SYNPOSE: A LARGE-SCALE AND DENSELY ANNOTATED SYNTHETIC DATASET FOR HUMAN POSE ESTIMATION IN CLASSROOM

Zefang Yu, Yangcheng Li, Yicheng Liu, Ting Liu, Yuzhuo Fu

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China

ABSTRACT

Deep learning-based methods for human pose estimation require large volumes of training data to achieve superior performance. However, data acquisition in classroom environments raises privacy concerns, which will undoubtedly hinder the development of the latest deep learning techniques in education domain. Due to the absence of large, richly annotated classroom datasets, research into classroom observation has had to be done by manually collecting and annotating datasets. Unfortunately, the annotation of such data is time-consuming and challenging in over-crowded classrooms. To break through these limitations, we open source SynPose, a large, densely labeled synthetic dataset specifically designed for crowded human pose estimation in classroom and meeting scenarios. Moreover, we propose a novel CTGAN to bridge the domain gap. Comprehensive experiments on real-world classroom images show that our proposed dataset and method deliver important performance benefits compared to existing datasets, revealing the potential of SynPose for future studies.

***Index Terms*—** Synthetic dataset, human pose estimation, domain adaptation, student behavior, classroom

1. INTRODUCTION

The analysis of students' behavior in classrooms can reflect students' engagement and provide an important basis for optimizing teaching strategies[1, 2, 3]. In recent years, various studies[4, 5, 6, 7, 8] have applied deep learning-based human behavior detection algorithms to classroom scenarios. Although these approaches have achieved impressive results, we argue that there still exist some limitations.

First, the aforementioned studies are basically based on self-built datasets. They manually collected and labelled images of real classrooms, which are time-consuming and labour-intensive. Besides, care must be taken with the data acquisition for reasons of privacy. In fact, the European Union already passed laws such as General Data Protection Regulations(GDPR)[9] to guarantee data privacy. Recently, a popular person re-identification dataset, DukeMTMC[10], was taken offline for privacy concerns. So far, none of the aforementioned self-built datasets are publicly available. Such phenomenon of *data silos* makes deep learning-based classroom observation costly and difficult to reproduction.

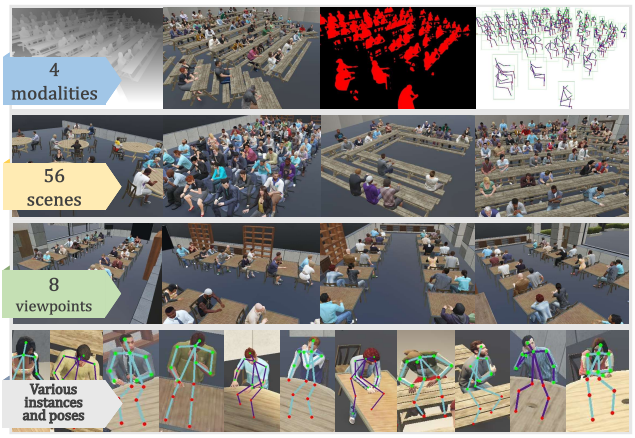

Fig. 1. Illustration of the proposed SynPose dataset about different scenes, annotations, viewpoints, modalities and poses.

Furthermore, compared to object-detection-based methods, the application of pose-estimation-based methods in classroom observation has not been well studied. As an important branch of human behavior analysis, the pose-estimation-based methods firstly generate body-joints in the image, then perform behavior analysis algorithms on the generated skeleton sequence, which is more robust to background, resolution and irrelevant objects. However, publicly available 2D multi-person pose estimation datasets[11, 12, 13, 14] contain on average less than 10 people per image, which is far from a real classroom scenario where students are densely distributed, seated and covered by a large number of desks and chairs. Self-built datasets cannot solve this problem since labeling skeletons of students in highly crowded classrooms can be challenging even for human annotators and may introduce errors in training data.

In this paper, we attempt to remedy the above problems by employing the virtual world. Firstly, we build **SynPose**, a large, densely labeled synthetic dataset specifically designed for crowded human pose estimation in classroom and meeting scenarios. SynPose consists of images in several generated classroom-like scenes captured by varying

This research was supported by the National Natural Science Foundation of China under Project (Grant No.61977045).

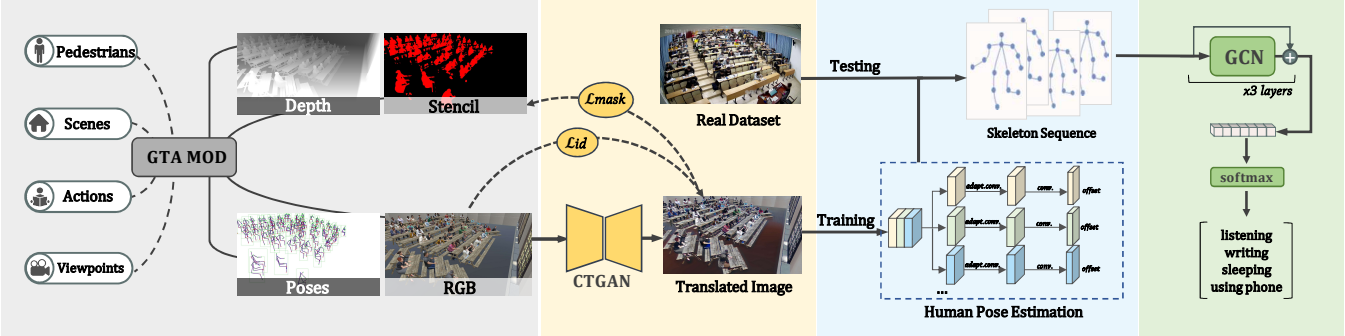

Fig. 2. The procedure of our proposed pose-estimation-based student behavior analysis framework, which consists of (1) synthetic dataset generation, (2) domain adaptation, (3) human pose estimation, and (4) student behavior classification.

camera viewpoints. Moreover, every image comes with accurate ground-truth annotations of visible and occluded body pose coordinates along with bounding-boxes for each instance. Secondly, we propose a novel **C**lassroom-style **T**ransfer **G**enerative **A**dversarial **N**etwork(CTGAN) to reduce the domain gap by transferring synthetic images to realistic classroom style. Furthermore, we design a pose-estimation-based student behavior analysis approach by exploiting the proposed SynPose dataset. Experimental results show that our dataset and domain adaptation method can effectively improve the performance of estimating human poses in real-world classroom scenes, especially in images with heavy occlusion and non-frontal viewpoints.

To summarize, the main contributions of this paper are the following: (1) We open source SynPose, a large synthetic dataset for crowded human pose estimation in classroom and meeting scenarios with accurate human pose and bounding-box annotations.(2) We propose a novel CTGAN by which the domain gap between synthetic and real data can be significantly reduced.(3) We design a pose-estimation-based student behavior analysis approach. Experimental results demonstrate the effectiveness of our dataset and method.

2. SYNPOSE DATASET

2.1. Dataset generation

To generate SynPose, we develop a data collector and labeler by exploiting Grand Theft Auto V (GTA V), a highly photo-realistic open-world game developed by Rockstar North. We wrote a custom mod using Script Hook V[15] to construct various classroom-like scenes via exploiting the objects of virtual world[1]. Then, the collector captures images from the constructed scenes in different camera viewpoints. Finally, by analyzing the data from rendering stencil, the labeler automatically annotates the accurate skeleton coordinates and bounding-boxes of each instance without any manpower.

[1]The publisher of GTA V allows for non-commercial and research uses in PC single mode[16].

2.2. Properties of SynPose

SynPose dataset comprises 60,480 images of 2,129k instances, with resolution of 1920×1080. Table 1 compares the basic information of SynPose and existing 2D human pose estimation datasets. For each dataset, we report the numbers of annotated images, instances per image, joints number and the availability of different modalities. We summarize the properties of SynPose into the following aspects:

Scenes. We built up 56 different classroom-like scenes by developing a GTA V custom map. Each scene has 8 surveillance cameras equipped with different locations and rotations to increase the diversity of viewpoints of captured images.

Instances. Instances in SynPose are from 551 pedestrian models provided by the GTA V. Each instance randomly performed one of the 80 sitting movements (e.g., using a computer, reading, etc.).

Joints. We obtained 59 joints of each instance by calling GTA V script native functions, including 6 face joints, 21 body joints and 32 finger joints.

Annotations. Every image comes with precise annotations of visible and occluded body joints, bounding boxes, stencil mask labels for instances, and depth maps.

We leave further details on our dataset homepage due to space limitation. Please check https://yuzefang96.github.io/SynPose/ for more information about dataset visualization, demonstration video, download link and ethical considerations.

3. METHOD

Figure 2 illustrates the process of our pose-estimation-based student behavior analysis method. The CTGAN transfers synthetic images to realistic classroom style. These translated synthetic images are then used to train a pose estimation model to boost performance. Finally, we estimate human poses on real-world classroom images and feed the extracted poses into a graph classification network for student behavior analysis.

Table 1. Overview of the publicly available datasets for 2D multi-person pose estimation

Dataset	Images	Inst. per img	Joints	Modalities
MPII[11]	25k	1.6	16	RGB
COCO[12]	82k	2.6	17	RGB
AIC-HKD[13]	300k	2.3	14	RGB
Crowdpose[14]	20k	4	14	RGB
SynPose	**60k**	**35.2**	**59**	**RGB, stencil, depth**

3.1. Domain adaptation via CTGAN

Domain translation. Given two domains S and R (Synthetic and Real), our purpose is to learn a translation mapping function $G : S \rightarrow R$. As we do not have corresponding pairs between our synthetic dataset and real classroom domains, we adapt the technique of CycleGAN[17] to circumvent this problem. Rather than learning a single mapping, there exists an opposite mapping $F : R \rightarrow S$ which attempts to learn a mapping function from domain R to S. Additionally, two adversarial discriminators D_S and D_R are trained for S and R, respectively. The final objective function of CycleGAN can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{CycleGAN}\left(G, F, D_S, D_R\right) = \mathcal{L}_{GAN}\left(G, D_R, S, R\right) \\ + \mathcal{L}_{GAN}\left(F, D_S, R, S\right) \\ + \alpha\mathcal{L}_{cyc}(G, F)\end{aligned} \tag{1}$$

Where $\mathcal{L}_{GAN}$ represents the standard adversarial loss[18], and $\mathcal{L}_{cyc}$ represents the cycle consistency loss. For more details of those loss functions, please refer to [17].

Semantic shift regularization. In the above formulation, there is no constraint on the colour distribution of the generated image $G(s_i)$ and the origin s_i. In our application, it is possible for the colour of a person's clothes or even skin to shift dramatically under G. To address this problem, we use the inside-domain identity constraint[19] as an aid to image translation, forcing the network to learn the identity mapping when samples from the target domain are provided as input to the generator. The constraint is written as:

$$\begin{aligned}\mathcal{L}_{id}(G, F) = \mathbb{E}_{r_i \sim p_{\text{data}}(r)}\left[\|G(r_i) - r_i\|_1\right] \\ + \mathbb{E}_{s_i \sim p_{\text{data}}(s)}\left[\|F(s_i) - s_i\|_1\right]\end{aligned} \tag{2}$$

Artifact removal. GAN-based image-to-image translation models often introduce artifacts when there is a large scene layout distribution gap between source and target datasets. In our case, back-propagating the gradient from the discriminator to the generator encourages the generator to add texture noise to synthetic images, making them appear more 'realistic'(See Fig 3). To alleviate the artifacts, we design a mask loss written as:

$$\mathcal{L}_{mask}(G) = \mathbb{E}_{s_i \sim p_{\text{data}}(s)}\left[\left\|\left(G(s_i) - s_i\right) \odot \widetilde{M}(s_i)\right\|_2\right] \tag{3}$$

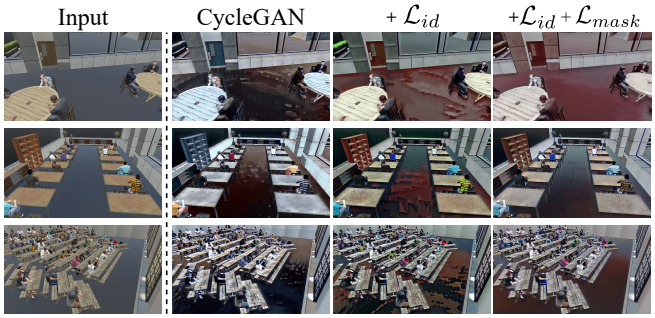

Fig. 3. Comparison of different constraints for translating synthetic images to real-classroom domain.

where $M(s_i)$ represents the corresponding stencil mask of the input image s_i. We only adopt the $\mathcal{L}_{mask}$ on G because the reverse mapping $F : R \rightarrow S$ is not what we focus on.

After employing semantic shift regularization and artifact removal, our full objective loss is:

$$\mathcal{L}_{CTGAN} = \mathcal{L}_{CycleGAN} + \beta\mathcal{L}_{id} + \gamma\mathcal{L}_{mask} \tag{4}$$

where $\beta = 10$ and $\gamma = 1$ in our experiments. Figure 3 illustrates sample results of the domain translation process.

3.2. Student behavior classification

We regard student behavior classification as a graph classification task. Specifically, we construct an undirected spatial temporal graph $G = (V, E)$ where node set $V = \{v_i | i = 1, ...N\}$ includes N joints of the extracted skeleton sequence and edge set $E = \{v_i v_j | (i, j) \in H\}$ featuring intra-body connection. Given a skeleton graph, we adopt GCN[20] as feature extractor, written as:

$$\mathbf{f}_{\text{out}} = \mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{f}_{\text{in}}\,\mathbf{W} \tag{5}$$

where $\Lambda^{ii} = \sum_j \left(A^{ij} + I^{ij}\right)$. $\mathbf{f}$ stands for the feature map and $\mathbf{W}$ means weight matrix. $\mathbf{A}$ is an adjacency matrix representing intra-body connections of joints and $\mathbf{I}$ is the identity matrix representing self-connections. We design the network architecture and training procedure inspired by the spatial graph convolution stream in [21].

4. EXPERIMENTAL EVALUATION

4.1. Datasets and implementation details

Real-classroom dataset. We collected 1000 hours of video from classroom surveillance videos at Shanghai Jiao Tong University and extracted 5000 frames from them to build a real dataset[2]. Bounding-boxes and action labels are annotated for each student.

[2]The acquisition of the videos has been licensed and the students' faces are blurred. Images are only used to visualize experimental results and will not be open-sourced.

Table 2. Comprehensive experiment results of domain adaptation, human pose estimation and student behavior classification. Specially, † indicates datasets generated after the corresponding domain adaptation method.

Training dataset	FID↓	AP (%)↑					Top1 (%)↑
		AP_s	AP_m	AP_l	AP_f	AP_b	
Crowdpose	245	60.48	39.00	12.29	19.12	14.31	55.46
SynPose14	211	54.79	31.38	10.15	15.93	12.47	51.78
SynPose14+CycleGAN†	162	65.70	34.04	19.62	24.41	17.68	62.16
SynPose14+CycleGAN+$\mathcal{L}_{id}$†	151	70.41	46.63	45.40	46.96	34.25	64.53
SynPose14+CycleGAN+$\mathcal{L}_{id}$+$\mathcal{L}_{mask}$(CTGAN)†	**147**	**72.69**	**48.19**	**46.27**	**48.02**	**38.41**	**66.42**
COCO	360	86.67	67.98	23.46	33.92	18.60	66.38
SynPose17	210	68.40	56.43	17.32	26.26	15.49	57.98
SynPose17+CycleGAN†	163	87.92	68.71	35.29	43.42	25.21	67.80
SynPose17+CycleGAN+$\mathcal{L}_{id}$†	150	89.12	69.25	46.73	52.59	37.40	68.23
SynPose17+CycleGAN+$\mathcal{L}_{id}$+$\mathcal{L}_{mask}$(CTGAN)†	**148**	**91.27**	**71.40**	**47.62**	**53.75**	**40.33**	**69.51**

SynPose subsets. We generated SynPose17 and SynPose14, which have the same joint number and training data volume as COCO and CrowdPose respectively to make a fair comparison in the following experiments.

Metric. We borrow the evaluation metric of MSCOCO, using average precision (AP) evaluate the result. We replace the object keypoint similarity (OKS) to the IoU between estimated poses and annotated bounding-boxes.

Implementation. The human pose estimation experiments are based on DEKR[22], a SOTA bottom-up method. We train DEKR on different datasets and test on real-classroom images to evaluate the effectiveness of our SynPose dataset. We use HRNet-W32 as backbone and follow the training settings in[22]. The max epoch number is set to 100.

4.2. Evaluation on real-classroom images

Realism evaluation. Table 2 reports the Fréchet Inception Distance(FID)[23] between different training datasets and real classroom dataset. The SynPose14 and SynPose17 subsets get the lower FID compared with COCO and Crowdpose, which means our proposed dataset has a closer distribution with real classroom images.

Pose estimation evaluation. Columns 3-7 of Table 2 report the pose estimation performance of DEKR tested on real-classroom images but trained with different training datasets. AP_s,AP_m and AP_l are AP of different images captured in small, medium and large classrooms. AP_f and AP_b represent the precision of images captured from front and back viewpoint. The models trained with our SynPose dataset achieved the best results on all measures. Remarkably, we got a huge boost especially on AP_l and AP_b thanks to the realistic scenes and diverse capture viewpoints of our dataset. Figure 4 shows some visualization results of pose estimation.

Student behavior classification. Table 2 shows the top1 accuracy on student behavior classification task. Here we only use poses estimated in small classrooms, where 1.5K images were used for training and 0.5K for testing. Results show that the poses estimated by the models trained on Synpose subsets are accurate enough to classify student behaviors.

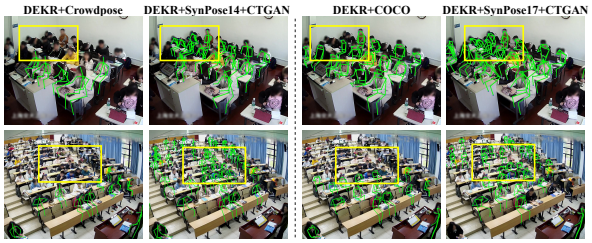

Fig. 4. Qualitative results of pose estimation evaluation. Please check our dataset homepage for more visualizations.

Ablation study. We study the contributions of semantic shift regularization and artifact removal in domain adaptation. Experiment results in Table 2 indicate that by adding $\mathcal{L}_{id}$ and $\mathcal{L}_{mask}$, the generated dataset can effectively reduce FID and improve the performance of pose estimation and behavior classification tasks.

5. CONCLUSION AND FUTURE WORK

In this paper, we attempt to relieve the dilemma of lacking open-source classroom observation dataset. To this end, we present SynPose, a large-scale synthetic dataset specifically designed for crowded human pose estimation in classroom scenarios. Additionally, a novel CTGAN is introduced to bridge the gap between synthetic and realistic images. Quantitative and qualitative experimental results demonstrated that our dataset and method are able to estimate more poses in real classroom scenarios. Results in a student behavior classification task reflect not only the quantity but also the quality of these estimated poses. We believe this paper will pave the road for research in classroom observation and student behavior analysis. In the future, we will continue to enrich the dataset and explore the potential of extending it to other tasks such as student behavior tracking and action localization.

6. REFERENCES

[1] Sidney K D'Mello, Andrew M Olney, Nathan Blanchard, Borhan Samei, Xiaoyi Sun, Brooke Ward, and Sean Kelly, "Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 557–566.

[2] Robin Cosbey, Allison Wusterbarth, and Brian Hutchinson, "Deep learning for classroom activity detection from audio," in *ICASSP*. IEEE, 2019, pp. 3727–3731.

[3] Hang Li, Yu Kang, Wenbiao Ding, Song Yang, Songfan Yang, Gale Yan Huang, and Zitao Liu, "Multimodal learning for classroom activity detection," in *ICASSP*. IEEE, 2020, pp. 9234–9238.

[4] Rui Zheng, Fei Jiang, and Ruimin Shen, "Intelligent student behavior analysis system for real classrooms," in *ICASSP*. IEEE, 2020, pp. 9244–9248.

[5] Wen Li, Fei Jiang, and Ruimin Shen, "Sleep gesture detection in classroom monitor system," in *ICASSP*. IEEE, 2019, pp. 7640–7644.

[6] Jiaojiao Lin, Fei Jiang, and Ruimin Shen, "Hand-raising gesture detection in real classroom," in *ICASSP*. IEEE, 2018, pp. 6453–6457.

[7] Feng-Cheng Lin, Huu-Huy Ngo, Chyi-Ren Dow, Ka-Hou Lam, and Hung Linh Le, "Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection," *Sensors*, vol. 21, no. 16, pp. 5314, 2021.

[8] Yunfang Xie, Su Zhang, and Yingdi Liu, "Abnormal behavior recognition in classroom pose estimation of college students based on spatiotemporal representation learning.," *Traitement du Signal*, vol. 38, no. 1, 2021.

[9] "2018 reform of eu data protection rules.," https://gdpr-info.eu, 2018.

[10] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCV*. Springer, 2016, pp. 17–35.

[11] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014, pp. 3686–3693.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[13] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al., "Ai challenger: A large-scale dataset for going deeper in image understanding," *arXiv preprint arXiv:1711.06475*, 2017.

[14] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," in *CVPR*, 2019, pp. 10863–10872.

[15] Script Hook V, "library for script native functions in custom plugins.," http://www.dev-c.com/gtav/scripthookv/, 2018.

[16] Rockstar Games, "Pc single-player mods.," https://support.rockstargames.com/articles/200153756/Policy-on-posting-copyrighted-Rockstar-Games-material, 2018.

[17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, Oct 2017.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[19] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *CVPR*, 2018, pp. 994–1003.

[20] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[21] 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.

[22] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang, "Bottom-up human pose estimation via disentangled keypoint regression," in *CVPR*, June 2021, pp. 14676–14686.

[23] 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," *Advances in neural information processing systems*, vol. 30, 2017.