Towards Efficient COVID-19 Detection with Skeleton-Based Action Synthesis

Yifan Jiang
School of Electrical and Computer Engineering
Korea University, Seoul, Korea
yfjiang@korea.ac.kr

Han Chen
School of Electrical and Computer Engineering
Korea University, Seoul, Korea
jessicachan@korea.ac.kr

Hanseok Ko
School of Electrical and Computer Engineering
Korea University, Seoul, Korea
hsko@korea.ac.kr

Abstract

Over the past two years, Coronavirus disease 2019 (COVID-19) caused a significant loss on both the economy and public health. As a result, researchers began to notice the importance of detecting a novel disease at an early stage. Leveraging from the rapid development of artificial intelligence (AI), many AI-based COVID-19 detection approaches have been presented recently. However, a well-trained AI model always demands high-quality and sufficient data, which is hard to obtain during an early spreading stage of the novel disease. In this paper, we present a novel skeleton-based action synthesis method, which can effectively generate high-quality and diverse sequential actions corresponding to the most common COVID-19 symptoms. In order to obtain synthetic action sequences, we propose a GAN inversion method that can generate a series of synthetic action sequences by editing a real action sequence. We evaluate the proposed method on a well-known action recognition dataset in two levels: skeleton quality and action recognition. The results show that our method outperforms the baseline approaches, and it can also be effectively deployed to action recognition modules with significant performance improvement.

1. Introduction

As of 29 July 2021, there had been more than 195 million COVID-19 [1] confirmed cases, and more than 4 million deaths all around the world [2]. The huge damage caused by the pandemic reminds everyone that if some early precautions have been taken during the initial stage of transmission, we may stop the pandemic from the early outbreak. There are many evidences show that the precaution measures like mass testing [3], social distance [4], wearing a mask [5] and transportation restriction [6] can be helpful to flatten the curve on COVID-19.

In the case of mass testing, testing efficiency and accuracy are the keys to success. Recently, the development of surveillance technology makes it possible to remotely detect potential COVID-19 patients. Action recognition is one of the novel surveillance technologies which is reinforced by AI research. Comparing to the traditional rRT-PCR, Action recognition techniques can reduce the testing time from 24 hours to 1 minute without any direct contact with the potential patient [7]. On the other hand, action recognition-based COVID-19 detection methods can operate 24/7 with more precise diagnostic performance to a human expert, leveraging from AI and computer vision technology supports [8]. Therefore, AI-based COVID-19 action recognition detection methods are getting more attention due to their high efficiency and sensitivity recently.

However, there are always not easy to obtain a well-performed AI diagnostic model due to many objective reasons. One of the most important factors is the data. It is obvious that a large-scale and well-labeled dataset can significantly improve the performance of AI-based action recognition algorithms, but it is always hard to build at especially the early stage of the outbreak. The reasons are summarized as follow: (1) the patient number keeps low, it limits the amount and the diversity of the data samples; (2) The novel disease does not cause alert among people, that means no sufficient resource is putting in the data collection work; (3) Building a large scale action recognition dataset requires the great effort of field operation. Therefore, how to obtain enough data at the early outbreak becomes a crucial problem before we are ready to face the next possible crisis.

In this paper, we propose a novel action synthesis algorithm, which is designed for improving COVID-19 and medical condition-related action recognition performance under the data-shortage condition. Our proposed method basically works as a GAN inversion algorithm that can generate a series of augmented action sequences with a single reference. With certain input of COVID-19 symptom action sequence, we can iteratively optimize a StyleGAN2 [9] generator and feature mapping network with this input so that we can obtain a bunch of action sequences that are similar to the reference but sightly different. Then, the action recognition algorithm can leverage these synthetic samples and find a significant improvement of its COVID-19 detection task. In particular,

(1) we propose a GAN inversion approach for improving COVID-19 and medical condition related action recognition performance. It is specially designed and optimized for handling the data-shortage problem at the very early stage of a novel disease.

(2) The proposed method can effectively generate a high-quality skeleton sequence, also brings diversity to the training set. The synthetic skeleton sequences are evaluated with image quality metrics, and the proposed method outperforms the baseline StyleGAN2 methods.

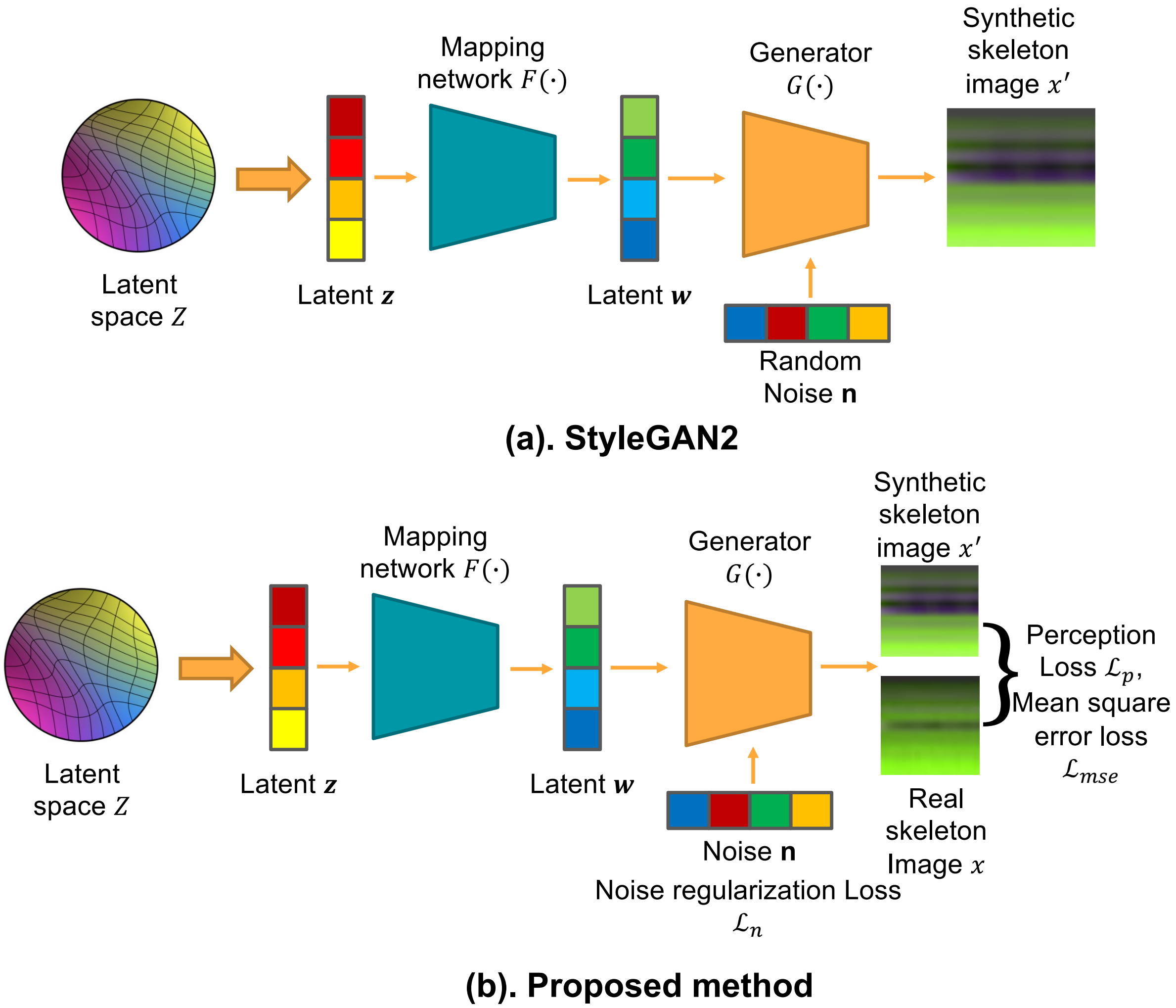
(3) The proposed method proves the synthetic action sequences can be useful for downstream tasks, for instance, COVID-19 detection and medical condition evaluation. Furthermore, the proposed method has the potential to be applied to other novel disease detection tasks during the early outbreak stage.

2. Related works

Generative adversarial networks. There are many efforts haven done following the invention of generative adversarial networks (GANs) [10]. The main focus lands on both the unconditional image modeling task and the conditional image modeling task. In the case of the unconditional image modeling task. StyleGAN [11] and its variant StyleGAN2 [9] achieve great success on image synthesis task with the capability of generating photo-realistic natural images, leveraging the advanced architecture and refined object functions. Moreover, BigGAN [12] is able to produce high-quality and diverse object images by training on large scale datasets like ImageNet [13]. For the conditional image modeling task, the networks take extra information as input in order to make the image modeling process more controllable. Text-to-image methods [14, 15] utilize texts to generate high-quality image with fine details. Pix2PixHD [16], and SEAN [17] conditions the model with a segmentation mask to fine control both the location and shape of a specific object. Due to competitive performance and high flexibility, conditional GANs (cGAN) are becoming a new research trend. However, the high data quality requirement limits the practical deployment of cGAN models.

Skeleton-based action recognition. Nowadays, action recognition gains more and more attention and has become an active area due to its high efficiency and the development of deep learning technology. With the rapid evolution of pose estimation methods, skeleton-based action recognition approaches are boosted by the high-quality skeleton data, which is obtained from advanced pose estimation methods. For skeleton-based action recognition approaches, there are three mainstreams, which are RNN based methods [18, 19, 20, 21, 22], CNN based methods [23, 24, 25, 26] and GCN based methods [27, 28, 29]. In the case of RNN based methods, they mainly replied on RNN structure as a long-term temporal learner, which is able to obtain long-range temporal information from the input videos. Ref [19] is a Siamese structure that can take both spatial and temporal information at the same time. Liu et al. [18] tried to learn the relationship from one dataset to the other. More recently, [22] combine attention mechanism with the RNN model and designed a special temporal attention module which is used for grabbing attention information from the temporal domain of input skeleton sequences. For CNN-based methods, ref [23] brought a new encoding strategy for skeleton data and mapped them into images. Ref [25] also focused on the encoding method, and this work considered both joint motion and temporal information from video together. An end-to-end manner based method [24] was brought in order to utilize different level feature representations. GCN based methods are the most popular stream of action recognition domain. ST-GCN [27] began to use a graph to represent the spatial and temporal information of skeleton joints. The potential is not limited by predicting current action, and ref [28, 29] enabled to predict next action from current skeleton inputs.

Handling data shortage for COVID-19 detection. Although the data collection process of COVID-19 CT scans started at the very beginning of the outbreak, the number of well-labeled data (for segmentation task) stays thousand-level until recently [30]. On the one hand, some efforts have been made in order to overcome the data shortage issue by adjusting the model structure or the training scheme to a few-shot condition. Chen et al. [31] applied contrastive learning to pre-train a Siamese encoder then use the pre-trained encoder to extract features, the features above are used to train a prototypical network for few-shot COVID-19 diagnostic task. Lai et al. [32] presented a deep neural network structure that can effectively handle the few-shot diagnostic task using predicted lesion segmentation. Our previous work [33] proposed a Siamese network structure that can handle the few-shot COVID-19 diagnostic task using a series of cross-domain losses. On the other hand, some works have been done for generating high-quality synthetic


Latent space Z
Latent $\boldsymbol{z}$
Mapping network $F(\cdot)$
Latent $\boldsymbol{w}$
Generator $G(\cdot)$
Synthetic skeleton image x'
Random Noise **n**
Latent space Z
Latent $\boldsymbol{z}$
Mapping network $F(\cdot)$
Latent $\boldsymbol{w}$
Generator $G(\cdot)$
Synthetic skeleton image x'
Perception Loss $\mathcal{L}_p$,
Mean square error loss $\mathcal{L}_{mse}$
Noise **n**
Noise regularization Loss $\mathcal{L}_n$
Real skeleton Image x

(a). StyleGAN2

(b). Proposed method

Figure 1. Overview of the proposed method. Subfigure (a) indicates the original StyleGAN2 generation approach. Subfigure (b) indicates the proposed method.

COVID-19 CT images in order to overcome the data shortage problem. Liu et al. [34] proposed a conditional generative adversarial network (cGAN) based method that utilized a 3D noise mask to conditionally transform a normal CT scan to an abnormal one. Li et al. [35] presented a cGAN based COVID-19 CT image synthesizer that can perform the synthesis procedure with a pair of COVID-19 and non-COVID-19 CT scans, which is connected by an elastic registration algorithm. Furthermore, our previous work [36] showed a possible way to create a large scale COVID-19 CT dataset with an advanced cGAN backbone. Even though there are already some successes that have been achieved for handling data shortage problems for COVID-19 diagnosis using medical images until now, the data shortage problem is still a critical concern that may limit the action recognition methods when face with novel diseases in the future.

3. Towards Efficient COVID-19 Detection with Skeleton-Based Action Synthesis

We introduce the proposed method in this section with details. In Figure 1, we demonstrate the overview of the proposed method and compare it to the original StyleGAN2 generation approach. Basically, a traditional StyleGAN2 generation process starts from sampling a latent vector z from latent space Z and map it into latent vector w with a mapping network $F(\cdot)$. Then, generator $G(\cdot)$ takes both latent vector w and random noise vector n and synthesizes a fake skeleton image. The drawback of the traditional StyleGAN2 generation process is that it is not controllable for certain action categories, for instance, coughing and blowing the nose. Comparing to the traditional StyleGAN2 generation process, the proposed method takes a reference input of a real skeleton image. Rather than generating a fake image with a random category, the proposed method can generate synthetic images with a certain category that is conditioned by the reference input.

3.1. StyleGAN2 preliminaries

In this paper, we utilize the fundamental structure of StyleGAN2. We demonstrate the generation process of StyleGAN in subfigure (a) of 1. In the very beginning, the generation process starts with a sampling operation that collects a latent vector z from latent space Z. Then, a feature mapping network $F(\cdot)$ is used to map latent vector z to another latent space W, and we can get a new latent vector w. At this time, the generator of StyleGAN2 takes inputs of latent vector w, random noise n, and randomly generates a synthetic skeleton image x'. Therefore, the whole generation process is formulated as $x' = G(w) = G(F(z))$.

3.2. Proposed method

The critical downside of traditional StyleGAN2 generation is it can not produce synthetic images with a specific category. It can only randomly generate synthetic samples with random categories and relatively lower quality. Above drawback limit StyleGAN2 being applied to skeleton-based action synthesis task. Therefore, we propose a novel method that tackles the problems of StyleGAN2 by developing a controllable action synthesis algorithm with a StyleGAN2 backbone.

3.2.1 Initialization for feature mapping network and generator

In the first step, we initialize both feature mapping network $F(\cdot)$ and $G(\cdot)$ through a pretraining procedure. Below object function is used to optimize two networks:

$$\mathcal{L}(x, z, F, G) = l_1(G(F(z)), x) + \mathcal{L}_p(G(F(z)), x) \quad (1)$$

where l_1 indicates L1 distance and $\mathcal{L}_p$ indicates LPIPS distance [37]. Therefore, the optimized feature mapping network F^* and optimized generator G^* can be obtained by following optimization procedures:

$$F^* = arg \min_F \mathbb{E}_{z,x} \mathcal{L}(x, z, F^*, G) \quad (2)$$

$$G^* = arg \min_G \mathbb{G}_{z,x} \mathcal{L}(x, z, F, G^*) \quad (3)$$

3.2.2 Controllable generation process

In subfigure (b) of 1, we demonstrate the controllable generation process of the proposed method. After the initialization for the feature mapping network and generator, we can obtain an optimized feature mapping network F^* and optimized generator G^*, and they are ready for the next step. For a specific input x, we treat it as a reference input, and our proposed method can generate a synthetic image that has the same label as the reference input has.

First, the generation process starts with sampling a latent vector z and mapping it to latent vector w. Then, generator G takes noise vector n and latent vector w to generate synthetic image x'.

Here, we introduce three loss functions to manage the difference between reference image x and synthetic image x'. Perception loss $\mathcal{L}_p$ is the same as we used in the stage of initialization. This loss is to manage the perceptual (high-level semantic information) similarity between real images and synthetic images. It is computed between two patches (x, x_0) using the l_2 distance in the channel dimension and average across spatial dimensions and given convolutional layers of different networks, and the details are shown in Equation 4. It ensures that synthetic skeleton image is sightly different on tiny aspect but with the same category.

$$\mathcal{L}_p = \sum_l \frac{1}{H_l W_l} ||w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})||^2_2 \quad (4)$$

where H and W represent the height and width of the feature map, $\hat{y}^l$ is the feature map from of the l^{th} layer.

A mean square error loss $\mathcal{L}_{mse}$ is utilized to maintain the action category of the synthetic image through managing the content (low-level structure information) difference between real image and synthetic image. It is a l_2 loss which is shown as in Equation 6:

$$\mathcal{L}_{mse} = \frac{1}{HWC} ||y - y_0||^2_2 \quad (5)$$

where H, W, and C represent the height, width and channel number of the feature map, respectively.

The third loss is a noise normalization loss $\mathcal{L}_n$ which follows StyleGAN2 in order to avoid the optimization from sneaking actual signal into noises. To sum up, the objective function is shown in Equation **??**.

$$\mathcal{L}_{overall} = \mathcal{L}_p + \alpha \mathcal{L}_{mse} + \beta \mathcal{L}_n \quad (6)$$

where $\alpha$and β represent the weight of the mean square error loss and noise normalization loss, respectively.

We run 200 iterations optimization process for each specific reference and generate the synthetic skeleton image after the optimization process.

4. Experiments

4.1. Dataset and experimental settings

In the experiments, we conduct all the evaluations under the NTU RGB+D 120 [38] dataset, which is a large-scale

Table 1. Skeleton quality evaluation results of synthetic action sequences. (The best evaluation score is marked in bold. $\uparrow$ means higher number is better, and $\downarrow$ indicates lower number is better.)

Settings	COVID-19-NTU_X-Set		COVID-19-NTU_X-Sub		Med-NTU_X-Set		Med-NTU_X-Sub	
Metrics	PR ($\uparrow$)	FID ($\downarrow$)	PR ($\uparrow$)	FID ($\downarrow$)	PR ($\uparrow$)	FID ($\downarrow$)	PR ($\uparrow$)	FID ($\downarrow$)
OURS	0.2667 / **0.0118**	**53.91**	**0.6169** / 0.0	**36.64**	**0.4493** / 0.0	**30.40**	**0.5627 / 0.0013**	**37.33**
StyleGAN2 [9]	**0.3161** / 0.0011	54.49	0.4688 / 0.0	85.71	0.3091 / 0.0	89.79	0.4462 / 0.0	116.29

Table 2. Dataset settings of action recognition evaluation. (The best evaluation score is marked in bold. $\uparrow$ means higher number is better, and $\downarrow$ indicates lower number is better.)

Evaluation	Training set	Projection set	Test set
A	100% training set	-	100%-n% test set
B	100% training set + n% data from test set	-	100%-n% test set
C	100% training set + n% data from synthetic data	n% data from test set	100%-n% test set
D	100% training set + 3n% data from synthetic data	n% data from test set	100%-n% test set
E	100% training set + 5n% data from synthetic data	n% data from test set	100%-n% test set

video dataset for the action recognition task. It contains a total of 120 different action categories, 114,480 video clips ranging from daily actions to two-person interactions. In this paper, we focus on medical condition related categories, which yield A41: sneeze/cough, A42: staggering, A43: falling down, A44: headache, A45: chest pain, A46: back pain, A47: neck pain, A48: nausea/vomiting, A49: fan self, A103: yawn, A104: stretch oneself, A105: blow nose. Besides, we also focus on COVID-19 common symptoms [39], which are A41: sneeze/cough, A42: staggering, A43: falling down, A44: headache, A48: nausea/vomiting, A105: blow nose. Furthermore, we follow NTU RGB+D 120 [38] to set up two benchmarks: Cross-Setup and Cross-Subject. The former split dataset into a training set and testing set by considering the subject's height and recording distance. While the latter split dataset on subjects' level. To sum up, there are four categories of evaluation: COVID-19-NTU_X-Set, COVID-19-NTU_X-Sub, Med-NTU_X-Set and Med-NTU_X-Sub. Where COVID-19 represents six categories of COVID-19 symptoms, and Med means twelve categories of medical conditions. X-Set and X-Sub represent Cross-Setup evaluation and Cross-Subject evaluation, respectively.

4.2. Quantitative Results

We divide the experiments into two parts: skeleton quality evaluation and action recognition evaluation.

4.2.1 Skeleton quality evaluation

In this study, we utilize two metrics to evaluate the skeleton quality, the metrics are Precision and Recall (PR) [40], and Fréchet inception distance (FID) [41]. The comparison results are demonstrated in Table 1.

The proposed method outperforms StyleGAN2 in most of the experimental settings. That shows the skeleton quality of synthetic data is reliable and has better quality than StyleGAN2.

4.2.2 Action recognition evaluation

In this part of the experiments, we evaluate how much the synthetic action sequences can boost action recognition approach performance. We set up six evaluations with different dataset settings, which are denoted as evaluation A to evaluation F. The dataset splits are shown in Table 2. We use $n = 10$ to randomly sample 10% data from the test set in order to form the dataset for the above five evaluations. Furthermore, we use accuracy (Acc) and F1 score (F1) metrics to evaluate the action recognition performance. The action recognition model here is MobileNetV2 [42].

The action recognition results are presented in Table 3. We can learn that the action recognition model achieves the best performance in evaluation D in the case of COVID-19 settings. That means synthetic data can effectively help to improve action recognition performance. In the case of Medical condition settings, synthetic data show that it maintains similar high-quality comparing to real data since comparable results are conducted in evaluation C to E.

4.3. Qualitative Results

In order to intuitively present synthetic skeleton results, we show two examples in Figure 2. We show coughing and blowing nose action categories, and frame-by-frame results tell that synthetic action reflects action characteristics well. Also, the skeleton is stable and diverse. It ensures the skeleton quality and the improvement of the downstream task.

5. Conclusions

In this paper, we introduced a novel skeleton-based action synthesis method for improving the performance of the action recognition approach. The proposed method was designed for generating COVID-19 symptoms and medical-

Table 3. Action recognition evaluation results. (The best evaluation score is marked in bold. ↑ means higher number is better, and ↓ indicates lower number is better.)

Settings	COVID-19-NTU_X-Set		COVID-19-NTU_X-Sub		Med-NTU_X-Set		Med-NTU_X-Sub	
Metrics	Acc (↑)	F1 (↑)	Acc (↑)	F1 (↑)	Acc (↑)	F1 (↑)	Acc (↑)	F1 (↑)
Evaluation A	60.87	61.90	62.33	62.20	**52.40**	**52.48**	52.83	51.48
Evaluation B	63.31	63.98	61.19	61.70	50.69	50.83	**56.20**	**55.37**
Evaluation C	60.50	60.55	63.24	62.15	51.30	52.12	53.17	50.17
Evaluation D	**66.26**	**66.11**	**65.75**	**66.61**	48.84	48.12	49.70	48.38
Evaluation E	63.97	64.56	65.64	63.82	49.05	49.03	52.52	51.36

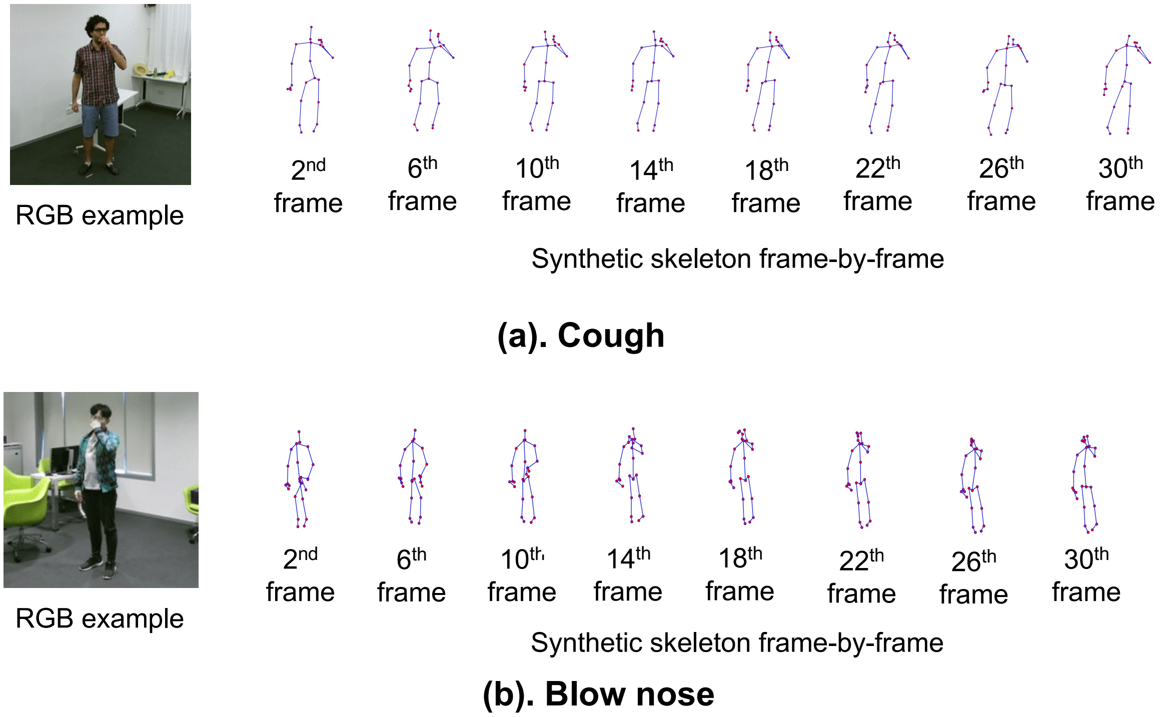


Figure 2. Two examples of synthetic action sequences. Subfigure (a) indicates the Cough action category, while Subfigure (b) indicates the Blow nose action category.

condition related action sequences. It took a single skeleton image as a reference so that it is able to generate a series of synthetic action sequences, which were similar to the reference but different on detail level. With the experiments on skeleton quality and action recognition, the proposed method proved its capability of high-quality action sequence synthesis tasks. Also, the synthetic action sequences can help to boost the action recognition performance. Furthermore, the proposed method has potential to be utilized in other downstream task likes pose estimation for fighting against COVID-19 pandemic.

6. Acknowledgment

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA2386-19-1-4001.

References

[1] Coronavirus disease (covid-19) pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019. 1

[2] Johns hopkins coronavirus resource center. https://coronavirus.jhu.edu/map.html. 1

[3] Luís Carlos Lopes-Júnior, Emiliana Bomfim, Denise Sayuri Calheiros da Silveira, Raphael Manhães Pessanha, Sara Isabel Pimentel Carvalho Schuab, and Regina Aparecida Garcia Lima. Effectiveness of mass testing for control of covid-19: a systematic review protocol. *BMJ open*, 10(8):e040413, 2020. 1

[4] Chanjuan Sun and Zhiqiang Zhai. The efficacy of social distance and ventilation effectiveness in preventing covid-19 transmission. *Sustainable cities and society*, 62:102390, 2020. 1

[5] Cornelia Betsch, Lars Korn, Philipp Sprengholz, Lisa Felgendreff, Sarah Eitze, Philipp Schmid, and Robert Böhm. Social and behavioral consequences of mask policies during the covid-19 pandemic. *Proceedings of the National Academy of Sciences*, 117(36):21851–21853, 2020. 1

[6] Jin Shen, Hongyang Duan, Baoying Zhang, Jiaqi Wang, John S Ji, Jiao Wang, Lijun Pan, Xianliang Wang, Kangfeng Zhao, Bo Ying, et al. Prevention and control of covid-19 in public transportation: experience from china. *Environmental pollution*, page 115291, 2020. 1

[7] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, and Liming Xia. Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: a report of 1014 cases. *Radiology*, page 200642, 2020. 1

[8] Yicheng Fang, Huangqi Zhang, Jicheng Xie, Minjie Lin, Lingjun Ying, Peipei Pang, and Wenbin Ji. Sensitivity of chest ct for covid-19: comparison to rt-pcr. *Radiology*, page 200432, 2020. 1

[9] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 2, 5

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[11] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2

[12] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[14] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020. 2

[15] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019. 2

[16] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 2

[17] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization, 2019. 2

[18] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer, 2016. 2

[19] Hongsong Wang and Liang Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 499–508, 2017. 2

[20] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1012–1020, 2017. 2

[21] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7912–7921, 2019. 2

[22] Ce Li, Chunyu Xie, Baochang Zhang, Jungong Han, Xiantong Zhen, and Jie Chen. Memory attention networks for skeleton-based action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 2

[23] Bo Li, Yuchao Dai, Xuelian Cheng, Huahui Chen, Yi Lin, and Mingyi He. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 601–604. IEEE, 2017. 2

[24] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055*, 2018. 2

[25] Carlos Caetano, Jessica Sena, François Brémond, Jefersson A Dos Santos, and William Robson Schwartz. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019. 2

[26] Yanshan Li, Rongjie Xia, Xing Liu, and Qinghua Huang. Learning shape-motion representations from geometric algebra spatio-temporal model for skeleton-based action recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1066–1071. IEEE, 2019. 2

[27] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 2

[28] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3595–3603, 2019. 2

[29] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019. 2

[30] Covid-19 ct segmentation dataset. `https://medicalsegmentation.com/covid19/`. 2

[31] Xiaocong Chen, Lina Yao, Tao Zhou, Jinming Dong, and Yu Zhang. Momentum contrastive learning for few-shot covid-19 diagnosis from chest ct images. *Pattern Recognition*, 113:107826, 2021. 2

[32] Yaoming Lai, Guangming Li, Dongmei Wu, Wanmin Lian, Cheng Li, Junzhang Tian, Xiaofen Ma, Hui Chen, Wen Xu, Jun Wei, et al. 2019 novel coronavirus-infected pneumonia on ct: A feasibility study of few-shot learning for computerized diagnosis of emergency diseases. *IEEE Access*, 8:194158–194165, 2020. 2

[33] Yifan Jiang, Han Chen, David K Han, and Hanseok Ko. Few-shot learning for ct scan based covid-19 diagnosis. *arXiv preprint arXiv:2102.00596*, 2021. 2

[34] Siqi Liu, Bogdan Georgescu, Zhoubing Xu, Youngjin Yoo, Guillaume Chabin, Shikha Chaganti, Sasa Grbic, Sebastian Piat, Brian Teixeira, Abishek Balachandran, et al. 3d tomographic pattern synthesis for enhancing the quantification of covid-19. *arXiv preprint arXiv:2005.01903*, 2020. 3

[35] Heng Li, Yan Hu, Sanqian Li, Wenjun Lin, Peng Liu, Risa Higashita, and Jiang Liu. Ct scan synthesis for promoting computer-aided diagnosis capacity of covid-19. In *International Conference on Intelligent Computing*, pages 413–422. Springer, 2020. 3

[36] Yifan Jiang, Han Chen, MH Loew, and Hanseok Ko. Covid-19 ct image synthesis with a conditional generative adversarial network. *IEEE Journal of Biomedical and Health Informatics*, 2020. 3

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4

[38] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2020. 4, 5

[39] Coronavirus disease (covid-19) symptoms. `https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html`. 5

[40] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *arXiv preprint arXiv:1904.06991*, 2019. 5

[41] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 5

[42] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 5