A Robust Mutual-Reinforcing Framework for 3D Multi-Modal Medical Image Fusion Based on Visual-Semantic Consistency

Hao Zhang[1*], Xuhui Zuo[1*], Huabing Zhou[2], Tao Lu[2], Jiayi Ma[1†]

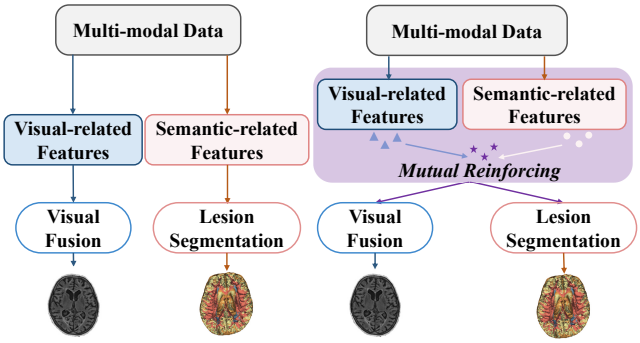
[1]Electronic Information School, Wuhan University, Wuhan 430072, China
[2]School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China
zhpersonalbox@gmail.com, zuoxh2001@163.com, zhouhuabing@gmail.com, lutxyl@gmail.com, jyma2010@gmail.com

Abstract

This work proposes a robust 3D medical image fusion framework to establish a mutual-reinforcing mechanism between visual fusion and lesion segmentation, achieving their double improvement. Specifically, we explore the consistency between vision and semantics by sharing feature fusion modules. Through the coupled optimization of the visual fusion loss and the lesion segmentation loss, visual-related and semantic-related features will be pulled into the same domain, effectively promoting accuracy improvement in a mutual-reinforcing manner. Further, we establish the robustness guarantees by constructing a two-level refinement constraint in the process of feature extraction and reconstruction. Benefiting from full consideration for common degradations in medical images, our framework can not only provide clear visual fusion results for doctor's observation, but also enhance the defense ability of lesion segmentation against these negatives. Extensive evaluations of visual fusion and lesion segmentation scenarios demonstrate the advantages of our method in terms of accuracy and robustness. Moreover, our proposed framework is generic, which can be well-compatible with existing lesion segmentation algorithms and improve their performance. The code is publicly available at https://github.com/HaoZhang1018/RMR-Fusion.

Introduction

Multi-modal medical image fusion aims to combine the different-attribute information of the body, giving visually more informative fused results (Xu and Ma 2021; Tang et al. 2022) or locating the lesions (Zhou et al. 2022; Fang and Wang 2022) more accurately. According to its intended use, the broad concept of multi-modal medical image fusion can be categorized into two specific types. i) **Visual fusion** (Zhang et al. 2020a; Li et al. 2023). Its target audience is the medical doctor, which integrates multi-modal medical images into a single image or cube data, aiming to provide high-quality visual results with sufficient tissue structure information and significant functional pathological distribution. Then, the doctors can observe the generated visual fusion results to make a diagnosis with the support of long-term accumulated experience. ii) **Semantic fusion**.



(a) Separate manner (b) Proposed method

Figure 1: Existing methods treating visual fusion and lesion segmentation as (a) separate issues, and our proposed (b) mutual-reinforcing framework.

It serves intelligent medical diagnostic machines, conducting lesion analysis by combining multi-modal medical images from the semantic perspective. **Lesion segmentation** is a typical semantic fusion technology, which is dedicated to pixel-level localization and classification of lesions (Zhang et al. 2023). Without loss of generality, we continue the following discussion using lesion segmentation as a representative of semantic fusion.

In recent years, a lot of deep multi-modal medical image fusion methods have been proposed to solve the problems of visual fusion (Ma et al. 2020) and lesion segmentation (Li et al. 2020). However, most of these methods treat them as two separate issues, as shown in Fig. 1 (a). For these methods, the consistent way to achieve performance improvements is to design better distance measures (Xue et al. 2021) for optimization guidance and deeper network architectures (Dolz et al. 2019) for nonlinear fitting within their own domain of knowledge. Nevertheless, without introducing new priors for model solving, performance bottlenecks in both visual fusion and lesion segmentation will inevitably arise under such a separate paradigm. In addition, the rigor of the medical field requires that related methods must be quite tolerant of image degradations. Otherwise, a slight perturbation to the source images may lead to serious medical accidents. Unfortunately, few existing multi-modal medical image fusion methods (Xu et al. 2022b; Liu et al. 2022b)

*These authors contributed equally.
†Corresponding author

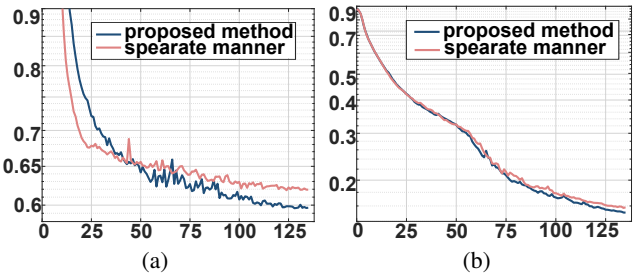


Figure 2: The proposed framework can facilitate better loss optimization than the separate manner. (a) Curves of visual fusion loss. (b) Curves of lesion segmentation loss.

consider the factor of image degradations, which means low robustness and reliability in real scenes. The above two challenges of existing methods can be summarized as *accuracy bottleneck* and *robustness limitation*.

To break the accuracy bottleneck and establish robustness guarantees for visual fusion and lesion segmentation, we propose a robust mutual-reinforcing framework for 3D multi-modal medical image fusion, improving their performance by exploring visual-semantic consistency.

First, we consider the improvement of accuracy, which relies on better optimization for these two tasks. Based on the visual-semantic consistency, we establish a mutual-reinforcing mechanism between visual fusion and lesion segmentation, which are new solution priors for each other, as shown in Fig. 1 (b). *It involves a philosophical question of whether intelligent machines and doctors make diagnoses on the same or positively related basic features.* Considering that some studies (Tang, Yuan, and Ma 2022; Zhu et al. 2021) have demonstrated the unidirectional facilitation of semantics and vision, we further think that visual features and semantic features can be unified into one domain to some extent, which is proved by the subsequent experiments. Therefore, we use a shared feature fusion module to connect the visual head and the semantic head, and couple the optimization of the visual fusion loss and lesion segmentation loss. As a result, the models of visual fusion and lesion segmentation can be optimized better, as shown in Fig. 2.

Second, we establish robustness guarantees for our fusion framework, increasing its tolerance to various degradations. Specifically, we design a powerful two-branch autoencoder using the Swin Transformer for feature extraction and reconstruction, in which three types of negative samples containing blur, noise, and structure loss are considered. Under the two-level refinement constraints in feature and image spaces, our method can effectively defend against various degradations, ensuring the performance of visual fusion and lesion segmentation.

The major contributions of this work are summarized as follows: **i)** We propose a novel 3D fusion framework, which can be used as a general architecture to break the performance bottleneck of visual fusion and lesion segmentation. **ii)** A new idea that considers the consistency between vision and semantics is explored, which derives a mutual-reinforcing mechanism to provide additional solution priors for visual fusion and lesion segmentation by integrating visual-related and semantic-related features into a unified domain. **iii)** We develop an autoencoder with strict two-level refinement constraints, improving the robustness of our framework to common medical image degradations greatly. **iv)** Experiments on visual fusion and lesion segmentation scenarios demonstrate the advantages of our method in terms of accuracy and robustness.

Related Work

Deep Multi-modal Medical Visual Fusion. Deep learning technology has driven great progress in multi-modal medical visual fusion. Early deep methods (Yin et al. 2019; Lahoud and Süsstrunk 2019; Liu et al. 2017) only participate in subparts of visual fusion, *e.g.*, feature fusion. However, these methods cannot fully release the ability of the neural network, and other hand-crafted parts still limit the performance of visual fusion. Realizing this limitation, researchers start to develop end-to-end visual fusion models based on various network architectures. Notably, since there is no ground truth, these methods (Zhang et al. 2020a; Zhang and Ma 2021; Zhou et al. 2020b; Tang et al. 2022) have to preserve efficient information based on perceptual preference. Nevertheless, these methods essentially still follow the conventional paradigm of constructing similarity constraints between source images and the fused image, and do not introduce a new solution prior. Therefore, the performance bottleneck of visual fusion is still difficult to break. Besides, most visual fusion methods can only deal with two-dimensional slice images and cannot be directly applied to real medical volumetric images, limiting their practical application value.

Deep Multi-modal Lesion Segmentation. Depending on whether segmentation labels are used, existing deep methods for multi-modal lesion segmentation can be classified into supervised (Sun et al. 2020; Zhou et al. 2020a) and unsupervised (Wu et al. 2021; Lu, Zheng, and Gupta 2022). Currently, the supervised method is still the mainstream route of the community, which can usually achieve better performance than unsupervised ones due to explicit optimization constraints by labels. Because of the explicit labels, the loss function in supervised methods is generally fixed as the distance between the prediction and the label. Therefore, supervised methods generally increase the segmentation accuracy by designing better network structures. Typically, the improvement of the network structure is carried out from two dimensions, *i.e.*, the communication between multi-modal features, and the interaction between shallow and deep features (Hatamizadeh et al. 2022a; Dolz et al. 2019; Liu et al. 2020). Notably, these methods only rely on the conventional supervised loss, lacking the development of new regularization terms that are helpful for solving. Thus, a natural optimization bottleneck prevents further improvement of segmentation accuracy (Ding et al. 2022; Li et al. 2019). Besides, complicating and deepening the networks will lead to the demand for larger training data, which is not friendly enough to the medical field with scarce data.

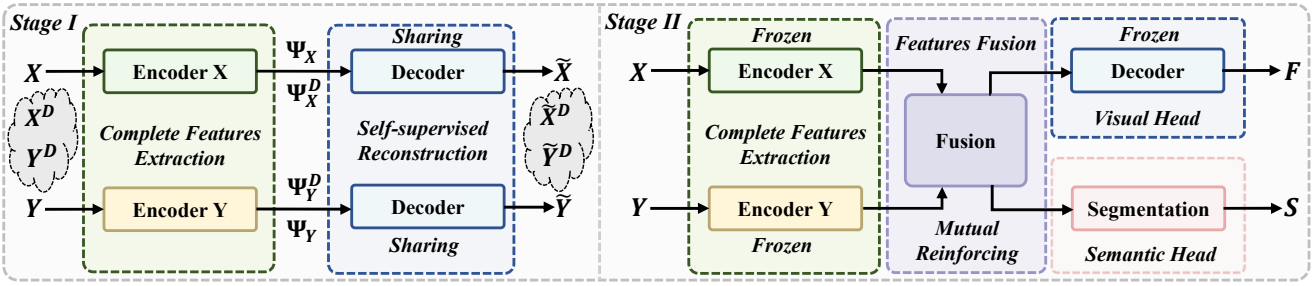


Figure 3: Overall pipeline of our proposed framework. $\{X, Y\}$ and $\{X^D, Y^D\}$ are paired clean and degraded multi-modal images, $\{\Psi_X, \Psi_Y, \Psi_X^D, \Psi_Y^D\}$ are extracted and purified features, $\{\widetilde{X}, \widetilde{Y}, \widetilde{X}^D, \widetilde{Y}^D\}$ are reconstructed multi-modal images by the shared autoencoder, F and S are produced visual fusion result and lesion segmentation result.

Proposed Method

To address the challenges mentioned earlier, we propose to explore the potential consistency between visual perception (Huang et al. 2020) and semantic decision (Hatamizadeh et al. 2022a). Therefore, a robust 3D fusion framework is derived to establish a mutual-reinforcing mechanism between visual fusion and lesion segmentation. The overall procedure is shown in Fig. 3, which consists of two stages. **First**, an autoencoder is trained, which can drive the encoder to fully extract complete features from multi-modal medical images, and constrain the decoder to have the ability of good visual reconstruction. In it, the consideration to refine degraded images greatly makes our framework robust to various negatives. **Second**, we freeze the parameters of the encoder and decoder, and introduce a fusion module to integrate multi-modal features. Then, the visual head and semantic head achieve visual fusion and lesion segmentation based on the shared fused features. Benefiting from the coupled optimization of visual fusion loss and lesion segmentation loss, the fusion module can integrate visual-related features and semantic-related features into a unified domain, establishing visual-semantic consistency to promote their performance in a mutual-reinforcing way.

Degradation-robust Autoencoder

According to the function setting, the role of our proposed degradation-robust autoencoder includes two aspects. First, it should be effectively trained to achieve complete feature extraction and good visual reconstruction in the absence of explicit supervision by the visual ground truth, which is the basis for the later implementation of high-quality visual fusion and lesion segmentation. Second, during the encoding-decoding process, it must be able to effectively filter out various degradation factors contained in the medical images, ensuring the robustness of our proposed framework.

For the first goal, the self-supervised constraints of the autoencoder can naturally guide the effective feature extraction and visual reconstruction process. As for the second goal, we introduce a data augmentation strategy and construct two-level refinement constraints of feature and image spaces, suppressing the transfer expression of degradation factors. Therefore, we give the specific architecture of the degradation-robust autoencoder, as shown in Fig. 4.

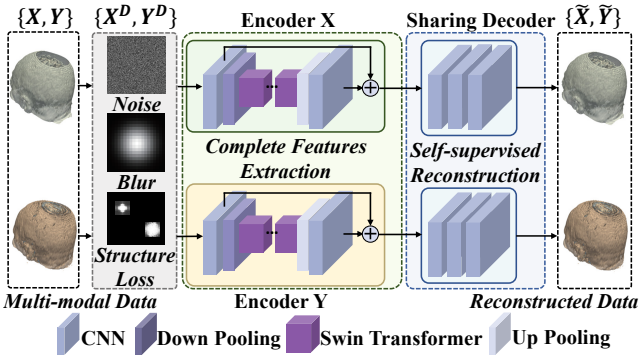


Figure 4: Architecture of degradation-robust autoencoder.

Formally, we first construct negative samples $\{X^D, Y^D\}$ by adding three common degradation factors to the clean multi-modal images $\{X, Y\}$, including noise, blur, and structure loss. Then, two non-shared homogeneous encoders are used to implement feature extraction, obtaining mapped features:

$$\begin{aligned} \{\Psi_X, \Psi_Y\} &= \{EN_X(X), EN_Y(Y)\}, \\ \{\Psi_X^D, \Psi_Y^D\} &= \{EN_X(X^D), EN_Y(Y^D)\}, \end{aligned} \tag{1}$$

where $EN_X(\cdot)$ and $EN_Y(\cdot)$ are the functions of encoders, $\{\Psi_X, \Psi_Y\}$ and $\{\Psi_X^D, \Psi_Y^D\}$ are features from clean $\{X, Y\}$ and dirty $\{X^D, Y^D\}$, respectively. Subsequently, a shared decoder is adopted to implement visual reconstruction, mapping $\{\Psi_X, \Psi_Y, \Psi_X^D, \Psi_Y^D\}$ from the feature space to image space. The reconstruction process is formulated as:

$$\{\widetilde{X}, \widetilde{Y}, \widetilde{X}^D, \widetilde{Y}^D\} = DE(\{\Psi_X, \Psi_Y, \Psi_X^D, \Psi_Y^D\}) \tag{2}$$

where $DE(\cdot)$ indicates the function of the shared decoder, $\{\widetilde{X}, \widetilde{Y}, \widetilde{X}^D, \widetilde{Y}^D\}$ are reconstructed images.

To efficiently train our proposed autoencoder and filter out the contained degradations, we propose a new two-level refinement constraint. First, we design an image-level refinement loss L_{IR} in the image space, which is formulated as:

$$\mathcal{L}_{\text{IR}} = Q(\{\widetilde{X}, \widetilde{Y}\}, \{X, Y\}) + Q(\{\widetilde{X}^D, \widetilde{Y}^D\}, \{X, Y\}), \tag{3}$$

where $Q(\cdot)$ represents the distance function, and we specify it as three types of similarity metrics, including the mean

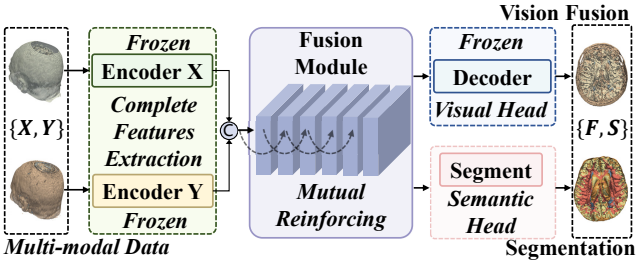


Figure 5: Architecture of mutual-reinforcing fusion module.

square error (MSE), total variation (TV) (Hou et al. 2020) and structural similarity index (SSIM) (Wang et al. 2004) as in our work. Intuitively, L_{IR} requires that the autoencoder can reconstruct clean images regardless of whether the input is clean or degraded. However, L_{IR} is to constrain the filtering of degradations from an overall perspective, which does not take into account the cleanliness in the feature space. Thus, the subsequent lesion segmentation based on features may still suffer from degradations. To address this challenge, we further develop a feature-level refinement loss L_{FR}, which is defined as:

$$\mathcal{L}_{\text{FR}} = Q(\{\Psi_X^D, \Psi_Y^D\}, \{\Psi_X, \Psi_Y\}). \tag{4}$$

L_{FR} requires that the encoders can extract clean features regardless of whether the input is clean or degraded, so as to guarantee the robustness of subsequent lesion segmentation. The final two-level refinement loss $\mathcal{L}_{\text{TR}}$ is obtained by the weighted summation of the image- and feature-level refinement losses, which are balanced by the hyper-parameter α:

$$\mathcal{L}_{\text{TR}} = \mathcal{L}_{\text{IR}} + \alpha \mathcal{L}_{\text{FR}}. \tag{5}$$

Our degradation-robust autoencoder is essentially a denoising autoencoder (Gondara 2016), but it is different from existing ones in two aspects. First, our degradation-robust autoencoder establishes a new two-level refinement loss, which not only requires the cleanliness of the reconstructed image like existing denoising autoencoder, but also additionally requires the purification of encoding features. This is crucial for the semantic segmentation task that relies on features to make decisions. Second, unlike the single-modal denoising autoencoder, ours is a cross-modal autoencoder with two encoders and a shared decoder. In particular, the shared decoder can effectively reconstruct a clean image from the purified features, regardless of which modality the features come from. This property is the basis for generating visually fused images from fused features.

Network Architecture. Our degradation-robust autoencoder consists of two encoders and one decoder, as shown in Fig. 4. First, these two encodes are homogeneous but non-shared. In them, the 3D CNN is first used to extract shallow features with a local receptive field. Then, we use the down-pooling operation to reduce the scale, and introduce Swin Transformer (Liu et al. 2021) on a small scale to capture spatial long-distance dependencies. Finally, we perform the up-pooling to restore the original scale, and utilize the 3D CNN to produce the final encoded features. While in the decoder, we use the pure 3D CNN to process encoded features, to fulfill the expected visual reconstruction.

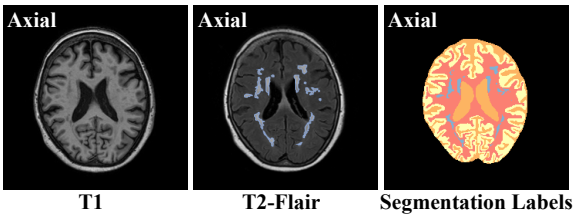


Figure 6: An example of MRI T_1 and $T_{2-Flair}$ (2D axial slices for better visualization).

Mutual-reinforcing Fusion

After learning in the first stage, we already have a powerful encoder and decoder, which can achieve complete feature extraction and high-quality visual reconstruction while filtering out various degradation factors. Now, we can integrate the extracted complete features, and use the coupled visual head and semantic head to jointly modulate this integration process, as shown in Fig. 5. Specifically, we develop a mutual-reinforcing fusion module FN to fuse the multi-modal features $\{\Psi_X, \Psi_Y\} = \{EN_X(X), EN_Y(Y)\}$ that are extracted by the frozen encoders, obtaining the fused features Ψ_F with high expressive ability: $\Psi_F = FN(\Psi_X, \Psi_Y)$. Then, visual fusion and lesion segmentation are implemented based on the fused features simultaneously:

$$F = DE(\Psi_F), \quad S = SH(\Psi_F), \tag{6}$$

where F is the visual fusion result, S is the lesion segmentation result, $DE(\cdot)$ denotes the function of frozen decoder, and $SH(\cdot)$ indicates the function of lesion segmentation network. Because both visual-related and semantic-related attributes must be taken into account in the fused features, a mutual-reinforcing mechanism based on visual-semantic consistency is naturally established.

Next, we describe the definitions and design motivations of the visual fusion loss and the lesion segmentation loss in detail. First, we consider the visual fusion loss. Typically, medical images can be divided into two types according to their characterization information: functional and structural. The former relies on the significance of contrast to reflect abnormal metabolic information such as lesions, while the latter describes the physical form of tissue structures. Taking the medical images used in this paper as an example, we specify X as MRI T_1 that contains rich tissue texture, and Y as $T_{2-Flair}$ in which significant white regions clearly indicate the white matter lesions, as shown in Fig. 6. A good visual fusion result for doctors should maintain these salient (lesion) regions while preserving sufficient tissue textures. Therefore, the visual fusion loss $\mathcal{L}_{\text{VF}}$ is defined as:

$$\mathcal{L}_{\text{VF}} = \|F - Y\|_1 + \delta \|\nabla F - \max(\nabla X, \nabla Y)\|_1, \tag{7}$$

where $\|\cdot\|_1$ designates the ℓ_1 norm, ∇ represents the 3D Sobel operator for seeking gradients, $\max(\cdot)$ is the maximum function, and δ is a hyper-parameter for the trade-off. The first term guarantees the maintenance of salient lesion information in the intensity domain, while the second term achieves the preservation of tissue texture based on the global maximum gradient approximation. Second, we use

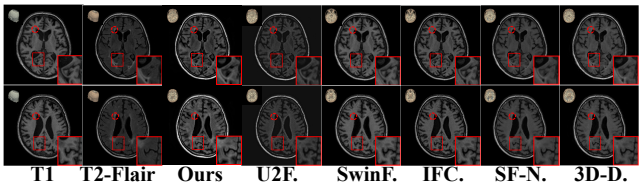


Figure 7: Results of visual fusion on clean images. U2F., SwinF., IFC., SF-N. and 3D-D. denote U2Fusion, SwinFusion, IFCNN, SF-Net and 3D-DTCWT, respectively.

Methods	EN	FMI	AG	SCD
3D-DTCWT	<u>4.3262</u>	0.8928	4.2412	1.1541
SF-Net	4.1902	**0.9000**	3.8287	0.9664
IFCNN	4.1913	0.8942	<u>4.8879</u>	0.9608
SwinFusion	4.2966	0.8907	4.8508	<u>1.2182</u>
U2Fusion	3.8856	0.8922	3.8439	0.6234
Ours	**4.3779**	<u>0.8994</u>	**5.8144**	**1.3529**

Table 1: Quantitative results of visual fusion on clean images. Bold/underline denotes the best/second best. These metrics are calculated on the whole testing set of the MRBrainS MRI dataset, and the same in the following tables.

the popular Dice Loss (Milletari, Navab, and Ahmadi 2016) to define the lesion segmentation loss $\mathcal{L}_{\text{LS}}$:

$$\mathcal{L}_{\text{LS}} = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I} S_{i,j}L_{i,j}}{\sum_{i=1}^{I} S_{i,j} + \sum_{i=1}^{I} L_{i,j}}, \tag{8}$$

where J denotes the total number of classes, and I is the total number of voxels. $S_{i,j}$ and $L_{i,j}$ indicate the probability of the lesion segmentation result and the one-hot encoded labels for j-th class at i-th voxel, respectively. Now, we couple the visual fusion loss and lesion segmentation loss to obtain the final mutual-reinforcing loss $\mathcal{L}_{\text{MF}}$:

$$\mathcal{L}_{\text{MF}} = \mathcal{L}_{\text{VF}} + \eta\mathcal{L}_{\text{LS}}, \tag{9}$$

where η is a hyper-parameter for the trade-off. The visual-semantic consistency implied by the mutual-reinforcing loss can also be perceived in Fig. 6. On the one hand, the preservation of salient lesion regions by the visual head can help the semantic head to better implement small target segmentation. On the other hand, the segmentation of different regions by the semantic head can also help the visual head to achieve fine integration of the texture details.

Network Architecture. In our mutual-reinforcing fusion module, we first perform the concatenation operation on the multi-modal features, and then utilize the pure 3D CNN to process the added features. Notably, the dense connection is used multiple times to promote feature communication, which has been shown to be very effective in feature fusion (Li and Wu 2019; Xu et al. 2020). The visual head is designated as the frozen decoder, which has a good visual reconstruction ability after autoencoder training at the previous stage. The semantic head can be specified as an existing lesion segmentation model, which is selected as the state-of-the-art Swin Unetr (Hatamizadeh et al. 2022a) in this paper.

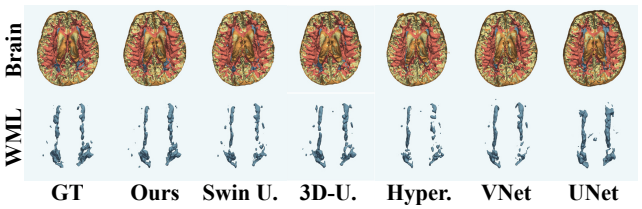


Figure 8: Results of lesion segmentation on clean images. Swin U., 3D-U. and Hyper. denote Swin Unetr, 3D-UX-Net and HyperDenseNet, respectively.

Experiments

Implementation Details. Our method is implemented using PyTorch and the MONAI (Cardoso et al. 2022) framework, running on an NVIDIA TITAN RTX GPU and a 2.20 GHz Intel Xeon Platinum 8273CL CPU. Evaluation is performed on the MRBrainS MRI dataset[1] (Mendrik et al. 2015). Data augmentation is applied to expand the source data, with a training-testing ratio of $4:1$. Autoencoder and mutual-reinforcing fusion module training use a batch size of 2 for 200 and 140 epochs, respectively. The model has 63.09 M parameters and is optimized using the Adam Optimizer (Kingma and Ba 2014) with an initial learning rate of $1e^{-4}$. Hyperparameters α, δ, and η are set to 1, 0.08, and 0.10, respectively.

Accuracy Evaluation

First, we implement the comparison on clean data.

Visual Fusion Five state-of-the-art visual fusion methods are selected for comparison, including 3D-DTCWT (Kushwaha et al. 2015), SF-Net (Liu et al. 2022a), IFCNN (Zhang et al. 2020b), SwinFusion (Ma et al. 2022), and U2Fusion (Xu et al. 2022a). The 2D slices are demonstrated in Fig. 7 for better visualization. Our method can simultaneously preserve the salience of lesion regions and integrate tissue structures, while other methods can not. Further, we use four common non-reference metrics to evaluate visual fusion results, including entropy (EN) (Roberts, Van Aardt, and Ahmed 2008), feature mutual information (FMI) (Haghighat, Aghagolzadeh, and Seyedarabi 2011), average gradient (AG) (Cui et al. 2015), and the sum of the correlations of differences (SCD) (Aslantas and Bendes 2015). As reported in Table 1, our method achieves three best and one second-best rankings, indicating that our results contain the most information, preserve the richest structures, and efficiently incorporate features from source images.

Lesion Segmentation We use five state-of-the-art segmentation methods for comparison, including UNet (Çiçek et al. 2016), VNet (Milletari, Navab, and Ahmadi 2016), 3D-UX-Net (Lee et al. 2023), HyperDenseNet (Dolz et al. 2019), and Swin Unetr (Hatamizadeh et al. 2022a). The visualization results of brain segmentation are presented in Fig. 8. Our method can produce segmentation results that are most consistent with the ground truth, having obvious

[1]https://mrbrains13.isi.uu.nl

Methods	CSF	GM	WM	WML	Average
UNet	0.7462	0.7533	0.7632	0.5122	0.6937
VNet	0.7607	0.7868	0.7661	0.5524	0.7165
HyperDenseNet	0.8129	0.8348	0.8008	0.5804	0.7572
3D-UX-Net	0.7884	0.8168	0.8140	0.6311	0.7626
Swin Unetr	0.7985	0.8309	0.8102	0.6187	0.7646
Ours	0.7946	0.8256	0.8197	0.6994	**0.7848**

Table 2: Quantitative results (Dice similarity coefficient) of lesion segmentation on clean images.

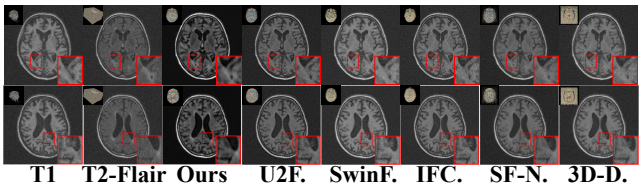


Figure 9: Visual fusion results on degraded images.

Methods	EN	FMI	AG	SCD
3D-DTCWT	6.5812	0.8845	7.1057	0.6296
SF-Net	6.5342	0.8909	7.1491	0.5844
IFCNN	6.6110	0.8834	**8.6804**	0.6000
SwinFusion	**6.6137**	0.8832	7.7014	0.7746
U2Fusion	6.2087	0.8859	6.7426	0.6613
Ours	4.7145	**0.8948**	5.5052	**1.1275**

Table 3: Quantitative visual fusion on degraded images.

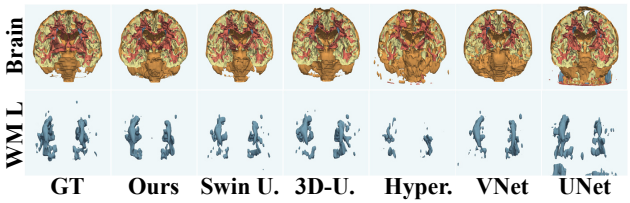


Figure 10: Lesion segmentation results on degraded images.

Methods	CSF	GM	WM	WML	Average	Drop
UNet	0.6776	0.7258	0.6805	0.2326	0.5791	0.1146
VNet	0.7427	0.7533	0.7134	0.4796	0.6722	0.0443
HyperDenseNet	0.7161	0.5764	0.5417	0.3433	0.5444	0.2128
3D-UX-Net	0.7844	0.7924	0.7746	0.5576	0.7273	0.0402
Swin Unetr	0.7834	0.8015	0.7767	0.5281	0.7224	0.0422
Ours	0.7780	0.8035	0.7909	0.6568	**0.7573**	**0.0275**

Table 4: Quantitative segmentation on degraded images.

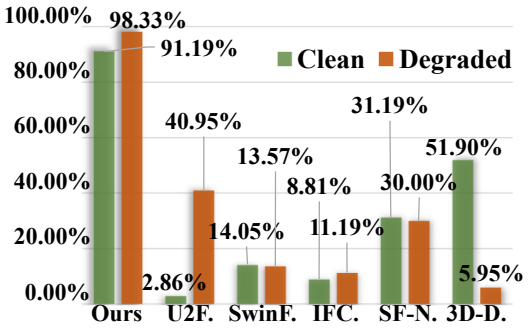


Figure 11: The probability of each method entering the top two based on the psychophysical study.

advantages over other methods, especially in white matter lesion regions. Furthermore, we introduce the Dice similarity coefficient (DSC) (Milletari, Navab, and Ahmadi 2016) to objectively assess the segmentation accuracy on four classes, including cerebrospinal fluid (CSF), gray matter (GM), white matter (WM), and white matter lesions (WML). As reported in Table 2, our method achieves the highest average segmentation accuracy, which is about 0.02 higher than our selected backbone Swin Unetr. This can prove that the proposed mutual-reinforcing framework can indeed promote the precision of semantic decisions. Meanwhile, it is worth noting that our method achieves far superior segmentation accuracy in our focused white matter lesion regions.

Robustness Evaluation

Next, we verify the robustness of these methods to various degradations. Specifically, we simulate the negatives introduced during the imaging process by adding a mixture of randomly sampled noise ($std = 0.04$, $mean = 0$), Gaussian blurring ($std = 0.6$, $mean = 0$), and structure loss (size of $(4, 16, 16)$'s random window with Fourier broken).

Visual Fusion The qualitative results are shown in Fig. 9. These comparative methods cannot remove degradations contained in the source images. Inevitably, their results of visual fusion suffer from information loss, which will not be conducive to the doctor's observation and analysis. In comparison, our method shows strong robustness, which effectively recovers useful information by removing degradation, presenting a clear visual fusion appearance. The quantitative results are further reported in Table 3. Since EN and AG are metrics highly correlated with high-frequency noise, our method does not achieve the best scores on them. For the other two metrics FMI and SCD, our method ranks first. Overall, these results demonstrate the robustness of our method to degradations on the visual fusion task.

Lesion Segmentation Then, we present qualitative results of lesion segmentation on degraded data in Fig. 10. Clearly, the addition of degradations is disastrous for some comparative methods, such as UNet and HyperDenseNet. For Swin Unetr, 3D-UX-Net, and VNet, some significant decision misjudgments also occur in their results. Comparatively, our approach successfully defends the interference of degradations at the semantic level, still providing results that are more consistent with the ground truth. Further quantitative results are reported in Table 4. The average segmentation accuracy of our method on degraded data is only 0.0275 lower than that on clean data, while the performance drop of other methods is generally above 0.04. All these fully demonstrate the robustness of our proposed method.

Psychophysical Study

The visual fusion performance cannot be perfectly evaluated due to the lack of ground truth. Even though some non-reference metrics are used, some degradation factors may in-

Visual Fusion	EN	FMI	AG	SCD	Segmentation	CSF	GM	WM	WML	Average
w/o $\mathcal{L}_{\text{LS}}$	4.0446	**0.9019**	4.1453	0.4616	w/o $\mathcal{L}_{\text{VF}}$	0.7963	0.8240	0.8166	0.6835	0.7801
Ours	**4.3779**	0.8994	**5.8144**	**1.3529**	Ours	0.7946	0.8256	0.8197	0.6994	**0.7848**

Table 5: Quantitative results of ablation studies.

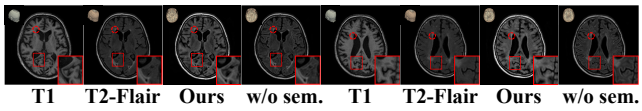


Figure 12: Qualitative impact of semantics on visual fusion. w/o sem. means removing the semantics head.

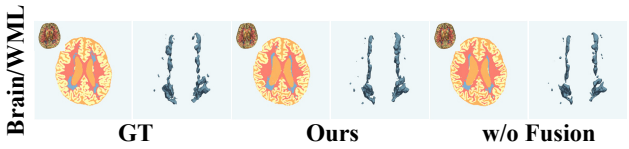


Figure 13: Qualitative impact of visual fusion on semantics.

terfere with their objectivity, such as EN and AG in Table 3. To address this, we invite 35 computer vision researchers in a psychophysical study. They are instructed to evaluate based on specific criteria: a good visual fusion result should exhibit rich structural information and salient lesions while suppressing degradations. Each researcher chooses the two best samples from each group to avoid randomness while ensuring fairness. Statistical results in Fig. 11 on both clean and degraded images reveal that 91.19% considered our results the best or second best on clean data, and this advantage becomes more evident on degraded data, with a voting rate of 98.33%. These statistics affirm the promising performance of our method in visual fusion.

Ablation Studies

Impact of Semantics on Visual Fusion In our framework, we utilize the underlying features crucial for semantic decisions to enhance appearance in visual fusion. By excluding the lesion segmentation loss $\mathcal{L}_{\text{LS}}$, we guide our framework to solely optimize visual fusion. Qualitative results in Fig. 12 reveal that without $\mathcal{L}_{\text{LS}}$, our method's ability to integrate brain structures weakens. This is further confirmed by the quantitative results in the left column of Table 5, indicating that semantics play a positive role in enhancing visual fusion performance.

Impact of Visual Fusion on Semantics Similarly, we remove the visual fusion loss $\mathcal{L}_{\text{VF}}$ to force the fusion module to be guided only by the lesion segmentation loss $\mathcal{L}_{\text{LS}}$. The qualitative results are demonstrated in Fig. 13. Obviously, the visual fusion loss can make segmentation voxels more complete and accurate, especially in the white matter lesion regions. The 2D slice reflects this difference more clearly. The quantitative results are reported in the right column of Table 5. It can be seen that the segmentation performance decreases after removing $\mathcal{L}_{\text{VF}}$. These results demonstrate that visual fusion can facilitate segmentation.

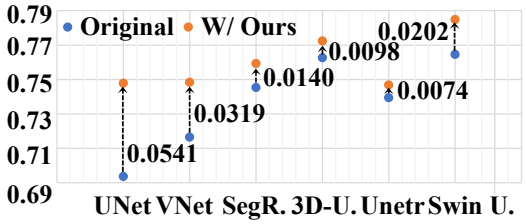


Figure 14: Segmentation gains brought by our framework.

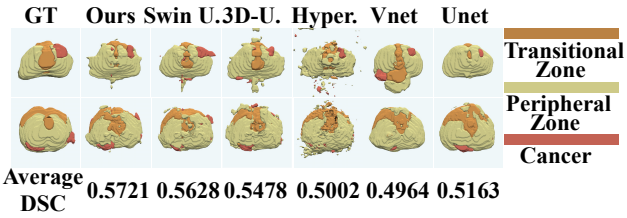


Figure 15: Results of prostate cancer segmentation.

Universality of the Proposed Framework

Our method is more of a general framework than a specific lesion segmentation model. By replacing the Swin Unetr backbone in the semantic head with models like UNet, VNet, SegResNet (Myronenko 2018), 3D-UX-Net, and Unetr (Hatamizadeh et al. 2022b), our framework consistently improves the performance of these segmentation models, as shown in Fig. 14. Notably, our framework requires the semantic head (*i.e.*, segmentation model) to use features integrated by our fusion module for decision-making. Hence, it may not be compatible with segmentation methods like HyperDenseNet, which primarily emphasize network structures for interacting multi-modal features.

Extended Application

We further apply the proposed framework to the Prostate158 dataset (Adams et al. 2022), for achieving prostate cancer segmentation. We demonstrate the qualitative and quantitative results in Fig. 15. It can be seen that our method still achieves a higher average segmentation accuracy than other methods, indicating its applicability and effectiveness.

Conclusion

In this paper, we designed a robust mutual-reinforcing 3D multi-modal medical image fusion framework. First, we proposed a Swin Transformer-based autoencoder with two-stage refinement for robustness against degradations. Second, a feature fusion module was designed to couple visual fusion and lesion segmentation to mutually promote their accuracy. Extensive experiments have revealed its superiority and compatibility to existing lesion segmentation methods.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62276192).

References

Adams, L. C.; Makowski, M. R.; Engel, G.; Rattunde, M.; Busch, F.; Asbach, P.; Niehues, S. M.; Vinayahalingam, S.; van Ginneken, B.; Litjens, G.; et al. 2022. Prostate158-An expert-annotated 3T MRI dataset and algorithm for prostate cancer detection. *Computers in Biology and Medicine*, 148: 105817.

Aslantas, V.; and Bendes, E. 2015. A new image quality metric for image fusion: The sum of the correlations of differences. *Aeu-International Journal of Electronics and Communications*, 69(12): 1890–1896.

Cardoso, M. J.; Li, W.; Brown, R.; Ma, N.; Kerfoot, E.; Wang, Y.; Murrey, B.; Myronenko, A.; Zhao, C.; Yang, D.; et al. 2022. MONAI: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*.

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 424–432.

Cui, G.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2015. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications*, 341: 199–209.

Ding, W.; Abdel-Basset, M.; Hawash, H.; and Pedrycz, W. 2022. Multimodal infant brain segmentation by fuzzy-informed deep learning. *IEEE Transactions on Fuzzy Systems*, 30(4): 1088–1101.

Dolz, J.; Gopinath, K.; Yuan, J.; Lombaert, H.; Desrosiers, C.; and Ayed, I. B. 2019. HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation. *IEEE Transactions on Medical Imaging*, 38(5): 1116–1126.

Fang, L.; and Wang, X. 2022. Brain tumor segmentation based on the dual-path network of multi-modal MRI images. *Pattern Recognition*, 124: 108434.

Gondara, L. 2016. Medical image denoising using convolutional denoising autoencoders. In *Proceedings of the IEEE International Cconference on Data Mining Workshops*, 241–246.

Haghighat, M. B. A.; Aghagolzadeh, A.; and Seyedarabi, H. 2011. A non-reference image fusion metric based on mutual information of image features. *Computers & Electrical Engineering*, 37(5): 744–756.

Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H.; and Xu, D. 2022a. Swin Unetr: Swin Transformers for semantic segmentation of brain tumors in MRI images. *arXiv preprint arXiv:2201.01266*.

Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H. R.; and Xu, D. 2022b. Unetr: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Applications of Computer Vision*, 574–584.

Hou, R.; Zhou, D.; Nie, R.; Liu, D.; Xiong, L.; Guo, Y.; and Yu, C. 2020. VIF-Net: An unsupervised framework for infrared and visible image fusion. *IEEE Transactions on Computational Imaging*, 6: 640–651.

Huang, J.; Le, Z.; Ma, Y.; Fan, F.; Zhang, H.; and Yang, L. 2020. MGMDcGAN: Medical image fusion using multi-generator multi-discriminator conditional generative adversarial network. *IEEE Access*, 8: 55145–55157.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kushwaha, A.; Khare, A.; Prakash, O.; Song, J.-I.; and Jeon, M. 2015. 3D medical image fusion using dual tree complex wavelet transform. In *Proceedings of the International Conference on Control, Automation and Information Sciences*, 251–256.

Lahoud, F.; and Süsstrunk, S. 2019. Zero-learning fast medical image fusion. In *Proceedings of the International Conference on Information Fusion*, 1–8.

Lee, H. H.; Bao, S.; Huo, Y.; and Landman, B. A. 2023. 3D UX-Net: A large kernel volumetric convNet modernizing hierarchical transformer for medical image segmentation. In *Proceedings of the International Conference on Learning Representations*.

Li, H.; and Wu, X.-J. 2019. DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5): 2614–2623.

Li, J.; Liu, J.; Zhou, S.; Zhang, Q.; and Kasabov, N. K. 2023. GeSeNet: A General Semantic-Guided Network With Couple Mask Ensemble for Medical Image Fusion. *IEEE Transactions on Neural Networks and Learning Systems*.

Li, J.; Yu, Z. L.; Gu, Z.; Liu, H.; and Li, Y. 2019. MMAN: Multi-modality aggregation network for brain segmentation from MR images. *Neurocomputing*, 358: 10–19.

Li, K.; Yu, L.; Wang, S.; and Heng, P.-A. 2020. Towards cross-modality medical image segmentation with online mutual knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 775–783.

Liu, L.; Hu, X.; Zhu, L.; Fu, C.-W.; Qin, J.; and Heng, P.-A. 2020. ψ-net: Stacking densely convolutional lstms for subcortical brain structure segmentation. *IEEE Transactions on Medical Imaging*, 39(9): 2806–2817.

Liu, Y.; Chen, X.; Cheng, J.; and Peng, H. 2017. A medical image fusion method based on convolutional neural networks. In *Proceedings of the International Conference on Information Fusion*, 1–7.

Liu, Y.; Mu, F.; Shi, Y.; and Chen, X. 2022a. Sf-net: A multi-task model for brain tumor segmentation in multimodal mri via image fusion. *IEEE Signal Processing Letters*, 29: 1799–1803.

Liu, Y.; Shi, Y.; Mu, F.; Cheng, J.; and Chen, X. 2022b. Glioma segmentation-oriented multi-modal MR image fusion with adversarial learning. *IEEE/CAA Journal of Automatica Sinica*, 9(8): 1528–1531.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of*

the IEEE/CVF International Conference on Computer Vision, 10012–10022.

Lu, C.; Zheng, S.; and Gupta, G. 2022. Unsupervised domain adaptation for cardiac segmentation: Towards structure mutual information maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2588–2597.

Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; and Ma, Y. 2022. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7): 1200–1217.

Ma, J.; Xu, H.; Jiang, J.; Mei, X.; and Zhang, X.-P. 2020. DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29: 4980–4995.

Mendrik, A. M.; Vincken, K. L.; Kuijf, H. J.; Breeuwer, M.; Bouvy, W. H.; De Bresser, J.; Alansary, A.; De Bruijne, M.; Carass, A.; El-Baz, A.; et al. 2015. MRBrainS challenge: Online evaluation framework for brain image segmentation in 3T MRI scans. *Computational Intelligence and Neuroscience*, 2015: 1–1.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the International Conference on 3D Vision*, 565–571.

Myronenko, A. 2018. 3D MRI brain tumor segmentation using autoencoder regularization. *arXiv preprint arXiv:1810.11654*.

Roberts, J. W.; Van Aardt, J. A.; and Ahmed, F. B. 2008. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1): 023522.

Sun, L.; Ma, W.; Ding, X.; Huang, Y.; Liang, D.; and Paisley, J. 2020. A 3D spatially weighted network for segmentation of brain tissue from MRI. *IEEE Transactions on Medical Imaging*, 39(4): 898–909.

Tang, L.; Yuan, J.; and Ma, J. 2022. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82: 28–42.

Tang, W.; He, F.; Liu, Y.; and Duan, Y. 2022. MATR: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Transactions on Image Processing*, 31: 5134–5149.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Wu, X.; Bi, L.; Fulham, M.; Feng, D. D.; Zhou, L.; and Kim, J. 2021. Unsupervised brain tumor segmentation using a symmetric-driven adversarial network. *Neurocomputing*, 455: 242–254.

Xu, H.; and Ma, J. 2021. EMFusion: An unsupervised enhanced medical image fusion network. *Information Fusion*, 76: 177–186.

Xu, H.; Ma, J.; Jiang, J.; Guo, X.; and Ling, H. 2022a. U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 502–518.

Xu, H.; Ma, J.; Le, Z.; Jiang, J.; and Guo, X. 2020. Fusiondn: A unified densely connected network for image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12484–12491.

Xu, H.; Ma, J.; Yuan, J.; Le, Z.; and Liu, W. 2022b. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19679–19688.

Xue, Z.; Li, P.; Zhang, L.; Lu, X.; Zhu, G.; Shen, P.; Shah, S. A. A.; and Bennamoun, M. 2021. Multi-modal co-learning for liver lesion segmentation on PET-CT images. *IEEE Transactions on Medical Imaging*, 40(12): 3531–3542.

Yin, M.; Liu, X.; Liu, Y.; and Chen, X. 2019. Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain. *IEEE Transactions on Instrumentation and Measurement*, 68(1): 49–64.

Zhang, H.; and Ma, J. 2021. SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129: 2761–2785.

Zhang, H.; Xu, H.; Xiao, Y.; Guo, X.; and Ma, J. 2020a. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12797–12804.

Zhang, Y.; Liu, Y.; Sun, P.; Yan, H.; Zhao, X.; and Zhang, L. 2020b. IFCNN: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54: 99–118.

Zhang, Y.; Peng, C.; Tong, R.; Lin, L.; Chen, Y.-W.; Chen, Q.; Hu, H.; and Zhou, S. K. 2023. Multi-Modal tumor segmentation with deformable aggregation and uncertain region inpainting. *IEEE Transactions on Medical Imaging*.

Zhou, C.; Ding, C.; Wang, X.; Lu, Z.; and Tao, D. 2020a. One-pass multi-task networks with cross-task guided attention for brain tumor segmentation. *IEEE Transactions on Image Processing*, 29: 4516–4529.

Zhou, T.; Fu, H.; Chen, G.; Shen, J.; and Shao, L. 2020b. Hi-net: Hybrid-fusion network for multi-modal MR image synthesis. *IEEE Transactions on Medical Imaging*, 39(9): 2772–2781.

Zhou, T.; Ruan, S.; Vera, P.; and Canu, S. 2022. A Tri-Attention fusion guided multi-modal segmentation network. *Pattern Recognition*, 124: 108417.

Zhu, L.; Ji, D.; Zhu, S.; Gan, W.; Wu, W.; and Yan, J. 2021. Learning statistical texture for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12537–12546.