



**Providing Choice & Value**

Generic CT and MRI Contrast Agents



**FRESENIUS  
KABI**

**CONTACT REP**

# AJNR

This information is current as  
of July 29, 2025.

## **Leveraging Physics-Based Synthetic MR Images and Deep Transfer Learning for Artifact Reduction in Echo-Planar Imaging**





Catalina Raymond, Jingwen Yao, Bryan Clifford, Thorsten  
Feiweier, Sonoko Oshima, Donatello Telesca, Xiaodong  
Zhong, Heiko Meyer, Richard G. Everson, Noriko Salamon,  
Timothy F. Cloughesy and Benjamin M. Ellingson

*AJNR Am J Neuroradiol* 2025, 46 (4) 733-741

doi: <https://doi.org/10.3174/ajnr.A8566>

<http://www.ajnr.org/content/46/4/733>

# Leveraging Physics-Based Synthetic MR Images and Deep Transfer Learning for Artifact Reduction in Echo-Planar Imaging

 Catalina Raymond,  Jingwen Yao,  Bryan Clifford,  Thorsten Feiweier, Sonoko Oshima, Donatello Telesca, Xiaodong Zhong,  Heiko Meyer, Richard G. Everson,  Noriko Salamon,  Timothy F. Cloughesy, and  Benjamin M. Ellingson



## ABSTRACT

**BACKGROUND AND PURPOSE:** This study utilizes a physics-based approach to synthesize realistic MR artifacts and train a deep learning generative adversarial network (GAN) for use in artifact reduction on EPI, a crucial neuroimaging sequence with high acceleration that is notoriously susceptible to artifacts.

**MATERIALS AND METHODS:** A total of 4,573 anatomical MR sequences from 1,392 patients undergoing clinically indicated MRI of the brain were used to create a synthetic data set using physics-based, simulated artifacts commonly found in EPI. By using multiple MRI contrasts, we hypothesized the GAN would learn to correct common artifacts while preserving the inherent contrast information, even for contrasts the network has not been trained on. A modified *Pix2PixGAN* architecture with an *Attention-R2UNet* generator was used for the model. Three training strategies were employed: (1) An “all-in-one” model trained on all the artifacts at once; (2) a set of “single models”, one for each artifact; and a (3) “stacked transfer learning” approach where a model is first trained on one artifact set, then this learning is transferred to a new model and the process is repeated for the next artifact set. Lastly, the “Stacked Transfer Learning” model was tested on ADC maps from single-shot diffusion MRI data in  $N = 49$  patients diagnosed with recurrent glioblastoma to compare visual quality and lesion measurements between the natively acquired images and AI-corrected images.

**RESULTS:** The “stacked transfer learning” approach had superior artifact reduction performance compared to the other approaches as measured by Mean Squared Error (MSE = 0.0016), Structural Similarity Index (SSIM = 0.92), multiscale SSIM (MS-SSIM = 0.92), peak signal-to-noise ratio (PSNR = 28.10), and Hausdorff distance (HAUS = 4.08mm), suggesting that leveraging pre-trained knowledge and sequentially training on each artifact is the best approach this application. In recurrent glioblastoma, significantly higher visual quality was observed in model predicted images compared to native images, while quantitative measurements within the tumor regions remained consistent with non-corrected images.

**CONCLUSIONS:** The current study demonstrates the feasibility of using a physics-based method for synthesizing a large data set of images with realistic artifacts and the effectiveness of utilizing this synthetic data set in a “stacked transfer learning” approach to training a GAN for reduction of EPI-based artifacts.

**ABBREVIATIONS:** BTIP = brain tumor imaging protocol; GAN = generative adversarial network; HAUS = Hausdorff distance; MS-SSIM = multiscale structural similarity index; MSE = mean square error; NAWM = Normal Appearing White Matter; PSNR = peak signal to noise ratio; RANO = Response Assessment in Neuro Oncology; RAS = Right, Anterior, Superior; SSIM = structural similarity index

EPI is one of the most widely used MRI pulse sequences for neuroimaging applications due to its high efficiency of image acquisition up to 10 times faster than conventional MRI sequences.

Received August 5, 2024; accepted after revision October 1.

From the UCLA Brain Tumor Imaging Laboratory (C.R., J.Y., S.O., B.M.E.), Departments of Radiological Sciences (C.R., J.Y., S.O., X.Z., N.S., B.M.E.), Neurosurgery (R.G.E.), Neurology (T.F.C.), and Psychiatry and Biobehavioral Sciences (B.M.E.), David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California; Siemens Medical Solutions USA, Inc. (B.C.), Los Angeles, California; Siemens Healthineers AG (T.F., H.M.), Erlangen, Germany; and Departments of Biostatistics (D.T.), and Bioengineering (X.Z., B.M.E.), Henry Samueli School of Engineering and Applied Science, University of California, Los Angeles, Los Angeles, California.

Funding Information: Siemens Healthcare (Ellingson, Raymond), NIH NCI R01CA270027 (Ellingson, Cloughesy), NIH NCI R01CA279984 (Ellingson), DoD CDMRP CA220732 (Ellingson, Cloughesy), NIH NCI P50CA211015 (Ellingson, Cloughesy).

This acceleration facilitates the measurement of essential parameters, encompassing perfusion, microstructural, functional, and other physiologic aspects in clinically feasible time frames. Unfortunately, this high acquisition efficiency can come at a significant cost, as EPI is prone to various imaging, affecting the quality and diagnostic utility of the images. Key artifacts include off-resonance effects

Please address correspondence to Benjamin M. Ellingson, Ph.D., Director and Professor, UCLA Brain Tumor Imaging Laboratory (BTIL), Departments of Radiological Sciences, Psychiatry, Bioengineering and Neurosurgery, David Geffen School of Medicine, University of California, Los Angeles, 924 Westwood Blvd., Suite 615, Los Angeles, CA 90024, e-mail: bellingson@mednet.ucla.edu

 Indicates article with online supplemental data.

<http://dx.doi.org/10.3174/ajnr.A8566>

(e.g. fat-water shift, geometric distortions from  $B_0$  inhomogeneities from the patient, and signal loss due to dephasing), Nyquist ghosting from poor shimming, gradient coil heating, receiver filter asymmetry, susceptibility, or induction of eddy currents in coils and/or magnet housing in response to the rapidly changing gradients<sup>1</sup>. These artifacts distort image geometry and affect contrast, potentially leading to misinterpretations of critical anatomical features and interfering with quantitative analyses. Numerous traditional (e.g.<sup>2-5</sup>) and deep learning approaches<sup>6-11</sup> have been proposed to mitigate EPI artifacts; however, these often reduce a few specific types of artifacts, require calibration scans, advanced offline reconstruction of raw data, or involve computationally intensive post-processing that isn't feasible in real-time clinical settings.

In the current work, we propose a novel artifact reduction methodology to address these limitations. First, we provide an alternative to gathering large data sets by utilizing MR physics to synthesize realistic acquisition artifacts. Data-driven methods such as deep learning benefit from large diverse data sets<sup>12</sup>, however, access to large, heterogeneous medical imaging data sets with various degrees of image artifacts for deep learning-based artifact correction can be difficult, expensive, and time-consuming to obtain. A physics-based approach that generates synthetic training data with various degrees and types of image artifacts using retrospectively collected high-resolution anatomical MRI images provides a framework for using known physical principles to create a realistic data set for deep learning model development.

Additionally, the current study aims to test the hypothesis that a "Stacked Transfer Learning" strategy provides better performance by using a sequence of models progressively trained on increasingly complex artifacts compared with a single model with identical architecture trained on a data set containing all artifacts in a single session as well as individual models trained for each artifact. Importantly, this methodology is only feasible through the utilization of a synthetic data set where artifacts can be added individually, and their influence precisely controlled.

Furthermore, by leveraging MR physics-based artifact synthesis, our strategy eliminates the necessity for specialized acquisition, capitalizing on both large retrospective data and a priori anatomical knowledge to enhance reconstruction. By incorporating prior anatomical knowledge, we can impose realistic anatomical priors on the reconstruction process, guiding the algorithm toward solutions that are anatomically plausible. Previous studies have demonstrated the effectiveness of incorporating anatomical constraints into reconstruction methods<sup>13,14</sup>. Notably, these methods have proven successful in maintaining the fidelity and anatomical accuracy of reconstructed images. Our approach builds upon this established principle, integrating it into the artifact correction algorithm, leading to more faithful representations of the underlying anatomy.

Finally, we assess the efficacy of the model applied to ADC maps from single-shot diffusion MRI data in patients with recurrent glioblastoma. By including multiple image contrasts in our training data set, we hypothesized the network would be able to learn to correct EPI-based artifacts, while preserving the inherent information of in different image contrasts, even in image contrasts it has not been exposed to previously. This comprehensive

strategy aligns with the overarching goal of enhancing the diagnostic quality of highly accelerated imaging protocols in neuroimaging applications.

To summarize, the aim of this study is to develop a novel artifact correction methodology for EPI, leveraging two technical advancements: (1) synthesis of realistic artifacts based on MR physics, which allows for the generation of diverse training data sets for deep learning models; (2) "Stacked Transfer Learning" approach where models are trained progressively on increasingly complex artifacts. This approach is designed to address a broad range of artifacts simultaneously, improving overall image quality and preserving diagnostic accuracy. The efficacy of the model will be evaluated using single-shot diffusion MRI data in patients with recurrent glioblastoma.

## MATERIALS AND METHODS

This study aimed to develop and validate a deep learning-based approach for artifact reduction in MRI images, particularly focusing on EPI sequences where artifacts are prevalent. The methodology was divided into 4 parts:

- **Artifact data set Synthesis:** We created a synthetic data set from artifact-free MRI images using physics-based simulations to introduce common artifacts such as magnetic susceptibility, chemical shift, fat saturation, and N/2 Nyquist aliasing.
- **Model Training:** The data set was then used to train a modified GAN architecture designed for artifact correction. The training process involved preprocessing the data, introducing anatomical constraints, and optimizing the model by comparing 3 training strategies.
- **Model Evaluation:** The trained model was evaluated using a combination of quantitative metrics and qualitative assessments.
- **Clinical Validation:** The final model was applied to clinical ADC maps derived from single-shot diffusion-weighted EPI data to assess its performance in a real-world setting in a cohort of patients with recurrent glioblastoma.

### Artifact Data Set Synthesis

A synthetic data set with physics-based simulated artifacts was created from 4,573 artifact-free acquisitions from 1,392 patients undergoing routine MRI at our institution using the standardized BTIP protocol<sup>15</sup> between January 2017 and December 2020. The mean age of the cohort is 52.4 years (standard deviation 18.0 years). Age spanned from 6 months to 102 years, and the gender distribution was balanced with 52% male and 48% female patients. The data set included 1,341 T1-weighted pre-contrast, 932 T1-weighted post-contrast, 1,313 T2-weighted, and 987 T2-weighted FLAIR exams (Supplemental Table 1). Table 1 highlights the demographic characteristics of each of the datasets.

**Magnetic Susceptibility ( $\Delta\chi$ ) Artifacts.** To simulate the off-resonance artifacts, including geometric distortion, signal loss, and signal pile-up, induced by the susceptibility differences at boundaries between air and various tissue types, a synthetic susceptibility map was created using the segmentation of CT image voxels into clusters corresponding to air, tissue, and bone. We used one single CT image, obtained from The Cancer Imaging Archive (<https://www.cancerimagingarchive.net/>, data set OPC-Radiomics,

**Table 1: Demographic characteristics of each dataset**

Demographic Characteristic	Training Set		Validation Set		Test Set		Qualitative ADC Clinical Validation		Quantitative ADC Clinical Validation	
Number of patients	1088		153		151		49		22	
Number of scans	3573		500		500		49		22	
Age (years), mean (SD)	52.03	( $\pm 18.19$ )	53.57	( $\pm 17.42$ )	53.53	( $\pm 16.76$ )	60.32	( $\pm 9.37$ )	59.95	( $\pm 6.72$ )
Age range (years)	1.5	—91	0.5	—102	6	—93	35	—85	49	—78
Sex										
Male	575	53%	79	51%	73	48%	29	59%	13	59%
Female	513	47%	73	48%	78	52%	20	41%	9	41%
Unknown	0	0%	1	1%	0	0%	0	0%	0	0%

**Note:**—1 patient lacked demographic information on their DICOM headers due to anonymization procedures.

ID OPC-00056), which covers the head, neck, and shoulder. Coverage of the head, neck, and shoulder was specifically desired due to considerable magnetic field inhomogeneities from tissue interfaces near the nasal cavity, mouth, sphenoid sinus, temporal bones, as well as the air/tissue interfaces at the shoulders<sup>16</sup>. This CT image was registered to MNI stereotaxic space and performed threshold-based segmentation of air, tissue and bone for subsequent use, with air  $< -200$  HU, bone  $> 200$  HU, and tissue mask containing the remaining voxels. A simulated susceptibility distribution was subsequently computed from the segmentations, using susceptibility values of 0.40, -8.44, and -9.04 ppm assigned to air, bone, and soft tissue, respectively.

With the synthetic three-dimensional susceptibility distribution and assuming a static magnetic field strength of 3T, the field map was calculated using the Fourier-based calculation method proposed by Bouwman et al.<sup>17</sup>. To mimic a more realistic head positioning and introduce more variability, we rotated the 3D patient data set registered to MNI space along a random axis by  $\pm 10^\circ$ . Lastly, the global field inhomogeneity in the brain was removed by subtracting the mean value of  $\Delta B_0$  within the brain from the  $\Delta B_0$  maps. Using simulated field map  $\Delta B_0$ , we performed off-resonance artifact simulation using the Fourier-based Off-REsonanCe Artifact simulation in the STeady-state (FORECAST) algorithm<sup>18</sup>.

**Chemical Shift Artifact and Fat Saturation.** Artifacts were simulated as a shift of the pixels containing the fat-containing non-brain tissue. The brain was first segmented using the Brain Extraction Tool (FSL, Functional Magnetic Resonance Imaging of the Brain Software Library v6.0, Oxford, UK). A fat-containing tissue mask for each patient was determined by subtracting the brain mask and then applying a threshold to remove the background pixels. The fat saturation artifact was simulated by ensuring that these shifts also accounted for the variations in signal intensity due to fat suppression techniques, thus reflecting the combined effects of chemical shift and fat saturation. The fraction of signal intensity experiencing pixel shift was a random variable between 0 and 1, representing the fat fraction. The pixel shift was determined using the fat-water frequency difference at 3T (430 Hz) and the pixel bandwidth along the phase-encoding direction. The pixel bandwidth was determined by a random variable of echo-spacing between 0.6 to 1.3 ms. A total of 128 pixels and a GeneRalized Autocalibrating Partially Parallel Acquisitions (GRAPPA) factor of 3 along the phase-encoding direction was

assumed. After calculating the size of the chemical-shift artifact, a fractional pixel shift using linear interpolation was performed.

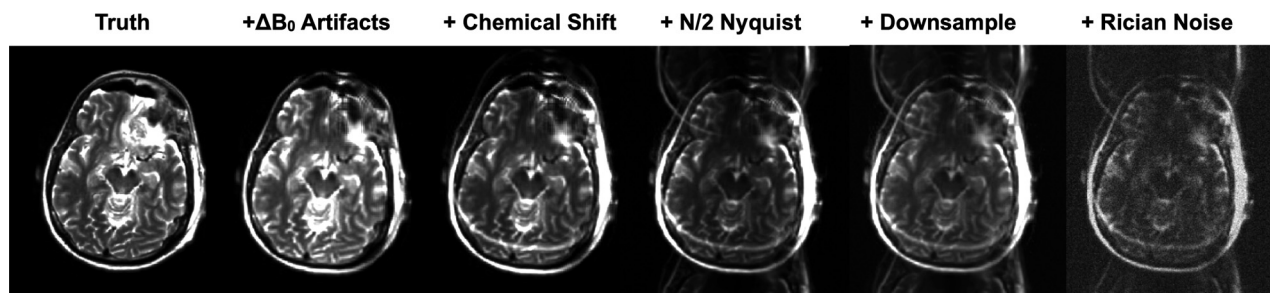
**N/2 Nyquist Aliasing Artifact.** N/2 Nyquist artifact was simulated as a linear phase difference between the data sampled in even and odd k-space lines. The artifact-free anatomical images were first transformed to the frequency domain and the phase artifact was added along the readout direction ( $k_x$ ) to the even k-space lines in the phase-encoding direction ( $k_y$ ). The linear phase ramp was simulated using a constant term  $b$  and a first-order term  $a$ :  $\phi(k_x) = (ak_x + b)\pi$ . Both parameters  $a$  and  $b$  were randomly drawn from normal distributions centered at zeroes, with standard deviations of 0.2 and 0.04, respectively. The directions of the frequency- and phase-encodings were also randomized for variability, with an 80% chance of phase-encoding direction being anterior-posterior and a 20% chance of being left-right. After adding the linear phase artifact to the even k-space lines, the data were transformed back to the image domain.

Gibbs artifacts were simulated by sampling the central sections of k-space with a rectangular window of varying sizes, ranging between 60% and 30% of the original size, effectively altering the image resolution. In addition to this, Rician noise was introduced separately as part of the data augmentation process during training. This noise was added randomly to the images to increase the diversity and robustness of the training data set, but it was not involved in the simulation of Gibbs artifacts.

All images underwent preprocessing to improve model training efficiency and ensure data consistency. Intensity values were normalized to a standard range, and images were resized to a common dimension of 256x256x128. To facilitate consistent spatial interpretation, the orientation of all images was set to RAS. To confirm fidelity and “realism” when compared to real-world EPI artifacts, we consulted with radiologists, physicists from both industry and academia, neurosurgeons, and neuro-oncologists. For robustness of the model, we also included combinations of artifacts that are somewhat not-realistic (e.g. chemical shift in one direction and Nyquist ghosting in a different direction), as a greater depth of complex artifacts should allow the model to better correct less complex artifacts observed in the real-world. **Figure 1** illustrates an example of a T2-weighted image with the sequential addition of various EPI-based artifacts used for the current project.

To incorporate prior anatomical knowledge and guide artifact correction towards anatomically plausible reconstructions, we employed edge detection. Utilizing a Canny edge detection





**FIG 1.** Example of simulated EPI artifacts on a T2-weighted image of a patient with a glioma.

algorithm<sup>19,20</sup>, we generated edge maps from scans separate from the artifact-containing input data. This approach aimed to minimize information leakage during training, where the model might exploit edge information directly related to the ground truth. For example, if the input was a T1-weighted image with artifacts, the edge map was derived from a co-registered T2-weighted image. This strategy assisted the model to learn inherent anatomical features rather than artifact-specific edges for reconstruction. The synthetic data set was divided into subsets, with 3,573 images allocated for training and two additional sets of 500 images designated for validation and testing purposes.

### Model Training

A modified GAN architecture based on *Pix2Pix*<sup>21</sup> was used in the current study due to its efficacy in image-to-image translation tasks (Fig 2A). The discriminator in the *Pix2Pix* GAN consists of a *PatchGAN* network with a classification matrix output (Fig 2B). The generator consists of a whole image-to-image auto-encoder network with U-Net skip connections to generate better image quality at higher resolutions. We modified the generator to a U-Net to include residual units<sup>22</sup>, as well as Recurrent Convolutional layers<sup>23</sup> with gates<sup>24</sup> (*AttentionR2UNet*) (Fig 2C). This architectural modification offers several advantages. First, the inclusion of residual units addresses the issue of vanishing gradients commonly encountered in deep models, proving especially beneficial for the training of deep architectures. Second, the utilization of recurrent convolutional layers facilitates improved feature representation through feature accumulation. Third, the integration of attention gates enables the network to concentrate on the salient areas within the images.

Due to GPU limitations, the images were processed in smaller batches of 16 slices at a time from the reformatted 256×256×128 images. This approach allowed us to manage computational constraints while still utilizing 3D spatial information across slices. The input to the network was structured as a 2-channel image, where the first channel contained the artifact image, and the second channel included an edge detection image.

In the 3D PatchGAN architecture, each convolutional layer has a kernel size of 3×3×3 with a stride of 2 and padding of 1. Across four layers, this configuration results in a receptive field of 31×31×16. The effective receptive field in the third dimension is constrained to 16, ensuring that the network effectively captures the entire input volume along this axis, enabling the network to learn intricate details over a substantial area of the image while maintaining focus on the complete depth of the input.

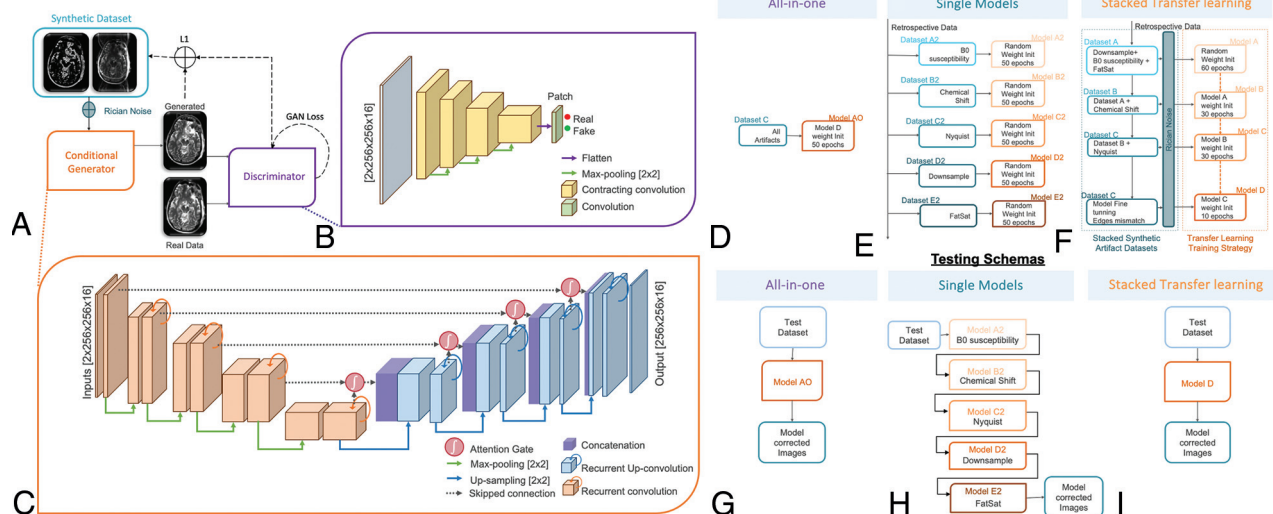
The GAN was trained using the *Adam* optimizer and binary cross-entropy adversarial loss function in conjunction with L1 reconstruction loss (weights 1:200). The models were implemented in Python 3.10.12 with the deep learning framework Pytorch 1.11.0. The training process was executed using a GPU cluster comprised of 44 NVIDIA GPUs: 20 Quadro RTX 8000, 8 Tesla V100, and 16 GeForce RTX 2080 Ti.

We leveraged the ability to create multiple synthetic data sets to train multiple models, thereby enabling comprehensive performance analysis of diverse training strategies (Fig 2D–I). All strategies were allowed to train for a maximum of 130 epochs total or until the model converged. Loss curves for the training process are presented in Supplemental Fig. 1 showing all models adequately converged and would not benefit from additional training. The models used for the current study included:

- 1) **“All-in-one” model:** Utilizes a single data set that includes all artifacts in a single training session (Fig 2D). The training was stopped once the model converged 50 epochs, with a batch size of 4. Testing consisted of direct application of the input images to the single “all-in-one” model (Fig 2G).
- 2) **Single models:** Utilizes different data sets, each with only one simulated artifact, to train independent models for each artifact (Fig 2E). The final images represent a data set that has been sequentially passed through each model (Fig 2H). Each model was trained for 50 epochs, with a batch size of 4. All models converged before training was stopped.
- 3) **Stacked Transfer Learning approach:** Utilizes three different data sets, where each data set included a new artifact added onto the previous data set (data sets A–C). A series of models were trained on these data sets with sequentially added artifacts, with the weights of each trained model used as the initialization parameters for the subsequent model (Fig 2F). A final fine-tuning step is used to improve edge mismatch. We theorized this strategic initialization process would enable the network to leverage knowledge gained from prior training stages, facilitating the effective removal of complex image artifacts in a cumulative manner. Testing consisted of direct application of the input images to a final model (Fig 2I). The proposed methodologies adhered to the guidelines outlined in the CLAIM checklist.

### Model Evaluation

Three key image quality metrics were used to compare the performance of the algorithms for image quality: MSE, SSIM<sup>25</sup>, and



**FIG 2.** A, GAN schematic. B, PatchGan discriminator architecture. C, Attention-R2UNet generator Network architecture. D, Sequence diagram for the training portion of the all-in-one model trained with a data set of all artifacts. E, Sequence diagram for the training portion of the single models method, where separate models are trained individually with each set of artifacts. F, Sequence diagram for the training portion of the stacked transfer learning approach, where a model is first trained on 1 artifact set, then this learning is transferred to a new model, and the process is repeated for the next artifact set. G, Sequence diagram for testing the all-in-one model trained with a data set of all artifacts. H, Sequence diagram for testing the single models method, where the data are passed through each single model sequentially. I, Sequence diagram for the testing of the stacked transfer learning approach.

MS-SSIM<sup>26</sup>. SSIM is a measure of image similarity that considers the luminance, contrast, and structure of the image while MSE is a measure of the pixel-wise difference between two images. For evaluating noise removal effectiveness while preserving prediction fidelity, PSNR was calculated. Finally, the HAUS was used to quantify improvements in geometric distortion arising from  $\Delta B_0$  susceptibility artifacts<sup>27</sup>. This performance metric assesses how closely the surfaces align by measuring the maximum distance between corresponding points on the surfaces. In our case, we extracted the edges of the major structures in the brain using Canny edge detection ( $\sigma = 3$ ) and calculated the HAUS between these edges and their counterparts in the ground truth images.

While the primary goal of artifact reduction models is to improve image quality by removing geometric distortion, testing them solely on artifact-laden images could be misleading. Evaluating the various models' performance on clean, artifact-free images offers crucial insights beyond its ability to remove artifacts, revealing potential unintended consequences on healthy image content. Therefore, we assessed the performance of the three models on both synthetic data sets with artifacts present as well as undistorted images (ground truth of the test set).

### Clinical Validation Using Single-Shot Diffusion-Weighted EPI

To assess the performance of the final "Stacked Transfer Learning" model, we applied it to ADC maps from single-shot diffusion-weighted EPI data in a retrospective cohort of patients diagnosed with recurrent glioblastoma who received cytotoxic chemotherapy (lomustine, temozolomide, or carboplatin) between 2004 and 2022 at our institution. All patients provided written informed consent to participate in this study, which was approved by our IRB<sup>28</sup>. The study included only patients who had baseline pre-treatment scans with contrast-enhanced T1-weighted images

performed within a month before initiating second-line therapy. Importantly, patients did not receive anti-angiogenics and had no intervening surgeries or treatment interventions before RANO-defined disease progression<sup>29</sup>. 56 patients met the inclusion criteria (mean age 60.32 years ( $\pm$  9.37 standard deviation); 29 males) and 49 had ADC imaging at baseline and 22 scans that followed BTIP. NAWM normalization of the ADC images was done with three spheres in the centrum semiovale<sup>30</sup>.

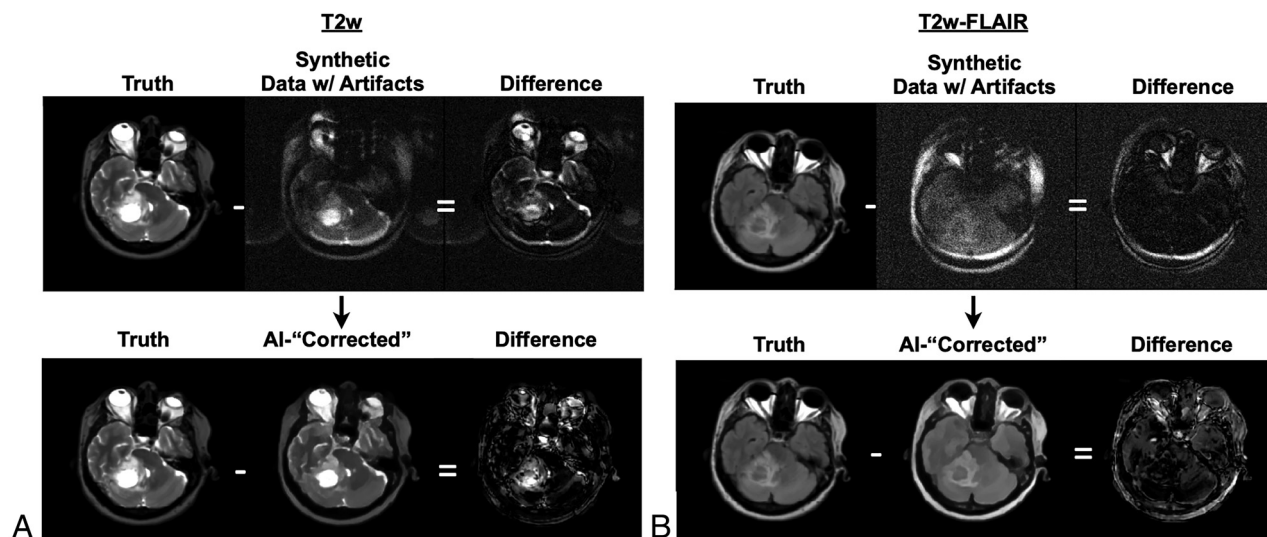
To assess the effectiveness of the model in removing artifacts from EPI images, a blinded evaluation was performed by two independent readers with expertise in neuro-oncology (Reader 1: radiologist with 11 years of experience in neuroradiology; Reader 2: imaging scientist with 15 years of brain tumor imaging and clinical trial experience) following the quality assurance methodology for diffusion MR images proposed by Ellingson *et al.*<sup>31</sup>. The ratings were based on a categorical scale from 1 to 5, where a score of 1 indicated that the image was unusable due to artifacts significantly affecting the tumor area, and a score of 5 indicated no distortion or artifacts. Each radiologist evaluated both the original (unprocessed) and model corrected images for a total of 98 images. Demographic details for these patients are in Table 1.

### Statistical Analysis

Spearman correlation of image quality assessment between Readers A and B before and after artifact correction was calculated to assess the heterogeneity. The scores for the original ADC images were compared to the scores on the model predicted ADC images using the Wilcoxon Signed-Rank Test. Additionally, Spearman correlation coefficients were calculated to assess the association between the expert ratings of the original and model predicted images. Cohen's weighted kappa was used to evaluate inter-rater reliability, quantifying the level of agreement between two readers while

**Table 2: Mean (STD) performance of the training frameworks for the synthetic artifact test set and artifact-free test set (ground truth). MSE and HAUS, lower is better. SSIM, MS-SSIM, PSNR higher is better.**

Synthetic Artifact Test Set					
Model	MSE	SSIM	MS-SSIM	PSNR	HAUS [mm]
All-in-one	0.0024 ( $\pm$ 0.0021)	0.88 ( $\pm$ 0.03)	0.88 ( $\pm$ 0.07)	27.12 ( $\pm$ 2.96)	4.15 ( $\pm$ 2.57)
Single Models	0.0023 ( $\pm$ 0.0015)	0.89 ( $\pm$ 0.0258)	0.90 ( $\pm$ 0.06)	26.56 ( $\pm$ 2.46)	4.46 ( $\pm$ 2.59)
Stacked Transfer Learning	0.0016 ( $\pm$ 0.0017)	0.92 ( $\pm$ 0.02)	0.92 ( $\pm$ 0.06)	28.10 ( $\pm$ 3.56)	4.08 ( $\pm$ 2.67)
Artifact-Free Test Set (ground truth)					
All-in-one	0.0024 ( $\pm$ 0.0015)	0.91 ( $\pm$ 0.01)	0.92 ( $\pm$ 0.06)	27.11 ( $\pm$ 2.90)	
Single Models	0.0005 ( $\pm$ 0.0003)	0.96 ( $\pm$ 0.00)	0.98 ( $\pm$ 0.02)	33.31 ( $\pm$ 2.43)	
Stacked Transfer Learning	0.0004 ( $\pm$ 0.0002)	0.96 ( $\pm$ 0.02)	0.98 ( $\pm$ 0.04)	33.99 ( $\pm$ 1.62)	



**FIG 3.** A representative test set case of T2-weighted (A) and T2-weighted FLAIR images (B) in a patient with recurrent glioma. The top rows reflect the ground truth images (*left*), the ground truth images with the addition of synthetic EPI artifacts (*middle*), and the difference between the truth and the images with synthetic artifacts (*right*). The bottom row highlights the ground truth images (*left*), the corrected images by using the stacked transfer learning model applied to the images with synthetic artifacts (*middle*), and the difference between the truth and the model-corrected images (*right*).

accounting for the possibility of agreement occurring by chance. To further assess the model's impact on tumor regions, ROIs were created for both the contrast enhancing and T2 hyperintense lesions using the NS-HGlio artificial intelligence software from Neosoma (Neosoma Inc, Groton, MA, <https://neosomainc.com>)<sup>32</sup>. The normalized average ADC values within the ROIs of 22 patients in the cohort that followed the BTIP protocol were used to evaluate the consistency of measurements between native and predicted ADC images using paired *t*-test. Additionally, Pearson correlations were performed between the native and "corrected" measurements in the T2 hyperintense as well in the contrast enhancing ROIs.

## RESULTS

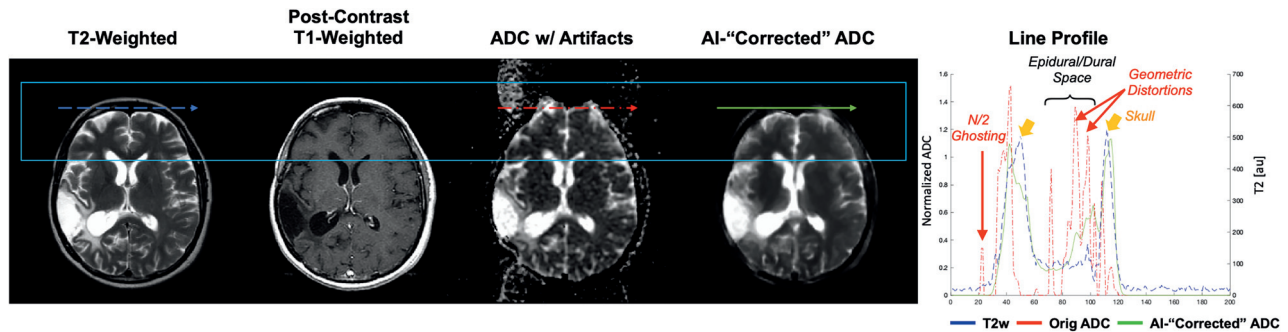
Results suggested the "Stacked Transfer Learning" approach had significantly better performance when examining the testing data set with synthetic artifacts present (Table 2), with the lowest MSE (0.0016 vs. 0.0024 and 0.0023 for "all-in-one" and "single models", respectively), highest SSIM (0.92 vs. 0.88 and 0.89), and highest MS-SSIM (0.92 vs. 0.88 and 0.90). This indicates the model generated images using the "Stacked Transfer Learning" approach better resembles the reference images in terms of pixel values and better captures anatomical details at different scales. Additionally,

this approach showed superior performance in preserving relevant information while removing noise, as demonstrated by exhibiting the highest PSNR value (28.19 vs. 27.12 and 26.56). Finally, the "Stacked Transfer Learning" approach had the lowest HAUS (4.08 mm vs. 4.15 mm and 4.46 mm), indicating the best mitigation of geometric distortion artifacts among the various approaches.

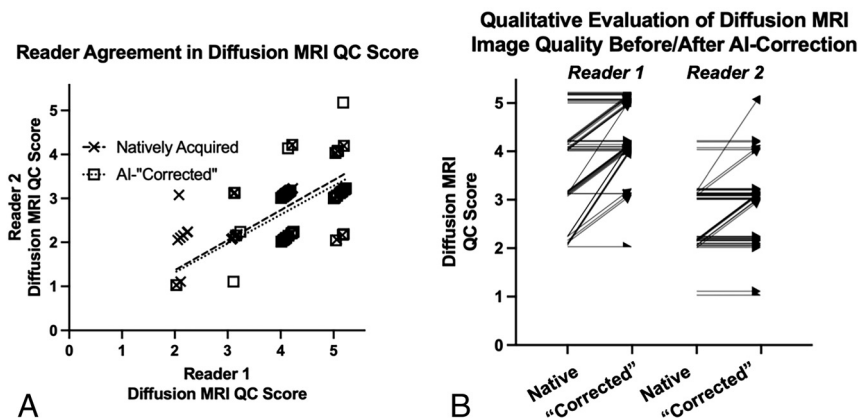
When examining the test set *without* the addition of artifacts, the "Stacked Transfer Learning" approach achieved the lowest pixel-wise error ( $MSE = 0.0004$  vs. 0.0024 and 0.0005), indicating the closest resemblance to the original images, as well as the highest PSNR (33.99 vs. 27.11 and 33.31), suggests exceptional noise reduction compared to the other training approaches. The "single model" structural similarity scores ( $SSIM = 0.96$ ,  $MS-SSIM = 0.98$ ) and "Stacked Transfer Learning" ( $SSIM = 0.96$ ,  $MS-SSIM = 0.98$ ) were similar. Finally, the "all in one" training approach appeared to score worst among all measures of performance, indicating weaker fidelity, structural preservation, and noise reduction when a model is trained directly with all the artifacts at once.

Figure 3 illustrates a representative case of T2-weighted (Fig 3A) and T2-weighted FLAIR images (Fig 3B) in a patient with recurrent glioma. The top rows reflect the ground truth images (*left*), the ground truth images with the addition of synthetic EPI artifacts (*middle*), and the difference between the truth and the images with





**FIG 4.** Clinical validation example ADC maps acquired from single-shot diffusion-weighted EPI in a patient with recurrent glioblastoma. Geometric distortion and N/2 Nyquist ghosting is noted on the original ADC map in the frontal lobe, distal from the site of tumor. Line profiles represent ADC or arbitrary pixel-values along segments of interest, allowing a visual comparison between landmark locations before and after artifact removal with respect to a T2-weighted image as a reference.



**FIG 5.** Comparison of diffusion MRI QC scores between 2 independent readers. A, Correlation of QC between reader 1 and reader 2 before and after AI-correction showing high correlation between readers. (B) Change in QC scores for reader 1 and reader 2 after AI correction. (Note: reader scores were offset slightly for visualization purposes).

synthetic artifacts (right). The bottom row highlights the ground truth images (left), the corrected images using the “Stacked Transfer Learning” model applied to the images with synthetic artifacts (middle), and the difference between the truth and the model-corrected images (right), highlighting the ability effectively suppress various artifacts while preserving important anatomic details.

When applying the model to a unique data set for clinical validation using single-shot diffusion-weighted EPI in patients with recurrent glioblastoma (Fig 4; Supplemental Fig. 2), results appeared to similarly demonstrate the ability to adequately correct for obvious artifacts – particularly the commonly encountered geometric distortions and signal dropout in the frontal and occipital regions, as well as Nyquist artifacts. Line profiles through key regions of the head with the most pronounced distortions appear to show better alignment with anatomic landmarks between model-corrected ADC maps and T2-weighted images, including the edge of the brain and skull. Additionally, generated ADC maps had notably higher SNR compared with the ADC maps with artifacts present. However, it is important to note that the model appeared to exhibit some elevated blurring and reduction in image detail in some cases, which warrants future investigation.

To generally assess the visual quality of the model-generated ADC maps, Spearman correlation analysis was performed

between QC scores assessed by Readers 1 and 2. Results showed high correlation in rank before ( $p=0.7505$ ,  $p<0.001$ ) and after ( $p=0.7938$ ,  $p<0.001$ ) model correction (Fig 5A), as well as high rank consistency between pre- and post-correction assessments by each reader (Reader 1,  $\rho=0.7505$ ,  $p<0.001$ ; Reader 2,  $\rho=0.7907$ ,  $p<0.001$ ). The Cohen’s weighted kappa statistic showed only slight agreement between readers in terms of the absolute QC score both before ( $\kappa=0.148$ ) and after ( $\kappa=0.054$ ) model correction as there was a bias toward a lower QC score for Reader 2 (Reader 1= $4.06\pm0.94$ , Reader 2= $2.77\pm0.77$ ,  $P<0.0001$ ). However, the average QC scores for the original ADC images were significantly lower for both readers when compared to the scores

on the AI-“corrected” ADC images (Fig 5B; Reader 1= $3.60\pm1.02$  vs.  $4.27\pm0.73$ ,  $p<0.0001$ ; Reader 2= $2.51\pm0.71$  vs.  $2.78\pm0.80$ ,  $p=0.0005$ ). Together, these results appear to confirm a significant improvement in image quality with the model-corrected ADC maps compared to ADC maps prior to correction.

Lastly, normalized ADC measurements from different tumor regions were compared to determine whether quantitative measurements would be significantly impacted by model-based artifact removal (Supplemental Fig. 3). Results did not demonstrate a significant difference in normalized ADC measurements in the T2 hyperintense regions (Supplemental Fig. 3A,  $p=0.2490$ ) or contrast enhancing tumor areas (Supplemental Fig. 3C;  $p=0.1390$ ). A strong linear correlation was observed between the natively acquired and model-corrected ADC measurements in areas of T2 hyperintensity (Supplemental Fig. 3B;  $r=0.9500$ ,  $p<0.0001$ ) as well as within contrast enhancing tumor regions (Supplemental Fig. 3D;  $r=0.9044$ ,  $p<0.0001$ ).

## DISCUSSION

The current study utilized stacked synthetic data sets with increasingly complex distortions to fully leverage the benefits of transfer learning in an artifact reduction model. Importantly, the



present study illustrates a novel approach for improved image quality using a combination (1) a physics-based method for generating an extensive data set of realistic synthesized images, and (2) the effective use of this synthetic data to train deep learning models using a “stacked transfer learning” approach.

Our findings indicate that adopting a physics-based methodology for synthesizing realistic image artifacts enhances accessibility to extensive training data sets, particularly in clinical settings where acquiring artifact-free ground truth data is not feasible. Traditional data collection methods to train such models require acquisition of extensively large training sets, each with unique or combined image artifacts, which would be costly and time-consuming. Furthermore, “stacked transfer learning” would not be possible with traditional approaches for data collection, as it would be extremely difficult and impractical to layer on each required artifact at various levels. The ability to synthesize a range of artifacts with varying intensities and combinations allows for generation of extensive data sets of images that can more efficiently train new and more powerful models.

It is important to point out some critical limitations of the current study. First, the data used in the current study was from a single institution, with the vast majority of data acquired on 3T scanners (>99%). This can conceivably introduce biases related to patient demographics, imaging protocols, and equipment characteristics specific to that site. Limited demographic information restricts access to race and ethnicity for the data set which could also contribute to unknown biases. Extrapolating findings to a broader population should be undertaken cautiously, recognizing the potential limitations inherent in a site-specific data set. To address this, future work should incorporate multi-center data sets, encompassing a broader range of scanner types (e.g., 1.5T and 7T scanners), protocols, and patient populations, including diverse racial and ethnic groups. This would improve the generalizability of the model and allow for a more comprehensive assessment of its performance across various clinical settings.

Additionally, the current study focused on a subset of artifacts commonly associated with single-shot EPI, as well as a single acceleration factor and no simultaneous multi-slice acquisition. While the proposed concept, design, and implementation successfully addressed identified artifacts (magnetic susceptibility, chemical shift, N/2 Nyquist aliasing, Gibbs artifact, and Rician noise), extending the methodology to other artifacts with different acquisition schemes remains unexplored. Future research could investigate additional artifacts such as coil profile artifacts, parallel imaging artifacts, ramp sampling, eddy-current induced geometric distortions, B<sub>1</sub>-sensitivity profile effects, metal implants, “blinds” artifacts or missing slices, and motion. An additional limitation in this work was the use of a single CT image to simulate magnetic field inhomogeneities from tissue interfaces. Susceptibility distributions might vary significantly with head anatomy, previous surgical interventions and metal implants. Future work should focus on including a more heterogeneous cohort allowing to capture of a broader range of scenarios and patient characteristics that would further validate the model’s adaptability to real-world clinical scenarios.

It is important to acknowledge that aggressive artifact reduction can sometimes lead to unintended alterations in the images.

This is a known challenge using generative A.I., and underscores the need for further evaluation beyond a single study. Synthetic images generated by A.I., including those produced by our model, should be used with caution in the clinical setting. Importantly, the current study did not observe any pseudo-lesions, where lesions appeared where they were not previously present or lesions disappeared that were previously present, in any of our validation data as confirmed by our clinical evaluators and investigators. However, it is essential to continually evaluate these images and to incorporate feedback from radiologists and other medical professionals to ensure that they meet the necessary criteria for the intended diagnostic use.

Finally, the evaluation of the proposed methodology was limited to a single EPI modality as a proof of concept. Extending the testing in future work to EPI sequences, including perfusion MRI, fMRI, and other MR image modalities, including non-proton MR, is essential for a comprehensive understanding of its applicability across different multiple imaging contrasts.

## CONCLUSIONS

The current study demonstrates feasibility of using a physics-based method for synthesizing a large data set of images with realistic EPI-based artifacts and the effectiveness of utilizing this synthetic data set in a “stacked transfer learning” approach to training a GAN for the purposes of artifact reduction.

**Disclosure forms** provided by the authors are available with the full text and PDF of this article at [www.ajnr.org](http://www.ajnr.org).

## REFERENCES

1. Franz Schmitt MKS, Turner R. *Echo-Planar Imaging: Theory, Technique, and Application*. Springer-Verlag; 1998:179–200
2. Buonocore MH, Gao L. **Ghost artifact reduction for echo planar imaging using image phase correction**. *Magn Reson Med* 1997; 38:89–100 [CrossRef Medline](#)
3. Chen N, Wyrwicz AM. **Removal of EPI Nyquist ghost artifacts with two-dimensional phase correction**. *Magn Reson Med* 2004;51:1247–53 [CrossRef Medline](#)
4. Bruder H, Fischer H, Reinfelder HE, et al. **Image reconstruction for echo planar imaging with nonequidistant k-space sampling**. *Magn Reson Med* 1992;23:311–23 [CrossRef Medline](#)
5. Poser BA, Barth M, Goa PE, et al. **Single-shot echo-planar imaging with Nyquist ghost compensation: interleaved dual echo with acceleration (IDEA) echo-planar imaging (EPI)**. *Magn Reson Med* 2013;69:37–47 [CrossRef Medline](#)
6. Bilgic B, Chatnuntawech I, Manhard MK, et al. **Highly accelerated multishot echo planar imaging through synergistic machine learning and joint reconstruction**. *Magn Reson Med* 2019;82:1343–58 [CrossRef Medline](#)
7. Kawamura M, Tamada D, Funayama S, et al. **Accelerated acquisition of high-resolution diffusion-weighted imaging of the brain with a multi-shot echo-planar sequence: deep-learning-based denoising**. *Magn Reson Med Sci* 2021;20:99–105 [CrossRef Medline](#)
8. Lee J, Han Y, Ryu JK, et al. **k-Space deep learning for reference-free EPI ghost correction**. *Magn Reson Med* 2019;82:2299–313 [CrossRef Medline](#)
9. Cui L, Song Y, Wang Y, et al. **Motion artifact reduction for magnetic resonance imaging with deep learning and k-space analysis**. *PLoS one* 2023;18:e0278668 [CrossRef Medline](#)
10. Duong ST, Phung SL, Bouzerdoum A, et al. **An unsupervised deep learning technique for susceptibility artifact correction in reversed**

- phase-encoding EPI images. *Magn Reson Imaging* 2020;71:1–10 [CrossRef Medline](#)
11. Hu Z, Wang Y, Zhang Z, et al. **Distortion correction of single-shot EPI enabled by deep-learning.** *Neuroimage* 2020;221:117170 [CrossRef Medline](#)
12. Luo G, Wang X, Blumenthal M, et al. **Generative image priors for MRI Reconstruction Trained from Magnitude-Only Images.** *arXiv preprint* 2023;230802340
13. Constantinides CD, Weiss RG, Lee R, et al. **Restoration of low resolution metabolic images with a priori anatomic information: 23Na MRI in myocardial infarction.** *Magn Reson Imaging* 2000;18:461–71 [CrossRef Medline](#)
14. Halder JP, Hernando D, Song SK, et al. **Anatomically constrained reconstruction from noisy data.** *Magn Reson Med* 2008;59:810–18 [CrossRef Medline](#)
15. Ellingson BM, Bendszus M, Boxerman J, et al; Jumpstarting Brain Tumor Drug Development Coalition Imaging Standardization Steering Committee. **Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials.** *Neuro Oncol* 2015;17:1188–98 [CrossRef Medline](#)
16. Truong TK, Clymer BD, Chakeres DW, et al. **Three-dimensional numerical simulations of susceptibility-induced magnetic field inhomogeneities in the human head.** *Magn Reson Imaging* 2002;20:759–70 [CrossRef Medline](#)
17. Bouwman JG, Bakker CJG. **Alias subtraction more efficient than conventional zero-padding in the Fourier-based calculation of the susceptibility induced perturbation of the magnetic field in MR.** *Magn Reson Med* 2012;68:621–30 [CrossRef Medline](#)
18. Zijlstra F, Bouwman JG, Braškutė I, et al. **Fast Fourier-based simulation of off-resonance artifacts in steady-state gradient echo MRI applied to metal object localization.** *Magn Reson Med* 2017;78:2035–41 [CrossRef Medline](#)
19. Rong W, Li Z, Zhang W, et al. **An improved CANNY edge detection algorithm.** In: *2014 IEEE International Conference on Mechatronics and Automation* 2014:577–82
20. Deriche R. **Using Canny's criteria to derive a recursively implemented optimal edge detector.** *Int J Comput Vision* 1987;1:167–87 [CrossRef](#)
21. Isola P, Zhu JY, Zhou T, et al. **Image-to-Image Translation with Conditional Adversarial Networks.** In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2017:5967–76 [CrossRef](#)
22. He K, Zhang X, Ren S, et al. **Deep residual learning for image recognition.** In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016:770–78 [CrossRef](#)
23. Alom MZ, Hasan M, Yakopcic C, et al. **Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation.** 2018
24. Oktay O, Schlemper J, Folgoc LL, et al. **Attention u-net: learning where to look for the pancreas.** *arXiv preprint* 2018:180403999
25. Wang Z, Bovik AC, Sheikh HR, et al. **Image quality assessment: from error visibility to structural similarity.** *IEEE Trans Image Process* 2004;13:600–12 [CrossRef](#)
26. Wang Z, Simoncelli EP, Bovik AC. **Multiscale structural similarity for image quality assessment.** In: *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003* 2003:1398–1402
27. Huttenlocher DP, Klanderman GA, Rucklidge WJ. **Comparing images using the Hausdorff distance.** *IEEE Trans Pattern Anal Machine Intell* 1993;15:850–63 [CrossRef](#)
28. Oshima S, Hagiwara A, Raymond C, et al. **Change in volumetric tumor growth rate after cytotoxic therapy is predictive of overall survival in recurrent glioblastoma.** *Neurooncol Adv* 2023;5:vdad084 [CrossRef Medline](#)
29. Wen PY, Macdonald DR, Reardon DA, et al. **Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group.** *J Clin Oncol* 2010;28:1963–72 [CrossRef Medline](#)
30. Cho NS, Hagiwara A, Sanvito F, et al. **A multi-reader comparison of normal-appearing white matter normalization techniques for perfusion and diffusion MRI in brain tumors.** *Neuroradiology* 2023;65:559–68 [CrossRef Medline](#)
31. Ellingson BM, Kim E, Woodworth DC, et al. **Diffusion MRI quality control and functional diffusion map results in ACRIN 6677/RTOG 0625: a multicenter, randomized, phase II trial of bevacizumab and chemotherapy in recurrent glioblastoma.** *Int J Oncol* 2015;46:1883–92 [CrossRef Medline](#)
32. Abayazeed AH, Abbassy A, Müller M, et al. **NS-HGlio: a generalizable and repeatable HGG segmentation and volumetric measurement AI algorithm for the longitudinal MRI assessment to inform RANO in trials and clinics.** *Neurooncol Adv* 2023;5:vdac184 [CrossRef Medline](#)