TransLiver: A Hybrid Transformer Model for Multi-phase Liver Lesion Classification

Xierui Wang[1], Hanning Ying[2], Xiaoyin Xu[3], Xiujun Cai[2], and Min Zhang[1](✉)

[1] College of Computer Science and Technology,
Zhejiang University, Hangzhou, China
min_zhang@zju.edu.cn

[2] Sir Run Run Shaw Hospital (SRRSH), affiliated with the Zhejiang University
School of Medicine, Hangzhou, China

[3] Brigham and Women's Hospital, Harvard Medical School, Boston, USA

Abstract. Early diagnosis of focal liver lesions (FLLs) can decrease the fatality rate of liver cancer, which remains a big challenge. We designed a deep learning approach based on CT to assess and differentiate FLLs. To achieve high accuracy, CTs in different phases are integrated to provide more information than single-phase images. While most of the related studies use convolutional neural networks, we exploit the Transformer for multi-phase liver lesion classification. We propose a hybrid model called TransLiver, which has a transformer backbone and complementary convolutional modules. Specifically, we connect modified transformer blocks with convolutional encoder and down-samplers. For multi-phase fusion, we utilize cross phase tokens to reinforce the phases communication. In addition, we introduce a pre-processing unit to resolve realistic annotation issues. Extensive experiments are conducted, in which we achieve an overall accuracy of 90.9% on an in-house dataset of four CT phases and seven liver lesion classes. The results also show distinct advantages in comparison to state-of-art approaches in classification. The code is available at https://github.com/sherrydoge/TransLiver.

Keywords: Focal liver lesion · Multi-phase fusion · Transformer

1 Introduction

Liver cancer is one of the most deadly cancers and has the second highest fatality rate [17]. Focal liver lesions (FLLs) are the most common lesions found in liver cancer, yet FLLs are challenging to diagnose because they can be either benign lesions, such as focal nodular hyperplasia (FNH), hepatic abscess (HA), hepatic

X. Wang and H. Ying—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43895-0_31.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 329–338, 2023.
https://doi.org/10.1007/978-3-031-43895-0_31

hemangioma (HH), and hepatic cyst (HC) or malignant tumors, such as intrahepatic cholangiocarcinoma (ICC), hepatic metastases (HM), and hepatocellular carcinoma (HCC). Accurate early diagnosis of FLLs is thus critical to increasing the 5-year survival rate, a task that remains challenging as of today. Dynamic contrast-enhanced CT is a common technique for liver cancer diagnosis, where four different phases of imaging, namely, non-contrast (NC), arterial (ART), portal venous (PV), and delayed (DL) provide complementary information about the liver. Different types of FLLs acquired in the four phases are shown in Fig. 1.

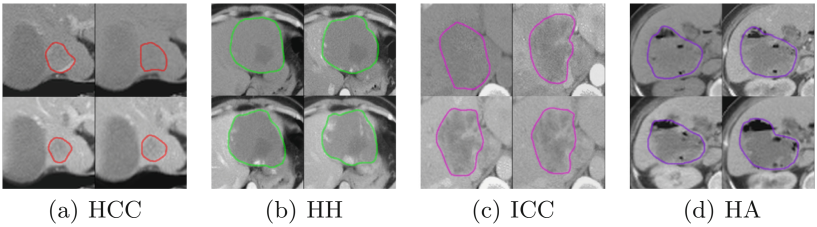
(a) HCC (b) HH (c) ICC (d) HA

Fig. 1. Representative types of FLLs shown in different CT phases, where contours show the annotated lesion boundaries. In each image, the phase sequence from left to right and top to bottom is NC, ART, PV, and DL, respectively.

With the development of deep learning, computer-aided liver lesion diagnosis has attracted much attention [5,8,16] in recent years. Romero et al. [16] presented an end-to-end framework based on Inception-V3 and InceptionResNet-V2 to discriminate liver lesions between cysts and malignant tumors. Heker et al. [8] combined liver segmentation and classification using transfer learning and joint learning to increase the performance of CNN. As a manner to elevate the accuracy of CNNs, Frid-Adar et al. [5] designed a GAN-based network to generate synthetic liver lesion images, improving the classification performance based on CNN. It is reported in many studies [9,18] that using multi-phase data, like most professionals do in practice, can help the network get a more accurate result, which also acts in liver lesion classification [15,23,24]. Yasaka et al. [24] proposed multi-channel CNN to extract features from multi-phase liver CT by concatenation. Roboh et al. [15] proposed an algorithm based on CNNs to handle 3D context in liver CTs and utilized clinical context to assist the classification. Xu et al. [23] constructed a knowledge-guided framework to integrate liver lesion features from three phases using self-attention and fused them with a cross-feature interaction module and a cross-lesion correlation module.

A single-phase lesion annotation means the annotation of both lesion position and its class. In hospitals, collected multi-phase CTs are normally grouped by patients rather than lesions, which makes single-phase lesion annotation insufficient for feature fusion learning. However, the number of lesions inside a single patient can vary from one to dozens and they can be of different types in realistic cases. Multi-phase CTs are also not co-registered in most cases, therefore, it is necessary to make sure the lesions extracted from different phases are somehow aligned for feature fusion, which is called as multi-phase lesion annotation.

Moreover, while most works have attached much importance to liver lesion segmentation [2], its outcome is usually organized at a single-phase level. Additional effort will be needed when consolidating segmentation and multi-phase classification.

Self-attention based transformers [19] have shown strong capability in natural language processing tasks. Meanwhile, vision transformers (ViT) [4] have been shown to replace CNN with a transformer encoder in computer vision tasks and can achieve obvious advantages on large-scale datasets. To the best of our knowledge, we find no study using ViT backbone network in liver lesion classification. The reason for this is twofold. First, pure ViT has several limitations itself [6], including ignoring local information within each patch, extracting only single-scale features, and lacking inductive bias. Second, no complete open liver lesion classification datasets exist. Most relevant studies are based on private datasets, which tend to be small in size and cause overfitting in learning models.

In this paper, we construct a hybrid framework with ViT backbone for liver lesion classification, **TransLiver**. We design a pre-processing unit to reduce the annotation cost, where we obtain lesion area on multi-phase CTs from annotations marked on a single phase. To alleviate the limitations of pure transformers, we propose a multi-stage pyramid structure and add convolutional layers to the original transformer encoder. We use additional cross phase tokens at the last stage to complete a multi-phase fusion, which can focus on cross-phase communication and improve the fusion effectiveness as compared with conventional modes. While most multi-phase liver lesion classification studies use datasets with no more than three phases (without DL phase for its difficulty of collection) or no more than six lesion classes, we validate the whole framework on an in-house dataset with four phases of abdominal CT and seven classes of liver lesions. Considering the disproportion of axial lesion slice number and the relatively small scale of the dataset, we adopt a 2-D network in classification part instead of 3-D in pre-processing part and achieve a 90.9% accuracy.

2 Method

Figure 2 illustrates the overall architecture of TransLiver, where activation layers and batch normalization layers are omitted. Multi-phase liver lesion CTs are converted from single-phase annotation to multi-phase annotation by a pre-processing unit including a registration network and a lesion matcher.

For each phase, a convolutional encoder extracts preliminary lesion features on axial slices. As the backbone of the whole framework, transformer encoder employs a 4-stage pyramid structure extracting multi-scale features, with each stage connected by a convolutional down-sampler. There are two types of transformer blocks, single-phase liver transformer block (SPLTB) and multi-phase liver transformer block (MPLTB). The former is phase-specific, while the latter is in charge of multi-phase fusion. Extracted features from different phases are averaged and classified by two successive fully connected networks. A voting strategy about slices is applied to decide the classification results.

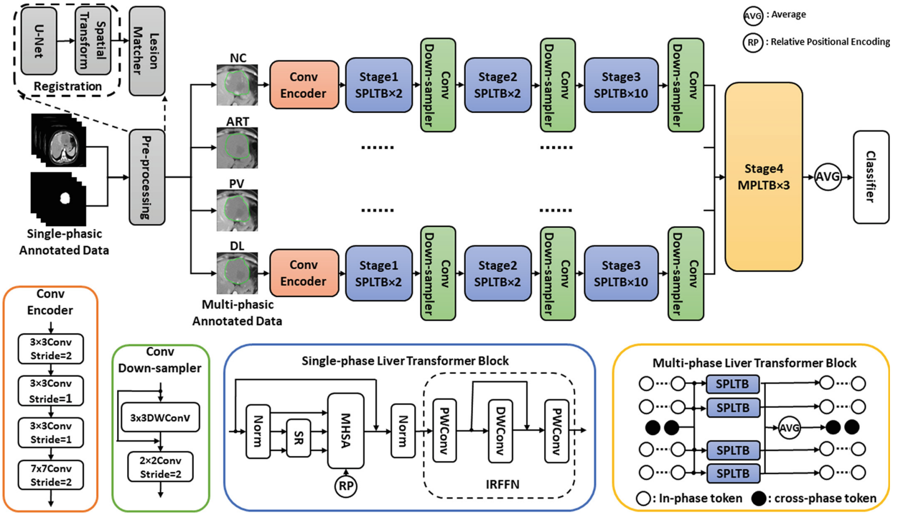

Fig. 2. The overall architecture of the proposed TransLiver model.

2.1 Pre-processing Unit

The single-phase annotated lesion has the position and class labels in all phases but they are not aligned, so we could have difficulty finding out which lesions in different phases are the same with 2 or more lesions in one patient. To reduce errors caused by unregistered data and address the situation that one patient has multiple lesions of different types, we pre-process the multi-phase liver CTs registered and grouped by lesions.

The registration network is based on Voxelmorph [1], with a U-Net learning registration field and moving data transformed by the field. We also use auxiliary Dice loss function between fixed image lesion masks and moved image lesion masks to help the registration field learning. In [1], the network needs to specify an atlas image, otherwise, pairs of images will be registered to each other. But in our work, we need to register the original data in a cross-phase form. We choose an atlas phase ART as suggested by clinicians and other phases of CTs are registered to the ART phase of every patient.

After registration, a lesion matcher finds the same lesions in different phases. We generate a minimum circumscribed cuboid with padding as the lesion window for each lesion to keep the surrounding information. The windows are then converted to 0–1 masks to calculate Dice coefficient. Lesions with the maximal window Dice coefficient that is no less than a set threshold are considered the same. Only lesions completely found in all phases will be used in the following classification network.

2.2 Convolutional Encoder and Convolutional Down-Sampler

In pure vision transformer, input images are converted to tokens by patch embedding and added with positional encoding to keep the positional information. Patch embedding consists of a linear connected layer or a convolutional layer, which does not enable to construct local relation [13]. Absolute positional encoding destroys the translation variance [10] that keeps the rotation and shift operations from altering the final results [6].

So, we construct a convolutional encoder without absolute positional encoding to replace the original embedding layer. For an input image $X \in \mathbb{R}^{B \times H \times W \times 1}$, B is the batch size, and $H \times W$ is the size of the input. The module contains four convolutional layers playing different roles. The first layer, Conv_1, with a kernel size of 3, stride of 2, and output channels of 32, reduces the size to $\frac{H}{2} \times \frac{W}{2}$. Next two layers, Conv_2 and Conv_3, each with a kernel size of 3, stride of 1, and the same output channel as Conv_1, extract local information. Conv_1, Conv_2, and Conv_3 each is followed by a GeLU activation layer and a batch normalization. Considering the design of PVTv2 [21], an overlapped convolutional layer, Conv_4, with a kernel size of 7, stride of 2, and output channels of 64, is used to strengthen the connection among patches. It is followed by layer normalization. The output Z is then reshaped from $\mathbb{R}^{B \times \frac{H}{4} \times \frac{W}{4} \times 64}$ to $\mathbb{R}^{B \times \frac{H \times W}{16} \times 64}$ to finish the tokenization of transformer.

We add convolutional down-samplers between stages of transformer encoder so that they can produce hierarchical representation like CNN structure. Each convolutional down-sampler contains a residual structure with a 3×3 depthwise convolution to increase the locality of our model. We also utilize a convolutional layer with a kernel size of 2 and stride of 2, which halves the image resolution and doubles the number of channels.

2.3 Single-Phase Liver Transformer Block

Vision Transformers can get excellent performance on large-scale datasets such as ImageNet [4], but they are also prone to overfit on small datasets such as private hospital datasets. We adopt the spatial reduction structure proposed in PVT [20] to largely reduce the computational overhead by reducing the size of K and V using depthwise convolution. Following [6,12], a learnable relative positional encoding is added here to replace the absolute positional encoding. The self-attention module can be written as:

$$\text{Attn}\left(Q, K, V\right) = \text{Softmax}\left(\frac{Q \times \text{SR}\left(K\right)}{\sqrt{d_h}} + P\right) \text{SR}\left(V\right) \tag{1}$$

where Q, K, V are the same with original ViT, d_h is the head dimension, and P is the relative positional encoding. Spatial reduction SR consists of a $k \times k$ depthwise convolution with a stride of k and a batch normalization, where k is the spatial reduction ratio set in each stage.

Feed forward network (FFN) is designed for a better capacity of representation. We use the module designed in [6] IRFFN (Inverted Residual FFN) with three convolutions instead of two linear layers in the initial vision transformer. The first and third convolutions are pointwise for dimension translation, which has a similar effect to the original linear layers. The second convolution with a shortcut connection extracts local information in a higher dimension and improves the gradient propagation ability across layers [6]. The structure also has two GeLU activation layers between convolutional layers and three batch normalizations after the GeLUs and the last convolutional layer for better performance.

2.4 Multi-phase Liver Transformer Block

Single-phase liver transformer block (SPLTB) is phase-specific, which means the model parameters of each phase are independent. It can fully extract features in different phases before fusion. Inspired by [14], in stage 4, we design a multi-phase liver transformer block (MPLTB) for communication between phases. MPLTB introduced some new parameters that are not in the original transformer. These parameters are randomly initialized, concatenated with the corresponding phase tokens respectively, and updated in phase-specific SPLTB. Then, they are separated and averaged for the next layer. The whole module is defined as:

$$\begin{aligned} \text{Concat}\left(X_i^{l+1}, t_i^l\right) &= \text{SPLTB}\left(\text{Concat}\left(X_i^l, t^l\right)\right) \\ t^{l+1} &= \text{Avg}\left(t_i^l\right) \end{aligned} \tag{2}$$

where X_i^l is phase tokens of the ith phase and the lth layer and t^l is cross phase tokens of the lth layer. Because of the phase-specific SPLTB, t_i^l represents the corresponding cross phase tokens output of the ith phase in the lth layer. Cross phase tokens need negligible extra cost and can force the information to concentrate inside the tokens [14]. Compared to the direct fusion of input images or output features like average and concatenation, cross phase tokens can also reduce fusion granularity to sufficiently explore the relationship among phases. It is worth noting that these tokens will be removed provisionally when reshaping the phase tokens in SPLTB for right execution. The fusion is conducted in deep layers because the semantic concepts are learned in higher layers which benefits the cross phase connection.

3 Experiments

3.1 Liver Lesion Classification

Dataset. The employed single-phase annotated dataset is collected from Sir Run Run Shaw Hospital (SRRSH), affiliated with the Zhejiang University School of Medicine, and has received the ethics approval of IRB. The collection process can be found in supplementary materials. The size of each CT slice is decreased to 224×224 using cubic interpolation. After the pre-processing unit with window Dice threshold of 0.3, we screen 761 lesions from 444 patients with four phases of CTs, seven types of lesions (13.2% of HCC, 5.3% of HM, 11.3% of ICC, 22.6% of HH, 31.1% of HC, 8.7% of FNH, and 7.8% of HA), and totally 4820 slices. To handle the imbalance of dataset, we randomly select 586 lesions as the training and validation set with no more than 700 axial slices in each lesion type, and the rest 175 lesions constitute the test set. Lesions from the same patient are either assigned to the training and validation set or the test set, but not both.

Implementations. The training and validation set is randomly divided with a 4:1 ratio. The data is augmented by flip, rotation, crop, shift, and scale. We initialize the backbone network using pre-trained weights of CMT-S [6]. Our models are implemented by Pytorch1.12.1 and Timm0.6.13 [22]. Then, they are trained on four NVIDIA Tesla A100 GPUs for 200 epochs using cross-entropy loss function with label smoothing and SGD optimizer with learning rate warmup and cosine annealing. The batch size is 32 and the learning rate is 1e-3. We measured performance by precision (Pre.), sensitivity (Sen.), specificity (Spe.), F1-score (F1), area under the curve (AUC), and accuracy (Acc.).

Results. We first compare the class-wise accuracy of our model against other advanced methods applying different architectures in multi-phase liver lesion classification with more than four lesion types [3,11,15,23,24]. TransLiver gets the highest overall accuracy of 90.9% classifying the most lesion types of seven (HCC 90.9%, HM 62.5%, ICC 73.7%, HH 91.7%, HC 100.0%, FNH 100.0%, and HA 93.3%). In the results of our method, HM has a relatively low performance of 62.5%, mainly due to its low proportion in our dataset. The details can be found in supplementary materials.

Because the sources of data are different among the methods compared above and to the best of our knowledge, no relevant study based on transformers was found, we further train some SOTA normal classification models on our dataset. Considering the fairness, all the models below are initialized with pre-trained weights and adopt 2-D structures using the same slice-level classification strategy. For completeness, we concatenate the multi-phase features to execute the fusion. As illustrated in Table 1, our proposed TransLiver model gets better performance than other models in all metrics. Behind our model, CMT-S achieves the best performance, indicating the effect of convolutional structures in transformer.

Table 1. Performance of TransLiver and other SOTA classification methods.

Method	Pre.	Sen.	Spe.	F1	AUC	Acc.
ResNet-18 [7]	71.7	72.6	96.1	71.2	92.6	77.1
ViT-S [4]	79.6	79.4	97.2	78.6	92.9	82.9
Swin-S [12]	77.7	78.1	97.1	77.3	93.8	82.3
CMT-S [6]	80.5	80.5	97.6	80.0	94.1	85.7
TransLiver	**88.7**	**87.4**	**98.5**	**87.3**	**95.1**	**90.9**

3.2 Ablation Study

To verify the improvement of our modules, we conduct three baseline experiments for comparison. Here convolutional encoder, convolutional down-sampler, and SPLTB as a whole is called c-SPLTB. Baseline 0 does not use c-SPLTB or cross phase tokens in MPLTB but replaces them with pure vision transformer and output feature concatenation respectively. Baseline 1 adds the c-SPLTB and Baseline 2 adds the cross phase tokens. A 3-D version of Baseline 2 utilizing 3-D patch embedding is also studied in Baseline 3 to validate the advantage of our 2-D model. The result shown in Fig. 3 demonstrates that our design choice is appropriate. It is worth mentioning that the 2-D structure is prone to redundancy between axial slices and ignores the relation between slices compared with the 3-D structure but gets observably higher accuracy. We suppose the reason is twofold. Most of lesions in our dataset having few slices weakens the redundancy between slices in 2-D pipeline, while the number of slices is still obviously larger than the number of lesions, alleviating the overfitting issue. Furthermore, vision transformers are mostly pretrained in 2-D images, causing poor performance when transferring to 3-D pipeline.

We also evaluate the model performance under different phase combinations by cutting the branch of certain phases. It shows that information from various phases can significantly influence the classification performance. A missing phase can cause an accuracy drop of about 10% and complete four-phase model outperforms single-phase model by nearly 20%. Figure 4 contains average results of phase number and details with all phase combinations can be found in supplementary materials.

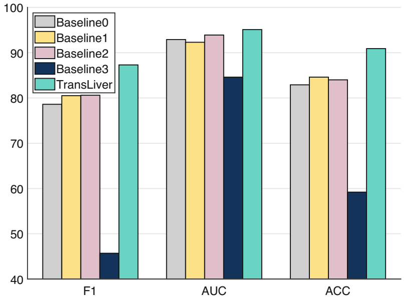

Fig. 3. Ablation study of modules.

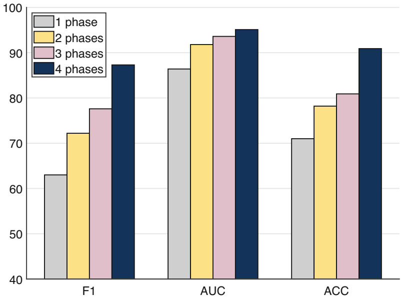

Fig. 4. Ablation study of phase number.

4 Conclusion

We have presented a hybrid architecture for multi-phase liver lesion classification in this paper. The lesion features are extracted by transformer backbone with several auxiliary convolutional modules. Then, we fuse the features from different phases through cross phase tokens to enhance their information exchange.

To handle the issues in realistic cases, we design a pre-processing unit to acquire multi-phase annotated lesions from single-phase annotated ones. We report performance of an overall 90.9% classification accuracy on a four-phase seven-class dataset through quantitative experiments and show obvious improvement compared with SOTA classification methods. In future work, we will extend classification to instance segmentation and provide an end-to-end effective model for liver lesion diagnosis.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (grant numbers 62202426) and Ministry of Science and Technology of China (grant numbers 2022AAA010502).

References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging **38**(8), 1788–1800 (2019)
2. Bilic, P., et al.: The liver tumor segmentation benchmark (LiTS). Med. Image Anal. **84**, 102680 (2023)
3. Chen, X., et al.: A cascade attention network for liver lesion classification in weakly-labeled multi-phase CT images. In: Wang, Q., et al. (eds.) DART/MIL3ID -2019. LNCS, vol. 11795, pp. 129–138. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33391-1_15
4. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, 3–7 May 2021. OpenReview.net (2021)
5. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using GAN for improved liver lesion classification. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 289–293. IEEE (2018)
6. Guo, J., et al.: CMT: Convolutional neural networks meet vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12175–12185 (2022)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Heker, M., Greenspan, H.: Joint liver lesion segmentation and classification via transfer learning. arXiv preprint arXiv:2004.12352 (2020)
9. Isen, J., et al.: Non-parametric combination of multimodal MRI for lesion detection in focal epilepsy. NeuroImage Clin. **32**, 102837 (2021)
10. Kayhan, O.S., Gemert, J.C.: On translation invariance in CNNs: convolutional layers can exploit absolute spatial location. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14274–14285 (2020)
11. Liang, D., et al.: Combining convolutional and recurrent neural networks for classification of focal liver lesions in multi-phase CT images. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 666–675. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_74

12. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
13. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157. IEEE (1999)
14. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. Adv. Neural. Inf. Process. Syst. **34**, 14200–14213 (2021)
15. Raboh, M., Levanony, D., Dufort, P., Sitek, A.: Context in medical imaging: the case of focal liver lesion classification. In: Medical Imaging 2022: Image Processing, vol. 12032, pp. 165–172. SPIE (2022)
16. Romero, F.P., et al.: End-to-end discriminative deep network for liver lesion classification. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1243–1246. IEEE (2019)
17. Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A.: Cancer statistics, 2022. CA: a cancer J. Clin. **72**(1), 7–33 (2022)
18. Subramanian, V., Syeda-Mahmood, T., Do, M.N.: Multimodal fusion using sparse CCA for breast cancer survival prediction. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1429–1432. IEEE (2021)
19. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
20. Wang, W., et al.: Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578 (2021)
21. Wang, W., et al.: PVT v2: improved baselines with pyramid vision transformer. Comput. Visual Med. **8**(3), 415–424 (2022)
22. Wightman, R.: Pytorch image models. https://github.com/rwightman/pytorch-image-models (2019). https://doi.org/10.5281/zenodo.4414861
23. Xu, X., et al.: A knowledge-guided framework for fine-grained classification of liver lesions based on multi-phase ct images. IEEE J. Biomed. Health Inform. **27**(1), 386–396 (2022)
24. Yasaka, K., Akai, H., Abe, O., Kiryu, S.: Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. Radiology **286**(3), 887–896 (2018)