Learning Video Representations of Human Motion from Synthetic Data

Xi Guo*
Beihang University
guoxi@buaa.edu.cn

Wei Wu*
SenseTime Research
wuwei@sensetime.com

Dongliang Wang
SenseTime Research
wangdongliang@sensetime.com

Jing Su
SenseTime Research
sujing@sensetime.com

Haisheng Su
SenseTime Research
suhaisheng@sensetime.com

Weihao Gan
SenseTime Research
ganweihao@sensetime.com

Jian Huang
Beihang University
hj@buaa.edu.cn

Qin Yang
Beihang University
yangqin@buaa.edu.cn

Abstract

*In this paper, we take an early step towards video representation learning of human actions with the help of large-scale synthetic videos, particularly for human motion representation enhancement. Specifically, we first introduce an automatic action-related video synthesis pipeline based on a photorealistic video game. A large-scale human action dataset named GATA (**GTA** **A**nimation **T**ransformed **A**ctions) is then built by the proposed pipeline, which includes 8.1 million action clips spanning over 28K action classes. Based on the presented dataset, we design a contrastive learning framework for human motion representation learning, which shows significant performance improvements on several typical video datasets for action recognition, e.g., Charades, HAA 500 and NTU-RGB. Besides, we further explore a domain adaptation method based on cross-domain positive pairs mining to alleviate the domain gap between synthetic and realistic data. Extensive properties analyses of learned representation are conducted to demonstrate the effectiveness of the proposed dataset for enhancing human motion representation learning.*

1. Introduction

Spatiotemporal semantic features are important for video understanding. Early works [41, 44] adopt two-stream networks to extract appearance features and motion information separately. However, the extraction of optical flow is expensive in both time and space. And the flow of objects and background are also retained, which introduce scene bias and thus affect the human motion representation learning. [20] adopts 3D convolutions to capture the spatiotemporal features directly from raw videos. However, stacked 3D convolutions require tremendous parameters and the motion dynamics are capture implicitly. Recently, many researchers [21, 25, 49] attempts to design elaborate architectures to extract motion features explicitly through neighboring feature-level difference, which provide complementary features for action recognition and achieve convincing performance. Therefore, learning a strong motion representation is essential for human action understanding. In contrast to the task-specific architecture designs for motion modeling, we try to solve this problem from the data perspective.

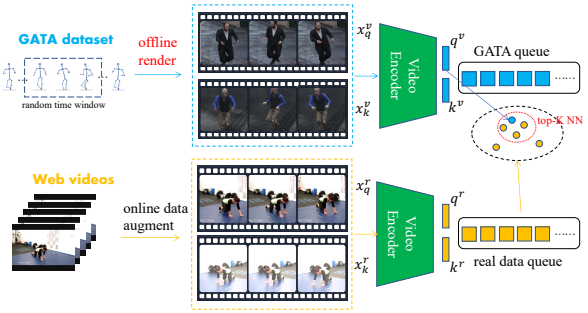

Figure 1. The proposed contrastive learning framework for our GATA dataset. Unlike previous methods, we construct positive pairs *offline*, which are different views rendered using the same semantic 3D skeleton action sequence with diverse backgrounds, human appearances and camera views. Our method samples *real* views rather than simple traditional data augmentation for contrastive unsupervised learning.

To analyze the human motion extraction process, a large-scale motion-oriented human action dataset is essential. However, existing public datasets, e.g., Kinetics [7, 8] and YouTube-8M [3], fail to effectively support the motion rep-

*Equal Contribution.

resentation learning due to the overwhelming bias of scene context [11]. That is said, correct action predictions can be made merely based on scene context instead of human actors. For example, a classroom environment or a whiteboard usually indicates the action of giving a lecture, while the actual activity in the scene is underrepresented. There are also some motion-oriented datasets, e.g., Charades [40] and NTU-RGB [39]. However, the size and diversity of these datasets are limited. Therefore, with the help of a high-performance automatic data collection pipeline based on GTA, we collect a large-scale synthetic video dataset named **GATA**, which contains $\sim$ 8.1M action instances covering $\sim$28k classes. In this dataset, an action class is defined by a specific character animation or a pose sequence. Randomized human characters are controlled to play this action under diverse scene settings with random camera views. In short, scene bias is weakened or even eliminated in the proposed dataset. Figure 2 illustrates some examples of our GATA.

Based on the proposed GATA dataset, a general and robust human motion representations can be learned using a ready-made action recognition model (e.g., SlowFast [13], TSM [26], etc.). And the encoded knowledge can be easily transferred to down-stream action understanding tasks. As shown in Figure 1, we design a contrastive learning framework for GATA, where a skeleton sequence is euqivalent to a sample, and the action instance is equivalent to a view in the tranditional contrastive learning setting. However, this view is not generated by simple online data augmentation but is rendered and stored offline by the CG pipeline. Furthermore, we analyze the learned representations through confusion matrix, Nearest-Neighbor retrieval and Class Activation Maps (CAMs) [50]. And we can find that our model tends to focus on human motion to recognize actions, while the model trained with Kinetics [8] is more inclined to recognize actions through scenes and surrounding objects. Surprisingly, we also find that our synthetic GATA and web-crawled videos are complementary. By jointly training with Kinetics and HAA500, the model can learn a more comprehensive representation for scenes, objects, and human motion. Furthermore, we propose a domain adaptation method based on cross-domain positive pair mining to alleviate the domain gap between synthetic and realistic data.

In summary, our contributions are three folds:

- We introduce an automatic high-performance data collection pipeline and synthesize a large-scale human action dataset. The videos of an action class are transformed from a specific character animation with the help of modern graphics technology, which is essential for human motion modeling.

- We formalize the GATA training process with a contrastive learning framework and design a joint contrastive learning strategy together with realistic videos for a more comprehensive video representation.

- Detailed experiments are conducted to learn and analyze the human motion representation with the help of proposed GATA, which shows considerable performance improvement on downstream tasks and evident enhancement of motion modeling by training solely or jointly with our proposed GATA.

2. Related Works

Action Datasets. Recently, many datasets have been proposed, including UCF101 [42], Kinetics [8], ActivityNet [5], Moments-in-Time [31], and others [6, 17, 18, 23, 29, 37, 45, 47, 48]. However, they suffer from server scene bias. Charades [40] collects videos of daily indoor activities, which has no scene bias but is small. Something-Something [15] and Jester [27] are typically temporally related datasets, but they are not universal enough. Something-Something focuses on the interaction between hand and object, and Jester is a gesture dataset.

Synthetic Datasets. Data synthesis based on computer graphics is an inexpensive way to obtain high-quality data for deep learning. [22, 34, 35] collect synthetic scenes based on GTA-V. Specifically, [35] develops a fast annotation method based on the rendering pipeline. [22] presents a method to analyze the internal engine buffers according to the depth information, which can produce accurate object masks. [34] proposes an approach to extract data without modifying the source code and content from GTA-V, which can provide six types of ground truth. [38] exploits the Unity Engine to construct synthetic street scene data for autonomous driving, which generates pixelwise segmentation labels and depth maps. [36] defines some actions and then render videos through procedural generation. We can obtain some semantic action categories by this way. But it is difficult to define a larger categoriy set.

Action Recognition and Motion Representation Learning. Early action recognition methods have focused on learning spatiotemporal or motion features. [43] propose 3D CNN to learn spatiotemporal features, while [41] employ an independent temporal stream to learn motion features from precomputed optical flows. [43, 46] propose decomposing 3D convolution filters into 2D spatial and 1D temporal filters. [51] propose studying mixed 2D and 3D networks with the frame sampling method of temporal segment networks (TSNs) [44]. [26] propose the temporal shift module (TSM) that simulates 3D convolution using 2D convolution with a part of input feature channels shifted along the temporal axis. [21] propose a module to extract motion features by spatial shift and subtraction operations between appearance features. In contrast, we propose a dataset without scene bias for better motion learning.

Unsupervised Contrastive Learning. Contrastive learning has demonstrated great potential in unlabeled data. Thanks to contrastive learning approaches, the model can be empowered to distinguish samples from separate domains without labels. There are some prior works in this area. [16] propose a momentum dictionary to store and exclude learned features on the fly for input samples so that the number of stored features can be heavily expanded. [9] propose a simplified contrastive learning framework including only major components that benefit the learned representation. However, these methods all rely on spatial or temporal data augmentation [32, 33] to construct separate views of input samples. In this paper, we achieve this by rendering the video background and human body based on computer graphics technology rather than simple data augmentation.

3. The GATA Dataset

In this section, we introduce how to automatically collect large-scale synthetic action videos and the details of the proposed GATA dataset, analyze its features and compare it with the related datasets.

3.1. Data Collection

The center block of our data collection pipeline relies on modern computer graphics technology for efficient realistic video content synthesis. We leverage the video game Grand Theft Auto V (GTA-V) [1] as this block because of 1) its real-time rendering capability for photorealistic video content synthesis; 2) a large virtual world with all kinds of urban scenes, weather, lighting conditions and pedestrian character models with optional clothing/equipment; 3) the feasibility to control elements of the virtual world including scene settings, human character and animation by game Mods; and 4) most importantly, its enormous amount of high-quality human character animations, created by motion capture and refined by artists.

With all the aforementioned advantages, we develop a high-performance automatic data collection pipeline extended from JTA [2, 12] for large-scale human action video dataset synthesis. To collect a dataset for representation learning, we tend to render all the available animations in diverse randomized scene settings. The large number of rendering tasks motivates us to have a pipeline designed with automation and efficiency. The pipeline consists of three main parts handled by different computing nodes: scene setting generation, scene rendering and postprocessing. The scene setting generation module executed by a manager server automatically generates diverse random scene settings. Scene rendering is conducted by multiple worker machines in parallel to render video frames and write them out with annotations according the received scene settings. The postprocessing step handled by a data server gathers all the synthesis data and filters out failed data samples. We detail the pipeline in supplementary.

There are two aspects of factors to guide the generation which is configured in scene settings : environment and human subject. The environment is determined by scene location, weather, time of day and camera view. The human subject variation includes character gender, stature,clothing and so on. Figure 2 shows these factors.

3.2. Database statistics, Comparison and Analysis

As shown in Table 1, we discuss the characteristics of several typical datasets.

Database size. Our dataset provides approximately 8.1M action clips with 27,814 fine-grained labels. An action clip is a tracklet of a subject playing a specific action animation. We place multiple subjects in the same scene for the parallel rendering of the same animation to scale up data samples efficiently. The 27814 fine-grained labels mean different animation instances. These instances are merged and filtered from game animation assets of GTA-V, which has more than 100K animation asset items in total. Despite the semantic labels being defined at animation-instance-level, the unprecedented scale of GATA as a synthetic video action dataset is much larger than those representative real video datasets and synthetic dataset like PHAV [36], which gives the possibility to learning video representation through the synthetic dataset. More dataset-related details can be found in supplementary.

Data source. Our dataset is generated by the Computer Graphic engine, which is easier to generate large-scale data with more accurate, noise-free annotations than web datasets. The previous synthetic datasets are mostly used for image tasks such as object detection, semantic segmentation and depth estimation [34], while GATA focuses on human motion representation, a more difficult task, where large-scale and diverse videos are necessary.

Clues for classification. There are three clues for classifying an action: scene, object and human motion. Many categories can be identified by scene and object in Kinetics. Charades, NTU-RGB and HAA500 decouple scenes and actions to a great extent. But a model classifies an action by understanding the objects and human motion because many labels are combinations of a verb and a noun, which may result in representation bias because it is difficult to enumerate and collect data for all verb-noun combinations [19, 28]. Different from them, human motion is the only discriminative clue in GATA.

Other features of GATA. As shown in Figure 2, thanks to the synthesis engine, our dataset contains many diversity factors, such as scene, view, time of day and weather, which is very beneficial for robust representation learning. In addition, GATA provides informative annotations except for action classes, such as 2D/3D bounding boxes, key points with visibility and camera parameters. (These pieces of in-

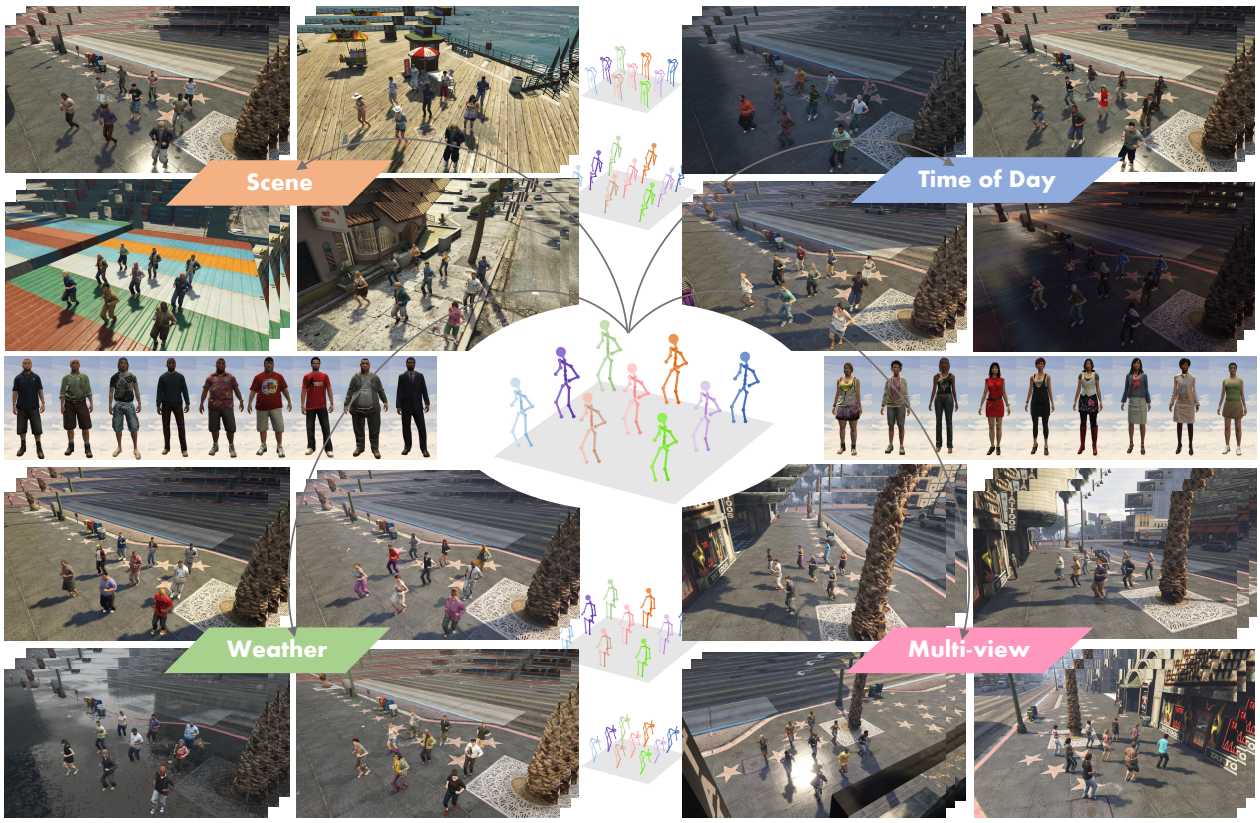

Figure 2. Overview of the proposed GATA dataset (zoom for better view). As the naming of the dataset, based on the rich animations of the video game GTA-V, we generate a large-scale human action dataset with randomized scenes, time of day, weather, camera views and human body configurations. A character animation related to an action clip is defined by a 3D skeleton motion sequence essentially in CG pipeline.

Table 1. Detailed comparisons of our GATA with several existing video datasets.

Dataset	#Clip	#Class	Source	Clues for classification		
				scene	object	motion **only**
Kinetics 400	0.2M	400	Web	✓	✓	×
Kinetics 700	0.5M	700	Web	✓	✓	×
Charades	66k	157	Actors	×	✓	×
NTU RGB	114k	120	Lab	×	✓	×
HAA500	50k	500	Web	×	✓	×
GATA (Ours)	**8.1M**	**27814**	CG	×	×	✓

formation are not displayed in this paper, which will be provided in the public dataset for research of related areas.)

4. Contrastive Learning Framework for GATA

In this section, we first describe a common unsupervised contrastive learning framework, MoCo. Then, we implement the framework on our GATA. Last, we propose a joint contrastive learning framework that incorporates real video data into the training process to obtain a universal representation of human action.

4.1. MoCo Review

Momentum Contrast (MoCo) provides a dictionary lookup for contrastive learning. Given an encoded query q and encoded keys $\{k_0, k_1, k_2, ...\}$ in a queue, the contrastive loss of MoCo can be written as follows:

$$L_q = -log\frac{exp(q \cdot k_+/\tau)}{\sum_{i=0}^{K} exp(q \cdot k_i/\tau)} \quad (1)$$

where τ is a scalar. The sum is over one positive and K negative sample. This loss tends to classify q as k_+ via a softmax classification process. The query q is the representation of an input sample via the encoder network, while the

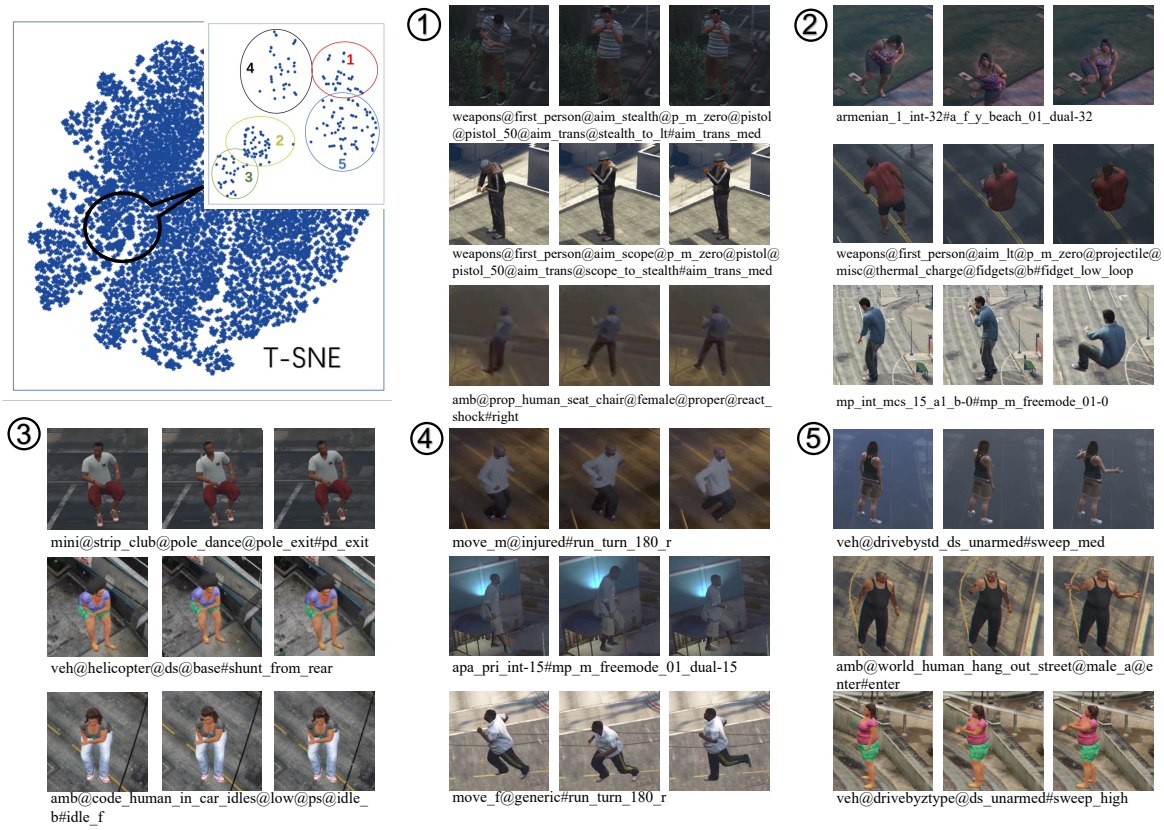

Figure 3. Examples of human action clips. Based on the trained model, we visualize the feature distribution of the classes in GATA. We randomly display 3 classes in 5 clusters and their label text which is name of the game animation assets.

keys k_i are the representations of the other training samples in the queue.

The core of momentum contrast is to dynamically maintain the queue. The samples in the queue are progressively replaced following an FIFO scheme. After computing the contrastive loss in Equation 1, the encoder is updated via gradients, while the momentum encoder is updated as a moving average of the encoder weights.

4.2. Contrastive Learning based on GATA

Our contrastive learning framework is shown in Figure 1. Different from other unsupervised video contrastive learning tasks, e.g., [33], where two clips from the same video are usually regarded as a positive pair, which is clearly not true under our setting. For example, the first half and the second half of a sequence generally represent completely different semantics. Therefore, given a skeleton sequence, we will randomly select a time window and then sample several frames in the same window in two rendered instances to form a positive pair.

We use MoCo V2 [10] as our learning algorithm. The training process is described in Algorithm 1.

Algorithm 1 Contrastive learning algorithm for GATA.

Input: skeleton sequence set X, x_i is the i-th rendered instance of sequence x.
Output: well-trained model.

while $iter < iter_{max}$ **do**
  x = loader.next() # load a minibatch x with N skeleton sequences.
  s, e = random_window(x) # the start and end frame index.
  x_q = sample(x, s, e) # sample a rendered instance.
  x_k = sample(x, s, e) # sample another rendered instance.
  MoCo(x_q, x_k) # core algorithm of MoCo.
end while
return trained model

4.3. Joint Contrastive Learning with Web Videos

Since the synthetic data are different from the real video in appearance, we introduce HAA500 and Kinetics-Tracklet [24] for joint training. HAA500 is a single-person action dataset that is matched with GATA. For the Kinetics-Tracklet dataset, we crop the person tracklets in training,

so it can also be regarded as a single-person action dataset. These three datasets are named the JNT (joint) dataset.

When training the JNT dataset, we control the ratio of synthetic data and real data in each minibatch to $1:1$. For web video, we still use data augmentation methods to obtain views of a sample. In addition, to avoid the model using domain clues to discriminate negatives during training, we adopt two disjoint feature queues for synthetic data and web video data and compute loss independently.

Due to the domain gap between synthetic data and real data, and our goal is to improve on real data, we design an effective method to alleviate the gap. In detail, for a feature of synthetic q, we will find the top-k nearest neighborhoods in the real feature queue and treat them as positive instances. For a query of real data, we still compute loss by Equation 1 while using the Multi Instance InfoNCE loss [30] for a query of synthetic video:

$$L_{q^v} = -log\frac{exp(q^v \cdot k_+^v/\tau) + \sum_{p^r \in P_{q^v}^r} exp(q^v \cdot p^r/\tau)}{\sum_{i=0}^{K} exp(q \cdot k_i/\tau) + \sum_{p^r \in P_{q^v}^r} exp(q^v \cdot p^r/\tau)} \quad (2)$$

where $P_{q^v}^r$ is the positive set from the real feature queue for query q^v, which is defined as:

$$P_{q^v}^r = \{K_i^r | i \in topK(q^v \cdot k), k \in K^r, K_i^r > th\} \quad (3)$$

where th is a similarity threshold for stable training, which is set as 0.7 for all experiments. Intuitively, it is cross-domain positive pairs mining for a synthetic query in the real feature queue while forcing the latent representation of synthetic data to be closer to the real data representation, as shown in Figure 3.

5. Experiments

5.1. Datasets

Charades contains 9848 videos with an average length of 30s. In each video, a person can perform one or more actions. The task is to recognize all the actions in the video without localization. We merge the labels according to the verb to form the Charades-Motion dataset for evaluating the quality of motion modeling.

Kinetics-Tracklet is a subset of Kinetics 700. To provide localized action labels on a wider variety of visual scenes, the researchers provide AVA action labels on videos from Kinetics-700, such that we can crop the region of a person according to the bounding box annotation.

HAA500 is comprised of $10k$ human-centric action videos from 500 fine-grained classes, with a high average of 69.7% detectable joints. The actions are from distinct areas, including sports, instrument performance and daily actions.

NTU-RGB is built by actors performing specified actions in different scenes. The dataset is collected using multicamera and multiview methods. In addition to RGB, there are other modalities, such as depth and skeleton. We only study RGB-based video representation. The dataset provides 3 settings: cross-setup (i.e., cross-scene), cross-subject and cross-view. We adopt cross-setup setting to evaluate the learned representations (i.e., X-setup in Table 2).

5.2. Unsupervised Training Details

Since the scale of the human body in the GATA dataset is almost invariant, which is not conducive to the robustness of the representation, we adopt random scale augmentation to train the models. For an input clip, we randomly sample a spatial scaling factor $\lambda \in [0.5, 1.0]$, resize the clip to $T \times \lambda H \times \lambda W$ and pad the clip to $H \times W$. Here, T, H and W are the input time, height and width, respectively.

We train the GATA dataset using three action recognition methods: TSM, TimeSFormer [4] and SlowOnly [13]. For TSM and TimeSFormer, given a video segment (limited by the time window from the 3rd line in Algorithm 1), we first divide it into T segments of equal duration. Then, we randomly sample one frame from each segment to obtain the input sequence with T frames. In addition to random scale, we perform random cropping flipping as data augmentation during training time. The input size $T \times H \times W$ is set as $8 \times 112 \times 112$. For SlowOnly, we densely sample a clip with T frames from the video segment with a stride of 2 frames. We train the two models with 16 GPUs, and each GPU processes a minibatch of 8 video clips. We start with a learning rate of 0.05 and reduce it to 0.0001 by a cosine schedule. We also use a linear warm-up strategy [14] in the first $8k$ iterations. We use momentum of 0.9 and wight decay of 10^{-4}. A dropout of 0.5 is used before the final FC layer in the cross entropy loss setting, but dropout is closed in the contractive learning setting. We train for 80k iterations on GATA dataset. Unless otherwise specified, we adopt ResNet-50 as the backbone of all models.

In the joint training experiment of GATA and other datasets (we name the joint dataset JNT), we do not perform random scale augmentation on other datasets but randomly crop 112×112 from a clip or its flip version, with a shorter side randomly sampled in $[128, 160]$ pixels. We use two individual heads and queues for the two domains. The final loss is the average of the two losses. We train the JNT dataset for 100k iterations. Other settings are the same as those in the GATA training independently. More training and testing details can be seen in supplementary.

5.3. Main Results and Observations

As shown in Table 2, we use 3 settings to conduct experiments: supervised trained over Kinetics-400, trained over GATA (the model with "CE" indicates training using Cross Entropy loss), trained over JNT (GATA + HAA500 + Kinetics-Tracklet) by contrastive learning.

Effects of pretraining over GATA. Compared with the K400 pretrained model, our model is weak on the Charades

Table 2. Main Results on Downstream tasks.

Model	Pretrain	Charades	Charades-Motion	HAA500	NTU-RGB, X-setup
SlowOnly	K400 sup.	35.5	24.5	23.4	40.5
TimeSFormer	K400 sup.	33.8	21.1	61.3	37.0
TSM	K400 sup.	34.5	24.0	**65.4**	41.7
SlowOnly	GATA	32.0	28.2	21.7	53.4
SlowOnly (CE)	GATA	31.5	27.9	21.4	52.0
TimeSFormer	GATA	30.1	27.2	57.0	51.5
TSM	GATA	31.2	28.0	58.7	51.8
SlowOnly	JNT unsup.	34.8	28.1	24.7	60.7
TimeSFormer	JNT unsup.	32.6	27.8	57.4	55.2
TSM	JNT unsup.	**35.7**	**31.0**	**61.7**	64.6
SlowOnly 101	JNT unsup.	**35.7**	29.9	28.3	**65.3**

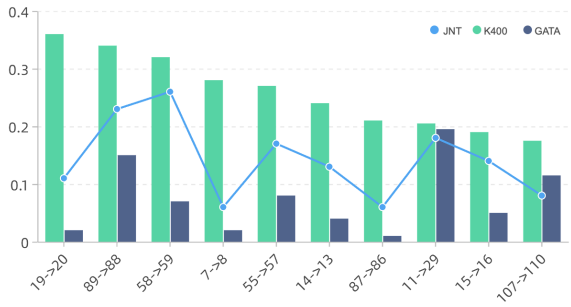

Figure 4. **Confusion Matrix Analysis**. 19→20: put on a hat/cap→take off a hat/cap. 89→88: take object out of bag→put object into bag. 58→59: walking towards→walking apart. 7→8: sit down→stand up. 55→57: giving object→shaking hands. 14→13: take off jacket→put on jacket. 87→86: take off bag→put on bag. 11→29: writing→type on a keyboard. 15→16: put on a shoe→take off a shoe. 107→110: knock over→step on foot.

and HAA500 datasets, which is in line with expectations, because some actions in these two datasets need to be distinguished by objects, such as labels in Charades: *Holding a Box* vs *Holding a Laptop*. However, the Charades-Motion and NTU-RGB datasets pay more attention to specific human motion, and the model pretrained on GATA outperforms the one with K400 under the same training setting. In particular, for the Charades-Motion dataset, the SlowOnly model pretrained over GATA (mAP=28.2) is 3.7 better than the baseline K400 model (mAP=24.5). Therefore, a motion-oriented dataset generated by an automatic graphic engine provides an even better human motion representation pretraining.

Effects of joint pretraining. We can see that joint pretraining can bring further improvement. On the one hand, HAA500 and Kinetics-Tracklet make up for the lack of GATA's ability to model objects and scenes. On the other hand, joint training narrows the domain gap with the real world. In addition, increasing the scale of the model can further improve the performance. SlowOnly 101 can perform better than SlowOnly 50 on all four tasks. Besides, we

Table 3. Ablation study of the domain adaptation operation.

Pretrained Dataset	GATA	Charades	HAA500
GATA	79.8	31.2	58.7
JNT	76.5	35.2	61.3
JNT (w/ DA)	74.2	**35.7**	**61.7**

Table 4. MAP on Charades of models trained over different dataset combinations.

Model	GATA	HAA500	Kinetic-Tracklet	MAP
TSM	✓			31.2
	✓	✓		34.5
	✓	✓	✓	**35.7**

demonstrate that both HAA 500 and Kintetics are beneficial for the joint training in Table 4.

Effects of cross-domain positive mining. Based on TSM-R50, we remove the domain adaptive strategy during joint training. Here, we define the accuracy on the GATA dataset based on the KNN algorithm. For a video, we set 3 seconds as the time window with a stride of 2 and then use the trained model to extract clip-level normalized features. We calculate the dot similarity of two videos, and the two sets of clip-level features are. If 3 or more of the top 5 similar videos are from the same skeleton sequence as it, it is regarded as a correct prediction; otherwise, it is viewed as an error. As shown in Table 3, the accuracy of GATA in the joint training is lower than that of the independent training GATA. After adding the domain adaptation strategy, it drops by 2.3 points. However, the performance on the downstream dataset is better.

5.4. Analysis of the Learned Representations

In this section, we freeze the backbone to fine-tune the downstream task to analyze the learned representation, which is the most direct way to analyze the learned representation. We use NTU-RGB dataset to achieve this. The reason is that NTU-RGB provides a cross-scene setting,

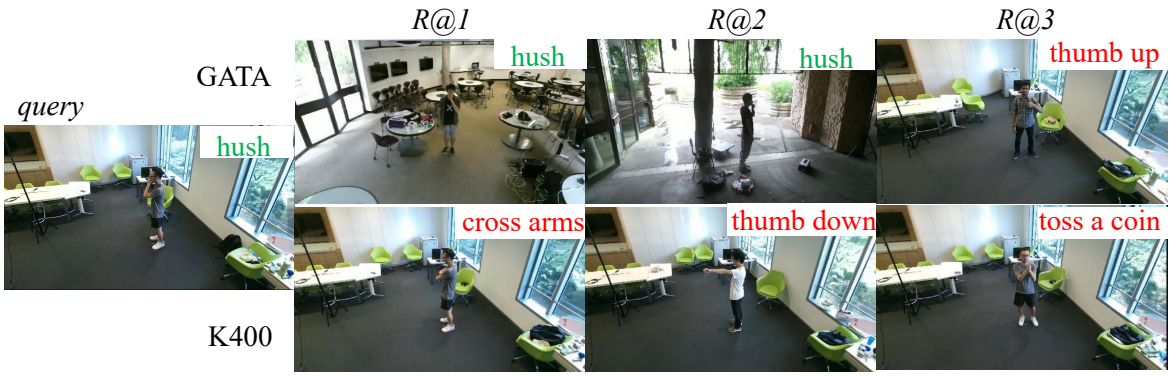

Figure 5. Video retrieval examples. The two rows are the top-3 nearest neighbors of the query video provided by the GATA model and K400 model. Green indicates the same class as the query video, while Red indicates a different class. For brevity, we choose a keyframe to represent a video.

where almost the same actions exist in various scenarios, which is helpful to reflect the scene bias of a model. We use TSM to train the GATA and then freeze the backbone to evaluate the learned representation.

Linear Evaluation Performance on NTU RGB. Table 2 shows the linear evaluation performance on the NTU-RGB dataset, i.e., training a new FC layer to classify with frozen backbone. We can see that the model trained over GATA is clearly better than the K400 model. Furthermore, training with the JNT dataset can further improve the performance, which verifies the complementarity of the two datasets.

Confusion Matrix Analysis. Figure 4 shows the top-10 error pairs of the K400 model on NTU-RGB and the performance of other two models. We can see that the K400 model is particularly prone to make mistakes on opposite verb pairs, such as the top-1 error pair *put on a hat/cap* $\rightarrow$ *take off a hat/cap*. In contrast, the GATA model performs much better on this pair (2% vs 36% error rate).

Video Retrieval. We visualize the nearest neighbor (NN) of the video segments in the feature space in Figure 5. In detail, one video is uniformly sampled from each video, and the spatiotemporal feature is extracted and pooled into a vector. Then, the feature vector is used to compute the L2 distance. Note that the network does not receive any class label during training. It can be seen that the K400 model prefers to encode some scene semantics, as the top-3 nearest neighbors are all from the same scene. In contrast, our model has actually learned the human motion representation. For example, our model can discover videos with the same class *hush* or a similar class *thumb up*. Notably, the scenes of these neighbors are different from the query video, which shows that our model indeed characterizes human motion rather than scenes.

Visualization. To further demonstrate the efficacy of our dataset, we show class activation maps (CAMs) [50] from the two models in Figure 6. We can see that the model trained with K400 usually focuses on the irrelevant region. However, the model trained with GATA truly discriminates an action based on where the person is located.

Figure 6. Class activation maps (CAMs) of models trained using GATA (first row) and K400 (second row) on NTU-RGB. For clarity, we show only the 5-th frame of a clip with 8 frames because the middle moment is usually when the action is most salient. The video representation trained with GATA can be more concentrated in the region of human.

6. Conclusions and Discussions

We propose a new synthetic action database, which is defined only by human motion. This dataset has a large volume of data, and the number of categories is much greater than that of the existing datasets. Pretraining on this dataset can enable a model to have a strong representation of human motion. In addition, we analyze the complementary effect of this dataset and real datasets. Joint training over these datasets achieves a substantial improvement. In addition, we propose a hard positive pairs mining-based domain adaptation method, which further enhances the ability to represent real human motion. GATA now covers single-player action only. Meanwhile, we only use single-player real datasets for joint training. Multiplayer action and human-object interaction can be simulated for general scenarios in the future.

Acknowledgements. This Research was partly supported by China Geological Survey (Project DD20190637) and the Beijing Municipal Science and Technology Project (Project Z201100008120005).

References

[1] https://www.rockstargames.com/V/. 3

[2] https://github.com/fabbrimatteo/JTA-Mods. 3

[3] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 1

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021. 6

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 2

[6] Fabian Caba Heilbron, Joon-Young Lee, Hailin Jin, and Bernard Ghanem. What do i annotate next? an empirical study of active learning for action localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 199–216, 2018. 2

[7] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 1

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 2

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 5

[11] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *Advances in Neural Information Processing Systems*, pages 853–865, 2019. 2

[12] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019. 2, 6

[14] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 6

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, volume 1, page 5, 2017. 2

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3

[17] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 2

[18] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013. 2

[19] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020. 3

[20] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012. 1

[21] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. Stm: Spatiotemporal and motion encoding for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2000–2009, 2019. 1, 2

[22] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*, 2016. 2

[23] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011. 2

[24] Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020. 5

[25] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 909–918, 2020. 1

[26] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7083–7093, 2019. 2

[27] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video

dataset of human gestures. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[28] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2020. 3

[29] Pascal Mettes, Jan C Van Gemert, and Cees GM Snoek. Spot on: Action localization from pointly-supervised proposals. In *European conference on computer vision*, pages 437–453. Springer, 2016. 2

[30] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 6

[31] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. 2

[32] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11205–11214, 2021. 3

[33] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 3, 5

[34] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2213–2222, 2017. 2, 3

[35] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 2

[36] Cesar Roberto de Souza, Adrien Gaidon, Yohann Cabon, and Antonio Manuel Lopez. Procedural generation of videos to train deep action recognition networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4757–4767, 2017. 2, 3

[37] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 2

[38] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 2

[39] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 2

[40] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 2

[41] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 1, 2

[42] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[43] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 2

[44] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 1, 2

[45] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Human action localization with sparse spatial supervision. *arXiv preprint arXiv:1605.05197*, 2016. 2

[46] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018. 2

[47] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126(2-4):375–389, 2018. 2

[48] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8668–8678, 2019. 2

[49] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 1

[50] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2, 8

[51] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video

understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 695–712, 2018. 2