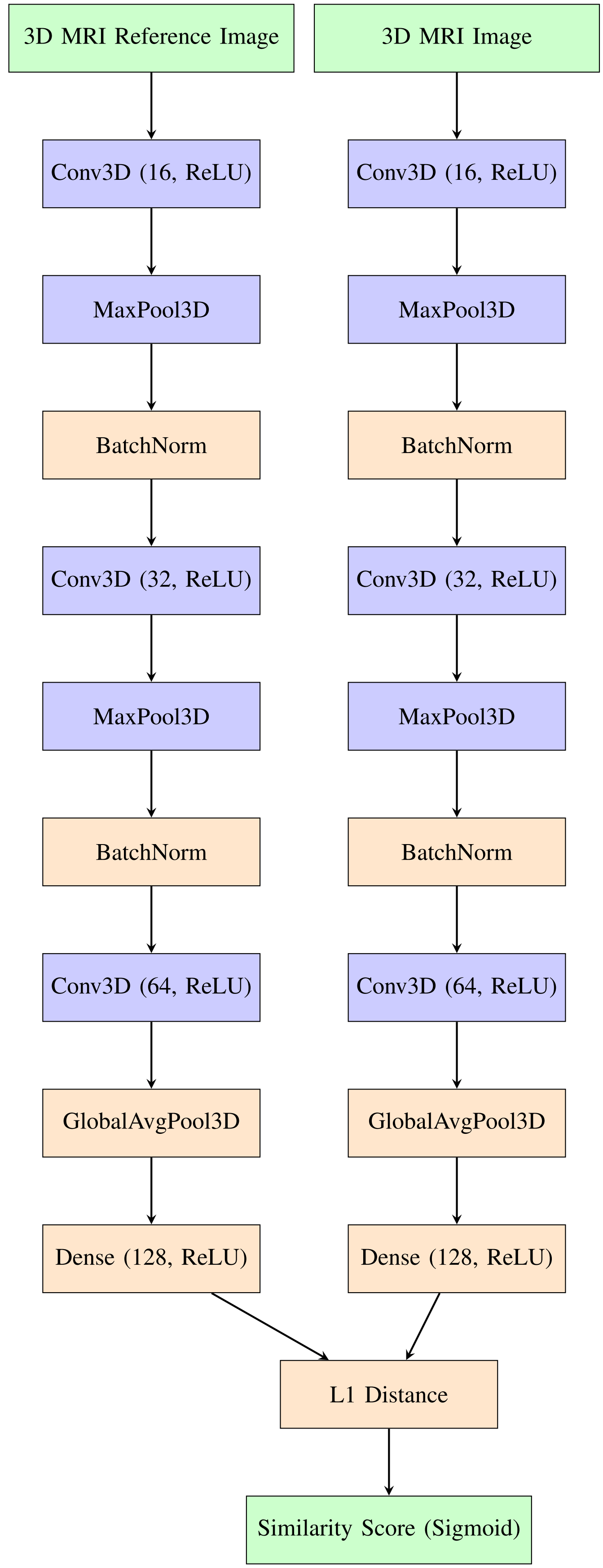

Fig. 2. Siamese Neural Network with 3D MRI Volumes

different level of blurriness than Gamma Knife surgery does. For example, if this model was implemented in a lab that focuses on identifying brain tumors in MRIs the reference image could be replaced with an MRI with a lower blurriness score. By changing the reference image used, we are able to adjust the desired threshold of blurriness to classify an image as blurry.

V. Results

A. Support Vector Machine

Label	Precision	Recall	F1 Score	Support
Blurry	0.20	0.33	0.25	3
Sharp	0.83	0.71	0.77	14
Accuracy			0.65	17
Macro Avg	0.52	0.52	0.51	17
Weighted Avg	0.72	0.65	0.68	17

Fig. 3. Pre-Tuning SVM Classification Report

The model was ran with a 70/30 training-test split. Initial modeling results pre-tuning held an accuracy of 65% with a weighted F1-Score of 0.68 as seen in Figure 3. The pre-tuned version correctly guessed that a blurry image was blurry about 20% of the time. However, it correctly guessed an image was sharp roughly 71% of the time. Post-tuning, the model slightly improved, correctly guessing that an image was sharp 93% and blurry 33% of their respective times as shown in Figure 4. While modeling results are not exactly stellar, it shows promise for future explorations.

Label	Precision	Recall	F1 Score	Support
Blurry	0.50	0.33	0.40	3
Sharp	0.87	0.93	0.90	14
Accuracy			0.82	17
Macro Avg	0.68	0.63	0.65	17
Weighted Avg	0.80	0.82	0.81	17

Fig. 4. Post-Tuning SVM Classification Report

B. Siamese Neural Network

The SNN was run using an 80/20 training testing split. Our initial SNN had an accuracy of 71% and precision of 74% on the training data, as shown in Figure 5. However, once we ran the SNN model on the test data we saw a drop-off in performance, with the accuracy falling to 58%. We believe this drop-off is likely due to the model overfitting to the training data. This hypothesis is supported by Figures 6 and 7 below, which show how accuracy increases and loss decreases as the model is trained for additional epochs.

After incorporating additional data augmentation techniques, as outlined earlier in this paper, our accuracy still hovered just under 60 percent. Although these results are less than ideal, we believe that creating a custom distance metric that assigns more weight to features that represent blurriness, such as the number of edges, has the potential to improve our model results.

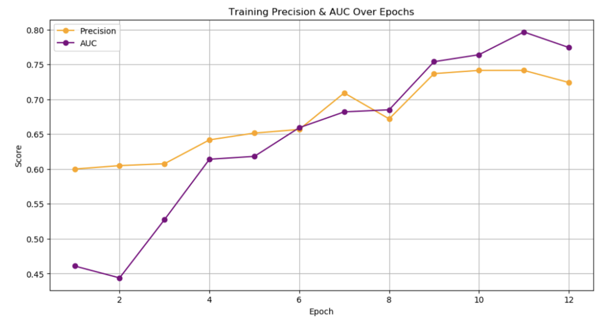

Fig. 5. Siamese Neural Network AUC and Precision

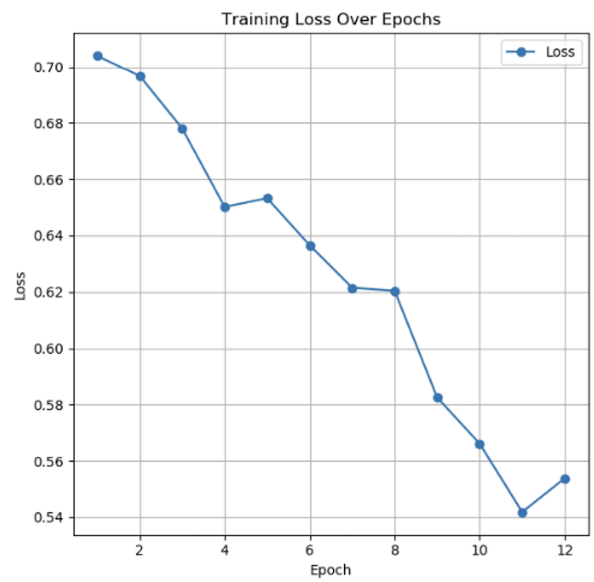

Fig. 6. Siamese Neural Network Training Loss

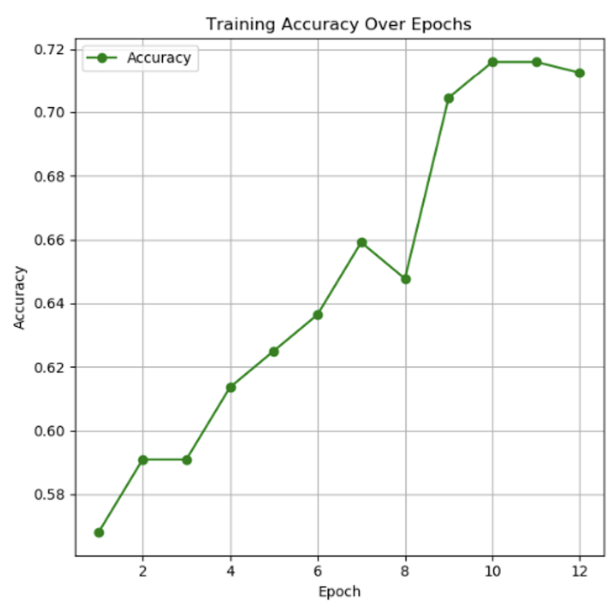

Fig. 7. Siamese Neural Network Training Accuracy

VI. Discussion

The results from both the SVM and SNN suggest that while our approach shows promise in identifying blurry MRIs, significant challenges remain. The SVM achieved relatively high precision and recall for sharp images, but performed poorly on detecting blurry ones, particularly after tuning. This limitation highlights a major flaw: severe class imbalance, compounded by the limited availability of blurry training samples.

Despite the small dataset, our use of image metrics provided meaningful input features for the SVM. However, these features alone appear insufficient to capture the variability introduced by motion artifacts and inconsistencies in clinical data. Data augmentation helps address this to some extent, but may have introduced artifacts not representative of real-world blurring, which affects model generalizability.

The SNN offered a theoretically better fit learning to compare the images rather than classify them outright. However, its performance plateaued around 60%, likely due to overfitting and sensitivity to the chosen reference image.

While our models can begin to classify image quality, they are not yet robust or reliable enough to support clinical decision making without further study. Overall, this project demonstrates feasibility of an automated image quality assessment model for MRI scans, but also spotlights the need for a more expansive dataset, improved feature representations, and higher quality tuning further outlined in Future Work.

VII. Conclusion

This work demonstrates the potential need for an automated system that can detect blurry MRI scans in real-time, potentially saving patients significant time and financial resources. However, substantial improves in model accuracy and precision are required prior to moving on to real world application. Expanding the training dataset and allocating more computational resources would improve progress towards desired performance benchmarks. The follow-up work outlined below presents a potential plan to improve model performance.

VIII. Future Work

This project would not only benefit from having a larger data set to work from, but having images organized by region would impact model performance. Multiple features the team wanted to extract ended up not making sense as each MRI would have different regions of the brain/eyes etc. For example, PSNR and SSIM both require a reference image to be determined. However, if we are looking at two completely different areas of the brain, then the comparison loses its intent. Having region-related buckets of data to work from would allow data scientists to actually consider these data points in models.

Leveraging third-party models through transfer learning would significantly expand the capabilities of this project beyond the current limited pre-trained models capabilities. Pre-trained deep learning models such as ResNet or Vision Transformers, are already optimized on a large data pool of medical images. As such, using these pre-trained models could provide substantial improvements to the feature extraction of the limited data set of patient MRIs. This approach would enable more sophisticated blur detection by transferring knowledge about image quality assessment from a broader medical imaging data pool. Therefore, this would substantially improve classification accuracy while reducing overfitting, which can occur when using a SNN. Furthermore, using a pre-trained architecture will improve memory allocation, allowing for faster training while producing results with higher accuracy.

After solidifying a model that both accurately and consistently identifies if an MRI is blurry or not, it would be extremely interesting and beneficial to try to recover a non-blurry image from the blurry images. This would prevent patients from having to have repeat MRIs done, because even if a blurry MRI was taken, a refined non-blurry image could be recovered. Some potential models to do this include the MIMO-UNet, which can be trained to learn the de-blurring function for 3D MRI [9]. However, this method will likely be computationally heavy and require more training data than is currently available.

IX. Acknowledgments

We would like to thank David J. Schlesinger, Ph.D and David R. Penberthy, MD for their insightful feedback and continuous encouragement throughout the course of this research. Their clinical knowledge greatly improved the clarity of this research. In addition, thank you for the support and sponsorship from the UVA Gamma Knife Center.

Additionally, we would like to thank Abbas Kazemipour, Ph.D for his mentorship and insightful feedback throughout the development of this work.

References

[1] A. Fatemi, S. Taghizadeh, C. C. Yang, M. R. Kanakamedala, B. Morris, and S. Vijayakumar, "Machine-Specific Magnetic Resonance Imaging Quality Control Procedures for Stereotactic Radiosurgery Treatment Planning," Cureus, vol. 9, no. 12, pp. e1957, Dec. 2017.

[2] K. Vecchiato et al., "Evaluation of DISORDER: Retrospective Image Motion Correction for Volumetric Brain MRI in a Pediatric Setting," Amer. J. Neuroradiol., vol. 42, no. 4, pp. 774-781, Apr. 2021.

[3] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese Neural Networks for One-shot Image Recognition," in Proc. 32nd Int. Conf. Machine Learning, Lille, France, 2015.

[4]] A. K. Allam, H. Sharma, M. B. Larkin, and A. Viswanathan, "Trigeminal Neuralgia: Diagnosis and Treatment," Neurol. Clinics, vol. 41, no. 1, pp. 107-121, Feb. 2023.

[5] S. Pertuz, D. Puig, and M. A. Garcia, "Analysis of focus measure operators for shape-from-focus," Pattern Recognition, vol. 46, no. 5, pp. 1415-1432, May 2013.

[6] O. Vincent and O. Folorunso, "A Descriptive Algorithm for Sobel Image Edge Detection," in Proc. Informing Science and IT Education Conf., Macon, GA, USA, 2009, pp. 97-107.

[7] M. Cordea, V. Manta, and C. Florea, "Comparison of focus measures in face detection and recognition," in Proc. Int. Conf. Computer Vision Theory and Applications (VISAPP), Barcelona, Spain, 2007, pp. 74-77.

[8] U. Sara, M. Akter, and M. S. Uddin, "Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative Study," J. Computer and Communications, vol. 7, no. 3, pp. 8-18, Mar. 2019.

[9] Cho, Sungmin, and Seungjun Lee. "Rethinking Coarse-to-Fine Approach in Single Image Deblurring." Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4641–4650.