



# SFusion: Self-attention Based N-to-One Multimodal Fusion Block

Zecheng Liu<sup>1</sup>, Jia Wei<sup>1(✉)</sup>, Rui Li<sup>2</sup>, and Jianlong Zhou<sup>3</sup>

<sup>1</sup> School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

`msaiyan@mail.scut.edu.cn, csjwei@scut.edu.cn`

<sup>2</sup> Golisano College of Computing and Information Sciences, Rochester Institute of Technology, Rochester, NY, USA

`rxlics@rit.edu`

<sup>3</sup> Data Science Institute, University of Technology Sydney, Ultimo, NSW 2007, Australia

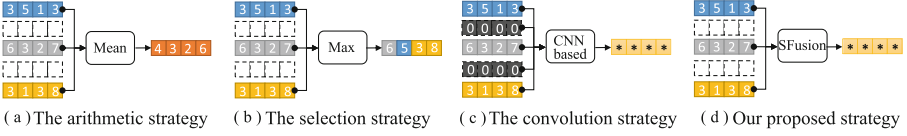
`jianlong.zhou@uts.edu.au`

**Abstract.** People perceive the world with different senses, such as sight, hearing, smell, and touch. Processing and fusing information from multiple modalities enables Artificial Intelligence to understand the world around us more easily. However, when there are missing modalities, the number of available modalities is different in diverse situations, which leads to an N-to-One fusion problem. To solve this problem, we propose a self-attention based fusion block called SFusion. Different from preset formulations or convolution based methods, the proposed block automatically learns to fuse available modalities without synthesizing or zero-padding missing ones. Specifically, the feature representations extracted from upstream processing model are projected as tokens and fed into self-attention module to generate latent multimodal correlations. Then, a modal attention mechanism is introduced to build a shared representation, which can be applied by the downstream decision model. The proposed SFusion can be easily integrated into existing multimodal analysis networks. In this work, we apply SFusion to different backbone networks for human activity recognition and brain tumor segmentation tasks. Extensive experimental results show that the SFusion block achieves better performance than the competing fusion strategies. Our code is available at <https://github.com/scut-cszcl/SFusion>.

**Keywords:** Multimodal fusion · Missing modalities · Brain tumor segmentation · Human activity recognition

## 1 Introduction

People perceive the world with signals from different modalities, which often carry complementary information about varying aspects of an object or event of interest. Therefore, collecting and utilizing multimodal information is crucial for Artificial Intelligence to understand the world around us. Data collected



**Fig. 1.** Fusion strategies. \* denotes the value is automatically learned.

from various sensors (e.g., microphones, cameras, motion controllers) are used to identify human activity [4]. Moreover, multimodal medical images obtained from different scanning protocols (e.g., Computed Tomography, Magnetic Resonance Imaging) are employed for disease diagnosis [12]. Satisfactory performances have been achieved with these multimodal data.

In practical application, however, modality missing is a common scenario. Wirelessly connected sensors may occasionally disconnect and temporarily be unable to send any data [3]. Medical images may be missing due to artifacts and diverse patient conditions [11]. In these unexpected situations, any combinatorial subset of available modalities can be given as input. To handle this, one intuitive solution is to train a dedicated model on all possible subsets of available modalities [6, 14, 23]. However, these methods are ineffective and time-consuming. Another way is to predict missing modalities and perform with the completed modalities [20]. But, these approaches also require additional prediction networks for each missing situation, and the quality of the recovered data directly affects the performance, especially when there are only a few available modalities. Recently, fusing the available modalities into a shared representation received wide attention. However, it is particularly challenging due to the varying number of input modalities, which results in the N-to-One fusion problem.

Currently, existing fusion strategies to tackle this challenge can be broadly grouped into three categories: the arithmetic strategy, the selection strategy and the convolution strategy. As shown in Fig. 1(a), in the arithmetic strategy, feature representations of available modalities are merged by an arithmetic function, such as averaging, computing the first and second moments or other designed formulas [10, 13, 17]. For the selection strategy, as shown in Fig. 1(b), each value of fused representation is selected from the values at the corresponding position of the inputs. The selection rule can be defined as max, min or probability-based [2, 8, 19]. Although the above two fusion strategies are easily scalable to various data missing situations, their fusion operation is hard-coded. All available modalities contribute equally and their latent correlations are neglected. Unlike hard-coding the fusion operation, in the convolution strategy, the convolutional fusion network automatically learns how to fuse these feature representations, which is beneficial to exploiting the correlation between multiple modalities. However, as shown in Fig. 1(c), this fusion strategy needs a constant number of data to meet the requirements of the input channels in the convolutional network. Therefore, it has to simulate missing data by crudely zero-padding or replacing it with similar modalities, which inevitably introduces a bias in computation and causes performance degradation [5, 18, 25].

Transformer has achieved success in the field of computer vision, demonstrating that self-attention mechanism has the ability to capture the latent correlation of image tokens. However, no work has explored the effectiveness of self-attention mechanism on the N-to-One fusion, where N is variable during training, rather than fixed. Furthermore, the calculation of self-attention does not require a fixed number of tokens as input, which represents a potential for handling missing data. Therefore, we propose a self-attention based fusion block (SFusion) to tackle the problems of the above fusion strategies. As shown in Fig. 1(d), SFusion can handle any number of input data instead of fixing its number. In addition, SFusion is a learning-based fusion strategy that consists of two components: the correlation extraction (CE) module and the modal attention (MA) module. In the CE module, feature representations extracted from available modalities are projected as tokens and fed into the self-attention layers to learn multimodal correlations. Based on these correlations, a modal softmax function is proposed to generate weight maps in the MA module. Finally, it builds a shared feature representation by fusing the varying inputs with the weight maps.

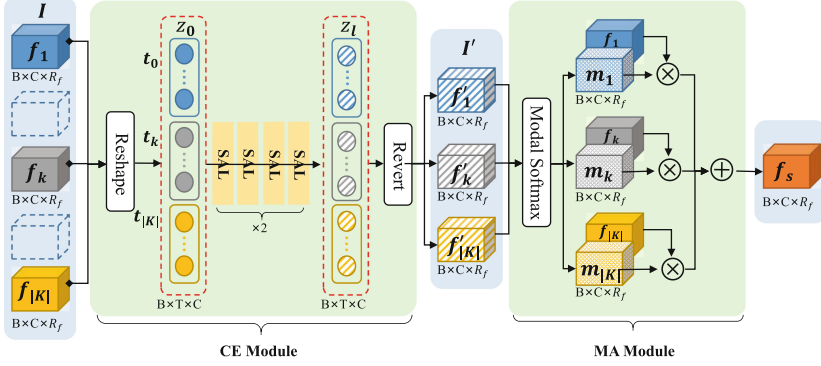
The contributions of this work are:

- We propose SFusion, which is a data-dependent fusion strategy without impersonating missing modalities. It can learn the latent correlations between different modalities and builds a shared representation adaptively.
- The SFusion is not limited to specific deep learning architectures. It takes inputs from any kind of upstream processing model and serves as the input of the downstream decision model, which enables applying the SFusion to various backbone networks for different tasks.
- We provide qualitative and quantitative performance evaluations on activity recognition with the SHL [22] dataset and brain tumor segmentation with the BraTS2020 [1] dataset. The results show the superiority of SFusion over competing fusion strategies.

## 2 Methodology

### 2.1 Method Overview

For multiple modalities, let  $k \in K \subseteq \{1, 2, \dots, S\}$  index a specific modality, within the available modality set of  $K$ , where  $S$  is the number of all possible modalities. Given an input  $f_k \in \mathbb{R}^{B \times C \times R_f}$ ,  $B$  and  $C$  denote the batch size and the number of channels, respectively.  $R_f$  represents the shape of feature representation extracted from the  $k$ -th modality of a sample data, which can be 1D (L), 2D (H×W), 3D (D×H×W) or higher-dimensional. In addition,  $I = \{f_k | k \in K\}$  denotes the input set of feature representations from all the available modalities. Our goal is to learn a fusion function  $F$  that can project  $I$  into a shared feature representation  $f_s$ , denoted as  $F(I) \rightarrow f_s$ . To achieve the goal, we design an N-to-One fusion block, SFusion. The architecture is shown in Fig. 2, which consists of two modules: correlation extraction (CE) module and modal attention (MA) module.



**Fig. 2.** The illustration of SFusion.  $R_f$ : L or  $H \times W$  or  $D \times H \times W$  (shape of feature representation);  $T = L \cdot |K|$  or  $H \cdot W \cdot |K|$  or  $D \cdot H \cdot W \cdot |K|$  (number of tokens).

## 2.2 Correlation Extraction

Given the feature representation  $f_k \in \mathbb{R}^{B \times C \times R_f}$ , we first flatten the  $R_f$  dimensions of  $f_k$  into one dimension and get a  $B \times C \times R$  feature representation, where  $R = L$  (1D),  $R = H \times W$  (2D),  $R = D \times H \times W$  (3D), etc. It can be viewed as  $B \times R$   $C$ -dimensional tokens  $t_k$ . Then, we obtain the concatenation of all the tokens  $z_0 \in \mathbb{R}^{B \times T \times C}$ , where  $T = R \times |K|$ , and  $|K|$  denotes the number of available modalities.

Given  $z_0$ , the stack of eight self-attention layers (SAL) are introduced to learn the latent multimodal correlations. Each layer includes a multi-head attention (MHA) block and a fully connected feed-forward network (FFN) [21]. Layer normalization (LN) is applied before every block. The outputs of the  $x$ -th ( $x \in [1, 2, \dots, 8]$ ) layer can be describe as:

$$z'_x = MHA(LN(z_{x-1})) + z_{x-1} \quad (1)$$

$$z_x = FFN(LN(z'_x)) + z'_x \quad (2)$$

Therefore, we get  $z_l \in \mathbb{R}^{B \times T \times C}$ , which is the last SAL output. By reverting  $z_l$  to the size of  $|K| \times B \times C \times R_f$ , we obtain the output  $I' = \{f'_k | k \in K\}$  of CE as:

$$I' = split(r(z_l)) \quad (3)$$

where  $r(\cdot)$  and  $split(\cdot)$  are the reshape and split operations, and  $I'$  is the set of calculated feature representations  $f'_k \in \mathbb{R}^{B \times C \times R_f}$  which contains multimodal correlations and has the same size as the original input  $f_k$ .

## 2.3 Modal Attention

Given the calculated feature representations set  $I'$ , the weight map  $m_k$  is generated with the modal attention mechanism. Feature representations extracted

from different modalities are expected to have different weights for fusion at the voxel level. Therefore, we introduce a modal-wise and voxel-level softmax function to generate the weight maps from  $I'$ , as shown in Fig. 3.

We denote the  $i$ -th voxel of  $f'_k$  and  $m_k$  as  $v_k^i$  and  $m_k^i$ , respectively.  $e$  is the natural logarithm. The value of weight map  $m_k$  can be defined as:

$$m_k^i = \frac{e^{v_k^i}}{\sum_{j \in K} e^{v_j^i}} \quad (4)$$

By element-wise multiplying input feature map  $f_k$  with the corresponding weight map  $m_k$  and summing all the modalities, we can obtain a fused feature map  $f_s$  as:

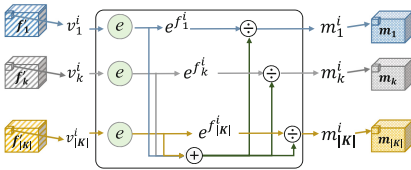
$$f_s = \sum_{k \in K} f_k \cdot m_k \quad (5)$$

Since the sum of  $m_1^i, \dots, m_{|K|}^i$  is 1, the value range of fused feature representation  $f_s$  remains stable to improve the robustness for variable input modalities. Moreover, the relative sizes of  $v_1^i, \dots, v_{|K|}^i$  (contain the latent multi-modal correlations learned from the CE module) are retained in the corresponding weights. In particular, when only one modality is available, all the values of the weight map are 1, which means  $f_s = f_k$  ( $k \in K, |K| = 1$ ). In this case, the input feature representation remains unchanged. It enables the backbone network (the upstream processing model and the downstream decision model) to enhance its capability to encode and decode information from different modalities rather than relying on a particular one. It is crucial for variable multimodal data analysis.

### 3 Experiments and Results

#### 3.1 Datasets

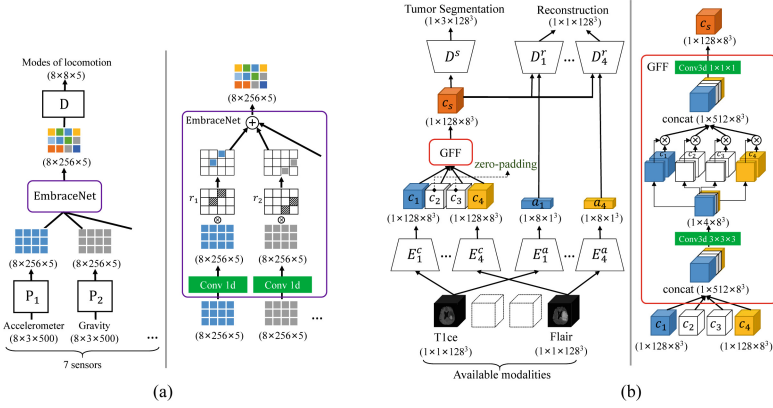
**SHL2019.** The SHL (Sussex-Huawei Locomotion) Challenge 2019 [22] dataset provides data from seven sensors of a smartphone to recognize eight modes of locomotion and transportation (activities), including *still*, *walking*, *run*, *bike*, *car*, *bus*, *train*, and *subway*. The sensor data are collected from smartphones of a



**Fig. 3.** The illustration of modal attention mechanism.

**Table 1.** Evaluation on SHL2019. *w/o* means without. † denotes results from [9].

Accuracy(%)	Bag	Hips	Torso	Hand	All
Early†	–	–	–	–	46.73
Intermediate†	–	–	–	–	63.87
Late†	–	–	–	–	63.85
Confidence†[7]	–	–	–	–	63.60
EmbraceNet† [9]	63.68	67.98	81.58	47.63	65.22
SFusion	<b>67.41</b>	<b>68.91</b>	<b>85.22</b>	<b>48.35</b>	<b>67.47</b>
SFusion <i>w/o</i> CE	56.82	63.14	74.69	46.70	60.33
SFusion <i>w/o</i> MA	65.01	67.95	83.49	47.52	65.99



**Fig. 4.** (a) Activity recognition with EmbraceNet; (b) Bran tumor segmentation with GFF. ( $B \times C \times R_f$ ) is given, where  $B$ ,  $C$  and  $R_f$  denotes the batch size, channels and data shape, respectively.

person with four locations, including the bag, trousers front pocket, breast pocket and hand. Each location is called “Bag”, “Hips”, “Torso”, and “Hand”, respectively. Data acquired from the locations except the “Hand” are given in the train subset, while the validation subset provides the data of all four locations. In the test subset, only unlabeled “Hand” location data are available.

**BraTS2020.** The BraTS2020 [1] dataset provide four modality scans: T1ce, T1, T2, FLAIR for brain tumor segmentation. It contains 369 subjects. To better represent the clinical application tasks, there are three mutually inclusive tumor regions: the enhancing tumor (ET), the tumor core (TC), and the whole tumor (WT) [1]. We select 70% data as training data, while 10% and 20% as validation and test data respectively. To prevent overfitting, two data augmentation techniques (randomly flip the axes and rotate with a random angle in  $[-10^\circ, 10^\circ]$ ) are applied during training. We apply z-score normalization [15] to the volumes individually and randomly crop  $128 \times 128 \times 128$  patches as inputs to the networks.

### 3.2 Baseline Methods

**EmbraceNet.** In the experiments on activity recognition, we compare SFusion with EmbraceNet [9], which employs a selection strategy (shown in Fig. 1 (b)) by generating feature masks ( $r_1, r_2, \dots, r_7$ ) with the rule of giving equal chances to all available modalities during each value selection. For a fair comparison, as shown in Fig. 4 (a), we adopt the same processing (P) and decision (D) model as used in [9]. We obtain the performance of our fusion strategy by replacing *EmbraceNet* with *SFusion*. Following [9] setting, the batch size is set to 8. A cross-entropy loss and the Adam optimization method [16] with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  are employed. The learning rate is initially set to  $1 \times 10^{-4}$  and reduced by a factor of 2 at every  $1 \times 10^5$  steps. A total of  $5 \times 10^5$  training steps are executed.

**GFF.** In the experiments on brain tumor segmentation, we compare SFusion with a gated feature fusion block (GFF) [5], which belongs to the convolution strategy (shown in Fig. 1(c)). As shown in Fig. 4 (b), a feature disentanglement architecture is employed. Multimodal medical images are decomposed into the modality-invariant content and the modality-specific appearance code by encoders  $E^c$  and  $E^a$ , respectively. The content codes (e.g.,  $c_2$  and  $c_3$ , shown in Fig. 4 (b)) of missing modalities are simulated with zero values. Then, all content codes are fused into a shared representation  $c_s$  by GFF. Given  $c_s$ , the tumor segmentation results are generated by the decoder  $D^s$ . For a fair comparison, we adopt the same encoders ( $E_i^c$  and  $E_i^a$ ) and decoders ( $D^s$  and  $D_i^t$ ) as used in [5]. *We obtain the performance of our fusion strategy by replacing GFF with SFusion and removing the zero-padding operation.* The training max\_epoch is set to 200. Following [5] setting, the batch size is set to 1. Adam [16] is utilized with a learning rate of  $1 \times 10^{-4}$  and progressively multiplies it by  $(1 - \text{epoch} / \text{max\_epoch})^{0.9}$ . Losses of  $\mathcal{L}_{KL}$ ,  $\mathcal{L}_{rec}$  and  $\mathcal{L}_{seg}$  are employed as [5]. During training, to simulate real missing modalities scenarios, each training patient's data is fixed to one of 15 possible missing cases. For a comprehensive evaluation, we test the performance of all 15 cases for each test patient.

Our implementations are on an NVIDIA RTX 3090(24G) with PyTorch 1.8.1.

### 3.3 Results

**Activity recognition.** We compare SFusion with the EmbraceNet [9] on SHL2019. As shown in Table 1, we also compare the results of other fusion methods, which use the same processing (P) model and decision (D) model as [9]. (1) In the early fusion method, the data of seven sensors are concatenated along their

**Table 2.** Dice(%) performance for MRI modalities being either absent (○) or present (●). \* denotes significant improvement provided by a Wilcoxon test ( $p$ -values  $< 0.05$ ).

Modalities				WT		TC		ET	
T1ce	T1	T2	Flair	GFF	SFusion	GFF	SFusion	GFF	SFusion
●	○	○	○	68.24	<b>69.75*</b>	73.27	<b>75.63*</b>	69.30	<b>71.94*</b>
○	●	○	○	64.45	<b>69.11*</b>	46.93	<b>53.86*</b>	23.74	<b>29.71*</b>
○	○	●	○	<b>79.78</b>	79.61	58.27	<b>61.99*</b>	<b>36.13</b>	35.87
○	○	○	●	81.82	<b>83.97</b>	50.53	<b>52.84</b>	29.50	<b>34.40*</b>
●	●	○	○	74.99	<b>75.30</b>	75.89	<b>80.35*</b>	72.09	<b>74.90*</b>
●	○	●	○	83.93	<b>84.27*</b>	79.55	<b>81.48*</b>	72.87	<b>74.74*</b>
●	○	○	●	<b>87.34</b>	87.32	79.01	<b>79.06</b>	74.89	<b>75.82</b>
○	●	●	○	81.76	<b>81.78</b>	59.75	<b>66.67*</b>	36.50	<b>40.38*</b>
○	●	○	●	85.86	<b>86.39</b>	61.92	<b>62.31</b>	37.52	<b>38.22</b>
○	○	●	●	86.99	<b>87.50*</b>	61.92	<b>66.38*</b>	38.94	<b>41.46*</b>
●	●	●	○	84.48	<b>84.59</b>	79.83	<b>82.32*</b>	73.74	<b>74.78</b>
●	●	○	●	88.03	<b>88.04</b>	80.50	<b>82.04*</b>	74.53	<b>75.44</b>
●	○	●	●	88.75	<b>89.11*</b>	81.60	<b>82.06</b>	74.43	<b>74.91</b>
○	●	●	●	86.84	<b>87.63*</b>	65.38	<b>68.76*</b>	40.90	<b>43.53*</b>
●	●	●	●	88.65	<b>88.93</b>	81.29	<b>82.18</b>	<b>74.55</b>	73.76
Average				82.13	<b>82.89*</b>	69.04	<b>71.86*</b>	55.31	<b>57.32*</b>

**Table 3.** Ablation experiments.

Dice(%)	w/o CE	w/o MA	SFusion
WT	82.42	82.76	<b>82.89</b>
TC	70.39	70.93	<b>71.86</b>
ET	55.65	55.56	<b>57.32</b>

**Table 4.** † denotes results from [24].

Dice(%)	WT	TC	ET	Overall
U-HVED† [10]	75.8	63.2	40.7	59.9
ACNet† [23]	52.5	46.9	41.8	47.1
D <sup>2</sup> -Net† [24]	76.2	66.5	42.3	61.7
SF_FDGf	<b>82.1</b>	<b>69.2</b>	<b>54.9</b>	<b>68.7</b>

$C$  dimension. The prediction results are obtained by inputting the concatenation into a network of P and D in series. (2) For the intermediate fusion approach, the EmbraceNet is replaced with the concatenation of feature representations along their  $R_f$  dimension. (3) In the late fusion method, an independent network of P and D in series is trained for each sensor, and then the decision is made from the averaged softmax outputs. (4) In the confidence fusion model, the EmbraceNet is replaced with the confidence calculation and fusion layers in [7]. The results of different fusion methods on the validation data are presented in Table 1. Our proposed SFusion outperforms the EmbraceNet in all four smartphone locations and improves the overall accuracy from 65.22% to 67.47%.

**Brain Tumor Segmentation.** The quantitative segmentation results are shown in Table 2. Compared with GFF, the network integrated with SFusion achieves better average performance over the 15 possible combinations in all three tasks. In particular, SFusion outperforms GFF for all the possible combinations in TC segmentation. Overall, SFusion achieves better Dice scores in most situations (13,15,13 situations for WT, TC and ET segmentation, respectively). In addition, we conduct the statistical significance analysis. The number of situations with significant improvement are 6, 10 and 8 for WT, TC and ET, respectively. It is provided by a Wilcoxon test ( $p$ -values  $< 0.05$ ). Besides, we find no significant drop in performance caused by SFusion. In addition, we compare the SF\_FDGf (where GFF is replaced by SFusion) with current state-of-the-art methods. Table 4 presents the average dice of 15 situations. For a fair comparison, we conduct experiments on BraTS2018, adopt the same data partition as [24], and cite the results in [24]. SF\_FDGf achieves the best performance and verifies the effectiveness of the SFusion.

**Ablation Experiments.** The correlation extraction (CE) module and the modal attention (MA) module are two key components in SFusion. We evaluate the SFusion without CE and MA, respectively. SFusion without CE denotes that feature representations are directly fed into the MA module (Fig. 2). SFusion without MA means that we directly add the calculated feature representations ( $I'$ ) up to get the fusion result. As shown in Table 1, we can find that SFusion without CE performs worse than other methods. Compared with EmbraceNet, the improvement of SFusion without MA is inconspicuous. As shown in Table. 3, we present the averaged performance over the 15 possible combinations on BraTS2020. It shows that both the CE and MA module lead to performance improvement across all the tumor regions. Therefore, ablation experiments on two different tasks show that both CE and MA play an important role in SFusion.



## 4 Conclusion

In this paper, we propose a self-attention based N-to-One fusion block SFusion to tackle the problem of multimodal missing modalities fusion. As a data-dependent fusion strategy, SFusion can automatically learn the latent correlations between different modalities and builds a shared feature representation. The entire fusion process is based on available data without simulating missing modalities. In addition, SFusion has compatibility with any kind of upstream processing model and downstream decision model, making it universally applicable to different tasks. We show that it can be integrated into existing backbone networks by replacing their fusion operation or block to improve activity recognition and achieve brain tumor segmentation performance. In particular, by integrating with SFusion, SF\_FDGF achieves the state-of-the-art performance. In the future, we will explore other tasks related to variable multimodal fusion with SFusion.

**Acknowledgements.** This work is supported in part by the Guangdong Provincial Natural Science Foundation (2023A1515011431), the Guangzhou Science and Technology Planning Project (202201010092), the National Natural Science Foundation of China (72074105), NSF-1850492 and NSF-2045804.

## References

1. Bakas, S., Menze, B., Davatzikos, C., Kalpathy-Cramer, J., Farahani, K., et al.: MICCAI Brain Tumor Segmentation (BraTS) 2020 Benchmark: Prediction of Survival and Pseudoprogression (Mar 2020). <https://doi.org/10.5281/zenodo.3718904>
2. Chartsias, A., Joyce, T., Giuffrida, M.V., Tsiftaris, S.A.: Multimodal mr synthesis via modality-invariant latent representation. *IEEE Trans. Med. Imaging* **37**(3), 803–814 (2018). <https://doi.org/10.1109/TMI.2017.2764326>
3. Chavarriaga, R., et al.: The opportunity challenge: a benchmark database for on-body sensor-based activity recognition. *Pattern Recogn. Lett.* **34**(15), 2033–2042 (2013)
4. Chen, C., Jafari, R., Kehtarnavaz, N.: Utd-mhad: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: 2015 IEEE International conference on image processing (ICIP), pp. 168–172. IEEE (2015)
5. Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., Heng, P.-A.: Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 447–456. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-32248-9\\_50](https://doi.org/10.1007/978-3-030-32248-9_50)
6. Chen, C., Dou, Q., Jin, Y., Liu, Q., Heng, P.A.: Learning with privileged multimodal knowledge for unimodal segmentation. *IEEE Trans. Medical Imaging* (2021). <https://doi.org/10.1109/TMI.2021.3119385>

7. Choi, J.H., Lee, J.S.: Confidence-based deep multimodal fusion for activity recognition. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, pp. 1548–1556 (2018)
8. Choi, J.H., Lee, J.S.: Embracenet: a robust deep learning architecture for multimodal classification. *Information Fusion* **51**, 259–270 (2019)
9. Choi, J.H., Lee, J.S.: Embracenet for activity: a deep multimodal fusion architecture for activity recognition. In: Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, pp. 693–698 (2019)
10. Dorent, R., Joutard, S., Modat, M., Ourselin, S., Vercauteren, T.: Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 74–82. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-32245-8\\_9](https://doi.org/10.1007/978-3-030-32245-8_9)
11. Graves, M.J., Mitchell, D.G.: Body mri artifacts in clinical practice: a physicist's and radiologist's perspective. *J. Magn. Reson. Imaging* **38**(2), 269–287 (2013)
12. Guo, Z., Li, X., Huang, H., Guo, N., Li, Q.: Deep learning-based image segmentation on multimodal medical imaging. *IEEE Trans. Radiation Plasma Med. Sci.* **3**(2), 162–169 (2019)
13. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: HeMIS: hetero-modal image segmentation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 469–477. Springer, Cham (2016). [https://doi.org/10.1007/978-3-319-46723-8\\_54](https://doi.org/10.1007/978-3-319-46723-8_54)
14. Hu, M., et al.: Knowledge distillation from multi-modal to mono-modal segmentation networks. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 772–781. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-59710-8\\_75](https://doi.org/10.1007/978-3-030-59710-8_75)
15. Isensee, F., et al.: nnu-net: self-adapting framework for u-net-based medical image segmentation. arXiv preprint [arXiv:1809.10486](https://arxiv.org/abs/1809.10486) (2018)
16. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
17. Lau, K., Adler, J., Sjölund, J.: A unified representation network for segmentation with missing modalities. arXiv preprint [arXiv:1908.06683](https://arxiv.org/abs/1908.06683) (2019)
18. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML (2011)
19. Ouyang, J., Adeli, E., Pohl, K.M., Zhao, Q., Zaharchuk, G.: Representation disentanglement for multi-modal brain MRI analysis. In: Feragen, A., Sommer, S., Schnabel, J., Nielsen, M. (eds.) IPMI 2021. LNCS, vol. 12729, pp. 321–333. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-78191-0\\_25](https://doi.org/10.1007/978-3-030-78191-0_25)
20. Shen, L., et al.: Multi-domain image completion for random missing input data. *IEEE Trans. Med. Imaging* **40**(4), 1113–1122 (2021). <https://doi.org/10.1109/TMI.2020.3046444>
21. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
22. Wang, L., Gjoreski, H., Ciliberto, M., Mekki, S., Valentin, S., Roggen, D.: Enabling reproducible research in sensor-based transportation mode recognition with the sussex-huawei dataset. *IEEE Access* **7**, 10870–10891 (2019)

23. Wang, Y., et al.: ACN: adversarial co-training network for brain tumor segmentation with missing modalities. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12907, pp. 410–420. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-87234-2\\_39](https://doi.org/10.1007/978-3-030-87234-2_39)
24. Yang, Q., Guo, X., Chen, Z., Woo, P.Y., Yuan, Y.: D2-net: dual disentanglement network for brain tumor segmentation with missing modalities. *IEEE Trans. Med. Imaging* (2022)
25. Zhou, T., Canu, S., Vera, P., Ruan, S.: Latent correlation representation learning for brain tumor segmentation with missing mri modalities. *IEEE Trans. Image Process.* **30**, 4263–4274 (2021)