# SynQ-ViT: Synthetic Quality Assessment for CT Calibration with Vision Transformer

**Abstract.** Optimising the Diagnostic Imaging System (DIS) requires considerate evaluation of image quality assessment (IQA) in CT scans using contrast media. In this paper, We present a novel methodology, called SynQ-ViT, for evaluating CT image quality using a novel vision transformer architecture. Specifically, we propose a hybrid architecture combining DenseNet, Vision Transformers, and reparameterization techniques to efficiently learn local and global features while maintaining computational feasibility. This approach maintains image fidelity, particularly in low-resolution contexts. SynQ-ViT involves restructuring the model for distinct training and inference phases to optimize memory usage and processing speed. The dataset comprises CT images of a Perspex phantom, injected with gold nanoparticle contrast media, collected with varying exposure settings. SynQ-ViT, achieved an R-squared value of 0.8892, outperforming other state-of-the-art models with fewer parameters, demonstrating its efficiency in predicting signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR) values. The effectiveness of various contrast agents, including Iodine-based agents and more recent substitutes like gold nanoparticles (AuNPs), is compared using IQA techniques.

**Keywords:** CT image quality · SNR · CNR · SynQ-ViT · Contrast agent

## 1 Introduction

X-ray imaging was the first imaging method used in medicine. However, as technology developed rapidly, other clinically used techniques were developed as well, such as computed tomography (CT), positron emission tomography (PET), PET/CT. Image quality is one of the main factors benefiting from advances in CT technology. The image quality metrics and application-specific parameters, including contrast, noise, resolution, signal-to-noise ratio, and depth-related image quality, differ between these modalities. This involves adjusting the CT equipment to correct any errors that could negatively impact image quality. Both objective and subjective techniques can be used to evaluate the quality of an image in DIS. The objective image quality evaluations might be dependent on equipment, such as modulation transfer functions or noise analysis, SNR, and CNR [1]. On the other hand, in the subjective methods such as visual grading analysis (VGA) and receiver operating characteristic (ROC), radiologists usually evaluate and judge the diagnostic images subjectively, and the performance of different imaging systems is compared using receiver operator characteristics [2].

In this study, we introduce a novel method called SynQ-ViT for automated CT IQA using machine learning. Our approach leverages a hybrid architecture combining DenseNet, Vision Transformers, and reparameterization techniques. This modification enables the model to learn local and global features effectively. We then added a branch for the regression task to predict critical image quality metrics such as CNR and SNR from the images. The data used in this study are CT images of a phantom manufactured from Perspex (polymethyl methacrylate (PMMA)), injected with gold nanoparticle contrast media at a concentration of 0.005 mg/ml. The data were collected with various exposure factor settings including Kiloboltage peak (KvP), which describes the maximum voltage that passes across the X-ray tube, Milliampere-seconds(mAs) which Tube current (mA) multiplied by exposure duration (s) and Rotation Time(RT) which is the amount of time it takes for the X-ray tube to circle the patient 360 times. The models used in the study have proven effective predict SNR/CNR value with CT images as input. The highest result was achieved with our proposed architecture, with an R-square value of 0.8892.

## 2 Related work

Based on the literature now available, AuNPs may serve as effective contrast agents in CT imaging, offering better contrast and image quality than traditional Iodine-based agents. Additionally, the work highlights how important it is to comprehend AuNP detectability limits in various phantom setups and improve imaging settings [3]. According to Taghavi et al. [4] AuNPs' ability to enhance contrast with iodinated contrast media was tested using a five-hole, 16 cm-diameter phantom. To scan the phantoms, various tube voltages and currents were used. The findings showed that AuNPs produced a higher CNR than iodinated contrast media at all tested concentrations and energies: the maximum CNR for AuNPs was recorded at 130 kVp. The results of these investigations demonstrate that AuNPs are a useful contrast agent in CT imaging when used with phantoms that have holes in them. They can improve image quality and offer better contrast than conventional Iodine-based agents, particularly at higher KVp [5].

Recent advancements in CT IQA emphasize the importance of metrics like SNR and CNR. A framework aligned with the International Commission on Radiation Units and Measurements (ICRU) Report 87 guidelines [6] includes measuring indices such as Modulation Transfer Function (MTF) and Noise Power Spectrum (NPS), along with noise level, CT number accuracy, and CNR. Chun et al. [7] improved CT IQA with fully automated techniques, focusing on noise level, Structure Sharpness Index (SSI), and Structure Alteration Index (SAI) using images from 120 patients and various reconstruction methods. Su et al. [8] introduced the Multi-view Multi-task Image Quality Assessment (M2IQA) for chest CT images, automating evaluations with a multi-view fusion strategy on images from 327 patients. Gao et al. [9] developed a Quality Assessment (QA) tool combining objective and subjective methods, identifying quality issues in the National Lung Screening Trial (NLST) dataset. Thiago V. M. Lima et al. [10] created the Machine Learning Tool for Image Quality Assessment in Computed Tomography (MAFIA-CT), using deep learning to score image quality based on human observations of low-contrast lesions. These studies highlight the impact of machine learning and deep learning in CT IQA.

While recent advancements in CT IQA have improved precision and automation, limitations still remain. The use of CNR and SNR for CT IQA, which are valuable for calibration, requires extensive data collection and varies across clinical environments. Studies often depend on instrument dependency, specific datasets and computationally intensive models, limiting real-time clinical application. Tools like the MAFIA-CT platform reduce human scoring variability but face limitations with non-modulated protocols and phantom design constraints.

## 3 Materials and methods

### 3.1 Data acquisition

The dataset used in this study comprises CT images of a specifically designed Perspex phantom. The phantom was scanned using a CT scanner (Biograph Vision 600) shown in the supplement document section CT scan features). It was injected with (AuNPs 0.005mg/ml) were also, and the phantom injected with Iodine with different concentrations (see supplement document section images for AuNPs(0.005mg/ml and images for Iodine contrast media at different concentrations), the phantom was scanned under a variety of exposure settings to rigorously evaluate image quality metrics. The exposure factors considered in this dataset include a range of kilovolt peak (kVp) values set at 80, 100, 120, and 140, along with milliampere-seconds(mAs) values of 215 and 429. Additionally, rotation times of 0.5 and 1 second were employed. For each unique combination of these exposure settings, 25 consecutive CT images were acquired, culminating in a comprehensive dataset consisting of 500 phantom CT images in total (see supplementary document attached figures in Images for AuNPs(0.005mg/ml)). The phantom was manufactured on-site from 10 cm-thick Perspex with 169 cylindrical holes in 13x13 matrix extruded into the phantom medium at varying diameters (4,4.5,5,5.5,6,6.5,7) mm to a constant depth of 5 cm. The distance between hole centres was fixed at 1 cm both horizontally and vertically. The justification behind using varying diameters for these tests is because the fundamental spatial resolution of a CT scanner is primarily hardware-dependent and influenced by factors of size and number of detector elements, the X-ray focus spot's size and distances from the source to object to the detector.

The 500 CT images were systematically cropped to isolate these individual holes, resulting in 45,500 cropped holes, each labeled with corresponding SNR and CNR indices. Figure 1 shows randomly selected images from our dataset.

According to Dong et al. [11], clinical CT was employed in a study to investigate AuNPs at the same concentrations and sizes using a cylindrical phantom with 12 holes (16 mm in diameter). It was discovered that attenuation rates

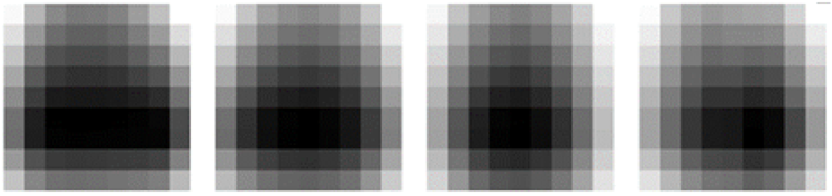

**Fig. 1.** Four random sample images from our dataset. Each of these images represents a hole in phantom images (each phantom image has 91 holes).

rose linearly with AuNP concentration, and AuNP size had an impact on image quality metrics such as CNR.

### 3.2 Methodology

To tackle the quality assessment task with CT images of phantoms as inputs, we propose employing a deep learning model that outputs continuous values, treating the task as a regression problem. Given the small size of the input images $X$, where the size of the input image ranges from 6 to 9 pixels we require an architecture that performs well with a low number of parameters while capturing both global and local features. In the context of low resolution image analysis, Talab et al. [12] highlight the importance of both global and local features. Extracting distinctive features from low resolution images requires advanced techniques that can effectively capture critical details.

Our proposed model (SynQ-ViT) focuses on learning both local and global features linearly. For local features, we employ dense blocks from DenseNet [13] due to their ability to effectively learn through feature reuse. The reduced connectivity and use of multi-scale convolution aggregation enhance the efficiency of feature reuse [14]. For global features, the Attention mechanism from Vision Transformers (ViT) is highly effective, as it enables the model to capture long-range dependencies and contextual information across the entire image [15]. To achieve efficient token mixing, reduce computational overhead through structural reparameterization, and enhance the model's capacity to learn complex patterns, we utilize the RepMixer block introduced by Vasu et al. [16].

SynQ-ViT architecture is divided into two distinct phases: training and inference. During training, the model learns parameters from data, passing outputs sequentially through layers to facilitate learning. Once trained, the model is restructured for inference, leveraging these learned parameters. This optimized structure enhances prediction efficiency, reducing memory overhead and increasing processing speed. Figure 2 illustrates overview of SynQ-ViT architecture with 2 separate train-inference architectures including a) Architecture of dense block, b) transition block, c) RepMixer used as cost assess reduction which reparametrizes a skip connection at inference d) Attention block use attention mechanism as token mixing. e) Explainability process where Grad-CAM is applied to the final convolutional layers of the model.

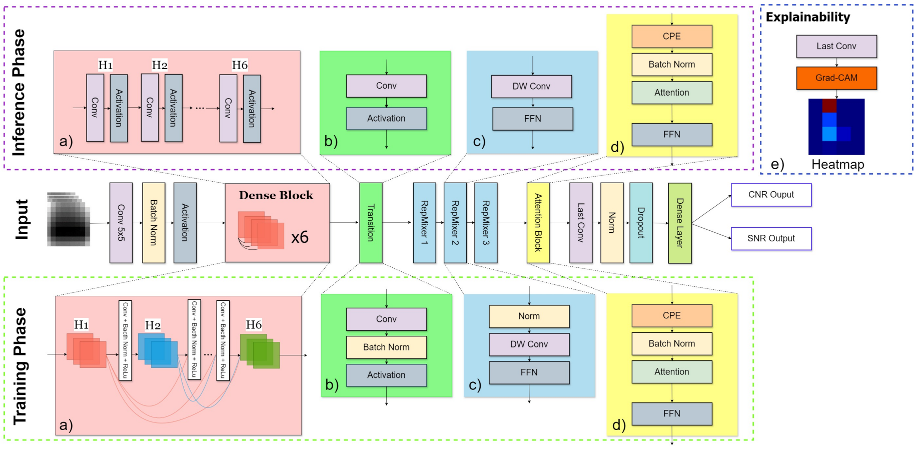

**Fig. 2.** Overview of SynQ-ViT architecture with 2 separate train-inference architectures. a) Architecture of dense block, b) transition block, c) RepMixer used as cost assess reduction which reparametrizes a skip connection at inference d) Attention block use attention mechanism as token mixing. e) Explainability process where Grad-CAM is applied to the final convolutional layers of the model.

**Dense block** (Fig. 2a) enhances information flow and gradient propagation between layers through dense connectivity. Each layer in the Dense block is directly connected to every other layer in a feed-forward manner, promoting feature reuse and alleviating the vanishing gradient issue. The output of the $l-th$ layer $x_l$ in Dense block is given by:

$$X_l = H_l([x_0, x_1, x_2, ..., x_l]) \tag{1}$$

where $H_l$ represents a composite function of operations at layer $l$, and $x_i$ are the concatenation of feature maps produced by layers 0 to $l-1$. Specifically, $H_l$ includes Batch Normalization (BN), Rectified Linear Units (ReLU), and a Convolution (Conv), ie: $H_l = Conv(ReLU(x))$.

**RepMixer block** (Fig 2c): This block operates as a token mixing block that leverages structural reparameterization to optimize performance during inference. RepMixer's design simplifies the network by reparameterizing skip connections, effectively eliminating them at inference time to reduce memory access costs and latency. Then, FFN (Feed Forward Neural) applied in the output to process and transform the token representations. At training phase, output of the layer computed as:

$$Yrep_{train} = DWConv(BatchNorm(Xrep)) + Xrep \tag{2}$$

Where $Yrep_{train}$ is output of RepMixer during training phase, $Xrep$ is the input to the RepMixer and DWConv is the depthwise convolution. During inference phase, the skip connection is reparameterized into the depthwise convolution

$$Yrep_{inf} = DWConv(Xrep) \tag{3}$$

**Attention block** (Fig 2d): This block operates as a mechanism to focus on the most relevant parts of the input data to capture long-range dependencies and contextual information. The Attention Block consists of a Conditional Positional Encoding (CPE), a multi-head self-attention (MHSA) mechanism, and a Feed Forward Network (FFN). With the input $Xattn$, CPE will enhance positional information by assigning tokens. The encoded input will then be linearly transformed into Q,K,V matrices representing query, key, and value to calculate MHSA for each head (relationships between tokens). The output of the attention mechanism is passed through a FFN, it can be summarized as:

$$Yattn = FFN(MHSA(CPE(Xattn))) + Xattn \tag{4}$$

During the inference phase, the same process is followed, utilizing the trained weight matrices to perform the attention.

In the explainability process (Fig 2e, we apply Grad-CAM (Gradient-weighted Class Activation Mapping) [17] block to the final convolutional layers of the model to visualize and interpret the model's decision-making process without changing the parameters. Finally, to generate the model's predictions, we use a Dense layer to transform the final feature representations into the desired continuous outputs. Specifically, the output layer is designed to produce two outputs, where each corresponds to a specific target value: SNR and CNR.

To train the model, we use the Mean Squared Error (MSE) for the combined regression tasks of SNR and CNR, and the coefficient of determination ($R^2$) as an additional performance metric. Let $\hat{y}_i = (\hat{y}_{i1}, \hat{y}_{i2})$ denote the predicted values for SNR and CNR, and $y_i = (y_{i1}, y_{i2})$ denote the true values. The MSE loss function measures the average squared difference between the predicted and true values:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}[(y_{i1} - \hat{y}_{i1})^2 + (y_{i2} - \hat{y}_{i2})^2] \tag{5}$$

Where $n$ is the number of samples, $(\hat{y}_{i1}, \hat{y}_{i2})$ denote the predicted values for SNR and CNR, and $(y_{i1}, y_{i2})$ denote the true values.

The $R^2$ metric assesses the proportion of variance in the dependent variables that is predictable from the independent variables:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}[(y_{i1} - \hat{y}_{i1})^2 + (y_{i2} - \hat{y}_{i2})^2]}{\sum_{i=1}^{n}[(y_{i1} - \bar{y}_1)^2 + (y_{i2} - \bar{y}_2)^2]} \tag{6}$$

Where $\bar{y}_1$ and $\bar{y}_2$ are the mean values of the true SNR and CNR, respectively.

Eventually, we utilized heatmap from Explainability process to visualize and analyze the critical regions identified by our models during prediction, particularly focusing on the ROI, which includes the holes within the phantom.

## 4 Experiment study

### 4.1 Data preprocessing

The image data in TIFF format, along with the associated SNR and CNR target values, were gathered for analysis. Each image was resized to 9 × 9 pixels to standardize the input dimensions for the deep learning model. To maintain the fidelity and scale of the SNR and CNR measurements, the pixel values were preserved in their original $16-bit$ format throughout the preprocessing stages. During the preprocessing phase, a thorough filtering process was implemented to remove any negative values present in the SNR and CNR data. These negative values typically arise from errors in data acquisition, often due to artifacts such as airs that occur when injecting contrast media into the phantom. By eliminating these erroneous values, we ensure that the dataset is clean and that the subsequent analysis is based on reliable and accurate data. Normalization of the SNR and CNR values was also performed to ensure that these metrics fall within a comparable range for better model training and evaluation. The SNR values in the dataset ranged from 0.006 to 130.968, while the CNR values ranged from 0.132 to 1291.274. To avoid value close to zero and to put both SNR and CNR on the same scale, these values were normalized to a range of 1 to 100.

### 4.2 Model training

To optimize the performance of our model, we conducted an extensive hyperparameter random search. The number of blocks within each stage was varied between 1 and 4. The growth rate was varied between 12 and 48 in increments of 12. The layer scale parameter was varied logarithmically between 0.000001 and 0.0001. The drop connect rate was adjusted between 0.0 and 0.5, while the dropout rate was varied in the same range. The learning rate for the Adam optimizer was varied logarithmically from $1 \times 10^{-4}$ to $1 \times 10^{-2}$. Here, *Gelu* was the activation function that we used. The tuning process involved using a Random Search strategy, where 20 trials were executed to explore the hyperparameter space. Early stopping was implemented to monitor the validation loss, with a patience of 10 epochs. If the validation loss did not improve for 5 consecutive epochs, a callback was employed to decrease the learning rate by a factor of *log*. In addition to introducing SynQ-ViT, we experimented with various other models to ensure a comprehensive evaluation. These models included FastViT [16] , ResNet [18], RNN [10], SE-ResNet [20], UNet-NILM [21], SqueezeNet [22]. Each model was subjected to a similar hyperparameter tuning process to ensure optimal performance and fairness.

## 5 Result

The training results of SynQ-ViT for the regression problem of predicting SNR and CNR values with CT images of a phantom compared to other experiment models are presented in Table 1. The model achieved optimal performance, attaining an R-squared value of approximately 0.89 and an MSE of 16.03 after 61 epochs on the validation set. Notably, convergence was reached quite early, around the 5th epoch. This early convergence and the model's strong performance with low-resolution input images can be attributed to our's architecture, which effectively reuses learned features from previous layers. In comparison with other robust computer vision models, as outlined in section 3, we evaluated additional metrics, including the number of parameters calculated during training, with results detailed in Table 1. It is evident that our proposed model surpasses other models in terms of performance. Conversely, the ResNet model, despite having fewer parameters, exhibits a lower R-squared value by approximately 0.05 and a slower convergence rate, taking 87 epochs.

**Table 1.** Experiment detailed result.

| Architecture | Epochs | Parameters | MSE | $R^2$ |
|---|---|---|---|---|
| SynQ-ViT (ours) | 61 | 145,902 | 16.0347 | 0.8892 |
| FastViT | 36 | 3,210,530 | 18.5105 | 0.8714 |
| RNN | 64 | 334,082 | 28.4053 | 0.8074 |
| Resnet | 87 | 118,002 | 22.0120 | 0.8487 |
| SqueezeNet | 32 | 13,026 | 17.5592 | 0.8785 |
| SE-ResNet | 44 | 8,672,624 | 17.3220 | 0.8798 |
| UNet-NILM | 25 | 2,032,578 | 20.2350 | 0.8600 |

We employed Grad-CAM to our best-performing model to visualize the areas of focus during predictions. Figure 3 illustrates the heatmaps generated from randomly selected images, highlighting the regions the model relied on for its predictions. The red areas in the heatmap correspond to the ROI in the CT images. These visualizations confirm that the model accurately focuses on the critical areas of the input images, thereby validating its effectiveness and reliability in predicting SNR and CNR values.

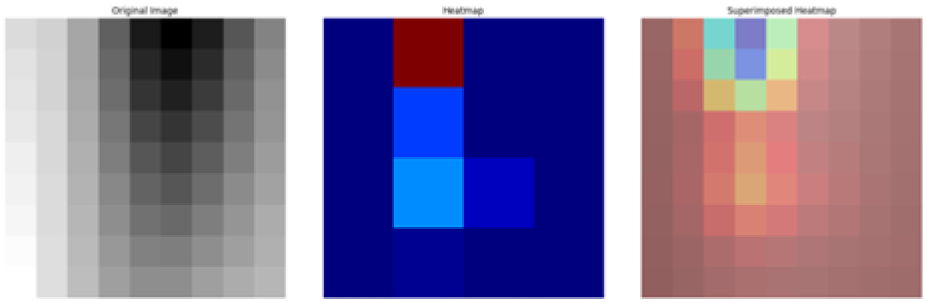

**Fig. 3.** Grad-CAM illustration of our model. First is the original image, the second is the heatmap image generated through the final convolutional layer of the model, indicating the areas the model focuses on to predict SNR and CNR, and the final image is the heatmap superimposed on the original image.

## 6 Conclusions and Discussion

In conclusion, the high concentration of contrast agents affects the kidney and live tissue, making it impossible for the Iodine contrast medium or AuNPs to leave the body. (see supplementary document - clinical context explanations). Moreover, people with kidney diseases could discover that using AuNPs at low concentrations is a safer option than using conventional Iodine-based contrast agents. This is because AuNPs may produce good imaging at lower dosages and the atomic number for AuNPs is 79 which is higher than Iodine with an atomic number of 53. Here, we introduced a model called SynQ-ViT that combines DenseNet, attention mechanisms, and reparameterization techniques to efficiently learn local and global features. SynQ-ViT addresses the regression problem with the input being CT images of phantom holes and the output being the SNR and CNR values to assess the quality of the CT images. SynQ-ViT can deal with small images and has achieved good performance in terms of both efficiency and computational cost. Throughout the training process, we discovered that the SNR and CNR values are strongly correlated with the holes in the CT images, even at low resolution.

## References

1. Ullah, F., Lee, J., Jamil, S. and Kwon, O.J., 2023. Subjective assessment of objective image quality metrics range guaranteeing visually lossless compression. Sensors, 23(3), p.1297.
2. Nocum, D.J., Robinson, J., Halaki, M., Båth, M., Mekiš, N., Liang, E., Thompson, N., Moscova, M. and Reed, W., 2022, April. Visual grading characteristic (VGC) analysis of uterine artery embolisation (UAE) image quality assessment by interventional radiologists and interventional radiographers. In Medical Imaging 2022: Image Perception, Observer Performance, and Technology Assessment (Vol. 12035, pp. 116-122). SPIE.
3. Oumano, M., Russell, L., Salehjahromi, M., Shanshan, L., Sinha, N., Ngwa, W. and Yu, H., 2021. CT imaging of gold nanoparticles in a human-sized phantom. Journal of applied clinical medical physics, 22(1), pp.337-342.
4. Taghavi, H., Bakhshandeh, M., Montazerabadi, A., Moghadam, H.N., Shahri, S.B.M. and Keshtkar, M., 2020. Comparison of gold nanoparticles and iodinated contrast media in radiation dose reduction and contrast enhancement in computed tomography. Iranian Journal of Radiology, 17(1).
5. Sibuyi, N.R.S., Moabelo, K.L., Fadaka, A.O., Meyer, S., Onani, M.O., Madiehe, A.M. and Meyer, M., 2021. Multifunctional gold nanoparticles for improved diagnostic and therapeutic applications: a review. Nanoscale Research Letters, 16, pp.1-27.
6. Pahn, G., Skornitzke, S., Schlemmer, H.P., Kauczor, H.U. and Stiller, W., 2016. Toward standardized quantitative image quality (IQ) assessment in computed tomography (CT): A comprehensive framework for automated and comparative IQ analysis based on ICRU Report 87. Physica Medica, 32(1), pp.104-115.
7. Chun, M., Choi, J.H., Kim, S., Ahn, C. and Kim, J.H., 2022. Fully automated image quality evaluation on patient CT: Multi-vendor and multi-reconstruction study. PLoS One, 17(7), p.e0271724.

8. Su, J., Li, M., Lin, Y., Xiong, L., Yuan, C., Zhou, Z. and Yan, K., 2023. Deep learning-driven multi-view multi-task image quality assessment method for chest CT image. BioMedical Engineering OnLine, 22(1), p.117.
9. Gao, R., Khan, M.S., Tang, Y., Xu, K., Deppen, S., Huo, Y., Sandler, K.L., Massion, P.P. and Landman, B.A., 2021. Technical report: quality assessment tool for machine learning with clinical CT. arXiv preprint arXiv:2107.12842.
10. Lima, T.V., Melchior, S., Özden, I., Nitzsche, E., Binder, J. and Lutters, G., 2021. Mafia-ct: Machine learning tool for image quality assessment in computed tomography. In Medical Image Understanding and Analysis: 25th Annual Conference, MIUA 2021, Oxford, United Kingdom, July 12–14, 2021, Proceedings 25 (pp. 472-487). Springer International Publishing.
11. Dong, Y.C., Hajfathalian, M., Maidment, P.S., Hsu, J.C., Naha, P.C., Si-Mohamed, S., Breuilly, M., Kim, J., Chhour, P., Douek, P. and Litt, H.I., 2019. Effect of gold nanoparticle size on their properties as contrast agents for computed tomography. Scientific reports, 9(1), p.14912.
12. Talab, M.A., Awang, S. and Ansari, M.D., 2020. A Novel Statistical Feature Analysis-Based Global and Local Method for Face Recognition. International Journal of Optics, 2020(1), p.4967034.
13. Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).
14. Hess, A., 2018. Exploring feature reuse in DenseNet architectures. arXiv preprint arXiv:1806.01935.
15. Li, Y., Wang, J., Dai, X., Wang, L., Yeh, C.C.M., Zheng, Y., Zhang, W. and Ma, K.L., 2023. How does attention work in vision transformers? A visual analytics attempt. IEEE transactions on visualization and computer graphics, 29(6), pp.2888-2900.
16. Vasu, P.K.A., Gabriel, J., Zhu, J., Tuzel, O. and Ranjan, A., 2023. FastViT: A fast hybrid vision transformer using structural reparameterization. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5785-5795).
17. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).
18. Targ, S., Almeida, D. and Lyman, K., 2016. Resnet in resnet: Generalizing residual architectures. arXiv preprint arXiv:1603.08029.
19. Grossberg, S., 2013. Recurrent neural networks. Scholarpedia, 8(2), p.1888.
20. Thiruppathi, K., Selvakumar, K. and Shenbagavel, V., 2023. SE-RESNET: Monkeypox Detection Model. International Journal of Advanced Computer Science and Applications, 14(9).
21. Faustine, A., Pereira, L., Bousbiat, H. and Kulkarni, S., 2020, November. UNet-NILM: A deep neural network for multi-tasks appliances state detection and power estimation in NILM. In Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring (pp. 84-88).
22. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J. and Keutzer, K., 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360.
23.