Article

FuseVis: Interpreting Neural Networks for Image Fusion Using Per-Pixel Saliency Visualization

Nishant Kumar * and **Stefan Gumhold**

Chair of Computer Graphics and Visualisation, Faculty of Computer Science, Technische Universität Dresden, Nöthnitzer Straße 46, 01187 Dresden, Germany; stefan.gumhold@tu-dresden.de

* Correspondence: nishant.kumar@tu-dresden.de; Tel.: +49-351-463-38354

check for updates

Abstract: Image fusion helps in merging two or more images to construct a more informative single fused image. Recently, unsupervised learning-based convolutional neural networks (CNN) have been used for different types of image-fusion tasks such as medical image fusion, infrared-visible image fusion for autonomous driving as well as multi-focus and multi-exposure image fusion for satellite imagery. However, it is challenging to analyze the reliability of these CNNs for the image-fusion tasks since no groundtruth is available. This led to the use of a wide variety of model architectures and optimization functions yielding quite different fusion results. Additionally, due to the highly opaque nature of such neural networks, it is difficult to explain the internal mechanics behind its fusion results. To overcome these challenges, we present a novel real-time visualization tool, named *FuseVis*, with which the end-user can compute per-pixel saliency maps that examine the influence of the input image pixels on each pixel of the fused image. We trained several image fusion-based CNNs on medical image pairs and then using our FuseVis tool we performed case studies on a specific clinical application by interpreting the saliency maps from each of the fusion methods. We specifically visualized the relative influence of each input image on the predictions of the fused image and showed that some of the evaluated image-fusion methods are better suited for the specific clinical application. To the best of our knowledge, currently, there is no approach for visual analysis of neural networks for image fusion. Therefore, this work opens a new research direction to improve the interpretability of deep fusion networks. The FuseVis tool can also be adapted in other deep neural network-based image processing applications to make them interpretable.

Keywords: image fusion; convolutional neural networks (CNN); saliency visualization; jacobians; graphical user interface (GUI)

1. Introduction

The recent development of state-of-the-art imaging modalities has revolutionized the way we perform our everyday activities. For example, in self-driving cars, infrared images from the respective camera sensors positioned on the vehicle help to detect obstacles such as pedestrians at night. Remote sensing satellites, on the other hand, acquires multi-spectral and multi-resolution images that are needed for object detection and recognition from high altitudes. In medical diagnosis and treatment, Magnetic Resonance Images (MRI) provides a detailed view of the internal structures of the human brain such as white and gray matter whereas Positron Emission Tomography (PET) and single photon emission computed tomography (SPECT) images provide functional information such as glucose metabolism and extent of cerebral blood flow (CBF) or perfusion activity in the specific regions of the brain. However, it is challenging to analyze such complementary information provided by these image modalities individually. Multimodal image fusion-based image processing technique solves this problem by combining two or more pre-registered images from single or multiple imaging modalities

into a fused feature space. Ideally, a fusion algorithm should have low computational overload and it should also preserve input features for usage in the aforementioned real-world applications.

In the early years of image-fusion research, most of the methods had been proposed in a three-step approach. First, the input images were converted into several feature maps using an appropriate image transformation method such as pyramid decomposition, wavelet transformation, and sparse representation. Then, the coefficients of the multimodal feature maps were combined using a pre-defined fusion strategy to get fused feature maps. Finally, the fused image was reconstructed using the inverse transformation applied to the fused feature maps. However, for an ideal fusion model design, these methods mainly focused on enhancing the transformation and fusion strategies to pursue good perceptual results by defining some intricate design rules. Recently, several unsupervised machine learning-based CNNs have been proposed to achieve real-time image-fusion results. However, these image fusion-based neural networks lacks the trust of the end users since there are no tools to analyze the quality of the fused images as these networks act as black boxes.

To overcome this challenge, the neural network-based image-fusion methods evaluate the quality of the results using performance metrics such as Structural Similarity Index (SSIM) [1] that compares the perceptual similarity between the input images and the fused image. However, the shortcoming of such evaluation is that these metrics do not visualize the influence of input image features on the features of the fused image as it does not consider the underlying heuristics of the hidden network layers thereby not analyzing the internal mechanics of these networks. Therefore, image fusion using neural networks that have been evaluated on these performance metrics have low interpretability.

Recently, several visualization techniques have been proposed that broadly try to interpret the neural network decisions. For example, the gradient-based visualization methods backpropagate the gradients through the hidden layers of the neural network to interpret the per-pixel influence of the input image on the output predictions. However, such methods are specifically suited for problems such as image classification where the aim is to visually explain the class decision made by the neural network using jacobian-based saliency map for the class score with respect to the input image. On the contrary, the primary goal for interpreting per-pixel decisions in a fusion problem is to compute the jacobian-based saliency map for each pixel of the fused image with respect to each pixel of the input image. However, such analysis leads to very high computational overload since the number of backpropagation iterations is equivalent to the number of fused pixels. To circumvent this challenge, the deep learning-based software frameworks such as PyTorch and TensorFlow implicitly aggregate the per-pixel gradients from the input tensor elements to compute a single gradient value for each of the output tensor elements. In image fusion, this is not helpful since the gradient information related to how each pixel in the input image influences the machine decision is lost. Hence, the per-pixel saliency visualization by computing gradients of a neural network-based fusion model remains untapped in popular literature.

Therefore, in this paper, we present a novel per-pixel saliency-based visualization approach with real-time capabilities that will guide in the interpretability of image fusion-based neural networks. The major novelties of this work are summarized below:

- We trained several state-of-the-art fusion-based unsupervised CNNs under the same learning objective. Considering the importance of understanding the fusion black boxes in life sensitive domain such as medical imaging, we focused on interpreting the trained neural networks specifically for MRI-PET medical image fusion.
- We performed fast computation of per-pixel jacobian-based saliency maps for the fused image with respect to input image pairs. To the best of our knowledge, it is a first-of-its-kind technique to visualize fusion networks by considering the backpropagation heuristics that helps it to be more transparent in a real-time setup.
- We constructed guidance images for each input modality by using gradients of the fused pixel with respect to the input pixel at the corresponding location in the input image. We also interpreted the gradient values in each of the guidance images with the grayscale intensities of the fused image by combining the images in the color channels of an RGB image.

- We computed scatter plots between the gradients of the guidance images which provides a visual overview of the correlation between the influence of each of the input modalities. For example, a positive correlation will show that the input modalities influence the fused image equally.
- We developed an interactive Graphical User Interface (GUI) named FuseVis, that combines all the visual interpretation tools in an efficient way. The FuseVis tool allows the computation of saliency maps in real time on the mouseover at the pixel pointed to by the mouse pointer. Our code is available at https://github.com/nish03/FuseVis.
- Finally, we performed clinical case studies on MRI-PET image pairs using our FuseVis tool and visually interpreted the fusion results obtained from several different neural networks. We showed the usefulness of FuseVis in identifying the capability of the evaluated neural networks to solve clinically relevant problems.

Section 2 provides a detailed literature review of the existing works relevant to the problem of image fusion and visualization of neural networks. In Section 3, we explain the visual analysis goals, the proposed visualization concepts, and our FuseVis tool layout in detail. In Section 4, we provide the experimental details of the training setup, the hardware used, the architecture of the neural networks for image fusion, loss function used, and the hyperparameter studies conducted. In Section 5, we present a clinical application of MRI-PET medical image fusion and discuss key visualization requirements for case studies in this field. In Section 6, we interpret the visualization results obtained from each of the fusion-based neural networks using our FuseVis tool. In the same section, we also show the frame rates achieved during the mouseover operation that supports the real-time capabilities of our tool. Finally, in Section 7, we summarize the major contributions of this work and discuss its applicability in the field of real-time visualization of neural networks in general.

2. Related Work

In this section, we provide an overview of the literature relevant to this paper. For this, we divide this section into five categories. The first subsection summarizes the classical image-fusion approaches that involve non-machine learning-based techniques proposed in the recent past along with some relevant review papers related to image-fusion problem. The second subsection gives an overview of the methods that applied the classical image-fusion approach along with pre-trained deep neural networks to attain the fusion results. The next subsection discusses recent methods in the field of unsupervised end-to-end deep learning-based image fusion specifically in the area of multi-focus and multimodal image fusion. The fourth subsection gives a detailed overview of the recent state-of-the-art visualization techniques that have been proposed for interpreting and explaining the black-box neural networks in general. Finally, we end the related work discussion with current literature on the fast computation of gradients in a neural network setup.

2.1. Classical Image-Fusion Approaches

Many review papers in the past such as [2–4] provided an overview of different methods for multimodal and multi-focus image fusion. It is quite pertinent from these papers that the classical method for the fusion of multimodal images involves activity level measurement and pre-defined fusion rules. Activity level measurement aims to transform the input images to obtain salient features by either performing coefficient-based activity measurement [5], window-based activity measurement [6], or region-based activity measurement [7]. The transformation strategies were mainly categorized into domains namely multi-scale decomposition, sparse representation, and hybrid transformation, among others. This was followed by pre-defined fusion rules to output a fused image. Subsequently, the quality of the fused image was evaluated by using performance metrics that involve human visual perception and some objective statistical assessments.

Multi-scale decomposition (MSD) [5,7–21] separates the source image into low and high-frequency sub-bands to separately analyze the base and detail level features. Some of the popular transformation

strategies under multi-scale image decomposition has been pyramids [5,8], discrete wavelets [9–11], dual-tree complex wavelets [7], contourlets [12] and shearlets [13,14]. Multi-scale geometric analysis-based MSD methods such as non-subsampled contourlet (NSCT) [15,16] and non-subsampled shearlet transform (NSST) [17–19] have also been proposed due to their effectiveness with image representation. Edge preserving filters such as bilateral filters [20] and guided filters [21] have been popular in multi-scale decomposition-based literature since they retain the salient features from the input images. Sparse representation (SR)-based transformation strategies [22–27] do not decompose the original image into low and high-frequency bands but instead assume that both frequency bands have similar sparse coefficients. Hybrid transformation strategies aim to define more than one transformation procedure such as curvelet-wavelet [28], combination of multi-scale decomposition and sparse representation- based methods [29–31], Intensity-Hue-Saturation (IHS) and Principal Component Analysis (PCA) [32] among others.

Finally, the image-fusion rules after the image transformation have traditionally been defined to combine the features obtained from multi-scale transform approaches into a single fused image and been a crucial factor that helps to provide state-of-the-art fusion performance [4]. The three different components which together comprised a robust fusion rule in the past are coefficient grouping, coefficient combination as well as consistency verification. The coefficient grouping and combination strategies include choose-maximum rules [20], weighted-averaging [5] and guided-filtering-based weighted averaging [21]. Subsequently, the consistency verification aimed to ensure that the neighborhood coefficients are fused using the same fusion rule by refining the calculated weight map based on some priors [33]. At last, the fused image obtained from the classical methods were evaluated with several performance metrics [1,34–37] since there is no gold standard for an ideal fused image.

2.2. Mixture of Classical and Deep Learning-Based Fusion Approaches

In recent years, there have been several works that use deep learning-based fusion networks [38] to further enhance the fusion image quality. However, these works focused on using pre-trained neural networks to extract high and low-level image features for activity level measurement while the fusion rules were still manually defined. For example, in [39], a CNN for medical image fusion was used to perform activity level measurement and based on these measurements, coefficient grouping and coefficient combination were performed manually to obtain the fused image. On the other hand, in [40], the infrared and visible input images were first separated into low and high-frequency features using a pyramid-based multi-scale decomposition approach, and then the high-frequency features were fed into a pre-trained deep learning-based framework to extract multi-channel high-frequency features. Finally, the low-frequency features were fused using weighted averaging while the high-frequency multi-channel features were fused using a pre-defined max-selection rule. In [41], low-resolution grayscale images were divided into high and low-frequency features using wavelet transformation, and the high-frequency features were fed into a trained neural network which provided the high-resolution version of the input. Then, the absolute-max strategy was applied to the high-resolution outputs while the simple averaging was done to fuse the low-frequency features. Finally, the inverse wavelet transformation was done to retrieve a high-resolution output image. Ref. [42] on the other hand, used multi-scale convolutional neural networks to obtain multi-scale feature maps that are fused and then post-processed to obtain the decision maps which are segmentation results of the fused image using watershed transformation. This is done to evaluate the quality of the fused image from different image-fusion algorithms.

2.3. Unsupervised End-to-End Deep Learning-Based Fusion Approaches

All the fusion methods discussed up to now use the traditional approach of defining the transformation and fusion strategies while some use deep neural networks to extract feature maps from its hidden layers to perform the activity level measurement. Based on the feature maps generated, a final fusion rule is defined. The second observation with these methods is that the network is trained on image distributions which possess properties quite distinct from the inferred images. Though such usage of convolutional neural networks has provided good results in these approaches, it undermines

the true ability of a standalone end-to-end learning algorithm capable of providing a final fused result by eliminating pre- or post-processing steps such as traditionally used transformation and fusion strategies. Recently, there are end-to-end deep learning-based image-fusion methods that introduce an optimization strategy using a neural network that models all the transformation, fusion, and reconstruction strategies in its hidden network layers and learns a fused image in an unsupervised manner using loss functions such as SSIM [1]. Refs [43–45] were some of the early works in the field of multi-focus image fusion which used end-to-end convolutional neural networks that jointly generated activity level measurements and fusion rules. However, these methods trained the network in a supervised environment by leveraging the training data with available groundtruth from image classification databases. Ref. [46] on the other hand, proposed a simple encoder-decoder-based architecture for directly mapping the source images with a single-camera focus to get a multi-focus fused image while [47,48] proposed a Conditional Generative adversarial network (CGAN)-based approach for multi-focus image fusion. In the field of multimodal medical image fusion, Ref. [49] proposed a novel end-to-end fusion network to fuse MRI and PET image pairs by extracting high and low-frequency features within the network layers with SSIM as the loss function. Ref. [50] used a similar approach for fusing multi-exposure RGB images. However, they first converted the RGB images into YCbCr color space and used an end-to-end neural network to fuse only the luma component Y of the input images while the Cb and Cr components were fused using weighted-averaging strategy. Finally, the fused YCbCr color space was converted into the RGB fused color space. There have been several works in the field of visible and infrared image fusion as well that use unsupervised deep learning-based fusion framework. Ref. [51] trained an end-to-end convolutional neural network using visible and infrared pedestrian images from both day and night to attain robust fusion results. Ref. [52] used a different approach to fuse infrared and visible images where they used only a single modality at once to optimize the common weights. Secondly, the feature extraction and feature reconstruction layers involved the trainable weights while the fusion layer did not contain any trainable parameters. Ref. [53] went a step further and defined a trainable fusion layer. Refs [54–56] applied the GAN approach for fusing infrared and visible images using end-to-end neural networks. Ref. [57] was another such GAN-based approach but they used a single modality and fused multi-resolution input images. Finally, there were some works [58] which used a novel densely connected network using unsupervised learning strategy to fuse source images for multiple fusion tasks whereas [59] proposed a new GAN-based method that combined the identity of one input image and the shape of another input image.

2.4. Visualization Techniques

Recently, there have been several methods that interpreted the decisions made by neural networks by visualizing the model predictions. Ref. [60] proposed a method for visualization of classification models by computing the gradients of the class predictions with respect to the input images. Ref. [61] on the other hand, computed the gradients for activation neurons in the hidden layers of the network with respect to the input image by visualizing the specific regions of the image that activates a given neuron the most. [62] extended this work and generalized the gradient-based visual explanations proposed in [61] for different types of neural network architectures. Ref. [63] proposed an alternative visualization method where they used a relevance score approach that visualized the contributions of each pixel of the input image to the classifier predictions. Ref. [64] presented a method inherently similar to [63] but they also visualized the output prediction by calculating the contributions of all neurons in the network to every feature of the input image.

Perturbation-based visualization algorithms, however, do not need backpropagation heuristics and is a model agnostic approach. Ref. [65] was one of the first perturbation-based approaches that tried to implement occlusion in the input image and examined the output of the classification-based neural networks. The results showed that the model was not localizing the objects in the image as the probability of the correct class reduced significantly when the object was occluded. Ref. [66] was another model agnostic method where the perturbations were performed in the input image around its neighborhood

and the behavior of the model's predictions was later analyzed. Finally, the method weights the perturbed data with respect to its proximity to the original image and learns an interpretable model based on the old and the new predictions. Ref. [67] interpreted the model predictions by changing the pixel intensities of the input images with some noise such as blurring or occlusion and modeled the change in the prediction probability of the output image. Ref. [68] showed that there are two important characteristics of a visualization method to properly explain the regions of the input that are important for the model's decisions. These characteristics are gradients and implementational invariance. To satisfy both the properties, the method integrated the gradients from multiple inputs due to which the method was able to satisfy the sensitivity and implementational invariance of a visualization method. Therefore, the method was able to visualize the prediction of a neural network with respect to its inputs and required just a few instances of gradient computations. [69] proposed a visualization approach specifically for medical image fusion where they computed a heat map showcasing the mutual information between the source and the fused image. However, this method cannot interpret and explain the learning-based fusion networks itself since there are no backpropagation heuristics involved. Although these perturbation-based visualization approaches are generally applicable to a wide range of the problem, they still need several iterations to get the visualization results due to which they are not yet suitable for real-time deployment.

Crucially, none of the visualization methods can be generalized for fusion-based deep neural networks since, in a fusion problem, the number of output predictions is equal to the number of output pixels. This has huge memory and time requirements due to the very high number of backpropagation heuristics involved. Ref. [70] efficiently computed the gradients of the loss function by conducting the backpropagation heuristics for per-training example. However, the idea only works on a neural network with a single class prediction, and therefore, it cannot be applied to an image-fusion problem where there are a large number of predictions in the form of fused pixels. Therefore, an efficient method is needed to decrease the computational cost for the visualization of high-dimensional output predictions provided by deep neural networks.

3. Method

In this section, we will present the key visual analysis goals of a neural network for image fusion. Next, we will describe the proposed visualization concepts along with their mathematical formulations that help in achieving these visual analysis goals. Finally, we will present an overview of our user interface, FuseVis, and describe its user interaction capabilities.

3.1. Visual Analysis Goals

A fused image provides important information related to the input image modalities in combined feature space by preserving features from both the input images. However, by visualizing only the fused images, it could be challenging to interpret the sensitivity of the fusion methods for the input images. It is especially difficult to interpret which of the fusion methods reproduce features from which input image in a better way and how sensitive it is to changes in pixel intensities of the input images. Hence, this analysis is important to analyze the stability of a fusion network to the pixel intensity changes in the specific regions of the input images.

3.2. Visualization Concepts

We present several visualization strategies that help to understand the fusion results. The most important of those strategies is to compute gradient-based saliency heat maps that indicate the relevance of each input pixel to a single pixel in the fused image. These heatmaps are particularly valuable since they allow for an easy and intuitive investigation of what the respective fusion model perceived to be important in the input image features. Therefore, gradient-based saliency heatmaps are useful in interpreting fusion black boxes and understanding the characteristics of image fusion-based neural networks. Using these saliency-based heat maps, we can also estimate whether a neural network for image fusion is robust to

new examples. Therefore, to improve the interpretability of such image fusion-based neural networks, we present several visualization techniques that help to understand the fusion results.

3.2.1. Jacobian Images

We define per-pixel saliency heat maps, known as Jacobian images, which highlight the regions in the input images that most prominently influenced the per-pixel fusion results. Let us interpret a fusion operator as a mapping $f : \mathbb{R}^{2n} \to \mathbb{R}^n$ that maps two input images x_1 and x_2 to a fused image $y = f(x_1, x_2)$, where all images have the same dimensions and contain n pixels. Then, the saliency analysis is focused on a single pixel $i \in \{1, \ldots, n\}$ of the fused image that can be selected by the user interactively and is called principle pixel, $y^{(i)}$, where we use a superscript enclosed in parenthesis to denote the pixel of an image. For each input image $x \in \{x_1, x_2\}$, we define the jacobian image of $y^{(i)}$ as the image of the partial derivatives with respect to the input image x:

$$\frac{\partial y^{(i)}}{\partial x} = \left(\frac{\partial y^{(i)}}{\partial x^{(1)}}, \frac{\partial y^{(i)}}{\partial x^{(2)}}, \ldots, \frac{\partial y^{(i)}}{\partial x^{(i)}}, \ldots, \frac{\partial y^{(i)}}{\partial x^{(n)}} \right) \tag{1}$$

The jacobian computation involves backpropagating the gradients through the hidden layers of the neural network using chain rule-based automatic differentiation. A jacobian image visualizes the extent up to which each pixel element of the input image x influences the principle pixel $y^{(i)}$. Therefore, the jacobian images reveal the sensitivity of the principle pixel $y^{(i)}$ in the fused image to the changes in the pixel intensities of the input images. An example of a jacobian image is shown in Figure 1e. It can be expected that the fused principle pixel $y^{(i)}$ is highly sensitive to changes in the grayscale intensities at $x^{(i)}$. Additionally, the local neighborhood of the pixel element $x^{(i)}$ might also have a direct influence on the fused principle pixel $y^{(i)}$. The jacobian image-based visualization concept also helps to compare the sensitivity of the fused principle pixel with respect to changes in the pixel intensities of each of the two input images as well as estimate which of the two input images has a greater neighborhood influence on the fused pixel.

3.2.2. Guidance Images

Based on the observation that the pixel $x^{(i)}$ has by far the largest saliency value, we developed a visualization concept named Guidance images, which only considers the gradients of the principle pixel $y^{(i)}$ with respect to the input pixel element $x^{(i)}$ located at the corresponding coordinate in the input image:

$$\frac{\partial y}{\partial x} = \left(\frac{\partial y^{(1)}}{\partial x^{(1)}}, \frac{\partial y^{(2)}}{\partial x^{(2)}}, \ldots, \frac{\partial y^{(i)}}{\partial x^{(i)}}, \ldots, \frac{\partial y^{(n)}}{\partial x^{(n)}} \right) \tag{2}$$

Although jacobian images allow per-pixel interpretation of fusion outputs, guidance image aims to provide for each input image, a static overview of its influence on the fused image. An example of a guidance image is shown in Figure 1c. The two static guidance images provide an overview of the jacobian images and allow us to compare the major gradient values on an absolute scale. This fosters comparison of the same pixel location in the two input modalities as well as different pixel locations in the same input image. The changes in the pixel intensities of the input image region with high gradients will significantly change the pixel intensities in the corresponding region of the fused image. On the contrary, it will require significant changes in the pixel intensities of the input image region with low gradients to have a similar effect on the fused features in the same region. Due to this, the guidance images reveal which of the two input modalities have higher influence on the fusion result in specific regions of interest as well as where each input image has a large influence on the fusion result.

3.2.3. Guidance RGB Images

During the interactive visual analysis with jacobian and guidance images, a typical analysis task is to compare the values in the guidance images of the two modalities to find out which modality

influences the fused image more. Furthermore, one typically wants to compare a guidance image with the fused image. To simplify these local comparisons, we provide a single Guidance RGB image that encodes the two normalized guidance images as well as the fused image in the three color channels red, green, and blue. We perform a max-min normalization scaling of the guidance images to re-scale the features with a distribution value between 0 and 1. For every guidance image, the minimum value of that image gets transformed into 0, and the maximum value gets transformed into 1. An example of a Guidance RGB image is shown in Figure 1i. In red are the regions where the MRI gradient is high, the PET gradient is low, and the fused image has relatively low pixel intensities. By increasing fused pixel intensities, the color changes from red to magenta. Similarly, green and cyan correspond to high PET gradients with low MRI gradient values. In blue regions, both gradient values are low, and the fused pixels are bright. The yellow depicts regions where both gradients are high. In this way, the Guidance RGB image allows us to find different gradient constellations fast and the selection of the principle pixel is also possible on the Guidance RGB image.
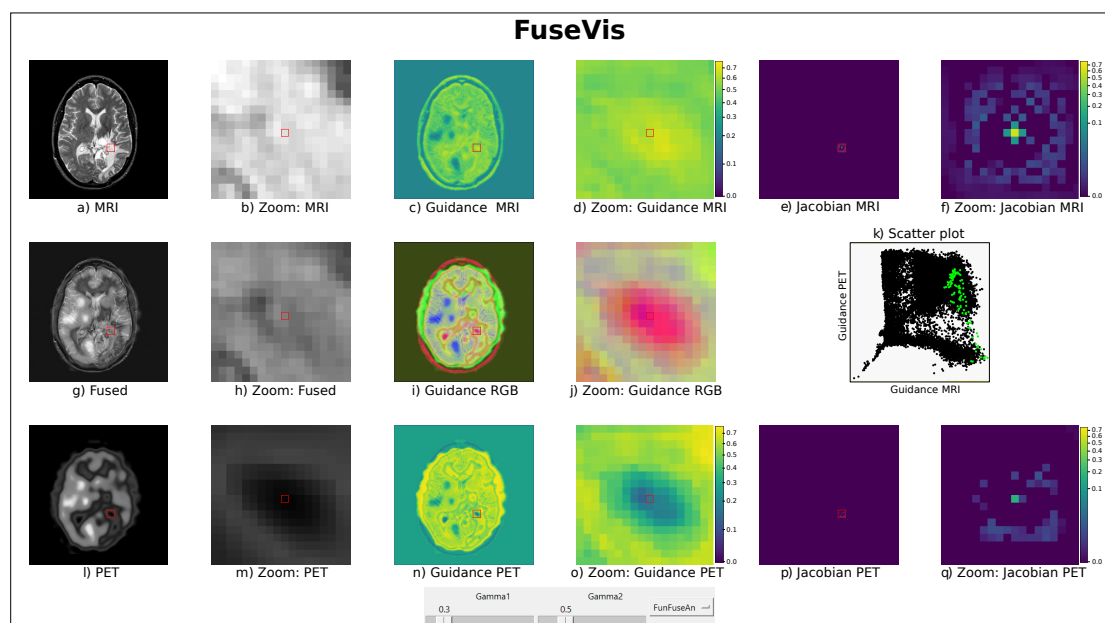

Figure 1. The figure shows the layout of FuseVis's user interface.

3.2.4. Scatterplot

As the Guidance RGB image gives a local overview of how the gradients of each modality behave in comparison to the fused image, a scatterplot can depict the correlation and more advanced statistical relationships between the gradient values of the guidance images. For example, a positive correlation between these gradients would mean that the input modalities equally influence the fused pixels while a negative correlation will show that an increase in the influence of one modality will lead to a decrease in the influence of the other modality. For this statistical analysis of the gradient values in the guidance images, we show a scatterplot with MRI gradients along the x-axis and PET gradients along the y-axis with an example provided in Figure 1k. The overall scatterplot in black shows a rather complicated relationship between the gradients. Therefore, we illustrated the points corresponding to pixels around the principle pixel in green. In this way, the local statistical relationships can be explored interactively.

3.2.5. Gamma Correction

The jacobian images inherently have low luminance since the gradients of the principle pixel is extremely bright compared to the gradient of its neighborhood pixels. Therefore, to better visualize the small gradient values in the jacobian images, we used the gamma correction, γ_{corr1}, on jacobian images as $\left(\frac{\partial y^{(i)}}{\partial x}\right)^{\gamma_{corr1}}$. To also adjust the luminance of the guidance images, we defined another gamma correction,

γ_{corr2}, on guidance images as $\left(\frac{\partial y}{\partial x}\right)^{\gamma_{corr2}}$ in an 8-bit range of RGB images. We chose two separate gamma corrections for jacobian and guidance images to independently fine-tune the luminance of these images. Gamma correction is an efficient method to improve the luminance since it only has a single parameter to fix for better visualization of the underlying gradient intensities. Currently, the range of the gamma correction values varies between 0.1 and 2 in our FuseVis tool. An example of the effect of γ_{corr1} and γ_{corr2} on jacobian and guidance images respectively is shown in Figure 2. It can be seen from the images that as the γ_{corr1} and γ_{corr2} decreased below 1, the luminance of the images increased due to which the small gradients became brighter while γ_{corr} above 1 decreased the luminance of the small gradients.
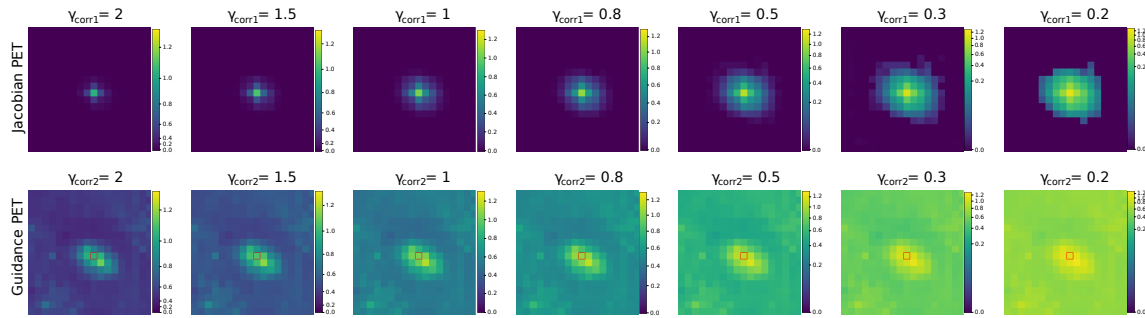

Figure 2. The first and second row shows the effect of the gamma correction on the jacobian PET and guidance PET images of MaskNet network with γ_{corr1} and γ_{corr2} varying between 0.1 and 2.0.

3.3. Overview of FuseVis Tool

An overview of the FuseVis tool is shown in Figure 1 where the images related to the MRI are placed at the top row, the images related to PET at the bottom row whereas all the modality combining images are in the middle row. We placed input and fused images on the left, guidance images in the middle, and the jacobian images on the right side of the tool. The tool also consists of a dropdown menu to select the fusion method and two separate slider widgets, namely *Gamma*1 that represents γ_{corr1} to fix the gamma correction of the jacobian images and *Gamma*2 that represents γ_{corr2} to fix the gamma correction of the guidance images. The principle pixel can be chosen in any of the input images, guidance images, or the fused image using the mouseover operation. The user performs the mouseover interaction by pressing and holding the left button of the mouse and hovering around the image. While performing this operation, the current mouse cursor position defines the principle pixel coordinate and the local environment around the principle pixel is displayed in the zoomed-in versions of the images which is always placed to the right of the respective images. The local environment of the principle pixel is defined based on the red squared bounding box and the current principle pixel coordinate, i, is positioned around the center of the bounding box which is also shown as the red square in the zoomed images.

4. Experimental Setup

As there are multiple real-world applications for image fusion, it is not possible to interpret fusion-based neural networks for each of these experimental settings. Therefore, we focus on the evaluation of fusion-based neural networks on a specific clinical application that deals with MRI-PET fusion for tumor resection planning. For this, we first comparably train the different neural networks. Then, we present medical case studies where we mention key visualization requirements of a robust fusion method. In the same section, we select the features in the input images that are most important for the clinical diagnosis. Finally, we use our tool to find out which of the fusion methods is best suited for this specific clinical application.

4.1. Fusion Networks

We visualized multiple end-to-end unsupervised learning-based fusion approaches available in popular image-fusion literature using our FuseVis tool. The image fusion-based neural networks

that we evaluated are FunFuseAn [49], MaskNet [71], DeepFuse [50] and DeepPedestrian [51]. All these fusion-based neural networks have distinct network architectures (as shown in Figure A1), built specifically for different types of image-fusion applications. In addition to these four end-to-end unsupervised fusion-based neural networks, we also evaluated a non-machine learning-based fusion approach by implementing a simple weighted-averaging technique and compared its visualization results with those from these fusion-based neural networks. The weighted averaging-based image fusion is a pixel-level method that assigns higher weights to the sharpest pixels in each input image. Assuming pixel $x_1^{(i)}$ from one input image to be extremely bright compared to pixel $x_2^{(i)}$ from the other input image, then the weighted-averaging method will give higher weightage $w_1^{(i)}$ to $x_1^{(i)}$ and lower weightage $w_2^{(i)}$ to $x_2^{(i)}$, thereby resulting in a bright fused pixel element $y^{(i)}$:

$$w_1^{(i)} = \frac{x_1^{(i)}}{x_1^{(i)} + x_2^{(i)}}, \quad w_2^{(i)} = \frac{x_2^{(i)}}{x_1^{(i)} + x_2^{(i)}}, \quad y^{(i)} = w_1^{(i)} * x_1^{(i)} + w_2^{(i)} * x_2^{(i)} \tag{3}$$

4.2. Hardware-Software Setup

The work was performed with Python (version 2.7) and deep learning library PyTorch was used to implement fusion-based neural networks. The Python wrapper of the GUI library, Tkinter, was used to create the user interface of our FuseVis tool. All the visualization results were derived from FuseVis tool using a 64-bit Windows operating system equipped with a single GeForce GTX 1080 Ti GPU and 12 Intel Core i7-8700K CPU @ 3.70 GHz with 64-GB RAM. The same hardware-software setup was used for the real-time computation of the jacobian images. However, to fasten the training of the neural networks and to speed up the static computations of guidance images, we used a more powerful hardware setup containing six NVIDIA Tesla V100 GPUs.

4.3. Training Dataset

For training each of the fusion networks in a uniform setup, we obtained several MRI-T2 and PET image pairs of 50 different patients publicly available at the Alzheimer's Disease Neuroimaging Initiative (ADNI) [72] with the age of patients varying between 55–90 years among both genders. The patients were chosen in such a way that the training dataset covers mild, moderate, and severe stages of Alzheimer's disease. All images were extracted as axial slices with a voxel size of $1.0 \times 1.0 \times 1.0$ mm^3. The MRI-T2 images were N3m MPRAGE sequences while PET images were co-registered, averaged, and standardized with a uniform resolution for each of the subjects. We aligned the MRI-PET image pairs using the Affine transformation tool of the 3D Slicer registration library.

4.4. Network Architectures

The architectures of the evaluated fusion-based neural networks are shown in Figure A1. A Conv $3 \times 3 \times 9 \times 16$ block, for example, means a convolution operation performed on an input with nine channels that results in an output with 16 channels. For this, 16 kernel filters each with size 3×3 and nine channels are used. Each channel of the kernel filter is convoluted with the corresponding input channel and subsequently summed up to generate a single feature map. This operation is repeated for all the kernel filters to extract output feature maps equal to the number of filters. The weights of the kernel filter have been initialized based on a uniform distribution while the striding and the padding operation have been performed in such a way that there is no downsampling in any hidden layer. A BatchNorm means a batch normalization operation and LReLU means a leaky ReLU activation function with a slope 0.2. It is noticeable that each of the networks has different architectures due to varying image-fusion tasks that they solve. The architecture of FunFuseAn [49] is inspired by the classical non-machine learning-based image-fusion technique of feature extraction, fusion, and feature reconstruction with specific application to the MRI and PET image pairs. MaskNet [71] concatenates the input images and performs a dense feature fusion in the initial layers of the network before constructing a fused image. DeepFuse [50] defines an

architecture where the feature extraction layers have tied weights for each of the input image modalities. This helps in increasing the brightness of an underexposed input image by using the luminance of an overexposed input image. DeepPedestrian [51] also concatenated the input images and used a deep architecture with multiple residual blocks specifically for infrared and visible image fusion. Since we aim to interpret the behavior of these image fusion-based neural networks by performing a per-pixel gradient-based saliency visualization using our FuseVis tool, we modified the original training setup of these fusion networks and defined a common strategy to train these networks.

4.5. Loss Function

An image-fusion approach aims to preserve as much of the input image features as possible in the fused image. Therefore, the training objective should be defined in such a way it could measure the similarity between the input and the fused image and optimize the network parameters to achieve perceptually similar results. Structural Similarity Index (SSIM) [1] is an efficient similarity metric that is also differentiable due to which it can be adopted as a primary loss function for training each of the fusion-based neural networks. However, empirically setting only SSIM as the loss function leads to the change of brightness in the final fused image. The ℓ_2 loss helps to overcome this shortcoming by preserving the luminance component better. Therefore, we define a combination of SSIM and ℓ_2 losses to train each of the neural networks as shown in Equation (4). The x_1 and x_2 are MRI and PET training images respectively, the regularization hyperparameter λ controls the weightage of the total L_{SSIM} and total L_{ℓ_2} losses while the hyperparameters γ_{ssim} and γ_{ℓ_2} controls the weightage between the individual SSIM and ℓ_2 losses of each input image modality, respectively.

$$L_{SSIM}^{MRI} = 1 - SSIM(x_1, y), \quad L_{SSIM}^{PET} = 1 - SSIM(x_2, y)$$
$$L_{SSIM} = \gamma_{ssim} * L_{SSIM}^{MRI} + (1 - \gamma_{ssim}) * L_{SSIM}^{PET}$$
$$L_{\ell_2}^{MRI} = ||y - x_1||_2, \quad L_{\ell_2}^{PET} = ||y - x_2||_2 \tag{4}$$
$$L_{\ell_2} = \gamma_{\ell_2} * L_{\ell_2}^{MRI} + (1 - \gamma_{\ell_2}) * L_{\ell_2}^{PET}$$
$$L_{total} = \lambda * L_{SSIM} + (1 - \lambda) * L_{\ell_2}$$

The loss convergence curves of all the evaluated unsupervised learning-based fusion networks are shown in Figure 3. Each of the fusion networks was trained for 200 epochs, the batch size was fixed as 2 and the Adam optimizer was used as the optimization function during the backpropagation step with a learning rate of 2×10^{-3}.

4.6. Hyperparameter Tuning

Ideally, a fused image should preserve features from both the input images equally and there should be minimal addition of the brightness artifacts. For achieving this, we followed a hyperparameter tuning strategy where we aimed to find a plausible balance between the L_{SSIM}^{MRI} and L_{SSIM}^{PET} as well as $L_{\ell_2}^{MRI}$ and $L_{\ell_2}^{PET}$ loss curves by training a single fusion network with multiple combinations of λ, γ_{ssim} and γ_{ℓ_2} values. Once a good balance was found between these curves, we evaluated the best λ configuration that leads to the least brightness artifacts by analyzing the resultant fused images. We initially trained the FunFuseAn network with $\lambda = 0.01, 0.2, 0.5, 0.8$ and 0.99 with γ_{ssim} and γ_{ℓ_2} each equal to $0.1, 0.3, 0.5, 0.7$ and 0.9. Therefore, a total of 125 instances of FunFuseAn network were learned in separate training environments. For each λ configuration, we analyzed the combination of γ_{ssim} and γ_{ℓ_2} that nearly balances the SSIM and ℓ_2 loss curves for both MRI and PET images. Then, we fine-tuned γ_{ssim} and γ_{ℓ_2} parameters for each λ by training several more instances of the FunFuseAn network in the vicinity of the previous best γ_{ssim} and γ_{ℓ_2} to find the next best configuration that balances the loss curves even better.

After fixing the best γ_{ssim} and γ_{ℓ_2} configuration for each λ, we determined which of the λ values leads to negligible brightness artifacts in the fused image. Therefore, we trained new network instances with additional values of $\lambda = 0.9, 0.999, 0.9999, 0.99999, 0.999999$ and 1.0 to properly evaluate the corner cases of our loss function. We observed that the brightness artifact in the fused images worsened

as the values of λ increased above 0.999 with $\lambda = 1.0$ having the most intense brightness artifact in the fused image. $\lambda = 0.99$ was one configuration where we observed negligible change in brightness of the fused image which meant a weightage of 0.01 to ℓ_2 was sufficient to remove the brightness artifact problem from the SSIM loss function. $\lambda = 0.999$ also had negligible brightness addition but the fused image from this configuration lost some PET features due to which we fixed $\lambda = 0.99$ as the best hyperparameter for the FunFuseAn network. Since $\lambda = 0.99$ was sufficient to overcome the brightness artifact problem, we used the same $\lambda = 0.99$ value for all the other fusion networks and then evaluated several instances of these networks to obtain the best γ_{ssim} and γ_{ℓ_2} that balances the loss curves. The loss curves in Figure 3 were generated with the best hyperparameter configuration for each of the fusion-based neural networks. The final hyperparameter values for the fusion networks can be seen in Table 1. It is observable that all the networks were able to balance the loss curves while DeepFuse network converged at higher loss values compared to the other fusion networks.
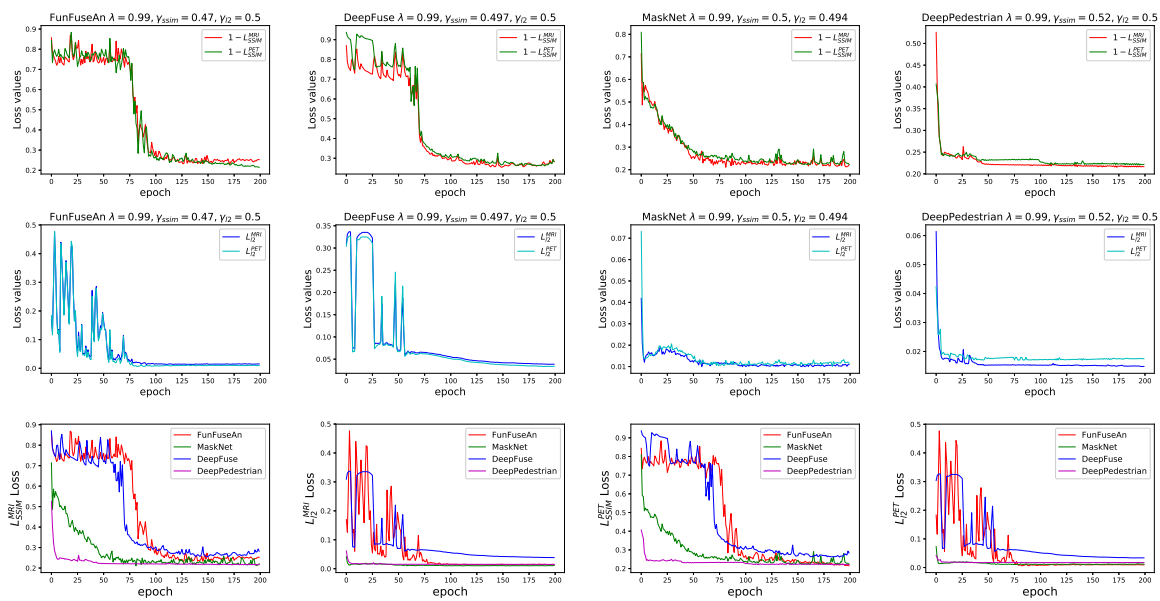

Figure 3. Loss curves of the evaluated fusion based neural networks trained on MRI-PET image pairs.

Table 1. The table shows the partial loss values of the trained fusion networks after 200 epochs and the fine-tuned hyperparameter configurations.

Network	λ	γ_{ssim}	γ_{ℓ_2}	L_{SSIM}^{MRI}	L_{SSIM}^{PET}	$L_{\ell_2}^{MRI}$	$L_{\ell_2}^{PET}$
FunFuseAn	0.99	0.47	0.5	0.2524	0.2147	0.0148	0.0094
MaskNet	0.99	0.5	0.494	0.2208	0.2236	0.0109	0.0115
DeepFuse	0.99	0.497	0.5	0.2824	0.2830	0.0385	0.0338
DeepPedestrian	0.99	0.52	0.5	0.2164	0.2219	0.0148	0.0175

5. Medical Case Studies

In this section, we discuss the specific clinical application of the MRI-PET image fusion in detail. Then, we perform case studies on pre-registered MRI and PET image pairs by providing the pathological information about the patients from whom these images were acquired. Finally, we will present key visualization requirements that will make a fusion approach suitable for usage in this specific clinical application.

5.1. Glioma and Its Pathological Features

Glioma is a type of brain tumor that originates in the glial cells that surround and support the neurons in the brain. MRI enables the clinicians to estimate the size and the location of glioma, thereby acting as an initial marker for malignancy. For example, in an MRI-T2 grayscale image, brighter pixel intensities imply

hyperintense signal abnormalities which are observable in the brain regions with a tumorous lesion such as glioma. The grayscale intensities in these tumorous lesions could be similar to the grayscale intensities in the lateral ventricle region of the brain that mostly contains cerebrospinal fluid (CSF) with high water content. However, there could be several sub-regions within the boundary of high-grade glioma with varying pathological features, which is challenging to visualize by only interpreting the MRI images as the MRI contrast is quite uniform within the tumor boundary. Hence, the aggressiveness and increased potential of the sub-tumor tissues cannot be adequately determined.

Functional imaging modalities such as PET and SPECT are better equipped to differentiate between the sub-regions of glioma by visualizing the functional information like glucose metabolism and the extent of CBF or perfusion activity in these sub-regions. The innermost part of glioma is the necrotic core that contains dead tissues and there is no possibility of CBF and/or glucose metabolism due to which it has very dark PET features. The region that surrounds necrotic tissues is called as an enhancing tumor region, which is filled with inflammation fluid and the blood-brain barrier leaks in this region due to cancerous angiogenesis. Hence, this region has glucose metabolism with blood supply to support the cancerous cell growth resulting in lighter dark PET features. The necrotic core and enhancing tumor region constitute the bulk of glioma that clinicians want to visualize using fused MRI and PET features and perform a precise resection of these two regions. The non-enhancing solid tumor region is the outermost region of glioma which is generally not resected since the blood-brain barrier is still intact in these tissues. The bright regions in the PET image resemble healthy tissues with a high level of blood perfusion and normal glucose metabolism.

5.2. Clinical Test Examples

The clinical test examples were acquired from Harvard Whole Brain Atlas database [73] with the combination of MRI-T2 and PET/SPECT images from patients suffering from different types of glioma. The test examples were disjoint to the training dataset and the network has not seen such clinical pathology during its training. The first clinical MRI-PET image pair is shown in the first four columns of Figure 4. The scans are of a patient who was suffering from Anaplastic Astrocytoma, a rare and malignant brain lesion classified under the category of high-grade glioma. A lesion in the right and the left side of the brain is visible in the MRI image and has bright grayscale intensities. On the other hand, bright regions in the PET image convey normal blood flow while very dark regions suggest no blood flow in the necrotic tissues. In the second clinical MRI-PET image pair as shown in the last four columns in Figure 4, the patient had a long history of tobacco usage and was originally suffering from Metastatic Bronchogenic Carcinoma, which is a type of lung cancer. The patient began having headaches and the scans revealed brain metastases that occurred due to the spread of cancer cells present in the lungs to the brain resulting in the diagnosis of glioma. A lesion in the right side of the brain with bright features is visible in the MRI image while very dark PET features reveal no blood flow in the necrotic region.

5.3. Visualization Requirements

As the anatomical MRI image cannot precisely estimate the sub-tumor boundaries within glioma and the functional PET/SPECT image is unable to model the overall extent of the tumor; therefore, one of the goals of MRI-PET image fusion is to better delineate the sub-peripheries of glioma by envisioning the metabolic and/or CBF characteristics of its sub-regions. However, it is challenging to interpret the suitability of these fused images for the mentioned clinical application, if the results are not supported by supplementary visualization tools. This secondary visual analysis becomes critically important when several fusion methods provide similar results. Additionally, the recent image-fusion methods are developed using unsupervised learning-based deep neural networks that are non-transparent (i.e., a 'black box') due to its difficulty to retrace the fusion decision in light of huge parameter space. This shortcoming is especially problematic in sensitive medical domains such as MRI-PET image fusion where understanding and explaining fusion results obtained from neural networks is required for a robust clinical decision. Therefore, the following visual tools are required to understand the suitability of a fusion approach for usage in the clinical application:

- A fusion approach should assist clinicians in visualizing the extent of hyper dark PET regions resembling necrotic core with no blood flow being superimposed on the bright anatomical boundary of the whole tumor mass. This information is important for clinicians to estimate the extent up to which a tumor resection is required. For example, in the first and the fifth column of Figure 4, the principle pixel in this very dark PET region was chosen for visual analysis.

- A fusion approach should preserve the very bright PET features which convey high blood perfusion and normal metabolism in healthy brain tissues as it helps clinicians in visualizing the regions with high brain activity due to external stimuli at a particular time. For example, in the third and seventh column of Figure 4, the principal pixel in the bright PET region was chosen for visual analysis.

- A fusion approach should be stable and less sensitive to changes in input features from a clinically less significant modality. For example, the change in grayscale MRI intensities within the necrotic core shall not highly influence the fused grayscale intensities as it might corrupt the clinically important dark PET features. For example, the MaskNet and DeepPedestrian networks are less sensitive to changes in the MRI features which can be visualized in the guidance MRI and guidance PET images of these networks shown in Figure 5.

- A fusion approach should be less sensitive to the changes in grayscale pixel intensities located in one sub-region of glioma (say enhancing tumor) when the principle pixel is in the other regions of glioma (say necrotic core). Therefore, a fusion method should have a negligible influence of the neighborhood pixels exterior to a local feature with the principle pixel interior to the local feature. For example, the fusion methods such as Weighted Averaging and MaskNet have no or very low gradients in the neighborhood pixels which are outside the very dark PET features resembling necrotic core as shown in the jacobian images in Figure 7.
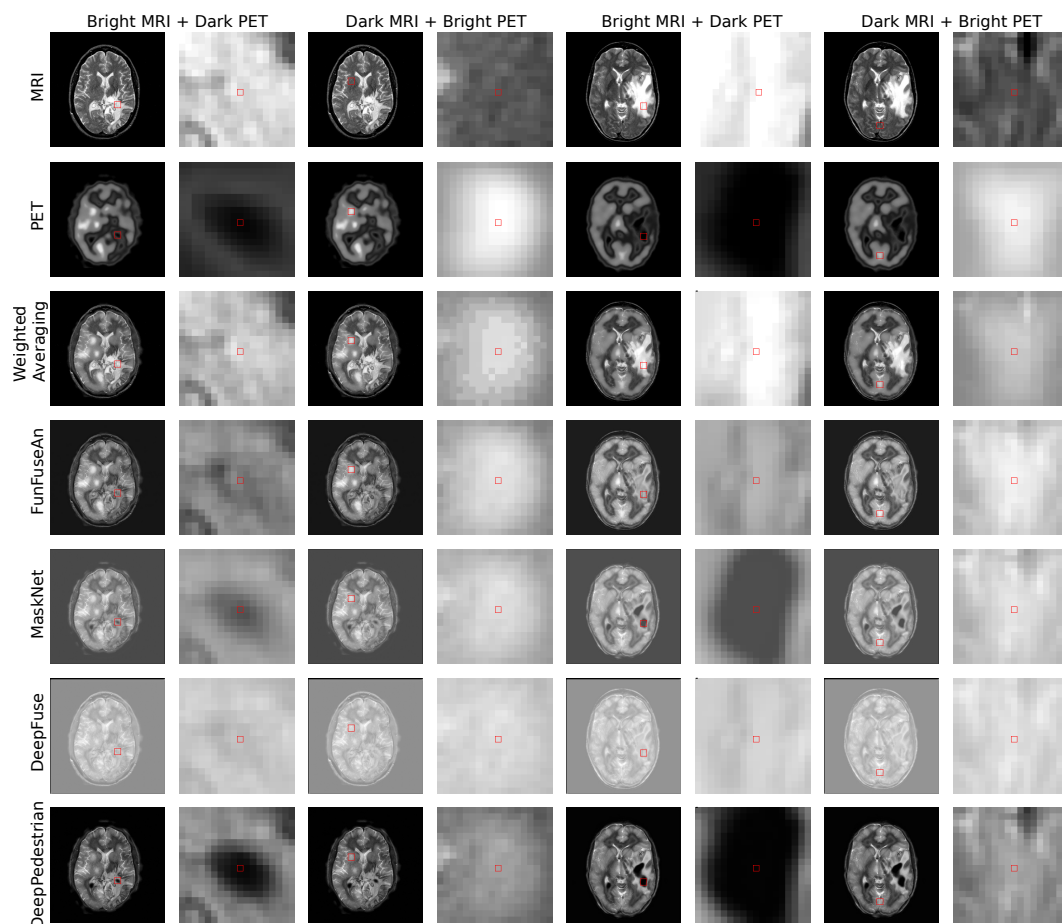

Figure 4. The figure shows the fusion results for each of the fusion methods. The zoomed image within the clinical region of interests is always placed on the right of the unzoomed image.
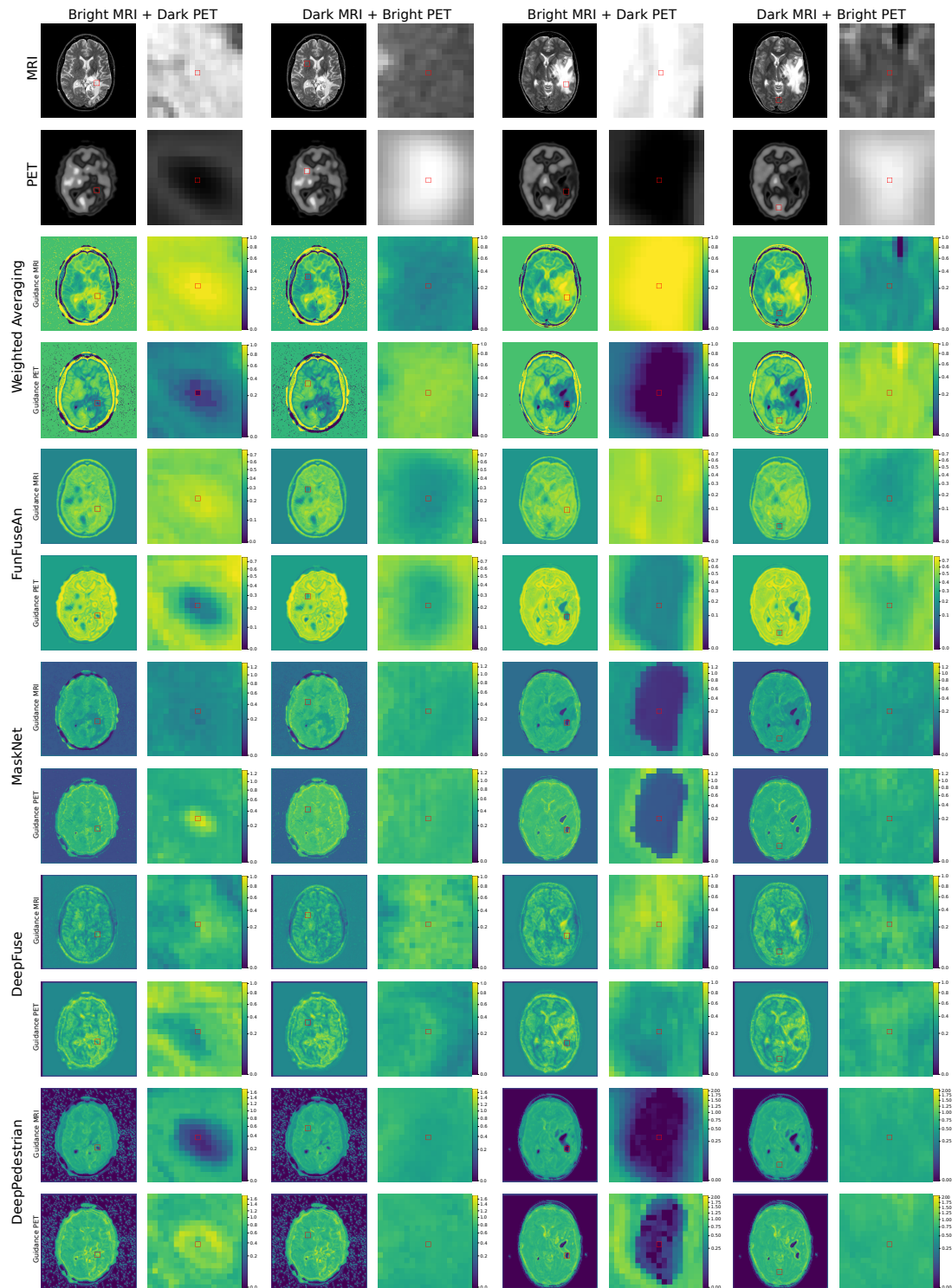
Figure 5. The figure shows the guidance MRI and guidance PET images for each of the fusion methods in the clinical region of interests. The γ_{corr2} was fixed at 0.5 for all the guidance images.

6. Results and Discussion

In this section, we present a detailed visual analysis of the fusion-based neural networks using our FuseVis tool. First, we will analyze the fused images from each of the fusion methods and discuss their clinical significance. Then, we will discuss the visualization insights by performing the saliency analysis and examine the feasibility of the fusion methods for usage in the clinical application. Finally, we will present the timing results of our FuseVis tool and show the tool's real-time capabilities.

6.1. Fused Images

The fused images of the Weighted Averaging method as shown in the third row of Figure 4 gives very high weightage to the bright input intensities in the regions where intensities in the other modality are darker. Therefore, we can perceive a reproduction of features from relatively brighter regions of the MRI and PET images due to which Weighted Averaging method is unable to preserve the clinically important dark PET features related to the necrotic core but favorably preserves the bright PET features resembling healthy tissues. The fused images of the FunFuseAn as shown in the fourth row of Figure 4 are comparatively similar to the Weighted Averaging approach where it is not able to preserve the very dark PET features. However, the fused images from FunFuseAn look relatively dull compared to the fused images from Weighted Averaging since the FunFuseAn network mixes both MRI and PET features even though the boundary information about the necrotic region is lost.

The analysis of the fused images from MaskNet in the fifth row of Figure 4 shows a significant loss of anatomical edges from MRI such as the brain skull. However, contrary to Weighted Averaging and FunFuseAn, MaskNet preserved the PET features better in both the dark and bright regions resembling the necrotic core and healthy tissues, respectively. However, the dark PET features are slightly blurred due to changes in the overall brightness of the fused image. The fused image obtained from DeepFuse as shown in the sixth row of Figure 4 has an even higher shift in grayscale intensities due to which the method is unable to preserve the dark PET features even though the overall anatomical MRI structures are well preserved. The change in brightness can be explained by the fact that the $L_{\ell_2}^{MRI}$ and $L_{\ell_2}^{PET}$ converged at higher loss values compared to other networks due to which the brightness component of SSIM was not properly optimized. Additionally, the architecture of DeepFuse network has been crafted for adding exposure to underexposed images by using the brightness component from each of the input image modalities and employ it to generate very bright fusion results. The fused image from DeepPedestrian as shown in the seventh row of Figure 4 clearly delineates the boundary of the necrotic core by preserving the very dark PET features, which is of high clinical significance for medical professionals. However, the anatomical edges from MRI are lost in the fused image which is also one of the main artifacts in the MaskNet network. Another important observation is that the bright PET features resembling healthy tissues appear to be not well preserved due to an overall brightness shift.

Summary

The analysis of the fused images shows that even though SSIM was used as a natural loss function for training the fusion networks in a common training setup, each of the fusion methods displayed very different fusion results. The MaskNet and DeepPedestrian methods provide clinically relevant results where the boundaries of the necrotic core of glioma are perceivable with the dark PET features clearly superimposed on the anatomical tumor features of MRI in the fused image. However, these methods do not provide loss-free fusion results as there are missing anatomical information in both these approaches while the dark PET features are not entirely preserved in the MaskNet network. Crucially, it is not clear upto what extent, the changes in the bright MRI features can affect the final fusion results within the clinically significant dark PET regions. This analysis is important for visualizing the stability of these networks to changes in the MRI features which is a critical visualization requirement. Hence, we will perform this analysis using the visualization concepts in the subsequent sub-sections.

6.2. Guidance Images

In this section, we analyze the fusion methods with respect to its sensitivity to feature level changes in the input images by visualizing the Guidance MRI and Guidance PET images in Figure 5 and Guidance RGB images in Figure 6 for the two clinically relevant regions. We chose the first region as bright MRI with very dark PET features since it is the region that contains the necrotic core of glioma which clinicians are interested to operate for resection. The second region of interest has dark MRI with bright PET features resembling healthy tissues which is interesting for the visualization of high brain

activity due to external stimuli. Ideally, a guidance MRI image should have low gradients meaning a fusion method should be less sensitive to changes in both dark and bright MRI regions.
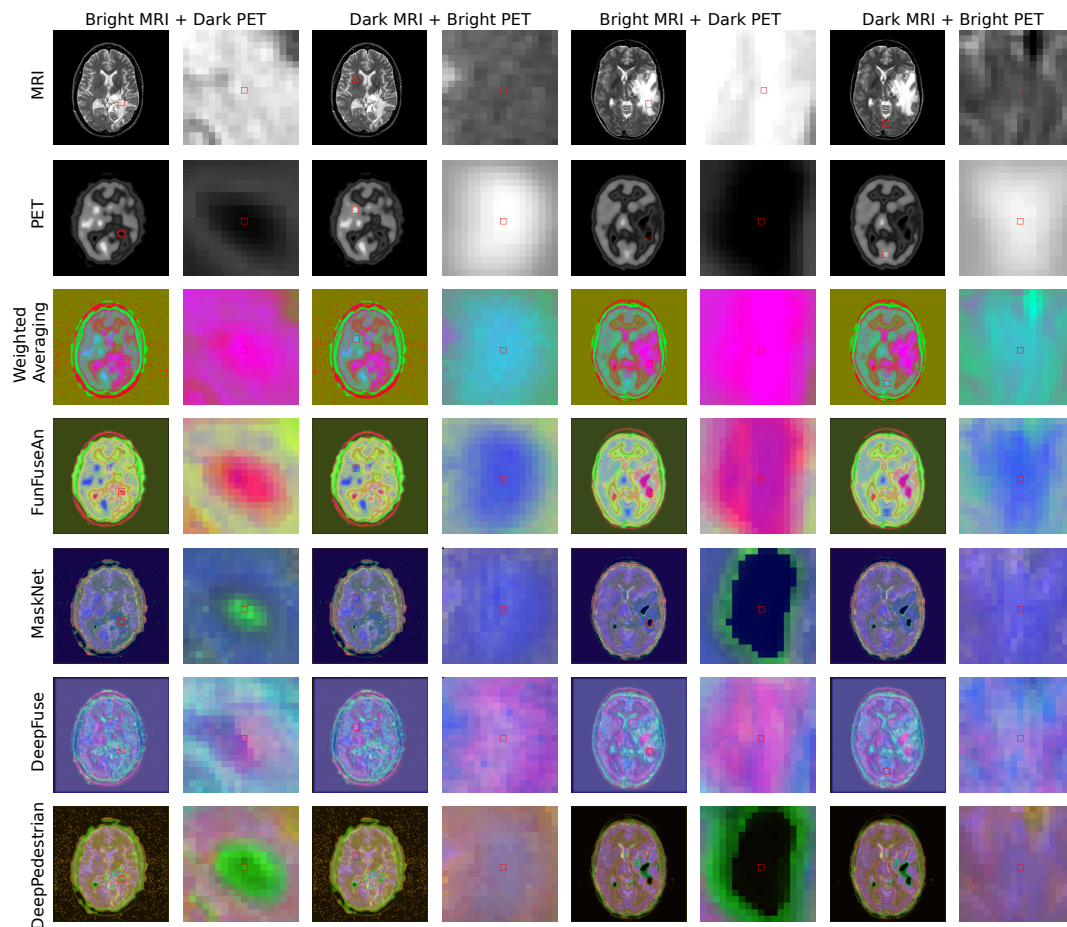

Figure 6. The figure shows the guidance RGB images for each of the fusion methods.

6.2.1. Weighted Averaging

For the region with bright MRI and dark PET intensities in both the test examples, it can be observed that the guidance MRI image is significantly brighter than the guidance PET image, revealing higher sensitivity to changes in MRI features. This suggests that a small change in the MRI pixel intensities will sharply modify the fused pixel intensities while it requires a substantial alteration in the PET pixel intensities to hold a considerable effect on the fused pixel intensities. This is an undesired outcome as a fusion method should be stable to changes in the pixel intensities of MRI. The guidance RGB image shows magenta color in these regions as the red channel containing guidance MRI image has high gradients and the blue channel containing the fused image have high pixel intensities compared to the green channel containing the guidance PET image with low gradients. On the other hand, the region with dark MRI and bright PET intensities shows an adverse effect where the guidance PET image is relatively brighter compared to the guidance MRI image since higher weightage is now given to the bright PET grayscale values. The guidance RGB image in this region shows cyan color since the green channel with guidance PET image has high gradients while the blue channel with the fused image also has higher intensities.

6.2.2. FunFuseAn

For the region with bright MRI and dark PET intensities, there are higher gradients in the guidance MRI image compared to the guidance PET image. This conveys that the fused pixels within the necrotic core have a higher sensitivity to changes in the MRI pixels. The guidance RGB image also shows red

color within the necrotic region indicating only the guidance MRI have high intensities compared to the images in other color channels. For the region with dark MRI and bright PET intensities, the guidance MRI and guidance PET images have low gradients. This means that the fused pixels in this region are quite stable to changes in the MRI and PET pixel intensities. Hence, the color in the guidance RGB image is predominantly blue as the fused pixel intensities are higher than the gradient values in both the guidance images. Therefore, it is challenging to interpret which of the two input modalities has a higher influence on the fused image in this particular region.

6.2.3. DeepFuse

In the bright MRI and dark PET region resembling necrotic core, there are relatively high gradient values in the guidance MRI image compared to the guidance PET image. This is the reason for the shades of magenta in the guidance RGB image. For the dark MRI and bright PET intensity region in the first test example, the pattern is similar where there are higher gradients in the guidance MRI image compared to the guidance PET image due to which magenta color is visible in the guidance RGB image. For the second test example, however, the guidance images from both MRI and PET have low gradients indicating a low influence of change in input pixel intensities within this region. Hence, the guidance RGB image also appears to be light blue in color in this region.

6.2.4. MaskNet and DeepPedestrian

For the necrotic region with bright MRI and dark PET intensities in the first test example, the guidance PET image is relatively brighter than the guidance MRI image for both MaskNet and DeepPedestrian networks. This conveys that both the networks are highly sensitive to the changes in very dark PET intensities of the necrotic region compared to the changes in MRI intensities. Since both the guidance MRI and the fused image have low intensities compared to the guidance PET image in this region, the guidance RGB image exhibits green color. However, for the second test example, both the guidance images have very low gradients in the same region, showcasing the stability of the networks for pixel intensity changes in this region. Since all the color channels in the guidance RGB image have low intensities, the guidance RGB image does not have a fixed color scheme and is dark. For the region with dark MRI and bright PET intensities, however, the behavior is similar in both the test examples where both the guidance images have low gradients due to which it is challenging to differentiate the influence of each of the modalities while the MaskNet network has learned the bright PET features better, leading to blue color in the guidance RGB image.

6.2.5. Summary

The guidance images provided a static overview of the influence of input principle pixel on the fused principle pixel and assisted in visualizing which of the two modalities has higher influence in the clinical region of interest. Therefore, it provided new insights related to the stability of fusion networks in these regions which were not perceivable by looking at the fused images. The guidance images clearly show that MaskNet and DeepPedestrian networks performed very differently compared to other fusion methods. For the region with the dark PET features resembling necrotic core, all the methods except MaskNet and DeepPedestrian were sensitive to changes in the bright MRI intensities which is not suitable for a reliable analysis of the necrotic tumor boundary since the dark PET features might not be properly preserved in the fused image due to changes in the MRI features. Both MaskNet and DeepPedestrian preserved the dark PET features in the fused image and were quite stable to the changes in the MRI pixel intensities due to which both these methods are far more suitable for the clinical application than any other fusion methods. However, the guidance images do not reveal the influence of neighborhood pixels on the fused principle pixel which is important for estimating a better fusion approach between the two of these networks.

6.3. Jacobian Images

The jacobian images of the selected principle pixel is shown in Figure 7. The analysis of the jacobian images will help in understanding the influence of the neighborhood pixels located around the principle pixel in the clinically relevant regions of the input image. It will also assist in visually comparing the neighborhood influence of MRI pixels with PET pixels.
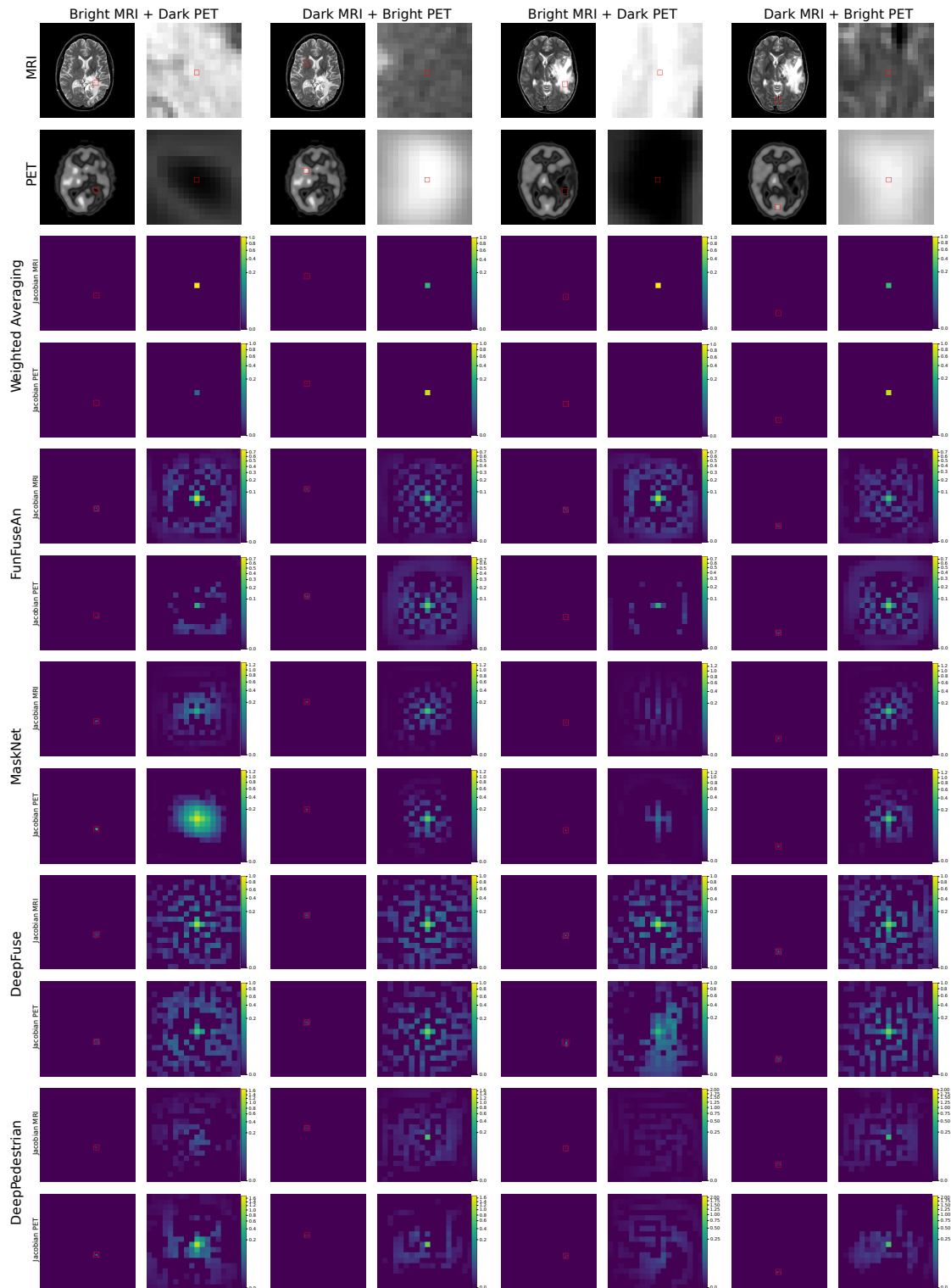

Figure 7. The figure shows the jacobian MRI and jacobian PET images for each of the fusion methods in the clinical region of interests. The γ_{corr1} was fixed at 0.3 for all the jacobian images.

6.3.1. Weighted Averaging

The jacobian MRI and jacobian PET images within the region with bright MRI and dark PET intensities for both the test examples show a higher gradient of the principle pixel in the jacobian MRI image compared to the jacobian PET image. For the dark MRI and bright PET intensities, there is a relatively high gradient in the principle pixel of the jacobian PET image showcasing sensitivity to changes in the bright PET pixel. One crucial observation from each of the jacobian images is that there is no influence of the neighborhood pixels on the prediction of the fused principle pixel since Weighted Averaging performs the per-pixel computation of fused pixel intensities.

6.3.2. FunFuseAn

After analyzing the jacobian MRI and jacobian PET images within the region of bright MRI and dark PET intensities, it can be observed that similar to the Weighted Averaging method, there is a higher gradient value for the principle pixel of the jacobian MRI image. The immediate neighborhood pixels in the jacobian MRI image also have significant gradient values meaning these pixels also influence the outcome of fused principle pixel. This is clinically useful since the change in intensities of pixels located within the necrotic core should affect the fused principle pixel. Additionally, there are more positive gradient values for the outer neighborhood pixels in the jacobian MRI image than the jacobian PET image. For the combination of dark MRI and bright PET intensities, however, there are low gradients for the principle pixel in both jacobian MRI and jacobian PET images. This conveys that the fused principle pixel is relatively stable to small changes in the bright PET and dark MRI pixel intensities. Hence, visualizing principle pixel in the jacobian images does not give much information about which input principle pixel has a higher influence on the fused principle pixel. On the other hand, the neighborhood pixel influence on the fused principle pixel is quite widespread due to the large number of positive gradients in both the jacobian MRI and jacobian PET images.

6.3.3. MaskNet

For the combination of bright MRI and dark PET intensities, the jacobian images in the first test example convey quite distinct characteristics compared to the jacobian images in the second test example. For the first test example, the jacobian PET image has a much higher gradient for the principle pixel compared to the jacobian MRI image. This reveals a high sensitivity of the fused principle pixel within the necrotic tumor core with respect to small changes in the pixel intensity of the PET principle pixel whereas the fused principle pixel is quite stable to changes in the MRI principle pixel. Additionally, the influence of neighborhood pixels of PET within the dark necrotic tumor core on the fused principle pixel is highly intense compared to neighborhood pixels of MRI. This conveys that the changes in the pixel intensities of the PET inside the dark necrotic core can greatly affect the fused principle pixel while the PET pixels outside the necrotic core boundary have a negligible effect on the fused principle pixel. However, in the second test example, both the jacobian MRI and jacobian PET images have low gradients for the principle pixel as well as in the neighborhood region within the necrotic core. This demonstrates that the network requires quite significant changes in these features to change the fusion output. For the combination of dark MRI and high PET intensities in both the test examples, the jacobian MRI and jacobian PET have similar properties to FunFuseAn where there are low gradients for the principle pixel in both the images; therefore, equally influencing the fused principle pixel. However, there are fewer neighborhood pixels in the jacobian MRI and jacobian PET images that influences the principle pixel when compared to the FunFuseAn network.

6.3.4. DeepFuse

The jacobian images in both the test examples, when visualized in the regions of bright MRI and dark PET intensities show a significantly high number of neighborhood pixels influencing the fused principle pixel while a high gradient value for the principle pixel can be seen in the jacobian MRI

image compared to the jacobian PET image. Therefore, the fused principle pixel is highly sensitive to changes in the MRI pixel while it is stable to changes in the PET pixel. By analyzing the jacobian images for the combination of dark MRI and bright PET intensities, it is observable that the gradient values at the principle pixel are low; therefore, showcasing stability to the pixel intensity changes from both the modalities while widespread neighborhood pixels are influencing the fused principle pixel.

6.3.5. DeepPedestrian

The combination of bright MRI and dark PET intensities in the first test example shows different jacobian results compared to the second test example. It can be noticed that the gradient value of the principle pixel in the jacobian MRI image is low while the gradient for the principle pixel is high in the jacobian PET image. This suggests that the fusion network is very stable to changes in the MRI pixel while sensitive to changes in the PET pixel within the necrotic tumor core. There are also positive gradients for the neighborhood pixels within the clinically important necrotic sub-region of glioma in the jacobian PET images due to which these pixels also influence the fusion principle pixel prediction. In the jacobian PET image, there is some influence from the neighborhood pixels outside the necrotic tumor core, but they have low gradients compared to the gradients of the neighborhood pixels within the necrotic core. The jacobian images for the bright MRI and dark PET intensities in the second test example have unique characteristics since both the jacobian MRI and jacobian PET images have very low gradients for the principle pixel as well as for the neighborhood pixels. Therefore, the network is very stable to changes in pixel intensities in this specific region of interest. For the combination of dark MRI and bright PET intensities, both the test examples show similar results where the neighborhood pixel influence on the fused principle pixel are relatively low in both the input modalities while the gradients for the principle pixel are also low meaning that the network does not get influenced by changes in pixels of each of the input modalities.

6.3.6. Summary

The analysis of jacobian images helped to visually compare the influence of the input neighborhood pixels from each of the input modalities on the fused principle pixel. It was observed that most of the fusion-based neural networks have a significant influence on neighborhood pixels. This visual analysis was not possible by only interpreting the guidance and fused images due to which jacobian images provide additional interpretation towards the suitability of a particular fusion method in a clinical setup. It was observed that Weighted Averaging has no neighborhood influence which is understandable by the fact that it is a per-pixel computation scheme. The MaskNet network, however, provides clinically favorable results as it has a high influence of neighborhood pixels within the necrotic core and very low influence from the pixels outside the necrotic core boundary.

6.4. Scatterplots

The scatterplots shown in Figure 8 helps in understanding the relationship between the gradients of guidance MRI and guidance PET images where each point resembles a pixel with the gradient value in guidance MRI at the horizontal axis and the gradient value in guidance PET at the vertical axis. A positive correlation between these gradients would mean that both the input modalities influence the fused pixel with equal strength. A negative correlation will show that an increase in the influence of one modality will lead to a decrease in the influence of the other modality. For the cases, where the gradients from one of the modalities are constant and only the gradient from the other modality varies, then this conveys that the feature in the input modality where the gradient varies is localized only in a subset of the pixels. If there is no correlation, then we cannot draw any conclusion from the scatterplot.
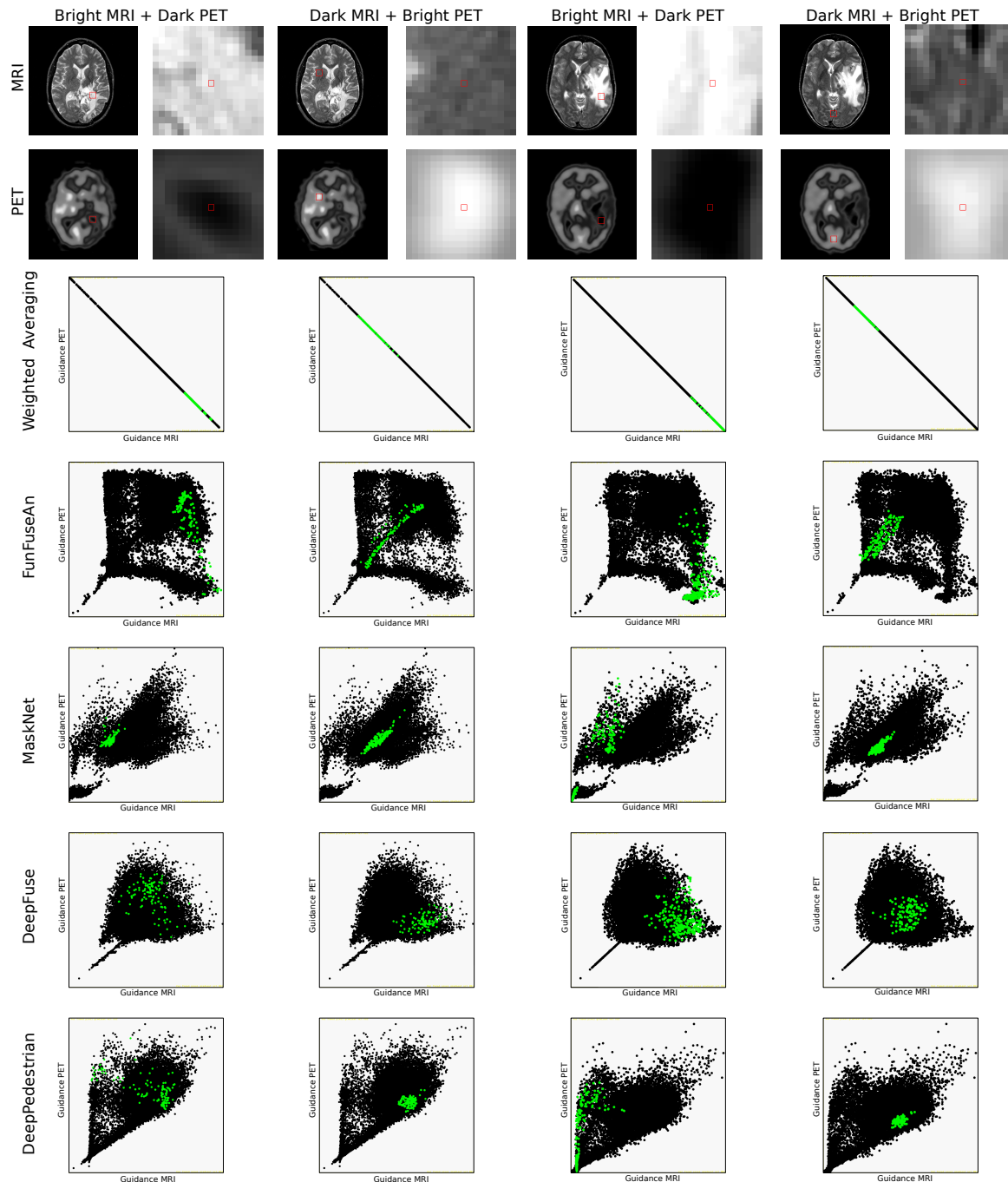
Figure 8. The figure shows the scatterplots between the gradients of the guidance MRI and guidance PET images for each of the fusion methods. The green scatter points are the gradients for the pixels in the zoomed region of interest.

By examining the scatterplots for each of the fusion methods, no clear relationship between the guidance MRI and guidance PET images can be observed, except for the positive correlation between the low gradient values of guidance MRI and guidance PET images in the FunFuseAn and DeepFuse networks as well as a negative correlation between the gradients in the Weighted Averaging method. However, by looking at the gradients within the zoomed region of interests shown by green points, a correlation patterns for a few of the fusion networks can be determined. The scatterplot for Weighted Averaging shows a negative correlation between the guidance MRI and guidance PET images. Therefore, for bright MRI and dark PET intensity setups in each of the two test examples, higher gradient values in guidance MRI causes the green scatter points to tilt towards the higher end

of the horizontal axis while for dark MRI and bright PET intensity setup in the two test examples, the scatter points tilt in the opposite direction. This behavior shows that an increase in the influence of one modality led to a decrease in the influence of the other and Weighted Averaging only prefers brightness as a feature which is clinically not interesting. The scatterplot of FunFuseAn shows a positive correlation between the lower gradient values of guidance MRI and guidance PET and there is no correlation between the other gradient values of guidance MRI and guidance PET images. However, by visualizing the green points for dark MRI and bright PET intensities, we can observe a positive correlation between the gradient values which means both input modalities equally influence the fused pixel in this region. For the bright MRI and dark PET intensity setup, the scatter points are distributed in such a way that gradients from guidance MRI are high and constant while the gradients from the guidance PET varies.

The scatterplots for the MaskNet network in bright MRI and dark PET intensity setup of the first test example show a positive correlation between the gradient values. In the second test example though, a positive correlation is seen only between the very low gradient values of guidance MRI and guidance PET images, while for higher gradients it is difficult to interpret a correlation. For the dark MRI and bright PET intensity setup, the gradients show a positive correlation in both the test examples. The scatterplots for DeepFuse does not show a clear correlation pattern between the gradients of guidance MRI and guidance PET images even after interpreting the green scatter points in the specific region of interests. For the second test example in the bright MRI and dark PET features of DeepPedestrian, the gradients of guidance PET image vary comprehensibly while the gradients of guidance MRI image remains low and constant. In the dark MRI and bright PET intensity setup for both the test examples, DeepPedestrian has gradients that are closely clustered, and it is challenging to interpret a correlation.

Summary

The scatterplots helped in interpreting the correlation between the gradients of the guidance MRI and guidance PET images for each of the fusion methods, which was difficult to estimate by visualizing only the jacobian and the guidance images. These correlations are important to visualize since an ideal fusion method should not have equal influence from both the input modalities in the regions where there are more clinically important features in one modality than the other. For example, in the region where there are dark PET features resembling the necrotic core, the gradients from guidance MRI should always be low and the scatterplot should not have a positive correlation between the gradients. The scatterplots of MaskNet and DeepPedestrian showed such characteristics but only for one of the test examples.

6.5. Memory and Frame Rates

Table 2 shows the timing results of the jacobian computations as well as the frame rates of our FuseVis tool for the visualization of jacobian images. As it can be seen from the results, the computation of jacobians is quite fast due to the powerful GPU hardware used during the experiments. The timings are averaged over 100 randomly selected principle pixels. The results unveil that FunFuseAn and DeepFuse are the fastest among the fusion-based networks. One reason is that these networks have far fewer parameters to be optimized. Networks such as MaskNet and DeepPedestrian have a high number of trained parameters and dense hidden layers due to which the automatic differentiation through backpropagation requires higher computational time. We also estimated frame rates within a static framework where we iteratively saved the jacobian images for each of the principle pixels without using our FuseVis tool. Then, we displayed these images in the FuseVis tool during the mouseover operation. For this, we measured a frame rate of almost 40 frames per second (fps) which is expected considering there were no jacobian computations with backpropagation heuristics involved for each change of principle pixel during the mouseover operation. However, the disadvantage of such

a setup is that it takes around 3 GB of memory to save jacobian images locally for each input modality at 100 dots per inch (DPI) resolution whereas our tool requires zero additional memory usage.

Table 2. The table shows the timing and frame rate results for each of the fusion-based neural networks.

Setting	FunFuseAn	MaskNet	DeepFuse	DeepPedestrian
Jacobian computations	0.003 s	0.004 s	0.003 s	0.005 s
FuseVis-Jacobian images	20 fps	15 fps	20 fps	10 fps
Guidance images	198 s	265 s	195 s	360 s

7. Conclusions

In this work, we presented a novel interactive approach to visually inspect fusion networks. To achieve this, we developed an easy-to-use FuseVis tool that enables the end-user to visualize jacobian images of the selected principle pixel in real time while performing a mouseover interaction. The real-time generation of jacobian images using an interactive user interface showcases the influence of pixels in the input images on a fused pixel. It is an intuitive idea that helps to overcome the huge computational complexity of generating per-pixel jacobian images that involves several backpropagation heuristics. The FuseVis tool also enables the user to visualize guidance images which estimate the sensitivity of image-fusion networks to the changes in input features. Therefore, the guidance images are a very useful technique to get a concise overview of the huge information content in the jacobian images by analyzing the influence of the input pixel to the fused pixel at the same location for all pixels. We currently train different fusion architectures in a similar environmental setting to foster a suitable comparison of black-box fusion algorithms for a critical application in medical diagnosis. However, our first-of-its-kind visualization tool can easily be used to interpret any neural network in its original experimental setup. By visually analyzing the neural networks for image fusion within a specific clinical application, it was found that only MaskNet and DeepPedestrian provided results that would be helpful to clinicians. However, these methods do not preserve all clinically relevant features efficiently due to which these methods cannot be used in a clinical setup yet. Additionally, it is not trivial to attain fusion groundtruth of an ideal fused image based on the specific clinical requirements. Therefore, there is a need to train fusion-based neural networks by using annotations provided by medical artists. We expect that our work should enable further research in the field of visual explanation and interpretation of fusion-based neural networks as well as neural networks for other image processing-based applications.
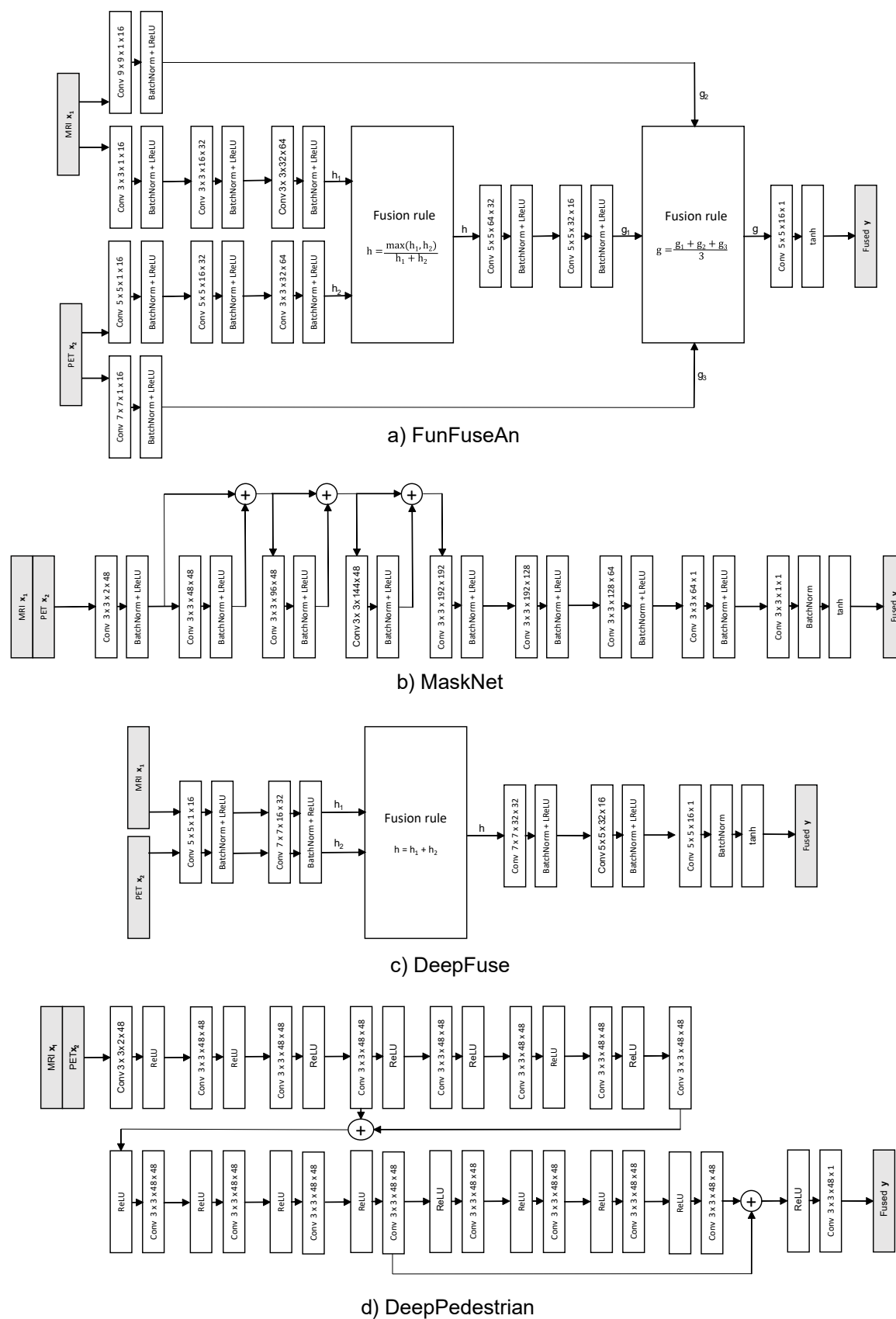
Appendix A

Figure A1. The figure shows different neural network architectures used in this work.

References

1. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
2. James, A.; Dasarathy, B. Medical Image Fusion: A survey of the state of the art. *Inf. Fusion* **2014**, *19*, 4–19. [CrossRef]
3. Li, S.; Kang, X.; Fang, L.; Hu, J.; Yin, H. Pixel-level image fusion: A survey of the state of the art. *Inf. Fusion* **2017**, *33*, 100–112. [CrossRef]
4. Du, J.; Li, W.; Lu, K.; Xiao, B. An overview of multi-modal medical image fusion. *Neurocomputing* **2016**, *215*, 3–20. [CrossRef]
5. Mertens, T.; Kautz, J.; Reeth, F. Exposure Fusion. In Proceedings of the 15th Pacific Conference on Computer Graphics and Applications (PG'07), Maui, HI, USA, 29 October–2 November 2007; pp. 382–390.
6. Yang, S.; Wang, M.; Jiao, L.; Wu, R.; Wang, Z. Image fusion based on a new contourlet packet. *Inf. Fusion* **2010**, *11*, 78–84. [CrossRef]
7. Lewis, J.; O'Callaghan, R.; Nikolov, S.; Bull, D.; Canagarajah, N. Pixel- and region-based image fusion with complex wavelets. *Inf. Fusion* **2007**, *8*, 119–130. [CrossRef]
8. Du, J.; Li, W.; Xiao, B.; Nawaz, Q. Union Laplacian pyramid with multiple features for medical image fusion. *Neurocomputing* **2016**, *194*, 326–339. [CrossRef]
9. Guihong, Q.; Dali, Z.; Pingfan, Y. Medical image fusion by wavelet transform modulus maxima. *Opt. Express* **2001**, *9*, 184–190. [CrossRef]
10. Li, H.; Manjunath, B.; Mitra, S. Multisensor Image Fusion Using the Wavelet Transform. In Proceedings of the 1st International Conference on Image Processing, Austin, TX, USA, 13–16 November 1994; Volume 57, pp. 235–245.
11. Liu, Z.; Yin, H.; Chai, Y.; Yang, S.X. A novel approach for multimodal medical image fusion. *Expert Syst. Appl.* **2014**, *41*, 7424–7435. [CrossRef]
12. Yang, L.; Guo, B.; Ni, W. Multimodality medical image fusion based on multiscale geometric analysis of contourlet transform. *Neurocomputing* **2008**, *72*, 203–211. [CrossRef]
13. Wang, L.; Li, B.; Tian, L. Multimodal Medical Volumetric Data Fusion Using 3-D Discrete Shearlet Transform and Global-to-Local Rule. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 197–206. [CrossRef] [PubMed]
14. Miao, Q.; Shi, C.; Xu, P.; Yang, M.; Shi, Y. A novel algorithm of image fusion using shearlets. *Opt. Commun.* **2011**, *284*, 1540–1547. [CrossRef]
15. Qu, X.; Yan, J.; Xiao, H.; Zhu, Z. Image Fusion Algorithm Based on Spatial Frequency-Motivated Pulse Coupled Neural Networks in Nonsubsampled Contourlet Transform Domain. *Acta Autom. Sin.* **2008**, *34*, 1508–1514. [CrossRef]
16. Bhatnagar, G.; Wu, Q.; Liu, Z. Directive Contrast Based Multimodal Medical Image Fusion in NSCT Domain. *IEEE Trans. Multimed.* **2013**, *15*, 1014–1024. [CrossRef]
17. Singh, S.; Gupta, D.; Anand, R.; Kumar, V. Nonsubsampled shearlet based CT and MR medical image fusion using biologically inspired spiking neural network. *Biomed. Signal Process. Control.* **2015**, *18*, 91–101. [CrossRef]
18. Yin, M.; Liu, W.; Zhao, X.; Yin, Y.; Guo, Y. A novel image fusion algorithm based on nonsubsampled shearlet transform. *Optik* **2014**, *125*, 2274–2282. [CrossRef]
19. Yin, M.; Liu, X.; Liu, Y.; Chen, X. Medical Image Fusion with Parameter-Adaptive Pulse Coupled Neural Network in Nonsubsampled Shearlet Transform Domain. *IEEE Trans. Instrum. Meas.* **2018**, 1–16. [CrossRef]
20. Hu, J.; Li, S. The multiscale directional bilateral filter and its application to multisensor image fusion. *Inf. Fusion* **2012**, *13*, 196–206. [CrossRef]
21. Li, S.; Kang, X.; Hu, J. Image Fusion with Guided Filtering. *IEEE Trans. Image Process.* **2013**, *22*, 2864–2875.
22. Yang, B.; Li, S. Pixel level image fusion with simultaneous orthogonal matching pursuit. *Inf. Fusion* **2012**, *13*, 10–19. [CrossRef]
23. Li, S.; Yin, H. Multimodal image fusion with joint sparsity model. *Opt. Eng.* **2011**, *50*, 1–11.
24. Liu, Y.; Chen, X.; Ward, R.; Wang, Z. Image Fusion with Convolutional Sparse Representation. *IEEE Signal Process. Lett.* **2016**, *23*, 1882–1886. [CrossRef]
25. Fei, Y.; Wei, G.; Zongxi, S. Medical Image Fusion Based on Feature Extraction and Sparse Representation. *Int. J. Biomed. Imaging* **2017**, *2017*, 1–11. [CrossRef] [PubMed]

26. Zhang, Q.; Liu, Y.; Blum, R.; Han, J.; Tao, D. Sparse Representation Based Multi-sensor Image Fusion for Multi-focus and Multi-modality Images. *Inf. Fusion* **2018**, *40*, 57–75. [CrossRef]

27. Das, S.; Kundu, M.K. A neuro-fuzzy approach for medical image fusion. *IEEE Trans. Biomed. Eng.* **2013**, *60*, 3347–3353. [CrossRef] [PubMed]

28. Li, S.; Yang, B. Multifocus image fusion by combining curvelet and wavelet transform. *Pattern Recognit. Lett.* **2008**, *29*, 1295–1301. [CrossRef]

29. Liu, Y.; Liu, S. A general framework for image fusion based on multi-scale transform and sparse representation. *Inf. Fusion* **2015**, *24*, 147–164. [CrossRef]

30. Wang, J.; Peng, J.; Feng, X.; He, G.; Wu, J.; Yan, K. Image fusion with nonsubsampled contourlet transform and sparse representation. *J. Electron. Imaging* **2013**, *22*, 1–15. [CrossRef]

31. Zhu, Z.; Yin, H.; Chai, Y.; Li, Y.; Qi, G. A novel multi-modality image fusion method based on image decomposition and sparse representation. *Inf. Sci.* **2018**, *432*, 516–529. [CrossRef]

32. He, C.; Liu, Q.; Li, H.; Wang, H. Multimodal medical image fusion based on IHS and PCA. *Procedia Eng.* **2010**, *7*, 280–285. [CrossRef]

33. Piella, G. A general framework for multiresolution image fusion: From pixels to regions. *Inf. Fusion* **2003**, *4*, 259–280. [CrossRef]

34. Wang, Q.; Shen, Y. Performances evaluation of image fusion techniques based on nonlinear correlation measurement. In Proceedings of the 21st IEEE Instrumentation and Measurement Technology Conference (IEEE Cat. No.04CH37510), Como, Italy, 18–20 May 2004; Volume 1, pp. 472–475.

35. Haghighat, M.B.A.; Aghagolzadeh, A.; Seyedarabi, H. A Non-reference Image Fusion Metric Based on Mutual Information of Image Features. *Comput. Electr. Eng* **2011**, *37*, 744–756. [CrossRef]

36. Piella, G.; Heijmans, H. A new quality metric for image fusion. In Proceedings of the International Conference on Image Processing (Cat. No.03CH37429), Barcelona, Spain, 14–17 September 2003; Volume 3, p. iii-173.

37. Han, Y.; Cai, Y.; Cao, Y.; Xu, X. A new image fusion performance metric based on visual information fidelity. *Inf. Fusion* **2013**, *14*, 127–135. [CrossRef]

38. Liu, Y.; Chen, X.; Wang, Z.; Wang, Z.J.; Ward, R.K.; Wang, X. Deep learning for pixel-level image fusion: Recent advances and future prospects. *Inf. Fusion* **2018**, *42*, 158–173. [CrossRef]

39. Liu, Y.; Chen, X.; Cheng, J.; Peng, H. A medical image fusion method based on convolutional neural networks. In Proceedings of the 20th International Conference on Information Fusion, Xi'an, China, 10–13 July 2017; pp. 1–7.

40. Li, H.; Wu, X.; Kittler, J. Infrared and Visible Image Fusion using a Deep Learning Framework. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 2705–2710.

41. Zhong, J.; Yang, B.; Li, Y.; Zhong, F.; Chen, Z. Image Fusion and Super-Resolution with Convolutional Neural Network. *Pattern Recognit.* **2016**, *663*, 78–88.

42. Du, C.; Gao, S. Image Segmentation-based Multi-focus Image Fusion through Multi-scale Convolutional Neural Network. *IEEE Access* **2017**, *5*, 15750–15761. [CrossRef]

43. Yan, X.; Gilani, S.Z.; Qin, H.; Mian, A. Unsupervised deep multi-focus image fusion. *arXiv* **2018**, arXiv:1806.07272.

44. Yang, B.; Zhong, J.; Li, Y.; Chen, Z. Multi-focus Image Fusion and Super-resolution with Convolutional Neural Network. *Int. J. Wavelets Multiresolut. Inf. Process.* **2017**, *15*, 1750037. [CrossRef]

45. Liu, Y.; Chen, X.; Peng, H.; Wang, Z. Multi-focus image fusion with a deep convolutional neural network. *Inf. Fusion* **2017**, *36*, 191–207. [CrossRef]

46. Ma, B.; Ban, X.; Huang, H.; Zhu, Y. SESF-Fuse:An unsupervised deep model for multi-focus image fusion. *arXiv* **2020**, arXiv:1908.01703.

47. Guo, X.; Nie, R.; Cao, J.; Zhou, D.; Mei, L.; He, K. FuseGAN: Learning to Fuse Multi-Focus Image via Conditional Generative Adversarial Network. *IEEE Trans. Multimed.* **2019**, *21*, 1982–1996. [CrossRef]

48. Huang, J.; Le, Z.; Ma, Y.; Mei, X.; Fan, F. ACGAN: A generative adversarial network with adaptive constraints for multi-focus image fusion. *Neural Comput. Appl.* **2020**, *32*, 15119–15129. [CrossRef]

49. Kumar, N.; Hoffmann, N.; Oelschlägel, M.; Koch, E.; Kirsch, M.; Gumhold, S. Structural Similarity Based Anatomical and Functional Brain Imaging Fusion. In *Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy. MBIA 2019, MFCA 2019*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2019; Volume 11846.

50. Prabhakar, K.R.; Srikar, V.S.; Babu, R.V. DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4724–4732.

51. Shopovska, I.; Jovanov, L.; Philips, W. Deep Visible and Thermal Image Fusion for Enhanced Pedestrian Visibility. *Sensors* **2019**, *19*, 3727. [CrossRef] [PubMed]

52. Li, H.; Wu, X. DenseFuse: A Fusion Approach to Infrared and Visible Images. *IEEE Trans. Image Process.* **2019**, *28*, 2614–2623. [CrossRef]

53. Hou, R.; Zhou, D.; Nie, R.; Liu, D.; Xiong, L.; Guo, Y.; Yu, C. VIF-Net: An Unsupervised Framework for Infrared and Visible Image Fusion. *IEEE Trans. Comput. Imaging* **2020**, *6*, 640–651. [CrossRef]

54. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2019**, *48*, 11–26. [CrossRef]

55. Ma, J.; Liang, P.; Yu, W.; Chen, C.; Guo, X.; Wu, J.; Jiang, J. Infrared and visible image fusion via detail preserving adversarial learning. *Inf. Fusion* **2020**, *54*, 85–98. [CrossRef]

56. Xu, D.; Wang, Y.; Xu, S.; Zhu, K.; Zhang, N.; Zhang, X. Infrared and Visible Image Fusion with a Generative Adversarial Network and a Residual Network. *Appl. Sci.* **2020**, *10*, 554. [CrossRef]

57. Ma, J.; Xu, H.; Jiang, J.; Mei, X.; Zhang, X. DDcGAN: A Dual-Discriminator Conditional Generative Adversarial Network for Multi-Resolution Image Fusion. *IEEE Trans. Image Process.* **2020**, *29*, 4980–4995. [CrossRef]

58. Xu, H.; Ma, J.; Le, Z.; Jiang, J.; Guo, X. FusionDN: A Unified Densely Connected Network for Image Fusion. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.

59. Joo, D.; Kim, D.; Kim, J. Generating a fusion image: One's identity and another's shape. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1635–1643.

60. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034.

61. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for simplicity: The all convolutional net. ICLR (workshop track). *arXiv* **2015**, arXiv:1412.6806.

62. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.

63. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Muller, K.R.; Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagatiom. *PLoS ONE* **2015**, 10, e0130140. [CrossRef] [PubMed]

64. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning important features through propagating activation differences. *arXiv* **2017**, arXiv:1704.02685.

65. Zeiler, M.D.; Fergus, R. Visualising and understanding convolutional networks. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 818–833.

66. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why should I trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD'16, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.

67. Fong, R.C.; Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbations. In Proceedings of the The IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.

68. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning ICML'17, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 3319–3328.

69. Kumar, N.; Hoffmann, N.; Kirsch, M.; Gumhold, S. Visualisation of Medical Image Fusion and Translation for Accurate Diagnosis of High Grade Gliomas. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; pp. 1–5.

70. Goodfellow, I. Efficient Per-Example Gradient Computations. *arXiv* **2015**, arXiv:1510.01799.

71. Xu, H.; Fan, F.; Zhang, H.; Le, Z.; Huang, J. A Deep Model for Multi-Focus Image Fusion Based on Gradients and Connected Regions. *IEEE Access* **2020**, *8*, 26316–26327. [CrossRef]

72. Clifford, R.J. The Alzheimer's Disease neuroimaging initiative (ADNI). *J. Magn. Reson. Imaging* **2008**, *27*, 685–691.

73. Johnson, K.; Becker, J. The Whole Brain Atlas. Available online: http://www.med.harvard.edu/AANLIB/home.html (accessed on 9 December 2020).

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.