MUVF-YOLOX: A Multi-modal Ultrasound Video Fusion Network for Renal Tumor Diagnosis

Junyu Li[1,2,3], Han Huang[1,2,3], Dong Ni[1,2,3], Wufeng Xue[1,2,3], Dongmei Zhu[4(✉)], and Jun Cheng[1,2,3(✉)]

[1] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, China
chengjun583@qq.com

[2] Medical UltraSound Image Computing (MUSIC) Lab, Shenzhen University, Shenzhen, China

[3] Marshall Laboratory of Biomedical Engineering, Shenzhen University, Shenzhen, China

[4] Department of Ultrasound, The Affiliated Nanchong Central Hospital of North Sichuan Medical College, Nanchong, China
zdm596987@gmail.com

Abstract. Early diagnosis of renal cancer can greatly improve the survival rate of patients. Contrast-enhanced ultrasound (CEUS) is a cost-effective and non-invasive imaging technique and has become more and more frequently used for renal tumor diagnosis. However, the classification of benign and malignant renal tumors can still be very challenging due to the highly heterogeneous appearance of cancer and imaging artifacts. Our aim is to detect and classify renal tumors by integrating B-mode and CEUS-mode ultrasound videos. To this end, we propose a novel multi-modal ultrasound video fusion network that can effectively perform multi-modal feature fusion and video classification for renal tumor diagnosis. The attention-based multi-modal fusion module uses cross-attention and self-attention to extract modality-invariant features and modality-specific features in parallel. In addition, we design an object-level temporal aggregation (OTA) module that can automatically filter low-quality features and efficiently integrate temporal information from multiple frames to improve the accuracy of tumor diagnosis. Experimental results on a multicenter dataset show that the proposed framework outperforms the single-modal models and the competing methods. Furthermore, our OTA module achieves higher classification accuracy than the frame-level predictions. Our code is available at https://github.com/JeunyuLi/MUAF.

Keywords: Multi-modal Fusion · Ultrasound Video · Object Detection · Renal Tumor

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 642–651, 2023.
https://doi.org/10.1007/978-3-031-43904-9_62

1 Introduction

Renal cancer is the most lethal malignant tumor of the urinary system, and the incidence is steadily rising [13]. Conventional B-mode ultrasound (US) is a good screening tool but can be limited in its ability to characterize complicated renal lesions. Contrast-enhanced ultrasound (CEUS) can provide information on microcirculatory perfusion. Compared with CT and MRI, CEUS is radiation-free, cost-effective, and safe in patients with renal dysfunction. Due to these benefits, CEUS is becoming increasingly popular in diagnosing renal lesions. However, recognizing important diagnostic features from CEUS videos to diagnose lesions as benign or malignant is non-trivial and requires lots of experience.

To improve diagnostic efficiency and accuracy, many computational methods were proposed to analyze renal US images and could assist radiologists in making clinical decisions [6]. However, most of these methods only focused on conventional B-mode images. In recent years, there has been increasing interest in multi-modal medical image fusion [1]. Directly concatenation and addition were the most common methods, such as [3,4,12]. These simple operations might not highlight essential information from different modalities. Weight-based fusion methods generally used an importance prediction module to learn the weight of each modality and then performed sum, replacement, or exchange based on the weights [7,16,17,19]. Although effective, these methods did not allow direct interaction between multi-modal information. To address this, attention-based methods were proposed. They utilized cross-attention to establish the feature correlation of different modalities and self-attention to focus on global feature modeling [9,18]. Nevertheless, we prove in our experiments that these attention-based methods may have the potential risks of entangling features of different modalities.

In practice, experienced radiologists usually utilize dynamic information on tumors' blood supply in CEUS videos to make diagnoses [8]. Previous researches have proved that temporal information is effective in improving the performance of deep learning models. Lin et al.[11] proposed a network for breast lesion detection in US videos by aggregating temporal features, which outperformed other image-based methods. Chen et al. [2] showed that CEUS videos can provide more detailed blood supply information of tumors allowing a more accurate breast lesion diagnosis than static US images.

In this work, we propose a novel multi-modal US video fusion network (MUVF-YOLOX) based on CEUS videos for renal tumor diagnosis. Our main contributions are fourfold. (1) To the best of our knowledge, this is the first deep learning-based multi-modal framework that integrates both B-mode and CEUS-mode information for renal tumor diagnosis using US videos. (2) We propose an attention-based multi-modal fusion (AMF) module consisting of cross-attention and self-attention blocks to capture modality-invariant and modality-specific features in parallel. (3) We design an object-level temporal aggregation (OTA) module to make video-based diagnostic decisions based on the information from multi-frames. (4) We build the first multi-modal US video datatset containing

B-mode and CEUS-mode videos for renal tumor diagnosis. Experimental results show that the proposed framework outperforms single-modal, single-frame, and other state-of-the-art methods in renal tumor diagnosis.

2 Methods

2.1 Overview of Framework

The proposed MUVF-YOLOX framework is shown in Fig. 1. It can be divided into two stages: single-frame detection stage and video-based diagnosis stage. (1) In the single-frame detection stage, the network predicts the tumor bounding box and category on each frame in the multi-modal CEUS video clips. Dual-branch backbone is adopted to extract the features from two modalities and followed by the AMF module to fuse these features. During the diagnostic process, experienced radiologists usually take the global features of US images into consideration [20]. Therefore, we modify the backbone of YOLOX from CSP-Darknet to Swin-Transformer-Tiny, which is a more suitable choice by the virtue of its global modeling capabilities [15]. (2) In the video-based diagnosis stage, the network automatically chooses high-confidence region features of each frame according to the single-frame detection results and performs temporal aggregation to output a more accurate diagnosis. The above two stages are trained successively. We first perform a strong data augmentation to train the network for tumor detection and classification on individual frames. After that, the first stage model is switched to the evaluation mode and predicts the label of each frame in the video clip. Finally, we train the OTA module to aggregate the temporal information for precise diagnosis.

2.2 Dual-Attention Strategy for Multimodal Fusion

Using complementary information between multi-modal data can greatly improve the precision of detection. Therefore, we propose a novel AMF module to fuse the features of different modalities. As shown in Fig. 1, the features of each modality will be input into cross-attention and self-attention blocks in parallel to extract modality-invariant features and modality-specific features simultaneously.

Taking the B-mode as an example, we first map the B-mode features F_B and the CEUS-mode features F_C into (Q_B, K_B, V_B) and (Q_C, K_C, V_C) using linear projection. Then cross-attention uses scaled dot-product to calculate the similarity between Q_B and K_C. The similarity is used to weight V_C. Cross-attention extracts modality-invariant features through correlation calculation but ignores modality-specific features in individual modalities. Therefore, we apply self-attention in parallel to highlight these features. The self-attention calculates the similarity between Q_B and K_B and then uses the similarity to weight V_B. Similarly, the features of the CEUS modality go through the same process in parallel. Finally, we merge the two cross-attention outputs by addition

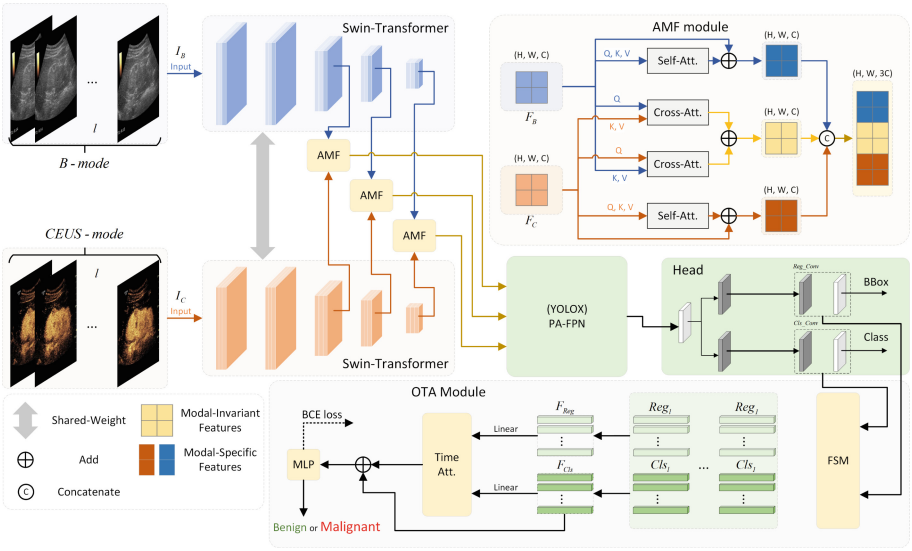

Fig. 1. Framework of MUVF-YOLOX. AMF module is used to fuse multi-modal features. OTA module is used to classify the tumor as benign or malignant based on videos. FSM means feature selection module.

since they are both invariant features of two modalities and concatenate the obtained sum and the outputs of the two self-attention blocks. The process mentioned above can be formulated as follows:

$$F_{invar} = Softmax(\frac{Q_B K_C^T}{\sqrt{d}})V_C + Softmax(\frac{Q_C K_B^T}{\sqrt{d}})V_B \tag{1}$$

$$F_{B-spec} = Softmax(\frac{Q_B K_B^T}{\sqrt{d}})V_B + F_B \tag{2}$$

$$F_{C-spec} = Softmax(\frac{Q_C K_C^T}{\sqrt{d}})V_C + F_C \tag{3}$$

$$F_{AMF} = Concat(F_{B-spec}, F_{invar}, F_{C-spec}) \tag{4}$$

where, F_{invar} represents the modality-invariant features. F_{B-spec} and F_{C-spec} represent the modal-specific features of B-mode and CEUS-mode respectively. F_{AMF} is the output of the AMF module.

2.3 Video-Level Decision Generation

In clinical practice, the dynamic changes in US videos provide useful information for radiologists to make diagnoses. Therefore, we design an OTA module that aggregates single-frame renal tumor detection results in temporal dimension for diagnosing tumors as benign and malignant. First, we utilize a feature selection

module [14] to select high-quality features of each frame from the Cls_conv and Reg_conv layers. Specifically, we select the top 750 grid cells on the prediction grid according to the confidence score. Then, 30 of the top 750 grid cells are chosen by the non-maximum suppression algorithm for reducing redundancy. The features are finally picked out from the Cls_conv and Reg_conv layers guided by the positions of the top 30 grid cells. Let $F_{Cls} = \{Cls_1, Cls_2, ...Cls_l\}$ and $F_{Reg} = \{Reg_1, Reg_2, ...Reg_l\}$ denote the above obtained high-quality features from l frames. After feature selection, we aggregate the features in the temporal dimension by time attention. F_{Cls} and F_{Reg} are mapped into $(Q_{Cls}, K_{Cls}, V_{Cls})$ and (Q_{Reg}, K_{Reg}) via linear projection. Then, we utilize scaled dot-product to compute the attention weights of V_{Cls} as:

$$Time\_Att. = [Softmax(\frac{Q_{Cls}K_{Cls}^T}{\sqrt{d}}) + Softmax(\frac{Q_{Reg}K_{Reg}^T}{\sqrt{d}})]V_{Cls} \quad (5)$$

$$F_{temp} = Time\_Att. + F_{Cls} \quad (6)$$

After temporal feature aggregation, F_{temp} is fed into a multilayer perceptron head to predict the class of tumor.

3 Experimental Results

3.1 Materials and Implementations

We collect a renal tumor US dataset of 179 cases from two medical centers, which is split into the training and validation sets. We further collect 36 cases from the two medical centers mentioned above (14 benign cases) and another center (Fujian Provincial Hospital, 22 malignant cases) to form the test set. Each case has a video with simultaneous imaging of B-mode and CEUS-mode. Some examples of the images are shown in Fig. 2. There is an obvious visual difference between the images from the Fujian Provincial Hospital (last column in Fig. 2) and the other two centers, which raises the complexity of the task but can better verify our method's generalization ability. More than two radiologists with ten years of experience manually annotate the tumor bounding box and class label at the frame level using the Pair annotation software package (https://www.aipair.com.cn/en/, Version 2.7, RayShape, Shenzhen, China) [10]. Each case has 40–50 labeled frames, and these frames cover the complete contrast-enhanced imaging cycle. The number of cases and annotated frames is summarized in Table 1.

Weights pre-trained from ImageNet are used to initialize the Swin-Transformer backbone. Data augmentation strategies are applied synchronously to B-mode and CEUS-mode images for all experiments, including random rotation, mosaic, mixup, and so on. All models are trained for 150 epochs. The batch size is set to 2. We use the SGD optimizer with a learning rate of 0.0025. The weight decay is set to 0.0005 and the momentum is set to 0.9. In the test phase, we use the weights of the best model in validation to make predictions. All Experiments are implemented in PyTorch with an NVIDIA RTX A6000 GPU. AP_{50} and AP_{75} are used to assess the performance of single-frame detection. Accuracy and F1-score are used to evaluate the video-based tumor diagnosis.

Table 1. The details of our dataset. Number of cases in brackets.

Category	Training	Validation	Test
Benign	2775(63)	841(16)	875(14)
Malignant	4017(81)	894(19)	1701(22)
Total	6792(144)	1735(35)	2576(36)

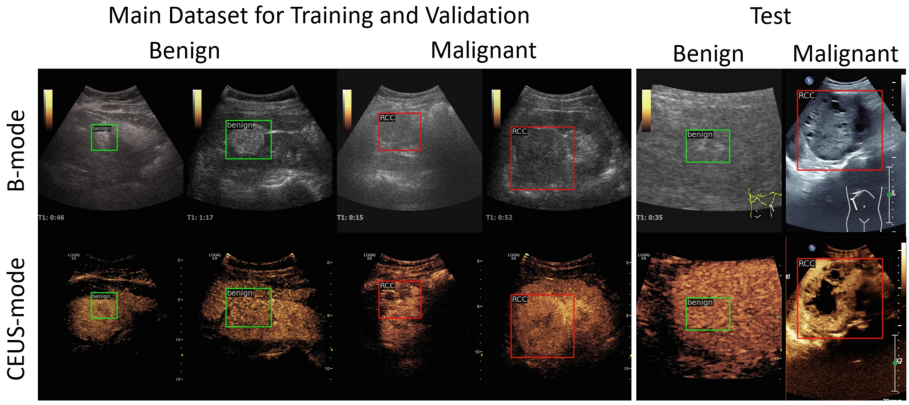

Fig. 2. Examples of the annotated B-mode and CEUS-mode US images.

3.2 Ablation Study

Single-Frame Detection. We explore the impact of different backbones in YOLOX and different ways of multi-modal fusion. As shown in Table 2, using Swin-Transformer as the backbone in YOLOX achieves better performance than the original backbone while reducing half of the parameters. The improvement may stem from the fact that Swin-Transformer has a better ability to characterize global features, which is critical in US image diagnosis. In addition, we explore the role of cross-attention and self-attention blocks in multi-modal tasks, as well as the optimal strategy for combining their outputs. Comparing row 5 with row 7 and row 8 in Table 2, the dual-attention mechanism outperforms the single cross-attention. It indicates that we need to pay attention to both modality-invariant and modality-specific features in our multi-modal task through cross-attention and self-attention blocks. However, "CA+SA" (row 6 in Table 2) obtains inferior performance than "CA" (row 5 in Table 2). We conjecture that connecting the two attention modules in series leads to the entanglement of modality-specific and modality-invariant information, which would disrupt the model training. On the contrary, the "CA//SA" method, combining two attention modules in parallel, enables the model to capture and digest modality-specific and modality-invariant features independently. For the same reason, we concatenate the outputs of the attention blocks rather than summing, which further avoids confusing modality-

specific and modality-invariant information. Therefore, the proposed method achieves the best performance.

Table 2. The results of ablation study. "CA" and "SA" denote cross-attention and self-attention respectively. "//" and "+" mean parallel connection and series connection.

Modal	Network	Validation		Test		Flops (GLOPS)	Params (M)
		AP_{50}	AP_{75}	AP_{50}	AP_{75}		
B-mode	YOLOX [5]	**60.7**	**39.7**	48.5	19.6	140.76	99.00
	Swin-YOLOX	59.6	38.0	**58.1**	**22.1**	**61.28**	**45.06**
CEUS-mode	YOLOX [5]	52.3	29.7	49.1	**17.8**	140.76	99.00
	Swin-YOLOX	**60.1**	**30.6**	**52.1**	14.1	**61.28**	**45.06**
Multi-modal	CA (CMF [18])	81.4	54.2	75.2	35.2	103.53	51.26
	CA+SA (TMM [9])	80.8	52.7	74.3	37.0	109.19	57.46
	CA//SA (Ours w/o Concat)	82.0	56.9	74.6	35.0	109.19	57.46
	Ours	**82.8**	**60.6**	**79.5**	**39.2**	117.69	66.76

Video-Based Diagnosis. We investigate the performance of the OTA module for renal tumor diagnosis in multi-modal videos. We generate a video clip with l frames from annotated frames at a fixed interval forward. As shown in Table 3, gradually increasing the clip length can effectively improve the accuracy. This suggests that the multi-frame model can provide a more comprehensive characterization of the tumor and thus achieves better performance. Meanwhile, increasing the sampling interval tends to decrease the performance (row 4 and row 5 in Table 3). It indicates that continuous inter-frame information is beneficial for renal tumor diagnosis.

Table 3. The results of video-based diagnosis.

Clip Length	Sampling Interval	Validation		Test	
		Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
1	1	81.6	81.6	90.3	89.2
2	1	82.1	82.1	90.5	89.3
2	2	82.9	82.9	90.0	88.7
4	1	83.7	83.7	**91.0**	**90.0**
4	2	82.6	82.6	90.8	89.7
8	1	**84.0**	**84.0**	90.9	89.9

3.3 Comparison with Other Methods

The comparison results are shown in Table 4. Compared to the single-modal models, directly concatenating multi-modal features (row 3 in Table 4) improves

AP_{50} and AP_{75} by more than 15%. This proves that complementary information exists among different modalities. For a fair comparison with other fusion methods, we embed their fusion modules into our framework so that different approaches can be validated in the same environment. CMML [19] and CEN [17] merge the multi-modal features or pick one of them by automatically generating channel-wise weights for each modality. They score higher AP in the validation set but lower one in the test set than "Concatenate". This may be because the generated weights are biased to make similar decisions to the source domain, thereby reducing model generalization in the external data. Moreover, CMF only highlights similar features between two modalities, ignoring that each modality contains some unique features. TMM focuses on both modality-specific and modality-invariant information, but the chaotic confusion of the two types of information deteriorates the model performance. Therefore, both CMF [17] and TMM [9] fail to outperform weight-based models. On the contrary, our AMF module prevents information entanglement by conducting cross-attention and self-attention blocks in parallel. It achieves $AP_{50} = 82.8$, $AP_{75} = 60.6$ in the validation set and $AP_{50} = 79.5$, $AP_{75} = 39.2$ in the test set, outperforming all competing methods while demonstrating superior generalization ability. Meanwhile, the improvement of the detection performance is beneficial to our OTA module to obtain lesion features from more precise locations, thereby improving the accuracy of benign and malignant renal tumor diagnosis.

Table 4. Diagnosis results of different methods.

Fusion Methods	Validation				External Test			
	AP_{50}	AP_{75}	Accuracy	F1-score	AP_{50}	AP_{75}	Accuracy	F1-score
B-mode	59.6	38.0	76.4	76.3	58.1	22.1	80.4	79.1
CEUS-mode	60.1	30.6	78.2	78.1	52.1	14.1	70.5	69.3
Concatenate	78.8	50.5	79.6	79.5	76.8	38.8	86.8	85.7
CMML [19]	80.7	54.4	80.1	80.1	76.0	37.2	87.4	86.2
CEN [17]	81.4	56.2	83.0	83.0	74.3	36.3	85.1	83.8
CMF [18]	81.4	54.8	79.7	79.7	75.2	35.2	87.8	86.8
TMM [9]	80.8	52.7	80.1	80.1	74.3	37.0	84.4	83.2
Ours	**82.8**	**60.6**	**84.0**	**84.0**	**79.5**	**39.2**	**90.9**	**89.9**

4 Conclusions

In this paper, we create the first multi-modal CEUS video dataset and propose a novel attention-based multi-modal video fusion framework for renal tumor diagnosis using B-mode and CEUS-mode US videos. It encourages interactions between different modalities via a weight-sharing dual-branch backbone and automatically captures the modality-invariant and modality-specific information by the AMF module. It also utilizes a portable OTA module to aggregate

information in the temporal dimension of videos, making video-level decisions. The design of the AMF module and OTA module is plug-and-play and could be applied to other multi-modal video tasks. The experimental results show that the proposed method outperforms single-modal, single-frame, and other state-of-the-art multi-modal approaches.

Acknowledgment. Our dataset was collected from The Affiliated Nanchong Central Hospital of North Sichuan Medical College, Shenzhen People's Hospital, and Fujian Provincial Hospital hospitals. This study was approved by local institutional review boards. This work is supported by the Guangdong Basic and Applied Basic Research Foundation (No. 2021B1515120059), National Natural Science Foundation of China (No. 62171290), and Shenzhen Science and Technology Program (No. SGDX20201103095613036 and 20220810145705001).

References

1. Azam, M.A., et al.: A review on multimodal medical image fusion: compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. Comput. Biol. Med. **144**, 105253 (2022)
2. Chen, C., Wang, Y., Niu, J., Liu, X., Li, Q., Gong, X.: Domain knowledge powered deep learning for breast cancer diagnosis based on contrast-enhanced ultrasound videos. IEEE Trans. Med. Imaging **40**(9), 2439–2451 (2021)
3. Chen, H., Li, Y., Su, D.: Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. Pattern Recogn. **86**, 376–385 (2019)
4. Fang, J., et al.: Weighted concordance index loss-based multimodal survival modeling for radiation encephalopathy assessment in nasopharyngeal carcinoma radiotherapy. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part VII, vol. 13437, pp. 191–201. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_19
5. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOx: exceeding YOLO series in 2021. arXiv preprint arXiv:2107.08430 (2021)
6. George, M., Anita, H.: Analysis of kidney ultrasound images using deep learning and machine learning techniques: a review. Pervasive Comput. Soc. Networking Proc. ICPCSN **2021**, 183–199 (2022)
7. Huang, H., et al.: Personalized diagnostic tool for thyroid cancer classification using multi-view ultrasound. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part III, vol. 13433, pp. 665–674. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_64
8. Kapetas, P., et al.: Quantitative multiparametric breast ultrasound: application of contrast-enhanced ultrasound and elastography leads to an improved differentiation of benign and malignant lesions. Invest. Radiol. **54**(5), 257 (2019)
9. Li, X., Ma, S., Tang, J., Guo, F.: TranSiam: fusing multimodal visual features using transformer for medical image segmentation. arXiv preprint arXiv:2204.12185 (2022)

10. Liang, J., et al.: Sketch guided and progressive growing GAN for realistic and editable ultrasound image synthesis. Med. Image Anal. **79**, 102461 (2022)
11. Lin, Z., Lin, J., Zhu, L., Fu, H., Qin, J., Wang, L.: A new dataset and a baseline model for breast lesion detection in ultrasound videos. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part III, vol. 13433, pp. 614–623. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_59
12. Liu, Y., Chen, X., Cheng, J., Peng, H.: A medical image fusion method based on convolutional neural networks. In: 2017 20th International Conference on Information Fusion (Fusion), pp. 1–7. IEEE (2017)
13. Ljungberg, B., et al.: European association of urology guidelines on renal cell carcinoma: the 2019 update. Eur. Urol. **75**(5), 799–810 (2019)
14. Shi, Y., Wang, N., Guo, X.: YOLOV: making still image object detectors great at video object detection. arXiv preprint arXiv:2208.09686 (2022)
15. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: TransBTS: multimodal brain tumor segmentation using transformer. In: de Bruijne, M., et al. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021, Proceedings, Part I 24, pp. 109–119. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_11
16. Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y., Huang, J.: Deep multimodal fusion by channel exchanging. Adv. Neural. Inf. Process. Syst. **33**, 4835–4845 (2020)
17. Wang, Y., Sun, F., Huang, W., He, F., Tao, D.: Channel exchanging networks for multimodal and multitask dense image prediction. IEEE Trans. Pattern Anal. Mach. Intell. **45**(5), 5481–5496 (2022)
18. Xu, J., et al.: RemixFormer: a transformer model for precision skin tumor differential diagnosis via multi-modal imaging and non-imaging data. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part III, vol. 13433, pp. 624–633. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_60
19. Yang, Y., Wang, K.T., Zhan, D.C., Xiong, H., Jiang, Y.: Comprehensive semi-supervised multi-modal learning. In: IJCAI, pp. 4092–4098 (2019)
20. Zhu, J., et al.: Contrast-enhanced ultrasound (CEUS) of benign and malignant renal tumors: distinguishing CEUS features differ with tumor size. Cancer Med. **12**(3), 2551–2559 (2022)