Moving In Sync: Self-Supervised Learning of n-Human Interactions

CSE 303 Final Project

Sonal Sannigrahi

Under the supervision of Prof. Vicky Kalogeiton
École Polytechnique, France
13 December, 2020

Moving In Sync: Self-Supervised Learning of n-Human Interactions

Sonal Sannigrahi[1]

Abstract— **The objective of this paper is to provide a self-supervised learning framework for video representation learning in order to classify human interactions. The following contributions are made: (i) we introduce a novel architecture, dubbed "Sync3D", that aims to synchronise human tracks to predict interactions by exploring contrastive loss on learnt I3D feature vectors; (ii) we provide a data sampling strategy based on temporal and spatial alignment to create positive and negative samples; (iii) lastly, we evaluate the quality of our proposed learnt representation on the downstream task of human interaction (or action) classification. We achieve a final top1 classification accuracy 75.5% on our validation set, outperforming I3D's classification accuracy on the TV Human Interaction dataset.**

I. INTRODUCTION

Interest in areas of self supervised learning, as opposed to fully supervised methods, have seen a steady growth over the past years. This interest primarily stems from the explosion of data freely available on the web, as such manually annotating large amounts of data seems like a tedious task that could be avoided with self supervision. Videos are an especially popular form of data as the abundance of such data coupled with the variety, from sources such as YouTube, is quite large.

One challenging task with video data is to detect human interactions in a given video. This task is useful in video annotation, automated surveillance, and content-based video retrieval for quick search results [1]. The challenge occurs because this task involves both tracking humans as well as learning the semantics of the interaction taking place. Previous approaches study CNN architectures, learn feature representations or decouple the visual and temporal aspects of human interactions by using LSTMs [3]. However, these methods suffer in training scenarios with low amounts of annotated data as learning feature representations require large amounts of annotated training data.

In this work, we propose a novel self-supervised method (termed "Sync-3D") that learns spatio-temporal video embeddings to enable the detection of human interactions. Our work combines the I3D architecture used for action localisation [2] and the SyncNet architecture for video-audio synchronisation [4] and casts the problem of human interaction detection as one of motion synchronisation both spatially and temporally.

[1] École Polytechnique

II. RELATED WORK

In this section, we introduce the three key areas related to our problem: action recognition, self-supervised representation learning, and the study of human interaction in a visual context. While many works in the recent past have studied action recognition, there are few that target human interactions [5], [6]. The existing works can mainly be seen as three different categories: no interaction (action or motion recognition), human-object interaction, and lastly human-human interaction.

A. *Action Recognition*

The actions of the humans in each frame are indicative of potential interactions between them. There has been significant progress made in the areas of video action recognition [7], [8], [9]. In these approaches, a notable improvement from pre-training was reported ; therefore, in this work we adopt this pre-training and use an action recognition module pre-trained on ImageNet. Furthermore, in these variants of a two stream architecture, optical flow has been reported to be a powerful representation showing great improvements in the task as compared to using only RGB frames. While in this work we focus solely on RGB frames, we argue that optical flow can increase the robustness of representation learning and it can be an interesting line of future work.

B. *Self-Supervision*

Due to the availability of large-scale data, the use of deep neural networks has shown significant success in representation learning over the past years. For both images and videos, using convolutional architectures has been widespread due to desirable properties such as shift-invariance and hierarchical learning. However, annotating these copious amounts of data is a cumbersome task, making self-supervised training of these architectures a necessity. The recent advances on self-supervised learning can be split into two categories based on the application domain: learning from images or learning from videos. In the areas of image classification, self-supervision has seen steady progress [10] through tasks such as inpainting [11], image colourisation [12] and jigsaw puzzle solving [13]. In the video domain, recent approaches aim to learn distinct instances from each other through contrastive learning [14], [15].

C. *Human Interaction*

Several works have studied human interactions in videos in terms of human-object interaction as well as a few in human-human interactions [16], [17]. A popular line of work has included the use of deep neural architectures

to learn relationships between humans by extracting rich feature representations. Past works such as LAEO-Net [18] address the problem of studying human-human interaction by studying head-orientation however using deep neural approaches goes back to using well annotated data. To tackle this problem, in this paper we prose a combination of self supervised learning with action recognition to allow for human interaction recognition in videos without the need for large scale annotation.

III. MOTION SYNCHRONISATION

In this section, we describe the model architecture, the learning framework, the self-supervised strategy, and the curriculum training used to learn progressively more difficult negatives.

A. Architecture

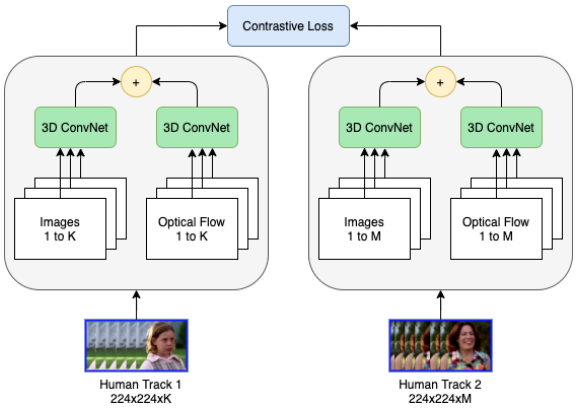

Fig. 1. **Sync3D:** Architecture to perform motion synchronisation. The Sync3D architecture uses the base structure from SyncNet itself but uses I3D to perform action recognition on inputs of human tracks. We chose to use a pre-trained ImageNet on the Kinetics-400 dataset [ref] for each of the two tracks.

We propose a novel architecture to learn video representations to perform motion synchronisation, called Sync3D. Here, we treat the problem of learning human interactions in video as one of both spatial and temporal synchronisation between two human tracks. As such, our base architecture is derived from SyncNet [4]; however, instead of syncing audio and video components, we sync two human tracks. In order to perform feature extraction on these input tracks, we turn to action recognition. In the proposed model, we choose the I3D architecture, introduced in [2], as our desired action recognition engine.

B. Learning Framework

The goal of Sync3D is to learn interaction labels for humans in videos through action recognition, where these labels are derived from verifying temporal and spatial alignment of tracks. As discussed in the previous sub-section, we use a two-stream convolutional neural architecture which takes as input two human tracks, with each branch being an I3D network. This architecture gives as output an interaction label of 1 if the two tracks are predicted to be synchronized and 0 otherwise. In this section, we discuss the loss used for training and the sampling strategy to create positive and negative data pairs.

1) Contrastive Loss: The training objective for the network is to report a target closer to one for tracks in which humans are interacting, and a target closer to zero otherwise. Contrastive Loss generally refers to the training paradigm where similarity scores of positive labels are boosted above those of the negative labels, making this loss suitable for our task. As done in SyncNet, we minimise the following objective function, first introduced in [19] to train Siamese networks like ours:

$$L = \frac{1}{2N}\sum_{i=1}^{N} y_n \times d_n^2 + (1 - y_n) max(margin - d_n, 0)^2$$

Here, we compare the similarity between the feature vectors of track 1 and track 2 after taking their output from the I3D action classifier. Then using the loss function, we minimise the distance between them if they are positive pairs otherwise we penalise the entry.

2) Data Sampling: Here we detail the sampling strategy to produce data pairs from a single video clip. We also classify three categories of negative samples, allowing for a curriculum learning strategy to train Sync3D.

Preprocessing of Video Clips: In order to create the data pairs, some preprocessing steps are required. From each video clip, we crop every frame to bounding boxes containing the upper bodies of the humans present. We then enlarge this crop by 20% to include background as part of our track. Finally, this cropped frame is placed on a neutral 224×224 canvas.

Positive Pairs: We define human interaction as consisting of two components: spatial and temporal alignment. In our use case, we measure spatial alignment with a simple metric based on a threshold: Intersection-over-Union (IOU). Thus, positive pairs are formed with pairs of human tracks from the same video such that their mean IOU over the temporal length is larger than a threshold which we set to be 0.1.

Negative Pairs: Using the complement of how we have defined positive pairs, negative pairs can be placed into three categories as follows:

- **Easy Negatives:** Here we consider videos that are neither temporally nor spatially aligned by considering tracks from two distinct videos. These negative pairs are considered easy to learn due to no or little similarity between each pair of sequences of frames, thus resulting in low similarity between predicted and ground-truth features.

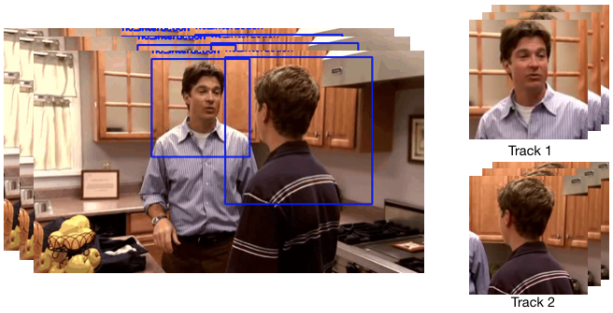

Fig. 2. A sample clip for which the preprocessing stage has been shown. As the mean IOU computed for the tracks is larger than the threshold, this is counted as a positive sample

- **Medium Negatives:** Now, we consider tracks that are temporally aligned but are not spatially aligned. This corresponds to considering tracks from the same video that do not meet the IOU threshold.

- **Hard Negatives:** Lastly, we consider tracks that are spatially aligned but are not temporally aligned. Here, we consider the tracks we classify as interacting based on the IOU threshold and apply a temporal shift of n frames, producing two tracks that are not temporally aligned however do maintain spatial alignment. These are considered as hard negatives due to how close their score will be to the positive pairs. In our experiments we set n=25 frames.

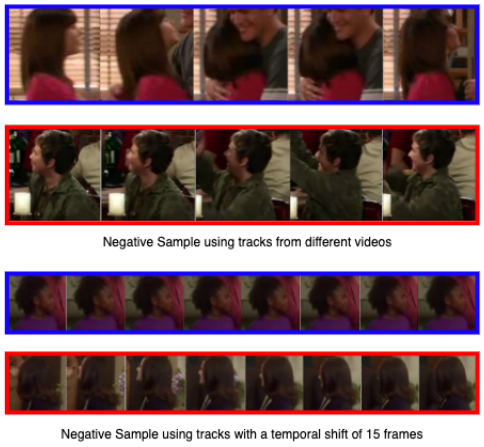

Fig. 3. Sampling strategy to obtain negative pairs. **Top** shows an easy negative pair, **Bottom** shows a hard negative

Curriculum Learning Strategy

At the training step, we use a $\frac{1}{4}$ and $\frac{3}{4}$ sampling probability for positive and negative pairs respectively. However, with the three different categories of negative pairs we have much more than $3N$ pairs of negatives where N represents the number of positive pairs. Thus, we must decide on a second sampling strategy to choose negative pairs that allow Sync3D to perform the best.

We introduce a curriculum learning strategy by continuously increasing the number of hard negatives after certain steps. Initially, we only incorporate the easy negatives, i.e. sample negative pairs from different videos. After the network has learnt this task, we incorporate the medium and hard negatives. By doing so, we force the network to learn how to distinguish between interactions (or not) in increasingly similar videos, thus gradually increasing the difficultly of the learning process. In this case, the network learns representations for the interactions themselves rather than trying to distinguish with, for instance, the background of the videos.

IV. EXPERIMENTS AND ANALYSIS

In this section, we describe the dataset used and the implementation details for Sync3D training. We further describe the downstream task of action recognition to evaluate the representation learnt from our self-supervised methodology.

A. Dataset

We use the TV Human Interaction Data (TV-HID) for self-supervised Sync3D training, which contains 300 video clips extracted from 23 different TV shows split into four interaction classes: hand shake, high five, hug, and kiss with each class having 50 videos [20]. In addition to these interaction classes, the remaining 100 videos are composed of negative samples where no interaction takes place. For each human in each frame in each video, we have the following annotations: interaction label, head orientation, and bounding box for the upper body. These annotations are used to perform the downstream evaluation. We chose to use the test/train split as provided by the dataset, which is a 50/50 split.

Interaction Class	% Positive	% Negative
Hand Shake	63.45	36.55
High Five	37.25	62.75
Hug	100	0
Kiss	89.10	10.90
Negatives	0	100

TABLE I

DETECTION ACCURACY FROM OUR SAMPLING STRATEGY

From Table 1, we note that our data sampling strategy in successful in three classes: kiss, hug, and negatives. Further, it performs with acceptable accuracy on the class Hand Shake, however it performs poorly on High Five. Hence, our self-supervised methodology misses some examples of positive interactions. In Fig. 4, we see an example from the interaction class "Hand Shake" but due to the IOU being 0, our strategy labels this pair as a negative.

B. Implementation Details

As discussed in Section III A, Sync3D uses the I3D architecture as the feature extractor. During its training, we use a non-linear projection layer which is later removed for downstream task evaluation as done in SimCLR [21]. For input, we take apply the sampling strategy to get two 25-frame RGB human tracks, at 25 fps, covering 1-second interactions. Each track is resized to 224×224 pixels.

Fig. 4. Here the bounding boxes do not intersect leading to a false negative according to our sampling strategy.

During the training step, we apply data augmentation techniques such as: random crops, random horizontal flips, image noise for each frame. These augmentations are applied consistently across each frame so that our network does not try to learn low level information. In order to maintain the 25-frame temporal window and to maximise the possibility of detecting positive samples, we choose to use the centre-most 25 frames as our target temporal window.

We train our network for 800 epochs using the SGD optimiser using an initial learning rate of 10^{-3} and a weight decay set to 10^{-5}, we also schedule our learning rate to drop to 10^{-4} and 10^{-5} at epochs 300 and 600 respectively. Due to sparsity of the data, we kept the batch size as relatively small but we applied batch normalisation on it as a regularisation method. We also applied dropout probability of 0.1 as an additional regularisation method. To apply our curriculum learning strategy, we describe two settings: Easy + Medium negatives and Hard negatives. In the first case, we do not incorporate any hard negatives in the data sampling so we samples from positives, and easy and medium negatives. In the second case, we include 33% hard negatives in the $\frac{3}{4}$ negative samples. All our models are trained end-to-end.

C. Evaluation Methodology

Our self-supervised architecture is first trained on TV-HID tracks to learn an interaction label of 1 or 0. This representation is then evaluated by its performance on the downstream task of **Interaction Recognition** on TV-HID. Here "Interaction Recognition" refers to the recognition of the human interaction classes of TV-HID. To evaluate this, we use a linear probe setting: the feature encoder is frozen, and a single layer perceptron is trained with cross-entropy loss. Since we are evaluating for human interaction recognition, we use the labels from TV-HID to give us five classes. We report top1 accuracy on both the downstream evaluation task of interaction classification in a fully supervised setting as well as the results from our self-supervised training. Recall that the top1 accuracy in the self-supervised setting refers to how often the network chooses the right interaction label and does not have any connection to the ground truth action classes. However, in the fully supervised setting, the top1 accuracy indicates the interaction classification accuracy on TV-HID.

D. Interaction Classifier

During the supervised learning stage, we pass 2 human tracks as input (the same as for the self-supervision training, each track is in $\mathbb{R}^{25\times224\times224\times3}$) and these tracks and then encoded as a sequence of feature maps using the frozen weights from the Sync3D training cycle. This encoded sequence is then passed onto a fully-connected layer and softmax for interaction classification. The classifier is trained using the SGD optimiser with an initial learning rate of 10^{-3} and weight decay of 10^{-5}. During the testing step, tracks from the validation set are sampled with the same procedure as in the training step ($\frac{1}{4}$ positive and $\frac{3}{4}$ negative) and no augmentations are applied. The softmax probabilities are averaged to give the final result.

E. Analysis

First, we present the results of the self-supervised learning by Sync3D on TV-HID. We achieve an overall accuracy of 50.4% in the training set and 42.6% in the validation set. As we will discuss in the next section, we faced significant issues in convergence of the loss function due to low training data; however, even with this minimal performance, we managed to receive interesting results on the downstream evaluation task.

In the downstream task, we compare our method against I3D on top1 classification accuracy. Here, our method outperforms the use of I3D features. Interestingly, this shows that while our methodology did not perform well on the self-supervised task, it still learnt a useful representation which improved its performance on human interaction classification. In the table below, we experiment with the use of our curriculum learning strategy and see that with the introduction of hard negatives, we are able to beat I3D's performance on this task (73.2% for I3D vs. 71.3% for Sync3D without curriculum learning and 75.5% with curriculum learning). In Fig. 5, we have displayed the confusion matrix for Sync3D and can make some noteworthy conclusions. Our learning representation seemed to have worked well for the classes Hug, Kiss, and Negative however we have poorer performance on Hand Shake and High Five. Interestingly, these are the same classes in which our data sampling strategy failed to perform well. Thus, improving the data sampling has potential to improve the classification accuracy and overall improve our learnt representation.

Method	setting	curr. learning	% top1 acc.
I3D	Normal	✗	73.2
Sync3D	Easy + Medium Neg	✗	71.3
Sync3D	Hard Neg	✓	75.5

TABLE II

TOP1 CLASSIFICATION ACCURACY FOR I3D AND SYNC3D USING DIFFERENT CURRICULUM LEARNING STRATEGIES

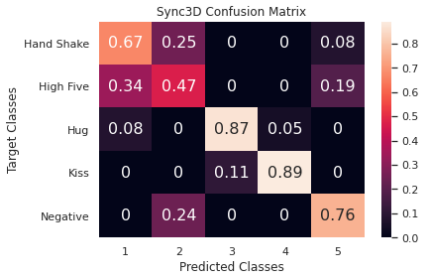

Fig. 5. Visualising the confusion matrix points to some interesting classification errors made by Sync3D. Note how the results point to more ambiguity in classifying handshakes against high fives, but both of these interactions consist of similar actions making these classifications somehow correct even while they are wrong.

V. DISCUSSION

Due to the small size of the dataset of human interactions, our current approach was not able to sufficiently learn from it. With a 50/50 split, a total of 216 data pairs were developed using the sampling strategy for training, out of which 54 pairs were labelled as interacting and 162 were negative pairs. As such the data is quite limited leading to difficulties in convergence during training. As we see in Fig. 6, while globally there is not much change observed, on a small scale we can see some fluctuations leading to minimal convergence. This leads us to believe that even in this low data scenario, Sync3D manages to learn some information about the input human tracks and their interactions.

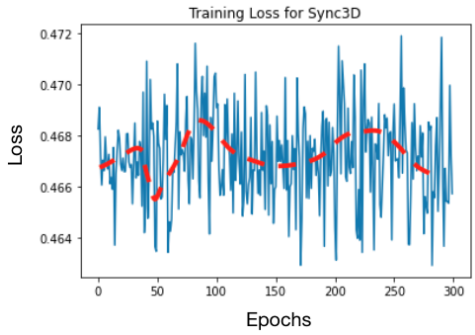

Fig. 6. Training loss achieved with training a three way classification (hug, kiss, negative) for the first 300 epochs.

Further, as we noted in Section IV A, our data sampling strategy also missed a significant portion of examples from the Hand Shake and High Five classes. To alleviate this error, one could use full body detections as opposed to upper body bounding boxes. In Fig. 7, we have displayed the two possible detection methods side by side to show the efficacy of detectron2 detections and the failure of our strategy on this simple example. We note that by using full body detections we are able to successfully get bounding box interaction as soon as the interaction is due to take place (in this example, as soon as the two characters approach each other for a hand shake). However, looking at the upper body bounding boxes, we note that even during the hand shake (i.e. when the action is taking place), the bounding boxes do not intersect leading to a false negative by our strategy. We leave this extension as future work.

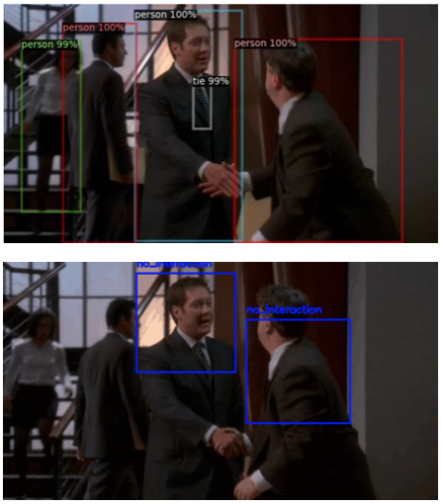

Fig. 7. **Top:** Detectron2 detection pre-Hand Shake, **Bottom:** Upper body bounding box during Hand Shake

VI. CONCLUSION

In this paper, we have presented a novel learning representation system based on self-supervision for human interactions. Here, we exploit the idea of video-audio synchronisation to perform human track synchronisation both temporally and spatially. To that extent, we propose the use of temporal and spatial alignment to sample positive and negative pairs of interacting tracks. Although the dataset led to limited accuracy on the self-supervision model itself, we note favourable performance on the downstream task of interaction recognition against the I3D features. For future work, we hope to incorporate the use of optical flow to encode frames as it has shown significant performance boost in related works and we also suggest the exploration of recurrent neural architectures to learn sequential dependence along with spatio-temporal information.

ACKNOWLEDGMENT

This project was completed under the advisory of Prof. Vicky Kalogeiton. I would like to thank her for the insightful discussions, support, and feedback in the duration of this project.

REFERENCES

[1] Yun, Kiwon, et al. "Two-person interaction detection using body-pose features and multiple instance learning." 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2012.

[2] Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset, in: Computer Vision and Pattern Recognition, (CVPR), pp. 4724–4733.

[3] MLA Stergiou, Alexandros, and Ronald Poppe. "Analyzing human–human interactions: A survey." Computer Vision and Image Understanding 188 (2019): 102799.

[4] Chung, Joon Son and Zisserman, Andrew. (2017). Out of Time: Automated Lip Sync in the Wild.

[5] Poppe, R., 2010. A survey on vision-based human action recognition. Image and vision computing 28, 976–990.

[6] Cheng, G., Wan, Y., Saudagar, A.N., Namuduri, K., Buckles, B.P., 2015. Advances in human action recognition: A survey. arXiv preprint arXiv:1501.05964 .

[7] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, 2014.

[8] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style)," IEEE Transl. J. Magn.Jpn., vol. 2, Aug. 1987, pp. 740–741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].

[9] M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.

[10] J. U. Duncombe, "Infrared navigationÑPart I: An assessment of feasibility (Periodical style)," IEEE Trans. Electron Devices, vol. ED-11, pp. 34–39, Jan. 1959.

[11] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In Proc. CVPR, 2016.

[12] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In Proc. ECCV, pages 649–666. Springer, 2016.

[13] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Proc. ECCV, pages 69–84. Springer, 2016.

[14] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020.

[15] K. He, H. Fan, A. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In Proc. CVPR, 2020.

[16] Wang, Tiancai, et al. "Learning Human-Object Interaction Detection using Interaction Points." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[17] Kerepesi, Andrea, et al. "Detection of temporal patterns in dog–human interaction." Behavioural processes 70.1 (2005): 69-79.

[18] Marín-Jiménez, Manuel, Kalogeiton, Vicky and Medina-Suárez, Pablo and Zisserman, Andrew. (2019). LAEO-Net: revisiting people Looking At Each Other in videos. 10.1109/CVPR.2019.00359.

[19] Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: Proc. CVPR. vol. 1, pp. 539–546. IEEE

[20] Patron-Perez, A., Marszalek, M., Zisserman, A., Reid, I.D., 2010. High five: Recognising human interactions in TV shows, in: British Machine Vision Conference (BMVC), p. 2.

[21] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709, 2020.