On the Relevance of Temporal Features for Medical Ultrasound Video Recognition

D. Hudson Smith[1](✉), John Paul Lineberger[1], and George H. Baker[2]

[1] Clemson University, Clemson, SC 29634, USA
{dane2,jplineb}@clemson.edu
[2] Medical University of South Carolina, Charleston, SC 29425, USA
baker@musc.edu

Abstract. Many medical ultrasound video recognition tasks involve identifying key anatomical features regardless of when they appear in the video suggesting that modeling such tasks may not benefit from temporal features. Correspondingly, model architectures that exclude temporal features may have better sample efficiency. We propose a novel multi-head attention architecture that incorporates these hypotheses as inductive priors to achieve better sample efficiency on common ultrasound tasks. We compare the performance of our architecture to an efficient 3D CNN video recognition model in two settings: one where we expect not to require temporal features and one where we do. In the former setting, our model outperforms the 3D CNN - especially when we artificially limit the training data. In the latter, the outcome reverses. These results suggest that expressive time-independent models may be more effective than state-of-the-art video recognition models for some common ultrasound tasks in the low-data regime. Code is available at https://github.com/MedAI-Clemson/pda_detection.

Keywords: Ultrasound · Video · Sample Efficiency · Attention

1 Introduction and Related Work

Ultrasound (US) is one of the most common imaging techniques in medical practice, with applications to fetal imaging, cardiac imaging, sports medicine, and more. With the rise of US for routine clinical care, there is a growing interest in applying computer vision techniques to automate or enhance the analysis of US imagery [13]. Many US examinations involve the collection of video clips showing different anatomical regions. The medical imaging community is in the early stages of applying techniques from the video recognition community to US recognition tasks. These applications face several challenges arising from the

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43895-0_70.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 744–753, 2023.
https://doi.org/10.1007/978-3-031-43895-0_70

nature of US as an imaging modality, differences between US imagery and natural imagery, and the lack of large representative datasets. To make matters worse, the collection of large medical datasets is often unethical or prohibitively costly. There is, therefore, a significant need for efficient methods that can produce high levels of performance using the minimum number of samples. In this work, we propose an efficient US video recognition architecture that takes advantage the nature of common US recognition tasks.

To design an efficient US recognition architecture, it is necessary to consider the space of US recognition tasks and evaluate the algorithmic structures needed to efficiently capture the semantics in those settings. We posit that many of these tasks amount to the identification of specific visual characteristics at key moments in the clip. The identification of the *standard plane* in fetal head US depends on recognizing key structures in fetal brain tissue [3,19]; the quality assessment of FAST clips [24] relies on the ability to recognize that key organs and other structures have been visualized in the clip; view identification relies on recognizing orientation of the anatomical structures in relation to one another [8,11]; and the quantification of heart function requires measurement of ventricular volumes at two key moments in the cardiac cycle [22]. Based on these observations, we propose a novel *US Video Network* (USVN) that treats frames as independent and unordered. USVN constructs expressive video representations by combining information from multiple frames using a novel multi-head attention mechanism. We demonstrate a setting in which USVN yields better performance and far better sample efficiency than a competing model that includes temporal features. We also demonstrate that, in a setting where temporal dependence is important, USVN lags behind the competing model. These contrasting outcomes demonstrate the importance of tailoring the model architecture to the structure of the US recognition task in data-constrained settings.

A large body of work has addressed video recognition tasks, including object tracking [14], temporal action localization [28], captioning [1], action recognition [30], and many others. Driven by the availability of large human action datasets, the field of action recognition has focused on the need to capture expressive spatiotemporal features. This has led to the development of two-stream networks using optical flow [21], the use of 3D convolutional networks [10,25], and, of course, the use of transformer-based architectures [15,18]. Our main point of departure with these methods is the importance placed upon temporal features. We posit that temporal features are not relevant in some common US tasks and that excluding these features leads to better sample efficiency. To explore this idea, we assume temporal independence *a priori*, placing our problem formulation in the format of a Multi-instance Learning (MIL) task.

Multi-instance learning (MIL) describes the situation where labels apply to bags of instances rather than to individual instances. Instances within a bag are assumed to be unordered and, conditional on the bag label, independent from one another [2]. Under our assumption that all video frames can be treated independently, video recognition can be viewed as MIL where the bag is the video, and the instances are the frames. MIL has a long history of applications to video recognition that predates deep learning [5,6,23,29]. In the classical

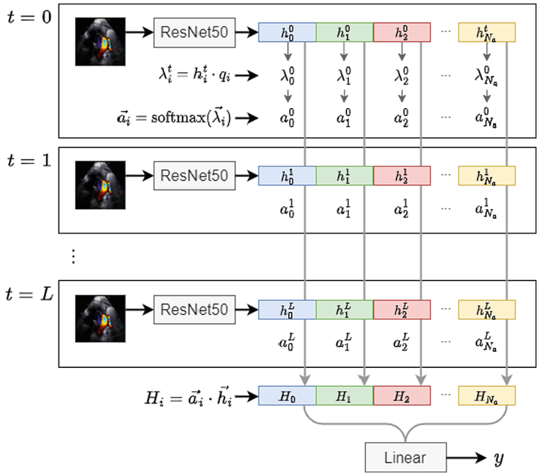

Fig. 1. Proposed video-recognition architecture. Frame representations from ResNet50 are partitioned into N_a equal-sized vectors, h_i^t, represented by the colored boxes at each time step. These are compared by dot product with global query vectors q_i to compute attention weights a_i^t. The video-level representation, H_i, is the attention-weighted sum of the partitions across frames. y is the video-level prediction.

formulation of MIL it is assumed that instances have unobserved labels, and the task is to extract these as latent variables and aggregate them to predict the bag-level label. In their paper *Attention-based deep multiple instance learning* Ilse, Tomczak, and Welling [9] depart from this classical perspective by aggregating embeddings rather than instance labels. We take a similar approach. Unlike their work, however, we use multiple attention heads focused on different subspaces of the image-level embeddings, with their work as a special case of ours. To our knowledge, we are the first to introduce a MIL formalism using multiple attention heads in this way.

There is growing interest in applying action recognition techniques to medical US video with applications to fetal [3,19,20], abdominal [11,24], and cardiac [4,8,17,22] US. Most existing applications make MIL assumptions but only apply a fixed pooling function to frame-level labels. Howard et al. [8] apply a range of techniques, including average pooling, two-stream networks, and 3D convolutions to identifying cardiac views. They conclude that two-stream networks yield the best performance. The authors do not test any methods that adaptively pool frame information in a time-independent manner. Lei et al. [12] specifically consider the detection of Patent Ductus Arteriosus (PDA). They make MIL assumptions by applying the video-level label to the individual frames and training a 2D CNN to estimate these noisy labels. Video-level labels are generated by applying a decision threshold to the frame-level predictions and then voting with equal weight across frames. Ouyang et al. [16] use 3D convolutions, specifically the R(2+1)D architecture [25], to predict ejection fraction from cardiac

US obtaining human-level performance. They do not assess the performance of any time-independent methods. Among these examples, we see a divide between methods that have no ability to adaptively weight different frames and those that can express arbitrary spatiotemporal features. We fill this gap by proposing a time-independent method that adaptively pools information from different moments in time.

2 Proposed Method

2.1 USVN

Architecture. Our video recognition architecture, shown in Fig. 1, pools information across frames using a multi-head attention mechanism. Like the attention mechanism in the transformer architecture [26], we compute attentions over subspaces of the frame-level representations. We hypothesize that US video recognition requires the detection of distinct visual features that may appear at different points of time in the video. The individual attention heads can function as detectors of these features. Unlike ordinary multi-head attention, the subspaces are not compared with other frames in the sequence but with a set of global query vectors inferred during training. The use of global query vectors arises from our inductive prior that the recognition task amounts to locating key pieces of information at any point in the sequence, and the inferred query vectors are representations of that key information.

Frames are first embedded into 2048-dimensional vectors using a CNN encoder. This encoder is initialized via ImageNet pretraining and fine-tuned during training. Rather than learn N_a projections from scratch for the attention weighting, we simply partition the frame representations into N_a vectors h_i^t each of size $d_a = 2048/N_a$ and rely on the final convolutional layers of the CNN to adapt. We then compute the un-normalized attention scores via dot product with the global query vectors: $\lambda_i^t = h_i^t \cdot q_i$. The resulting scores are normalized resulting in N_a attention vectors, $\boldsymbol{a_i} = \text{softmax}(\boldsymbol{\lambda_i})$, where the arrow notation represents vectorization in time. The video-level representation from the i^{th} head is then simply $H_i = \boldsymbol{a_i} \cdot \boldsymbol{h_i}$, and the full video representation is the concatenation $H = \text{concat}([H_1, H_2, \ldots, H_{N_a}])$. The video-level prediction can then be computed using a shallow fully-connected network, $y = f(H)$.

Augmentation by Frame Sampling. Because USVN treats all frames independently, it is not necessary to use contiguous spans of frames during training. Instead, we randomly sample fixed-size sets of frames from each video. This can have a regularizing effect by using novel frames for each training epoch. During evaluation we use all video frames. We accommodate the varying numbers of frames in each video by zero padding and masked attention.

Model Interpretability. We identify prototype frames for each attention head. These prototypes produce embedding subspace vectors h_i^t that are closely aligned with the corresponding query vector q_i. These prototype images can then be qualitatively evaluated by the clinical specialist (see Supplemental Material).

2.2 Benchmark Implementations

A simple and common approach for video recognition is to use fixed pooling functions to aggregate the frame-level representations across time, treating each element of the representation as a channel. We evaluate this approach using max and average pooling functions. Our attention-based method can implement average pooling by assigning equal weight to all frames for each attention head. Neglecting potential optimization challenges, this suggests that attention-based pooling should be at least as good as average pooling. On the other hand, our model can only approximate max pooling in the $N_a = 2048$ case by assigning very large, positive values to the single-element query vectors causing the attentions to become sharply concentrated at one time step. However, this solution pushes the softmax over time into regions with very small gradients. We conclude that max pooling can learn video representations that cannot be expressed by USVN (and vice versa).

R(2+1)D is a 3D CNN video recognition architecture that decomposes the spatial and temporal convolution into two successive steps [25]. First, a 2D convolution is applied over space then a 1D convolution is applied over time. Compared to its 3D ResNet counterparts on Sports-1M and Kinetics datasets, R(2+1)D is a very capable model that can learn complex features while having the same number of parameters in a more data-efficient way. We choose to benchmark against this architecture due to its efficiency and because this is the architecture used by Ouyang et al. to achieve human-level performance on the EchoNet-Dynamic US dataset [16].

3 Experimental Results

3.1 Datasets

Patent Ductus Arteriosus (PDA). PDA is an opening between the aorta and pulmonary artery that, in severe cases, can cause heart failure shortly after birth. Ultrasound imaging is the primary diagnostic tool for detecting and characterizing PDA. Specifically, doppler US imaging can visualize the motion of the blood through the PDA opening. This motion appears as a characteristic blob of color in the region of the PDA. Physicians are trained to recognize the color and shape of the blob as well as where it appears in relation to other visible anatomy. Superficially, this recognition task makes no reference to the dynamics of the video. We therefore expect that temporal features are not required for accurate PDA recognition. For this dataset we train USVN to predict whether or not an image indicates the presence of PDA. The model output, y, is therefore a single number interpreted as the log-odds of PDA.

We retrospectively collected a set of 1,145 doppler US clips from 165 distinct examinations involving 66 distinct patients. Each clip was labeled to indicate the presence (661 clips) or absence (484 clips) of PDA. Patients were divided into training (44), validation (11), and test (11) sets with stratification on the presence of PDA. These sets contained 755, 118, and 272 videos, respectively.

The large variation in the number of videos in the validation and test sets results from the fact that patients have a variable number of examinations ranging from 1 to 10.

Table 1. Model performance comparison. EchoNet benefits from modeling temporal features; PDA does not. Performance is measured on the test set.

Model	PDA (ROC AUC)	EchoNet (r^2)
R(2+1)D	0.816	**0.822**
Average Pool	0.837	0.679
Max Pool	0.835	0.657
USVN (Ours)	**0.855**	0.765

EchoNet-Dynamic. The Echonet Dynamic dataset consists of 10,030 apical-4 chamber echocardiograms downsampled to 112×112. Each study has clinical measurements: ejection fraction (EF), end systolic volume (ESV), and end diastolic volume (EDV). EF is commonly used to assess cardiac function and is computed from ESV and EDV as

$$\mathrm{EF} = 1 - \mathrm{ESV}/\mathrm{EDV}. \tag{1}$$

The echocardiograms were obtained by registered sonographers and level 3 echocardiographers. For each of these videos, a masking and cropping transformation was performed to remove text and instrument information from the scanning area.

For this dataset, we train USVN to predict ejection fraction. Rather than predict EF directly, we output a tuple of real numbers (y_1, y_2) and insert them in place of ESV and EDV in Eq. (1). This choice is motivated by the knowledge that ESV and EDV are determined from different phases of the cardiac cycle. We speculate that decomposing EF into ESV and EDV effectively linearizes the estimation of EF as a function of the video representation H with different attention heads responsible for estimating ESV and EDV.

3.2 Results

Model Performance. Table 1 summarizes the performance of USVN and our benchmark implementations on the PDA and EchoNet tasks. For PDA classification, we evaluate using the area under the ROC curve (ROC AUC). For EchoNet, we use the percent of variance explained (r^2). USVN results are based on $N_a = 16$ and $N_a = 128$ for PDA and EchoNet, respectively, based on a hyperparameter search (see Supplemental Material). For the PDA dataset, we expected that temporal features are not beneficial and, indeed, we see that R(2+1)D performs worse than all other methods, likely due to the unneeded capacity in the temporal convolutions and the relatively small size of the PDA dataset. USVN leads to a small benefit over average and max pooling for this task. The EchoNet task does benefit from modeling temporal features as indicated by R(2+1)D obtaining the highest score. However, USVN significantly outperforms the fixed pooling methods and is surprisingly close to R(2+1)D. This suggests that temporal features play a relatively small part in explaining the variability in the EchoNet dataset.

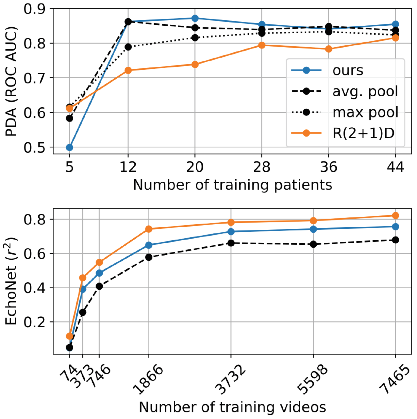

Fig. 2. Dependence on number of patients in training set for PDA classification (*top*) and EchoNet ejection fraction prediction (*bottom*). For PDA, we show patients, rather than videos along the x-axis due to the non-independence of videos from the same patient. For EchoNet, we omit the "max pool" variant because it failed to obtain positive r^2 values for several points along the x-axis. Performance is measured on the test set.

Sample Efficiency. In Fig. 2 we evaluate the sample efficiency of USVN by artificially limiting the amount of training data. In the case of PDA, we downsample the number of patients because videos from a single patient are correlated with one another. For EchoNet, we downsample the number of videos. In both cases, we use the full validation and test sets to better isolate variation due to limited training data from variation due to model selection and evaluation.

For PDA, R(2+1)D underperforms the time-independent methods, and the gap is larger for smaller numbers of training patients (see Fig. 2, top panel). Surprisingly, USVN and average pooling have very similar performance across samples and saturate for a small subset of the available patients. R(2+1)D needs all available patients to approach a similar level of performance. This result aligns with our expectation that the inductive prior of time independence can yield sample efficiency benefits when applied to the appropriate task.

R(2+1)D outperforms the time-independent models across all samples for the EchoNet task (see Fig. 2, bottom panel). Despite being a much simpler architecture than R(2+1)D and approaching similar levels of performance, USVN does not exhibit any sample efficiency benefits in the low-data regime for the EchoNet task. Solving the EchoNet task with spatial features alone may require more adaptation of the pretrained encoder than is required when solving with temporal features. For instance, it may be possible through extensive adaptation of the encoder network to recognize the visual characteristics associated with the end of diastole. However, the end of diastole may also manifest as, for example, an extremum in time of some visual characteristic. A model with access to temporal features such as R(2+1)D may be able to capture such an extremum with relatively little adaptation of the pretrained network.

3.3 Implementation Details

For the fixed pooling methods and USVN, we use an ImageNet-pretrained ResNet50 image encoder provided through the `timm` library [27]. We train using the `timm` implementation of the AdamP optimizer [7] with $\beta_{1,2} = 0.9, 0.999$, weight decay of 0.001, batch size of 20 clips, and initial learning rates of $3 \cdot 10^{-5}$ and 0.001 for PDA and EchoNet, respectively. We sample 32 frames per clip during training. We reduce the learning rate by a factor of 10 after 3 epochs with no improvement of the validation loss, and we terminate training after ten consecutive epochs of no improvement. We use 50% dropout on the inputs to the linear layer for each dataset.

To reproduce the results of R(2+1)D on Echonet Dynamic Dataset by Ouyang et al. [16], we cloned their github repo and re-ran their experiments with their best found hyperparameters. Our training runs show similar, if not better, results than stated in the original work. To adapt the model for PDA classification, we modified their data loader, training script, and the R(2+1)D model to allow PDA images. We also removed the manual bias term initialization, left over from predicting ejection fraction on the fully connected linear layer, and initialize it randomly instead. Finally, we replaced MSE loss with binary cross entropy with logits in the training loop. Every run was done for 45 epochs with a batch size of 20 for Echonet Dynamic dataset and 10 for PDA dataset. Model saving occurred for every epoch that showed improvement to the validation loss.

4 Conclusions and Discussion

The field of video recognition has been driven by large human action recognition datasets. Unlike videos of human actions, the accurate recognition of medical ultrasound images often only requires identifying key pieces of information at any point in the video and does not make reference to the sequence of events. The contrast between results for the PDA task (where USVN excels) and the EchoNet task (where USVN suffers) demonstrates the importance of tailoring the model architecture to the task at hand in data-constrained settings. Our results suggest that models developed for human action recognition are not optimal in some practical scenarios involving medical ultrasound and that models that assume temporal independence have better sample efficiency. We introduce an architecture, USVN, that is tailored to the medical ultrasound context and demonstrate a situation where the inductive prior of time independence leads to significant sample efficiency benefits. We also present a situation where temporal features are relevant and show that, even for very small datasets, USVN produces no efficiency benefits. Practitioners of deep learning who work with medical ultrasound in the low-data regime should take care to match the architecture choice to the nature of the recognition task.

Acknowledgement. We thank Clemson University for their generous allotment of compute time on the Palmetto Cluster.

References

1. Amirian, S., Rasheed, K., Taha, T.R., Arabnia, H.R.: Automatic image and video caption generation with deep learning: a concise review and algorithmic overlap. IEEE Access **8**, 218386–218400 (2020)
2. Carbonneau, M.A., Cheplygina, V., Granger, E., Gagnon, G.: Multiple instance learning: a survey of problem characteristics and applications. Pattern Recogn. **77**, 329–353 (2018)
3. Chen, H., et al.: automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 507–514. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24553-9_62
4. Dezaki, F.T., et al.: Deep residual recurrent neural networks for characterisation of cardiac cycle phase from echocardiograms. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 100–108. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_12
5. Ding, X., Li, B., Hu, W., Xiong, W., Wang, Z.: Horror video scene recognition based on multi-view multi-instance learning. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7726, pp. 599–610. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37431-9_46
6. Gu, Z., Mei, T., Hua, X.S., Tang, J., Wu, X.: Multi-layer multi-instance learning for video concept detection. IEEE Trans. Multimedia **10**(8), 1605–1616 (2008)
7. Heo, B., et al.: Adamp: slowing down the slowdown for momentum optimizers on scale-invariant weights. arXiv preprint arXiv:2006.08217 (2020)
8. Howard, J.P., et al.: Improving ultrasound video classification: an evaluation of novel deep learning methods in echocardiography. J. Med. Artif. Intell. **3** (2020)
9. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136. PMLR (2018)
10. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2012)
11. Kornblith, A.E., et al.: Development and validation of a deep learning strategy for automated view classification of pediatric focused assessment with sonography for trauma. J. Ultrasound Med. **41**(8), 1915–1924 (2022)
12. Lei, H., Ashrafi, A., Chang, P., Chang, A., Lai, W.: Patent ductus arteriosus (PDA) detection in echocardiograms using deep learning. Intell.-Based Med. **6**, 100054 (2022)

13. Liu, S., et al.: Deep learning in medical ultrasound analysis: a review. Engineering **5**(2), 261–275 (2019)
14. Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., Kim, T.K.: Multiple object tracking: a literature review. Artif. Intell. **293**, 103448 (2021)
15. Mazzia, V., Angarano, S., Salvetti, F., Angelini, F., Chiaberge, M.: Action transformer: a self-attention model for short-time pose-based human action recognition. Pattern Recogn. **124**, 108487 (2022)
16. Ouyang, D., et al.: Video-based AI for beat-to-beat assessment of cardiac function. Nature **580**(7802), 252–256 (2020)
17. Patra, A., Huang, W., Noble, J.A.: Learning spatio-temporal aggregation for fetal heart analysis in ultrasound video. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 276–284. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_32
18. Plizzari, C., Cannici, M., Matteucci, M.: Spatial temporal transformer network for skeleton-based action recognition. In: Del Bimbo, A., et al. (eds.) ICPR 2021. LNCS, vol. 12663, pp. 694–701. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68796-0_50
19. Pu, B., Li, K., Li, S., Zhu, N.: Automatic fetal ultrasound standard plane recognition based on deep learning and IIoT. IEEE Trans. Industr. Inf. **17**(11), 7771–7780 (2021)
20. Rasheed, K., Junejo, F., Malik, A., Saqib, M.: Automated fetal head classification and segmentation using ultrasound video. IEEE Access **9**, 160249–160267 (2021)
21. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Adv. Neural Inform. Process. Syst. **27** (2014)
22. Sofka, M., Milletari, F., Jia, J., Rothberg, A.: Fully convolutional regression network for accurate detection of measurement points. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 258–266. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_30
23. Stikic, M., Schiele, B.: Activity recognition from sparsely labeled data using multi-instance learning. In: Choudhury, T., Quigley, A., Strang, T., Suginuma, K. (eds.) LoCA 2009. LNCS, vol. 5561, pp. 156–173. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01721-6_10
24. Taye, M., Morrow, D., Cull, J., Smith, D.H., Hagan, M.: Deep learning for fast quality assessment. J. Ultrasound Med. **42**(1), 71–79 (2022)
25. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018)
26. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inform. Process. Syst. **30** (2017)
27. Wightman, R.: Pytorch image models. https://github.com/rwightman/pytorch-image-models (2019). https://doi.org/10.5281/zenodo.4414861
28. Xia, H., Zhan, Y.: A survey on temporal action localization. IEEE Access **8**, 70477–70487 (2020)
29. Yang, J., Yan, R., Hauptmann, A.G.: Multiple instance learning for labeling faces in broadcasting news video. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, pp. 31–40 (2005)
30. Zhang, H.B., et al.: A comprehensive survey of vision-based human action recognition methods. Sensors **19**(5), 1005 (2019)