Energy-induced Explicit quantification for Multi-modality MRI fusion

No Author Given

No Institute Given

1 Supplementary

1.1 Details of attention network

The details of attention-based fusion network in EHF are illustrated. In hierarchical aggregation, the same attention network is utilized to aggregate the multi-modality MRI. Here, the fusion on $tf4$, $c4$, $n4$, and $w4$ is taken for example. Through the attention-based fusion (Fig. 1), the representations of different modalities ($tf4$, $c4$, $n4$ and $w4$) are fused into $f4$. Firstly, the multi-modality representations with the size of $C \times D \times W \times H$ are added as a whole (A_f). Second, the fused representation is fed into local network to extract the local attention in channel dimension. The A_f is extracted by the convolution and normalization layers with the size of $C/r \times D \times W \times H$. Then the ReLU, convolution, and normalization layers are applied to achieves the local channel-wise attention ($C \times D \times W \times H$). Thirdly, the global attention is realized by a global-pooling and two convolution & normalization layers with the size of $C \times 1 \times 1 \times 1$. Fourth, the global and local attention are added and converted into sigmoid to achieve the attention in multi-modality MRI. Finally, the attention is applied to A_f for the attention-corrected fusion ($f4$). The above fusion process is applied to hierarchical multi-modality MRI representations.

1.2 The necessity of alignment

In proposed energy-regularized space alignment of E^2PA, the alignment of different modalities' representations is based on the representations of different modalities extracted by encoders belong to various spaces. To evaluate the necessity of alignment before the measurement of consistency in information flow, the representations of the aggregation and multi-modality MRI are calculated to find the corresponding vector spaces by QR decomposition. As shown in Fig. 2 (brain multi-modality MRI for the example), the purple dots and coordinate represent the aggregation representation and corresponding space ($[i_{fusion}, j_{fusion}]$), and the blue dots and coordinate represent the multi-modality MRI representation and and corresponding space ($[i_{tf}, j_{tf}]$) (taking T2f MRI as an example here). We take the purple (R^4_{fusion}) and blue (R^4_{tf}) dots outlined in red as an example for analysis. For the space of $[i_{fusion}, j_{fusion}]$, the coordinate of R^4_{fusion} is $[3, 4]$ and the coordinate of R^4_{tf} is $[0, 4]$. For the space of $[i_{tf}, j_{tf}]$, the coordinate of

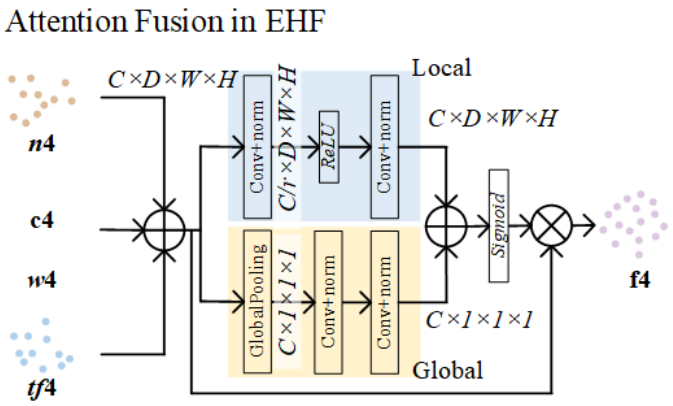

Fig. 1: The details of attention-based fusion network in EHF.

R^4_{fusion} is $[1.25, 1]$ and the coordinate of R^4_{tf} is $[1, 2]$. Without alignment, the distance between R^4_{fusion} and R^4_{tf} is calculated by:

$$Dis(R^4_{fusion}, R^4_{tf}) = ||Co(R^4_{tf}|[i_{tf}, j_{tf}]) - Co(R^4_{fusion}|[i_{fusion}, j_{fusion}])||^2 \quad (1)$$

where $Co(x|y)$ is the coordinate of x on the space of y. Under the above value, $Dis(R^4_{fusion}, R^4_{tf})$ is calculated ($|||[1, 2] - [3, 4]||^2 = \sqrt{8}$). However, the R^4_{tf} and R^4_{fusion} should be aligned to the same space to calculate the distance. The aligned distance calculation should be:

$$Dis(R^4_{fusion}, R^4_{tf}) = ||Co(R^4_{cine}|[i_{fusion}, j_{fusion}]) - Co(R^4_{fusion}|[i_{fusion}, j_{fusion}])||^2. \quad (2)$$

This equation represents that the distance between R^4_{tf} and R^4_{fusion} is calculated in the space of $[i_{fusion}, j_{fusion}]$. The distance can be easily calculated $|||[0, 4] - [3, 4]||^2 = 3$. From the different values of distance ($\sqrt{8}$vs.3), it can be easily found that the alignment of fusion representation and multi-modality MRI representations is necessary. In the energy-regularized vector space alignment of E^{2}PA, different modalities' representations are aligned with the aggregation in the energy calculation for the accurate measurement.

1.3 Experimental settings

To evaluate the performance of our E^{2}PA, we compare with various state-of-the-art methods in segmentation and classification. The segmentation methods include AWSNet, MyoPS-Net, NestedFormer, HyperDense, MAML, and MMSNet. The classification methods include TransMed, MRNet, ELNet, and MRPyrNet. In the model training, the data augmentation is adopted as the pre-processing,

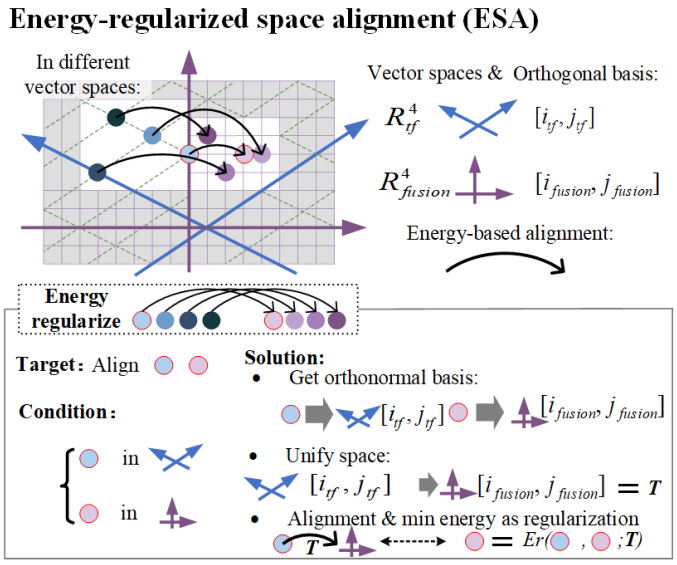

Fig. 2: The alignment process of ESA.

including rotation, mirror, resize, and Gaussian noise. Our E^{2}PA is optimized by Adam with the learning rate of 1×10^{-4} for 300 epochs on segmentation and classification tasks separately. The batchsize in segmentation task is 1, and 5 in classification task. In our data splitting strategy, the three-fold cross-validation is adopted to achieve the best model for direct testing. In the segmentation task, the combination of dice loss and cross-entropy loss is adopted as the target loss function. In the classification task, the cross-entropy loss is the target loss function. For the compared methods above, we follows the claimed in the papers to optimize the model on each multi-modality MRI dataset for the best model.

The BraTS dataset contains four modalities: native (T1n) and b) post-contrast T1-weighted (T1c), c) T2-weighted (T2w), and d) T2 Fluid Attenuated Inversion Recovery (T2f) volumes, and were acquired with different clinical protocols and various scanners from multiple data contributing institutions. The sub-regions considered for evaluation are the "enhancing tumor" (ET), the "tumor core" (TC), and the "whole tumor" (WT).

The MyoPS dataset contains balanced steady-state free precession (cine) sequence, late gadolinium enhancement (LGE) sequence, T2-weighted sequence (T2). The segmentation target contains the left ventricle (LV), right ventricle (RV), myocardium (MYO), scar, and edema.

The MRNet dataset conducts examinations by GE scanners (GE Discovery, GE Healthcare, Waukesha, WI) with standard knee MRI coil and a routine non-contrast knee MRI protocol that included the following sequences: coronal T1

weighted, coronal T2 with fat saturation, sagittal proton density (PD) weighted, sagittal T2 with fat saturation, and axial PD weighted with fat saturation.

1.4 More visual result in different diseases

We visualize more results on BraTS and MyoPS (Fig. 3). From the visual results in Fig. 3 and article, our E^{2}PA obtains the best performance. This indicates the superiority of E^{2}PA for various diseases' multi-modality aggregation. For the scar and edema region in cardiac MRI, other methods can not aggregate the information from LGE and T2 effectively. Our E^{2}PA aggregates the information of scar and edema by the explicit quantification from multi-modality MRI. The best performance also indicates that the superiority of aggregation strategy.

1.5 The visual inter-dependencies.

We evaluate the inter-dependencies in different diseases. In the article, we evaluate the inter-dependencies on MyoPS. Here, we visualize the feature maps of BraTS to evaluate the inter-dependencies on multi-modality MRI ($T1c$, $T1n$, $T2n$, $T2w$). As shown in Fig. 4, different modalities provides different information according to the inter-dependencies during aggregation. It can be found that the tumor core (TC), enhancing tumor (ET) and whole tumor (WT) are related to different modalities. The information of TC comes from $T1c$, $T2f$, $T2w$. Th $T1c$ and $T2f$ provide the information of ET. The left regions are supported by $T1n$ and $T2w$. This also proves that our E^{2}PA uncovers the inter-dependencies among different modalities.

1.6 Clarification of the superiority over baselines.

To further clarify the superiority over baselines, we also compare E^{2}PA with the simple weighting and other strategies in compare methods on the BraTS dataset (Tab. 1). It can be found that our E^{2}PA is superior to the existing image fusion methods.

Table 1: The comparison with the existing image fusion baselines on multi-modality MRI fusion.

Method	TC	ET	WT	AVG
Pixel fusion [6]	77.0	74.9	81.2	77.7
Feature fusion [1]	79.3	74.5	81.1	78.3
Decision fusion [3]	78.9	77.6	83.8	80.1
Ours	**91.0**	**87.3**	**93.5**	**90.6**

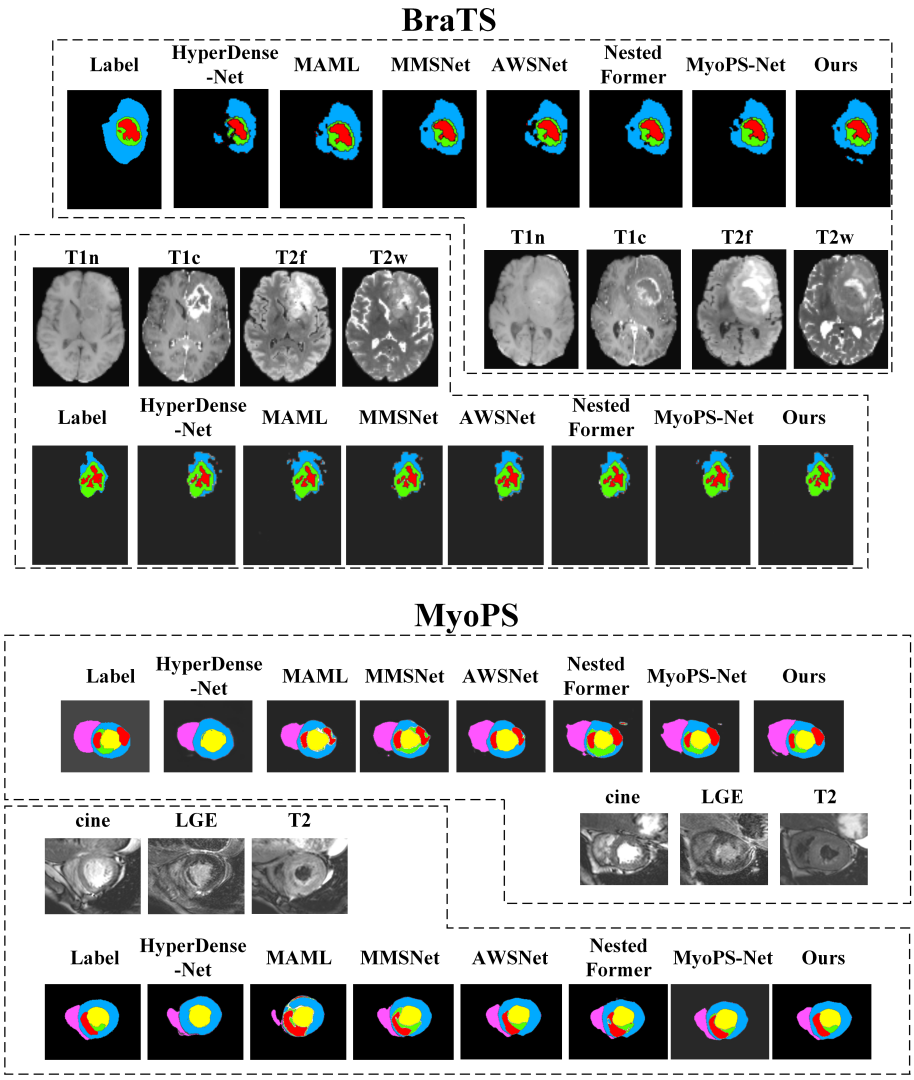

Fig. 3: More visual results on BraTS and MyoPS indicate the superiority of our E^{2}PA.

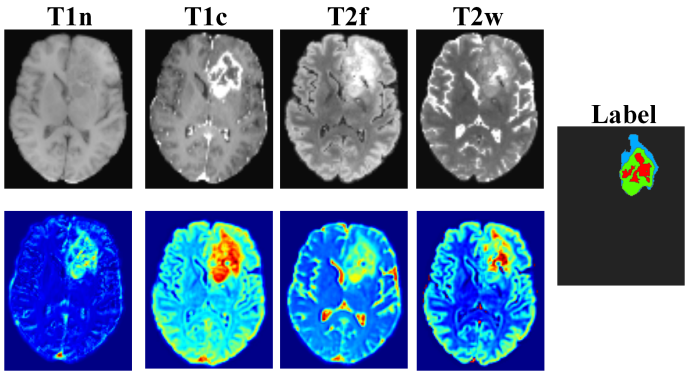

Fig. 4: The visual inter-dependencies of multi-modality MRI in BraTS.

1.7 Comparison with other energy model.

Trough the survey [5], there is currently no energy-based model that can be directly applied to multi-modal MRI fusion. To realize the evaluation, we modified the existing energy-based models [2,4] to replace the energy functions ($\mathcal{L}_{i-d}$ and $\mathcal{L}_r$) in our E^{2}PA (Tab. 2). It indicates that E^{2}PA is superior to other energy based models on multi-modality MRI fusion.

Table 2: The comparison with other energy based models replacing the energy functions ($\mathcal{L}_{i-d}$ and $\mathcal{L}_r$) in our E^{2}PA.

Method	TC	ET	WT	AVG
E1 [2]	82.8	84.4	86.0	84.4
E2 [4]	87.7	86.1	88.1	87.3
Ours	**91.0**	**87.3**	**93.5**	**90.6**

References

1. Bandara, W.G.C., Patel, V.M.: Hypertransformer: A textural and spectral feature fusion transformer for pansharpening. In: CVPR (2022) 4
2. Du, Y., Mordatch, I.: Implicit generation and modeling with energy based models. Advances in Neural Information Processing Systems **32** (2019) 6
3. Hermessi, H., Mourali, O., Zagrouba, E.: Multimodal medical image fusion review: Theoretical background and recent advances. Signal Processing (2021) 4
4. Isack, H., Boykov, Y.: Energy-based geometric multi-model fitting. International journal of computer vision **97**(2), 123–147 (2012) 6
5. Pang, B., Han, T., Nijkamp, E., Zhu, S.C., Wu, Y.N.: Learning latent space energy-based prior model. Advances in Neural Information Processing Systems **33**, 21994–22008 (2020) 6
6. Raudonis, V., Paulauskaite-Taraseviciene, A., Sutiene, K.: Fast multi-focus fusion based on deep learning for early-stage embryo image enhancement. Sensors (2021) 4