ColorMedGAN: A Semantic Colorization Framework for Medical Images

Shaobo Chen [1] , Ning Xiao [1], Xinlai Shi [2], Yuer Yang [3,4] , Huaning Tan [5] , Jiajuan Tian [1] and Yujuan Quan [1,2,*]

1. College of Information Science and Technology, Jinan University, Guangzhou 510632, China
2. Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Informatization, Jinan University, Guangzhou 510632, China
3. College of Cyber Security, Jinan University, Guangzhou 511436, China
4. School of Economics, Jinan University, Guangzhou 510632, China
5. International School, Jinan University, Guangzhou 511436, China
* Correspondence: quanyj@126.com

Abstract: Colorization for medical images helps make medical visualizations more engaging, provides better visualization in 3D reconstruction, acts as an image enhancement technique for tasks such as segmentation, and makes it easier for non-specialists to perceive tissue changes and texture details in medical images in diagnosis and teaching. However, colorization algorithms have been hindered by limited semantic understanding. In addition, current colorization methods still rely on paired data, which is often not available for specific fields such as medical imaging. To address the texture detail of medical images and the scarcity of paired data, we propose a self-supervised colorization framework based on CycleGAN(Cycle-Consistent Generative Adversarial Networks), treating the colorization problem of medical images as a cross-modal domain transfer problem in color space. The proposed framework focuses on global edge features and semantic information by introducing edge-aware detectors, multi-modal discriminators, and a semantic feature fusion module. Experimental results demonstrate that our method can generate high-quality color medical images.

Keywords: medical image colorization; CycleGAN; visualization

1. Introduction

Colorization has always been a popular topic in computer vision research [1] and there have been many challenges in general colorization research to date. Human organ and tissue images are often in grayscale, which creates a big challenge when applying colorization methods using supervised learning with paired data. However, color images of medical images can be better visualized [2,3] and can play a special role in aiding diagnosis and treatment for both professionals and non-professionals. For instance, color Doppler ultrasound measures blood flow velocity through the Doppler effect and displays it in the image through colorization [4]. In addition, modern medicine often employs three-dimensional reconstruction technology to observe human organs, and through colorization and three-dimensional display, their appearance can be made more realistic and intuitive, leading to better observation results [5].

Early image colorization algorithms have traditionally focused on modeling the correspondence between grayscale and color [6,7]. In 2001, Reinhard et al. [8] proposed a novel colorization concept based on the observation that channels in the CIELAB(L*a*b* color model developed by The International Commission on Illumination) color space are uncorrelated. They developed a set of color transfer formulas that can be applied to each color component, yielding promising results. Alternatively, some researchers have explored the diffusion of color from local to global regions. Horiuchi and Hirano [9] leveraged the local Markov property of the image to minimize the color difference between neighboring pixels and achieve colorization, while Musialski et al. [10] implemented a user-guided

approach with colored sketches to minimize user interaction. However, these approaches either heavily relied on user interaction or resulted in desaturated colorizations, limiting their practicality and applicability [11].

In recent years, the development of deep learning has led to impressive performance in colorization methods [12]. One of the pioneering approaches, Colorful Image Colorization, was initially proposed by Cheng et al. [11] and treats colorization as a classification problem. Several studies [12–14] have employed GAN (generative adversarial network) [15] for image colorization, which has produced impressive results but has not matched the performance of the method proposed by Zhang et al. [11]. The Pix2pix model [12] based on GAN has also achieved excellent performance in colorization tasks. To achieve the colorization of real-world objects, ColorGAN(Image Colorization with Generative Adversarial Networks) [16] was introduced, which utilizes conditional GANs. While methods such as Pix2Pix have shown promising results, they rely on paired datasets where inputs and labels are paired one-to-one. Architectures like CycleGAN [17], DualGAN(Unsupervised Dual Learning for Image-to-Image Translation) [18], and DiscoGAN(generative adversarial networks that learn to discover relations between different domains) [19] leverage two GANs to map features from one set to another, which is beneficial for datasets with unpaired images (a dataset with non-corresponding input and label data). Recently, Transformer architecture has gained attention in the field of image colorization. Kumar et al. [20] proposed a grayscale colorization network (Colorization Transformer), called Colorization Transformer, based on Transformer blocks, which enables the matching of color information to input grayscale images at low resolution. However, these methods have their own limitations. Architectures like Transformer require a large amount of data and computing resources, while self-supervised methods like CycleGAN are prone to producing artifacts during colorization, which may result in the loss of structural information and contours. Low-level semantic information such as contours and geometric shapes plays a crucial role in the colorization process. Therefore, the use of extracted features to calculate non-local similarity has become a popular method in recent years [21,22]. Ge et al. [23] have also incorporated local detail perception with semantic segmentation to improve the quality of colorization.

One of the few studies on colorizing medical images involved the use of transfer functions to map intensity values to colors, as explored by Ljung et al. [24]. However, these methods can be cumbersome and time-consuming, and often require manual intervention. Khan et al. [5] addressed this issue by embedding color information in each pixel of the subject image, resulting in the potential for better depiction of anatomical information compared to traditional monochrome images. Another approach, used by Zeng et al. [25], involved the use of VGG19 (Very Deep Convolutional NetWorks for Large-Scale Image Recognition) [26] and adaptive target image retrieval to generate high-quality color medical images. This method incorporated Y-loss and content-loss to maintain texture structure and image color loss to maintain color invariance between the generated image and the reference image. Li et al. [27] used an image colorization method based on Gabor filtering and Welsh colorization to extract local spatial and frequency domain information from grayscale images and render images with distinct texture features in the filtered images. However, these methods are often dependent on the performance of retrieval algorithms or the quality of reference images, and the resulting images can look more like pseudo-colored images than the actual situation. Mathur et al. [28] used 2D style migration for colorizing 3D medical images, demonstrating multi-modal body colorization using only one 2D style image. This approach can be used to create better 3D visualizations without the need for time-consuming transfer functions. Zhang et al. [2] proposed a fully automatic spatial mask-guided colorization with a generative adversarial network framework for medical image colorization. This method generates colorized images with fewer artifacts by introducing spatial masks, which encourage the network to focus on the colorization of foreground regions rather than the entire image. Although the above methods generate anatomical-style images from MRI or grayscale images, they may lose a significant amount

of information inherent in medical imaging. Additionally, some of these methods rely on one-to-one paired datasets, such as MRI and anatomical colored images in [28], which requires considerable manual screening work and carries the risk of misalignment.

Medical images are known for their low variation in tissue and high complexity in texture, making them prone to color bleeding and loss of texture details during colorization. Existing colorization methods for medical images often require paired data or manual alignment between different modalities, which can result in unclear boundaries and a single colorization effect. Our experiments demonstrate that current models cannot generate high-quality colorized medical images. To address these challenges, we propose a novel approach that treats medical image colorization as a cross-modal transfer problem in color space. Our method constrains color generation with rich semantic and edge information to eliminate the need for paired data and produce high-quality images. CycleGAN (The Cycle Generative Adversarial Network) has emerged as a powerful tool for medical image translation, and recent research has explored its potential for medical image augmentation and unsupervised translation. Recently, some studies [29,30] have combined CycleGAN's advantages with the constraints of local structure and feature details to improve medical image augmentation. Cycle GAN-based models are also widely applied to medical image translation, with Armanious et al. [31] proposing a novel unsupervised translation framework called Cycle-MedGAN. This framework utilizes a new non-adversarial cycle loss to guide the framework in minimizing texture and perceptual differences in the translated images. Other similar studies include [32–36]. Drawing inspiration from these works, we propose a CycleGAN-based unsupervised medical image colorization framework that pays attention to both low-level structural information and high-level semantic information. By doing so, we aim to generate high-quality colorized medical images without the need for paired data or manual alignment.

In this paper, we propose a pipeline based on CycleGAN [17] and ColorCycleGAN (Single Image Colorization via Modified CycleGAN) [37] for the automatic coloring of MRI images. Our approach formulates the colorization problem as a domain transfer task in the color domain, transferring the colors from the source domain (cryo-anatomical images) to the target domain (MRI images) while preserving the semantic colors of the source domain and the structural information of the target domain. To achieve this, we pre-train a segmentation network to fuse semantic features and introduce edge loss and multi-modal discriminators to guide the model to generate rich texture details. Our approach does not require paired image sets or manual alignment of images and can be trained using cross-modal unpaired datasets. We use the Brain MRI dataset Brats2018 [38] with instance segmentation annotations as the source domain data and a subset of The Human Project [39] and The Korean Human Project [40]'s cryo-anatomical data as the target domain data. In summary, this paper:

- Proposes a framework that can generate semantically colored images across modalities. This method can train the colorization model with an unpaired dataset and generate high-quality color images.
- Introduces an instance segmentation network branch in generating phase. The generator is guided to produce highly semantic and color-consistence images based on the attention mechanism.
- Introduces the edge loss and multi-modal discriminator based on edge-aware detectors which enables the generator to focus on the global edge information of the image, preserve more edge detail textures, and prevent the color bleeding problem.

2. Methodology

Our method can bypass the design of the image color transfer optimization function and directly generate realistic colorized images. It can be trained on cross-modal paired datasets, thereby avoiding the considerable workload generated by matching one-to-one datasets and automating the modal transformation operation during training CycleGAN, thus reducing the need for manual operations. The principle of our method is illustrated

in Figure 1. We employ a forward network G to transform the grayscale MR images into anatomical images in RGB format and then use a backward network F to recover the input.
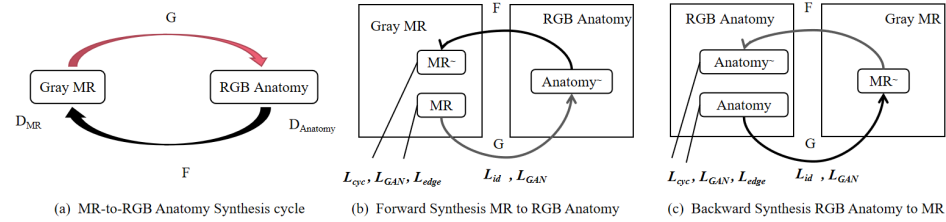

Figure 1. (**a**) illustrates the MR-to-anatomical synthesis cycle in color space. (**b**,**c**) primarily demonstrate the forward synthesis from MR to anatomical and the backward synthesis from anatomical to anatomical, with the inclusion of the edge and color object functions.

2.1. ColorMedGAN Architecture

Unlike traditional one-way GANs, CycleGAN is essentially two mirror-symmetric GANs. Inspired by this work and colorCycleGAN, we propose a CycleGAN-based colorization framework for unpaired datasets. A summary of our network structure is shown in Figure 2, in this method, the grayscale MR images as the source domain, and the RGB anatomical images as the target domain are input into the framework. Then, the features of the generator and the features captured by the segmentation network are weighted by the attention module and then continue to be input into the generator, and finally output the colorful MR images. In addition, this module also introduces an edge extraction module to constrain the edge and structural information in the generation through an edge loss and a multimodal discriminator. We aim to utilize cross-modal unpaired datasets to generate realistic color medical images, transfer colorful semantic information to MR images, and reduce the loss of structural and contour information in them.
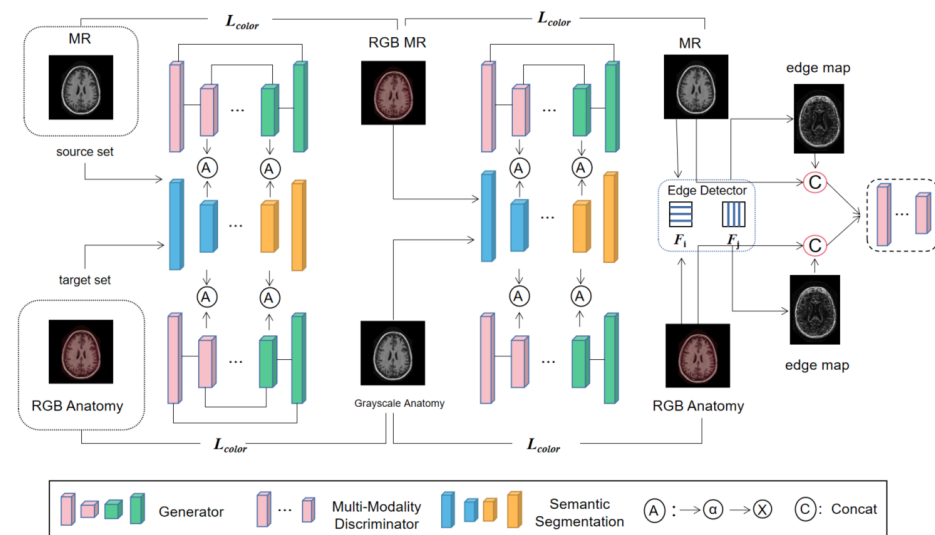

Figure 2. Framework of the proposed method. It takes grayscale MR images as the source domain and RGB anatomical images as the target domain, which are input into the framework. The generator features and segmentation network features are weighted by an attention module, which is then fed back into the generator, producing colorful MR images. Furthermore, the method includes an edge extraction module to constrain the edge and structural information in the generation by means of an edge loss and a multi-modal discriminator.

2.1.1. Backbone Network

The backbone network of the framework consists of three parts: generator **G**, discriminator **D** and semantic segmentation network **S**. The U-Net-like network structure [41] is

adopted for both the semantic segmentation network **S** and generative network **G**, where **G** takes in a one-dimensional grayscale image as input and generates a three-channel color image, while **S** pertains to output a feature map with semantic information. **S** and **G** are similarly chosen for the lightweight 6-layer encoder-decoder framework to fuse the consistency between features. As a colorization network, the U-structured encoder allows low-level information to pass through the network, and then the final RGB color image is predicted by resizing the learned feature maps back to the same spatial size as the input image deconvolution layer. Instance Normalization (IN) [42] is used in the network to remove instance-specific information and simplify the gray features before concatenation, resulting in high-quality generation performance.

For discriminator network **D**, we use a patch-level discriminator [43] to distinguish between samples from the real dataset and generated ones, as the discriminator task is relatively straightforward. The basic local features are encoded using two convolutional blocks for classification. A convolutional block containing a feature transform layer and a convolutional pixel-level regression layer is then used to obtain the classification response. The weights of the generator and discriminator's convolutional layers are spectrally normalized, except for their respective last convolutional layer.

In the colorization process, semantic information is of great importance as it provides valuable color guidance in combination with grayscale values to the colorization network. Therefore, accurate semantic mapping is crucial for our colorization network. To maintain consistency in fused feature representation, we use a semantic segmentation network trained offline with a UNet network [41] similar to the colorization network. To balance the trade-off between the layers of fused information and computational efficiency, in this study, we extract feature layers (*batch* $\times$ 128 $\times$ 128 $\times$ 128) from each of the encoder and decoder of the segmentation network to fuse with the generator, representing both shallow and deep semantic representations.

2.1.2. Attention Guided Module

Inspired by [44], we introduce an attention-guided generator that leverages edge semantic information to enhance the feature representation in the hierarchy. The attentional fusion structure is depicted in Figure 3a. We explore two feature fusion approaches (Figure 3) to transfer more efficient semantic information from the edge feature map F_e to the image feature map F_i. One is to concatenate the edge features with the colorization network features and input concat(F_e, F_i) to the encoder at the next layer; the other is to use the attention-guided approach to input the edge features to the activation layer and then multiply them per wise with the feature map from the generator. The latter approach is experimentally proven to be superior as shown in Table 1. The edge feature map is first processed by the sigmoid activation function to generate the corresponding attention map $F_a = \sigma(F_e)$. Then, we multiply the generated attention map with the corresponding image feature map to obtain a refined map containing local structure and details, which is further inputted into the next convolutional layer. The feature maps of the two fusion modes are presented in Figure 4. The right figure shows that our attention mode outperforms the concatenate mode as it preserves more information from the segmentation and reduces noise. The generated images (comparison of Figure 5c and Figure 5d) also demonstrate the superiority of the sigmoid mode

$$F_i^j = \text{Sigmoid}(F_e^j) \times F_i^j \tag{1}$$

where F_e^j denotes the feature map of the segmentation network, F_i^j denotes the feature map of the generator.
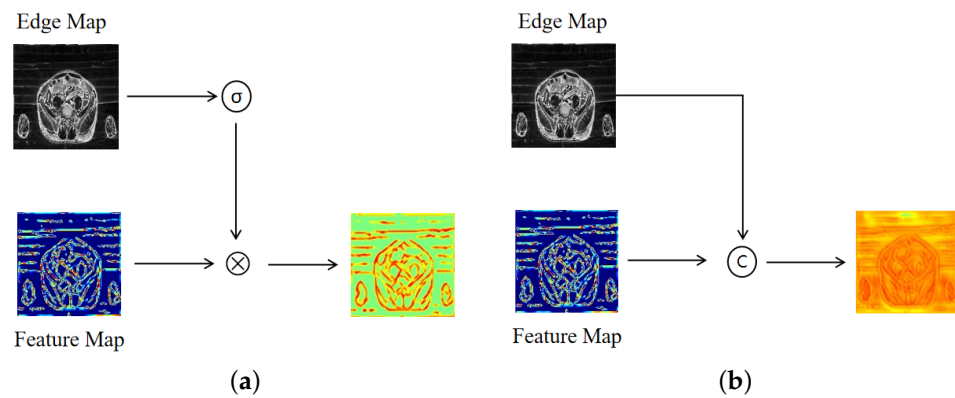
Figure 3. Two modes of feature fusion: (**a**) element-wise multiplication and Sigmoid activation function (**b**) channel-wise concatenate.
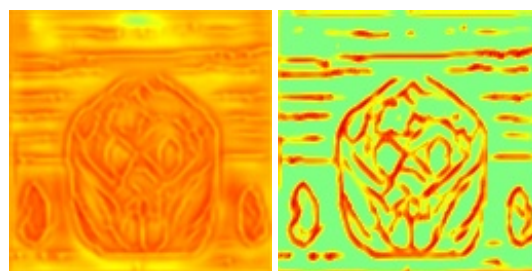

Figure 4. The feature map of two fusion modes. The left one presents the concatenate mode, the right one presents the sigmoid mode.
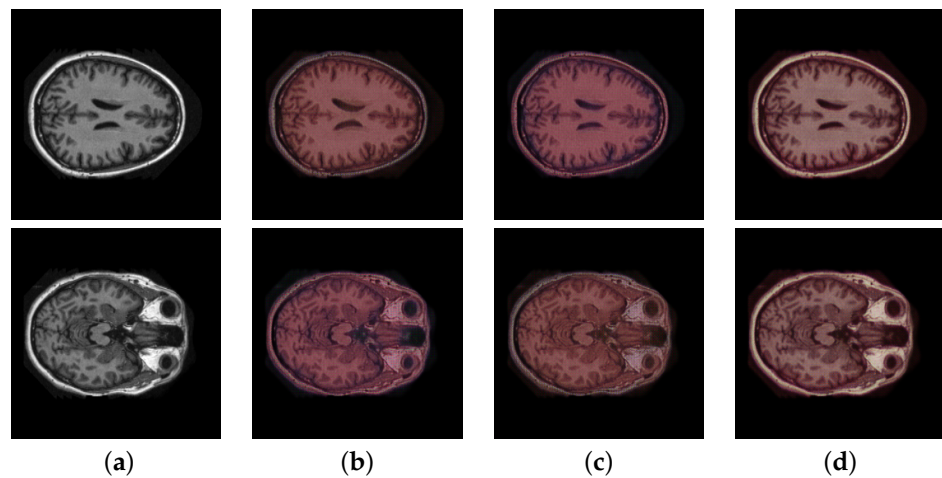

Figure 5. Comparison of the influence of semantic feature. (**a**) origin image; (**b**) without semantic information; (**c**) with semantic information in concatenate mode; (**d**) with semantic information in sigmoid mode.

Table 1. Comparison of Concatenate and Sigmoid mode.

Model	SSMI ↑	PSNR ↑	FID ↓
Concat mode	0.896	**33.325**	94.59
Sigmoid mode	**0.901**	33.224	**84.81**

Bold fonts indicate better results.

Then, we perform another convolution operation on F_0 to obtain a feature map $F_0 \in R(C + N + 3 + 3) \times H \times W$ to enhance the representativeness of this feature. In addition, F_0 is identical to the original input F. This makes the module flexible and inserted into other architectural structures without modifying other parts to refine the output. Finally,

the feature map F_0 is fed into the convolutional layer followed by the $\tanh()$ nonlinear activation layer to obtain the final result, allowing the generator to converge faster. The proposed semantic preservation module enhances the representational power of the model by adaptively recalibrating the semantic class-dependent feature maps and is in a similar spirit to style transfer and the recent works SENet (Squeeze-and-Excitation Networks) [45] and EncNet (Context Encoding Network) [46]. An intuitive example of this module's utility is in generating small object classes, which are easily lost in the generated results due to the loss of spatial resolution. Our scaling factor can emphasize small objects and help preserve them.

2.2. Loss Function

This section can be divided into two parts. The first part discusses the loss function used for color learning, while the second part focuses on the loss function that constrains the structure and contour information. The scheme for both parts is shown in Figure 1b,c.

2.2.1. ColorCycleGAN Loss Function

Consistent with the definition of CycleGAN, we define the colorization task as two sets of mapping directions: $G : X \rightarrow Y$ and $F : Y \rightarrow X$, where X, Y represent the grayscale and color domains. $x_i^N = 1$ where $x_i \in X$ and $y_j^M = 1$ where $y_j \in Y$. The segmentation feature is defined as $S_x = S(X)$, $S_y = S(Y)$. The data domain distribution is denoted as $x \sim \mathrm{Pdata}(x)$, $y \sim \mathrm{Pdata}(y)$. In addition, there are two discriminators, D_x and D_y, in the system, where D_x is used to distinguish the true X from the false X generated by F and D_y is used to distinguish the true Y from the false Y generated by G. Therefore, the loss function in this part can be defined as Equation (2).

$$\begin{aligned} \mathcal{L}_{\mathrm{cycleGAN}}(G, F, D_x, D_y, S) = \mathcal{L}_{\mathrm{GAN}}(G, S, D_y, X, Y) \\ + \mathcal{L}_{\mathrm{GAN}}(F, S, D_x, Y, X) \end{aligned} \tag{2}$$

where the first part denotes the loss functions of the forward synthesis, and the second part denotes the loss functions of the backward synthesis.

For the image x in the X domain, the image generated after passing into G and then put into F to generate the X domain image back should be consistent with the original image, $x \rightarrow G(x) \rightarrow F(G(x))$, y in the same way. So we can define the consistency loss function as Equation (3).

$$\begin{aligned} \mathcal{L}_{cyc}(G, F) = \mathbb{E}_{x \sim Pdata(x)}[||F(G(x)) - x||_1] \\ + \mathbb{E}_{y \sim Pdata(y)}[||G(F(y)) - y||_1] \end{aligned} \tag{3}$$

The consistency loss function is defined concerning the colorCycleGAN approach, using perceptual losses as computational objects instead of pixel points, thus enabling the style between X and Y to map, shown as Equation (4).

$$\begin{aligned} \mathcal{L}_{ide}(G, F) = \mathbb{E}_{y \sim Pdata(y)}[||VGG_l(G(y)) - VGG_l(y)||_1] \\ + \mathbb{E}_{x \sim Pdata(x)}[||VGG_l(F(x)) - VGG_l(x)||_1] \end{aligned} \tag{4}$$

The gray loss function, for any RGB image, we can get the corresponding grayscale images through the following Equation (5), also introduced by colorCycleGAN, keeps the image structure invariant and slightly adjusts the generated image by calculating the loss of gray intensity using Equation (5). We got Equation (6) finally. The first half indicates that the image $G(x)$ generated by grayscale MR is consistent with the gray intensity of

the original image, and the second half indicates that the image $F(y)$ generated by the anatomical image is consistent with the gray intensity of the input image.

$$Gray(\text{img}) = 0.299r + 0.576g + 0.114b \tag{5}$$

$$\mathcal{L}_{col}(G, F) = \mathbb{E}_{x \sim Pdata(x)}[||Gray(G(x)) - x||_1]$$
$$+ \mathbb{E}_{y \sim Pdata(y)}[||F(y) - Gray(y)||_1] \tag{6}$$

2.2.2. Edge Aware Loss Function

In addition to the aforementioned loss function, we also introduce an edge-aware loss function. It enables the generator to emphasize the generation of image texture edges by adding a feature extractor to the network, calculating the edges of the image using the Laplace operator, and calculating the edge loss between the original and the generated image, guiding the generator to generate an image with more prominent texture details in global space.

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \tag{7}$$

The operator is expressed by the mathematical formula shown as Equation (7), where x and y represent the horizontal and vertical coordinates on the Cartesian coordinate system xOy. For the polar coordinate system, our equation will be as shown in Equation (8). Furthermore, it is expressed on the pixel matrix as Equation (9), whose loss function is shown as Equation (10).

$$\Delta f = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial f}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 f}{\partial \theta^2} \tag{8}$$

$$\nabla^2 f(x, y) = f(x + 1, y) + f(x - 1, y)$$
$$+ f(x, y + 1) + f(x, y - 1) - 4f(x, y) \tag{9}$$

$$\mathcal{L}_{edge,x} = \mathbb{E}[||Lap(X) - X||] \tag{10}$$

where f in Equation (8) denotes the complex function of the real part r and imaginary part θ; x and y in Equation (9) represent the horizontal and vertical coordinates.

2.2.3. Multi-Modality Discriminator

In addition to the training loss of the Laplace operator generator, we also incorporate the effect of this operator on the discriminator into the adversarial training process. We calculate the product of the feature map extracted by the Laplace operator after activation with the original map and assign the image edge weights with attention focus before adding them to the discriminator calculation. Equation (11) shows the calculation above.

$$\mathcal{L}_D = \mathbb{E}[||D(\sigma(Lap(y)) \oplus y) - D(\sigma(Lap(G(x))) \oplus G(x))||] \tag{11}$$

where $\oplus$ denotes concatenate operation, σ denotes activation operation and Lap means the edge detector.

2.2.4. Full Objective

Overall, the total loss function can be defined as in Equation (12), which we optimize to Equation (13).

$$\mathcal{L}_{color}(G, F) = \mathcal{L}_{cycleGAN}(G, F, D_x, D_y)$$
$$+\alpha \mathcal{L}_{cyc}(G, F) + \beta \mathcal{L}_{ide}(G, F) \tag{12}$$
$$+\gamma \mathcal{L}_{col}(G, F) + \theta \mathcal{L}_{edge}(G, F, D_x, D_y)$$

$$G^* = \underset{G,F,D_x,D_y}{\operatorname{argmin}} \mathcal{L}_{colorCycleGAN}(G, F, D_x, D_y)$$
$$F^* = \underset{G,F,D_x,D_y}{\operatorname{argmax}} \mathcal{L}_{colorCycleGAN}(G, F, D_x, D_y) \tag{13}$$

where the scalars α, β, γ and θ are used to adjust the weights of the consistency loss and regularization terms and $\mathcal{L}_{edge}$ is a typical L1 loss which represents sparse regularization, making sparse features.

Therefore, we construct a color domain adaptive colorization method with edge constraints, which tries to minimize the difference between the overall distribution of the source domain and the target domain from the perspective of color and edge structure information.

2.3. Experimental Setup

Datasets: Three publicly available datasets were used to train our framework. We extracted a subset of the Visible Korean Human [40] and the Visible Human datasets [39] as color reference images for the target image domain. The Oasis Dataset [38] was used as the training data for the segmentation network and the source domain data for the colorization network. The colorization network was trained on 1707 Visible Human cryo-anatomical images and 50,094 MRI images, while the segmentation network was trained on 94,000 Oasis MRI images with their labels. Additionally, to prevent the image background from interfering with the target colorization object in the training process, we remove the background with the U2Net [47] pre-training segmentation model.

Metrics: To evaluate the performance of our framework, we used three calculated values: SSMI (structural similarity), PSNR(Peak Signal-to-Noise Ratio), and FID (Fréchet Inception Distance). Structural similarity is a full-reference image quality evaluation metric that measures image similarity in terms of brightness, contrast, and structure. PSNR is one of the most widely used objective image quality evaluation metrics. It is based on the error between corresponding pixels, that is, error-sensitive image quality evaluation. The three-channel color image generated in this experiment was converted to YCbCr format for comparison with the grayscale image, and then only the PSNR of the Y component, i.e., the luminance component, was calculated.

PSNR metric is commonly used to evaluate image quality, but it has limitations as it does not consider the visual characteristics of the human eye. For example, the human eye is more sensitive to contrast differences at lower spatial frequencies and more sensitive to luminance contrast differences than chromaticity. Additionally, the perception of an image by the human eye is affected by its surrounding neighboring regions. Therefore, the evaluation results are often inconsistent with subjective human perception. In addition to evaluating the peak signal-to-noise ratio and image structural similarity, mutual information was introduced as an evaluation index of the image in this experiment. According to the work presented in the paper [48], the similarity metric responds poorly to two graphics with high noise and two medical images with different modalities.

Fréchet Inception Distance (FID) [49] is another metric used to evaluate the performance and image quality of our method. We calculate it with the official implementation of it with torch [50]. FID evaluates the similarity between two datasets of images and has been shown to be closely related to the human judgment of visual quality. It is commonly used to assess the quality of the samples generated by the network. FID is calculated by calculating

the Fleischer distance between the two Gauses that fit into the Inception network feature representation.

3. Experiment and Result

In the section, we will focus on addressing three key research questions that were essential to achieving the goal of semantically coloring medical images:

- How to prevent color bleeding in the colorization stage?
- How to enrich the semantic information of color images?
- How to overcome the difference in modality during color migration?

To address the challenges of color bleeding, this paper introduces an edge loss function that can effectively constrain color bleeding during the colorization process. To enrich the semantic information in color images, we incorporate semantic information into the segmentation network to provide additional context for the colorization process. Additionally, to overcome the differences in image modality, a multimodal identifier is used to effectively align the source and target domains. To evaluate the effectiveness of our proposed method, we will compare it with other mainstream image colorization methods on publicly available datasets in the brain. The code is available online at: https://github.com/csbbbv/ColorMedGAN, accessed on 29 January 2023.

3.1. Ablations Analysis

This study performs ablation of the following components: edge Loss, semantic feature fusion, and the multi-modal discriminator. We conduct ablation experiments on these three modules, respectively, and compare the generated images with the evaluation indicators to explore the effect of each module on the generated images.

3.1.1. Effectiveness of Edge Loss

We conducted comparison experiments on two datasets, Oasis1 and Kaggle brain MR datasets [51]. The Kaggle image data has lower resolution and worse texture details, and texture loss is more severe when colorization on such data, with color overbleeding or edge loss. By introducing edge loss, it is possible to constrain the edge information globally in the image generation process. The experimental results in Figure 6 show that the edge loss is more prominent for poor-quality data sets. From the results of not adding edge loss constraints in the second column of Figure 6, the generated image outline color bleeding is severe, and the structural information of the image is seriously lost. After adding edge loss constraints (the third column of Figure 6), a clear image can be generated, and the color itself retains semantic information. Table 2 lists the relevant metrics. The image quality generated after introducing edge loss is better in each metric. The image has better similarity and better results at the level of abstract perception. On the one hand, it can prevent the color overflow phenomenon; on the other hand, adding the edge loss can guide the colorization network to pay more attention to the edge information in the global scope and generate higher-quality texture details in the shading process.

Table 2. Comparison of the influence of edge loss in ColorMedGAN.

Model	SSMI ↑	PSNR ↑	FID ↓
w/o edge loss	0.745	20.735	112.15
w edge loss	**0.757**	**21.010**	**110.32**

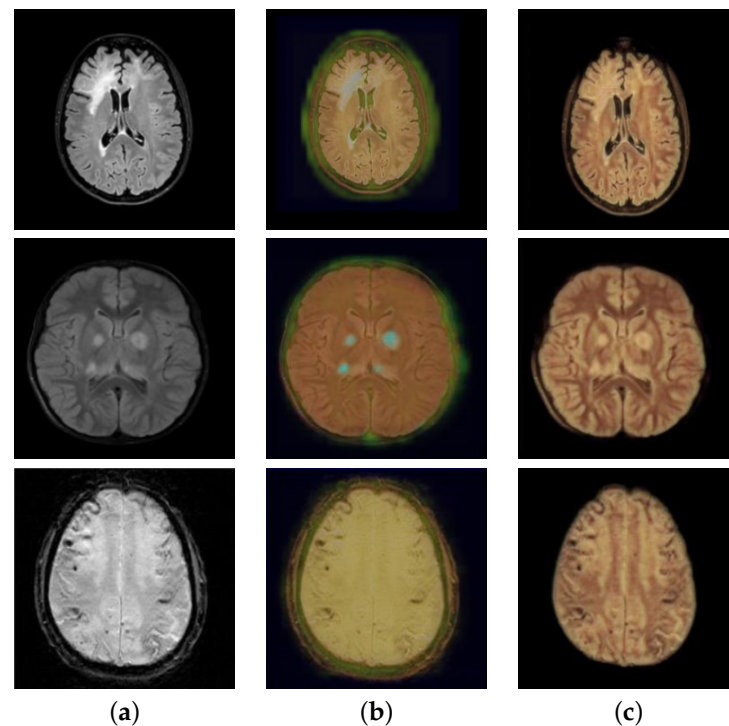Bold fonts indicate better results.

Figure 6. Comparison of the results with and without edge loss. (**a**) origin; (**b**) w/o edge loss; (**c**) w edge loss.

3.1.2. Performance of Semantic Module

We trained the semantic segmentation network of brain MR offline with a U-Net-like network with the same structure as the generator. Based on the consideration of semantic information and computational resources for different layers of features, we did feature fusion on the second and penultimate layers of 128×128 feature maps, firstly, to let the generator learn both shallow and deep semantic information, and secondly, to save computational effort and speed up the generation. We also focus on comparing the effect of different fusion methods. We compare the effect of two fusion methods: directly contacting the feature map and doing attention-guided fusion by multiplying the speech information map with the colorization feature map. The experiments as shown in Figure 5 and Table 3 indicate that the attention-guided fusion method is superior. From the comparison of Figure 5b,d, it can be seen that the color of the results in Figure 5d is more semantic, and the color contrast between different tissues is more vivid, which is reflected in the fact that the colors of the same tissue are closer and the colors of different tissues are distinguishable. At the same time, Table 3 also shows that the quality of the generated images is improved in each evaluation metric with the addition of the segmentation module.

Table 3. Comparison of the influence of segmentation module in ColorMedGAN.

Model	SSMI ↑	PSNR ↑	FID ↓
w/o segmentation module	0.757	21.010	105.23
w segmentation module	**0.857**	**29.698**	**84.81**

Bold fonts indicate better results.

3.1.3. The Multi-Modal Discriminator

This section investigates the performance of a multimodal discriminator designed to learn from cross-modal color space information. We discuss different fusion methods, and experiments in Figure 7 and Table 4 show that the multi-modal discriminator training effect is improved after introducing edge feature maps. In different fusion methods, the image quality generated via concatenation is superior, as demonstrated by Table 4 where modal:2

is better. We hypothesize that the edge feature map contains shallow semantic information, and the weight-attachment fusion method may lose some of this information. In contrast, the concatenation fusion method retains all the original and additional edge features in the input, enabling the convolution network to learn better. As illustrated in the detail figure of Figure 7, after introducing the multimodal discriminator, the small-scale details are generated more accurately and the texture details are preserved more completely.
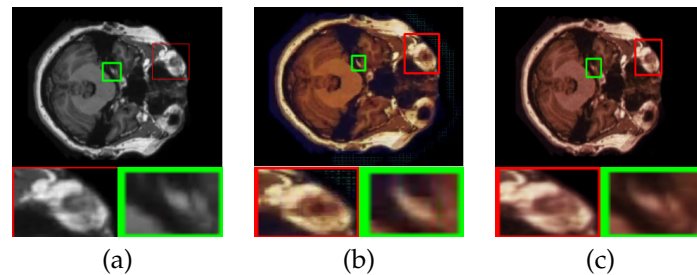

(a)   (b)   (c)

Figure 7. The results and details of comparison of the influence of multi-modal discriminator. (**a**) Origin; (**b**) Without; (**c**) With.

Table 4. Comparison of different multi-modal modes in ColorMedGAN. Multi modal:1 denotes element-wise addition after element-wise multiplication and Sigmoid activation function; Multi modal:2 denotes concatenate directly; Multi modal:3 denotes element-wise multiplication and Sigmoid activation function.

Model	SSMI ↑	PSNR ↑	FID ↓
Multi modal:1	**0.885**	33.40	223.14
Multi modal:2	0.884	**35.95**	**102.84**
Multi modal:3	0.871	28.34	215.95
w/o multi modal	0.875	28.01	156.72

Bold fonts indicate better results.

3.2. Comparisons with State-of-the-Art Networks

Colorization of medical images presents unique challenges due to the lack of paired black-and-white and color data, making it difficult to use supervised learning methods. Furthermore, current methods tend to suffer from color bleeding in medical images with texture details, as shown in Figures 8 and 9. In particular, Figure 8 displays the results of Zhang's method [11], which exhibits color bleeding and semantic discontinuity. Moreover, Figure 9 further shows the results of more advanced methods, all of which are affected by unnatural phenomena such as color bleeding, loss of structure and contour information, and discontinuity of color semantics.
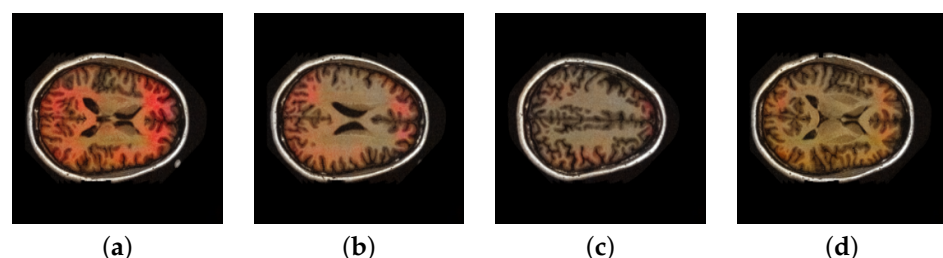

(**a**)   (**b**)   (**c**)   (**d**)

Figure 8. This is the result of Zhang's method being replicated in our dataset. (**a**–**d**) are the results of different slices of a patient's data, and it can be seen that the colors between the slices are not continuous and the colors lose their semantics.

In addition, due to the lack of semantic information for guiding the color of MRI tissues, the color information in the image cannot reflect the relationship between tissues, thus leading to color inconsistency between different slices of the same organ. In Figure 8a–d,

depict different slices of an organ, and as a result, the same tissue regions in these slices will have different colors. However, different regions may also have the same color. When comparing the four images horizontally, it can be observed that the colors of the tissues in different slices lack consistency. Such colorization is meaningless and fails to reflect the relationship between tissues.
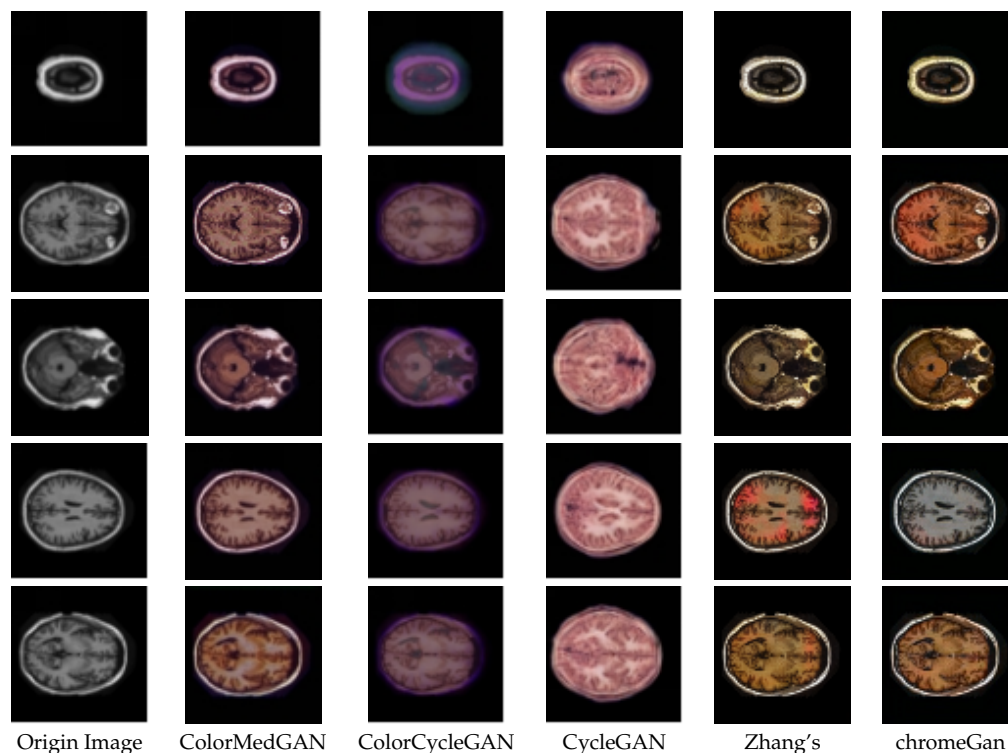

| Origin Image | ColorMedGAN | ColorCycleGAN | CycleGAN | Zhang's | chromeGan |

Figure 9. Results of different models.

We compared several unsupervised coloring models, which were also trained using the unpaired dataset constructed by OASIS1 [38] and HUMAN2 [39]. We selected 1000 oasis MR grayscale images and 610 images from 800 frozen anatomical images composed of HUMAN2 and Korean human datasets [40] for training. As shown in Figure 9, due to the small dataset and lack of intensity constraints, the result of Baseline CycleGAN [17] is far from expectation, the structure of generated images is easily migrated to the modality of frozen anatomical images, edges and contours are completely distorted, textures are almost lost, and the generated images are meaningless; the result of colorCycleGAN [37] shows that the model can ensure the consistency of intensity values, the structure of images can be retained, but due to the lack of global edge constraints, the generated images are obviously stained with colors, and the generated colors are semantically meaningless; Zhang's method [11] generates images with no semantic colors and serious color bleeding, which cannot learn the semantics of structures in medical images; the method of Chroma-GAN [52], a generative network is used to infer the chromaticity of a given grayscale image conditioned to semantic clues, is similar to Zhang's, and sometimes the color generation will fail, resulting in poor quality of generated images;From the results obtained by our method, it can be seen that the generated image is more semantically continuous, the texture details and contour edges are better maintained, the color contrast between the tissues is more realistic in the sense. By comparing the SSMI, PSNR and FID indices of the generated images, as shown in Table 5, it can be seen that our method generates medical images of higher quality in a self-supervised manner, and the three metrics all indicate that the quality of our generated images is better.

Table 5. Coparison of different models.

Model	SSMI ↑	PSNR ↑	FID ↓
colorCycleGAN	0.69	19.86	163.65
CycleGAN	0.28	11.01	266.58
Zhang's	0.89	32.45	104.72
ChromaGAN	0.89	27.48	125.65
Ours	**0.90**	**33.22**	**84.81**

Bold fonts indicate better results.

4. Discuss and Challenges

Our method provides a valuable reference for medical image colorization. By incorporating an edge-aware loss, our model is able to capture delicate texture information while avoiding common issues such as ghosting and color bleeding, all without requiring paired data. Additionally, our approach ensures a high degree of semantic continuity in the generated color images, which can be particularly useful for recognizing tissue information in organs. Even when generating images from large volumes of data piecemeal, our model retains its ability to maintain color consistency due to its retention of semantic information during training. The resulting high-quality color images provide valuable references for 3D visualization rendering, particularly for non-specialists. While extending color consistency to the 3D level may present a challenge due to limitations in computing power resources and input resolution, our model's scalability to 3D via the CycleGAN framework offers significant promise for future work in this field.

5. Conclusions

The task of colorizing medical images is challenging due to the limited availability of paired data. In this paper, our main contribution is to propose a self-supervised framework that aims to generate high-quality color images that are semantically informative and preserve texture details. We demonstrate how to construct a dataset with a colorization architecture, remove the image background to eliminate the effect of the background on the generated color and add a modal transformation module to the segmentation branch on the symmetric architecture of CycleGAN to enable the branch to function in a loop. The CycleGAN model shows good convergence and generalization on MR images and can map the color space of cryosection onto the data domain of cross-modal MR images. Our results demonstrate that our method generates high-fidelity and high-quality color images, despite not relying on paired data. This framework provides a promising approach for colorizing medical images and can facilitate the interpretation of medical images by non-specialists.

Author Contributions: Conceptualization, S.C. and Y.Q.; methodology, S.C.; software, S.C.; validation, S.C.; formal analysis, X.S.; investigation, Y.Y.; resources, H.T.; data curation, J.T.; writing—original draft preparation, S.C. and N.X.; writing—review and editing, S.C. and X.S.; visualization, N.X.; project administration, Y.Q.; funding acquisition, Y.Q. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The Visible Korean Human can be found here: https://archive.org/details/VisibleKoreanHumanSmallImages/, accessed on 29 January 2023; The Visible Human Project can be found here: https://www.nlm.nih.gov/databases/download/vhp.html, accessed on 29 January 2023; The Oasis Dataset can be found here: http://www.oasis-brains.org/, accessed on 29 January 2023.

References

1. Chen, S.Y.; Zhang, J.Q.; Zhao, Y.Y.; Rosin, P.L.; Lai, Y.K.; Gao, L. A review of image and video colorization: From analogies to deep learning. *Vis. Inform.* **2022**, *6*, 51–68. [CrossRef]
2. Zhang, Z.; Li, Y.; Shin, B.S. Robust Medical Image Colorization with Spatial Mask-Guided Generative Adversarial Network. *Bioengineering* **2022**, *9*, 721. [CrossRef]
3. Viana, M.S.; Morandin Junior, O.; Contreras, R.C. An improved local search genetic algorithm with a new mapped adaptive operator applied to pseudo-coloring problem. *Symmetry* **2020**, *12*, 1684. [CrossRef]
4. Williamson, T.H.; Harris, A. Color Doppler ultrasound imaging of theeye and orbit. *Surv. Ophthalmol.* **1996**, *40*, 255–267. [CrossRef] [PubMed]
5. Khan, M.U.G.; Gotoh, Y.; Nida, N. Medical image colorization for better visualization and segmentation. In Proceedings of the Annual Conference on Medical Image Understanding and Analysis, Edinburgh, UK, 11–13 July 2017; pp. 571–580.
6. Luo, W.; Lu, Z.; Wang, X.; Xu, Y.Q.; Ben-Ezra, M.; Tang, X.; Brown, M.S. Synthesizing oil painting surface geometry from a single photograph. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 885–892.
7. Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F.A.; Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv* **2018**, arXiv:1811.12231.
8. Reinhard, E.; Adhikhmin, M.; Gooch, B.; Shirley, P. Color transfer between images. *IEEE Comput. Graph. Appl.* **2001**, *21*, 34–41. [CrossRef]
9. Horiuchi, T.; Hirano, S. Colorization algorithm for grayscale image by propagating seed pixels. In Proceedings of the 2003 International Conference on Image Processing (Cat. No. 03CH37429), Barcelona, Spain, 14–18 September 2003; Volume 1.
10. Musialski, P.; Cui, M.; Ye, J.; Razdan, A.; Wonka, P. A framework for interactive image color editing. *Vis. Comput.* **2013**, *29*, 1173–1186. [CrossRef]
11. Zhang, R.; Isola, P.; Efros, A.A. Colorful image colorization. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 649–666.
12. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
13. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
14. Frans, K. Outline colorization through tandem adversarial networks. *arXiv* **2017**, arXiv:1704.08834.
15. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [CrossRef]
16. Nazeri, K.; Ng, E.; Ebrahimi, M. Image colorization using generative adversarial networks. In Proceedings of the International Conference on Articulated Motion and Deformable Objects, Palma de Mallorca, Spain, 12–13 July 2018; pp. 85–94.
17. Zhang, C.L.; Wu, J. Improving CNN linear layers with power mean non-linearity. *Pattern Recognit.* **2019**, *89*, 12–21. [CrossRef]
18. Yi, Z.; Zhang, H.; Tan, P.; Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2849–2857.
19. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; pp. 1857–1865.
20. Kumar, M.; Weissenborn, D.; Kalchbrenner, N. Colorization transformer. *arXiv* **2021**, arXiv:2102.04432.
21. Shi, M.; Zhang, J.Q.; Chen, S.Y.; Gao, L.; Lai, Y.; Zhang, F.L. Reference-based deep line art video colorization. *IEEE Trans. Vis. Comput. Graph.* **2022**, *20*, 1.
22. Siyao, L.; Zhao, S.; Yu, W.; Sun, W.; Metaxas, D.; Loy, C.C.; Liu, Z. Deep animation video interpolation in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6587–6595.
23. Ge, C.; Sun, H.; Song, Y.Z.; Ma, Z.; Liao, J. Exploring local detail perception for scene sketch semantic segmentation. *IEEE Trans. Image Process.* **2022**, *31*, 1447–1461. [CrossRef]
24. Ljung, P.; Krüger, J.; Groller, E.; Hadwiger, M.; Hansen, C.D.; Ynnerman, A. State of the art in transfer functions for direct volume rendering. *Comput. Graph. Forum* **2016**, *35*, 669–691. [CrossRef]
25. Zeng, X.; Tong, S.; Lu, Y.; Xu, L.; Huang, Z. Adaptive medical image deep color perception algorithm. *IEEE Access* **2020**, *8*, 56559–56571. [CrossRef]
26. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
27. Li, H.A.; Fan, J.; Yu, K.; Qi, X.; Wen, Z.; Hua, Q.; Zhang, M.; Zheng, Q. Medical image coloring based on gabor filtering for internet of medical things. *IEEE Access* **2020**, *8*, 104016–104025. [CrossRef]

28. Mathur, A.N.; Khattar, A.; Sharma, O. 2D to 3D Medical Image Colorization. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikola, HI, USA, 19–25 June 2021; pp. 2847–2856.

29. Hammami, M.; Friboulet, D.; Kéchichian, R. Cycle GAN-based data augmentation for multi-organ detection in CT images via YOLO. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 25–28 October 2020; pp. 390–393.

30. Ma, Y.; Liu, Y.; Cheng, J.; Zheng, Y.; Ghahremani, M.; Chen, H.; Liu, J.; Zhao, Y. Cycle structure and illumination constrained GAN for medical image enhancement. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, 4–8 October 2020; pp. 667–677.

31. Armanious, K.; Jiang, C.; Abdulatif, S.; Küstner, T.; Gatidis, S.; Yang, B. Unsupervised medical image translation using cycle-MedGAN. In Proceedings of the 2019 27th European Signal Processing Conference (EUSIPCO), A Coruña, Spain, 2–6 September 2019; pp. 1–5.

32. Cohen, J.P.; Luck, M.; Honari, S. Distribution matching losses can hallucinate features in medical image translation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2018: 21st International Conference, Granada, Spain, 16–20 September 2018; pp. 529–536.

33. Kong, L.; Lian, C.; Huang, D.; Hu, Y.; Zhou, Q. Breaking the dilemma of medical image-to-image translation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 1964–1978.

34. Cao, B.; Zhang, H.; Wang, N.; Gao, X.; Shen, D. Auto-GAN: Self-supervised collaborative learning for medical image synthesis. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10486–10493.

35. Yao, S.; Tan, J.; Chen, Y.; Gu, Y. A weighted feature transfer gan for medical image synthesis. *Mach. Vis. Appl.* **2021**, *32*, 22. [CrossRef]

36. Sun, L.; Wang, J.; Huang, Y.; Ding, X.; Greenspan, H.; Paisley, J. An adversarial learning approach to medical image synthesis for lesion detection. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2303–2314. [CrossRef] [PubMed]

37. Xiao, Y.; Jiang, A.; Liu, C.; Wang, M. Single image colorization via modified CycleGAN. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 3247–3251.

38. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med Imaging* **2015**, *34*, 1993–2024. [CrossRef] [PubMed]

39. Spitzer, V.M.; Whitlock, D.G. The Visible Human Dataset: The anatomical platform for human simulation. *Anat. Rec. Off. Publ. Am. Assoc. Anat.* **1998**, *253*, 49–57. [CrossRef]

40. Park, J.S.; Chung, M.S.; Hwang, S.B.; Shin, B.S.; Park, H.S. Visible Korean Human: Its techniques and applications. *Clin. Anat. Off. J. Am. Assoc. Clin. Anat. Br. Assoc. Clin. Anat.* **2006**, *19*, 216–224. [CrossRef] [PubMed]

41. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.

42. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv* **2016**, arXiv:1607.08022.

43. Li, C.; Wand, M. Precomputed real-time texture synthesis with markovian generative adversarial networks. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 702–716.

44. Tang, H.; Qi, X.; Xu, D.; Torr, P.H.; Sebe, N. Edge guided GANs with semantic preserving for semantic image synthesis. *arXiv* **2020**, arXiv:2003.13898.

45. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.

46. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.

47. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404. [CrossRef]

48. Wells III, W.M.; Viola, P.; Atsumi, H.; Nakajima, S.; Kikinis, R. Multi-modal volume registration by maximization of mutual information. *Med Image Anal.* **1996**, *1*, 35–51. [CrossRef]

49. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 25–34.

50. Seitzer, M. Pytorch-Fid: FID Score for PyTorch; Version 0.2.1. 2020. Available online: https://github.com/mseitzer/pytorch-fid (accessed on 15 June 2021).

51. Nickparvar, M. Brain Tumor MRI Dataset. 2021. Available online: https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection (accessed on 20 August 2021).

52. Vitoria, P.; Raad, L.; Ballester, C. Chromagan: Adversarial picture colorization with semantic class distribution. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 2445–2454.