Journal of Imaging Informatics in Medicine (2025) 38:629–645
https://doi.org/10.1007/s10278-024-01205-8

Let UNet Play an Adversarial Game: Investigating the Effect of Adversarial Training in Enhancing Low-Resolution MRI

Mohammad Javadi[1] · Rishabh Sharma[1] · Panagiotis Tsiamyrtzis[2,3] · Andrew G. Webb[4] · Ernst Leiss[5] · Nikolaos V. Tsekos[1]

Received: 7 February 2024 / Revised: 8 July 2024 / Accepted: 12 July 2024 / Published online: 31 July 2024

Abstract

Adversarial training has attracted much attention in enhancing the visual realism of images, but its efficacy in clinical imaging has not yet been explored. This work investigated adversarial training in a clinical context, by training 206 networks on the OASIS-1 dataset for improving low-resolution and low signal-to-noise ratio (SNR) magnetic resonance images. Each network corresponded to a different combination of perceptual and adversarial loss weights and distinct learning rate values. For each perceptual loss weighting, we identified its corresponding adversarial loss weighting that minimized structural disparity. Each optimally weighted adversarial loss yielded an average SSIM reduction of 1.5%. We further introduced a set of new metrics to assess other clinically relevant image features: Gradient Error (GE) to measure structural disparities; Sharpness to compute edge clarity; and Edge-Contrast Error (ECE) to quantify any distortion of the pixel distribution around edges. Including adversarial loss increased structural enhancement in visual inspection, which correlated with statistically consistent GE reductions (p-value $<<0.05$). This also resulted in increased Sharpness; however, the level of statistical significance was dependent on the perceptual loss weighting. Additionally, adversarial loss yielded ECE reductions for smaller perceptual loss weightings, while showing non-significant increases (p-value $>>0.05$) when these weightings were higher, demonstrating that the increased Sharpness does not adversely distort the pixel distribution around the edges in the image. These studies clearly suggest that adversarial training significantly improves the performance of an MRI enhancement pipeline, and highlights the need for systematic studies of hyperparameter optimization and investigation of alternative image quality metrics.

Keywords Generative Adversarial Networks · Adversarial loss · REAL-ESRGAN · UNet · Mixed effects modeling

Introduction

In many clinical implementations of low-field magnetic resonance imaging (MRI), under-sampling the spatial frequency domain (k-space) or acquiring low spatial resolution images are used to acquire images in a clinically acceptable time. These methods, however, can yield lower image quality and significant image artifacts [1, 2]. To overcome these problems, recent works have explored deep learning (DL) methodologies [3–5] to improve the signal-to-noise ratio

✉ Nikolaos V. Tsekos
nvtsekos@central.uh.edu

Mohammad Javadi
mjavadi@uh.edu

Rishabh Sharma
sharm32.rishabh@gmail.com

Panagiotis Tsiamyrtzis
pt@aueb.gr

Andrew G. Webb
a.webb@lumc.nl

Ernst Leiss
cosc1@central.uh.edu

1 Medical Robotics and Imaging Lab, Department of Computer Science, University of Houston, 501, Philip G. Hoffman Hall, 4800 Calhoun Road, Houston, TX 77204, USA

2 Department of Mechanical Engineering, Politecnico di Milano, Milan, Italy

3 Department of Statistics, Athens University of Economics and Business, Athens, Greece

4 C.J. Gorter Center for High Field MRI, Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands

5 Department of Computer Science, University of Houston, Houston, TX, USA

(SNR) of these images. The primary goal of these studies has been to establish an optimal mapping from the "lower-quality" (LQ) images to the "higher-quality" (HQ) images, effectively achieving what is known as super-resolution (SR).

Various studies have investigated a wide spectrum of methodologies, including the utilization of Convolutional Neural Networks (CNNs) [6, 7], Generative Adversarial Networks (GANs) [8–11], and Attention Networks [12–14], as well as work focused on exploring various loss functions [15–19]. The loss functions of these enhancement networks often comprise additions of several weighted terms. One term is a pixel-wise loss, which primarily reflects the reconstruction of the low-frequency details, one such loss being Mean Intensity Error (MIE) [9, 18, 19]. A second term is perceptual loss to address the reconstruction of high-frequency details [9, 20–22], for example, VGG loss [23]. A third term is the adversarial loss to improve visual realism, which has been used in both non-medical [22, 24–29] and medical [8, 11, 18, 19, 30–33] applications.

Evaluating the effect of adversarial loss has been overlooked in medical image enhancement, where this loss is just incorporated as part of a weighted loss function. Such evaluations have been conducted only in non-clinical research, where visual realism serves as the key image enhancement criterion [18, 22, 27, 30–37]. These evaluation methodologies are not applicable in clinical SR because, in such a context, other image features like structure and sharpness are determined as enhancement criteria, which are crucial for characterizing pathological foci.

Also, previous clinical works neglect to perform any systematic search to find the optimal weighting of the adversarial loss component [24–29, 38]; however, the optimal value of the loss weightings for each network is unique, i.e., each optimization problem has a unique search space of hyperparameters [39–42]. This lack of systematic study on the impact of adversarial training in medical imaging, especially in MRI, formed the primary research question of our work: can training an MRI enhancement network through pixel and perceptual losses alone sufficiently improve the enhancement performance, or would combining these loss terms with an adversarial loss within a systematic hyperparameter search yield even greater improvements?

To answer this question, we trained a network with varying loss term weightings. We then incorporated appropriate statistical tests to measure the significance of different loss weightings, particularly adversarial loss weight, across various evaluation metrics of clinical relevance.

We performed a systematic hyperparameter search using a composite loss function with a weighted addition of MIE, perceptual, and adversarial loss terms. We tried 6 different perceptual loss weights, and for each weight, we tested 18 different adversarial loss weights progressively increasing from 0 to 3. This adjustment of adversarial loss weight at each perceptual loss weight enables investigation of the effect of adversarial loss at each different level of its contribution to the total loss.

For evaluation criteria, this study required a set of metrics that capture the critical properties of a diagnostic image. Previous works in MRI SR use traditional metrics (e.g., PSNR, MIE, SSIM) [10, 11, 43–47] to measure structural enhancements, although they inherently lack correlation with the human visual system in highlighting structural disparities [48]. These metrics also tend to overlook the quality of high-frequency details, another pivotal property in the interpretability of a medical image across various regions of interest [49]. In response to these issues, we introduce a new set of metrics that measures both the structure and quality of high-spatial-frequency details: these are termed Gradient Error (GE), Sharpness, and Edge-Contrast Error (ECE). We follow the approach of previous works [32, 50, 51] and evaluate our metrics specifically on a lesion, unlike other works that perform the assessment over the entire enhanced image [8–10, 14, 43, 52].

In the context of statistical inference, a proper choice of significance tests on the previously outlined evaluation metrics is required to assess the clinical relevance of using adversarial training. The inference phase in previous works shows a strong inclination towards assessing mean differences [12, 18, 53, 54], often excluding a consideration of whether these means fall within the same standard deviation. This approach is also used in hyperparameter selection techniques, such as grid search, which selects the best hyperparameter configuration based on minimizing the mean of an error (e.g., MIE) or maximizing the mean of a reward (e.g., SSIM), without accounting for variation between the mean values or considering the effect of factors (e.g., learning rate) on a specified criterion [55, 56]. Some previous works tackled this issue by measuring the significance of any differences using analysis of variance (ANOVA) [57, 58]; however, this approach neglects the condition that the test population in MRI SR studies is measured repeatedly across different experimental setups, which violates the assumptions behind traditional ANOVA [39].

To address these issues, this study conducts a statistical analysis of factors, with a particular focus on adversarial loss weight, using mixed effects models (MEM) [39, 43, 59–61]. This statistical model effectively addresses the issue of having repeated measurement, since our test data comes from an identical pool of subjects evaluated under varying inference setups. Additionally, this MEM analysis accounts for variability between images as a random effect, aiming to ensure that any observed changes in our response variables are not confounded with differences within the tested images.

Materials and Methods

Network

There is a wide variety of GAN architectures in the SR domain [22, 27, 35, 62], from which we focused on *REAL-ESRGAN* [22]. Since the GAN generator in our study is relatively simple, we used *UNet* [63], which has shown a strong record of performance in medical applications [47, 64–69], as the generator of our GAN architecture. For the discriminator, a UNet with the inclusion of spectral normalization is employed to address per-pixel feedback to the generator [70]. The discriminator takes an enhanced image and computes the realness probability map of its pixels. An average is taken over this map to provide a score of the realness of the image and provide the discriminator gradient feedback to the generator. An overview of our network architecture is shown in Fig. 1.

Objective Function

The objective function used in the studied network incorporates the three aforementioned terms: mean intensity loss, perceptual loss, and adversarial loss. Each of these terms is discussed in detail in the subsequent sections.

Mean Intensity Loss this is one of the most popular cost functions used in SR tasks; it is measured by taking an average over the error map between generator output and the ground truth. Equation (1) formulates the MIE loss as:

$$J_{mie} = E(I_{LR}, I_{HR}) \| G(I_{LR}) - I_{HR} \|_1 \tag{1}$$

where $G(I_{LR})$ is the output of the generator G given I_{LR} as an LQ sample.

Perceptual Loss Regardless of the wide popularity of mean intensity loss in super-resolution applications, it leads the network to de-emphasize high-spatial frequency details and produce blurry outputs. To tackle this issue, as in our baseline (REAL-ESRGAN), we included perceptual loss as the second term of our objective function, which has been empirically found to help the upscaling network to reconstruct sharper images with more structural details. To compute the perceptual loss, we fed the image to a VGG-19 [71] network pre-trained on ImageNet [72]. A weighted sum with weights of {0.1, 0.1, 1, 1, 1} was then taken over the L2 norm of activation maps obtained from the first five convolution layers of the network. This loss can be formulated as:

$$J_{per} = \sum_{i=1}^{5} W_i * E(I_{LR}, I_{HR}) \| \phi_i(G(I_{LR})) - \phi_i(I_{HR}) \|_2 \tag{2}$$

where ϕ_i represents the feature map generated by the i_{th} layer of the VGG-19, and W_i determines the contribution of the i_{th} layer's loss value to the final perceptual loss.

Adversarial Loss The GAN discriminator is a classifier to distinguish real and generated (fake) samples. The more realistic the generator output is, the more likely the discriminator is to classify a generated sample (fake) as real. This explains the adversarial game scenario, where a trained discriminator is expected to be fooled by the generator and classify a generated sample as real [73]. If this fails, then the discriminator provides this failure as gradient feedback to the generator, so the generator updates its parameters and tries to fool the discriminator again. This failure, denoted as adversarial loss, is computed by taking a cross-entropy loss between the discriminator output on generated images and the real class' ground truth value:

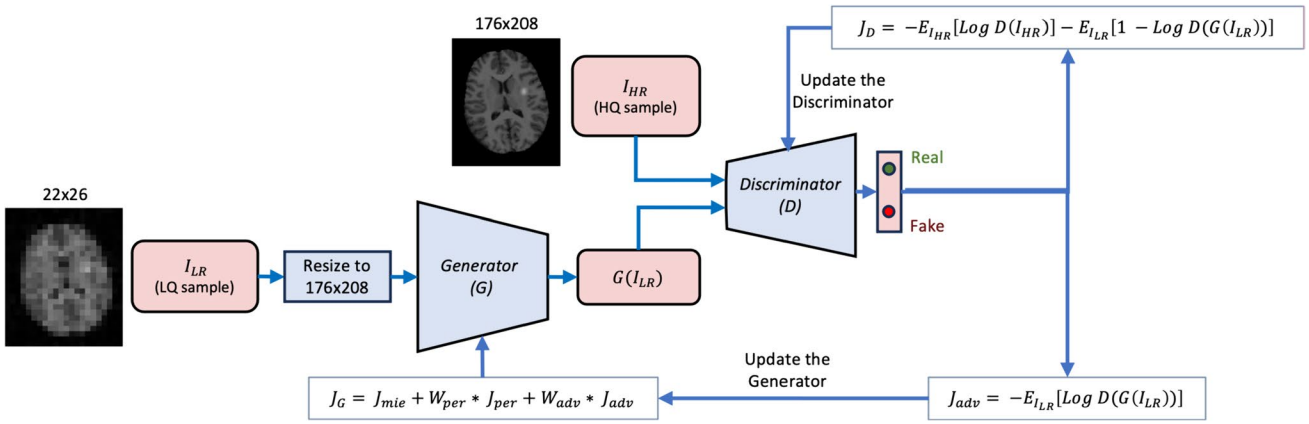


Fig. 1 An overview of our GAN architecture

$$J_{adv} = -E_{I_{LR}}\left[Log\, D\left(G(I_{LR})\right)\right] \quad (3)$$

and the discriminator objective function is defined as:

$$J_D = -E_{I_{HR}}\left[Log\, D\left(I_{HR}\right)\right] - E_{I_{LR}}\left[1 - Log\, D\left(G(I_{LR})\right)\right] \quad (4)$$

Total Loss The final objective function of the generator is:

$$J_{total} = J_{mie} + W_{per} * J_{per} + W_{adv} * J_{adv} \quad (5)$$

where W_{per} and W_{adv} determine the contribution of each loss term to the total loss. We examined the effect of these weights as fixed factors in our statistical analysis on the designated evaluation criteria. The W_{adv} coefficient takes 18 values {0, 0.0001, 0.001, …, 0.5, 1, 1.2, 1.5, 2, 3}, where the training scenario with $W_{adv} = 0$ is designated as "*VGG-Enhanced*", and the one with $W_{adv} > 0$ is denoted as "*VGG+GAN-Enhanced*". Increasing W_{adv} over this wide range affects the degree to which the generator factors in the discriminator's assessment of determining the realism of the generator output. This exploration of the W_{adv} coefficient determines the key step towards answering the question on the effect of adversarial training on final image quality.

The W_{per} value was originally set to 1.0 in REAL-ESRGAN. However, our experiment modified this hyperparameter across the range {0.5, 0.7, 1, 1.2, 1.5, 2}. Testing W_{per} at 6 different levels allowed us to measure how emphasizing high-spatial frequency information impacts the result of adversarial training.

Training and Testing Dataset

A Deep Learning-based pipeline requires a large number of LQ images with their corresponding HQ datasets. For our experiment, we trained and evaluated our networks on OASIS-1 [74] which is a widely used neuroimaging dataset of 207 × 176 (data matrix) T1-weighted brain images and provides MRI and demographic data from a group of healthy and cognitively impaired individuals. We used the 436 MR images from this dataset as our initial HQ samples and an image synthesis technique based on [43] to augment our HQ portion and subsequently generate their corresponding LQ data, to include structural, noise, and signal intensity variations.

Each sample in the OASIS-1 dataset was augmented through rotations, adding hyperintense lesions of random size/position to white and gray matter regions, applying elastic deformations, and combining tissue maps (white matter, gray matter, CSF) with random weightings to create synthetic higher-quality MRI images (HQ). Lower-quality images (LQ) were then derived from the HQ images by adding noise, and k-space truncation to 1/8 × 1/8 of the matrix size of the full k-space of the HQ images. A complete explanation of this synthesis pipeline is outlined in [43].

The new dataset contains a total of 2616 samples from the original patient data. From this dataset, we carefully curated the train/validation/test splits at the patient level, i.e., no patients appeared across multiple splits, to prevent any data leakage between the three splits that could lead to excessively optimistic performance estimates. The samples from 346 patients (2076 images) were selected for the training phase, with 20% of those used for validation during the model training process to prevent overfitting. The remaining 540 samples from 90 patients were retained as the test set for final evaluation. Importantly, we ensured that these 540 test samples represented a diverse range of data variations, thereby guaranteeing variability not only across the training and test sets, but also within the test set itself.

Network Implementation and Training

All the networks were trained on the Sabine Cluster of the University of Houston using PyTorch [75]. Following the training procedure of *Real-ESRGAN*, we initially trained the generator for 200 epochs under an L1 loss using the Adam [76] optimizer and learning rate (LR) of 2×10^{-4}. The weights obtained at this stage were then transferred to the generator of our network architecture (Fig. 1). We continued to train this network for 500 epochs under the composite loss function shown in Eq. (5), with a batch size of 8 using two distinct initial learning rates of 0.001 and 0.01. We included the learning rate as another factor in our statistical analysis alongside W_{per} and W_{adv}, meaning in total we have 216 different trained networks for evaluation: 12 networks in the VGG-Enhanced group and 204 networks in the VGG+GAN-Enhanced group.

Since the input and output size of the UNet architecture are the same, we preprocessed an LQ sample before feeding it to the network and resized it with nearest neighbor interpolation to the dimension of its HQ counterpart (176 × 208). We also normalized our LQ and HQ samples between 0 and + 1 to tackle the issue of gradient explosion.

Evaluation Metrics

We employed five metrics to compare the performance of our enhancement networks within the lesion area: (1) Gradient Error (GE); (2) Sharpness; (3) Edge-Contrast Error (ECE); (4) Mean Intensity Error (MIE); (5) Structural Similarity Index (SSIM). GE, MIE, and SSIM are calculated within a region cropped from the top-left to the bottom-right coordinates of the lesion. This process was facilitated by the fact that we had access to lesion coordinates in our database as they were synthesized. In the following sections,

we elaborate on the algorithms to compute GE, Sharpness, and ECE.

Gradient Error

Despite the widespread usage of SSIM to quantify the structural similarities between images, this is not the primary goal of this metric as it fundamentally lacks correlation with the human perception system [48]. In this study, we opted for gradient error (GE) as the metric to address the structural difference between an enhanced image and its ground truth. This metric evaluates the extent to which one image replicates the edges and contours present in another image. This idea is derived from the *Flip* metric [77], which employs gradient error as part of similarity measurement between two images. Our work quantifies the deviation of edges and structures for an enhanced image using this metric. Given images A and B, the Gradient Error between these images can be formulated as:

$$GE(A,\ B) = \frac{GE^x(A,\ B)\ +\ GE^y(A,\ B)}{2} \tag{6}$$

with $GE^a(A, B)$, in general, being the Gradient Error between A and B on axis a, and can be measured by:

$$GE^a(A,\ B) = \frac{\left|\ ||\ \nabla G_A^a\ ||_1\ -\ ||\ \nabla G_B^a\ ||_1\right|}{N} \tag{7}$$

where ∇G_A^a and ∇G_B^a are first-order derivatives of images A and B on a given axis a. A 3×3 Sobel filter is applied to an image in order to compute the first-order derivative on a given axis. N represents the number of non-zero elements in the elementwise multiplication of ∇G_A^a and ∇G_B^a.

Sharpness

This metric measures the quality of high-spatial frequency components in an image. The computation of Sharpness is based on Sigmoidal Modeling of Edges [49] which has been shown to be noise resilient. Let i be an enhanced image with g as its ground truth, the Sharpness of i can be computed as follows:

1. Retrieve the corresponding lesion map of g from the database as l.
2. Given coordinates of l, compute its center of mass c.
3. Locate E, a set of edge points in l that form angles of principal directions to c. The angle of edge point e to c is defined as $\theta(e)$.
4. For all the 8 edge pixels in E, extract their corresponding pixels in i and create a new set called E'.
5. For each pixel in E', create an intensity profile. The profile for an edge pixel e is a set of pixels along a 5-unit line segment that fall within the direction of $\theta(e)$ (Fig. 2b).
6. Fit a Sigmoid function f to each intensity profile, given by:

$$f\left(p;\ a_0,\ a_1,\ a_2,\ S\right) = a_1 + \frac{a_2}{1 + e^{S*(a_0 - x)}} \tag{8}$$

where p is an intensity profile for the edge pixel e, a_0 is the Sigmond's midpoint, a_1 defines the vertical offset (minimum value), a_2 determines the edge contrast, and S determines the sharpness of profile p or shows the growth rate of the fitted Sigmoid at e.
7. The Sharpness for image i can be computed by:

$$Sharpness(i) = \frac{1}{N} * \sum_{j=1}^{N} S_{ji} \tag{9}$$

where N determines the number of intensity profiles, which in our case is 8. S_{ji} is the S parameter estimated for the fitted sigmoid function on the intensity profile j in image i.

An overview of the steps for Sharpness measurement is shown in Fig. 2, with the first 4 steps being summarized in Fig. 2a, Step 5 in Fig. 2b, and Step 6 in Fig. 2c.

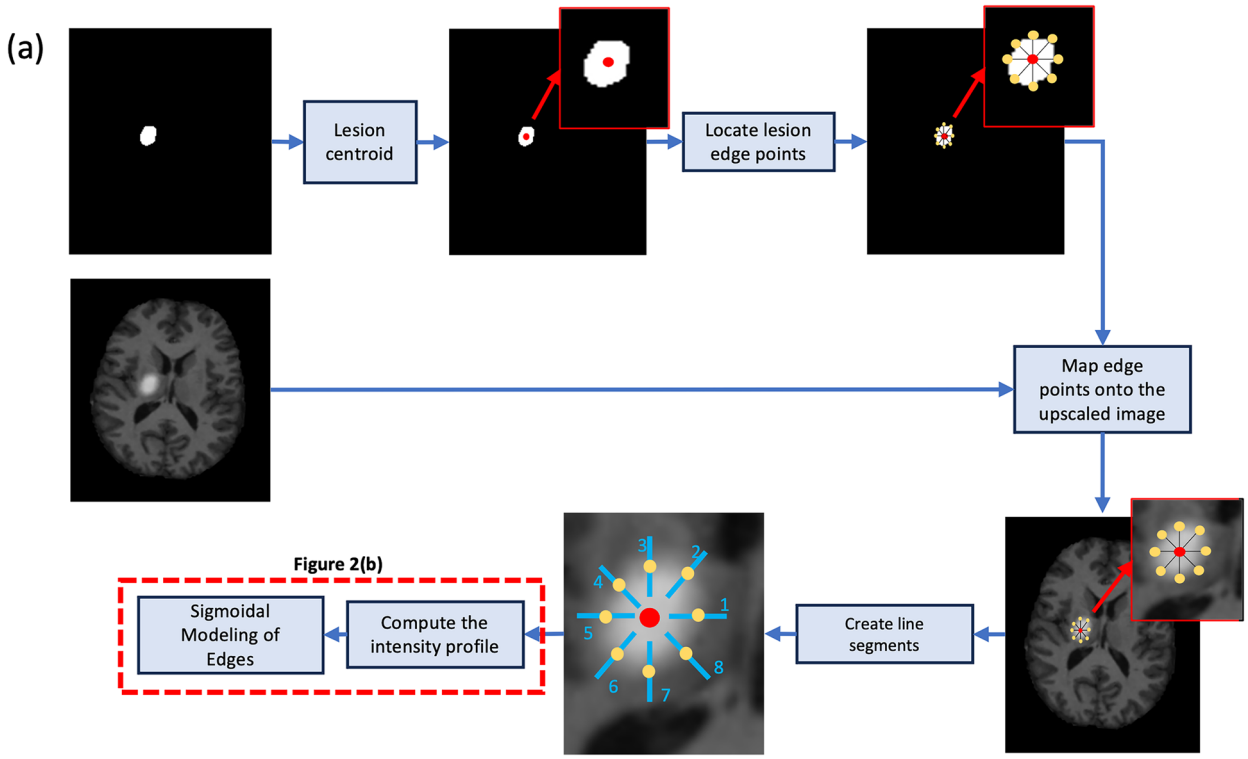
Edge-Contrast Error (ECE)

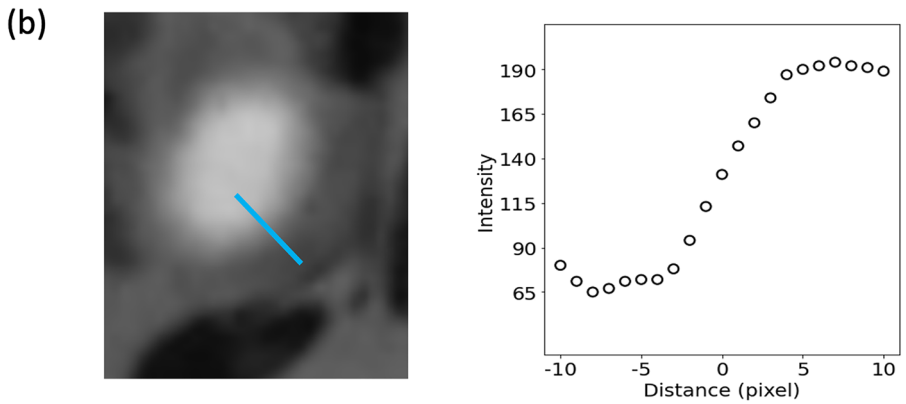
Even though Sharpness can measure the quality of high-spatial-frequency components, it should be noted that an increase in this metric can translate into a practical image improvement only if an accurate sharpness representation of the actual lesion is generated for the clinician. To compute the disparity between the sharpness of an enhanced image and the actual sharpness of the ground truth, we compare the pixel distribution around the lesion boundaries between an enhanced image and its ground truth by a metric called Edge-Contrast Error (ECE), with smaller values indicating an improvement in Sharpness. Let i be an enhanced image with g being its ground truth, the Edge-Contrast Error for image i can be computed as follows:

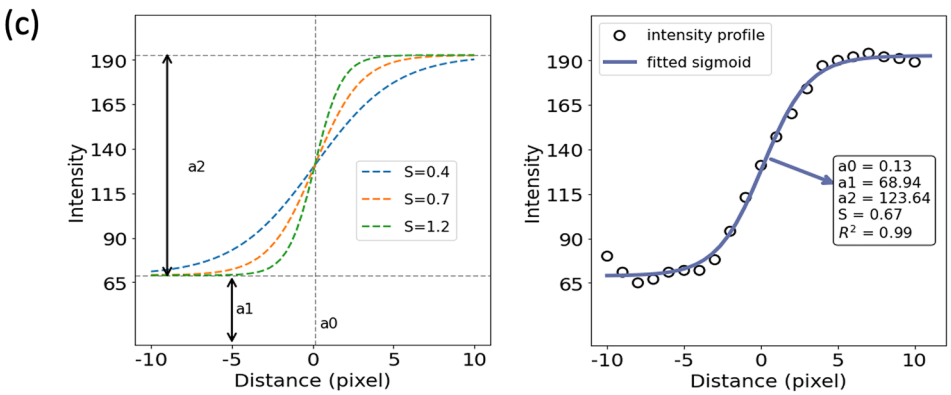
1. Following algorithm (2), extract the 8 intensity profiles for images i and g.
2. Let $P(X, Y)$ be the Y_{th} intensity profile estimated for image X, we have:

$$ECE(i,g) = \frac{1}{N * M} * \sum_{j=1}^{M} \left|\|P(i,j)\|_1 - \|P(g,j)\|_1\right| \tag{10}$$

where M denotes the total number of intensity profiles extracted for the lesion in image i, and N stands for the number of pixels in each intensity profile.


(a)
Lesion centroid
Locate lesion edge points
Map edge points onto the upscaled image
Create line segments
Compute the intensity profile
Sigmoidal Modeling of Edges
Figure 2(b)
1
2
3
4
5
6
7
8


(b)
Intensity
Distance (pixel)


(c)
a2
a1
a0
S=0.4
S=0.7
S=1.2
Intensity
Distance (pixel)
intensity profile
fitted sigmoid
a0 = 0.13
a1 = 68.94
a2 = 123.64
S = 0.67
R² = 0.99

◂**Fig. 2** The lesion sharpness measurement process. **a** The flowchart outlines the steps for sharpness measurement, starting with an enhanced image along with the corresponding lesion map of its ground truth. The last two steps are detailed in **b**. **b** Left: One of the eight potential line segments across the lesion. Right: the corresponding intensity profile p for that line segment. **c** Left: Various representations of the sigmoid function described by Eq. (6) for different values of S. Right: The sigmoid fitted to intensity profile p. The X axis of the graphs in **b** and **c** represents the distance of a point from the center of the line segment

Statistical Analysis

We employed a linear mixed effect model (MEM) analysis for an in-depth exploration of the different evaluation factors' impact on our evaluation criteria. We treated the three aforementioned variables in our study as fixed effect factors: the LR (learning rate) with 2 levels, the perceptual loss weight (W_{per}), with 6 levels, and the adversarial loss weight W_{adv} with 18 levels. In cases where the MEM model indicates significance, a post hoc *Tukey-HSD* test was used to evaluate the statistical significance of pairwise differences between factors at various levels.

Model adequacy checking was performed for all the MEMs and in some cases violations of the normality or/and homoscedasticity assumption were found. To resolve these issues, we searched and found an appropriate response transformation. Specifically, we used the natural logarithm (ln(.)) for MIE, GE, Sharpness, and ECE. Subsequently, the MEM was applied to the transformed response variables and the respective model adequacy checking indicated no assumption violations.

Data Analysis

Our analysis was conducted at two different levels:

(I) **Per W_{per},** which groups the trained networks by W_{per}. This level comprises the analysis of 6 distinct groups.

(II) **Per W_{per}-LR,** which applies to the networks with the same W_{per} and LR, resulting in 12 groups to study.

The analysis pipeline includes comparing the best VGG-Enhanced and VGG+GAN-Enhanced networks of each group using GE as the ranking criterion, where a network with a smaller GE is superior. This metric is specifically chosen for its ability to address structural disparities, a characteristic that other metrics such as MIE and SSIM may not capture effectively. Sharpness is another metric to consider for this matter; however, this property has a lower priority compared to enhancing the shape of the lesion.

Results

Figure 3 and Supplementary Figs. 1–5 illustrate the data distribution for our evaluation criteria. It is apparent that these distributions align consistently within a narrow range. This uniformity, however, presents a challenge in measuring the significance of changes induced by factors that we investigated, as the minimum, maximum, and mean differences between all distributions do not provide enough information about their shape and variability. This issue is addressed by the MEM interaction plots in Figs. 4 and 5 and Supplementary Figs. 6–11.

Following the pipeline outlined in the "Data Analysis" section, we extract the best networks for each group of the two analysis levels. At the first level, for each W_{per}, after identifying the best hyperparameters for each VGG-Enhanced and VGG+GAN-Enhanced network, we illustrate their estimated value across different response variables using red and green rectangles in Fig. 4 and Supplementary Figs. 6–10. The optimal VGG-Enhanced case is shown in red, while the best VGG+GAN-Enhanced is in green. Within each W_{per}, based on Fig. 4a and Supplementary Figs. 6a–10a, it is notable that several VGG+GAN-Enhanced cases yield smaller GE than the best VGG-Enhanced case (red rectangle).

This pattern extends to the second analysis level, where for each combination of W_{per} and LR, there is at least one VGG+GAN-Enhanced network with a smaller GE than that of its VGG-Enhanced case, the optimal case of which is illustrated by the arrow in Fig. 5 and Supplementary Fig. 11.

GE Analysis

Our statistical tests between the optimal cases of each W_{per} (red and green rectangles in Fig. 4a and Supplementary Figs. 6a–10a) reflect a significant reduction in GE values (p-value $<<0.05$ in Table 1). This result is the same at the second level and the best VGG+GAN-Enhanced network of each W_{per} and LR shows a significant reduction in GE (p-value $<<0.05$ in Table 1). The observation at both analysis levels can be confirmed by comparing the structural disparities between representative examples in Fig. 6. For instance, the zoomed area for $W_{per} = 0.7$ in Fig. 6a and b shows that $W_{adv} > 0$ has more low-intensity pixels (a sharp arrow-like shape), i.e., a closer pattern to the ground truth. This pattern in black pixels is identical for other W_{per} values (0.5 and 1.5) which share the same LR in Fig. 6a and b.

At both analysis levels, based on the interaction plots (Fig. 4a and Supplementary Figs. 6a–10a), we also note

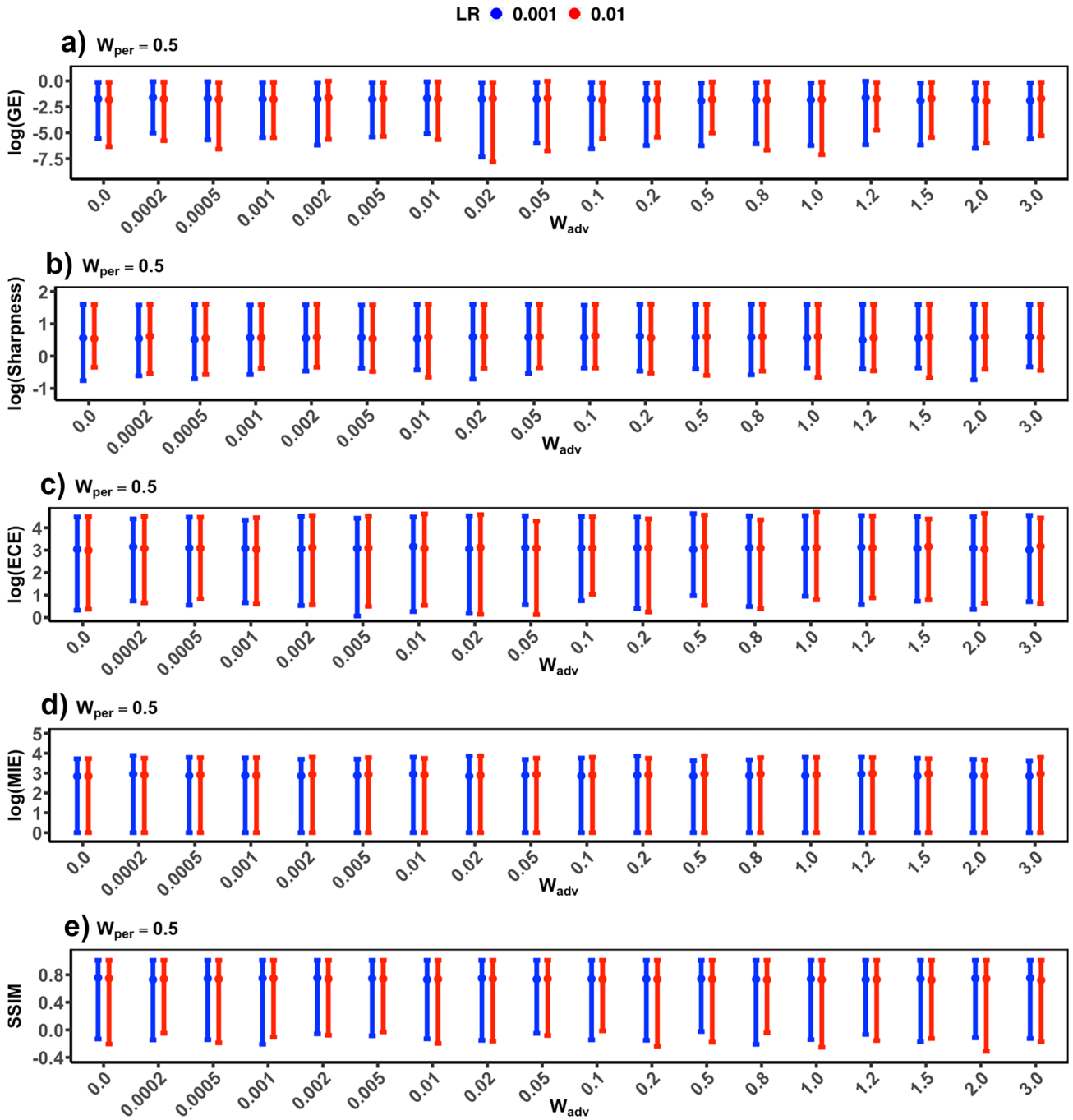


Fig. 3 Data distribution at different W_{adv} and LR combinations when $W_{per} = 0.5$ under (**a**–**e**). **a** log(GE), **b** log(Sharpness), **c** log(ECE), **d** log(MIE), **e** SSIM. Each distribution line is marked with a dot indicating its mean value

that the best VGG+GAN-Enhanced case emerges at large W_{adv} values. This pattern holds true not only within networks of the same W_{per}, but also those within the second level, where most of the optimal solutions feature a W_{adv} greater than 1, with many of them improving at the higher value of 3.

Sharpness Analysis

For all W_{per} values, the Sharpness for the optimal VGG+GAN-Enhanced network is larger than that of the VGG-Enhanced network, according to Fig. 4b and Supplementary Figs. 6b–10b. This increase in Sharpness is

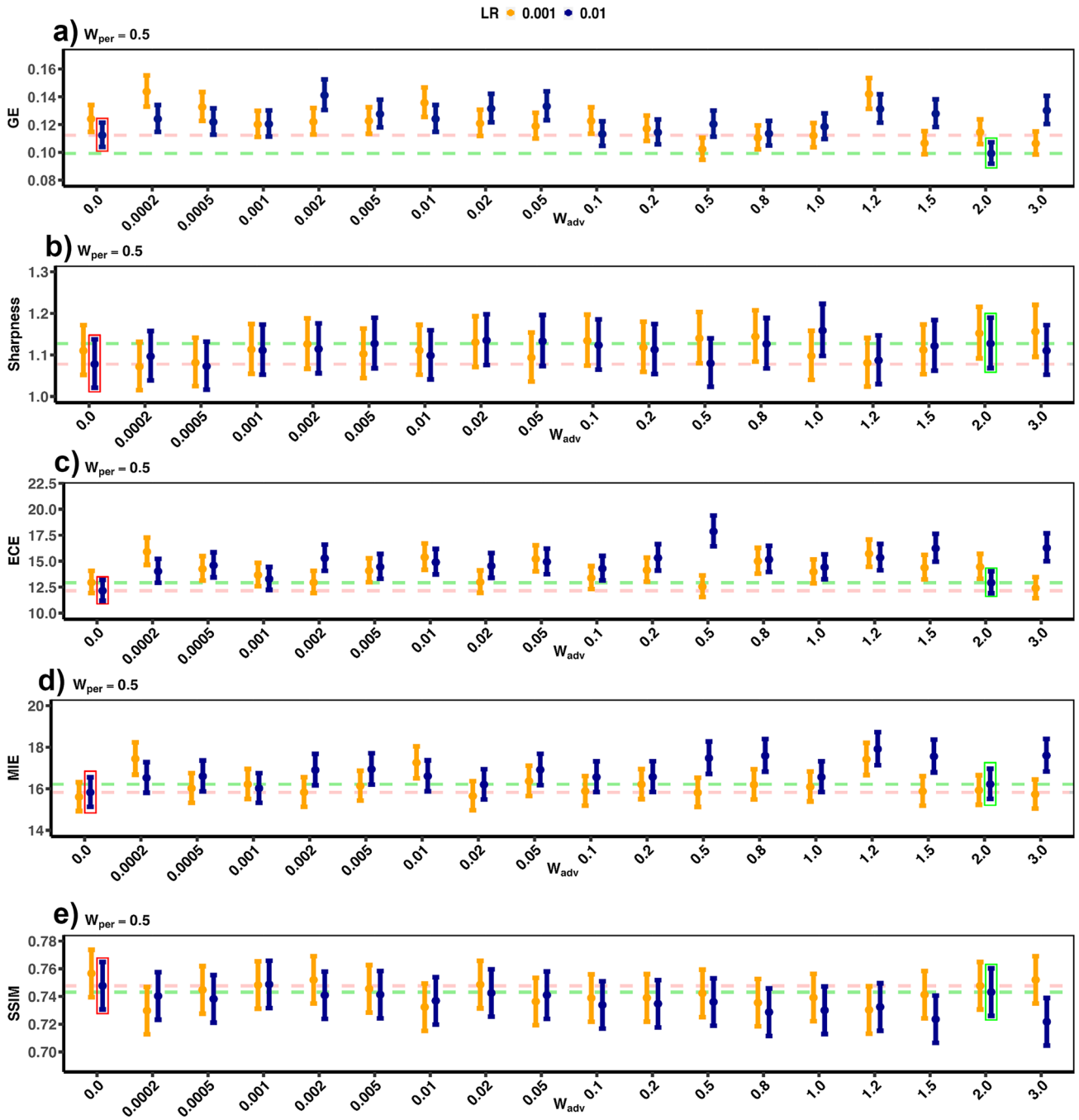


Fig. 4 Interaction plots for the mixed-effects model, indicating how W_{adv} and LR at $W_{per} = 0.5$ interact with each other under (**a**–**e**): **a** GE, **b** Sharpness, **c** ECE, **d** MIE, **e** SSIM. The upper and lower bound at each level show the confidence interval. The estimated value for each level is bulleted with a circle. Also, red and green rectangles point to the confidence interval of a response variable for the optimal network found at $W_{per} = 0.5$

the same for the cases with the same W_{per} and LR combination in Fig. 5b and Supplementary Fig. 11b, in which each optimal VGG+GAN-Enhanced case is predicted to have a higher Sharpness value. According to Table 1, this upward trend using adversarial training is estimated to be statistically significant, but does not occur in all cases, specifically when the networks are trained with a W_{per} value of 0.001.

This difference can be confirmed by a visual inspection of the representative examples in Fig. 6, and our case-by-case analysis reflects a discernible edge contrast difference upon the inclusion of adversarial loss, though not across all the

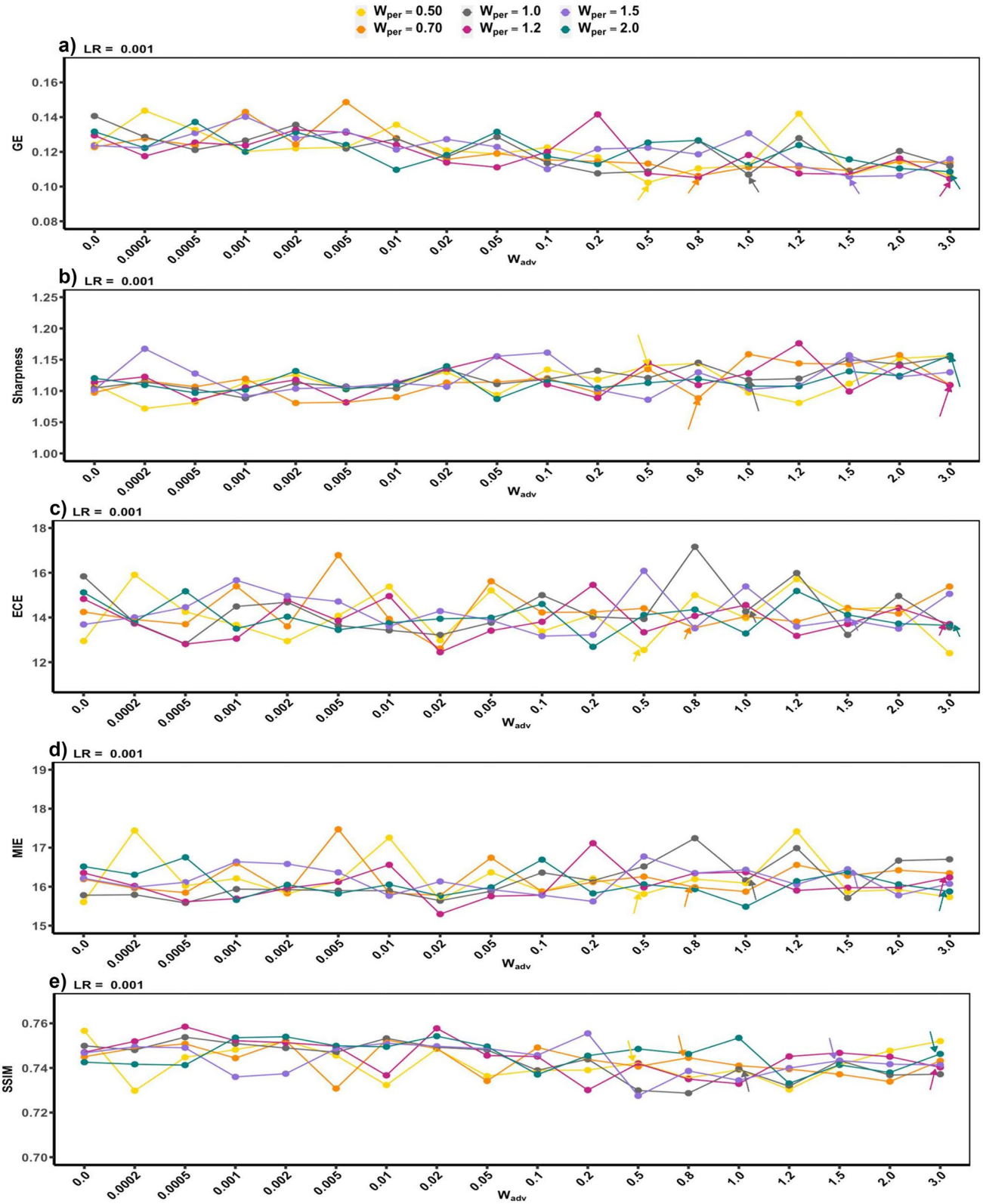


Fig. 5 Interaction plots for the mixed-effects model, indicating how W_{adv} and W_{per} at $LR = 0.001$ interact with each other under (**a**–**e**): **a** GE, **b** Sharpness, **c** ECE, **d** MIE, **e** SSIM. The estimated value for each level is bulleted with a circle. The optimal W_{adv} for each W_{per} is pointed with an arrow

Table 1 *F*-ratios and *p*-values (in parenthesis) to compare the top-performing VGG-Enhanced versus VGG+GAN-Enhanced networks across distinct response variables. For $Wper = 0.7$, the table contains 3 sets of learning rates compared, which is because this *Wper* value results in networks with different learning rates at the first analysis level

W_{per}	VGG-Enhanced	VGG+GAN-Enhanced		*F*-ratio (*p*-value)				
	LR	W_{adv}	LR	GE	Sharpness	ECE	MIE	SSIM
0.5	0.001	0.5	0.001	**45.01 (0.000)**	2.09 (0.149)	0.63 (0.425)	1.99 (0.158)	**15.13 (0.000)**
	0.01	2.0	0.01	**15.06 (0.000)**	5.46 (0.202)	2.35 (0.125)	**4.71 (0.030)**	0.98 (0.321)
0.7	0.001	1.0	0.001	**30.04 (0.000)**	0.318 (0.573)	2.53 (0.114)	2.22 (0.136)	0.01 (0.895)
	0.01	1.0	0.01	**61.65 (0.000)**	**9.97 (0.001)**	**12.79 (0.000)**	1.53 (0.215)	0.00 (0.976)
	0.001	1.0	0.01	**26.15 (0.000)**	**10.79 (0.001)**	0.34 (0.552)	**8.07 (0.004)**	1.84 (0.174)
1.0	0.001	1.0	0.001	**94.67 (0.000)**	0.59 (0.439)	**8.08 (0.004)**	**9.72 (0.001)**	**9.84 (0.001)**
	0.01	1.0	0.01	**32.28 (0.000)**	**21.33 (0.000)**	1.77 (0.184)	**8.32 (0.004)**	**10.67 (0.001)**
1.2	0.001	3.0	0.001	**57.85 (0.000)**	0.07 (0.784)	**5.79 (0.016)**	0.41 (0.517)	3.67 (0.056)
	0.01	1.5	0.01	**81.67 (0.000)**	**16.25 (0.000)**	0.002 (0.961)	1.12 (0.288)	1.55 (0.213)
1.5	0.001	1.5	0.001	**36.75 (0.000)**	**8.32 (0.004)**	0.20 (0.648)	**3.44 (0.064)**	0.97 (0.323)
	0.01	3.0	0.01	**24.31 (0.000)**	**9.86 (0.001)**	2.09 (0.149)	1.09 (0.296)	**9.70 (0.001)**
2.0	0.001	3.0	0.001	**51.81 (0.000)**	3.43 (0.064)	**9.54 (0.002)**	**17.40 (0.000)**	0.97 (0.323)
	0.01	1.5	0.01	**18.55 (0.000)**	0.60 (0.436)	**4.59 (0.033)**	**36.78 (0.000)**	**16.52 (0.000)**

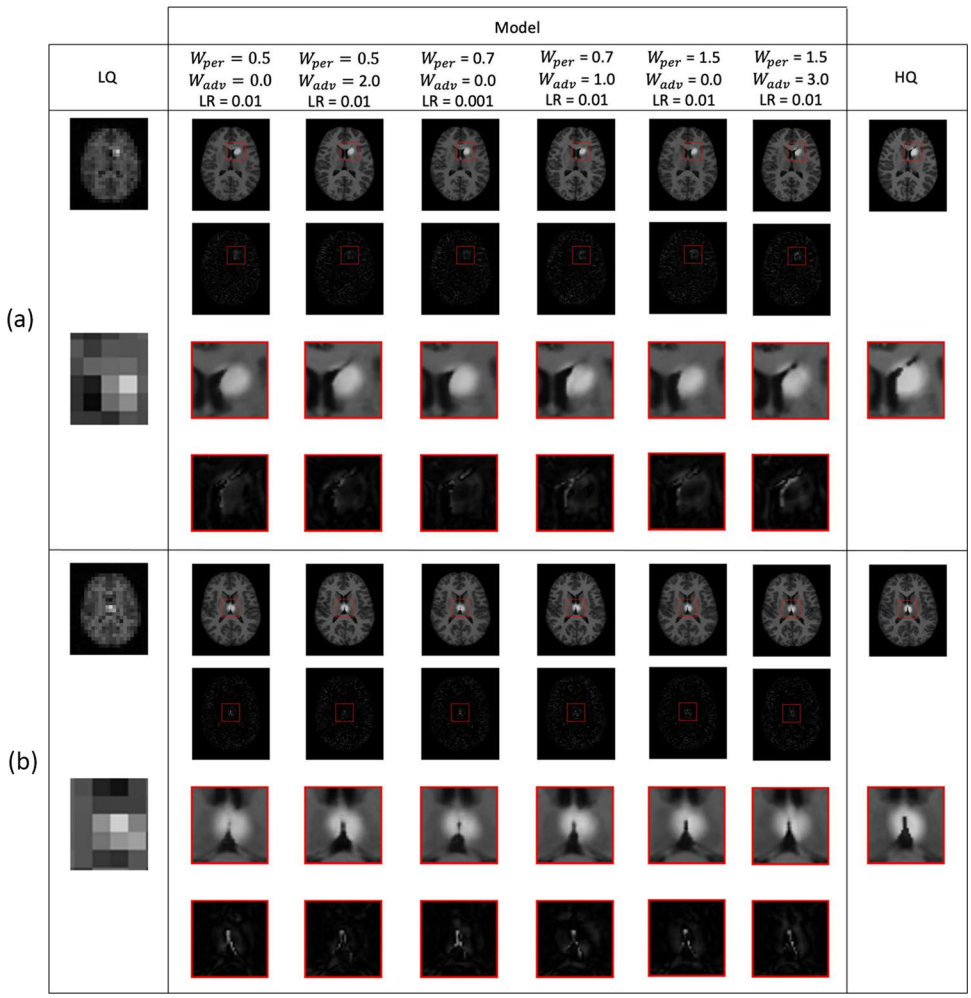


Fig. 6 Representative examples of enhanced images from LQ samples using various types of networks and learning rates. In the examples provided, VGG+GAN-Enhanced cases have a closer representation of the ground truth

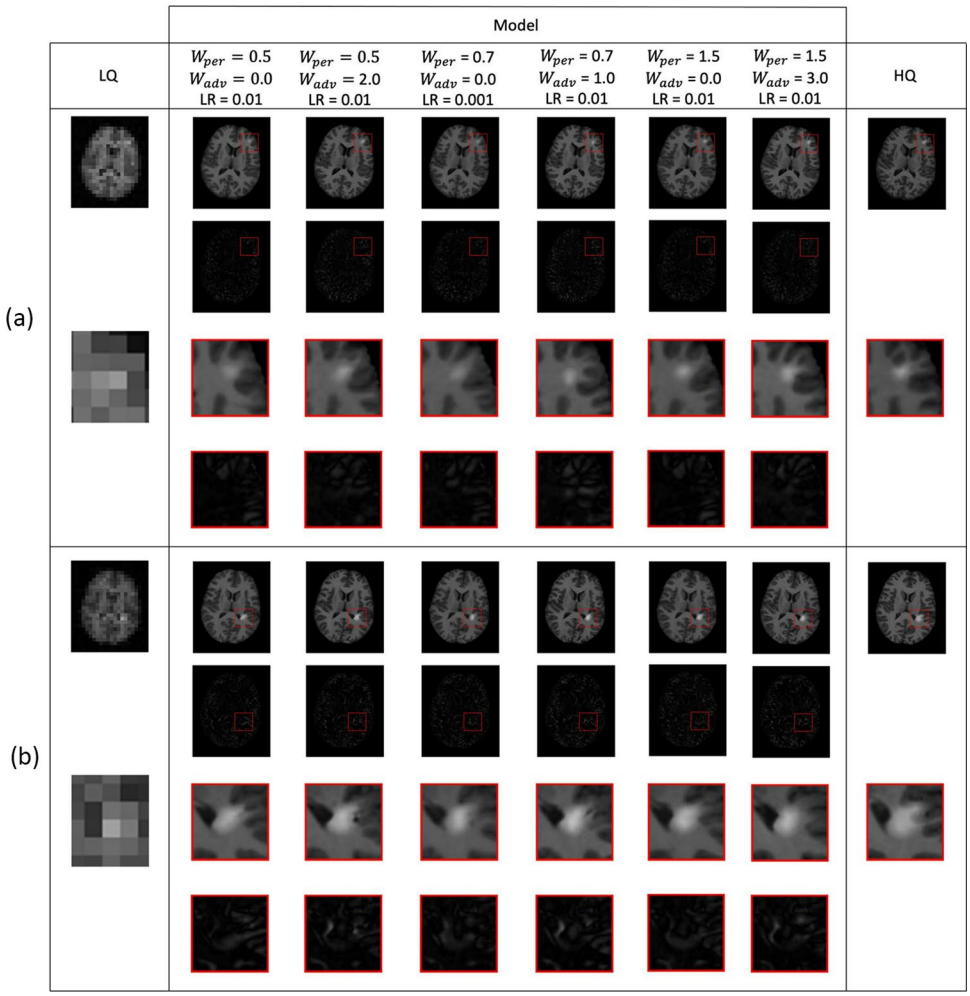


Fig. 7 Representative examples of enhanced images from LQ samples using various types of networks and learning rates. In the examples provided, VGG+GAN-Enhanced cases appear sharper

examples. This is verified based on the examples provided in Fig. 7a and b, where the edge contrast for the lesion map of VGG+GAN-Enhanced cases is higher than that of VGG-Enhanced cases. However, the examples in Fig. 6a and b do not show this effect, where all the lesions have a similar sharpness.

ECE Analysis

As shown in Fig. 4c and Supplementary Figs. 6c–10c, the optimal VGG+GAN-Enhanced networks found for each W_{per} yield a reduction in ECE when $0.7 \leq W_{per} \leq 1.2$, though not necessarily statistically significant, as outlined in Table 1. When using a fixed W_{per} and LR (second analysis level), the $LR = 0.001$ illustrates smaller ECE values for all the VGG+GAN-Enhanced cases. This is, nonetheless, different when using an $LR = 0.01$, and the optimal W_{adv} value found for some W_{per} yields an increase in ECE. In terms of significance, Table 1 shows that some of these variations are substantial (p-value $<<0.05$), though not consistent for a specific W_{per}, LR, or their combination. However, we found that each significant Sharpness variation comes with either a significant ECE reduction (p-value $<<0.05$) or a non-significant ECE jump (p-value >0.05). Also, nearly all the non-significant changes in Sharpness (p-value >0.05) correspond to a reduction in ECE.

MIE and SSIM Analysis

The findings across Fig. 4d and Supplementary Figs. 6d–10d indicate that a significant portion of optimal VGG+GAN-Enhanced scenarios exhibit higher MIE values. This performance decrease is similar in terms of SSIM, as observed in Fig. 4e and Supplementary Figs. 6e–10e, where the estimated SSIM decreases with the incorporation of adversarial training but not significantly over all the cases according to Table 1.

Discussion

The key findings of this work suggest that adversarial training can significantly contribute to optimizing an MRI SR network, enabling it to output structurally enhanced images that more closely represent the ground truth. This

conclusion holds true across both the visual inspection of Fig. 6 and the statistical analysis of metrics. In each network group categorized by W_{per}, or both W_{per} and LR, there is at least one VGG+GAN-Enhanced solution with a smaller GE compared to the optimal VGG-Enhanced network of that group. These reductions in GE, p-values $<<0.05$ in Table 1, are found to be statistically significant based on the used mixed effect model which is the most appropriate statistical test to account for the repeated measurement design used in our experiments.

Moreover, the VGG+GAN-Enhanced solutions of all groups exhibit higher Sharpness; however, their significance depends on the particular W_{per} and LR used. This increase in Sharpness, as mentioned before in the "Evaluation Metrics" section, needs further exploration to assess whether it actually translates into an improvement in lesion characterization or not. As evidenced by the p-values in Table 1, all the VGG+GAN-Enhanced images experienced either a non-significant rise or a significant reduction in ECE, meaning that the increased sharpness does not appear to interfere with the distribution of pixel intensities around lesion edges. This suggests that adversarial training enhances sharpness without compromising edge preservation; this is the manifestation of an improvement in sharpness.

It should be noted that while adversarial training improves perceptual quality, our observation in the "Results" section reflects that it may degrade performance on MIE and SSIM. This suggests that the use of these two metrics to distinguish subtle structural variations is not as powerful as commonly perceived, providing support to the hypothesis that they lack inherent correlation with the human perceptual system [48]. Measuring structural differences and matching them with our visual inspection was possible by incorporating GE, the idea of which is derived from the perceptually motivated Flip metric, a novel metric that is built on the principles of human perception. This emphasizes the conclusion that the community needs to come up with a new set of evaluation criteria for tracking very small changes, specifically when the input is at a high degradation order. Based on our experiments, relying only on SSIM and MIE would have led to concluding a negative impact of adversarial training on performance, which contradicts our actual conclusion.

Our findings also indicate that the optimal VGG+GAN-Enhanced solution corresponds to larger adversarial loss weights: however, not every W_{adv} value leads to an improved performance. This statement aligns with our statistics in Fig. 4a and Supplementary Figs. 6a–11a, where for some W_{adv} the GE value non-significantly decreases or even increases. For $W_{per} = 1.0$ in Supplementary Fig. 8 as an example, the network with $\{W_{adv} = 0.2$ and $LR = 0.01\}$ is neither the optimal solution nor does it have a smaller GE output compared to the $\{W_{adv} = 0.0$ and $LR = 0.01\}$.

It is also notable that a systematic search over W_{per}, W_{adv}, and LR is essentially required for training, as an optimal hyperparameter value such as learning rate found for one network is not necessarily extendable to another network, and each case needs to be optimized within its own specific hyperparameter search space. The network with $W_{per} = 0.7$ is an example in this case, where the $LR = 0.001$ is optimal for the best VGG-Enhanced case, while for $W_{adv} > 0$ the optimum has a faster learning rate ($LR = 0.01$). Also, for the cases with the same W_{per} and LR, the positive impact of adversarial loss suggests that proceeding with the optimization of a pre-trained VGG-Enhanced network using adversarial training would lead to a better answer, although not necessarily the global one.

This adaptive weight tuning over a systematic search also offers the ability to dynamically balance the preservation of pixel intensity and structural details based on clinical requirements. In certain applications, accurately reconstructing pixel intensities may be the priority to enable visualization of specific intensity patterns [78–80]. In other clinical scenarios, preserving the overall structural and delineation of anatomical boundaries takes precedence over pixel-wise intensity values [81, 82]. The adaptive loss weighting strategy allows radiologists to negotiate this tradeoff optimally before any deployments in the clinical flow.

For instance, in a functional image where monitoring metabolic activity and perfusion in tumors is crucial [83], a higher emphasis on the pixel loss term can enhance the depiction of subtle intensity variations associated with contrast accumulation. On the other hand, in radiation therapy where the radiation needs to target tumor and minimize damage to the surrounding tissue [84], the model can be tuned to prioritize the perceptual and adversarial losses, yielding images with improved sharpness and edge preservation, enabling better characterization of pathological findings in relation to surrounding anatomy.

The clinical application of this work focuses on highlighting the fact that one model does not fit all clinical problems. For any patient, a clinical decision is made on their specific case. Similarly, we suggest that based on the requirements and diagnosis, one model may be a better choice over another one and there is no model/hyperparameter that fits all solution. This work creates a foundation for future research to select model parameters with more than one aspect of the evaluation metric that can be used in a clinical setting. These include, but are not limited to, pixel-level quantification, structural definition of region of interest, edge contrast, and overall image quality to name a few. Future practice must use multiple evaluation metrics along with advanced statistical evaluation methods to find the correct

tradeoffs between the best method under fixed variables of evaluation metrics and clinical diagnosis.

The impact of this systematic hyperparameter search, however, is particularly noticeable when working with highly degraded data, which allowed for more substantial quality improvement and easier measurement of performance variations. In scenarios with input data of relatively high quality, the abundance of spatial frequency information may make the incorporation of new methods less impactful, and the integration of pixel and perceptual losses might suffice for considerable enhancement.

Future Works

Future research could focus on utilizing datasets with realistic lesion maps (e.g., BRaTS [85]) that have more complex properties, and generating LQ data that consider the physical attributes of MR scanners to better emulate real-world scenarios. Additionally, exploring deeper network architectures for the GAN generator may yield further improvements, although achieving stronger structural enhancement primarily relies on the quality of the LQ data. Furthermore, extending this research to include other modalities such as CT scans or conducting similar experiments on different body parts can provide more insights into the generalizability of our findings on the applicability of GANs in medical image enhancement.

Conclusion

We explored the influence of adversarial training within an MR image enhancement pipeline, with REAL-ESRGAN as the foundational GAN architecture and UNet as its generator. This effect was specifically examined within the lesion, a region of clear importance for medical applications. A mixed effect model was utilized for statistical inference, allowing us to study the effect of adversarial training as a factor on our evaluation metrics. Our results demonstrated that adversarial training significantly assists pixel and perceptual loss functions in finding solutions with further enhancements in terms of structure, sharpness, and contrast, both visually and quantitatively. This conclusion was found by transitioning beyond the conventional metrics used in MRI image enhancement studies, addressing the necessity of adopting alternative metrics capable of more accurately capturing visual disparities.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10278-024-01205-8.

Author Contribution Mohammad Javadi and Rishabh Sharma share equal credit for their contributions to the creation of this paper.

Declarations

Ethics Approval The dataset owner mandated the signing of a copyright agreement for research purposes in this study. Ethics approval was not necessary beyond this requirement.

Conflict of Interest The authors declare no competing interests.

References

1. Peters DC, Korosec FR, Grist TM, et al. Undersampled projection reconstruction applied to MR angiography. *Magn Reson Med*. 2000;43(1):91-101. https://doi.org/10.1002/(SICI)1522-2594(200001)43:1<91::AID-MRM11>3.0.CO;2-4
2. Hu R, Kleimaier D, Malzacher M, Hoesl MAU, Paschke NK, Schad LR. X-nuclei imaging: Current state, technical challenges, and future directions. *Journal of Magnetic Resonance Imaging*. 2020;51(2):355-376. https://doi.org/10.1002/jmri.26780
3. Yang H, Wang Z, Liu X, Li C, Xin J, Wang Z. Deep learning in medical image super resolution: a review. *Applied Intelligence*. 2023;53(18):20891-20916. https://doi.org/10.1007/s10489-023-04566-9
4. Sandino CM, Cheng JY, Chen F, Mardani M, Pauly JM, Vasanawala SS. Compressed Sensing: From Research to Clinical Practice with Deep Neural Networks: Shortening Scan Times for Magnetic Resonance Imaging. *IEEE Signal Process Mag*. 2020;37(1):117-127. https://doi.org/10.1109/MSP.2019.2950433
5. Qiu D, Cheng Y, Wang X. Medical image super-resolution reconstruction algorithms based on deep learning: A survey. *Comput Methods Programs Biomed*. 2023;238. https://doi.org/10.1016/j.cmpb.2023.107590
6. Chen R, Tang X, Zhao Y, et al. Single-frame deep-learning super-resolution microscopy for intracellular dynamics imaging. *Nat Commun*. 2023;14(1). https://doi.org/10.1038/s41467-023-38452-2
7. Kim YB, Van Le T, Lee JY. Lightweight brain MR image super-resolution using 3D convolution. *Multimed Tools Appl*. 2024;83(3):8785-8795. https://doi.org/10.1007/s11042-023-15969-8
8. Chen Y, Shi F, Christodoulou AG, Xie Y, Zhou Z, Li D. Efficient and accurate MRI super-resolution using a generative adversarial network and 3D multi-level densely connected network. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. ; 2018:91–99.
9. Lyu Q, Shan H, Steber C, et al. Multi-Contrast Super-Resolution MRI through a Progressive Network. *IEEE Trans Med Imaging*. 2020;39(9):2738-2749. https://doi.org/10.1109/TMI.2020.2974858
10. Guerreiro J, Tomás P, Garcia N, Aidos H. Super-resolution of magnetic resonance images using Generative Adversarial Networks. *Computerized Medical Imaging and Graphics*. 2023;108. https://doi.org/10.1016/j.compmedimag.2023.102280
11. Wang Q, Mahler L, Steiglechner J, Birk F, Scheffler K, Lohmann G. DISGAN: Wavelet-informed Discriminator Guides GAN to MRI Super-resolution with Noise Cleaning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. ; 2023:2452–2461.
12. Huang S, Liu X, Tan T, et al. TransMRSR: transformer-based self-distilled generative prior for brain MRI super-resolution. *Visual Computer*. 2023;39(8):3647-3659. https://doi.org/10.1007/s00371-023-02938-3
13. Li G, Lv J, Tian Y, et al. Multicontrast MRI Super-Resolution via Transformer-Empowered Multiscale Contextual Matching

and Aggregation. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol 2022-June. IEEE Computer Society; 2022:20604–20613. https://doi.org/10.1109/CVPR52688.2022.01998
14. Feng CM, Yan Y, Yu K, et al. Exploring separable attention for multi-contrast MR image super-resolution. *IEEE Trans Neural Netw Learn Syst*. Published online 2024.
15. Pawar K, Chen Z, Shah NJ, Egan GF. Suppressing motion artefacts in MRI using an Inception-ResNet network with motion simulation augmentation. *NMR Biomed*. 2022;35(4). https://doi.org/10.1002/nbm.4225
16. Muckley MJ, Ades-Aron B, Papaioannou A, et al. Training a neural network for Gibbs and noise removal in diffusion MRI. *Magn Reson Med*. 2021;85(1):413-428. https://doi.org/10.1002/mrm.28395
17. Park S, Gach HM, Kim S, Lee SJ, Motai Y. Autoencoder-Inspired Convolutional Network-Based Super-Resolution Method in MRI. *IEEE J Transl Eng Health Med*. 2021;9. https://doi.org/10.1109/JTEHM.2021.3076152
18. Yu M, Guo M, Zhang S, et al. RIRGAN: An end-to-end lightweight multi-task learning method for brain MRI super-resolution and denoising. *Comput Biol Med*. 2023;167. https://doi.org/10.1016/j.compbiomed.2023.107632
19. Zou B, Ji Z, Zhu C, Dai Y, Zhang W, Kui X. Multi-scale deformable transformer for multi-contrast knee MRI super-resolution. *Biomed Signal Process Control*. 2023;79. https://doi.org/10.1016/j.bspc.2022.104154
20. Song J, Yi H, Xu W, Li X, Li B, Liu Y. ESRGAN-DP: Enhanced super-resolution generative adversarial network with adaptive dual perceptual loss. *Heliyon*. 2023;9(4). https://doi.org/10.1016/j.heliyon.2023.e15134
21. Wang Q, Mahler L, Steiglechner J, Birk F, Scheffler K, Lohmann G. A three-player gan for super-resolution in magnetic resonance imaging. In: *International Workshop on Machine Learning in Clinical Neuroimaging*. ; 2023:23–33.
22. Wang X, Xie L, Dong C, Shan Y. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. ; 2021:1905–1914.
23. Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*. ; 2016:694–711.
24. Zhu X, Zhang L, Zhang L, et al. GAN-Based Image Super-Resolution with a Novel Quality Loss. *Math Probl Eng*. 2020;2020. https://doi.org/10.1155/2020/5217429
25. Zhang Y, Liu S, Dong C, Zhang X, Yuan Y. Multiple cycle-in-cycle generative adversarial networks for unsupervised image super-resolution. *IEEE Transactions on Image Processing*. 2020;29:1101-1112. https://doi.org/10.1109/TIP.2019.2938347
26. Yang Q, Liu Y, Yang J. Two-branch crisscross network for realistic and accurate image super-resolution. *Displays*. 2023;80. https://doi.org/10.1016/j.displa.2023.102549
27. Wang X, Yu K, Wu S, et al. Esrgan: Enhanced super-resolution generative adversarial networks. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. ; 2018:0.
28. Park J, Son S, Lee KM. Content-aware local GAN for photo-realistic super-resolution. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. ; 2023:10585–10594.
29. Chen D, Liang J, Zhang X, Liu M, Zeng H, Zhang L. Human guided ground-truth generation for realistic image super-resolution. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. ; 2023:14082–14091.
30. Altun Güven S, Talu MF. Brain MRI high resolution image creation and segmentation with the new GAN method. *Biomed Signal Process Control*. 2023;80. https://doi.org/10.1016/j.bspc.2022.104246
31. Li H, Xuan Z, Zhou J, Hu X, Yang B. Fast and accurate super-resolution of MR images based on lightweight generative adversarial network. *Multimed Tools Appl*. 2023;82(2):2465-2487. https://doi.org/10.1007/s11042-022-13326-9
32. de Farias EC, di Noia C, Han C, Sala E, Castelli M, Rundo L. Impact of GAN-based lesion-focused medical image super-resolution on the robustness of radiomic features. *Sci Rep*. 2021;11(1). https://doi.org/10.1038/s41598-021-00898-z
33. Umirzakova S, Ahmad S, Khan LU, Whangbo T. Medical image super-resolution for smart healthcare applications: A comprehensive survey. *Information Fusion*. 2024;103. https://doi.org/10.1016/j.inffus.2023.102075
34. Wicaksono KP, Fujimoto K, Fushimi Y, et al. Super-resolution application of generative adversarial network on brain time-of-flight MR angiography: image quality and diagnostic utility evaluation. Published online 2022. https://doi.org/10.1007/s00330-022-09103-9/Published
35. Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. ; 2017:4681–4690.
36. Chen Y, Christodoulou AG, Zhou Z, Shi F, Xie Y, Li D. MRI super-resolution with GAN and 3D multi-level DenseNet: smaller, faster, and better. *arXiv preprint arXiv:200301217*. Published online 2020.
37. Borji A. Pros and cons of GAN evaluation measures: New developments. *Computer Vision and Image Understanding*. 2022;215:103329.
38. Wang Y, Hu Y, Yu J, Zhang J. Gan prior based null-space learning for consistent super-resolution. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol 37. ; 2023:2724–2732.
39. Sharma R, Tsiamyrtzis P, Webb AG, Leiss EL, Tsekos N V. Learning to deep learning: statistics and a paradigm test in selecting a UNet architecture to enhance MRI. *Magnetic Resonance Materials in Physics, Biology and Medicine*. Published online 2023. https://doi.org/10.1007/s10334-023-01127-6
40. Adam SP, Alexandropoulos SAN, Pardalos PM, Vrahatis MN. No free lunch theorem: A review. In: *Springer Optimization and Its Applications*. Vol 145. Springer International Publishing; 2019:57–82. https://doi.org/10.1007/978-3-030-12767-1_5
41. Kendall A, Gal Y, Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. ; 2018:7482–7491.
42. Wolpert DH, Macready WG. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*. 1997;1(1):67-82.
43. Sharma R, Tsiamyrtzis P, Webb AG, et al. A Deep Learning Approach to Upscaling "Low-Quality" MR Images: An In Silico Comparison Study Based on the UNet Framework. *Applied Sciences (Switzerland)*. 2022;12(22). https://doi.org/10.3390/app122211758
44. Islam KT, Zhong S, Zakavi P, et al. Improving portable low-field MRI image quality through image-to-image translation using paired low- and high-field images. *Sci Rep*. 2023;13(1). https://doi.org/10.1038/s41598-023-48438-1
45. de Leeuw den Bouter ML, Ippolito G, O'Reilly TPA, Remis RF, van Gijzen MB, Webb AG. Deep learning-based single image super-resolution for low-field MR brain images. *Sci Rep*. 2022;12(1). https://doi.org/10.1038/s41598-022-10298-6

46. Koonjoo N, Zhu B, Bagnall GC, Bhutto D, Rosen MS. Boosting the signal-to-noise of low-field MRI with deep learning image reconstruction. *Sci Rep*. 2021;11(1). https://doi.org/10.1038/s41598-021-87482-7
47. Lin H, Figini M, Tanno R, et al. Deep learning for low-field to high-field MR: image quality transfer with probabilistic decimation simulator. In: *Machine Learning for Medical Image Reconstruction: Second International Workshop, MLMIR 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 2*. ; 2019:58–70.
48. Nilsson J, Akenine-Möller T. Understanding ssim. *arXiv preprint arXiv:200613846*. Published online 2020.
49. Ahmad R, Ding Y, Simonetti OP. Edge sharpness assessment by parametric modeling: Application to magnetic resonance imaging. *Concepts Magn Reson Part A Bridg Educ Res*. 2015;44(3):138-149. https://doi.org/10.1002/cmr.a.21339
50. Ren S, Jain DK, Guo K, Xu T, Chi T. Towards efficient medical lesion image super-resolution based on deep residual networks. *Signal Process Image Commun*. 2019;75:1-10. https://doi.org/10.1016/j.image.2019.03.008
51. Zhu J, Yang G, Lio P. How can we make GAN perform better in single medical image super-resolution? A lesion focused multi-scale approach. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. ; 2019:1669–1673.
52. Do WJ, Seo S, Han Y, Ye JC, Choi SH, Park SH. Reconstruction of multicontrast MR images through deep learning. *Med Phys*. 2020;47(3):983-997. https://doi.org/10.1002/mp.14006
53. Yang G, Yu S, Dong H, et al. DAGAN: Deep De-Aliasing Generative Adversarial Networks for Fast Compressed Sensing MRI Reconstruction. *IEEE Trans Med Imaging*. 2018;37(6):1310-1321. https://doi.org/10.1109/TMI.2017.2785879
54. Luo G, Zhao N, Jiang W, Hui ES, Cao P. MRI reconstruction using deep Bayesian estimation. *Magn Reson Med*. 2020;84(4):2246-2261. https://doi.org/10.1002/mrm.28274
55. Liashchynskyi P, Liashchynskyi P. Grid search, random search, genetic algorithm: a big comparison for NAS. *arXiv preprint arXiv:191206059*. Published online 2019.
56. Shekar BH, Dagnew G. Grid search-based hyperparameter tuning and classification of microarray cancer data. In: *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*. ; 2019:1–8.
57. Cuocolo R, Comelli A, Stefano A, et al. Deep Learning Whole-Gland and Zonal Prostate Segmentation on a Public MRI Dataset. *Journal of Magnetic Resonance Imaging*. 2021;54(2):452-459. https://doi.org/10.1002/jmri.27585
58. Wahlang I, Maji AK, Saha G, et al. Brain Magnetic Resonance Imaging Classification Using Deep Learning Architectures with Gender and Age. *Sensors*. 2022;22(5). https://doi.org/10.3390/s22051766
59. Schading S, Seif M, Leutritz T, et al. Reliability of spinal cord measures based on synthetic T1-weighted MRI derived from multiparametric mapping (MPM). *Neuroimage*. 2023;271. https://doi.org/10.1016/j.neuroimage.2023.120046
60. Dror R, Shlomov S, Reichart R. Deep dominance-how to properly compare deep neural models. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ; 2019:2773–2785.
61. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:181112808*. Published online 2018.
62. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision*. ; 2017:2223–2232.
63. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. ; 2015:234–241.
64. Ding PLK, Li Z, Zhou Y, Li B. Deep residual dense U-Net for resolution enhancement in accelerated MRI acquisition. In: SPIE-Intl Soc Optical Eng; 2019:14. https://doi.org/10.1117/12.2513158
65. Guan S, Khan AA, Sikdar S, Chitnis P V. Fully Dense UNet for 2-D Sparse Photoacoustic Tomography Artifact Removal. *IEEE J Biomed Health Inform*. 2020;24(2):568-576. https://doi.org/10.1109/JBHI.2019.2912935
66. Masutani EM, Bahrami N, Hsiao A. Deep learning single-frame and multiframe super-resolution for cardiac MRI. *Radiology*. 2020;295(3):552-561. https://doi.org/10.1148/radiol.2020192173
67. Cai S, Tian Y, Lui H, Zeng H, Wu Y, Chen G. Dense-unet: A novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network. *Quant Imaging Med Surg*. 2020;10(6):1275-1285. https://doi.org/10.21037/QIMS-19-1090
68. Chatterjee S, Sarasaen C, Rose G, Nürnberger A, Speck O. DDoS-UNet: Incorporating temporal information using Dynamic Dual-channel UNet for enhancing super-resolution of dynamic MRI. *arXiv preprint arXiv:220205355*. Published online 2022.
69. Chatterjee S, Sciarra A, Dünnwald M, et al. ShuffleUNet: Super resolution of diffusion-weighted MRIs using deep learning. In: *2021 29th European Signal Processing Conference (EUSIPCO)*. ; 2021:940–944.
70. Schonfeld E, Schiele B, Khoreva A. A u-net based discriminator for generative adversarial networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. ; 2020:8207–8216.
71. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: *3rd International Conference on Learning Representations (ICLR 2015)*. ; 2015.
72. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. ; 2009:248–255.
73. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. *Commun ACM*. 2020;63(11):139-144. https://doi.org/10.1145/3422622
74. Marcus DS, Fotenos AF, Csernansky JG, Morris JC, Buckner RL. Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *J Cogn Neurosci*. 2010;22(12):2677-2684.
75. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, d Alché-Buc F, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems*. Vol 32. Curran Associates, Inc.; 2019. https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf
76. Kingma D, Ba J. Adam: A Method for Stochastic Optimization. In: *International Conference on Learning Representations (ICLR)*. ; 2015.
77. Andersson P, Nilsson J, Akenine-Möller T, Oskarsson M, Åström K, Fairchild MD. FLIP: A Difference Evaluator for Alternating Images. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*. 2020;3(2). https://doi.org/10.1145/3406183
78. Wang J, Weygand J, Hwang KP, et al. Magnetic Resonance Imaging of Glucose Uptake and Metabolism in Patients with Head and Neck Cancer. *Sci Rep*. 2016;6. https://doi.org/10.1038/srep30618
79. Reimer P, Schneider G, Schima W. Hepatobiliary contrast agents for contrast-enhanced MRI of the liver: Properties, clinical

development and applications. *Eur Radiol*. 2004;14(4):559-578. https://doi.org/10.1007/s00330-004-2236-1
80. Rivlin M, Perlman O, Navon G. Metabolic brain imaging with glucosamine CEST MRI: in vivo characterization and first insights. *Sci Rep*. 2023;13(1). https://doi.org/10.1038/s41598-023-48515-5
81. Garg N, Choudhry MS, Bodade RM. A review on Alzheimer's disease classification from normal controls and mild cognitive impairment using structural MR images. *J Neurosci Methods*. 2023;384. https://doi.org/10.1016/j.jneumeth.2022.109745
82. Vemuri P, Jack CR. Role of structural MRI in Alzheimer's disease. *Alzheimers Res Ther*. 2010;2:1-10.
83. Lau D, Corrie PG, Gallagher FA. MRI techniques for immunotherapy monitoring. *J Immunother Cancer*. 2022;10(9).
84. Symms M, Jäger HR, Schmierer K, Yousry TA. A review of structural magnetic resonance neuroimaging. *J Neurol Neurosurg Psychiatry*. 2004;75(9):1235-1244. https://doi.org/10.1136/jnnp.2003.032714
85. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;34(10):1993-2024. https://doi.org/10.1109/TMI.2014.2377694