DUAL REPRESENTATION LEARNING FROM FETAL ULTRASOUND VIDEO AND SONOGRAPHER AUDIO

Mourad Gridach[1] *Mohammad Alsharid*[1,2] *Jianbo Jiao*[1,3]
Lior Drukker [1] *Aris T. Papageorghiou*[1] *J. Alison Noble*[1]

[1]University of Oxford, [2]Khalifa University, [3]University of Birmingham

ABSTRACT

This paper tackles the challenging problem of real-world data self-supervised representation learning from two modalities: fetal ultrasound (US) video and the corresponding speech acquired when a sonographer performs a pregnancy scan. We propose to transfer knowledge between the different modalities, even though the sonographer's speech and the US video may not be semantically correlated. We design a network architecture capable of learning useful representations such as of anatomical features and structures while recognising the correlation between an US video scan and the sonographer's speech. We introduce dual representation learning from US video and audio, which consists of two concepts: Multi-Modal Contrastive Learning and Multi-Modal Similarity Learning, in a latent feature space. Experiments show that the proposed architecture learns powerful representations and transfers well for two downstream tasks. Furthermore, we experiment with two different datasets for pretraining which differ in size and length of video clips (as well as sonographer speech) to show that the quality of the sonographer's speech plays an important role in the final performance.

***Index Terms*—** Ultrasound, Multi-Modal, Self-Supervised

1. INTRODUCTION

Fetal US is an essential part of pregnancy care to ensure maternal and fetal well being. The main purposes of the scan are to ensure pregnancy viability and assess the risk for chromosomal anomalies. Scans are carried out by a sonographer who controls the US probe to achieve a series of standard imaging planes. This process requires a high level of skill and is not available worldwide. Automated guidance in scanning may contribute to changing current practice, help a non-expert to detect abnormality and make US accessible worldwide.

CNNs were widely used in medical image analysis. However, most of the existing approaches rely on supervised learning when large-scale annotated datasets can be used for training. The annotation process needs extensive effort, is time-intensive, tedious and requires expertise, which limits the use of these models. In the meantime, learning useful and relevant representations from a large amount of unlabeled image data is an important step in medical image analysis.

Various methods have been proposed to learn representations from unlabeled data called "self-supervised learning (SSL)". It has shown great success in learning image-level features from large-scale unlabeled data without using any human-annotated labels. In SSL setting, a model (encoder) is pretrained on unlabeled data according to a specific task and then the trained encoder can be fine-tuned for training a supervised downstream task to evaluate the learned representations. These downstream tasks for images include classification [1], segmentation [2], and colourisation [3]. Recent methods have investigated learning from audio and video modalities [4, 5]. In this case, audio and video are assumed to be in dense correspondence. While they have considered cross-modal representation learning, they assume that there is a (strong) correlation between modalities. Nevertheless, in fetal US, the video and the sonographer speech may not always be semantically correlated (conversation about weather, and food).

In this paper, we tackle a challenging multi-modal representation learning problem using a SSL approach for US scans with corresponding sonographer speech without relying on annotations. Developing networks that can learn representations from unlabeled US raw data is difficult because the model must bridge the semantic gap between the different datatypes and deal with incompatible network architectures for each datatype. To address these challenges, we propose dual representation learning from fetal US videos and speech audios. Our approach consists of two main concepts: 1) Multi-Modal Contrastive Learning (MMCL), and 2) Multi-Modal Similarity Learning (MMSL) in a latent feature space. Learning anatomical features in an unsupervised manner while bridging the gap between different datatypes is the main goal of our approach. Within our framework, we consider three datatypes: video, speech, and image. In the dataset, we sample videos, corresponding speech, and an image randomly sampled from the video. Fig. 1 illustrates this setting. Our hypothesis is that including sonographer speech to the scans can contribute to better representation learning. Therefore, we collected two datasets for pretraining: the first is large but with short clips (2s), while the second is a medium-sized dataset but with longer clips (6s).

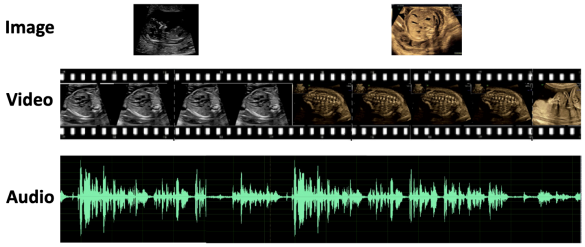


Fig. 1: The three datatypes used in our approach.

2. METHODOLOGY

In this work, we validate our hypothesis that useful representations can be learned from audio and video with the help of images to align the sonographer's speech with the US video scan. Fig. 2 illustrates the main idea of our proposed method.

2.1. Encoding Networks

Given a dataset $\mathcal{D} = \{V_i, A_i\}_{i=1}^{n}$ where there are N videos (V_i) each with audio (A_i). For image (I_{ij}), we randomly sample it from video V_i. Then, we extract the unimodal embeddings for each datatype using CNNs.

Image Encoder. To extract anatomical features from an US image, we use a 2D-CNN architecture. Given a video V_i, we randomly sample an image I_{ij} from it, then we use the 2D-CNN to extract the image features as follows: $F_i = W_{2D-CNN}(I_ij)$, where F_i is D_i-dimensional feature vector (image embedding).

Audio Encoder. Given an audio A_i corresponding to a video V_i, we extract the log-mel spectrogram and feed it to a pretrained 1D-CNN to get the audio features as follows: $F_a = W_{1D-CNN}(A_i)$, where F_a is D_a-dimensional feature vector (audio embedding) and W_{D-CNN} are parameters.

Video Encoder. We use a 3D CNN to extract video representations. The feature extractor has a stack of residual blocks with $(2+1)D$ convolutions. Given a video V_i represented as a volume of D×H×W×T, features are extracted as follows: $F_v = W_{3D-CNN}(V_i)$.

2.2. Multi-Modal Contrastive Learning (MMCL)

We hypothesize that two modalities, either the pair video-audio or video-image, share similar anatomical structures at the same timestamp, which will help to learn a cross-modal representation. Therefore, the model aims to successfully recognize whether a sonographer speech and an US video are correlated. The same idea can be applied to the pair of image and video scan. To reduce the possible semantic gap between datatypes, we propose to correct the speech audio and image embeddings by applying an efficient fusion method between the two pairs namely video-audio and video-image embeddings to reduce the semantic gap in the latent feature space. Given the video, speech audio and image embeddings represented as F_v, F_a, and F_i respectively, we fuse the two pairs

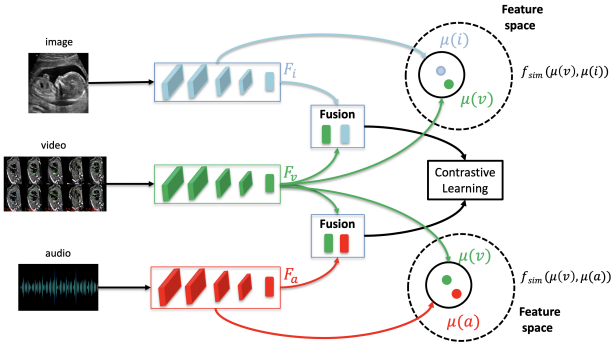


Fig. 2: Our proposed pretraining approach.

using the following equations:

$$U_{va}(F_v, F_a) = F_{va} = F_a + V_{\theta_{va}}(F_v, F_a) \quad (1)$$

$$U_{vi}(F_v, F_i) = F_{vi} = F_i + V_{\theta_{vi}}(F_v, F_i). \quad (2)$$

To obtain the function $V_\theta(.,.)$, we apply two main operations on the embeddings: concatenation and linear projection. Then, we apply contrastive learning in the latent feature space to enforce the positive pairs to be close and the negative pairs to be far apart. The positive pairs are images, videos, and speech audios from the same timestamp (as illustrated in Fig. 1). Therefore, the contrastive loss between the audio and video scan is defined as follows:

$$\mathrm{L}_{CL}(V_k, A) = -\log \frac{\exp(\psi(F_v^{(k)}, F_a^{(k)})/\tau)}{\sum_{j=1}^{M} 1_{[j\neq k]} \exp(\psi(F_v^{(k)}, F_a^{(j)})/\tau)}. \quad (3)$$

Also we can construct a contrastive loss between the video embeddings and video-audio embedding as follows:

$$\mathrm{L}_{CL}(V_k, VA) = -\log \frac{\exp(\psi(F_v^{(k)}, F_{va}^{(k)})/\tau)}{\sum_{j=1}^{M} 1_{[j\neq k]} \exp(\psi(F_v^{(k)}, F_{va}^{(j)})/\tau)}. \quad (4)$$

Similarly, we apply the same idea to image and video:

$$\mathrm{L}_{CL}(V_k, I) = -\log \frac{\exp(\psi(F_v^{(k)}, F_i^{(k)})/\tau)}{\sum_{j=1}^{M} 1_{[j\neq k]} \exp(\psi(F_v^{(k)}, F_i^{(j)})/\tau)}. \quad (5)$$

$$\mathrm{L}_{CL}(V_k, VI) = -\log \frac{\exp(\psi(F_v^{(k)}, F_{vi}^{(k)})/\tau)}{\sum_{j=1}^{M} 1_{[j\neq k]} \exp(\psi(F_v^{(k)}, F_{vi}^{(j)})/\tau)}. \quad (6)$$

The final contrastive losses between video/audio and video/image pairs are defined as follows:

$$L_a = \pi L_{CL}(F_v, F_a) + (1-\pi) L_{CL}(F_v, F_{va}) \quad (7)$$

$$L_i = \pi L_{CL}(F_v, F_i) + (1-\pi) L_{CL}(F_v, F_{vi}), \quad (8)$$

where π balances the two terms. Thus, the losses L_a and L_i are considered as similarity constraints to make the embeddings close together in the multi-modal latent feature space.

2.3. Multi-Modal Similarity Learning (MMSL)

As with the MMCL idea, we use the sonographer speech audio and the image as a summary of the video to understand the anatomical structures found in US video scans. Given the video feature vector F_v, the image feature vector F_i, and the audio feature vector F_a, we proceed to a projection in the latent feature space using affine transformations in the form of a fully connected (FC) layer. Batch normalization (BN) is used after the FC layer to ease training using: (1) $\mu(v) = BN(\theta_v F_v + b_v)$, (2) $\mu(a) = BN(\theta_a F_a + b_a)$, and (3) $\mu(i) = BN(\theta_i F_i + b_i)$, where θ_v, θ_a, and θ_i are parameters of each FC layer, while b_v, b_a, and b_i are biases. We use video-audio and video-image similarities to learn cross-modal representations between pairs of datatypes as follows:

$$f_{sim}(v,a) = \frac{\mu(v).\mu(a))}{\|\mu(v)\|.\|\mu(a)\|} \tag{9}$$

$$f_{sim}(v,i) = \frac{\mu(v).\mu(i))}{\|\mu(v)\|.\|\mu(i)\|}, \tag{10}$$

where $f_{sim}(v,a)$ is the similarity between video and audio while $f_{sim}(v,i)$ is the similarity between video and image. f_{sim} is the cosine similarity where its goal is to make the audio closer to the video when the sonographer speech is relevant to the scan, while pushing them far apart when they are irrelevant in the latent feature space. The same idea is applied to the video-image pair. Given a mini-batch of video-audio pairs, we use the improved triplet ranking loss [6], which penalizes the model according to the hardest negative samples. Consequently, given a video-audio pair (v,a) and video-image pair (v,i), we compute the losses as follows:

$$\begin{aligned} L_{MMSL}(v,a) = \max(0, \alpha + f_{sim}(v,a^-) - \\ + \max(0, \alpha + f_{sim}(v^-,a) - f_{sim}(v,a)) \end{aligned} \tag{11}$$

$$\begin{aligned} L_{MMSL}(v,i) = \max(0, \alpha + f_{sim}(v,i^-) - \\ f_{sim}(v,i)) + \max(0, \alpha + f_{sim}(v^-,i) - f_{sim}(v,i)), \end{aligned} \tag{12}$$

where α is the margin constant, while a^- and v^- respectively indicate a negative audio sample for the video v and a negative video sample for a. The same thing for i^- and v^- when we compute the similarity between video and image.

3. EXPERIMENTS AND ANALYSIS

Experimental Settings. In terms of backbone networks, we use the $2D-ResNet18$ for image, $1D-CNN14$ for audio, and $R(2+1)D-18$ for video. Since these networks' dimensions differ, we use linear projections to map the embeddings to the same dimension. The dimension of the latent feature space is 512. We use an Adam optimizer with the default initial learning rate of 3.10^{-3} and weight decay of 10^{-4}. The temperature τ is set to 0.1 and π is set to 0.5. The video clips are cropped to 128×128, where each clip contains 14 frames.

Table 1: Performance of our approach on SP detection.

Models	Precision	Recall	F1-score
ImageNet Init.	74.6 (1.8)	71.2 (1.9)	72.5 (1.8)
Rand. Init.	69.5 (1.9)	65.6 (1.5)	67.49 (1.1)
Ours (Large)	70.1 (1.2)	66.4 (1.4)	68.19 (1.2)
Ours (Medium)	**73.9** (1.3)	**74.8** (1.4)	**74.36** (1.1)

Models are trained using PyTorch and are optimized with the following final objective function:

$$L_{final} = L_a + L_i + L_{MMSL}(v,a) + L_{MMSL}(v,i) \tag{13}$$

Pretraining Dataset. The data used to pretrain our approach is from routine clinical fetal US exams with real-time sonographer gaze tracking data from the PULSE study[1]. We select 81 scans with speech data. On average, each video scan is about 55,000 frames with frame rate of 30 fps. To pretrain our approach, we collect two datasets: (1) a dataset with 44,173 videos with 2s each, and (2) a dataset with 10,457 video scans where the length of each is 6s.

Fine-Tuning Datasets. We fine-tuned our proposed method using two downstream tasks namely saliency prediction and standard plane detection, where we used a dataset of 135 scans with five-fold cross-validation (90/45 for train/test). In standard plane detection, we have 14 categories: heart three-vessel and trachea view, heart four-chamber view, abdomen, femur, brain transventricular plane, kidneys, heart left ventricular outflow tract, lips, profile, heart right ventricular outflow tract, brain transcerebellum plane, spine coronal plane, spine sagittal plane and background. For saliency prediction, we use the same dataset because it contains simultaneous eye-gaze tracking data from sonographers.

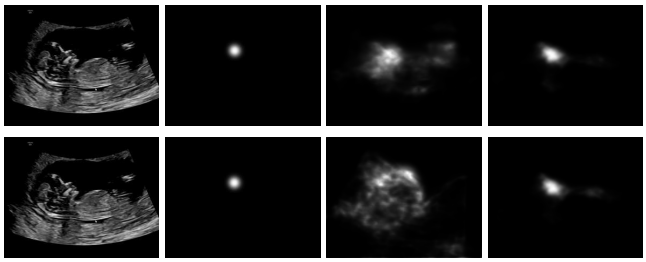
3.1. Standard Plane Detection

In this first experiment, we investigate two questions: (1) do the learned features improve the performance of standard plane detection, even in the situation of small datasets?; and (2) does the pretraining dataset size and the quality of speech audio play an important role in the final results? We compare the performance of our approach on two unlabeled datasets, which differ in the size, length, and the quality of sonographer speech. Also, we compare our approach with random weights, and pretrained weights on ImageNet, which is fully supervised. The quantitative results are shown in Table 1. It is clear that our proposed approach performs better than the other methods. More importantly, we notice that the model trained on the medium-sized dataset but with video scans of 6s performs better than the one trained on a larger dataset but with shorter video scans (2s). Moreover, we see that by leveraging the speech data, we improve the performance, showing that stronger representations are learned.

[1]This study was approved by the UK Research Ethics Committee (Reference 18/WS/0051) and the ERC ethics committee.

Table 2: Performance on saliency prediction.

Model	KLD ↓	SIM ↑	NSS ↑	CC ↑
ImageNet	3.95 (0.28)	0.08 (0.01)	1.72 (0.25)	0.14 (0.02)
Rand. Init.	4.02 (0.11)	0.07 (0.09)	1.76 (0.10)	0.13 (0.09)
Ours (Large)	3.96 (0.10)	0.09 (0.12)	1.77 (0.12)	0.14 (0.09)
Ours (Medium)	**3.02 (0.06)**	**0.13 (0.05)**	**2.72 (0.04)**	**0.25 (0.05)**



(a) US (b) GT (c) RI (d) Ours

Fig. 3: Qualitative performance on saliency prediction.

3.2. Ultrasound Saliency Prediction

As in standard plane detection, we use the pretrained weights and finetune on the current task. We modify the last layers of the network to predict the saliency map. We use Kullback-Leibler divergence (KL), Normalised Scanpath Saliency (NSS), Correlation Coefficient (CC) and Similarity (SIM) as the evaluation metrics [7]. Results are shown in Table 2. Our proposed approach outperforms the other alternative methods including ImageNet initialization, which was pre-trained with manual annotations. We notice that the model performs better when trained on the medium-sized dataset (video scans with 6s) rather than large dataset but with short video scans. Furthermore, we show qualitative results in Fig. 3 where we compare with random initialization (RI).

3.3. Ablation Study

We perform an ablation study to quantify the performance of each module and the image datatype. We select standard plane detection and the medium unlabeled dataset as our evaluation task. Table 3 shows the results with only the MMCL and the MMSL strategies and the contribution of the image. We can see that MMCL better contributes to the overall performance compared to MMSL. Combining the two approaches further improves the results. We notice that using the image datatype slightly boosts the performance (referred to as w/o image in Table 3).

4. DISCUSSION AND CONCLUSION

We present a novel dual representation learning approach from US video and sonographer speech. We show that audio plays an important role in learning strong representations. To validate our assumption that including sonographer speech with the visual data (video scans), we can learn strong representations and improve performance, we designed and tested models using two different datasets. For the first models we used a large dataset of short video clips (2s) that contain some redundancies such as video clips with sometimes no sonographer speech. The second model used a medium-sized dataset of longer video clips (6s). These video clips can contain complete commentary about the video clip. The experimental results confirm our hypothesis that we get improved performance when we train the models on the second dataset and fine-tune them on two challenging downstream tasks. Furthermore, we show that image datatype can help to improve the performance, especially when the sonographer speech is not correlated with the video scan, and validate our intuition that an image can summarize the whole video scan.

Table 3: Ablation study of the proposed approach.

Models	Precision.	Recall	F1-score
MMSL	69.8 (1.2)	71.5 (1.4)	70.63 (1.3)
MMCL	71.1 (1.3)	72.4 (1.2)	71.74 (1.2)
W/o image	73.1 (1.3)	74.3 (1.2)	73.69 (1.1)
Ours (Medium)	**73.9** (1.3)	**74.8** (1.4)	**74.34** (1.1)

5. ACKNOWLEDGMENTS

We acknowledge the ERC (ERC-ADG-2015 694581 project PULSE), the EPSRC (EP/MO13774/1), the Royal Society (IES\R3\223050), and the NIHR Oxford Biomedical Research Centre (BRC) funding scheme. The authors have no financial conflicts of interest to disclose related to this work.

6. REFERENCES

[1] Chen et al., "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[2] Hoyer et al., "Three ways to improve semantic segmentation with self-supervised depth estimation," in *CVPR*, 2021.

[3] Zhang et al., "Colorful image colorization," in *ECCV*. Springer, 2016, pp. 649–666.

[4] Afouras et al., "Deep audio-visual speech recognition," *TPAMI*, vol. 44, no. 12, pp. 8717–8727, 2018.

[5] Jiao et al., "Self-supervised contrastive video-speech representation learning for ultrasound," in *MICCAI*. Springer, 2020, pp. 534–543.

[6] Faghri et al., "VSE++: Improving visual-semantic embeddings with hard negatives," in *BMVC*, 2018.

[7] Bylinskii et al., "What do different evaluation metrics tell us about saliency models?," *TPAMI*, 2018.