Self-supervised Contrastive Video-Speech Representation Learning for Ultrasound

Jianbo Jiao[1], Yifan Cai[1], Mohammad Alsharid[1], Lior Drukker[2], Aris T. Papageorghiou[2], and J. Alison Noble[1]

[1] Department of Engineering Science, University of Oxford, Oxford, UK
jianbo.jiao@eng.ox.ac.uk
[2] Nuffield Department of Women's & Reproductive Health, University of Oxford, UK

Abstract. In medical imaging, manual annotations can be expensive to acquire and sometimes infeasible to access, making conventional deep learning-based models difficult to scale. As a result, it would be beneficial if useful representations could be derived from raw data without the need for manual annotations. In this paper, we propose to address the problem of self-supervised representation learning with multi-modal ultrasound video-speech raw data. For this case, we assume that there is a high correlation between the ultrasound video and the corresponding narrative speech audio of the sonographer. In order to learn meaningful representations, the model needs to identify such correlation and at the same time understand the underlying anatomical features. We designed a framework to model the correspondence between video and audio without any kind of human annotations. Within this framework, we introduce cross-modal contrastive learning and an affinity-aware self-paced learning scheme to enhance correlation modelling. Experimental evaluations on multi-modal fetal ultrasound video and audio show that the proposed approach is able to learn strong representations and transfers well to downstream tasks of standard plane detection and eye-gaze prediction.

Keywords: Self-supervised · Representation learning · Video-audio.

1 Introduction

Deep learning-based medical image analysis approaches rely heavily on annotated training data, which limits the progress in medical image analysis if a large dataset has to be manually annotated for every new task. Extracting meaningful representations directly from unlabelled data is therefore an important and interesting sub-topic in learning-based medical image analysis.

Several approaches in the literature have been explored to deal with the problem of learning from unlabelled data, which is usually termed as "self-supervised representation learning" (or unsupervised learning in some works). The common practice is to pre-train a model on unlabelled data according to a pretext task and then to fine tune the model with some specific target tasks to evaluate the learned representations. Typical pretext tasks include colourisation [19], and

rotation prediction [5] for images; and tracking [16], temporal ordering [10] for videos. Some recent studies propose to learn representations by contrastive predictive coding [7,11] or contrastive learning [4,6], and showed a powerful learning capability. There are also works have considered medical images, *e.g.*, predicting the distance between patches [15], Rubik's cube recovery [20] and anatomy-aware joint reasoning [9]. However, the above-mentioned approaches are single modality. Some recent approaches have investigated learning from natural audio and video modalities [1, 12], where the pretext task is designed as video-audio alignment. In this case the audio and video are assumed to be in dense correspondence. Such multi-modal learning has not been explored for medical data before. Audio data rarely exists for medical images and even when available, it is mostly narrative diagnosis/interpretation speech, which has a sparse correlation with the visual data, making the task rather challenging.

In this paper, we propose to address the problem of self-supervised cross-modal representation learning for ultrasound video with corresponding narrative speech audio, both of which are captured on-the-fly without any manual annotations. Unlike other medical imaging modalities that are with clear anatomical structures (*e.g.*, MRI and CT), ultrasound video is much more difficult to interpret by eye for humans although experts are adept at interpreting anatomy in the acoustic patterns. As a result, learning representations automatically from unlabelled ultrasound data is rather challenging. On the other hand, in our case we have synchronised narrative speech from the sonographer accompanied with the ultrasound video. The basic assumption here is that by leveraging cross-modal correlations, a useful representation can be learned. To this end, we propose to learn the anatomical representations from ultrasound video-speech data by identifying the affinity (*i.e.*, association strength) between the two modalities. Specifically, we first randomly sample video-speech samples from our dataset, from which positive and negative pairs are generated. Unlike prior works that use straightforward simple training pairs, in this work we instead leverage hard-negative as well as hard-positive pairs for training, in order to force the model to *learn harder*. Additionally, we further introduce cross-modal contrastive learning to encourage a positive pair to have strong affinity while a negative pair will have weaker affinity in a projected shared latent space. In our diagnostic narrative speech audio, we observe that background noise (*e.g.*, button clicks, air conditioning) and unrelated conversation (*e.g.*, about the weather, travel *etc.*) exist in the raw data, which degrades the representation learning. To mitigate this problem, we propose an affinity-aware self-paced learning curriculum over the representation learning process.

To evaluate the proposed self-supervised learning framework, we consider two ultrasound-related downstream tasks: standard plane detection and eye-gaze saliency prediction. Experimental results presented show that the proposed approach significantly improves the performance of downstream tasks without referring to any manual annotations. The experiments also reveal that by leveraging speech data, useful representations can be learned when transferring to other tasks, outperforming single-modal learning methods.
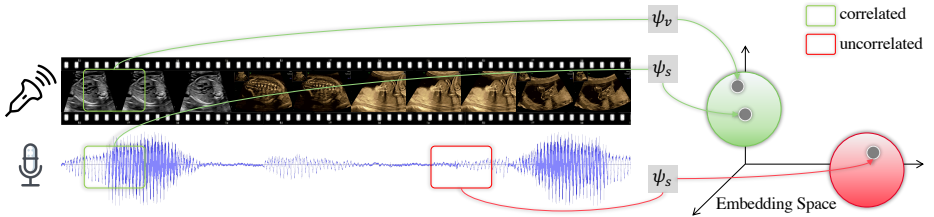
Fig. 1. Main idea of the proposed self-supervised video-speech representation learning framework. A model is trained to identify whether a sampled video-speech pair is anatomically correlated, and at the same time encourage the projected embeddings from correlated pair to lie on the same anatomical sphere (*e.g.*, the green one).

The main contributions of this paper are summarised as follows:

– We propose, to our knowledge, the first self-supervised video-speech representation learning approach for ultrasound data.
– We introduce cross-modal contrastive learning and affinity-aware self-paced learning for ultrasound video-speech cross-modal representation learning.
– The proposed approach is demonstrated to be effective for two downstream tasks.

2   Method

In this section, we present the proposed self-supervised contrastive video-speech representation learning approach in detail. The main idea of the proposed approach is to build the correlation between video and narrative speech data by both explicit optimisation and implicit regularisation. In this paper, we demonstrate the idea with fetal ultrasound video synchronised with corresponding speech. A novel self-supervised learning framework is proposed accordingly. Fig. 1 illustrates the main idea of the proposed framework.

2.1   Video-Speech Representation Learning

Speech audio, as an additional modality to visual data (*e.g.*, images and videos), is recorded on-the-fly without external annotations. The idea is to use the discriminative power of audio to understand visual data where ambiguity may exist. The basic assumption to learn a cross-modal (ultrasound video and corresponding speech from sonographer in our case) representation is that the two modalities share similar anatomical meanings at the same timestamp. In our case, if a model can successfully identify whether a video clip and a speech clip are correlated or not, it has learned to understand the underlying anatomical representations.

Based on the above assumption, we design a self-supervised learning framework to extract cross-modal representations with ultrasound video-audio data.
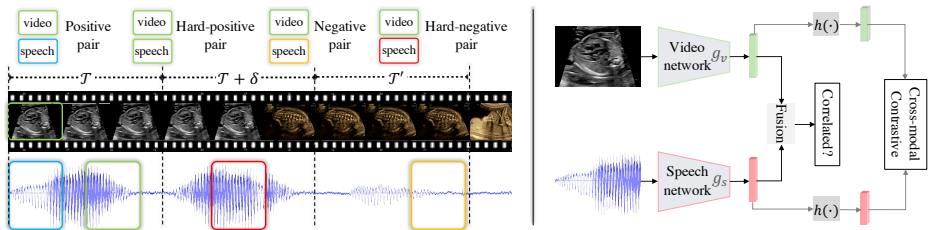
Fig. 2. Left: proposed training pair sampling scheme, where the hard-positive and hard-negative are illustrated. Right: proposed framework for self-supervised video-speech representation learning.

Specifically, we randomly sample positive and negative pairs from the original video-speech data, where a positive pair indicates that the considered video and speech are correlated, while a negative pair is uncorrelated. A deep model is then trained to identify such positive/negative pairs accordingly, resembling a classification problem. Unlike natural image video and its corresponding audio where highly-dense correlations present (*e.g.*, playing the violin, cooking), the speech audio from a sonographer and the medical ultrasound video are sparsely correlated and narratively presented, making the correlation identification more challenging. To address this issue, we first propose to force the model to *learn harder* by sampling hard-negative and hard-positive pairs (as illustrated in Fig. 2 left), so as to learn a more strongly correlated representations. Suppose the ultrasound video and speech clips at time interval $\mathcal{T}$ are $\mathcal{V}_{\mathcal{T}}$ and $\mathcal{S}_{\mathcal{T}}$, the speech clip at a shifted time interval $\mathcal{T}'$ is $\mathcal{S}_{\mathcal{T}'}$, $(\mathcal{V}_{\mathcal{T}}, \mathcal{S}_{\mathcal{T}})$ is considered as a positive pair and $(\mathcal{V}_{\mathcal{T}}, \mathcal{S}_{\mathcal{T}'})$ a negative pair. Instead of sampling $\mathcal{T}'$ from a different scan sequence or a diverse anatomy sub-sequence, we force the learning process to be constrained by sampling $\mathcal{T}'$ from a nearby segment. This does not have to be for a different anatomy as $\mathcal{T}$, *i.e.*, $\mathcal{T}' = \mathcal{T} + \delta$ where $\delta < \mathcal{D}$ is a short random timestep and $\mathcal{D}$ is the shift range. Furthermore, we sample ultrasound video frames and speech clips to increase the generalisability: $\{v_t, s_t, s_{t'} | v \in \mathcal{V}, s \in \mathcal{S}, t \in \mathcal{T}, t' \in \mathcal{T}'\}$. In terms of the positive pair, we also make it *harder* by perturbing the alignment within $\mathcal{T}$ so that v_t and s_t do not have to be exactly at the same timestamp t. Additionally, we randomly sample a group of positive/negative pairs within each interval to represent the whole clip. Finally, the hard-positive and hard-negative pairs used for training are defined as $\{(v_{ik}, s_{jk}) | k \in [1, K]\}$ and $\{(v_{ik}, s_{t'k}) | k \in [1, K]\}$ respecctively, where $i, j \in \mathcal{T}$ and K the number of sampling groups. An illustration of the proposed pair sampling scheme when $K = 1$ is shown in Fig. 2 (left).

With the above-mentioned sampling scheme, training pairs are fed into deep networks to extract the corresponding features. Suppose the ultrasound video and speech subnetworks as g_v and g_s, the extracted features are $g_v(v_t)$ and $g_s(\eta(s_t))$ where $\eta(\cdot)$ is a speech pre-processing function that converts the 1D signal to a 2D spectrogram (more details see Sec. 3.1). The video and speech features are fused for the correlation decision with a fusion function $f(v_t, s_t) =$

$g_v(v_t) \oplus g_s(\eta(s_t))$ where $\oplus$ represents feature concatenation. The correlation decision task is then modelled as a classification problem with the objective:

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{i}^{C} c_i^n log(f(v_t, s_t)_i^n), \tag{1}$$

where N is the total number of samples while $C = 2$ for our binary classification problem, c is the label indicating if the inputs are correlated or not.

2.2   Cross-modal Contrastive Learning

We observe that in the speech data, there exist medical-unrelated contents (*e.g.*, random talk), which deteriorate correlation learning. To address this, we introduce cross-modal contrastive learning, in addition to the afore-mentioned classification objective. The key idea here is to encourage the representations of video and speech from a (hard-)positive pair to be similar while repelling the (hard-)negative pair, at a projected embedding space. Specifically, before the feature fusion, $g_v(v_t)$ and $g_s(\eta(s_t))$ are further projected to a latent space by function $h(\cdot)$, where the cross-modal contrastive objective is applied. Suppose $y_v = h(g_v(v_t)), y_s = h(g_s(\eta(s_t)))$ are the projected embeddings, the cross-modal contrastive objective is defined as:

$$\mathcal{L}_{cont} = -log\frac{e^{sim(y_v, y_s)} - e^{sim(y_v, y_{s'})}}{\sum_{k=1}^{N} \mathbb{1}_{[k \neq v]} e^{sim(y_v, y_k)}}, \tag{2}$$

where $\mathbb{1} \in \{0, 1\}$ and $y_{s'} = h(g_s(\eta(s_{t'})))$. Function $sim(a, b) = a^\top b$ measures the similarity between two vectors. An illustration of the additional cross-modal contrastive learning is shown in Fig. 2 (right).

2.3   Affinity-Aware Self-paced Learning

The microphone used to record the speech is placed in an open environment without any specific voice filtering, which means that all the sound is recorded including background noise (*e.g.*, air conditioner, button-click), in addition to the main narrative content. Since our main focus is the speech with meaningful descriptions, the noise inevitably affects representation learning. To this end, we further propose a self-paced learning curriculum based on the affinity between video and speech data. Specifically, we divide the multi-modal data into different affinity levels and perform specific learning schemes accordingly. For simplicity, here we choose two affinity levels, *i.e.*, low-affinity and high-affinity, where low-affinity refers to the speech audio mostly consisting of noise and the rest is called high-affinity. The affinity is automatically detected according to an energy-based voice activity detection algorithm [14].

The proposed representation learning approach is only performed on the high-affinity fragments. In terms of the low-affinity data, instead of directly discarding it, we propose to leverage the video data which provides visual clues

for representation learning. As a result, we include an additional pretext task to extract representations from the whole video data. Inspired by [10], we randomly shuffle the frames in a video clip and train the model to predict the correct order. The assumption here is that the model can correct the shuffled frame order only if it understands the anatomical information within the video clip. Four frames are used to construct a clip for this task and the forward and backward sequences are considered to be the same (*e.g.*, 0-1-2-3 *v.s.* 3-2-1-0). Therefore, this pretext task is modelled as a 12-category classification problem with cross-entropy objective $\mathcal{L}_{ord}$. To avoid model cheating, the fan-shape of the ultrasound images is cropped out, keeping only the inner part.

2.4   Implementation and Training

The proposed model framework is illustrated in Fig. 2 (right). In terms of the backbone network architecture, we choose the ResNeXt-50 [17] with Squeeze-and-Excitation module [8] and dilated convolutions [18]. The video and speech subnetworks share the same architecture but are optimised separately with a joint objective function as defined in Eq. 3:

$$\mathcal{L} = \alpha\mathcal{L}_{cls} + \beta\mathcal{L}_{cont} + \gamma\mathcal{L}_{ord}, \tag{3}$$

where α, β, γ are weighting parameters and are empirically determined to be equal. The projection function $h(\cdot)$ is achieved by a multilayer perceptron (MLP) with a hidden layer and non-linear activation. The models are trained with the SGD optimizer with momentum set to 0.9 and weight decay as 5×10^{-4}. The learning rate is initialised to 10^{-3} and divided by 10 for every 20 epochs. The whole model is trained for 80 epochs. Gradient clipping is applied and the batch size is set to 32. The sampling group number $K = 2$ due to memory limitation and the interval skip range $\mathcal{D} = 5$. The whole framework is implemented using PyTorch on an NVIDIA RTX 2080Ti.

3   Experiments and Analysis

3.1   Data and Experimental Settings

The data used in this work is from a routine clinical fetal ultrasound dataset [1] with the scanned video and corresponding real-time speech as well as eye-gaze data from sonographers. In total, we have 81 scans with speech data. On average, each video scan is about 55,000 frames with frame rate of 30 fps. Each video clip $\mathcal{T}$ consists of 60 consecutive frames and we have 73,681 clips for model training. When sampling a training pair within each video clip, we extract 0.6s of the corresponding speech data and resample it to 24kHz. The speech is then converted to a 2D log-spectrogram representation of size 256×256, using a short-time Fourier transform (STFT) with 256 frequency bands, 10ms window length and

---

[1] UK Research Ethics Committee Reference 18/WS/0051.

Table 1. Evaluation results on standard plane detection (mean±std.[%]). Best performance is marked in **bold**. Note the methods on the right side are fully-supervised using external annotations.

	Rand.Init.	Video	Ours	ImageNet Init.	SonoNet
Precision	70.4±2.3	71.9±2.0	**72.7±1.8**	74.6±1.8	82.3±1.3
Recall	64.9±1.6	71.7±3.5	**73.3±2.4**	71.2±1.9	87.3±1.1
F1-score	67.0±1.3	71.5±2.4	**72.6±1.7**	72.5±1.8	84.5±0.9

Fig. 3. Confusion matrix on standard plane detection. Left: *Video*. Right: *Ours*. (Best viewed in digital form.)

5ms hop length. Two downstream tasks are included for the learned representation evaluation, where we use 135 scans with three-fold cross-validation (90/45 for train/test) and each scan is temporally down-sampled at a rate of 8.

3.2 Standard Plane Detection

To evaluate the learned representations, we first perform transfer learning on the standard plane detection task by fine-tuning the pre-trained weights from pretext tasks. We use 14 categories of heart three-vessel and trachea view (3VT), heart four-chamber view (4CH), abdomen, femur, brain transventricular plane (BrainTv.), kidneys, heart left ventricular outflow tract (LVOT), lips, profile, heart right ventricular outflow tract (RVOT), brain transcerebellum plane (BrainTc.), spine coronal plane (SpineCor.), spine sagittal plane (SpineSag.) and background. Standard plane labels are obtained from the dataset mentioned above. Initialisation from random weights, pre-trained weights only on video data (by the aforementioned frame order prediction task), pre-trained weights on ImageNet [13] and weights pre-trained with SonoNet [2] are included for comparison. Quantitative results are presented in Table 1. We see that the proposed self-supervised cross-modal representation approach performs better than the other alternative solutions on average. It also reveals that by leveraging the speech data with consideration of its correlation to the visual data, better performance is achieved, indicating stronger representations are learned. Note that our learned representation performs better than the ImageNet initialisation for most metrics, which also suggests that representation extracted from natural images do not transfer very well to medical data. In addition, we show the label confusion matrix in Fig. 3. It can be observed that our approach performs well on each category except the fine-grained ones like 3VT, 4CH, LVOT and RVOT (different views of fetal heart), which are challenging for sonographers too. When compared to the video-only representations (Fig. 3 left), we can see that our approach improves on almost all the categories, especially the abovementioned fine-grained ones. This is mainly due to the incorporation of visual-speech correlation where the additional speech representation reduces the ambiguity that exists in video data.
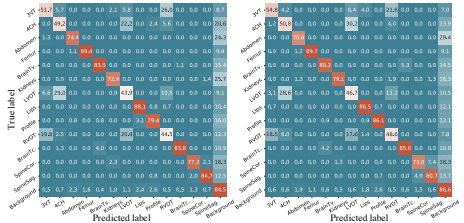
Table 2. Quantitative evaluation on eye-gaze saliency prediction. Best performance is marked in **bold**.

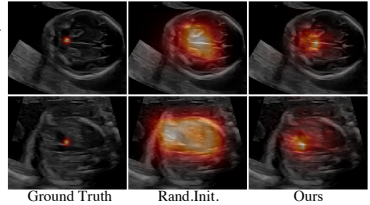|	KL↓	NSS↑	AUC↑	CC↑	SIM↑
Rand.Init.	3.94±0.18	1.47±0.24	0.90±0.01	0.12±0.02	0.05±0.01
Video	3.57±0.10	1.86±0.12	0.91±0.01	0.15±0.01	0.08±0.01
Ours	**3.05±0.04**	**2.68±0.05**	**0.95±0.00**	**0.22±0.00**	**0.11±0.00**
ImageNet Init.	3.95±0.28	1.72±0.25	0.89±0.02	0.14±0.02	0.08±0.01
SonoNet	3.14±0.02	2.62±0.03	0.94±0.00	0.21±0.00	0.12±0.00

Ground Truth   Rand.Init.   Ours

Fig. 4. Qualitative performance on eye-gaze saliency prediction.

3.3 Eye-Gaze Saliency Prediction

Since our dataset contains simultaneous eye-gaze tracking data from sonographers, in addition to the standard plane detection task as in Sec. 3.2, we further evaluate the effectiveness of the learned representations on a regression-based task, namely eye-gaze saliency prediction. Similarly, we load the pre-trained weights and fine-tune on the downstream task. A similar network architecture is used, with only the last layers modified to predict a 2D saliency map. Following [3], we use the KL divergence (KL), normalised scanpath saliency (NSS), area under curve (AUC), correlation coefficient (CC) and similarity (SIM) as the evaluation metrics. Quantitative and qualitative results are shown in Table 2 and Fig. 4 respectively, from which we can see that our approach again outperforms the alternative solutions, and even better than the approaches (ImageNet Init. and SonoNet) that were pre-trained with manual annotations.

3.4 Ablation Study

To better understand the effectiveness of the proposed learning strategies, we present an ablation study with corresponding performance shown in Table 3. We take the video-speech learning approach without cross-model contrastive learning (*CM.Contra.*) and affinity-aware self-paced learning as baseline. We can see from Table 3 that when including the cross-modal contrastive learning, the performance is improved by a large margin. Adding the affinity-aware learning scheme (*Ours*), further improves the model.

Table 3. Ablation study on each of the proposed strategies for two downstream tasks.

	Prec.[%]↑	Rec.[%]↑	F1[%]↑	KL↓	NSS↑	AUC↑	CC↑	SIM↑
Baseline	70.7±3.6	70.2±2.4	69.8±2.9	3.96±0.69	1.76±0.42	0.90±0.03	0.15±0.03	0.08±0.01
+ CM.Contra.	71.8±2.3	71.1±2.3	70.9±1.9	3.43±0.01	2.23±0.02	0.93±0.00	0.19±0.00	0.08±0.00
Ours	72.7±1.8	73.3±2.4	72.6±1.7	3.05±0.04	2.68±0.05	0.95±0.00	0.22±0.00	0.11±0.00

4   Conclusion

In this paper, we propose a self-supervised representation learning framework for ultrasound video-speech multi-modal data. To the best of our knowledge, this is the first attempt towards cross-modal representation learning without human annotations for ultrasound data. We designed a simple, yet effective, approach by modelling the affinity between these two modalities. To address the inherent sparse correlation and noise issues in the speech data, we propose a cross-modal contrastive learning scheme and an affinity-aware self-paced learning scheme. Experimental evaluation on two downstream tasks shows that the learned representations can transfer to fetal ultrasound standard plane detection and eye-gaze saliency prediction and improve the average performance accordingly. The proposed approach shows the potential to mitigate the need for laborious manual annotation work in deep learning-based applications for medical imaging via automated cross-modal self-supervision. Since the proposed approach is trained and evaluated on video data with narrative speech audio and is not specifically tailored for ultrasound, it could apply to other unseen video-audio data.

References

1. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: ICCV (2017)
2. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D.: Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE TMI **36**(11), 2204–2215 (2017)
3. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE TPAMI **41**(3), 740–757 (2018)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
5. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018)
6. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
7. Hénaff, O.J., Srinivas, A., De Fauw, J., Razavi, A., Doersch, C., Eslami, S., Oord, A.v.d.: Data-efficient image recognition with contrastive predictive coding. arXiv preprint arXiv:1905.09272 (2019)
8. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
9. Jiao, J., Droste, R., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Self-supervised representation learning for ultrasound video. In: ISBI (2020)
10. Lee, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Unsupervised representation learning by sorting sequences. In: ICCV (2017)

11. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
12. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: ECCV (2018)
13. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
14. Sakhnov, K., Verteletskaya, E., Simak, B.: Approach for energy-based voice detector with adaptive scaling factor. IAENG International Journal of Computer Science **36**(4) (2009)
15. Spitzer, H., Kiwitz, K., Amunts, K., Harmeling, S., Dickscheid, T.: Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. In: MICCAI (2018)
16. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: ICCV (2015)
17. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR (2017)
18. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
19. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)
20. Zhuang, X., Li, Y., Hu, Y., Ma, K., Yang, Y., Zheng, Y.: Self-supervised feature learning for 3d medical images by playing a rubik's cube. In: MICCAI (2019)