# Deep evidential fusion with uncertainty quantification and reliability learning for multimodal medical image segmentation

Ling Huang, Su Ruan, Pierre Decazes, Thierry Denœux

# Deep evidential fusion with uncertainty quantification and reliability learning for multimodal medical image segmentation

Ling Huang[a], Su Ruan[b], Pierre Decazes[c], Thierry Denœux[a,d]

[a] *Université de technologie de Compiègne, CNRS, Heudiasyc, Compiègne, France*
[b] *Université de Rouen Normandie, Quantif, LITIS, Rouen, France*
[c] *Université de Rouen Normandie, Centre Henri Becquerel, Rouen, France*
[d] *Institut universitaire de France, Paris, France*

**Abstract**

Single-modality medical images generally do not contain enough information to reach an accurate and reliable diagnosis. For this reason, physicians commonly rely on multimodal medical images for comprehensive diagnostic assessments. This study introduces a deep evidential fusion framework designed for segmenting multimodal medical images, leveraging the Dempster-Shafer theory of evidence in conjunction with deep neural networks. In this framework, features are first extracted from each imaging modality using a deep neural network, and features are mapped to Dempster-Shafer mass functions that describe the evidence of each modality at each voxel. The mass functions are then corrected by the contextual discounting operation, using learned coefficients quantifying the reliability of each source of information relative to each class. The discounted evidence from each modality is then combined using Dempster's rule of combination. Experiments were carried out on a PET-CT dataset for lymphoma segmentation and a multi-MRI dataset for brain tumor segmentation. The results demonstrate the ability of the proposed fusion scheme to quantify segmentation uncertainty and improve segmentation accuracy. Moreover, the learned reliability coefficients provide some insight into the contribution of each modality to the segmentation process.

*Keywords:* Dempster-Shafer theory, Evidence theory, Medical image processing, Deep learning, Decision-level fusion

## 1. Introduction

Recent advances in medical imaging technologies have facilitated the acquisition of multimodal data such as Positron Emission Tomography (PET)/Computed Tomography (CT) and multi-sequence Magnetic Resonance Imaging (MRI). Images from a single modality provide partial insight into cancer and other abnormalities within the human body. Multimodal medical image analysis, which integrates information from diverse medical imaging modalities, significantly contributes to a comprehensive understanding of intricate medical conditions [90]. It encompasses factors such as the location, size, and extent of pathological

structures. Medical image segmentation based on the fusion of multimodal medical information allows clinicians to better delineate anatomical structures, lesions and abnormalities, thus enhancing the effectiveness of disease detection, diagnosis, and treatment planning.

Multimodal medical image fusion strategies can be implemented at different levels [84]. At the lowest pixel level, multimodality images are concatenated as a single input. Alternatively, features can be extracted from different modalities and combined for further modeling and reasoning (feature-level fusion). Finally, in the decision-level approach, partial decisions are made independently based on each modality and aggregated to obtain a final decision. Though recent developments in multimodal medical image analysis have yielded promising experimental results, conventional multimodal medical image fusion strategies still suffer from some limitations. It is often difficult to explain why a given strategy works in a given context, and to quantify decision uncertainty in a reliable way. Moreover, most approaches are based on optimistic assumptions about data quality and, contrary to clinical knowledge, they treat images from different modalities as equally reliable when segmenting tumors, which may lead to biased or wrong decisions.

The success of information fusion depends on the relevance and complementarity of input information, the existence of prior knowledge about the information sources, and the expressive power of the uncertainty model employed [65, 24, 61]. Given that the quality of input information and prior knowledge is intricately tied to the data collection stage, a lot of work has been devoted to modeling uncertainties in a faithful way [39]. As a critical factor in the information fusion process [1, 38], accurate uncertainty quantification must be regarded as a primary objective to achieve precise multimodal medical image segmentation.

Early methods for quantifying uncertainty essentially relied on probabilistic models, often integrated with Bayesian inference or sampling techniques to estimate uncertainty across various parameters or variables [34, 56]. The advent of deep neural networks has sparked renewed interest in uncertainty estimation [1], leading to the development of methods such as Monte-Carlo dropout [29] and deep ensembles [45]. However, it is important to note that these probabilistic models rely on assumptions about the underlying data distribution, and improper distributions can result in inaccurate uncertainty estimations. Furthermore, uncertainty quantification via inference or sampling algorithms heavily relies on computational approximations and may lack rigorous theoretical justification [6, 7]. These and other limitations motivate the search for alternative approaches for uncertainty quantification for information fusion and decision-making applications.

Instead of making strong assumptions on actual data distribution, non-probabilistic methods use alternative mathematical frameworks or representations such as possibility theory [87, 21] and Dempster-Shafer theory (DST) [15, 69, 20] to quantify uncertainty. In particular, the latter formalism is an evidence-based information modeling, reasoning, and fusion framework that can be used with both supervised [16, 80, 79] and unsupervised learning [48, 19], providing an effective way to handle imperfect (i.e., imprecise, uncertain, and conflicting) data. Compared to possibility theory, DST allows the quantification of both aleatory and epistemic uncertainty while providing a powerful mechanism for combining multiple unreliable pieces of information [61].

In multimodal medical image segmentation, effectively combining uncertain information

from diverse sources presents a significant challenge. Some learning-based approaches propose addressing conflicting decisions by introducing learnable weights [50, 4, 71]. The term "weight" in those approaches usually refers to the importance of information. In contrast, reliability pertains to the trustworthiness of the information and needs to be carefully analyzed in different medical situations. Four major approaches have been used to provide reliability coefficients: 1) modeling the reliability of sources using a degree of consensus [14]; 2) modeling expert opinions using probability distributions [12]; 3) using external domain knowledge or contextual information to model reliability coefficients [26]; 4) learning the reliability coefficients from training data [25, 62], which is a very general approach that does not require any prior domain knowledge or expert opinions. In this work, we consider an even more flexible approach in which the reliability of each image modality is described by several coefficients, one for each ground truth value. The reliability coefficient for source $i$ and class $k$ is then defined as one's belief that the information from source $i$ is reliable, if the true class is $k$.

In this paper, we introduce a new approach to multimodal medical image segmentation combining DST with deep neural networks[1]. The proposed fusion scheme comprises multiple encoder-decoder-based feature extraction modules, DST-based evidence-mapping modules, and a multimodality evidence fusion module. The evidence-mapping modules transform the extracted features into mass functions representing the evidence from each imaging modality about the class of each voxel. These mass functions are then corrected by a contextual discounting operation, and the discounted pieces of evidence are combined by Dempster's rule of combination. The whole framework is trained end-to-end by minimizing a loss function quantifying the errors before and after the fusion of information from each modality. Our main contributions are, thus, the following:

1. We propose a new hybrid fusion architecture for multimodal medical images composed of feature extraction, evidence-mapping, and combination modules.
2. Within this architecture, we integrate mechanisms for (i) quantifying segmentation uncertainty using Dempster-Shafer mass functions, (ii) correcting these mass functions to account for the relative reliability of each imaging modality using context discounting, and (iii) combining corrected mass functions from different sources to reach final segmentation decisions.
3. We introduce an improved two-part loss function making it possible to optimize the segmentation performance of each individual source modality together with the overall performance of the combined decisions.
4. Through extensive experiments with two real medical image datasets, we show that the proposed decision-level fusion scheme improves segmentation reliability and quality as compared to alternative pixel-level methods for exploiting different image modalities.

[1]This paper is an extended version of the short paper presented at the 25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2022) [36]. This extended version includes a much more detailed description and explanation of the fusion framework, an improved optimization strategy with a two-part loss function, as well as extended results with a second dataset for lymphoma segmentation and an additional transformer-based feature-extraction module.

5. We show that the learned reliability coefficients provide some insight into the contribution of each imaging modality in the segmentation process.

The rest of this paper is organized as follows. Background information and related work are first recalled in Section 2. Our approach is then introduced in Section 3, and experimental results are reported in Section 4. Finally, Section 5 concludes the paper and presents some directions for further research.

## 2. Related work

The basic concepts of DST and its application to classification are first recalled in Section 2.1. The contextual discounting operation, which plays a central role in our approach, is described separately in Section 2.2. The evidential neural network model used in this paper is then introduced in Section 2.3, and related work on multimodal medical image fusion is briefly reviewed in Section 2.4.

### *2.1. Dempster-Shafer theory*

Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_K\}$ be the finite set of possible answers to some question, called the *frame of discernment*. Evidence about a variable taking values in $\Theta$ can be represented by a *mass function* $m : 2^\Theta \to [0, 1]$, such that

$$\sum_{A \subseteq \Theta} m(A) = 1 \quad \text{and} \quad m(\emptyset) = 0.$$

Each subset $A \subseteq \Theta$ such that $m(A) > 0$ is called a *focal set* of $m$. The mass $m(A)$ represents a share of a unit mass of belief allocated to focal set $A$, which cannot be allocated to any strict subset of $A$. The mass $m(\Theta)$ can be interpreted as a degree of ignorance. Full ignorance is represented by the *vacuous* mass function $m_?$ verifying $m_?(\Theta) = 1$. If all focal sets are singletons, then $m$ is said to be *Bayesian*; it is equivalent to a probability distribution.

*Belief and plausibility functions.* The information provided by a mass function $m$ can also be represented by a *belief function* $Bel$ or a *plausibility function* $Pl$ from $2^\Theta$ to $[0, 1]$ defined, respectively, as:

$$Bel(A) = \sum_{B \subseteq A} m(B)$$

and

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) = 1 - Bel(\bar{A}),$$

for all $A \subseteq \Theta$, where $\bar{A}$ denotes the complement of $A$. The quantity $Bel(A)$ can be interpreted as a degree of support for $A$, while $Pl(A)$ is a measure of lack of support against $A$. The *contour function* $pl$ associated to $m$ is the function that maps each element $\theta$ of $\Theta$ to its plausibility, i.e.,

$$pl(\theta) = Pl(\{\theta\}), \quad \forall \theta \in \Theta.$$

As shown below, this function can be easily computed when combining several pieces of evidence; it plays an important role in decision-making.

*Dempster's rule.* In DST, the beliefs about a certain question are established by aggregating independent pieces of evidence represented by belief functions over the same frame of discernment [69]. Given two mass functions $m_1$ and $m_2$ derived from two independent items of evidence, the mass function $m_1 \oplus m_2$ representing the pooled evidence is defined as

$$(m_1 \oplus m_2)(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C), \tag{1a}$$

for all $A \subseteq \Theta, A \neq \emptyset$, and $(m_1 \oplus m_2)(\emptyset) = 0$. The coefficient $\kappa$ is the *degree of conflict* between $m_1$ and $m_2$,

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C). \tag{1b}$$

This operation is called *Dempster's rule of combination*. It is commutative and associative. The combined mass function $m_1 \oplus m_2$ is called the *orthogonal sum* of $m_1$ and $m_2$. Mass functions $m_1$ and $m_2$ can be combined if and only if $\kappa < 1$. Let $pl_1$, $pl_2$ and $pl_1 \oplus pl_2$ denote the contour functions associated with, respectively, $m_1$, $m_2$ and $m_1 \oplus m_2$. The following equation holds:

$$\forall \theta \in \Theta, \quad (pl_1 \oplus pl_2)(\theta) = \frac{pl_1(\theta) pl_2(\theta)}{1-\kappa}. \tag{2}$$

The complexity of calculating the combined contour function using (2) is linear in the cardinality of $\Theta$, whereas computing the combined mass function using (1) has, in the worst-case, exponential complexity.

*Conditioning.* Given a mass function $m$ and a nonempty subset $A$ of $\Theta$ such that $Pl(A) > 0$, the conditional mass function $m(\cdot|A)$ is defined as the orthogonal sum of $m$ and the mass function $m_A$ such that $m(A) = 1$. Conversely, given a conditional mass function $m_0$ given $A$ (expressing one's beliefs in a context where it is only known that the truth lies in $A$), its *conditional embedding* [73] is the least precise mass function $m$ on $\Theta$ such that $m(\cdot|A) = m_0$; it is obtained by transferring each mass $m_0(C)$ to $C \cup \overline{A}$, for all $C \subseteq A$. Conditional embedding is a form of "deconditioning", i.e., it performs the inverse of conditioning.

*Plausibility-probability transformation.* Once a mass function representing the combined evidence has been computed, it is often used to make a decision. Decision-making methods in DST are reviewed in [18]. Here, we will use the simplest method [11], which consists in computing a probability distribution on $\Theta$ by normalizing the plausibilities of the singletons,

$$\forall \theta \in \Theta, \quad p(\theta) = \frac{pl(\theta)}{\sum_{k=1}^{K} pl(\theta_k)}. \tag{3}$$

Once probabilities have been computed, a decision can be made by maximizing the expected utility. We note that this method fits well with Dempster's rule, as the plausibility of the singletons can be easily computed from (2) without computing the whole combined mass function.

### 2.2. Modeling the reliability of evidence

In the DST framework, the reliability of a source of information can be taken into account using the *discounting* operation, which transforms a mass function into a weaker, less informative one and thus allows us to combine information from unreliable sources [69]. Let $m$ be a mass function on $\Theta$ and $\beta$ a real number in $[0, 1]$ interpreted as the degree of belief that the source mass function $m$ is reliable. The discounting operation [69] with discount rate $1 - \beta$ transforms mass function $m$ into a less informative one ${}^{\beta}m$ defined as a weighted sum of $m$ and the vacuous mass function $m_?$, with coefficients $\beta$ and $1 - \beta$:

$$ {}^{\beta}m = \beta\, m + (1 - \beta)\, m_?. \tag{4} $$

In the rest of this paper, we will refer to $\beta$ as a *reliability coefficient*. When $\beta = 1$, we accept the mass function $m$ provided by the source and take it as a description of our knowledge; when $\beta = 0$, we reject it and are left with the vacuous mass function $m_?$.

The discounting operation plays an important role in many applications of DST, where it makes it possible to take into account "meta-knowledge" about the reliability of a source of information. It can be justified as follows [74]. Assume that $m$ is provided by a source that may be reliable ($R$) or not ($\neg R$). If the source is reliable, we adopt its opinion as ours, i.e., we set $m(\cdot|R) = m$. If it is not reliable, then it leaves us in a state of total ignorance, i.e., $m(\cdot|\neg R) = m_?$. Furthermore, assume that we have the following mass function on $\mathcal{R} = \{R, \neg R\}$: $m_{\mathcal{R}}(\{R\}) = \beta$ and $m_{\mathcal{R}}(\mathcal{R}) = 1 - \beta$, i.e., our degree of belief that the source is reliable is equal to $\beta$. Then, combining the conditional embedding of $m(\cdot|R)$ with $m_{\mathcal{R}}$ yields precisely ${}^{\beta}m$ in (4), after marginalizing on $\Theta$.

*Contextual discounting.* In [58], the authors generalize the discounting operation using the notion of *contextual discounting*, which makes it possible to account for richer metaknowledge about the reliability of a source in different contexts, i.e., conditionally on different hypotheses regarding the variable of interest. In the corresponding refined model, $m(\cdot|R)$ and $m(\cdot|\neg R)$ are defined as before, but our beliefs about the reliability of the source are now defined by $K$ coefficients $\beta_1, \ldots, \beta_K$, one for each state in $\Theta$. More specifically, we have $K$ conditional mass functions defined by $m_{\mathcal{R}}(\{R\}|\theta_k) = \beta_k$ and $m_{\mathcal{R}}(\mathcal{R}|\theta_k) = 1 - \beta_k$, for $k = 1, \ldots, K$. In this model, $\beta_k$ is, thus, the degree of belief that the source of information is reliable, given that the true state is $\theta_k$. As shown in [58], combining the conditional embeddings of $m(\cdot|R)$ and $m_{\mathcal{R}}(\cdot|\theta_k)$ for $k = 1, \ldots, K$ by Dempster's rule yields the following discounted mass function,

$$ {}^{\boldsymbol{\beta}}m(A) = \sum_{B \subseteq A} m(B) \left( \prod_{\theta_k \in A \setminus B} (1 - \beta_k) \prod_{\theta_l \in \overline{A}} \beta_l \right) \tag{5} $$

for all $A \subseteq \Theta$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ is the vector of all reliability coefficients, and a product of terms is equal to 1 if the index set is empty. In many applications, we actually do not need to compute the whole mass function (5): we can compute only the associated

contour function ${}^{\boldsymbol{\beta}}pl$, which is all we need for decision-making. As shown in [58], this contour function is equal to

$$ {}^{\boldsymbol{\beta}}pl(\theta_k) = 1 - \beta_k + \beta_k pl(\theta_k), \quad k = 1, \ldots, K. \tag{6} $$

It can be computed in linear time with respect to the size of $\Theta$, instead of exponential time for ${}^{\boldsymbol{\beta}}m$. An evidential $k$ nearest neighbor rule based on the contextual discounting operation was introduced in [22].

**Example 1.** *Consider a simplified diagnostic problem in which a patient may have one of two diseases denoted by $\theta_1$ and $\theta_2$. Assume that $\theta_1$ is a heart disease while $\theta_2$ is a lung disease. A cardiologist examines the patient and describes his opinion by the following mass function on $\Theta = \{\theta_1, \theta_2\}$: $m(\{\theta_1\}) = 0.7$, $m(\{\theta_2\}) = 0.2$, $m(\Theta) = 0.1$, i.e., his degrees of belief in $\theta_1$ and $\theta_2$ are, respectively, 0.7 and 0.2. Furthermore, suppose that the cardiologist is fully reliable to diagnose heart diseases ($\beta_1 = 1$), i.e., if the true state of the patient is $\theta_1$, the physician's opinion can be fully trusted, whereas he is only 60% reliable to diagnose lung diseases ($\beta_2 = 0.6$), i.e., if $\theta_2$ is the true disease, there only is only 60% chance that the physician's diagnostic is relevant. Applying formula* (5) *to $m$ gives the following discounted mass function:*

$$ \begin{aligned} {}^{\boldsymbol{\beta}}m(\{\theta_1\}) &= \beta_2 m(\{\theta_1\}) = 0.42 \\ {}^{\boldsymbol{\beta}}m(\{\theta_2\}) &= \beta_1 m(\{\theta_2\}) = 0.2 \\ {}^{\boldsymbol{\beta}}m(\Theta) &= 1 - \beta_2 m(\{\theta_1\}) - \beta_1 m(\{\theta_2\}) = 0.38. \end{aligned} $$

*The contour function of the original mass function $m$ is*

$$ \begin{aligned} pl(\{\theta_1\}) &= 0.7 + 0.1 = 0.8 \\ pl(\{\theta_2\}) &= 0.2 + 0.1 = 0.3. \end{aligned} $$

*After contextual discounting, we get*

$$ \begin{aligned} {}^{\boldsymbol{\beta}}pl(\{\theta_1\}) &= 0.42 + 0.38 = 0.8 \\ {}^{\boldsymbol{\beta}}pl(\{\theta_2\}) &= 0.2 + 0.38 = 0.58. \end{aligned} $$

*We can check that ${}^{\boldsymbol{\beta}}pl(\{\theta_1\}) = 1 - 1 + 1 \times 0.8$ and ${}^{\boldsymbol{\beta}}pl(\{\theta_2\}) = 1 - 0.6 + 0.6 \times 0.3$, which is consistent with* (6).

### *2.3. Evidential neural network*

In [16], Denœux proposed a DST-based evidential neural network (ENN) classifier in which mass functions are computed based on distances between the input vector and prototypes. As shown in Figure 1, the ENN model comprises a prototype activation layer, a mass calculation layer, and a combination layer.

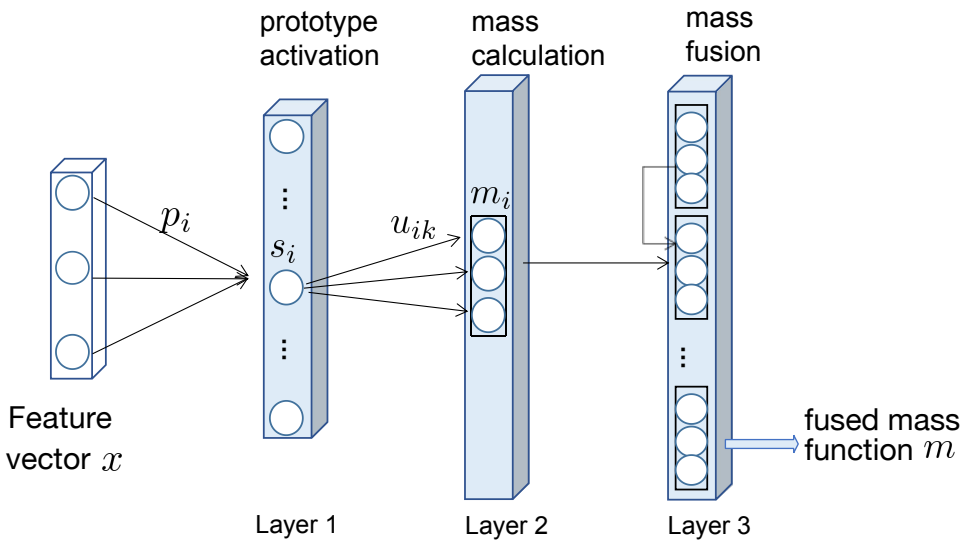


Figure 1: The evidential neural network model.

The prototype activation layer comprises $I$ units, whose weight vectors are prototypes $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_I$ in input space. The activation of unit $i$ in the prototype layer is

$$s_i = \alpha_i \exp(-\gamma_i \left\| \boldsymbol{x} - \boldsymbol{p}_i \right\|^2), \tag{7}$$

where $\gamma_i > 0$ and $\alpha_i \in [0, 1]$ are two parameters. Each quantity $s_i$ can be interpreted as a degree of similarity between input vector $\boldsymbol{x}$ and prototype $\boldsymbol{p}_i$.

The second hidden layer computes mass functions $m_i$ representing the evidence of each prototype $\boldsymbol{p}_i$, using the following equations:

$$m_i(\{\theta_k\}) = u_{ik} s_i, \quad k = 1, \ldots, K, \tag{8a}$$

$$m_i(\Theta) = 1 - s_i, \tag{8b}$$

where $u_{ik}$ is the membership degree of prototype $i$ to class $\theta_k$, and $\sum_{k=1}^{K} u_{ik} = 1$. The mass function $m_i$ can thus be seen as a discounted Bayesian mass function, with a discount rate $1 - s_i$; its focal sets are singletons and $\Theta$. The mass assigned to $\Theta$ increases with the distance between $\boldsymbol{x}$ and $\boldsymbol{p}_i$. Finally, the third layer combines the $I$ mass functions $m_1, \ldots, m_I$ using Dempster's rule (1). The output mass function $m = \bigoplus_{i=1}^{I} m_i$ is a discounted Bayesian mass function that summarizes the evidence of the $I$ prototypes.

The idea of applying the above model to features extracted by a convolutional neural network (CNN) was first proposed by Tong et al. in [79]. In this approach, the ENN module becomes an "evidential layer", which is plugged into the output of a CNN instead of the usual softmax layer. The feature extraction and evidential modules are trained simultaneously. Huang et al. applied the ENN model to medical image segmentation within a deep evidential segmentation network [37].

**Remark 1.** *The approach described in this section should not be confused with the "evidential deep learning" approach introduced in [68] and applied to brain tumor segmentation in*

*[93]. The latter approach is based on learning the parameters of a Dirichlet distribution that represents second-order uncertainty on the class probabilities. Although the parameters of the Dirichlet distribution can be formally identified to a mass function whose focal sets are the singletons $\{\theta_k\}$ and the whole frame $\Theta$, this is actually a Bayesian approach that learns a probability distribution over the class probabilities through a suitable loss function.*

### *2.4. Multimodal medical image fusion*

Multimodal medical image fusion can be performed at the pixel, feature or decision level. Pixel-level fusion is the traditional approach; it can be conducted directly in the spatial domain or indirectly through the application of transformations and representations. The fusion of high-level features is typically performed by a neural network learning a shared representation or a joint embedding space derived from multimodal features. Decision fusion consists in pooling decisions made independently from different image modalities; it can be performed with traditional or deep-learning approaches. In the following, we review previous work on multimodal medical image fusion, emphasizing the distinction between traditional and deep-learning approaches.

#### *2.4.1. Traditional approaches*

Traditional fusion methods aim at combining relevant information (either pixels themselves or low-level image features) from multiple images to produce a single fused image with enhanced features for further analysis. Four main approaches have been proposed: multi-scale transformation, sparse representation extraction, edge-preserving filters, and meta-heuristic optimization. The first three approaches focus on effective image representation, while the last one aims at combining the represented features efficiently.

The multi-scale transform approach decomposes images into different scales or frequency components using techniques such as wavelet transform [72], contourlet transforms [86], pyramid transforms [23] or curvelet transform [3], allowing relevant features from each source image to be combined. Sparse representation extraction assumes that multimodal images can be represented as a sparse linear combination of basis functions; search techniques such as dictionary learning [43] or sparse coding with dictionary learning [81] are used to obtain the sparse image representation and to merge images focusing on the most important features. Edge-preserving filters ensure the preservation of edges while smoothing images to ensure the fusion of critical features without blurring [75]. Commonly used filters include bilateral filters [47], guided filters [59], anisotropic diffusion [82], and total variation minimization [91]. The three above approaches can be used independently or in combination, which often yields better results. For example, in [35], Hu et al. propose a multimodal medical image fusion method based on separable dictionary learning and Gabor filtering; in [83], Wang et al. describe a multimodal medical image fusion method using Laplacian pyramid and adaptive sparse representations; in [51], Liu et al. introduce a general image fusion framework based on multi-scale transform and sparse representation.

In addition to studying effective image representations, a complementary research direction has been to design meta-heuristic optimization algorithms allowing one to find the best fusion parameters for combining features obtained by different transform, sparse or fitting

algorithms. Many approaches use meta-heuristic optimization techniques such as genetic algorithms [3], particle swarm optimisation [77] or ant colony optimisation [70].

### *2.4.2. Deep learning approaches*

Recent advances in deep learning have allowed breakthroughs in medical image fusion by making it to learn a joint embedding or a shared representation space from multiple features. Recent techniques include adversarial learning [67], co-training [89], multi-kernel learning [57], multi-task learning [52], etc. These methods exploit the ability of neural networks to extract meaningful representations and perform fusion in high-level feature spaces with learnable feature fusion rules. These approaches enable more sophisticated and robust image fusion, capable of handling complex relationships and producing high-quality fused image features. Here, we summarize three important models commonly used for multi-model medical image fusion.

*Convolutional Neural Networks.* Convolutional neural networks (CNNs) are widely used in image processing due to their strong feature representation capability. Within CNNs, various fusion operations can be used to effectively integrate information from different imaging modalities. Such operations include but are not limited to, concatenation, element-wise addition and multiplication, weighted sum, max pooling, etc. Fusion can occur at different stages of the network, i.e., early, middle, or late stages.

Early fusion stacks different modalities along a channel dimension and feeds into a single CNN [49]. This is the simplest operation but it requires high image registration quality. In the case of middle fusion, separate CNN branches are employed to extract features from each modality, which are subsequently concatenated at the feature level or fused in a particular common representation space. More recently, transformer-based CNN architectures, such as the Vision Transformer (ViT) [33], have also demonstrated considerable versatility in handling diverse types of data with the introduction of an attention mechanism [46]. CNNs can also be integrated with some traditional fusion ideas to obtain more robust fusion results using, e.g., the multiscale transformer [76] or multiscale residual pyramid attention network [28].

In contrast to the emphasis on image pixels or features in earlier fusion techniques, later fusion places greater importance on the aggregation of high-level decisions. It integrates information derived from preliminary classifications with the application of appropriate fusion rules. Approaches can be classified into two main categories: 1) *hard fusion* methods, which merge logical information membership values, such as model ensembling with majority or average voting [42]; and 2) *soft fusion* methods, where classifiers assign numerical values to reflect their confidence in decisions, as exemplified by fuzzy voting [32, 27].

*Encoder-Decoder Networks.* Encoder-decoder networks are another type of convolutional neural network commonly used for image segmentation and reconstruction. Within the encoder-decoder network, multiple encoders are used to extract deep features from each modality. These features are subsequently integrated either through a straightforward concatenation process or through a latent layer or learnt joint embedding space. The fused features are then passed to the decoder to produce the final image. Compared with CNNs,

the Encoder-decoder architecture offers a more structured and effective fusion framework with enhanced feature representation, precise spatial alignment, and flexible and effective fusion strategies. Multimodal Transformer (MMT) [85] is one of the most sophisticated forms of multimodal encoder-decoder networks that employ self-attention mechanisms to integrate and process multimodal data in an effective manner; nnFormer [92] has been identified as the most advanced model for multimodal MRI brain tumor segmentation.

*Generative Adversarial Networks.* Generative Adversarial Networks (GANs), composed of a generator and a discriminator, are capable of learning complex relationships between disparate modalities through the generation of highly realistic images via unsupervised adversarial training [30]. In the context of multimodal medical image fusion, the generator learns to generate a fused image that combines the semantic features of the inputs from different modalities. The discriminator guides the generator to produce high-quality fused images by distinguishing between the fused and the real images. GAN-based fusion methods are particularly useful for advanced medical image fusion tasks where the quality and realism of the fused image are of paramount importance, such as the combination of structural and functional imaging modalities. For example, in [88], the authors propose a conditional generative adversarial network with a transformer for multimodal image fusion by introducing a wavelet fusion module to maintain long-distance dependencies across domains; in [67], the authors introduce an unsupervised medical fusion generative adversarial network to generate an image with CT bone structure and MRI soft tissue contrast by fusing CT and MRI image sequences.

Although a lot of research has been devoted to the study of multimodal medical image segmentation and promising experimental results have been obtained, modeling the reliability of each modality in a given context and quantifying the uncertainty on the outcome of the fusion process remain challenging research questions. In this paper, we address these questions using a deep evidential fusion framework combining deep learning with DST, and taking into account the reliability of each of the modalities being combined. The proposed decision-fusion framework is described in detail in the following section.

## 3. Proposed framework

The main idea of this paper is to hybridize a deep evidential fusion framework with uncertainty quantification and reliability learning for multimodal medical image segmentation under the framework of DST. The architecture of the system is described in Section 3.1, and the loss function used to train the whole framework end-to-end is presented in Section 3.2.

### *3.1. Architecture*

The proposed framework is depicted in Figure 2. Features are first extracted from different modalities using independent encoder-decoder feature-extraction (FE) modules. The features from each modality are then transformed into mass functions using evidence mapping (EM) modules. Finally, mass functions are discounted and combined in a multi-modality evidence fusion (MMEF) module. These modules are described in greater detail below.

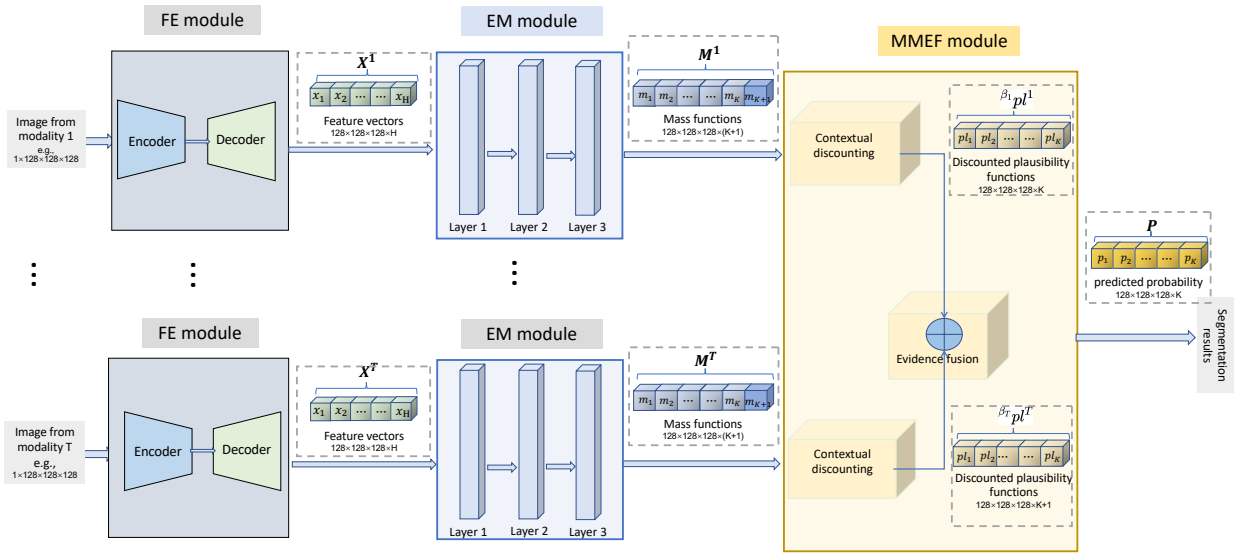


Figure 2: The proposed deep evidential fusion framework. It is composed of encoder-decoder *feature extraction (FE)* modules that represent images using deep features, *evidence mapping (EM)* modules that map deep features into mass functions, and a *multimodal evidence fusion (MMEF)* module that combines evidence from different modalities.

### *3.1.1. Feature-extraction (FE) module*

Deep neural network architectures have been shown to be very powerful for extracting relevant information from high-dimensional data. Our approach is compatible with any deep FE architecture. The baseline model considered in this paper is UNet [41], a foundational medical image segmentation model. As illustrated in Figure 3, a UNet-based feature extraction module incorporates residual connections within each layer, following the same architecture as in [37]. Each layer of the module comprises encoding and decoding paths, connected by skip connections. In the encoding path (represented by blue blocks), the data undergoes downsampling through stride convolutions, while the decoding path (represented by green blocks) employs stride transpose convolutions for upsampling. The bottom layer, represented by the gray block, serves as the base connection without performing any down or up-sampling of the data. In Section 4.3, in addition to UNet, we will also consider the more recent nnUNet [40] and nnFormer [92] models as alternative FE modules. The settings of these modules will be described in Section 4.1.

### *3.1.2. Evidence mapping (EM) module*

The EM module is based on the ENN architecture recalled in Section 2.3. It is identical to that described in [37]. As illustrated in Figure 2, we have one such module for each modality. The input to each module is a tensor containing the $H$ features extracted for each voxel. The prototypes are, thus, vectors in the $H$-dimensional space of features extracted from modality $t$ images by the FE module. As explained in Section 2.3, a prototype layer first computes the similarities between feature vectors and prototypes using (7). The next layer

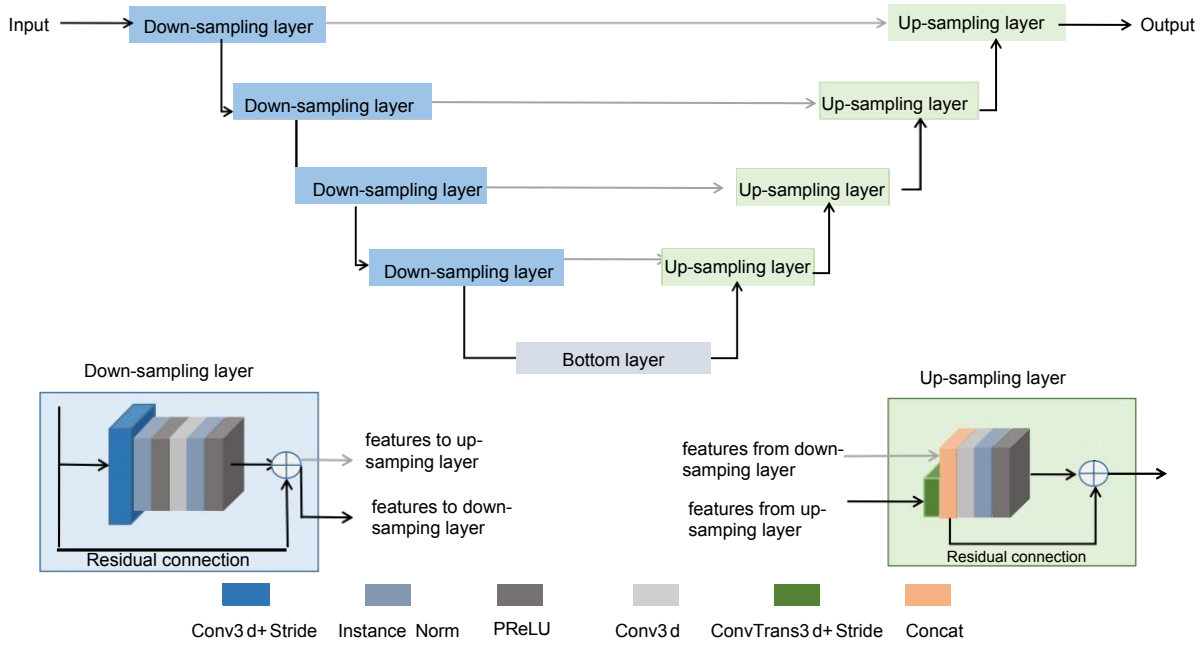


Figure 3: Schematic description of a UNet-based FE module. The network consists of a contracting path (down-sampling layers) and an expansive path (up-sampling layers), which gives it the u-shaped architecture. Reproduced based on [41].

computes mass functions for each prototype using (8) (see Figure 1). Finally, the prototype-based mass functions are combined by Dempster's rule (1) in a third layer. Denoting by $\Theta = \{\theta_1, \ldots, \theta_K\}$ the set of classes, the EM module thus computes, for each voxel $n$ and modality $t$, a mass function[2] $m_n^t$ with focal sets $\{\theta_k\}$, $k = 1, \ldots, K$ and $\Theta$. The mass $m_n^t(\Theta)$ is a measure of the segmentation uncertainty for classifying voxel $n$ in the image of modality $t$.

### *3.1.3. Multi-modality evidence fusion (MMEF) module*

This module first transforms the contour functions from the EM modules using the contextual discounting operation recalled in Section 2.2. The contour function for voxel $n$ and modality $t$ is obtained from mass function $m_n^t$ as

$$pl_n^t(\theta_k) = m_n^t(\{\theta_k\}) + m_n^t(\Theta), \quad k = 1, \ldots, K.$$

Using (6), the discounted contour function is given by

$$^{\boldsymbol{\beta}^t}pl_n^t(\theta_k) = 1 - \beta_k^t + \beta_k^t pl_n^t(\theta_k), \quad k = 1, \ldots, K, \tag{9}$$

[2]Throughout this paper, we use an upper index $t$ to denote modalities, and lower indices $n$ and $k$ to denote, respectively, voxels and classes.

where $\boldsymbol{\beta}^t = (\beta_1^t, \ldots, \beta_K^t)$ is the vector of discounting (reliability) coefficients for modality $t$. We recall that $\beta_k^t$ represents our degree of belief that the modality $t$ is reliable when it is known that the actual class of voxel $n$ is $\theta_k$. From (2), the combined contour function at voxel $n$ can then be computed up to a multiplicative constant by multiplying the contour functions for the $T$ modalities as

$$^{\boldsymbol{\beta}}pl_n(\theta_k) \propto \prod_{t=1}^{T} {}^{\boldsymbol{\beta}^t}pl_n^t(\theta_k), \quad k = 1, \ldots, K,$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}^1, \ldots \boldsymbol{\beta}^T)$ is the vector of $KT$ reliability coefficients for the the $K$ classes and $T$ modalities. Finally, the predicted probability distribution $P_n$ for voxel $n$ after combining evidence from the $T$ modalities is obtained from (3) as

$$^{\boldsymbol{\beta}}p_n(\theta_k) = \frac{^{\boldsymbol{\beta}}pl_n(\theta_k)}{\sum_{l=1}^{K} {}^{\boldsymbol{\beta}}pl_n(\theta_l)} = \frac{\prod_{t=1}^{T}(1 - \beta_k^t + \beta_k^t pl_n^t(\theta_k))}{\sum_{l=1}^{K}\prod_{t=1}^{T}(1 - \beta_l^t + \beta_l^t pl_n^t(\theta_l))}, \quad k = 1, \ldots, K. \tag{10}$$

The learnable parameters in this module are the $KT$ reliability coefficients in vector $\boldsymbol{\beta}$.

### *3.2. Loss function*

The whole framework is optimized by minimizing the following loss function,

$$\mathsf{loss} = \mathsf{loss}_s + \mathsf{loss}_f,$$

where

- The term $\mathsf{loss}_s$ is the Dice loss quantifying the segmentation performance of each source modality independently, with

$$\mathsf{loss}_s = \sum_{t=1}^{T}\left[1 - \frac{2\sum_{n=1}^{N}\sum_{k=1}^{K} m_n^t(\{\theta_k\}) \times G_{kn}}{\sum_{n=1}^{N}\sum_{k=1}^{K}(m_n^t(\{\theta_k\}) + G_{kn})}\right], \tag{11}$$

  where $N$ is the number of voxels, and $G_{kn} = 1$ if voxel $n$ belongs to class $\theta_k$, and $G_{kn} = 0$ otherwise;

- The term $\mathsf{loss}_f$ quantifies the segmentation performance after combination:

$$\mathsf{loss}_f = 1 - \frac{2\sum_{n=1}^{N}\sum_{k=1}^{K} {}^{\boldsymbol{\beta}}p_n(\theta_k) \times G_{kn}}{\sum_{n=1}^{N}\sum_{k=1}^{K} {}^{\boldsymbol{\beta}}p_n(\theta_k) + G_{kn}}, \tag{12}$$

  where ${}^{\boldsymbol{\beta}}p_n$ is the predicted probability distribution for voxel $n$ given by (10).

The learnable parameters are the weights of the FE module, the prototypes and associated parameters $\alpha_i$, $\gamma_i$ and $u_{ik}$ of the EM module, and the reliability coefficients $\beta_k^t$ in the MMEF module. Learning the reliability coefficients is an original feature of our approach. As shown in Sections 4.2 and 4.3, these coefficients can allow us to gain some insight into the multi-modality segmentation process.

## 4. Experiments and results

In this section, the proposed framework described in Section 3 is applied to two real multimodal medical image datasets. The experimental settings are first described in Section 4.1. The results on the two datasets are then reported in Sections 4.2 and 4.3.

### *4.1. Experimental settings*

*Datasets.* The proposed framework was tested on two multimodal medical image datasets.

The *PET-CT lymphoma dataset* contains 3D images from 173 patients who were diagnosed with large B-cell lymphomas and underwent PET-CT examination[3]. For lymphoma segmentation, PET imaging helps identify active tumor sites by highlighting areas of increased metabolic activity. In contrast, CT imaging provides anatomical information about the size, shape, location, and surrounding structures of lymphoma tumors. While PET image makes it possible to obtain functional information about the tumor and surrounding tissues, CT images provide complementary anatomical details allowing for more accurate segmentation. The lymphomas in mask images were delineated manually by experts and considered as ground truth. Figure 4 shows an example of PET and CT images of a patient with lymphomas. The PET and CT images and the corresponding mask images have different sizes and spatial resolutions due to the use of different imaging machines and operations. For CT images, the size varies from $267 \times 512 \times 512$ to $478 \times 512 \times 512$. For PET images, the size varies from $276 \times 144 \times 144$ to $407 \times 256 \times 256$.

The *multi-MRI brain tumor* dataset was made available for the BraTS2021 challenge [5]. The original BraTS2021 dataset comprises training, validation, and test sets with, respectively, 1251, 219, and 570 cases. There are four modalities: FLAIR, T1Gd, T1, and T2 with $240 \times 240 \times 155$ voxels. Figure 5 shows examples of four-modality MRI slices for one patient. The appearance of brain tumors varies in different modalities [5]. T1Gd MRI images are obtained following the administration of a gadolinium-based contrast agent that enhances areas with disrupted blood-brain barrier such as tumor regions, making tumors appear hyperintense (bright) and improving the visibility of tumor margins. FLAIR MRI images suppress the signal from cerebrospinal fluid (CSF), highlighting pathological changes while suppressing the CSF signal. T2 MRI images are sensitive to tissue water content and provide good contrast between soft tissues. Tumors with increased water content often appear hyperintense (bright) on T2 images. T1 MRI images are crucial for identifying tumor location and structural details by their excellent anatomical detail. Annotations of scans comprise gadolinium (GD)-enhancing tumor (ET), necrotic and non-enhancing tumor core (NRC/NET), and peritumoral edema (ED). The task of the BraTS2021 challenge was to segment the images into three overlapping regions: ET, tumor core (TC, the union of ET and NRC/NET), and whole tumor (WT, the union of ET, NRC/NET, and ED). In this work, we evaluated the segmentation performances with respect to these three overlapping regions to allow a fair comparison with other state-of-the-art methods. Additionally, we also

[3]The study was approved as a retrospective study by the Henri Becquerel Center Institutional Review Board.

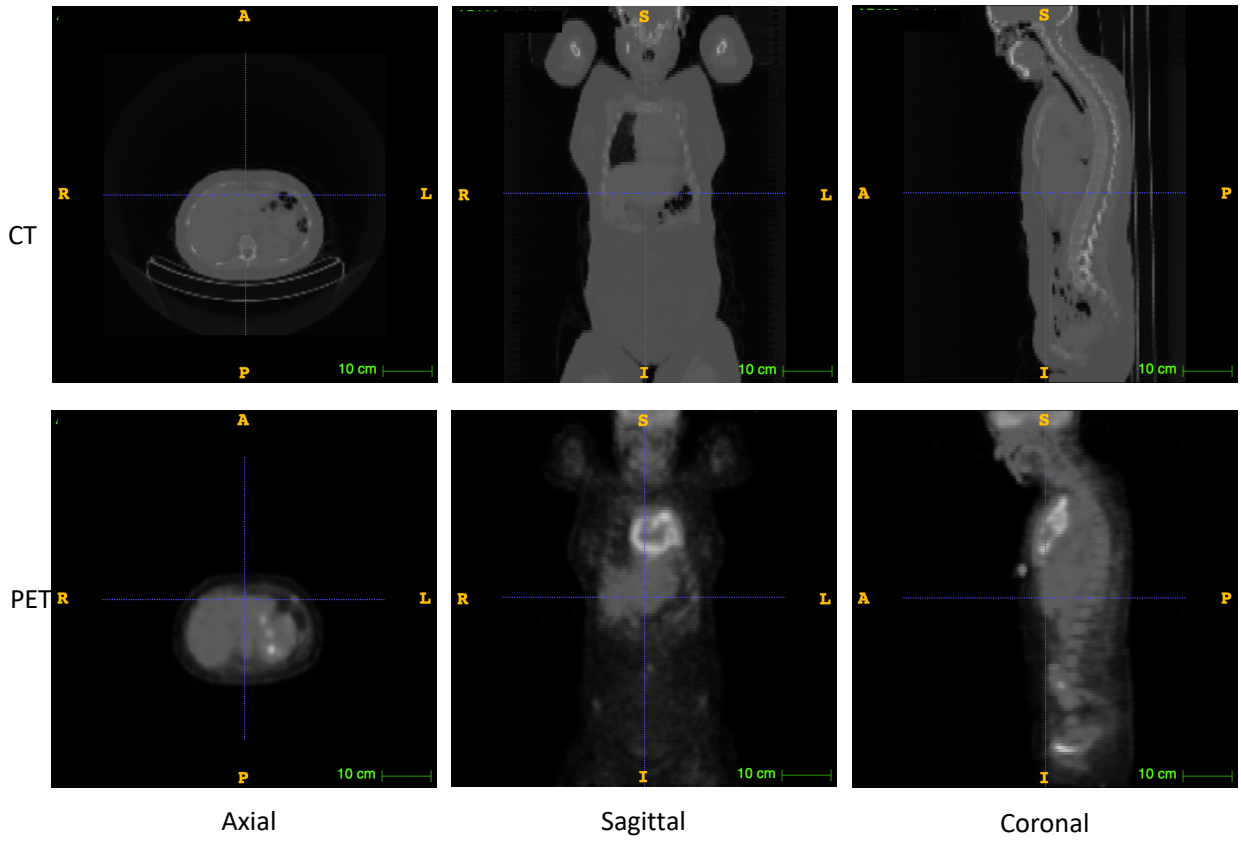


Figure 4: Example of a patient with lymphomas. The first and second rows showcase, respectively, CT and PET slices, depicting axial, sagittal, and coronal views. The lymphomas correspond to the bright regions in PET slices.

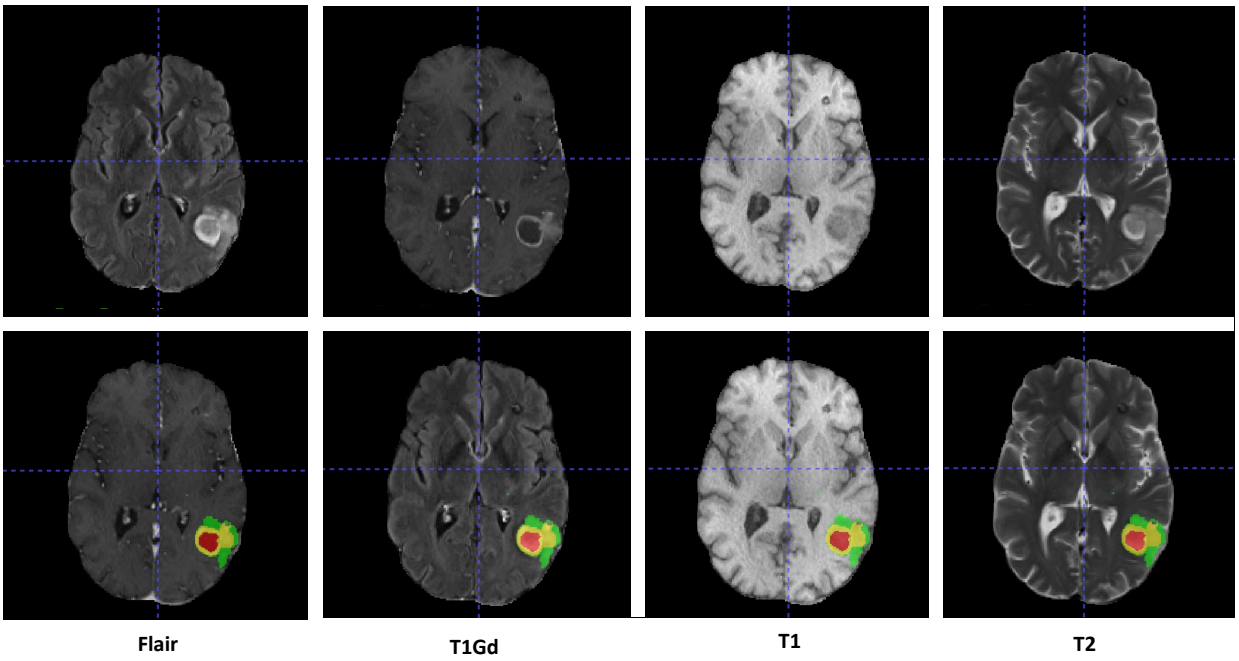


Figure 5: Examples of a patient with brain tumors in four MRI modalities: FLAIR, T1Gd, T1, and T2. The first and second rows show, respectively, the original images and the images with tumor masks for the three classes: peritumoral edema (ED, green), enhancing tumor (ET, yellow), and necrotic tumor core or non-enhancing tumor (NCR/NET, red).

compared the results with respect to the three original non-overlapping tumor regions to highlight the impact of contextual discounting on subregion segmentation.

*Pre-processing.* For the PET-CT dataset, we first normalized the PET, CT and mask images: (1) for PET images, we applied a random intensity shift and scale to each channel with a shift value of 0 and scale value of 0.1; (2) for CT images, the shift and scale values were set to 1000 and 1/2000; (3) for mask images, the intensity value was normalized into the $[0, 1]$ interval by replacing the outside value by 1. We then resized the PET and CT images to $256 \times 256 \times 128$ by linear interpolation and mask images to $256 \times 256 \times 128$ by nearest neighbor interpolation. Lastly, CT and PET images were registered using B-spline interpolation. Following [37], we randomly divided the 173 scans into subsets of size 138, 17, and 18 for, respectively, training, validation, and test. The training process was then repeated five times to test the stability of our framework, with different data used exactly once as the validation and test data.

For the BraTS2021 dataset, we used the same pre-processing operation as in [60]. We first performed a min-max scaling operation and clipped intensity values to standardize all volumes; we then cropped/padded the volumes to a fixed size of $128 \times 128 \times 128$ by removing the unnecessary background (the cropping/padding operation was only applied to training data). No data augmentation technique was applied, and no additional data was used in this study. Since the ground truth labels are unavailable for the validation and test sets, we trained and tested our framework with the training set. Following [60], we randomly divided the 1251 training scans into subsets of 834, 208, and 209 cases for

training, validation, and testing, respectively. The process was repeated five times to test the stability of our framework. All the preprocessing methods mentioned in this paper can be found in the SimpleITK [53] toolkit.

All the compared methods used the same dataset composition and pre-processing operations. They were implemented in Python with the PyTorch-based medical image framework MONAI[4].

*Parameter initialization and learning.* At the FE stage, the number of filters in UNet was set to $(8, 16, 32, 64, 128)$ with kernel size equal to five and convolutional strides equal to $(2, 2, 2, 2)$ for layers from left to right. For nnUNet used in Section 4.3, the kernel size was set to $(3, (1, 1, 3), 3, 3)$ and the upsample kernel size was set to $(2, 2, 1)$ with strides $((1, 1, 1), 2, 2, 1)$. For nnFormer used in Section 4.3, the crop size was set to $(128, 128, 128)$ with embedding dimension set to 96 and the number of heads was set to $(3, 6, 12, 24)$. The number of extracted features was $H = 2$ for the PET-CT lymphoma dataset and $H = 4$ for the multi-MRI BraTS2021 dataset.

To train our fusion framework, we proceeded in three steps. First, FE modules (i.e., UNet, nnUNet, or nnFormer) were pre-trained independently for each modality during 50 epochs. Then, the weights of the FE modules were fixed, and the parameters of the EM and MMEF modules were optimized. Finally, the whole framework was fine-tuned for a few epochs. The initial values of parameters $\alpha_i$ and $\gamma_i$ in the EM modules were set to 0.5 and 0.01, and the membership degrees $u_{ik}$ were initialized randomly by drawing uniform random numbers, and normalizing. We used, $I = 10$ prototypes for the PET-CT lymphoma dataset, and $I = 20$ prototypes for the more complex multi-MRI BraTS2021 dataset. These prototypes were randomly initialized from a normal distribution with zero mean and an identity covariance matrix. Details about the initialization of the EM module can be found in [37]. The reliability coefficients $\beta_k^t$ in the MMEF module were initialized at 0.5.

For both datasets, we used the Adam optimization algorithm with an early stopping strategy: training was stopped when there was no improvement in performance on the validation set during ten epochs. The initial learning rate was set to 0.01. The batch size was set to 4. For all the compared methods, the model with the best performance on the validation set was saved as the final model for testing[5].

*Evaluation criteria.* Although many authors have shown that segmentation performance can be improved by merging multimodal medical images into deep neural networks [63, 2], the reliability of information sources and the quality of uncertainty quantification have rarely been investigated. Here, the former issue will be addressed by analyzing the reliability coefficients $\beta_k^t$ defined in Section 3.1.3. To assess the quality of uncertainty quantification, we will use three metrics: the *Brier score* [9], the *negative log-likelihood* (NLL), and *Expected Calibration Error* (ECE) [31]. These metrics provide a robust evaluation framework for the

[4]More details about how to use those models can be found in MONAI core tutorials `https://monai.io/started.html##monaicore`.

[5]The code is available at `https://github.com/iWeisskohl/Deep-evidential-fusion`.

uncertainty of the segmentation results, with smaller values indicating better performance. Their definitions are recalled below.

The Brier Score and NLL are defined, respectively, as

$$\mathsf{BS} = \frac{1}{N}\sum_{n=1}^{N}(P_n - G_n)^2,$$

and

$$\mathsf{NLL} = -\sum_{n=1}^{N} G_n \log P_n + (1 - G_n)\log(1 - P_n),$$

where $G_n$ is the ground truth of voxel $n$, $P_n$ is the predicted probability of voxel $n$, and $N$ is the number of voxels.

The ECE measures the correspondence between predicted probabilities and ground truth. The output normalized plausibilities of the model are first discretized into equally spaced bins $E_b$, $b \in [1, B]$ ($B = 10$ in this paper). The accuracy of bin $E_b$ is defined as

$$\mathsf{acc}(E_b) = \frac{1}{\mid E_b \mid}\sum_{n \in E_b} \mathbf{1}(S_n = G_n),$$

where $S_n$ is the predicted class label for voxel $n$ and $\mathbf{1}(\cdot)$ is the indicator function. The average confidence of bin $E_b$ is defined as

$$\mathsf{conf}(E_b) = \frac{1}{\mid E_b \mid}\sum_{n \in E_b} P_n.$$

The ECE is the weighted average of the difference in accuracy and confidence of the bins:

$$\mathsf{ECE} = \sum_{b=1}^{B} \frac{\mid E_b \mid}{N} \mid \mathsf{acc}(E_b) - \mathsf{conf}(E_b) \mid .$$

A model is perfectly calibrated when $\mathsf{acc}(E_b) = \mathsf{conf}(E_b)$ for all $b \in \{1, ..., B\}$, in which case $\mathsf{ECE} = 0$.

Since our dataset has imbalanced foreground and background proportions, we only considered voxels belonging to the foreground or tumor region to calculate the above three indices. For the PET-CT lymphoma dataset, focusing only on the tumor region is not easy since the lymphomas are scattered throughout the whole body. Thus, we focused on the foreground region for this dataset. For the BraTS2021 dataset, we followed the suggestion from [66] to focus on the tumor region for the reliability evaluation. For each patient in the test set, we defined a bounding box covering the foreground or tumor region and calculated the corresponding values in this bounding box. For all segmentation performance criteria, the reported results were obtained by calculating the criteria for each test 3D scan and then averaging over the patients.

In addition to evaluating segmentation reliability, we also measured segmentation accuracy using the *Dice score*. In a segmentation task, the Dice score measures the volume of the overlapping region of the predicted object and the ground truth object as

$$\mathsf{Dice} = \frac{2\,TP}{FP + 2\,TP + FN},$$

where $TP$, $FP$, and $FN$ denote, respectively, the numbers of true positive, false positive, and false negative voxels.

### *4.2. Segmentation results on the PET-CT lymphoma dataset*

*Segmentation uncertainty.* The results concerning uncertainty estimation are reported in Table 1. Our model (MMEF-UNet) was compared to

1. UNet with a softmax decision layer (the baseline);
2. UNet with Monte-Carlo (MC) dropout [29] and deep ensemble [45], two popular techniques for improving the uncertainty quantification capabilities of probabilistic deep neural networks;
3. ENN-UNet, composed of UNnet as the FE module and the EM module in place of the softmax layer; this is the architecture studied in [37];
4. RBF-UNet, an alternative model composed of UNnet and a radial-basis function (RBF) module in place of the softmax layer; as shown in [37], this model makes it possible to compute output belief functions that are similar to those computed by ENN-UNet.

We can remark that approaches 1 to 4 above implement pixel-level fusion, whereas our approach is based on decision-level fusion. As for uncertainty quantification, UNet, UNet-MC and UNet-Ensemble are probabilistic methods. UNet only computes point estimates of class probabilities without taking into account second-order uncertainty. UNet-MC applies dropout during both training and inference, sampling multiple forward passes to estimate uncertainty by averaging the predictions. UNet-Ensemble quantifies uncertainty by averaging the predictions obtained from multiple independently-trained models. In contrast, ENN-UNet and RBF-UNet are evidential methods: they both calculate belief functions to represent segmentation evidence and uncertainty under the DST framework. For UNet-MC, the dropout rate was set to 0.2 and the number of samples was set to five; we averaged the five output probabilities at each voxel as the final output of the model. For UNet-ensembles, the number of samples was set to five; the five output probabilities were then averaged at each voxel as the final output of the model. The settings of ENN-UNet and RBF-ENN are the same as those reported in [37].

From Table 1, we can see that Monte-Carlo dropout and deep ensembles do not significantly improve the segmentation reliability as compared to the baseline UNet model, as shown, e.g., by the higher NLL values. In contrast, the addition of the EM module to the FE module, as implemented in ENN-UNet, brings a significant improvement, particularly according to NLL; the RBF-UNet model yields similar results. The decision-fusion framework MMEF-UNet brings an additional improvement according to all three criteria (ECE,

Table 1: Means and standard errors of segmentation quality and reliability measures for MMEF-UNet and the referenced uncertainty quantification methods on the lymphoma dataset. The best results are in bold and the second best are underlined.

| Model | ECE↓ | Brier score ↓ | NLL↓ | Dice score ↑ |
|---|---|---|---|---|
| UNet | $0.056 \pm 3.6 \times 10^{-3}$ | $0.065 \pm 3.9 \times 10^{-3}$ | $0.310 \pm 8.8 \times 10^{-2}$ | $0.770 \pm 3.2 \times 10^{-2}$ |
| UNet-MC | $0.053 \pm 4.6 \times 10^{-3}$ | $0.062 \pm 4.9 \times 10^{-3}$ | $0.400 \pm 8.7 \times 10^{-2}$ | $0.801 \pm 1.1 \times 10^{-2}$ |
| UNet-Ensemble | $0.063 \pm 7.6 \times 10^{-2}$ | $0.064 \pm 4.0 \times 10^{-3}$ | $0.343 \pm 7.2 \times 10^{-2}$ | $0.802 \pm 6.7 \times 10^{-3}$ |
| ENN-UNet | $\underline{0.050} \pm 3.5 \times 10^{-3}$ | $0.062 \pm 3.9 \times 10^{-3}$ | $\underline{0.191} \pm 1.4 \times 10^{-2}$ | $\underline{0.805} \pm 7.1 \times 10^{-3}$ |
| RBF-UNet | $0.051 \pm 3.3 \times 10^{-3}$ | $\underline{0.061} \pm 0.9 \times 10^{-3}$ | $0.193 \pm 1.3 \times 10^{-2}$ | $0.802 \pm 6.9 \times 10^{-3}$ |
| MMEF-UNet (ours) | $\mathbf{0.045} \pm 1.3 \times 10^{-3}$ | $\mathbf{0.056} \pm 2.7 \times 10^{-3}$ | $\mathbf{0.180} \pm 1.3 \times 10^{-2}$ | $\mathbf{0.811} \pm 3.4 \times 10^{-2}$ |

Brier score, NLL) and outperforms the other models: specifically, we observe decreases of 1.1%, 0.9%, and 13% in ECE, Brier score, and NLL, respectively, as compared to UNet. We can conclude that, compared to the baseline model, both the EM and MMEF modules contribute to a higher segmentation reliability.

These findings are, to some extent, confirmed by Figure 6, which shows the calibration plots (also known as reliability diagrams) for the compared methods on the lymphoma dataset. Calibration plots are graphical representations showing how well the probabilistic predictions of a segmentation model are calibrated, i.e., how well confidence matches accuracy. In the left graph of Figure 6, we can see that the curve corresponding to UNet-Ensemble is closer to the diagonal than those of UNet and UNet-MC, which indicates better calibration. Looking at the right graph in Figure 6, we can see that the three DST-based models, ENN-UNet, RBF-UNet, and MMEF-UNet, have better calibration performance than the probabilistic ones, as shown by their calibration curves closer to the diagonal. Among them, MMEF-UNet shows the best calibration performance as ENN-UNet is slightly overconfident, while RBF-UNet is slightly underconfident.

*Segmentation accuracy.* The segmentation accuracy was measured by the Dice score, as shown in Table 1. Compared with the baseline model UNet, our proposal MMEF-UNet significantly increases segmentation performance, as shown by a 4.1% increase in the Dice score. Compared with the two DST-based deep evidential segmentation methods, MMEF-UNet has a higher Dice score (although the difference with ENN-UNet is not statistically significant). Figure 7 shows an example of visualized segmentation results obtained by UNet, ENN-UNet, RBF-UNet, and MMEF-UNet. We can see that UNet and RBF-UNet are more conservative (they correctly detect only a subset of the tumor voxels), while ENN-UNet is more radical (some of the voxels that do not belong to tumors are predicted as tumors). In contrast, the tumor regions predicted by MMEF-UNet better overlap the ground-truth tumor region, especially for the isolated lymphomas, which is also reflected by the promising Dice score value. These conclusions are consistent with the calibration trends displayed in Figure 6.

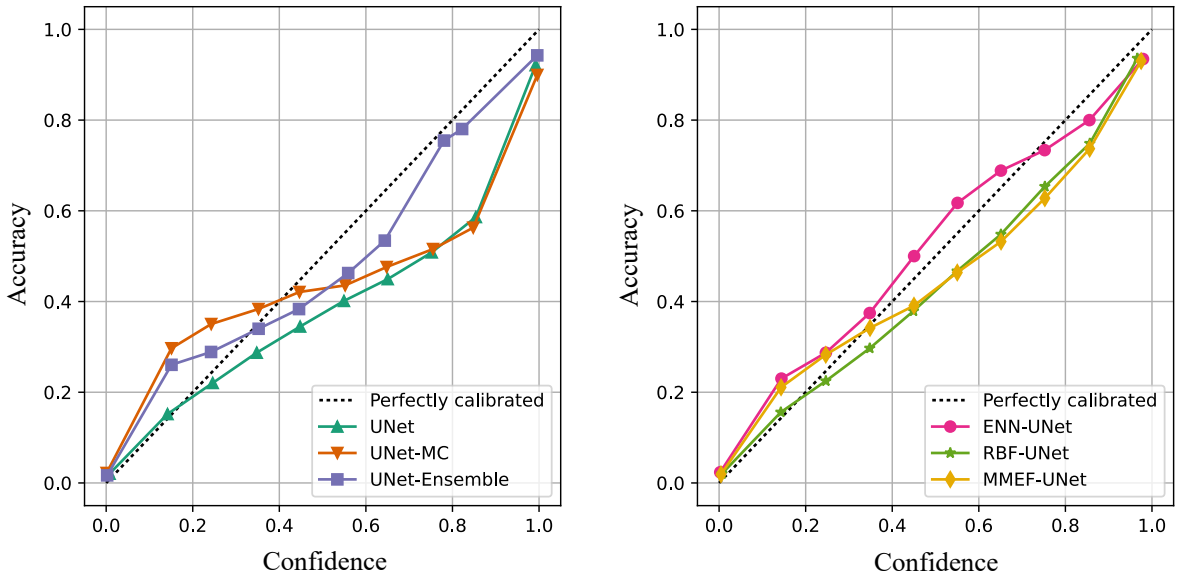


Figure 6: Calibration plots for probabilistic (left) and evidential (right) deep segmentation models.

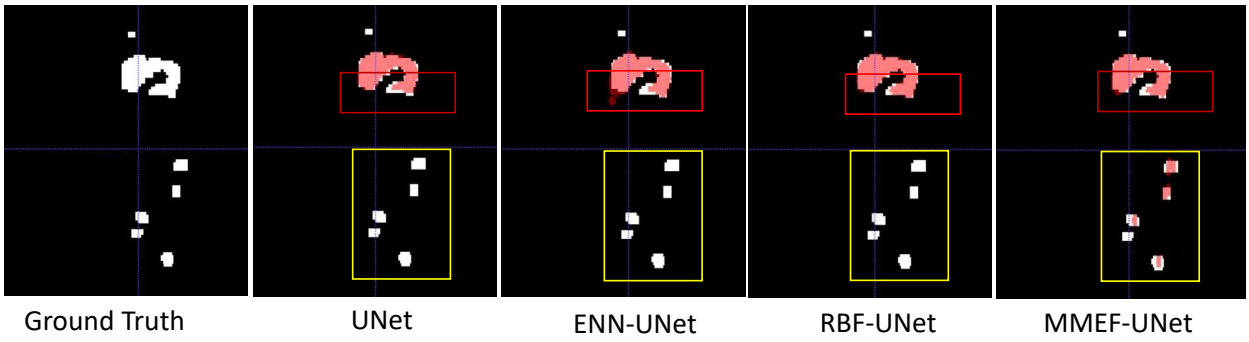


Figure 7: Examples of visualized segmentation results: from left to right, ground truth, and segmentation results obtained by UNet, ENN-UNet, RBF-UNet, and MMEF-UNet. The white and red regions represent, respectively, the ground truth and the segmentation result. Red and yellow boxes highlight the main differences in segmenting large and small isolated tumors.

Table 2: Estimated reliability coefficient $\beta_k^t$ (means and standard errors) after training for the background and lymphoma classes and the two modalities. Higher values correspond to greater contribution to the segmentation.

| $\beta_k^t$ | background | lymphomas |
|---|---|---|
| PET | $0.999 \pm 8.9 \times 10^{-3}$ | $0.996 \pm 4.5 \times 10^{-3}$ |
| CT | $0.863 \pm 1.8 \times 10^{-2}$ | $0.975 \pm 8.9 \times 10^{-3}$ |

Table 3: Segmentation quality and reliability of UNet and ENN-UNet applied to the lymphoma dataset with a single modality (CT or PET).

| Model | ECE↓ | Brier score ↓ | NLL↓ | Dice score ↑ |
|---|---|---|---|---|
| UNet (CT) | $0.133 \pm 4.9 \times 10^{-3}$ | $0.157 \pm 9.8 \times 10^{-3}$ | $0.571 \pm 3.6 \times 10^{-2}$ | $0.544 \pm 2.8 \times 10^{-2}$ |
| UNet (PET) | $0.060 \pm 4.0 \times 10^{-3}$ | $0.068 \pm 4.0 \times 10^{-3}$ | $0.348 \pm 8.2 \times 10^{-2}$ | $0.764 \pm 2.9 \times 10^{-2}$ |
| ENN-UNet (CT) | $0.131 \pm 8.5 \times 10^{-3}$ | $0.156 \pm 1.0 \times 10^{-2}$ | $0.521 \pm 3.2 \times 10^{-2}$ | $0.543 \pm 2.7 \times 10^{-2}$ |
| ENN-UNet (PET) | $0.050 \pm 4.9 \times 10^{-3}$ | $0.064 \pm 5.4 \times 10^{-3}$ | $0.195 \pm 2.2 \times 10^{-2}$ | $0.781 \pm 3.5 \times 10^{-2}$ |

*Analysis of reliability coefficients.* Table 2 reports the learned reliability coefficients. We can see that they are higher for the PET modality. This is consistent with domain knowledge, as mentioned in Section 4.1: PET images provide functional information about tumor activity and make it possible to identify active tumor sites, whereas CT images essentially provide detailed anatomical information (e.g., size, shape, and location) about lymph nodes and surrounding tissues and are used as a complement to PET images. This is also confirmed by the results presented in Table 3, showing that the performance of UNet and ENN-UNet with either CT alone or PET alone, the latter configuration yielding better results.

### *4.3. Segmentation results on the multi-MRI BraTS2021 dataset*

*Segmentation uncertainty.* For the BraTS2021 dataset, we tested the segmentation performance of our fusion framework with UNet as well as two alternative FE modules: nnUNet and nnFormer. The nnUNet model was reported to have the best performance in the BraTS2021 challenge [54] and nnFormer is now one of the state-of-the-art brain tumor segmentation models. The complete frameworks with nnUNet and nnFormer as a feature extractor are referred to, respectively, as MMEF-nnUNet and MMEF-nnFormer. We compared our results with three baseline models: UNet, nnUNet, and nnFormer, and three Monte Carlo-based uncertainty segmentation models: UNet-MC, nnUNet-MC, and nnFormer-MC. Since the results obtained in Section 4.2, as well as those reported in [37] have shown that ENN-UNet and RBF-UNet yield similar results, here we only compared the performance of the ENN-based models, i.e., ENN-UNet, ENN-nnUNet and ENN-nnFormer. Moreover, we did not test the performance of deep ensemble models because applying them to larger-scale datasets exceeds our computation resources.

As with the lymphoma dataset, we used the ECE, Brier score, and NLL metrics to assess segmentation uncertainty. The results with UNet, nnUNet and nnFormer in the FE module are presented, respectively, in Tables 4, 5 and 6. We can see that our fusion model consistently outperforms the baseline models with all three FE models and across all uncertainty evaluation metrics, although the differences are more significant when UNet is used as a feature extractor. Indeed, the fusion mechanism can be expected to have a smaller impact when information sources are more informative. Overall, MMEF-nnUNet achieves the highest segmentation reliability with the lowest ECE, Brier score, and NLL values, and MMEF-nnFormer yields the second-best results.

Table 4: Reliability measures (means and standard errors) for MMEF-UNet and the reference methods based on UNet on the BraTS2021 dataset. The best results are in bold and the second bests are underlined.

| Model | ECE↓ | Brier score ↓ | NLL↓ |
|---|---|---|---|
| UNet | $0.071 \pm 1.8 \times 10^{-3}$ | $0.141 \pm 1.8 \times 10^{-3}$ | $2.475 \pm 2.2 \times 10^{-3}$ |
| UNet-MC | $0.067 \pm 1.3 \times 10^{-3}$ | $0.135 \pm 4.5 \times 10^{-3}$ | $2.264 \pm 7.3 \times 10^{-2}$ |
| ENN-UNet | $\underline{0.065} \pm 1.3 \times 10^{-3}$ | $\underline{0.130} \pm 4.5 \times 10^{-3}$ | $\underline{2.250} \pm 3.6 \times 10^{-2}$ |
| MMEF-UNet (ours) | $\mathbf{0.060} \pm 1.3 \times 10^{-3}$ | $\mathbf{0.115} \pm 2.2 \times 10^{-3}$ | $\mathbf{2.189} \pm 4.1 \times 10^{-2}$ |

Table 5: Reliability measures (means and standard errors) for MMEF-nnUNet and the reference methods based on nnUNet on the BraTS2021 dataset. The best results are in bold. The best results are in bold, and the second-best results are underlined.

| Model | ECE↓ | Brier score ↓ | NLL↓ |
|---|---|---|---|
| nnUNet | $\underline{0.053} \pm 2.2 \times 10^{-3}$ | $0.109 \pm 4.5 \times 10^{-3}$ | $1.823 \pm 7.3 \times 10^{-2}$ |
| nnUNet-MC | $\mathbf{0.051} \pm 1.8 \times 10^{-3}$ | $\underline{0.107} \pm 4.5 \times 10^{-3}$ | $1.810 \pm 5.8 \times 10^{-2}$ |
| ENN-nnUNet | $\underline{0.053} \pm 1.8 \times 10^{-3}$ | $0.109 \pm 4.9 \times 10^{-3}$ | $\underline{1.804} \pm 8.2 \times 10^{-2}$ |
| MMEF-nnUNet (ours) | $\mathbf{0.051} \pm 1.3 \times 10^{-3}$ | $\mathbf{0.102} \pm 2.7 \times 10^{-3}$ | $\mathbf{1.748} \pm 5.9 \times 10^{-2}$ |

Table 6: Reliability measures (means and standard errors) for MMEF-nnFormer and the reference methods based on nnUNet on the BraTS2021 dataset. The best results are in bold and the second-best are underlined.

| Model | ECE↓ | Brier score ↓ | NLL↓ |
|---|---|---|---|
| nnFormer | $0.055 \pm 1.6 \times 10^{-3}$ | $0.111 \pm 3.2 \times 10^{-3}$ | $1.917 \pm 5.5 \times 10^{-2}$ |
| nnFormer-MC | $\underline{0.053} \pm 1.8 \times 10^{-3}$ | $\underline{0.107} \pm 3.6 \times 10^{-3}$ | $\mathbf{1.756} \pm 6.1 \times 10^{-2}$ |
| ENN-nnFormer | $0.055 \pm 1.4 \times 10^{-3}$ | $0.110 \pm 3.6 \times 10^{-3}$ | $1.907 \pm 7.0 \times 10^{-2}$ |
| MMEF-nnFormer (ours) | $\mathbf{0.052} \pm 0.6 \times 10^{-3}$ | $\mathbf{0.103} \pm 1.2 \times 10^{-3}$ | $\underline{1.787} \pm 2.2 \times 10^{-2}$ |

Table 7: Dice score (means and standard errors) for MMEF-UNet and the reference methods based on UNet on the BraTS2021 dataset. The best results are in bold and the second bests are underlined.

| Model | ET | TC | WT | Mean |
|---|---|---|---|---|
| UNet | $0.807 \pm 9.4 \times 10^{-3}$ | $0.825 \pm 8.5 \times 10^{-3}$ | $0.881 \pm 6.7 \times 10^{-3}$ | $0.837 \pm 7.2 \times 10^{-3}$ |
| UNet-MC | $0.812 \pm 1.3 \times 10^{-2}$ | $0.832 \pm 1.1 \times 10^{-2}$ | $0.886 \pm 6.3 \times 10^{-3}$ | $0.843 \pm 8.9 \times 10^{-3}$ |
| ENN-UNet | $\underline{0.810} \pm 1.3 \times 10^{-2}$ | $\underline{0.842} \pm 1.1 \times 10^{-2}$ | $\underline{0.896} \pm 5.4 \times 10^{-3}$ | $\underline{0.849} \pm 9.4 \times 10^{-3}$ |
| MMEF-UNet (ours) | $\mathbf{0.833} \pm 1.2 \times 10^{-2}$ | $\mathbf{0.854} \pm 7.2 \times 10^{-3}$ | $\mathbf{0.907} \pm 4.9 \times 10^{-3}$ | $\mathbf{0.864} \pm 5.8 \times 10^{-3}$ |

Table 8: Dice score (means and standard errors) for MMEF-nnUNet and the reference methods based on nnUNet on the BraTS2021 dataset. The best results are in bold and the second bests are underlined.

| Model | ET | TC | WT | Mean |
|---|---|---|---|---|
| nnUNet | $0.791 \pm 4.9 \times 10^{-3}$ | $0.850 \pm 5.8 \times 10^{-3}$ | $0.912 \pm 3.5 \times 10^{-3}$ | $0.851 \pm 4.4 \times 10^{-3}$ |
| nnUNet-MC | $0.802 \pm 4.4 \times 10^{-3}$ | $0.860 \pm 4.9 \times 10^{-3}$ | $\underline{0.916} \pm 4.4 \times 10^{-3}$ | $0.859 \pm 4.9 \times 10^{-3}$ |
| ENN-nnUNet | $\underline{0.807} \pm 9.8 \times 10^{-3}$ | $\underline{0.869} \pm 1.9 \times 10^{-2}$ | $0.915 \pm 5.4 \times 10^{-3}$ | $\underline{0.863} \pm 9.8 \times 10^{-3}$ |
| MMEF-nnUNet (ours) | $\mathbf{0.832} \pm 9.8 \times 10^{-3}$ | $\mathbf{0.873} \pm 2.6 \times 10^{-3}$ | $\mathbf{0.918} \pm 1.3 \times 10^{-3}$ | $\mathbf{0.875} \pm 4.4 \times 10^{-3}$ |

*Segmentation accuracy.* Segmentation accuracy was evaluated by the Dice score for the three overlapping regions, ET, TC, and WT, as well as by the mean Dice score. The results with UNet, nnUNet and nnFormer as feature extractors are reported, respectively, in Tables 7, 8 and 9. Again, we can see that our fusion strategy improves segmentation accuracy for all three FE models. Overall, the highest segmentation accuracy was achieved by MMEF-nnFormer, with an increase of 1.5 % in the mean Dice score compared with the second-best method, ENN-nnFormer.

We also report the Dice score for the segmentation of the three original tumor regions: ED, ET, and NRC/NET in Table 10. As we can see, the baseline nnFormer shows good performance for segmenting ED and ET, while it does not perform as well for segmenting NRC/NET. Indeed, the lack of clear contrast, the similar signal intensities to normal brain tissue, the infiltrative growth patterns, and the need for multi-modal data make the segmentation of NRC/NET inherently more challenging compared to ED and ET. When the MMEF-nnFormer approach was applied, the Dice scores for the ED, ET, and NRC/NET improved by 0.6%, 1.6%, and 6.5%, respectively. The substantial improvement in NRC/NET segmentation is particularly encouraging, as it demonstrates the effectiveness of the proposed fusion method for delineating fuzzy tumor boundaries and solving challenging segmentation tasks.

Figures 8 and 9 show two segmentation cases when using nnFormer as the feature extractor. Figure 8 shows an easy segmentation case where only one tumor type is present. Both the Flair and T1Gd images exhibit good segmentation performance with only a few mislabeled voxels. It is surprising to see that concatenating multimodal medical images as the input for nnFormer resulted in worse outcomes, with the most mislabeled voxels. This might

Table 9: Dice score (means and standard errors) for MMEF-nnFormer and the reference methods based on nnFormer on the BraTS2021 dataset. The best results are in bold and the second bests are underlined.

| Model | ET | TC | WT | Mean |
|---|---|---|---|---|
| nnFormer | $\underline{0.839}\pm3.8\times10^{-3}$ | $0.878\pm2.9\times10^{-3}$ | $\mathbf{0.915}\pm2.4\times10^{-3}$ | $0.877\pm1.7\times10^{-3}$ |
| nnFormer-MC | $0.837\pm3.7\times10^{-3}$ | $0.877\pm4.5\times10^{-3}$ | $\underline{0.914}\pm2.9\times10^{-3}$ | $0.876\pm2.3\times10^{-3}$ |
| ENN-nnFormer | $0.836\pm9.8\times10^{-3}$ | $\underline{0.882}\pm5.6\times10^{-2}$ | $\underline{0.914}\pm5.2\times10^{-3}$ | $\underline{0.878}\pm3.2\times10^{-3}$ |
| MMEF-nnFormer (ours) | $\mathbf{0.854}\pm7.5\times10^{-3}$ | $\mathbf{0.911}\pm5.4\times10^{-3}$ | $\underline{0.914}\pm2.3\times10^{-3}$ | $\mathbf{0.893}\pm4.8\times10^{-3}$ |

Table 10: Dice score (means and standard errors) for MMEF-nnFormer and nnFormer on the BraTS2021 dataset in segmenting detailed tumor class.

| Model | ED | ET | NRC/NET | Mean |
|---|---|---|---|---|
| nnFormer | $0.817\pm5.0\times10^{-3}$ | $0.839\pm3.8\times10^{-3}$ | $0.740\pm7.2\times10^{-3}$ | $0.799\pm2.5\times10^{-3}$ |
| MMEF-nnFormer (ours) | $\mathbf{0.823}\pm3.3\times10^{-3}$ | $\mathbf{0.855}\pm7.5\times10^{-3}$ | $\mathbf{0.805}\pm7.3\times10^{-3}$ | $\mathbf{0.828}\pm4.7\times10^{-3}$ |

be due to the hard fusion strategy of nnFormer, i.e., image concatenation, which cannot mitigate the impact of noisy information. Consequently, the fused results are sometimes not as good as those from single-modality inputs. The proposed MMEF-nnFormer approach achieves the best performance, with fewer mislabeled voxels compared to other methods. Figure 9 illustrates a challenging segmentation scenario involving a tumor with ED, ET, and NRC/NET components. We can remark that the FLAIR image alone provides sufficient information to accurately segment ED, which is consistent with domain knowledge. Overall, the MMEF-nnFormer model yields the best results in this case. This example illustrates the ability of our method to improve segmentation accuracy by appropriately weighting and combining information from different modalities.

*Analysis of reliability coefficients.* We first recall some clinical domain knowledge of MRI images in segmenting brain tumors:

1. T1Gd images are particularly useful for delineating tumor boundaries by making tumor regions hyperintense (bright);
2. FLAIR images help delineate tumor boundaries, assess tumor infiltration into surrounding brain tissue, and are particularly sensitive to peritumoral edema, which appears hyperintense (bright) on FLAIR sequences;
3. T2 images help delineate tumor extent, identify peritumoral edema, and assess the relationship between the tumor and surrounding brain structures;
4. Tumors typically appear hypointense (dark) on T1 images, while the contrast between the tumor and surrounding normal brain tissue may not always be sufficient for accurate segmentation.

Figure 10 shows the learned reliability coefficients $\beta_k^t$ estimated by MMEF-nnFormer, for the four modalities and the three tumor classes. It can be seen that the evidence from

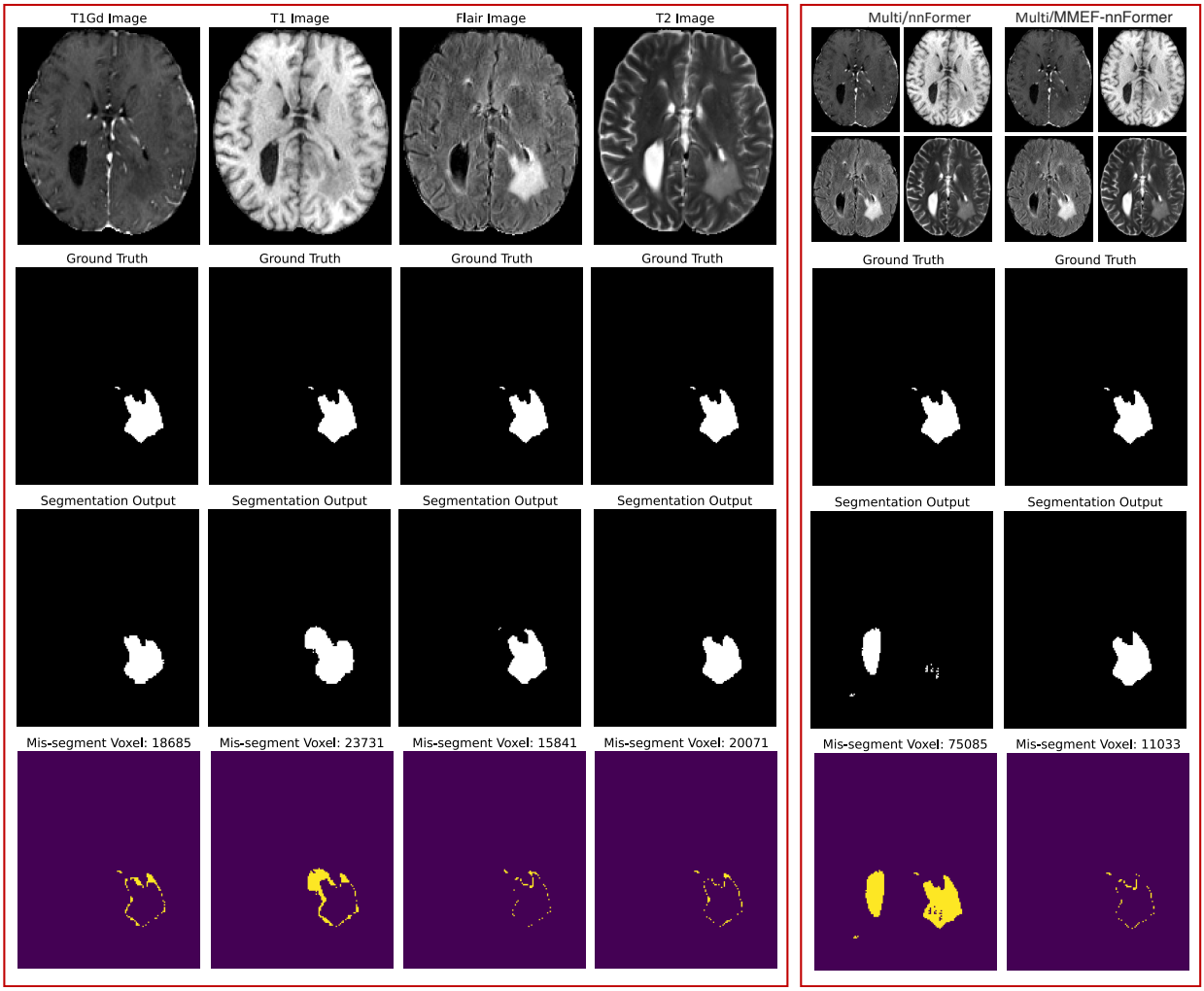


Figure 8: Examples of an easy tumor segmentation case. The first and second rows display the input modalities and the tumor ground truth, respectively. The third and last rows present the segmentation output and the mis-segmented voxels (highlighted in yellow). The left red block shows results from single-modality input using nnFormer, while the right red block compares results from multimodal input using nnFormer (left column) and MMEF-nnFormer (right column).

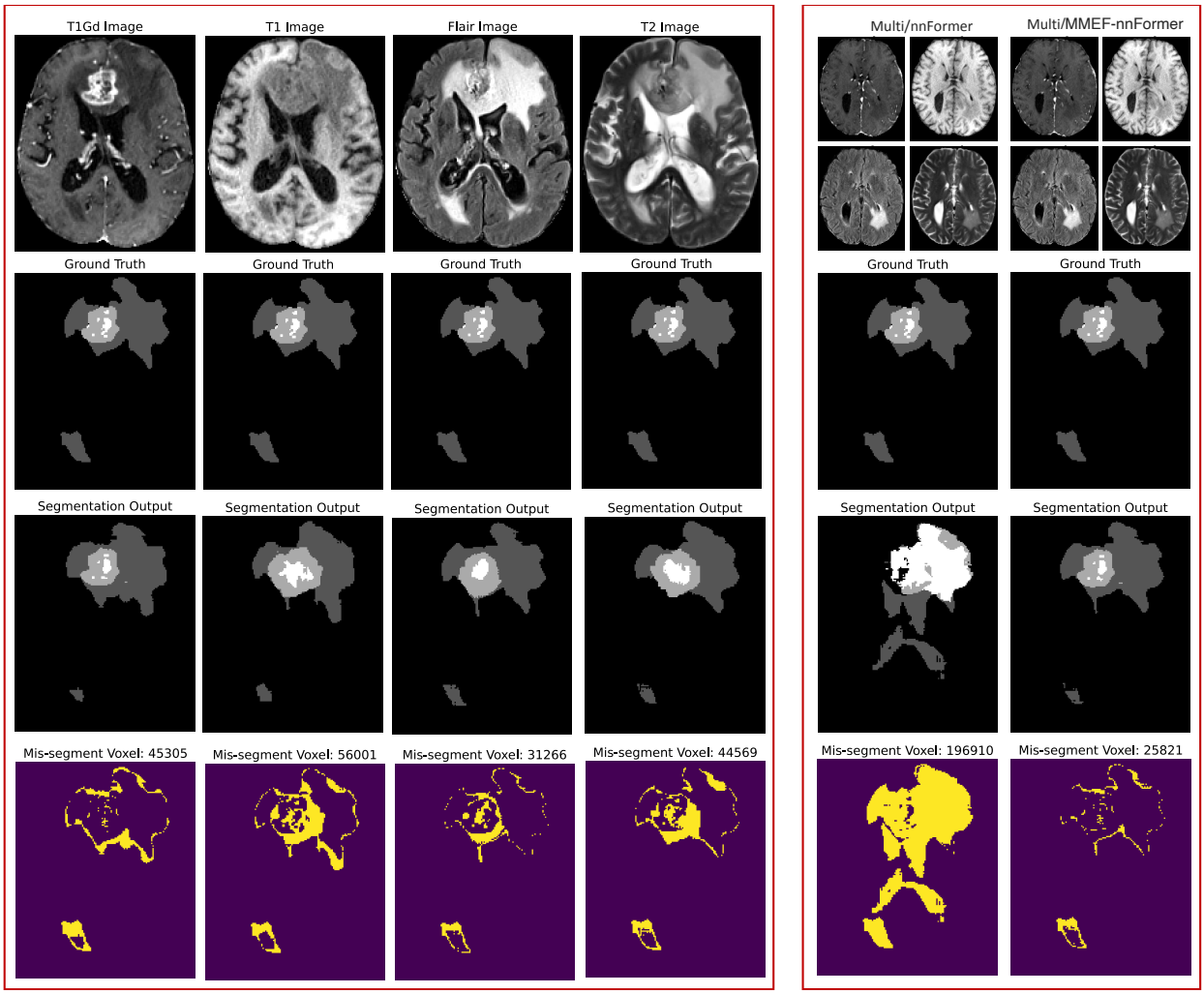


Figure 9: Examples of a challenging tumor segmentation case. The first and second rows display the input modalities and the tumor ground truth, respectively. The third and last rows present the segmentation output and the mis-segmented voxels (highlighted in yellow). The left red block shows results from single-modality input using nnFormer, while the right red block compares results from multimodal input using nnFormer (left column) and MMEF-nnFormer (right column).

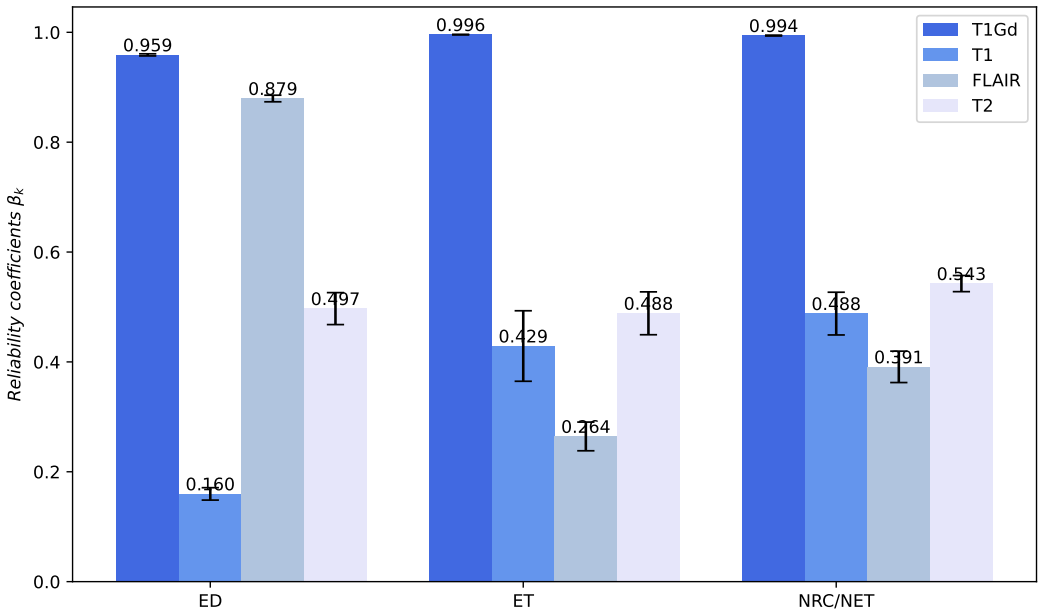


Figure 10: Estimated reliability coefficients $\beta_k$ (means and standard errors) after training of MMEF-nnFormer for classes ED, ET, and NRC/NET in the four modalities. Higher values correspond to greater contribution to the segmentation.

the T1Gd modality is reliable when the true class is ED, ET, or NRC/NET, with all the reliability values greater than 0.9. In contrast, the evidence from the FLAIR modality is more reliable for the ED class with a high-reliability coefficient of 0.879 against, respectively, 0.26 and 0.39 for ET and NRC/NET. The evidence from the T2 modality shows similar reliability in segmenting the three classes with a reliability coefficient of around 0.5. The evidence from the T1 modality is the least reliable one, compared with the other three MRI modalities. These results are consistent with domain knowledge about these modalities as reported in [5] and recalled at the beginning of this section, i.e., T1Gd images are useful for delineating tumor boundaries, FLAIR images are sensitive to ED, and T1 images are not sufficient for accurate tumor segmentation. This transparency and explainability of the decision-making process can be expected to enhance end-users' trust and can be seen as significant advantages of the proposed multimodal evidence fusion approach, as opposed to the "black box" nature of conventional deep learning segmentation models.

### *4.4. Discussion*

In the following, we provide some discussion about the generalizability, computational complexity, and limitations of our approach.

*Generalizability.* The main advantage of our framework is its ability to model and learn the reliability of each image modality, which can be crucial when dealing with diverse, potentially

noisy, or low-quality data. While multimodal medical image segmentation tasks are the focus of this paper, the proposed deep evidential fusion framework can be applied to a broader range of challenging medical tasks involving heterogeneous data sources. For instance, in medical tasks such as diagnosing dementia or Alzheimer's disease, various heterogeneous medical data are available [10]. These data can include lower-quality brain MRI images due to brain degeneration, textual data on disease history and progression, time-series data on blood-brain-barrier integrity, cerebrovascular information, and other relevant physiological measures. Traditional models struggle to effectively address this heterogeneous data within a single neural network [13], and recent work also proposed to address data heterogeneity with model ensembles and hard decision fusion [78]. Our deep evidential fusion framework could be well-suited to analyze such heterogeneous medical tasks. By learning the reliability coefficients for each of the modalities, our model can effectively combine the evidence from heterogeneous sources to reach a more informed and explainable diagnostic decision.

Beyond medical image processing, our approach could be applied to multimodal data fusion in other domains, such as reviewed in [44] and [8]. As examples of potential application domains where heterogeneous data need to be processed to make decisions, we can mention remote sensing and earth observations, in which light detection and ranging (LiDAR), synthetic aperture radar (SAR), and hyperspectral images need to be combined for, e.g., improved classification of objects. As noted in [44], SAR and LiDAR use different electromagnetic frequencies and thus interact differently with materials and surfaces. It would thus be beneficial to apply different discounting (reliability) coefficients to these sensor data depending on the nature of the objects of interest. This conjecture needs, of course, to be validated experimentally, which goes beyond the scope of this paper.

*Computational complexity.* Although the operations of DST have, in the worst case, exponential complexity, the mass functions computed in the EM module have only $K$ focal sets, where $K$ is the number of classes, and the contextual discounting operation computed in the MMEF is applied to the contour function, as explained in Section 3.1.3. Consequently, the number of operations performed in the EM and MMEF modules is only linear in the number of classes. More precisely, as shown in [16], each forward and backward propagation for one voxel and one modality in the EM module has complexity $O(I(H + K))$, where $I$ is the number of prototypes, $H$ is the number of features extracted by the FE module, and $K$ is the number of classes. In the MMEF module, the discounting of the $T$ mass functions for each voxel using (9) and their combination using (10) can be performed in $O(KT)$ operations, and the backward pass (gradient calculation) requires the same computational effort. Overall, the complexity of our model is, thus, similar to that performed in standard neural network architectures based on weighted sums. In terms of computing times, pre-training each of the FE modules with the nnFormer architecture took approximately one hour on our machine[6] for the BraTS2021 dataset, and training the whole system end-to-end took 2.3 hours. The total training time (6.4 hours) is slightly less than that of nnFormer with the four modalities (7.8 hours). As far as state-of-the-art uncertainty quantification techniques

[6]All models were trained on an NVIDIA A100-SXM4 graphics card with 40 GB GPU memory.

are concerned, Monte Carlo dropout does not significantly impact training time, while the deep ensemble method is notoriously time-consuming because it implies training several models. Overall, our framework based on DST and decision fusion is at least as efficient as alternative uncertainty quantification approaches.

*Limitations.* Our approach is based on combining high-level information extracted from each modality by the FE and EM modules in the form of mass functions. It, thus, has all the advantages and limitations of decision-level fusion approaches. On the plus side, it is highly modular and can still provide sensible results when only some of the modalities are available. This advantage is not crucial in multimodal image segmentation applications because all modalities are usually available, but it can matter in other potential applications such as remote sensing, as mentioned above. Another advantage of decision fusion is that the fusion process is simple and transparent, as already discussed in Sections 4.2 and 4.3. On the minus side, decision-level fusion is, at least in principle, suboptimal because it does not consider all input data globally: we can always construct a classification task in which a single classifier trained with a set of features will perform better than a combination of classifiers trained with each of the features. The good performances of our approach reported in Sections 4.2 and 4.3 show that this potential suboptimality is not an issue in the considered medical image segmentation applications, but it could be in other applications. Another limitation of our approach is that, to keep computations simple, we do not combine the whole discounted mass functions in the MMEF module, but only the contour functions. As a result, the output at each voxel is not a full mass function (with $2^K - 1$ focal sets), which prevents us from harnessing the full power of DST, such as some of the decision rules reviewed in [18]. This and other limitations will be addressed in future work.

## 5. Conclusion

We have proposed a deep decision-level fusion architecture for multi-modality medical image segmentation. In this approach, features are first extracted from each modality using a deep neural network such as UNet. An evidence-mapping module based on prototypes in feature space then computes a Dempster-Shafer mass function at each voxel. To account for the varying reliability of different information sources in different contexts, the mass functions are transformed using the contextual discounting operation before being combined by Dempster's rule. The whole framework is trained end-to-end by minimizing a loss function that quantifies prediction error both at the modality level and after fusion.

This model has been evaluated using two real-world datasets for lymphoma segmentation in PET-CT images and brain tumor segmentation in multi-MRI images. In both cases, our approach has been shown to allow for better uncertainty quantification and image segmentation as compared to various alternative schemes based on pixel-level fusion. In particular, as compared to UNet, nnUNet or nnFormer alone with a softmax layer, the introduction of the evidential mapping module (computing the mass functions) improves the results, and the decision-level fusion scheme with contextual discounting brings an additional improvement. Furthermore, the values found for the reliability coefficients are consistent with domain

knowledge, which suggests that these coefficients can provide useful insight into the fusion process.

This work can be extended in many directions. First, as discussed in Section 4.4, our DST-based fusion approach can be applied to a variety of learning tasks in which several sources of information must be combined. In the biomedical domain, it could be applied to fuse heterogeneous data such as signals, personal information, biomarkers, gene information, etc. In remote sensing, a potential application could be, e.g., the fusion of Lidar, SAR and hyperspectral data. References [44] and [8] mention many other applications in which multimodal data fusion plays an important role, including human-machine interaction, meteorological monitoring using weather radar and satellite data, or concrete structural monitoring through fusing ultrasonic, impact echo, capacitance, and radar. From a theoretical point of view, our approach could be extended in several directions. As mentioned in Section 4.4, we could combine not only the contour functions from the EM module but the whole mass functions, which would allow us to compute richer outputs that could be exploited within more sophisticated decision strategies such as partial classification [55], or further combined with other data. We could also consider other mass-function correction methods making it possible to account for more diverse meta-knowledge about information sources such as proposed, e.g. in [62], and/or other combination rules such as the cautions rule [17] or variants with learnable parameters as used in [64].

## Acknowledgements

This work was supported by the China Scholarship Council (No. 201808331005). It was carried out in the framework of the Labex MS2T, which was funded by the French Government, through the program "Investments for the Future" managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02).

## References

[1] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

[2] G. Andrade-Miranda, V. Jaouen, O. Tankyevych, C. C. Le Rest, D. Visvikis, and P.-H. Conze. Multimodal medical transformers: A meta-analysis for medical image segmentation in oncology. *Computerized Medical Imaging and Graphics*, 110:102308, 2023.

[3] M. Arif and G. Wang. Fast curvelet transform through genetic algorithm for multimodal medical image fusion. *Soft Computing*, 24(3):1815–1836, 2020.

[4] C. Asha, S. Lal, V. P. Gurupur, and P. P. Saxena. Multi-modal medical image fusion with adaptive weighted combination of NSST bands using chaotic grey wolf optimization. *IEEE Access*, 7:40782–40796, 2019.

[5] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

[6] M. Bauer, M. Van der Wilk, and C. E. Rasmussen. Understanding probabilistic sparse gaussian process approximations. *Advances in neural information processing systems*, 29, 2016.

[7] Y. Bengio, I. Goodfellow, and A. Courville. *Deep learning*, volume 1. MIT press Cambridge, MA, USA, 2017.

[8] R. Bokade, A. Navato, R. Ouyang, X. Jin, C.-A. Chou, S. Ostadabbas, and A. V. Mueller. A cross-disciplinary comparison of multimodal data fusion approaches and applications: Accelerating learning through trans-disciplinary information sharing. *Expert Systems with Applications*, 165:113885, 2021.

[9] G. W. Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[10] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. OConnell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[11] B. R. Cobb and P. P. Shenoy. On the plausibility transformation method for translating belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3):314–330, 2006.

[12] R. Cooke et al. *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press, 1991.

[13] Y. Dai, D. Qiu, Y. Wang, S. Dong, and H.-L. Wang. Research on computer-aided diagnosis of Alzheimers disease based on heterogeneous medical data fusion. *International Journal of Pattern Recognition and Artificial Intelligence*, 33(05):1957001, 2019.

[14] F. Delmotte, L. Dubois, and P. Borne. Context-dependent trust in data fusion within the possibility theory. In *1996 IEEE International Conference on Systems, Man and Cybernetics. Information Intelligence and Systems*, volume 1, pages 538–543. IEEE, 1996.

[15] A. P. Dempster. Upper and lower probability inferences based on a sample from a finite univariate population. *Biometrika*, 54(3-4):515–528, 1967.

[16] T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(2):131–150, 2000.

[17] T. Denœux. Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. *Artificial Intelligence*, 172:234–264, 2008.

[18] T. Denœux. Decision-making with belief functions: A review. *International Journal of Approximate Reasoning*, 109:87–110, 2019.

[19] T. Denœux. NN-EVCLUS: neural network-based evidential clustering. *Information Sciences*, 572:297–330, 2021.

[20] T. Denœux, D. Dubois, and H. Prade. Representations of uncertainty in artificial intelligence: Beyond probability and possibility. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial Intelligence Research*, volume 1, chapter 4, pages 119–150. Springer Verlag, 2020.

[21] T. Denœux, D. Dubois, and H. Prade. Representations of uncertainty in artificial intelligence: Probability and possibility. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial Intelligence Research*, volume 1, chapter 3, pages 69–117. Springer Verlag, 2020.

[22] T. Denœux, O. Kanjanatarakul, and S. Sriboonchitta. A new evidential k-nearest neighbor rule based on contextual discounting with partially supervised learning. *International Journal of Approximate Reasoning*, 113:287–302, 2019.

[23] J. Du, W. Li, B. Xiao, and Q. Nawaz. Union laplacian pyramid with multiple features for medical image fusion. *Neurocomputing*, 194:326–339, 2016.

[24] D. Dubois, W. Liu, J. Ma, and H. Prade. The basic principles of uncertain information fusion. an organised review of merging rules in different representation frameworks. *Information Fusion*, 32:12–39, 2016.

[25] Z. Elouedi, K. Mellouli, and P. Smets. Assessing sensor reliability for multisensor data fusion within the transferable belief model. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):782–787, 2004.

[26] S. Fabre, A. Appriou, and X. Briottet. Presentation and description of two classification methods using data fusion based on sensor management. *Information Fusion*, 2(1):49–71, 2001.

[27] P. H. Foo and G. W. Ng. High-level information fusion: An overview. *J. Adv. Inf. Fusion*, 8(1):33–72, 2013.

[28] J. Fu, W. Li, J. Du, and Y. Huang. A multiscale residual pyramid attention network for medical image fusion. *Biomedical Signal Processing and Control*, 66:102488, 2021.

[29] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[31] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[32] D. L. Hall, M. McNeese, J. Llinas, and T. Mullen. A framework for dynamic hard/soft fusion. In *2008 11th International Conference on Information Fusion*, pages 1–8. IEEE, 2008.

[33] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

[34] G. E. Hinton and D. Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, 1993.

[35] Q. Hu, S. Hu, and F. Zhang. Multi-modality medical image fusion based on separable dictionary learning and gabor filtering. *Signal Processing: Image Communication*, 83:115758, 2020.

[36] L. Huang, T. Denoeux, P. Vera, and S. Ruan. Evidence fusion with contextual discounting for multi-modality medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages 401–411. Springer, 2022.

[37] L. Huang, S. Ruan, P. Decazes, and T. Denœux. Lymphoma segmentation from 3d PET-CT images using a deep evidential network. *International Journal of Approximate Reasoning*, 149:39–60, 2022.

[38] L. Huang, S. Ruan, and T. Denoeux. Application of belief functions to medical image segmentation: A review. *Information fusion*, 91:737–756, 2023.

[39] L. Huang, S. Ruan, Y. Xing, and M. Feng. A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. *Medical Image Analysis*, 97:103223, 2024.

[40] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

[41] E. Kerfoot, J. Clough, I. Oksuz, J. Lee, A. P. King, and J. A. Schnabel. Left-ventricle quantification using residual u-net. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 371–380. Springer, 2018.

[42] S. U. R. Khan, M. Zhao, S. Asif, and X. Chen. Hybrid-net: A fusion of densenet169 and advanced machine learning classifiers for enhanced brain tumor diagnosis. *International Journal of Imaging Systems and Technology*, 34(1):e22975, 2024.

[43] M. Kim, D. K. Han, and H. Ko. Joint patch clustering-based dictionary learning for multimodal image fusion. *Information fusion*, 27:198–214, 2016.

[44] D. Lahat, T. Adali, and C. Jutten. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[45] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[46] H. Li and X.-J. Wu. Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, 103:102147, 2024.

[47] X. Li, F. Zhou, H. Tan, W. Zhang, and C. Zhao. Multimodal medical image fusion based on joint bilateral filter and local gradient energy. *Information Sciences*, 569:302–325, 2021.

[48] C. Lian, S. Ruan, T. Denœux, H. Li, and P. Vera. Joint tumor segmentation in PET-CT images using co-clustering and fusion based on belief functions. *IEEE Transactions on Image Processing*,

28(2):755–766, 2019.
[49] X. Liang, P. Hu, L. Zhang, J. Sun, and G. Yin. Mcfnet: Multi-layer concatenation fusion network for medical images fusion. *IEEE Sensors Journal*, 19(16):7107–7119, 2019.
[50] Y. Liu, X. Chen, J. Cheng, and H. Peng. A medical image fusion method based on convolutional neural networks. In *2017 20th international conference on information fusion (Fusion)*, pages 1–7. IEEE, 2017.
[51] Y. Liu, S. Liu, and Z. Wang. A general framework for image fusion based on multi-scale transform and sparse representation. *Information fusion*, 24:147–164, 2015.
[52] Y. Liu, F. Mu, Y. Shi, and X. Chen. Sf-net: A multi-task model for brain tumor segmentation in multimodal mri via image fusion. *IEEE Signal Processing Letters*, 29:1799–1803, 2022.
[53] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek. The design of simpleitk. *Frontiers in neuroinformatics*, 7:45, 2013.
[54] H. M. Luu and S.-H. Park. Extending nn-unet for brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 173–186, Cham, 2022. Springer International Publishing.
[55] L. Ma and T. Denœux. Partial classification in the belief function framework. *Knowledge-Based Systems*, 214:106742, 2021.
[56] D. J. MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.
[57] X. Meng, Q. Wei, L. Meng, J. Liu, Y. Wu, and W. Liu. Feature fusion and detection in Alzheimers disease using a novel genetic multi-kernel svm based on mri imaging and gene data. *Genes*, 13(5):837, 2022.
[58] D. Mercier, B. Quost, and T. Denœux. Refined modeling of sensor reliability in the belief function framework using contextual discounting. *Information fusion*, 9(2):246–258, 2008.
[59] C. Pei, K. Fan, and W. Wang. Two-scale multimodal medical image fusion based on guided filtering and sparse representation. *IEEE Access*, 8:140216–140233, 2020.
[60] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi. A robust volumetric transformer for accurate 3d tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages 162–172. Springer, 2022.
[61] F. Pichon, D. Dubois, and T. Denœux. Quality of information sources in information fusion. In É. Bossé and G. L. Rogova, editors, *Information Quality in Information Fusion and Decision Making*, pages 31–49. Springer International Publishing, Cham, 2019.
[62] F. Pichon, D. Mercier, E. Lefèvre, and F. Delmotte. Proposition and learning of some belief function contextual correction mechanisms. *International Journal of Approximate Reasoning*, 72:4–42, 2016.
[63] Z. Qu, Y. Li, and P. Tiwari. Qnmf: A quantum neural network based multimodal fusion system for intelligent diagnosis. *Information Fusion*, 100:101913, 2023.
[64] B. Quost, M.-H. Masson, and T. Denœux. Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules. *International Journal of Approximate Reasoning*, 52(3):353–374, 2011.
[65] G. L. Rogova and V. Nimier. Reliability in information fusion: literature survey. In *Proceedings of the seventh international conference on information fusion*, volume 2, pages 1158–1165, 2004.
[66] A.-J. Rousseau, T. Becker, J. Bertels, M. B. Blaschko, and D. Valkenborg. Post training uncertainty calibration of deep networks for medical image segmentation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1052–1056. IEEE, 2021.
[67] M. Safari, A. Fatemi, and L. Archambault. Medfusiongan: multimodal medical image fusion using an unsupervised deep generative adversarial network. *BMC Medical Imaging*, 23(1):203, 2023.
[68] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
[69] G. Shafer. *A mathematical theory of evidence*, volume 42. Princeton University Press, 1976.
[70] H. R. Shahdoosti and Z. Tabatabaei. Mri and pet/spect image fusion at feature level using ant colony

based segmentation. *Biomedical Signal Processing and Control*, 47:63–74, 2019.

[71] Y. Shi, C. Zu, P. Yang, S. Tan, H. Ren, X. Wu, J. Zhou, and Y. Wang. Uncertainty-weighted and relation-driven consistency training for semi-supervised head-and-neck tumor segmentation. *Knowledge-Based Systems*, 272:110598, 2023.

[72] R. Singh, M. Vatsa, and A. Noore. Multimodal medical image fusion using redundant discrete wavelet transform. In *2009 Seventh International Conference on Advances in Pattern Recognition*, pages 232–235. IEEE, 2009.

[73] P. Smets. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *International Journal of Approximate Reasoning*, 9:1–35, 1993.

[74] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.

[75] W. Tan, W. Thitøn, P. Xiang, and H. Zhou. Multi-modal brain image fusion based on multi-level edge-preserving filtering. *Biomedical Signal Processing and Control*, 64:102280, 2021.

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

[77] A. Tannaz, S. Mousa, D. Sabalan, and P. Masoud. Fusion of multimodal medical images using non-subsampled shearlet transform and particle swarm optimization. *Multidimensional Systems and Signal Processing*, 31:269–287, 2020.

[78] M. Tanveer, T. Goel, R. Sharma, A. Malik, I. Beheshti, J. Del Ser, P. Suganthan, and C. Lin. Ensemble deep learning for Alzheimers disease characterization and estimation. *Nature Mental Health*, pages 1–13, 2024.

[79] Z. Tong, P. Xu, and T. Denoeux. An evidential classifier based on Dempster-Shafer theory and deep learning. *Neurocomputing*, 450:275–293, 2021.

[80] Z. Tong, P. Xu, and T. Denœux. Evidential fully convolutional network for semantic segmentation. *Applied Intelligence*, 51:6376–6399, 2021.

[81] F. G. Veshki and S. A. Vorobyov. Coupled feature learning via structured convolutional sparse coding for multimodal image fusion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2500–2504. IEEE, 2022.

[82] Q. Wang, S. Li, H. Qin, and A. Hao. Robust multi-modal medical image fusion via anisotropic heat diffusion guided low-rank structural analysis. *Information fusion*, 26:103–121, 2015.

[83] Z. Wang, Z. Cui, and Y. Zhu. Multi-modal medical image fusion by laplacian pyramid and adaptive sparse representation. *Computers in Biology and Medicine*, 123:103823, 2020.

[84] Y. Weng, Y. Zhang, W. Wang, and T. Dening. Semi-supervised information fusion for medical image analysis: Recent progress and future perspectives. *Information Fusion*, page 102263, 2024.

[85] P. Xu, X. Zhu, and D. A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45(10):12113–12132, oct 2023.

[86] L. Yang, B. Guo, and W. Ni. Multimodality medical image fusion based on multiscale geometric analysis of contourlet transform. *Neurocomputing*, 72(1-3):203–211, 2008.

[87] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.

[88] J. Zhang, L. Jiao, W. Ma, F. Liu, X. Liu, L. Li, P. Chen, and S. Yang. Transformer based conditional gan for multimodal image fusion. *IEEE Transactions on Multimedia*, 25:8988–9001, 2023.

[89] S. Zhang, J. Zhang, B. Tian, T. Lukasiewicz, and Z. Xu. Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 83:102656, 2023.

[90] Y.-D. Zhang, Z. Dong, S.-H. Wang, X. Yu, X. Yao, Q. Zhou, H. Hu, M. Li, C. Jiménez-Mesa, J. Ramirez, et al. Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation. *Information Fusion*, 64:149–187, 2020.

[91] W. Zhao and H. Lu. Medical image fusion and denoising with alternating sequential filter and adaptive fractional order total variation. *IEEE Transactions on Instrumentation and Measurement*, 66(9):2283–2294, 2017.

[92] H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, and Y. Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 2023.

[93] K. Zou, X. Yuan, X. Shen, M. Wang, and H. Fu. Tbrats: Trusted brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 503–513. Springer, 2022.