Transformer-Based End-to-End Classification of Variable-Length Volumetric Data

Marzieh Oghbaie[1,2(✉)], Teresa Araújo[1,2], Taha Emre[2], Ursula Schmidt-Erfurth[1], and Hrvoje Bogunović[1,2]

[1] Christian Doppler Laboratory for Artificial Intelligence in Retina, Department of Ophthalmology and Optometry, Medical University of Vienna, Vienna, Austria
{marzieh.oghbaie,hrvoje.bogunovic}@meduniwien.ac.at
[2] Laboratory for Ophthalmic Image Analysis, Department of Ophthalmology and Optometry, Medical University of Vienna, Vienna, Austria

Abstract. The automatic classification of 3D medical data is memory-intensive. Also, variations in the number of slices between samples is common. Naïve solutions such as subsampling can solve these problems, but at the cost of potentially eliminating relevant diagnosis information. Transformers have shown promising performance for sequential data analysis. However, their application for long sequences is data, computationally, and memory demanding. In this paper, we propose an end-to-end Transformer-based framework that allows to classify volumetric data of variable length in an efficient fashion. Particularly, by randomizing the input volume-wise resolution(#slices) during training, we enhance the capacity of the learnable positional embedding assigned to each volume slice. Consequently, the accumulated positional information in each positional embedding can be generalized to the neighbouring slices, even for high-resolution volumes at the test time. By doing so, the model will be more robust to variable volume length and amenable to different computational budgets. We evaluated the proposed approach in retinal OCT volume classification and achieved 21.96% average improvement in balanced accuracy on a 9-class diagnostic task, compared to state-of-the-art video transformers. Our findings show that varying the volume-wise resolution of the input during training results in more informative volume representation as compared to training with fixed number of slices per volume.

Keywords: Optical coherence tomography · 3D volume classification · Transformers

1 Introduction

Volumetric medical scans allow for comprehensive diagnosis, but their manual interpretation is time consuming and error prone [19,21]. Deep learning methods

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 358–367, 2023.
https://doi.org/10.1007/978-3-031-43987-2_35

have shown exceptional performance in automating this task [27], often at medical expert levels [6]. However, their application in the clinical practice is still limited, partially because they require rigid acquisition settings. In particular, variable volume length, i.e. number of slices, is common for imaging modalities such as computed tomography, magnetic resonance imaging or optical coherence tomography (OCT). Despite the advantages of having data diversity in terms of quality and size, automated classification of dense scans with variable input size is a challenge. Furthermore, the 3D nature of medical volumes results in a memory-intensive training procedure when processing the entire volume. To account for this constraint and make the input size uniform, volumes are usually subsampled, ignoring and potentially hiding relevant diagnostic information.

Among approaches for handling variable input size, Multiple Instance Learning (MIL) is commonly used. There, a model classifies each slice or subgroup of slices individually, and the final prediction is determined by aggregating subdecisions via maximum or average pooling [16,17,23], or other more sophisticated fusion approaches [20,24]. However, they often do not take advantage of the 3D aspect of the data. The same problem occurs when stacking slice-wise embeddings [4,11,22], applying self-attention [5] for feature aggregation, or using principal component analysis (PCA) [9] to reduce the variable number of embeddings to a fixed size. As an alternative, recurrent neural networks (RNNs) [18] consider the volume as a sequence of arranged slices or the corresponding embeddings. However, their performance is overshadowed by arduous training and lack of parallelization.

Vision Transformers (ViTs) [8], on the other hand, allow parallel computation and effective analysis of longer sequences by benefiting from multi-head self-attention (MSA) and positional encoding. These components allow to model both local and global dependencies, playing a pivotal role for 3D medical tasks where the order of slices is important [13,14,26]. Moreover, ViTs are more flexible regarding input size. Ignoring slice positional information (bag-of-slices) or using sinusoidal positional encoding enables them to process sequences of arbitrary length with respect to computational resources. However, ViTs with learnable positional embeddings (PEs) have shown better performance [8]. In this case, the only restriction in processing variable length sequences is the number of PEs. Although interpolating the PE sequence helps overcome this restriction, the resultant sequence will not model the exact positional information of the corresponding slices in the input sequence, affecting ViTs performance [2]. Notably, Flexible ViT [2] (FlexiViT) handles patch sequences of variable sizes by randomizing the patch size during training and, accordingly, resizing the embedding weights and parameters corresponding to PEs.

Despite the merits of the aforementioned approaches, three fundamental challenges still remain. First, the model should be able to process inputs with variable volume resolutions, where throughout the paper we refer to the resolution in the dimension across slices (number of slices), and simultaneously capture the size-independent characteristics of the volume and similarities among the constituent slices. The second challenge is the scalability and the ability of the model

to adapt to unseen volume-wise resolutions at inference time. Lastly, the training of deep learning models with high resolution volumes is both computationally expensive and memory-consuming.

In this paper, we propose a late fusion Transformer-based end-to-end framework for 3D volume classification whose local-similarity-aware PEs not only improve the model performance, but also make it more robust to interpolation of PEs sequence. We first embed each slice by a spatial feature extractor and then aggregate the corresponding sequence of slice-wise embeddings with a Feature Aggregator Transformer (FAT) module to capture 3D intrinsic characteristics of the volume and produce a volume-level representation. To enable the model to process volumes with variable resolutions, we propose a novel training strategy, Variable Length FAT (VLFAT), that enables FAT module to process volumes with different resolutions both at training and test times. VLFAT can be trained with a proportionally few #slices, an efficient trait in case of training time/memory constraints. Consequently, even with drastic slice subsampling during training, the model will be robust against extreme PEs interpolation for high-resolution volumes at the test time. The proposed approach is model-agnostic and can be deployed with Transformer-based backbones. VLFAT beats the state-of-the-art performance in retinal OCT volume classification on a private dataset with nine disease classes, and achieves competitive performance on a two-class public dataset.

2 Methods

Our end-to-end Transformer-based volume classification framework (Fig. 1) has three main components: 1) Slice feature extractor (SFE) to extract spatial biomarkers and create a representation of the corresponding slice; 2) Volume feature aggregator (VFA) to combine the slice-level representations into a volume-level representation, and 3) Volume classification. Trained with the proposed strategy, our approach is capable of processing and classifying volumes with varying volume-wise resolutions. Let's consider a full volume $v \in \mathbf{R}^{(N \times W \times H)}$, where (N, H, W) are the #slices, its width and height respectively. The input to the network is a subsampled volume by randomly selecting n slices.

Slice Feature Extractor (SFE). To obtain the slice representations, we use ViT as our SFE due to its recent success in medical interpretation tasks [10]. ViT mines crucial details from each slice and, using MSA and PE, accumulates the collected information in a learnable classification token, constituting the slice-wise embedding. For each slice token, we then add a learnable 1D PE [8] to retain the position of each slice in the volume.

Volume Feature Aggregator (VFA). The output of the previous step is a sequence of slice-wise embeddings, to which we append a learnable volume-level classification token [7]. The resulting sequence of embedding vectors is

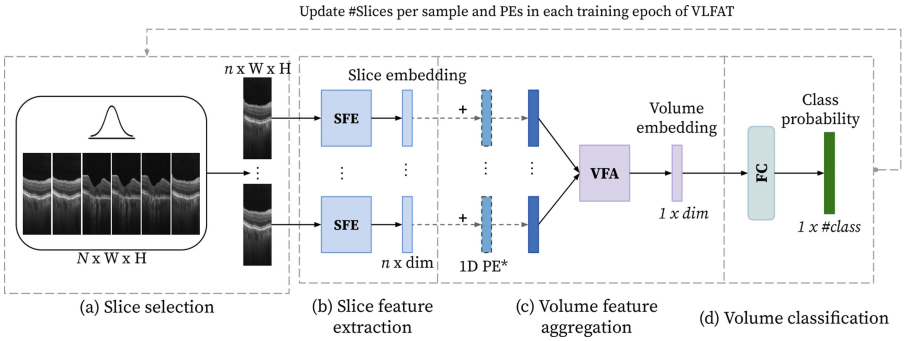

Fig. 1. The overview of the proposed Transformer-based approach for 3D volume classification. The shared SFE processes the input slices, and in line with VLFAT, both #slices and the PEs sequence are updated at each epoch. *1D PE is added to each slice embedding for FAT and VLFAT.

then processed by the FAT module to produce a volume-level embedding. In particular, we propose *VLFAT*, a FAT with enhanced learnable PEs, inspired on FlexiViT [2], where we modify #slices per input volume instead of patch sizes and correspondingly apply PEs interpolation. This allows handling arbitrary volume resolutions, which generally would not be possible except for an ensemble of models of different scales. Specifically, at initialization we set a fixed value, n, for #slices, resulting in PEs sequence with size $(n+1, dim)$, where an extra PE is assigned to the classification token and dim is the dimension of the slice representation. In each training step, we then randomly sample a new value for n from a predefined set and, accordingly, linearly interpolate the PEs sequence (Fig. 1), using the known adjacent PEs [3]. This allows to preserve the similarity between neighboring slices in the volume, sharing biomarkers in terms of locality, and propagating the corresponding positional information. The new PEs are then normalized according to a truncated normal distribution.

Volume Classification. Finally, the volume-level classification token is fed to a Fully Connected (FC) layer, which produces individual class scores. As a loss function, we employ the weighted cross-entropy.

3 Experiments

We tested our model for volume classification of macula-centered retinal OCT scans, where large variation in volume resolution (#B-scans) between samples is very common. For multiclass classification performance metrics, we relied on Balanced Accuracy (BAcc) and one-vs-all Area Under the Receiver Operating Curve (AUROC). The source code is available at: github.com/marziehoghbaie/VLFAT.

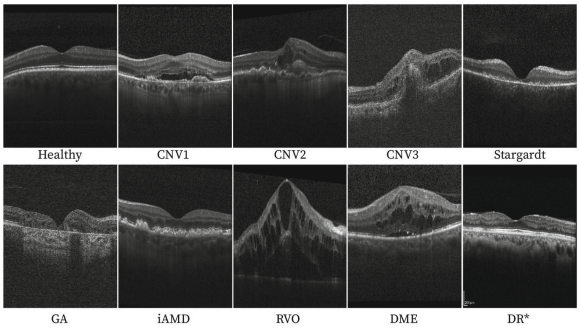

Fig. 2. Samples of central B-scan from all disease classes. *DR is only in OLIVES.

Datasets. We utilized three large retinal OCT volumetric datasets: *Duke* for pre-training all models, and *9C* and *OLIVES* for fine-tuning and testing purposes.

- *Duke:* Public dataset [22] with 269 intermediate age-related macular degeneration (iAMD) and 115 normal patients, acquired with a Bioptigen OCT device with a resolution of 100 B-scans per volume. Volumes were split patient-wise into 80% for training (384 samples) and 20% for validation (77 samples).
- *9C:* Private dataset with 4766 volumes (4711 patients) containing 9 disease classes: iAMD, three types of choroidal neovascularization (CNV1-3), geographic atrophy (GA), retinal vein occlusion (RVO), diabetic macular edema (DME), Stargardt disease, and healthy. Volumes were split patient-wise, for each class, into 70% training (3302 samples), 15% validation (742 samples), and 15% test (722 samples). The OCT volumes were acquired by four devices (Heidelberg Engineering, Zeiss, Topcon, Nidek), exhibiting large variation in #slices. Minimum, maximum, and average #slices per volume were 25, 261, and 81, respectively.
- *OLIVES:* Public dataset [15] with 3135 volumes (96 patients) labeled as diabetic retinopathy (DR) or DME. OCTs were acquired with Heidelberg Engineering device, and have resolutions of either 49 or 97, and were split patient-wise for each class into 80% training (1808 samples), 10% validation (222 samples), and 10% test (189 samples) (Fig. 2).

Comparison to State-of-the-Art Methods. We compared the performance of the proposed method with two state-of-the-art video ViTs (ViViT) [1], originally designed for natural video classification: 1) factorized encoder (FE) ViViT, that models the spatial and temporal dimensions separately; 2) factorised self-attention (FSA) ViViT, that simultaneously computes spatial and temporal interactions. We selected FE and FSA ViViTs as baselines to understand the importance of separate feature extractors and late fusion in our approach. FE ViViT, similar to ours, utilizes late fusion, while FSA ViViT is a slow-fusion model and processes spatiotemporal patches as tokens.

Ablation Studies. To investigate the contribution of SFE module, we deployed ViT and ResNet18 [26], a standard 2D convolutional neural network (CNN) in medical image analysis, with pooling methods as VFA where the quality of slice-wise features is more influential. For VFA, we explored average pooling (AP), max pooling (MP), and 1D convolution (1DConv). As MIL-based baselines, pooling methods can be viable alternatives to VLFAT for processing variable volume resolutions. In addition to learnable PE, we deployed sinusoidal PE (sinPE) and bag-of-slices (noPE) for FAT to examine the effect of positional information.

Robustness Analysis. We investigate the robustness of VFLAT and FAT to PEs sequence interpolation at inference time by changing the volume resolution. To process inputs with volume resolutions different from FAT's and VLFAT's input size, we linearly interpolate the sequence of PEs at the test time. For 9C dataset, we only assess samples with minimum #slices of 128 to better examine the PE's scalability to higher resolutions.

Implementation Details. The volume input size was $25 \times 224 \times 224$ for all experiments except for FSA ViViT where #slices was set to 24 based on the corresponding tublet size of $2 \times 16 \times 16$. During VLFAT training, the #slices varied between $\{5, 10, 15, 20, 25\}$, specified according to memory constraints. We randomly selected slices using a normal distribution with its mean at the central slice position, thus promoting the inclusion of the region near the fovea, essential for the diagnosis of macular diseases. Our ViT configuration is based on ViT-Base [8] with patch size 16×16, and 12 Transformer blocks and heads. For FAT and VLFAT, we set the number of Transformer blocks to 12 and heads to 3. The slice-wise and volume-wise embedding dimension were set to 768. The configuration of ViViT baselines was set according to the original papers [1]. Training was performed using AdamW [12] optimizer with learning rate of 6×10^{-6} with cosine annealing. All models were trained for 600 epochs with a batch size of 8. Data augmentation included random brightness enhancing, motion blur, salt/pepper noise, rotation, and random erasing [28]. The best model was selected based on the highest BAcc on the validation set. All experiments were performed using Pytorch 1.13.0+cu117 and timm library [25] on a server with 1 TB RAM, and NVIDIA RTX A6000 (48 GB VRAM).

4 Results and Discussion

In particular, on large 9C dataset our VLFAT achieved 21.4% and 22.51% BAcc improvement compared to FE ViViT and FSA ViViT, respectively. Incorporating our training strategy, VLFAT, improved FAT's performance by 16.12% on 9C, and 8.79% on OLIVES, which verifies the ability of VLFAT in learning more location-aware PEs, something that is also reflected in the increase of AUROCs (0.96→ 0.98 on 9C dataset and 0.95→ 0.97 on OLIVES). Per-class AUROCs are shown in Table 2. The results show that for most of the classes, our VLFAT has better diagnostic ability and collects more disease-specific clues from the volume.

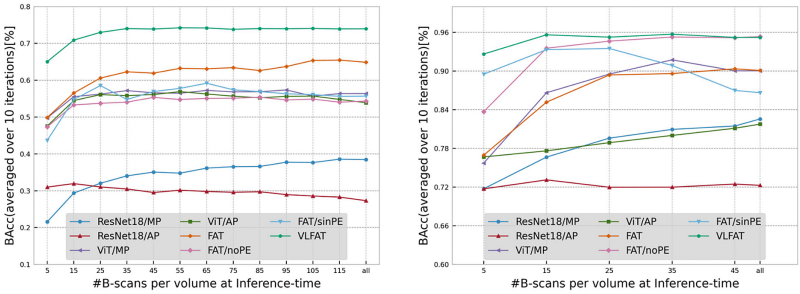

Fig. 3. Robustness analysis of VLFAT and vanilla FAT against PEs interpolation at the test time: (a) 9C dataset; (b) OLIVES

Table 1. Classification performance in terms of balanced accuracy (BAcc) and mean one-vs-all AUROC. FE: Factorised Encoder, FSA: Factorised Self-Attention, SFE: Slice Feature Extractor, VFA: Volume Feature Aggregator.

Method (SFE/VFA)	9C		OLIVES		#slices*
	BAcc	AUROC	BAcc	AUROC	
FE ViViT (baseline) [1]	0.64	0.96	0.93	0.98	25
FSA ViViT (baseline) [1]	0.63	0.95	0.92	0.98	24
ViT/1DConv	0.61	0.94	0.95	0.97	25
ResNet18/AP	0.34	0.73	0.72	0.83	all
ResNet18/MP	0.41	0.87	0.83	0.93	all
ViT/AP	0.59	0.95	0.82	0.97	all
ViT/MP	0.61	0.96	0.90	0.98	all
ViT/FAT (noPE)	0.58	0.93	**0.95**	**0.99**	all
ViT/FAT (sinPE)	0.62	0.93	0.87	0.96	all
ViT/FAT^{+}	0.67	0.96	0.88	0.95	25
ViT/VLFAT (ours)	**0.78**	**0.98**	0.95	0.97	all

Legend: *#slices at the test time;
$^{+}$the input length is fixed in both training and test time

The ablation study (Table 1) showed that each introduced component in the proposed model contributed to the performance improvement. In particular, ViT was shown as a better slice feature extractor compared to ResNet18, particularly on 9C dataset where the differences between disease-related biomarkers are more subtle. Additionally, the poor performance of the pooling methods as compared to FAT and 1DConv, emphasizes the importance of contextual volumetric information, the necessity of a learnable VFA, and the superiority of Transformers over 1DConv. Although, on OLIVES, less complicated VFAs (pooling/1DConv) and FAT (noPE) also achieved comparable results, which can be attributed primarily to DR vs. DME [15] being an easier classification task compared to the diverse disease severity in the 9C dataset. In addition, the competitive advantage

Table 2. Per-class classification performance (one-vs-all AUROC) on 9C dataset.

Method(SFE/VFA)	CNV1	CNV2	CNV3	DME	GA	Healthy	iAMD	RVO	Stargardt
FE ViViT [1]	0.93	0.91	0.95	0.94	0.99	0.95	0.92	0.95	0.99
FSA ViViT [1]	0.94	0.91	0.92	0.92	**1.0**	0.94	0.93	0.94	0.99
ViT/1DConv	0.88	0.92	0.94	0.91	0.98	0.92	0.92	0.92	**1.0**
ResNet18/AP	0.68	0.63	0.58	0.75	0.81	0.75	0.75	0.76	0.97
ResNet18/MP	0.78	0.77	0.79	0.87	0.91	0.84	0.84	0.84	0.91
ViT/AP	0.9	0.81	0.92	0.93	0.98	0.94	0.93	0.94	0.99
ViT/MP	0.95	0.85	0.95	0.94	0.98	0.94	0.93	0.96	0.99
ViT/FAT (noPE)	0.9	0.87	0.89	0.89	0.98	0.91	0.9	0.94	0.99
ViT/FAT (sinPE)	0.94	**0.93**	0.93	0.88	0.97	0.91	0.89	0.93	0.99
ViT/FAT	0.93	0.82	0.97	0.94	0.99	0.95	0.94	0.95	0.99
ViT/VLFAT (ours)	**0.98**	0.92	**0.98**	**0.98**	**1.0**	**0.99**	**0.98**	**0.98**	**1.0**

of VLFAT in handling different resolutions was not fully exploited in OLIVES since the large majority of cases had the same #slices. On 9C, however, the comparison of positional encoding strategies demonstrated that although ignoring PEs and sinusoidal approach provide deterministic predictions, the importance of learnable PEs in modeling the anatomical order of slices in the volume is crucial. The robustness analysis is shown in Fig. 3. VLFAT was observed to have more scalable and robust PEs when the volume-wise resolutions at the test time deviated from those used during training. This finding highlights the VLFAT's potential for resource-efficient training and inference.

5 Conclusions

In this paper, we propose an end-to-end framework for 3D volume classification of variable-length scans, benefiting from ViT to process volume slices and FAT to capture 3D information. Furthermore, we enhance the capacity of PE in FAT to capture sequential dependencies along volumes with variable resolutions. Our proposed approach, VLFAT, is more scalable and robust than vanilla FAT at classifying OCT volumes of different resolutions. On a large-scale retinal OCT datasets, our results indicate that this effective method performs in the majority of cases better than other common methods for volume classification.

Besides its applicability for volumetric medical data analysis, our VFLAT has potential to be applied on other medical tasks including video analysis (e.g. ultrasound videos) and high-resolution imaging, as is the case in histopathology. Future work would include adapting VLFAT to ViViT models to make them less computationally expensive. Furthermore, PEs in VLFAT could be leveraged for improving the visual interpretation of decision models by collecting positional information about the adjacent slices sharing anatomical similarities.

Acknowledgements. This work was supported in part by the Christian Doppler Research Association, Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development, and Heidelberg Engineering.

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: a video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6836–6846 (2021)
2. Beyer, L., et al.: Flexivit: one model for all patch sizes. arXiv preprint arXiv:2212.08013 (2022)
3. Blu, T., Thévenaz, P., Unser, M.: Linear interpolation revitalized. IEEE Trans. Image Process. **13**(5), 710–719 (2004)
4. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) ACCV 2016. LNCS, vol. 10112, pp. 87–103. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54184-6_6
5. Das, V., Prabhakararao, E., Dandapat, S., Bora, P.K.: B-scan attentive CNN for the classification of retinal optical coherence tomography volumes. IEEE Signal Process. Lett. **27**, 1025–1029 (2020)
6. De Fauw, J., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. **24**(9), 1342–1350 (2018)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Fang, L., Wang, C., Li, S., Yan, J., Chen, X., Rabbani, H.: Automatic classification of retinal three-dimensional optical coherence tomography images using principal component analysis network with composite kernels. J. Biomed. Opt. **22**(11), 116011–116011 (2017)
10. He, K., et al.: Transformers in medical image analysis: a review. Intell. Med. (2022)
11. Howard, J.P., et al.: Improving ultrasound video classification: an evaluation of novel deep learning methods in echocardiography. J. Med. Artif. Intell. **3** (2020)
12. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
13. Peiris, H., Hayat, M., Chen, Z., Egan, G., Harandi, M.: A robust volumetric transformer for accurate 3D tumor segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13435, pp. 162–172. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_16
14. Playout, C., Duval, R., Boucher, M.C., Cheriet, F.: Focused attention in transformers for interpretable classification of retinal images. Med. Image Anal. **82**, 102608 (2022)
15. Prabhushankar, M., Kokilepersaud, K., Logan, Y.Y., Corona, S.T., AlRegib, G., Wykoff, C.: Olives dataset: Ophthalmic labels for investigating visual eye semantics. arXiv preprint arXiv:2209.11195 (2022)
16. Qiu, J., Sun, Y.: Self-supervised iterative refinement learning for macular oct volumetric data classification. Comput. Biol. Med. **111**, 103327 (2019)

17. Rasti, R., Rabbani, H., Mehridehnavi, A., Hajizadeh, F.: Macular OCT classification using a multi-scale convolutional neural network ensemble. IEEE Trans. Med. Imaging **37**(4), 1024–1034 (2017)
18. Romo-Bucheli, D., Erfurth, U.S., Bogunović, H.: End-to-end deep learning model for predicting treatment requirements in neovascular AMD from longitudinal retinal OCT imaging. IEEE J. Biomed. Health Inform. **24**(12), 3456–3465 (2020)
19. Semivariogram and semimadogram functions as descriptors for AMD diagnosis on SD-OCT topographic maps using support vector machine. Biomed. Eng. Online **17**(1), 1–20 (2018)
20. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
21. Singh, S.P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., Gulyás, B.: 3D deep learning on medical images: a review. Sensors **20**(18), 5097 (2020)
22. Sun, Y., Zhang, H., Yao, X.: Automatic diagnosis of macular diseases from OCT volume based on its two-dimensional feature map and convolutional neural network with attention mechanism. J. Biomed. Opt. **25**(9), 096004–096004 (2020)
23. de Vente, C., González-Gonzalo, C., Thee, E.F., van Grinsven, M., Klaver, C.C., Sánchez, C.I.: Making AI transferable across oct scanners from different vendors. Invest. Ophthalmol. Visual Sci. **62**(8), 2118–2118 (2021)
24. Wang, J., Cherian, A., Porikli, F., Gould, S.: Video representation learning using discriminative pooling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1149–1158 (2018)
25. Wightman, R.: Pytorch image models (2019) https://doi.org/10.5281/zenodo.4414861. https://github.com/rwightman/pytorch-image-models
26. Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A.: Context-aware transformers for spinal cancer detection and radiological grading. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13433, pp. 271–281. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_26
27. Wulczyn, E., et al.: Deep learning-based survival prediction for multiple cancer types using histopathology images. PLoS ONE **15**(6), e0233678 (2020)
28. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 13001–13008 (2020)