Self-Supervised Human Activity Recognition by Augmenting Generative Adversarial Networks

Mohammad Zaki Zadeh
University of Texas at Arlington, USA
Arlington, USA
mohammad.zakizadehgharie@mavs.uta.edu

Ashwin Ramesh Babu
University of Texas at Arlington
Arlington, USA

Ashish Jaiswal
University of Texas at Arlington
Arlington, USA

Maria Kyrarini
University of Texas at Arlington
Arlington, USA

Fillia Makedon
University of Texas at Arlington
Arlington, USA

ABSTRACT

This article proposes a novel approach for augmenting generative adversarial network (GAN) with a self-supervised task in order to improve its ability for encoding video representations that are useful in downstream tasks such as human activity recognition. In the proposed method, input video frames are randomly transformed by different spatial transformations, such as rotation, translation and shearing or temporal transformations such as shuffling temporal order of frames. Then discriminator is encouraged to predict the applied transformation by introducing an auxiliary loss. Subsequently, results prove superiority of the proposed method over baseline methods for providing a useful representation of videos used in human activity recognition performed on datasets such as KTH, UCF101 and Ball-Drop. Ball-Drop dataset is a specifically designed dataset for measuring executive functions in children through physically and cognitively demanding tasks. Using features from proposed method instead of baseline methods caused the top-1 classification accuracy to increase by more then 4%. Moreover, ablation study was performed to investigate the contribution of different transformations on downstream task.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**.

KEYWORDS

cognitive assessment, human-computer interaction, computer vision, deep learning

ACM Reference Format:
Mohammad Zaki Zadeh, Ashwin Ramesh Babu, Ashish Jaiswal, Maria Kyrarini, and Fillia Makedon. 2021. Self-Supervised Human Activity Recognition by Augmenting Generative Adversarial Networks. In *Petra '21: The Pervasive Technologies Related to Assistive Environments, June 29– July 02, 2021, Corfu, Greece*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3453892.3453893

Petra '21, June 29– July 02, 2021, Corfu, Greece

ACM ISBN 978-1-4503-8792-7/21/06…$15.00
https://doi.org/10.1145/3453892.3453893

1 INTRODUCTION

Recent advances in Deep Learning [20] and the challenge of gathering huge amounts of labeled data have encouraged new research in unsupervised or self-supervised learning. In particular, Computer Vision tasks could greatly benefit from successful models that learned abstract low-dimensional features of images and videos without any supervision, because unlabeled images and video sequences can be gathered automatically without human intervention [1, 4, 17].

As a result, plenty of research has been focused on methods that can adapt to new conditions without expensive human supervision. The main focus of this paper is on self-supervised visual representation learning, which is a subclass of unsupervised learning. Self-supervised learning techniques have produced state of the art low-dimensional representations on most computer vision benchmarks [5, 6, 8, 26, 37].

In self-supervised learning framework, only unlabeled data is needed in order to formulate a learning task, such as predicting context [8] or image rotation [6] for which a target objective can be computed without supervision. These methods usually incorporate Convolutional Neural Networks (CNN) [18] which after training, their intermediate layers encode high-level semantic visual representations. The obtained representations can be used for solving downstream tasks of interest, such as object detection or human activity recognition. Moreover Self-supervised learning can be employed in finding internal representations of the environment, which is useful in model-based reinforcement learning settings [15].

While most of the research in application of self-supervised learning in computer vision is concentrated on still images, the focus of this paper is human activity recognition in videos. This work is motivated by the real-world ATEC (Activate Test of Embodied Cognition) system [3, 7, 32], which assesses executive function in children through physically and cognitively demanding tasks. The system requires manually labeling hundred of hours of videos by experts. Therefore, this work exploits self-supervised learning techniques to extract low-dimensional representations of the videos.

The method proposed in this work (Figure 1) is inspired by [6] that augments Generative Adversarial Networks (GAN) [10, 31] with self-supervised rotation loss in order to improve the representation capability of discriminator network. However, the proposed

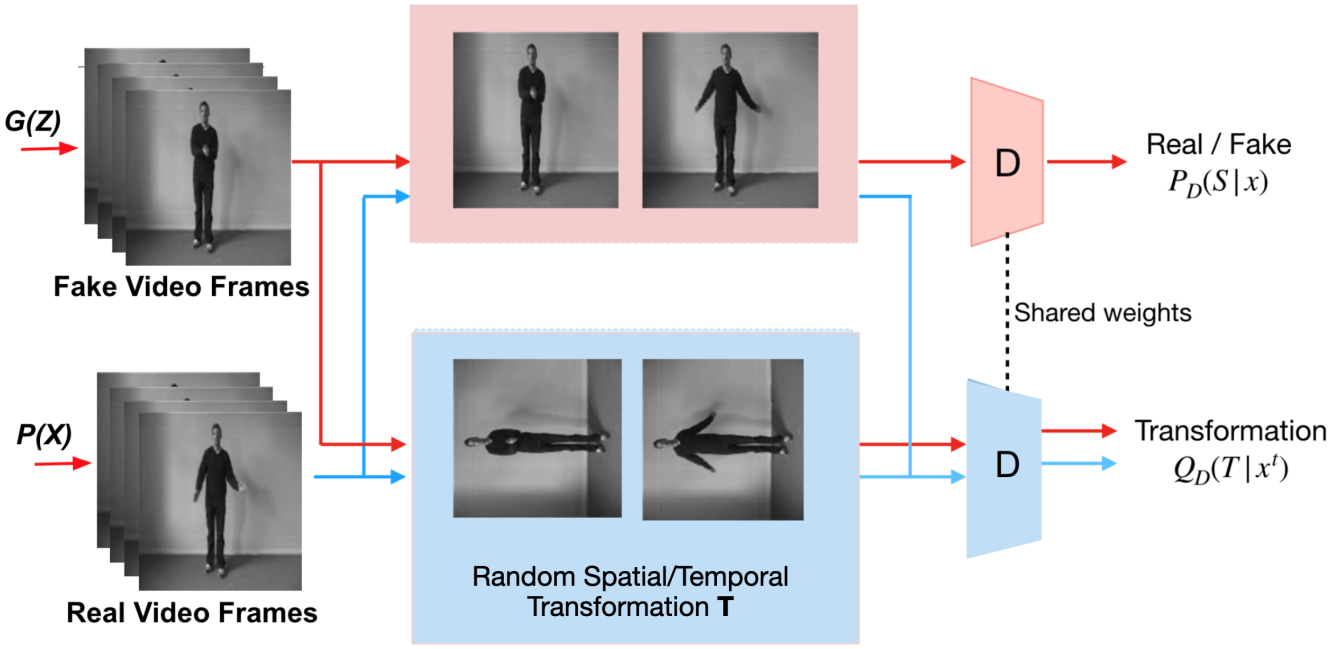

Figure 1: Proposed method architecture. Diagram is inspired by [6]

work has significant differences from the existing methods. First, the purpose of this work is to find a low-dimensional representation of videos rather than still images. Second, In [6] an auxiliary loss is added to the discriminator network to detect angles of random rotation applied on still images.

But in this work, the discriminator classifies among three different spatial transformations, such as rotation, translation or shearing and a temporal transformation that shuffles the temporal order of frames. All aforementioned transformations are randomly applied on video frames. Moreover, a thorough ablation study is performed which investigates the effect of each different transformation.

The results prove that in general the introduction of self-supervised transformation loss, improves the quality of representation provided by discriminator network. It also shows that inclusion of additional spatial transformations and temporal shuffling improves the downstream classification accuracy specially in Ball-Drop dataset. The rest of the paper is organized as follows: In section 2 a brief review of some of the recent self-supervised methods used in computer vision is provided. In section 3 the mathematical basis of self-supervised methods employed in this article is discussed. Then in section 4, results of proposed methods along with datasets and criteria used are presented. Finally, the last section includes the conclusion and directions for future research.

2 RELATED WORKS

In this section, some of the state of the art research works that employed self-supervised learning framework in computer vision are briefly summarized. Authors in [8] explore using spatial context as supervisory signal for learning image representation. For training, the authors selected two random pairs of patches from each image and tried to predict the relative position of the second patch in respect to the first one. Other efforts used colorizing gray-scale images [38] and reconstructing missing parts of an image (Image Inpainting) [30] as self-supervised task for learning features. Researchers in [27] counting the number of visual primitives in images is considered as a supervision signal. This signal is acquired without any manual annotation by using equivariance relations. Authors in [26] divided an image into 9 tiles and shuffled their position via a randomly chosen permutation from a predefined permutation set and then predicted the index of the chosen permutation. All of these patches are sent through the same network, then their representations are concatenated and passed through fully-connected multi-layer perceptrons for prediction.

Another useful supervision signal is rotation. In [9] the authors randomly rotated an input image and trained a deep convolutional neural networks to predict the rotation angle. In a similar fashion, authors in [6] augmented a generative adversarial network with a rotation loss that encourages discriminator to classify which rotation was performed on input image. Contrastive Learning [5, 14] is another interesting idea that has achieved state of the art results. In contrast to generative models that generate computationally expensive pixel-level images, contrastive learning methods perform self-supervised task in latent space. For example in [37] given masked-out patches in an input image, Contrastive Predictive Coding loss is used to learn to select the correct patch, among other distractor patches sampled from the same image to fill in the masked location. The authors employ a network of convolutional blocks to process patches followed by an attention pooling network to encode the content of unmasked patches before predicting masked

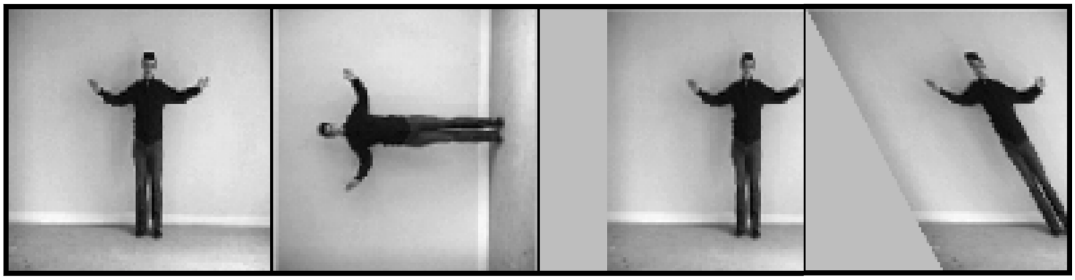

Figure 2: Examples of spatial transformation used. From left to right: Original Image, rotation, translation, shear.

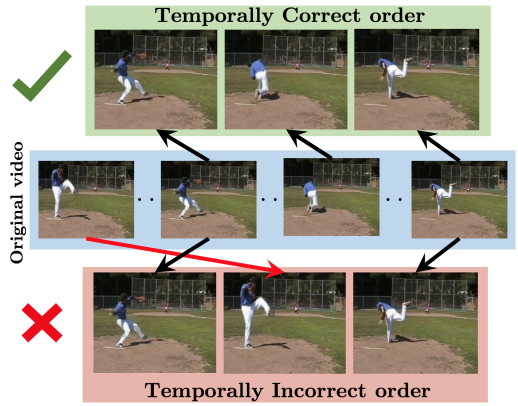

Figure 3: In temporal transformation used in proposed method, classifier tries to find if temporal order of videos frames are shuffled or not (Adopted from [24]).

ones. Furthermore self-supervised learning has been very successful in finding representation of videos. One of the most popular self-supervised task for video is prediction of future frames given past frames [2, 23]. In [12, 29] it was proven that predicting future frames or motion of objects can provide a compact representation of video that can be exploited in downstream tasks. Another interesting task is utilizing temporal order of frames by either detecting whether temporal order of frames is valid [24] or sorting shuffled frames [21].

The method proposed in this work is also focused on providing low-dimensional representation of videos. It tries to predict which random transformation, spatial or temporal has been applied on input video frames.

3 METHODOLOGY

In this section, first GAN is introduced which is the basis of methods used in this work. Subsequently the proposed self-supervised GAN is described in detail and how video representation (features) are extracted from it. These features are fed into a simple 2 layer multilayer perceptrons (MLP) network for downstream classification tasks such as human activity recognition.

3.1 Generative Adversarial Networks (GAN)

GAN [10, 31] is a framework for producing a model distribution that mimics a given target distribution, and it consists of a generator $G(z;\theta_g)$ that produces the model distribution and a discriminator $D(x;\theta_d)$ that distinguishes the model distribution from the target. Training data is denoted by x and input noise is z with probability distribution of $P_z(z)$.

In practice both generator and discriminator are implemented by differentiable CNNs with parameters: θ_g and θ_d. D is trained to maximize the probability of assigning the correct label to both training examples and samples from G. At the same time G is trained to minimize $log(1 - D(G(z)))$. In other words, D and G play the following two-player minimax game with value function $V(D, G)$:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}(x)}[logD(x)] + \mathbb{E}_{z\sim P_z(z)}[log(1 - D(G(z)))]. \quad (1)$$

But using GAN in practice is challenging because of instability occurring in training, mode collapsing, etc. However in recent years variety of novel techniques such as gradient penalty [11] or spectral normalization [25] have been proposed to solve some of the challenges.

3.2 Self-supervised Learning

One of the main problems with GANs that limits their ability for providing good representation is discriminator forgetting [6]. Because in practice as parameters of generator G varies so does the distribution P_G which causes learning process of discriminator to be non-stationary. In other words, the discriminator is not encouraged to keep a useful data representation as long as the current representation is useful to discriminate between the classes.

In order to alleviate the above problem, the discriminator network is augmented with a self-supervised task like predicting rotation angle [9] or counting objects in image [27] to motivate GAN to learn a useful compact representations. The method proposed in this work is based on spatial and temporal transformation of video frames. In this method (Figure 1), one transformation is randomly picked and applied on frames of input video. Then the self-supervised task is predicting the transformation used on video frames. As a result the loss function of both generator and discriminator are modified as follows:

Figure 4: Samples of generated images from custom Ball-Drop dataset.

$$L_G = -V(D,G) - \alpha \mathbb{E}_{x\sim P_G} \mathbb{E}_{t\sim T}[logQ_D(T=t|x^t)]$$
$$L_D = V(D,G) - \mathbb{E}_{x\sim P_{data}} \mathbb{E}_{t\sim T}[logQ_D(T=t|x^t)] \quad (2)$$

where $V(D,G)$ is the value function from Equation 1 and $t \in T$ is a transformation selected from a set of possible spatial and temporal transformations. x^t is input x transformed by transformation t, $Q_D(T|x^t)$ is discriminator distribution over possible transformations and α is self-supervised loss weight. For this method three different spatial affine transformations such as rotation, translation and shearing along with a temporal transformation, in which temporal order of video frames are shuffled, are chosen. Examples of spatial and temporal transformations are depicted in Figure 2 and 3 respectively.

For rotation only four classes were considered corresponding to rotation angles of 0°, 90°, 180° and 270°. Respectively, three classes for translation (vertical, horizontal and both), three for shearing (vertical, horizontal and both) and one class for temporal transformation (shuffled or not) were chosen. So in total eleven different transformation classes were selected.

As explained in [6], generator and discriminator are collaborative with respect to predicting the transformation task. Because for detecting the transformations, the discriminator is trained only on the true data thus the generator is motivated to generate images that are easy for discriminator to detect. As illustrated in Figure 1 the discriminator has two heads, which the former like normal GANs predicts whether non-transformed video frames are real or fake. The latter head on the other hand predicts the transformation class of transformed inputs.

After training is completed, output of the last layer before the heads is extracted as a compact representation of the input video. Then a simple 2 layer feed forward MLP is trained on extracted video representations for human activity recognition.

4 RESULT AND DISCUSSION

In this section, details of the datasets used in this experiment are discussed. This is followed by a discussion of how the neural network models are used and how they are trained. Finally, results of both baseline and proposed method are presented. It should be noted that in this article the focus is on providing compact representation of videos that can be exploited for activity recognition. Thus evaluating fidelity of generated image frames is outside scope of this paper and as a result criteria such as Frechet Inception Distance (FID) are not used.

4.1 Datasets

In this article in order to evaluate the performance of the proposed method for providing video representation useful for activity recognition three different video datasets were used. First two are publicly available video datasets like KTH [33] and UCF101 [34] containing short video clips of humans doing various activities. KTH dataset contains 6 types of human actions performed several times by 25 subjects in four different scenarios. UCF101 dataset consist of realistic action videos collected from YouTube, having 101 action categories. The third dataset which for simplicity is called Ball-Drop (Ball-Drop-to-the-Beat) is based on one of tasks designed for ATEC system (Activate Test of Embodied Cognition) to assess both audio and visual cue processing of children while performing upper-body movements. The ATEC is an assessment test designed to measure executive functions in children through physically and cognitively demanding tasks [3, 7, 32].

For this task the child is required to pass a ball from one hand to another, following audio and visual cues. Based on the instructions, the child has to pass the ball for Green-Light, keep the ball still for Red-Light, and move the ball up and down with the same hand for Yellow-Light. There are 10 different tasks based on how the light colors are presented audibly or visually. There are total of 30 subjects present in this experiment, each recording 2 versions

Table 1: Top-1 classification accuracy of using features extracted from different methods.

Method	KTH	UCF101	Ball-Drop
GAN	71.46 ± 2.5	64.68 ± 0.4	77.93 ± 2.7
GAN+Rotation	74.47 ± 2.5	66.86 ± 0.6	80.47 ± 2.5
GAN+Spatial	76.41 ± 2.0	66.95 ± 1.6	81.99 ± 4.5
GAN+Temporal	76.09 ± 3.2	70.88 ± 0.7	80.69 ± 3.7
GAN+SpatioTemporal	77.13 ± 3.6	69.17 ± 1.8	84.53 ± 3.0

Table 2: Investigating the effects of combination of different transformations on Top-1 classification accuracy.

Method	Ball-Drop
GAN	77.93 ± 2.7
GAN+Rotate	80.47 ± 2.5
GAN+Translate	80.04 ± 3.3
GAN+Shear	79.52 ± 3.3
GAN+Rotate+Translate	81.32 ± 5.1
GAN+Translate+Shear	80.33 ± 3.1
GAN+Rotate+Shear	81.01 ± 4.6
GAN+Rotate+Translate+Shear	81.99 ± 4.5

of all 10 different tasks. Each task consists of either 8 or 16 short segments that each one should be classified into 3 different classes of green-light, red-light and yellow-light. One of the main reasons that motivated authors of this article to pursue self-supervised learning is that manually annotating this dataset proved to be cumbersome and error prone.

All of datasets used in this article were divided into 3 different sets. First 80% of each dataset was considered as unlabeled and used solely for training self-supervised GANs. After training the remaining 20% (labeled) were fed into trained discriminator network to extract video representations (features). The features are extracted from penultimate layer of the discriminator network. Then again for activity recognition the features were divided into train and test set with ratio of 4 to 1.

4.2 Models

In self-supervised GANs for both generator and discriminator a 6 layer convolutional neural net (CNN) was used. Since the input is video, in discriminator the first 2 layer and for generator the last 2 layers employ 3D convolutional nets [36]. As discussed by [19, 22] performance of GANs depends on many different hyper-parameters and there is no set of hyper-parameters that guarantee superior performance on all datasets and finding one requires massive computational budget. Due to our limited computational budget, very deep complex networks such as densenet and resnet101 were avoided [13] and a small grid search was performed for tuning the hyper-parameters.

All the models, including baseline GAN and proposed self-supervised GAN were trained for 100 epochs using PyTorch framework [28] with ADAM [16] as optimizer with following parameters, which are selected empirically.

generator learning rate: 0.0001, generator learning rate: 0.0004, beta1: 0.5, beta2: 0.999. Spectral Normalization was used in all methods to stabilize the training process. And for self-supervised GAN parameter of α in equation 2 was chosen as 0.25. Finally for doing classification on extracted features a 2 layer MLP were trained with ADAM optimizer with similar hyper-parameters.

4.3 Experimental Results

In Figure 4 examples of generated images by proposed method are depicted. As stated at the start of this section, the quality of the generated images is not the focus of this paper. The real pictures of the children cannot be shown due to the privacy protection of the participants in Ball-Drop task. However, their generated images can be portrayed since faces of children are anonymized because they are blurred.

After training all the baseline and proposed methods including GAN, features (representation) of labeled video were extracted. Then, a supervised (MLP-based) human activity recognition method was trained on features and the average top-1 classification accuracy on test set was calculated by using 5-fold cross validation and presented in Table **??**. Baseline methods include GAN [10] and self-supervised GAN with only rotation as learning task (GAN+Rotation) [6] and proposed methods are self-supervised GAN with three different spatial transformations such as rotation, translation and shearing (GAN+Spatial), self-supervised GAN with only temporal transformation (shuffling) of video frames (GAN+Temporal) and finally self-supervised GAN with both spatial and temporal transformations (GAN-SpatioTemporal).

The experimental results prove superiority of the proposed method (GAN-SpatioTemporal) over baseline GAN and GAN+Rotation for providing a useful representation of videos, specially for Ball-Drop dataset which is the focus of this paper. It is also interesting to see that in UCF101 dataset, GAN+Temporal outperforms GAN+Spatial and even GAN-SpatioTemporal.

Next an ablation study is performed in order to investigate the effect of different spatial transformation used in proposed method on downstream classification accuracy. First, proposed method was trained using only one spatial transformation (rotation, translation or shearing). Then two transformation were used followed by all three. The top-1 classification accuracy of using features extracted from these methods applied on Ball-Drop dataset is shown in Table **??**. The results show that although rotation outperforms other transformation such as translation and shearing when used alone but combining different spatial transformation gains the best result.

5 CONCLUSION AND FUTURE WORKS

In this work, a novel method was proposed to augment GAN with a self-supervised task in order to improve its ability for generating useful representation of videos. The self-supervised task in this method consists of randomly picking a transformation and applying it on video frames. Subsequently, the discriminator is encouraged to predict the correct transformation that was used. The experimental results proved that in overall, introduction of new transformations in different modalities enhances the capability of baseline GAN [10] and outperforms rotation only self-supervised GAN [6] in providing a representation of videos useful for human activity recognition.

Next step would be using much deeper networks and applying this method on very large datasets, something that was beyond our computational budget at the moment. Another possible direction is to consider transformations as special case of policy in reinforcement learning [35]. Ability to find the policy that changed the state of the environment would be very useful in model-based reinforcement learning [15].

ACKNOWLEDGMENTS

This work was partially supported by National Science Foundation grants IIS 1565328 and IIP 1719031.

REFERENCES

[1] Unaiza Ahsan, Chen Sun, and Irfan Essa. 2018. DiscrimNet: Semi-Supervised Action Recognition from Videos using Generative Adversarial Networks. arXiv:1801.07230 [cs.CV]

[2] Sandra Aigner and Marco Körner. 2018. FutureGAN: Anticipating the Future Frames of Video Sequences using Spatio-Temporal 3d Convolutions in Progressively Growing GANs. arXiv:1810.01325 [cs.CV]

[3] A. R. Babu, M. Zakizadeh, J. R. Brady, D. Calderon, and F. Makedon. 2019. An Intelligent Action Recognition System to assess Cognitive Behavior for Executive Function Disorder. In *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*. 164–169. https://doi.org/10.1109/COASE.2019.8843199

[4] Javier Selva Castelló. 2018. *A Comprehensive Survey on Deep Future Frame Video Prediction.* Master's thesis. Universitat de Barcelona, The address of the publisher. An optional note.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709 [cs.LG]

[6] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. 2018. Self-Supervised GANs via Auxiliary Rotation Loss. arXiv:1811.11212

[7] Alex Dillhoff, Konstantinos Tsiakas, Ashwin Ramesh Babu, Mohammad Zakizadehghariehali, Benjamin Buchanan, Morris Bell, Vassilis Athitsos, and Fillia Makedon. 2019. An Automated Assessment System for Embodied Cognition in Children: From Motion Data to Executive Functioning. In *Proceedings of the 6th International Workshop on Sensor-Based Activity Recognition and Interaction* (Rostock, Germany) *(iWOAR '19)*. Association for Computing Machinery, New York, NY, USA, Article 9, 6 pages. https://doi.org/10.1145/3361684.3361693

[8] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. 2015. Unsupervised Visual Representation Learning by Context Prediction. arXiv:1505.05192

[9] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised Representation Learning by Predicting Image Rotations. arXiv:1803.07728 [cs.CV]

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[11] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved Training of Wasserstein GANs. arXiv:1704.00028

[12] Tengda Han, Weidi Xie, and Andrew Zisserman. 2019. Video Representation Learning by Dense Predictive Coding. arXiv:1909.04656 [cs.CV]

[13] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE. https://doi.org/10.1109/cvpr.2018.00685

[14] Ashish Jaiswal, Ashwin ramesh babu, Mohammad Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A Survey on Contrastive Self-Supervised Learning. *Technologies* 9 (12 2020), 2. https://doi.org/10.3390/technologies9010002

[15] Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Ryan Sepassi, George Tucker, and Henryk Michalewski. 2019. Model-Based Reinforcement Learning for Atari. arXiv:1903.00374

[16] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

[17] Yu Kong and Yun Fu. 2018. Human Action Recognition and Prediction: A Survey. arXiv:1806.11230 [cs.CV]

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[19] Karol Kurach, Mario Lucic, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. 2018. A Large-Scale Study on Regularization and Normalization in GANs. arXiv:1807.04720 [cs.LG]

[20] Y. LeCun, Y. Bengio, and G. Hinton. 2015. Deep Learning. *Nature* 521, 10 (2015), 436–444. https://doi.org/10.1038/nature14539

[21] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2017. Unsupervised Representation Learning by Sorting Sequences. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE. https://doi.org/10.1109/iccv.2017.79

[22] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. 2017. Are GANs Created Equal? A Large-Scale Study. arXiv:1711.10337 [stat.ML]

[23] Michael Mathieu, Camille Couprie, and Yann LeCun. 2015. Deep multi-scale video prediction beyond mean square error. arXiv:1511.05440 [cs.LG]

[24] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. 2016. Shuffle and Learn: Unsupervised Learning using Temporal Order Verification. arXiv:1603.08561 [cs.CV]

[25] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral Normalization for Generative Adversarial Networks. arXiv:1802.05957

[26] Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles, Vol. 9910. 69–84. https://doi.org/10.1007/978-3-319-46466-4_5

[27] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. 2017. Representation Learning by Learning to Count. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE. https://doi.org/10.1109/iccv.2017.628

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703 [cs.LG]

[29] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. 2016. Learning Features by Watching Objects Move. arXiv:1612.06370 [cs.CV]

[30] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. 2016. Context Encoders: Feature Learning by Inpainting. arXiv:1604.07379 [cs.CV]

[31] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv:1511.06434

[32] Ashwin ramesh babu, Mohammad Zadeh, Ashish Jaiswal, Alexis Lueckenhoff, Maria Kyrarini, and Fillia Makedon. 2020. A Multi-modal System to Assess Cognition in Children from their Physical Movements. https://doi.org/10.1145/3382507.3418829

[33] Christian Schüldt, Ivan Laptev, and Barbara Caputo. 2004. Recognizing human actions: A local SVM approach. *Proceedings - International Conference on Pattern Recognition* 3, 32 – 36 Vol.3. https://doi.org/10.1109/ICPR.2004.1334462

[34] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv:1212.0402 [cs.CV]

[35] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction.* A Bradford Book, Cambridge, MA, USA.

[36] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE. https://doi.org/10.1109/cvpr.2018.00675

[37] Trieu H. Trinh, Minh-Thang Luong, and Quoc V. Le. 2019. Selfie: Self-supervised Pretraining for Image Embedding. arXiv:1906.02940 [cs.LG]

[38] Richard Zhang, Phillip Isola, and Alexei A. Efros. 2016. Colorful Image Colorization. arXiv:1603.08511 [cs.CV]