Chaotic World: A Large and Challenging Benchmark for Human Behavior Understanding in Chaotic Events

Kian Eng Ong[1‡] Xun Long Ng[1‡] Yanchao Li[1‡] Wenjie Ai[1‡] Kuangyi Zhao[1]
Si Yong Yeo[1,2] Jun Liu[1*]

[1] Information Systems Technology and Design, Singapore University of Technology and Design, Singapore
[2] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

Abstract

Understanding and analyzing human behaviors (actions and interactions of people), voices, and sounds in chaotic events is crucial in many applications, e.g., crowd management, emergency response services. Different from human behaviors in daily life, human behaviors in chaotic events are generally different in how they behave and influence others, and hence are often much more complex. However, currently there is lack of a large video dataset for analyzing human behaviors in chaotic situations. To this end, we create the first large and challenging multi-modal dataset, Chaotic World, that simultaneously provides different levels of fine-grained and dense spatio-temporal annotations of sounds, individual actions and group interaction graphs, and even text descriptions for each scene in each video, thereby enabling a thorough analysis of complicated behaviors in crowds and chaos. Our dataset consists of a total of 299,923 annotated instances for detecting human behaviors for Spatiotemporal Action Localization in chaotic events, 224,275 instances for identifying interactions between people for Behavior Graph Analysis in chaotic events, 336,390 instances for localizing relevant scenes of interest in long videos for Spatiotemporal Event Grounding, and 378,093 instances for triangulating the source of sound for Event Sound Source Localization. Given the practical complexity and challenges in chaotic events (e.g., large crowds, serious occlusions, complicated interaction patterns), our dataset shall be able to facilitate the community to develop, adapt, and evaluate various types of advanced models for analyzing human behaviors in chaotic events. We also design a simple yet effective IntelliCare model with a Dynamic Knowledge Pathfinder module that intelligently learns from multiple tasks and can analyze various aspects of a chaotic scene in a unified architecture. This method achieves promising results in experiments. Dataset and code can be found at https://github.com/sutdcv/Chaotic-World.

1. Introduction

Chaotic situations can occur in many scenarios, often arising on the onset or aftermath of natural disasters or human-caused incidents such as accidents, crowd crush, crimes, or protests. How people behave, interact, and influence others in chaotic events can be very different from those in peaceful daily life situations (*e.g.*, [33, 54, 20, 11]), and are generally much more complex [43]. Also, the number of people in such scenarios can vary from a few individuals to crowd of thousands, and how they interact with one another is complex, which warrants detailed analysis of the situation before they evolve to become major public safety and security concerns, and subsequently resulting in damages to properties, injuries, or even death [31]. Even a peaceful or joyous event can suddenly turn chaotic [31] such as the recent unfortunate Itaewon crowd crush [56].

Ensuring public safety and order has become challenging, and costs the world at least USD$13.6 trillion each year [15, 16]. The probability of chaotic events occurring is likely to increase with the increased number of natural disasters due to climate change [66, 63], protests (which have tripled in the past 15 years) [42] and violence occurrences [15, 16] etc. Hence, the ability to closely monitor and accurately detect such sudden onset of chaotic situations and potential accidents; rapidly analyze (especially critical during time-sensitive emergencies) humans, their actions and interactions with others; intelligently acquire vital information about the chaotic scene; and then swiftly deploy humanitarian aid [68, 1] has become even more important than ever, and can tremendously minimize casualties and damages [25, 61]. This area of research has also attracted greater attention over the years [68, 50, 1, 70, 73, 8, 74, 31].

Besides being extremely critical for public safety [5], understanding and analyzing human behaviors in chaotic events are also significant in many applications [21, 67, 61, 8] such as the following: Firstly, event organisers, intelligence services, and emergency services can utilize the insights to intelligently deploy smart assistant systems, plan safe evacuation route, manage and control crowds [67, 61, 8]. Secondly, urban space designers can likewise design safer and better spaces [8]. Thirdly, psychologists, criminologists, and social behavioral scientists can learn about the complex human interactions in chaotic situations, so as to develop effective strategies to mitigate or handle chaotic situations or post-chaos trauma [5]. Lastly, content

[‡] Equal contributions. * Corresponding author.

providers or film regulators can automatically identify and tag scenes that may be uncomfortable to viewers (*e.g.*, violence) and provide appropriate content rating guidance [6].

Considering that chaotic scenes are often dynamic (and often with serious occlusions) and complex (including actions of individuals and interactions between individuals), analyzing human behaviors and interactions in such scenes while taking in account the different data modalities is currently challenging, and especially so for humans to screen through large volume of social media or surveillance footages with multiple views of the scene [1, 68], and often a large part of these footages may not contain the chaotic scene of interest [31]. Furthermore, given that many people can appear in the same frame in the chaotic scene, with each person performing one or more actions [10], and further compounded by the fact that actions can occur and change within split seconds, localizing relevant temporal segments and human behaviors becomes even more challenging. Thus, there is a demand to develop intelligent automated systems to analyze human behaviors and interactions in chaotic scenes, and pre-empt potential escalation in crowd tension or threats to public safety [5, 39, 31], thereby minimizing possible damage or casualty through alerts or mitigatory measures.

Before analyzing a chaotic situation, one would first be concerned about identifying the relevant chaotic scene or person of interest in long video streams, and this can be expedited by Spatiotemporal Event Grounding (Fig. 5). Next, identifying each individual's action is important (*e.g.*, uncovering intentions), and therefore there is a need to localize them through Spatiotemporal Action Localization (Fig. 3). Besides actions, analyzing sound can also be important as a sound (*e.g.*, explosion going off) or voice can affect the actions of one or many (*e.g.*, causing the crowd to scamper and leading to even more chaos). Thus, it is crucial to locate the source of sound in such a chaotic scene using Event Sound Source Localization (Fig. 6). On top of these, the action of a person can also affect the trajectories and actions of many people (*e.g.*, person shouting "Bomb" can cause the crowd to scamper or duck to the ground). Hence, analyzing the complex graph of one's interactions with others through Behavior Graph Analysis (Fig. 4) is crucial for one to understand relationships (*e.g.*, accomplices), sense crowd dynamics, detect changes in human behaviors, and analyze trends and development of the fluid chaotic situation. All of these tasks are crucial for analyzing human behaviors in chaotic situations, and it is also important to have a large and comprehensive dataset to develop, adapt, and evaluate deep learning models [24, 33].

Unfortunately, there is lack of a large video dataset that comprehensively analyzes human behaviors in chaotic events [31], thereby stifling deeper research and prospect of bringing much needed benefits to this underserved domain.

Motivated by this, we create Chaotic World, the *first* large and challenging multi-modal video dataset for analyzing different dimensions of a scene for holistic and detailed analysis of human behaviors in chaotic situations — from high-level information on the **(1)** *type of chaotic situations*, ranging from natural disasters (*e.g.*, earthquakes) to human-caused incidents (*e.g.*, accidents, crimes, protests) in different parts of the world, to mid-level details of the **(2)** *complex interaction graph* among individuals for Behavior Graph Analysis, and down to low-level details of **(3)** *actions of individuals* for Spatiotemporal Action Localization, and even **(4)** *sound and voice* for Event Sound Source Localization. Our dataset also contains **(5)** *descriptions of people in scenes-of-interest* that can be used to identify the relevant segment in long video streams through Spatiotemporal Event Grounding (Fig. 5). Our Chaotic World dataset possesses challenges with complex scenes of chaotic scenarios with crowds (up to 50 people in one frame) and serious occlusions, dynamic background (*e.g.*, moving objects in a flood) and motion blur (*e.g.*, camera shake during earthquake). There are also new, unique, and complex action and interaction patterns that are unique in chaotic scenarios (*e.g.*, postures and trajectories of someone walking in an earthquake that are not seen in normal situations). Also, the interactions among many people in chaotic scenarios form distinct and very complex chaotic patterns and large interaction graphs that differ significantly from everyday activities, because one person can affect many others, many may also affect one, and many can affect many in such scenarios, thus forming complicated graphs of actions, interactions, and sounds (Fig. 1) that are seldom handled in previous datasets.

With the rich annotations, our challenging dataset shall be able to attract and facilitate the community to develop, train, and evaluate various types of advanced methods for analyzing human behaviors in chaotic situations, and take advantage of highly correlated multi-modal information to assist the model in learning complex patterns associated with chaotic events so as to obtain detailed holistic analysis of human behaviors in chaotic events, thereby expediting the rate in which such research is conducted to benefit a wide range of future applications. We also present an effective IntelliCare model with a Dynamic Knowledge Pathfinder module that intelligently learns from multiple tasks and can analyze various aspects of a chaotic scene in a unified architecture to facilitate a more thorough understanding and analysis of human behaviors in chaotic events.

2. Related Works

Analyzing human behaviors is an important task. Below we review some of the notable human behavioral video datasets and methods. Readers are suggested to refer to the survey papers [34, 52, 60, 31] for a comprehensive list of human behavior datasets and methods.

Human Behavioral Analysis Video Datasets. Most of

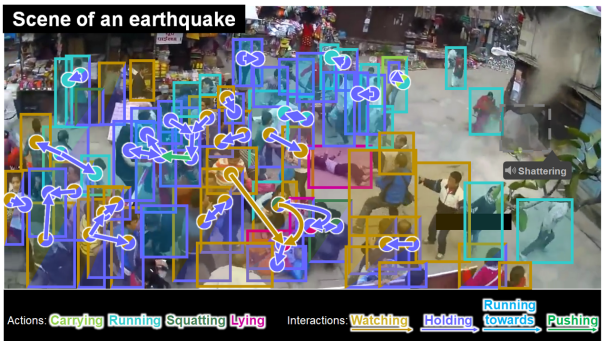

Figure 1. Our Chaotic World dataset contains complicated graphs of actions, interactions, and sound in complex scenes of chaotic scenarios with crowds, serious occlusions, complex action and interaction patterns *etc*.

the existing datasets (*e.g.*, [20, 35, 33, 55, 54, 29, 27, 58, 24, 19, 41, 40, 51, 38, 28, 64, 36, 37]) focus on common day-to-day activities, while some specially focus on group activities (*e.g.*, [69, 26]) or specific domains such as sports (*e.g.*, [26, 32]), campus activities (*e.g.*, [11]), or first-person egocentric view of human actions (*e.g.*, [19, 12, 53]).

Most of the human behavioral localization datasets (*e.g.*, [27]) provide only temporal, but not spatial, localization for each action. Unlike the rest, Atomic Visual Actions (AVA) [20] is a multi-label Spatiotemporal Action Localization dataset with 430 video clips and 80 types of fine-grained visual actions localized in both time and space.

Different from human behaviors in these peaceful and daily life situations, human behaviors in chaotic situations can be more different and complex (*e.g.*, trajectory and manner in which people run in normal situation VS after a shooting incident), and there can be significantly larger number of people in the same frame. There are also unique actions specific to chaotic events (*e.g.*, carrying casualty). Besides analyzing the actions of an individual, analyzing how the individual interacts with others (which results in the formation of extensive, intricate, and complex multi-person spatiotemporal interaction graphs) can also be important in understanding human behaviors. Furthermore, analyzing human behaviors and the interaction graphs in chaotic scenes can be much more complex and challenging with serious occlusions, motion blur (*e.g.*, camera shake during earthquake), dynamic background (*e.g.*, moving objects in a flood), and even trajectories or postures that are not seen in normal situations (*e.g.*, Fig. 2) [31]. Though there are some datasets on crowd surveillance and events analysis [3, 36, 37], incidents [68, 50], or violence [70, 23, 46], most of these datasets are small, typically focus on peaceful events or a specific type of scene, and merely provide simple and often single label for action classification of the entire scene. Thus, there is lack of a large and comprehensive human behavioral video dataset with detailed graph and different levels of details (*e.g.*, actions, interactions, sound, voice, and textual descriptions) for analyzing human behaviors in various types of chaotic situations.

Unlike existing datasets, our Chaotic World is the first large and comprehensive multi-modal dataset for analyzing human behaviors, human interaction graphs, and sound/voice in various chaotic situations, in a holistic manner with different levels of annotated details (*i.e.*, event category, action, interaction, and audio labels). With rich annotations, our dataset shall facilitate the community to develop and test various types of models, with a challenging suite of benchmarks, for holistic human behavior analysis of chaotic scenes.

Human Behavioral Analysis Methods. There are various types of existing human behavior analysis methods, such as C3D [65], I3D [7], and SlowFast [14]. They use Convolutional Neural Networks [65, 7, 13, 14, 11] and sometimes with Long-Short Term Memory [33, 22]. Some other methods use Transformer Networks [4, 35]. To spatially locate humans before identifying their actions, current Spatiotemporal Action Localization methods [20, 30, 59, 44] often use a two-stage process — first detecting humans using Fast-RCNN [18] before performing action recognition using methods such as those described earlier [7, 65, 13, 14]. To analyze interactions between human and objects (or at times between humans), existing Scene Graph Generation methods use scene graph feature banks [29], hierarchical relation tree [62], or multiple linear transformations for relationship prediction [9]. In our case for Spatiotemporal Action Localization and Behavioral Graph Analysis, we present a simple yet effective unified multi-task IntelliCare model with a Dynamic Knowledge Pathfinder module to intelligently learn from multiple tasks and select, for respective task, relevant blocks of learned task-shared and task-specific features.

3. Proposed Chaotic World Dataset

In this section, we introduce the significant properties of our dataset and annotated tasks that are helpful for analyzing chaotic scenes.

3.1. Description of Dataset

Our dataset contains 299,923 annotated instances for Spatiotemporal Action Localization, 224,275 instances for Behavior Graph Analysis, 336,390 instances for Spatiotemporal Event Grounding, and 378,093 instances for Event Sound Source Localization in chaotic scenes.

Multiple levels of details. Our dataset provides a holistic view of the chaotic event, with different levels of details annotated for various tasks — from the high-level details of the category of chaotic events, to the mid-level details of complex interactions between individuals in Behavior Graph Analysis, down to the low-level details of fine-grained actions of an individual in Spatiotemporal Action Localization, and even sound and voice for Event Sound Source Localization. These tasks facilitate a thorough un-

Figure 2. Examples of chaotic events in our dataset. Different from existing datasets, human behaviors in chaotic scenes that arise from the onset of natural disasters, accidents, protests *etc*. can be much more different and complex, in terms of how they move, behave and interact with others.

derstanding and analysis of human behaviors in chaotic events from various aspects.

Diverse range of events and behaviors. Our dataset encapsulates different types of chaotic events ranging from natural disasters (*e.g*., hailstorm) to accidents to violence and crimes (*e.g*., looting) to protests in different parts of the world. Human actions in chaotic events can be much more different and complex (*e.g*., running that can come in many forms), hence our dataset includes many types of actions (*e.g*., performing resuscitation, burning/setting fire) that are not found in most datasets. In total, our action vocabulary contains fine-grained behaviors, ranging from disruptive behaviors (*e.g*., waving flarestick, spraying aerosol) to aggressive behaviors (*e.g*., shooting/firing, punching) to criminal behaviors (*e.g*., stealing/looting) to life-saving behaviors (*e.g*., carrying casualty, extinguishing fire). Analyzing these behaviors will be useful to different groups of professionals (*e.g*., emergency response team and social behavioral scientists) in assessing and analyzing chaotic scenes. Given that human actions are highly varied, having a large and diverse dataset is important in understanding and analyzing human behaviors in dynamic real-world chaotic scenes, and essential for training various types of robust models.

Dynamic scene compositions with fine-grained multi-label actions and complex interactions. In contrast to most other human behavioral datasets [20, 33, 26, 69], our dataset contains much more complex scenes of massive and dynamic chaotic scenarios with crowds (up to 50 people in one frame), with different people performing different actions. Additionally, a person can perform more than one action at the same time. The same type of action can look differently when performed by the same person before, during, and after the chaotic situation (*e.g*., runners running before and after an explosion (Fig. 2)). Furthermore, the interactions between many individuals can be dynamic and form a complex graph. All of these inevitably result in multi-label actions and unequal long-tailed distribution of actions, which pose significant and practical challenges. In addition, our dataset possesses challenges such as serious occlusions, dynamic background (*e.g*., moving objects in a flood) and motion blur (*e.g*., camera shake during earthquake). For real-world human behavior analysis in chaotic events, having such a challenging dataset like ours containing scenes of both crowd and chaos, with multiple individuals sometimes performing subtly different actions, is necessary to develop and evaluate robust and accurate models to tackle the complex real-world scenarios, and hence shall encourage the community to develop advanced strategies to overcome these practical challenges.

Multiple data modalities. In addition to RGB video frames, our dataset contains audio information that can provide complementary information for behavior analysis in chaotic scenes. The diverse nature of our dataset provides variances in natural conversational features (*e.g*., variation of noisy backgrounds), thus presenting practical challenges in localizing the source of sound (voice), thereby facilitating future research into these areas.

Besides audio information, the textual description of the scene, person, behaviors, and interactions in the chaotic scene also provides different levels of details of the scene, which will be helpful in video and information retrieval. Utilizing all modalities (visual, audio, and text information about the chaotic scene) can also enrich the feature space, which enables multi-modality model training [2, 48, 72].

Diverse types of footages and viewpoints. The footages in our dataset are obtained from YouTube and are captured by the public, news agencies, and security agencies who use different equipment (*e.g*., phone camera, surveillance camera) and are captured at different angles and styles (*e.g*., live recording captured by the public, and news with the news anchor reporting live at the scene). Our dataset encapsulates a wide range of unscripted real-world chaotic scenes from all over the world with carefully annotated untrimmed long videos. This also means that our dataset contains a rich diversity of human subjects across ages, genders, languages, clothes and accessories, and locations (*e.g*., streets, parks, indoor). Therefore, having diverse representation of humans in a dataset to reflect the diversity in the real world would facilitate the development of robust and generalizable models to handle the challenging task of analyzing human behaviors in chaotic events. While all these inject realism and reflect the natural interactions between people in chaotic situations, the complex compositions and the spatiotemporal dynamism of the scene also present practical challenges for human behavior analysis.

Dynamic illumination and environmental conditions. The footages in our dataset contain different illumination conditions (*e.g*., night time or even smoky due to the fire started by protesters (Fig. 2) which leads to occlusion). All of these are practical challenges in chaotic events and can

affect the appearances of humans and present significant visual challenges (*e.g.*, low foreground-background contrast) to the level of details and actions that can be recognized. Having such diversity in our dataset is the first step towards building a robust model that can handle varied scenarios.

3.2. Dataset Tasks and Annotations

In this section, we introduce salient details of the key tasks (*i.e.*, Spatiotemporal Action Localization, Behavior Graph Analysis, Spatiotemporal Event Grounding, and Event Sound Source Localization) that can be crucial in analyzing human behaviors in chaotic scenarios. Carefully annotated by 15 individuals over a year, we manually identified and provided annotations of chaotic events, sounds, human actions, and interactions in the videos, and conducted 2 rounds of quality checks for the annotations. The video clips range from 10 seconds to 1 hour.

Task 1: Spatiotemporal Action Localization identifying individuals and their action(s) in chaotic scenes. In Spatiotemporal Action Localization (Fig. 3), a video clip is provided to the model as the input. The model first detects where (spatial) each person is, and then identifies the action when (temporal) it occurs in the chaotic scene, thereby localizing the person of interest and outputting his/her action(s). Our dataset contains 299,923 instances, with each action ranging from 5 to 200 seconds (average of 10 seconds). Our action vocabulary consists of unique fine-grained actions, ranging from disruptive behaviors (*e.g.*, waving flarestick) to aggressive behaviors (*e.g.*, shooting/firing) to criminal behaviors (*e.g.*, looting) to life-saving behaviors (*e.g.*, carrying casualty).

Figure 3. In Spatiotemporal Action Localization, the positions of the individuals are located and the actions of individuals are identified. Note that for clarity of presentation, we do not show all annotations of the whole scene in this figure.

Task 2: Behavior Graph Analysis describing the interactions between individuals in chaotic scenes. Similar to a graph with nodes connected by edges, a Behavior Graph is a graph that describes the interacting behavior (edge) between individuals (nodes). The interacting behavior (or termed as interaction in short) between individuals (Fig. 4) can provide much more contextual clues about the chaotic scene. Such interactions are predicted by Behavior Graph Analysis models (using Scene Graph Generation models [9, 62]). To enable such analysis of interactions, our dataset contains 224,275 instances of individuals interacting with others (*e.g.*, throwing object at someone, shouting at, arresting). In Behavior Graph Analysis, when given the video frame, the model predicts the interactions between individuals. Thus, all the interactions in the scene form a graph, with the interactions being the directed edges (see arrows in Fig. 4). This can be quite useful for comprehensively understanding the behavior relations and trend of the chaotic scenario. In the simplest Predicate Classification (PredCls) setting, the additional information of the bounding boxes of individuals are also provided to the model, so that the model only predicts the interactions between individuals. In contrast, in the more challenging Graph Detection (GDet) setting, besides predicting the interactions between individuals, the model also needs to predict the spatial locations (bounding boxes) of the individuals.

Figure 4. In Behavior Graph Analysis, when given the video frames, the model in the GDet setting locates the spatial positions of the individuals (shown in bounding boxes) and identifies the interactions between individuals (indicated by arrows between bounding boxes). In the PredCls setting, only the interactions between individuals are identified. In Behavior Graph Analysis task, we label the interactions. Note that we have also labelled the actions of each individual which is part of the Spatiotemporal Action Localization task (see Fig. 3).

Task 3: Spatiotemporal Event Grounding retrieving scenes of interest based on textual description. In Spatiotemporal Event Grounding (Fig. 5), very much like a video clip search engine using text data, the user types an input query sentence that describes the (chaotic) scene, description of persons, behaviors, and interactions of interest (*e.g.*, Fig. 5). The model will then output both the relevant temporal segment with the start and end timing, as well as the spatial position of the relevant person in the video frame. Spatiotemporal Event Grounding plays a critical role in expediting the video and information retrieval from long videos, because a large part of the long videos may not contain the footage of the individuals and behaviors of interest [31]. Hence, Spatiotemporal Event Grounding will be an extremely useful and convenient way for users to search for the relevant time segment in long videos by describing the scene, person, behavior, and interaction of interest. This can be of significant interest during times of emergencies, whereby rapid video and information retrieval is mission critical. To facilitate this, our dataset contains 336,390 instances for Spatiotemporal Event Grounding.

Task 4: Event Sound Source Localization outputting location of sound. In Event Sound Source Localization (Fig. 6), when given the video frames with the corresponding audio clip, the model locates the spatial position

Figure 5. Example of Spatiotemporal Event Grounding. After entering a textual search query, the model identifies the relevant segment with the start and end timing, as well as the location (shown in bounding box) of the person of interest.

(source) of the sound of interest (*e.g.*, shouting for help) in the video frames of a chaotic scene. During training via the unsupervised setting, the audio clip and the corresponding video frames are provided to train the model. In contrast, during training via the supervised setting, the information of the bounding box (*i.e.*, source of the sound of interest) is also provided when training the model. During inference, the audio clip and video frames are provided to the model for it to find the spatial location of the sound of interest in the video frames of a chaotic scene. This is especially useful in instances whereby audio data in the video provides additional informative clues to visual cues, that are helpful in identifying the scene, person, and behavior of interest (*e.g.*, finding a person shouting for help). To facilitate this, our dataset contains 378,093 instances with different types of sounds (*e.g.*, explosion, shouting, smashing).

Figure 6. Example of Event Sound Source Localization. When given video frames with the corresponding audio clip, the model locates the spatial position (visualized in the form of a heatmap) of the sound of interest in the video frames of the chaotic scene.

In a nutshell, our Chaotic World dataset is the first large, comprehensive, and challenging dataset for multi-modal multi-task learning with multiple levels of details about chaotic scenes. The four tasks complete a detailed graph of a chaotic scene, distilling the complex interactions between sound, voice, text, actions and interactions of people. Thus, our dataset shall provide a useful benchmark with practical challenges to facilitate further research into various types of advanced strategies for human behavioral analysis in chaotic events.

4. IntelliCare Model for Multi-task Learning

An in-depth understanding and analysis of human behaviors in chaotic events may require thorough analysis of various aspects – from the sound or voice (Event Sound Source Localization), to the spatiotemporal information of the scene or behaviors (Spatiotemporal Event Grounding), to behaviors (Spatiotemporal Action Localization) of individuals, and their interactions with others (Behavior Graph Analysis). While the inputs of these four tasks may not be identical (*e.g.*, Spatiotemporal Event Grounding uses text and visual inputs, while Event Sound Source Localization uses audio and visual inputs), the visual input is required and common across all the four tasks. Hence, by using a shared visual backbone, visual features containing task-shared knowledge can be learned and used in downstream tasks. As such, we design a unified architecture, which we term as IntelliCare model (Fig. 7), that uses a shared visual backbone and respective task heads to train the four tasks together. Such a unified model with multi-task learning can benefit from the visual features learned with the task-shared knowledge across all four tasks, thus enhancing model's generalization ability.

In addition to learning task-shared knowledge, learning task-specific knowledge is also important for the respective task's performance, since not all visual features are equally helpful and can be shared by all four tasks. Thus, we need to effectively and dynamically learn the relevant task-shared and task-specific knowledge for each task. As such, we design a simple yet effective Dynamic Knowledge Pathfinder module that automatically and intelligently chooses ('cares') the learning path in the network, with required task-shared and task-specific knowledge for each task. This module contains multiple dynamic layers, with each layer consisting of multiple knowledge blocks. For each task, the Pathfinder dynamically finds the best path that is made up of, over layers, the best knowledge blocks for the task. Specifically, we use I3D [7] as the shared visual backbone and build the Dynamic Knowledge Pathfinder with it. To construct the Dynamic Knowledge Pathfinder, we replace the last L layers of I3D with the dynamic layers. For each dynamic layer, we duplicate B copies of the block at the corresponding original I3D layer, and thus achieves B knowledge blocks. We also add an extra score module consisting of a linear layer and a Gumbel Softmax layer at each dynamic layer. This score module is used to produce the decision about which block at the current dynamic layer to be used for each task.

During training, we use four embeddings (*i.e.*, [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], and [0, 0, 0, 1]) as the task indicators of the four tasks. For each task, the score module at each layer takes in the corresponding task indicator, and outputs a one-hot vector (with length $1 \times B$) to represent the score (0 or 1) of each knowledge block at this layer. Thus, at each layer, the knowledge block with the score of 1 is selected to be the best knowledge block for the current task. This operation is done for each dynamic layer, as the Pathfinder needs to choose the best knowledge block for the task at each layer. During training, the four tasks are trained together, where the IntelliCare model with Dynamic Knowledge Pathfinder is trained in an end-to-end manner. Thus, after end-to-end training, the Dynamic

Knowledge Pathfinder is trained to find the optimal block at each layer for each task, so as to optimize the performance of this task. This will cause some blocks to be shared among some tasks (*i.e.*, reused by multiple tasks), while some may only be used by one task. As such, each task will have a unique sequence of knowledge blocks to construct a unique path, which contains task-shared and task-specific knowledge. During testing, the parameters of the score module are fixed, and thus each task takes a fixed optimal path for inference.

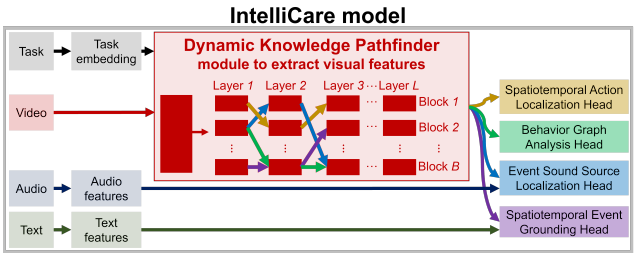

Figure 7. Our multi-task IntelliCare model with Dynamic Knowledge Pathfinder that intelligently selects relevant blocks of knowledge learned from different tasks. Each task will take in respective inputs and have a unique path (*i.e.*, arrows of different colors for different tasks) — a unique sequence of best knowledge blocks (*i.e.*, red boxes) for its task head.

Besides the visual input with the task embedding that is used as input of the shared visual feature encoder (*i.e.*, I3D with Dynamic Knowledge Pathfinder), other modality inputs (*i.e.*, audio and text), depending on the input requirement of each task, are sent to their respective modality encoders (*i.e.*, using the same audio encoder that is used in [57] for audio, and text encoder that is used in [71] for text). For the visual features of each task, we use our shared visual encoder (*i.e.*, I3D with Dynamic Knowledge Pathfinder). The visual features are then combined with other corresponding modality features, and are sent to the respective task heads. Here, the task heads are the same that are used as Actor-Centric Relation Network (ACRN) [59] for Spatiotemporal Action Localization, Target Adaptive Context Aggregation Network (TRACE) [62] for Behavior Graph Analysis, Self-Supervised Predictive Learning (SSPL) [57] for Event Sound Source Localization, and TubeDETR [71] for Spatiotemporal Event Grounding. More details of the network design can be found in Supplementary.

In summary, the IntelliCare model is simple yet effective, and can dynamically choose relevant knowledge blocks for shared or unshared feature learning for each task (instead of sharing all features and blocks over all tasks), hence achieving good generalization ability and performing well on all tasks. Also, it is convenient and easier to use a unified model to analyze different tasks for obtaining an in-depth analysis of the chaotic scene, as opposed to having to separately run different disparate models.

5. Experiments

Based on our dataset, we evaluate various methods such as ACRN [59], I3D [7], SlowFast [14], and Actor-Context-Actor Relation Network (ACAR) [45] for Spatiotemporal Action Localization; Spatial-Temporal Transformer (STTran) [9] and TRACE [62] for Behavior Graph Analysis; TubeDETR [71], Grounding of Textual Phrases in Images by Reconstruction (GroundeR) [47] with Temporal Activity Localization via Language Query (TALL) [17] for Spatiotemporal Event Grounding; Learning Sound Source in Visual Scenes (LSSVS) [49] and Self-Supervised Predictive Learning (SSPL) [57] for Event Sound Source Localization. We use their original codes and adapt them to our Chaotic World dataset.

In addition, we also evaluate our unified multi-task IntelliCare model with Dynamic Knowledge Pathfinder module, which can intelligently learn for multiple tasks. To evaluate the benefits of multi-task learning, we use two settings – IntelliCare without multi-task learning (*i.e.*, all the tasks are trained totally separately with the I3D visual backbone and their respective task heads), and IntelliCare without Dynamic Knowledge Pathfinder (*i.e.*, all tasks use the same I3D visual backbone with the dynamic layer consisting of only one block, *i.e.*, all blocks are shared by all tasks). To evaluate the benefits of using dynamic network in multi-task learning, we compare IntelliCare with and without Dynamic Knowledge Pathfinder.

For the experiments, we adopt the leave-k-out setting, whereby each video clip is assigned to either the training set (80% of samples), or test set (20% of samples). The details of the calculation of the evaluation metrics can be found in the Supplementary.

5.1. Spatiotemporal Action Localization Results

We follow [20] and use the commonly used $frame\text{-}mAP@spatial\text{-}IoU$=0.5 for Spatiotemporal Action Localization (STAL). mAP refers to the mean Average Precision. $spatial\text{-}IoU$ refers to spatial Intersection-over-Union (IoU) of the predicted and groundtruth bounding boxes of individuals. Our results are shown in Table 1.

Table 1. Results of Spatiotemporal Action Localization

Methods	frame-mAP
ACRN [59] (I3D backbone [7])	11.91
ACRN [59] (SlowFast backbone [14])	12.90
SlowFast [14]	13.24
ACAR [45]	13.99
IntelliCare (w/o Multi-task Learning)	12.06
IntelliCare (w/o Dynamic Knowledge Pathfinder)	14.10
IntelliCare (w/ Dynamic Knowledge Pathfinder)(Full model)	**16.25**

5.2. Behavior Graph Analysis Results

In line with [62, 9], we adopt $Recall@k$ ($R@k$) (k refers to top-k returned results) as our evaluation metrics for Behavior Graph Analysis (BGA), and use two evaluation settings: (1) Predicate Classification (PredCls) which predicts the interactions (*i.e.*, predicate in linguistic terms) between individuals, using the input of the bounding boxes of individuals; and (2) Graph Detection (GDet) which, on top

Table 2. Results of Behavior Graph Analysis

Methods	PredCls $R@10$	GDet $R@10$
STTran [9]	12.91	7.12
TRACE [62]	13.58	7.37
IntelliCare (w/o Multi-task Learning)	13.58	7.37
IntelliCare (w/o Dynamic Knowledge Pathfinder)	15.71	8.75
IntelliCare (w/ Dynamic Knowledge Pathfinder)(Full model)	**16.62**	**9.51**

of predicting the interactions between individuals, also predicts the locations of individuals, where the prediction of interaction is considered to be correct only if the predicted tubelets (*i.e.*, bounding box of each individual tracked over time) of both individuals have an IoU of at least 0.5 with the ground-truth tubelets. We present the results in Table 2.

5.3. Spatiotemporal Event Grounding Results

Following [71], we use different evaluation metrics to evaluate different aspects of Spatiotemporal Event Grounding (STEG). To evaluate the temporal grounding only, we use mean $temporal\text{-}IoU$ (m-$tIoU$), which is computed by taking the average of $tIoU$ of all videos. $tIoU$ is defined as Intersection-over-Union (IoU) of the predicted and groundtruth timestamps.

To evaluate both spatial and temporal video grounding (*i.e.*, Spatiotemporal Event Grounding), we use $spatiotemporal\text{-}IoU$ ($vIoU$), which is defined as $vIoU = \frac{1}{F_U}\sum_{t\in F_I} IoU(\hat{b}_t, b_t)$, where F_I (or F_U) comprises the set of frames in the intersection (or union) of the predicted and groundtruth timestamps, with $\hat{b}_t$ and b_t representing the predicted and groundtruth bounding boxes at time t respectively. $vIoU@\tau$ represents the percentage of samples whereby $vIoU \geq IoU$ threshold τ. The m-$vIoU$ (*i.e.*, mean $vIoU$) indicates the average of $vIoU$ of all videos. The results are presented in Table 3.

Table 3. Results of Spatiotemporal Event Grounding.

Methods	m-$tIoU$	$vIoU@0.3$	$vIoU@0.5$	m-$vIoU$
GrounderR [47] + TALL [17]	19.53	8.36	4.21	6.84
TubeDETR [71]	44.92	47.83	10.87	28.59
IntelliCare (w/o Multi-task Learning)	45.12	48.06	10.90	28.70
IntelliCare (w/o Dynamic Knowledge Pathfinder)	46.36	50.13	12.20	31.40
IntelliCare (w/ Dynamic Knowledge Pathfinder)(Full model)	**47.38**	**51.74**	**13.60**	**32.60**

5.4. Event Sound Source Localization Results

We follow [49, 57] and use consensus-IoU ($cIoU@0.5$), whereby the score of each pixel is computed based on the consensus of multiple annotations [49], as well as Area Under the Curve (AUC) [49] for Event Sound Source Localization (ESSL). Following [49], the bounding boxes of the source of sound are provided to the model for the supervised setting during training, but not provided to the model in the unsupervised setting.

Table 4. Results of Event Sound Source Localization

Methods	$cIoU$	AUC
LSS [49] (Unsupervised)	23.85	36.56
SSPL [57] (Unsupervised)	35.50	43.58
LSS [49] (Supervised)	41.85	45.41
SSPL [57] (Supervised)	52.58	48.54
IntelliCare (w/o Multi-task Learning (Supervised))	52.94	48.82
IntelliCare (w/o Dynamic Knowledge Pathfinder (Supervised))	54.66	49.38
IntelliCare (w/ Dynamic Knowledge Pathfinder)(Full model) (Supervised)	**57.37**	**51.44**

5.5. Ablation Studies

Our results in Tables 1 to 4, show that our IntelliCare model that leverages multi-task learning outperforms models that are trained on single task, whereby complementary information from multiple tasks can enhance the model's performance. This is observed by comparing IntelliCare (without Dynamic Knowledge Pathfinder) with IntelliCare (without Multi-task Learning), which has better performance across all tasks. More details on the effectiveness of multi-task learning can be found in Supplementary. With the dynamic design in our Dynamic Knowledge Pathfinder (with $L = 5$ dynamic layers and $B = 2$ blocks in each layer), our model achieves optimal performance. We also find that as B increases, the model's performance is improved (Table 5), and increase becomes marginal when $B > 2$. In addition, the model's performance also increases with the increase of number of L layers (Table 6). However, there is no further distinct increase in model's performance beyond 5 layers. In view of these findings, we use 5 layers with 2 blocks in our model.

Table 5. Results of ablation studies (Number (B) of blocks in each dynamic layer)

Task		STAL	BGA		STEG				ESSL	
			PredCLS	GDet						
Metric		frame-mAP	$R@10$	$R@10$	m-$tIoU$	$vIoU@0.3$	$vIoU@0.5$	m-$vIoU$	$cIoU$	AUC
Number (B) of blocks	1	14.10	15.71	8.75	46.36	50.13	12.20	31.40	54.66	49.38
	2	16.25	16.62	9.51	47.38	51.74	13.60	32.60	57.37	51.44
	3	16.49	17.18	10.20	48.89	53.24	14.50	33.10	57.53	51.68

Table 6. Results of ablation studies (Number (L) of dynamic layers)

Task		STAL	BGA		STEG				ESSL	
			PredCLS	GDet						
Metric		frame-mAP	$R@10$	$R@10$	m-$tIoU$	$vIoU@0.3$	$vIoU@0.5$	m-$vIoU$	$cIoU$	AUC
Number (L) of layers	3	15.73	15.97	8.93	45.23	49.32	12.60	30.80	54.79	50.20
	4	15.92	16.20	9.12	46.62	50.71	13.00	31.30	55.55	50.83
	5	16.25	16.62	9.51	47.38	51.74	13.60	32.60	57.37	51.44
	6	16.42	16.96	9.85	48.11	52.66	14.10	33.70	57.50	51.66

6. Conclusion

Our Chaotic World dataset provides a comprehensive yet challenging benchmark for analyzing human behaviors in chaotic scenes, with fine-grained and dense spatiotemporal annotations of group interaction graphs, individual actions, sounds, and even text descriptions for the scene in each video. We also demonstrate how a unified multi-task IntelliCare model, with Dynamic Knowledge Pathfinder that can intelligently learn and leverage the task-shared and task-specific features, can effectively analyze the chaotic scene. We believe our work will attract the community to further develop and evaluate various types of advanced strategies for human behavioral analysis in chaotic events.

Acknowledgements. This work is supported by Singapore Ministry of Education (MOE) AcRF Tier 2 under award number MOE-T2EP20222-0009 and SUTD SKI Project (SKI 2021-02-06). We would like to thank our annotators Foo Lin Geng, Goh Jet Wei, Hui Xiaofei, Li Rui, Lu Mingqi, Peng Duo, Qu Haoxuan, Shu Xiu, Wang Pengfei, Umali Mike Guil Anonuevo, Xu Li, Zhang Wenxiao for working on the annotations and conducting the quality checks.

References

[1] Firoj Alam, Ferda Ofli, and Muhammad Imran. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, June 2018. 1, 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *NeurIPS*, 2022. 4

[3] George Awad, Keith Curtis, Asad A. Butt, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, Lukas Diduch, Jeffrey Liu, Yvette Graham, and Georges Quénot. An overview on the evaluated video retrieval tasks at TRECVID 2022. In *TREC Video Retrieval Evaluation (TRECVID)*. NIST, USA, 2022. 3

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding? In *ICML*, 2021. 3

[5] Marcel Bouchard, Jennifer Haegele, and Henry Hexmoor. Crowd dynamics of behavioural intention: Train station and museum case studies. *Connection Science*, 27(2):164–187, 2015. 1, 2

[6] Kevin D Browne and Catherine Hamilton-Giachritsis. The influence of violent media on children and adolescents: A public-health approach. *The Lancet*, 365(9460):702–710, 2005. 2

[7] João Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, pages 4724–4733, 2017. 3, 6, 7

[8] Xiaowei Chen and Jian Wang. Entropy-based crowd evacuation modeling with seeking behavior of social groups. *IEEE Access*, 9:4653–4664, 2020. 1

[9] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-Temporal Transformer for Dynamic Scene Graph Generation. In *ICCV*, pages 16372–16382, 2021. 3, 5, 7, 8

[10] Mahsa Ehsanpour, Alireza Abedin, Fatemeh Saleh, Javen Shi, Ian Reid, and Hamid Rezatofighi. Joint learning of social groups, individuals action and sub-group activities in videos. In *ECCV*, pages 177–195. Springer, 2020. 2

[11] Mahsa Ehsanpour, Fatemeh Saleh, Silvio Savarese, Ian Reid, and Hamid Rezatofighi. JRDB-Act: A Large-Scale Dataset for Spatio-Temporal Action, Social Group and Activity Detection. In *CVPR*, pages 20983–20992, 2022. 1, 3

[12] Kristen Grauman et al. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *CVPR*, pages 18973–18990, 2022. 3

[13] Christoph Feichtenhofer. X3D: Expanding Architectures for Efficient Video Recognition. In *CVPR*, pages 200–210, 2020. 3

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 3, 7

[15] Organisation for Economic Co-operation and Development (OECD). *States of Fragility 2016*. Organisation for Economic Co-operation and Development (OECD), 2016. 1

[16] Organisation for Economic Co-operation and Development (OECD). *States of Fragility 2022*. Organisation for Economic Co-operation and Development (OECD), 2022. 1

[17] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017. 7, 8

[18] Ross Girshick. Fast R-CNN. In *ICCV*, 2015. 3

[19] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter N. Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *ICCV*, pages 5843–5851, 2017. 3

[20] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, pages 6047–6056, 2018. 1, 3, 4, 7

[21] Susan L Handy, Marlon G Boarnet, Reid Ewing, and Richard E Killingsworth. How the built environment affects physical activity: Views from urban planning. *American Journal of Preventive Medicine*, 23(2, Supplement 1):64–73, 2002. 1

[22] Alex Hanson, Koutilya Pnvr, Sanjukta Krishnagopal, and Larry S. Davis. Bidirectional Convolutional LSTM for the Detection of Violence in Videos. In *ECCV Workshops*, 2018. 3

[23] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *CVPR Workshops*, pages 1–6, 2012. 3

[24] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 2, 3

[25] Horacio Hojman, Rishi Rattan, Rob Osgood, Mengdi Yao, and Nikolay Bugaev. Securing the Emergency Department During Terrorism Incidents: Lessons Learned From the Boston Marathon Bombings. *Disaster Medicine and Public Health Preparedness*, 13(4):791–798, 2019. 1

[26] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A Hierarchical Deep Temporal Model for Group Activity Recognition. In *CVPR*, pages 1971–1980, 2016. 3, 4

[27] H. Idrees, A. R. Zamir, Y. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. The THUMOS challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 3

[28] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199, Dec. 2013. 3

[29] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action Genome: Actions As Compositions of Spatio-Temporal Scene Graphs. In *CVPR*, pages 10236–10247, 2020. 3

[30] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-

temporal action localization. In *ICCV*, pages 4405–4413, 2017. 3

[31] Abdolamir Karbalaie, Farhad Abtahi, and Mårten Sjöström. Event detection in surveillance videos: A review. *Multimedia Tools and Applications*, pages 1–39, 2022. 1, 2, 3, 5

[32] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale Video Classification with Convolutional Neural Networks. In *CVPR*, 2014. 3

[33] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 2, 3, 4

[34] Viet-Tuan Le, Kiet Tran-Trung, and Vinh Truong Hoang. A Comprehensive Review of Recent Deep Learning Techniques for Human Activity Recognition. *Computational Intelligence and Neuroscience*, 2022, 2022. 2

[35] Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The AVA-Kinetics localized human actions video dataset. In *CVPR Workshops*, 2020. 3

[36] Weiyao Lin, Huabin Liu, Shizhan Liu, Yuxi Li, Hongkai Xiong, Guojun Qi, and Nicu Sebe. HiEve: A Large-Scale Benchmark for Human-Centric Video Analysis in Complex Events. *IJCV*, 2023. 3

[37] Weiyao Lin, Huabin Liu, Shizhan Liu, Yuxi Li, Hongkai Xiong, Guojun Qi, and Nicu Sebe. HiEve: A Large-Scale Benchmark for Human-Centric Video Analysis in Complex Events. *IJCV*, pages 1–25, 2023. 3

[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. *TPAMI*, 42(10):2684–2701, 2019. 3

[39] Qi Meng, Tingting Zhao, and Jian Kang. Influence of Music on the Behaviors of Crowd in Urban Open Public Spaces. *Frontiers in Psychology*, 9, 2018. 2

[40] Mathew Monfort, Kandan Ramakrishnan, Alex Andonian, Barry A. McNamara, Alex Lascelles, Bowen Pan, Quanfu Fan, Dan Gutfreund, Rogério Schmidt Feris, and Aude Oliva. Multi-Moments in Time: Learning and Interpreting Models for Multi-Action Video Understanding. *TPAMI*, PP:1–1, 2019. 3

[41] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa M. Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in Time Dataset: One Million Videos for Event Understanding. *TPAMI*, 42:502–508, 2020. 3

[42] Isabel Ortiz, Sara Burke, Mohamed Berrada, and Hernán Saenz Cortés. *An Analysis of World Protests 2006–2020*, pages 13–81. Springer International Publishing, Cham, 2022. 1

[43] William O'Toole, Stephen Luke, Jason Brown, Andrew Tatrai, and Travis Semmens. *Crowd Management: Risk, Security and Health*. Goodfellow Publishers, 2019. 1

[44] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-Context-Actor Relation Network for spatio-temporal action localization. In *CVPR*, pages 464–474, 2021. 3

[45] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *CVPR*, pages 464–474, 2021. 7

[46] Mauricio Perez, Alex Chichung Kot, and Anderson Rocha. Detection of Real-world Fights in Surveillance Videos. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2662–2666, 2019. 3

[47] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of Textual Phrases in Images by Reconstruction. In *ECCV*, 2016. 7, 8

[48] Sebastian Ruder. An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv preprint arXiv:1706.05098*, 2017. 4

[49] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, pages 4358–4366, 2018. 7, 8

[50] Duygu Sesver, Alp Eren Gençoğlu, Çağrı Emre Yıldız, Zehra Günindi, Faeze Habibi, Ziya Ata Yazıcı, and Hazım Kemal Ekenel. VIDI: A Video Dataset of Incidents. *IEEE Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, 2022. 1, 3

[51] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *CVPR*, pages 1010–1019, 2016. 3

[52] Vijeta Sharma, Manjari Gupta, Anil Kumar Pandey, Deepti Mishra, and Ajai Kumar. A Review of Deep Learning-based Human Activity Recognition on Benchmark Video Datasets. *Applied Artificial Intelligence*, 36(1):2093705, 2022. 2

[53] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *CVPR*, pages 7396–7404, 2018. 3

[54] Gunnar A. Sigurdsson, Gül Varol, X. Wang, Ali Farhadi, Ivan Laptev, and Abhinav Kumar Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *ECCV*, volume abs/1604.01753, 2016. 1, 3

[55] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the Kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*, 2020. 3

[56] Josh Smith, Ellis Ng, Simon Scarr, Ju-min Park, Adolfo Arranz, and Jitesh Chowdhury. How a night of Halloween revelry turned to disaster in South Korea. 2022. 1

[57] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. Self-Supervised Predictive Learning: A Negative-Free Method for Sound Source Localization in Visual Scenes. In *CVPR*, pages 3222–3231, 2022. 7, 8

[58] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *Centre for Research in Computer Vision CRCV-TR-12-01*, abs/1212.0402, 2012. 3

[59] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-Centric Relation Network. In *ECCV*, pages 318–334, 2018. 3, 7

[60] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *TPAMI*, 2022. 2

[61] Anne Templeton and Fergus Neville. Modeling collective behaviour: Insights and Applications from Crowd Psychology. In *Crowd Dynamics, Volume 2*, pages 55–81. Springer, 2020. 1

[62] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target Adaptive Context Aggregation for Video Scene Graph Generation. In *ICCV*, pages 13688–13697, 2021. 3, 5, 7, 8

[63] Vinod Thomas. *Climate change and natural disasters: Transforming economies and policies for a sustainable future*. Taylor & Francis, 2017. 1

[64] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, pages 247–263, 2018. 3

[65] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, pages 4489–4497, 2015. 3

[66] Maarten K. van Aalst. The impacts of climate change on the risk of natural disasters. *Disasters*, 30 1:5–18, 2006. 1

[67] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. NWPU-Crowd: A Large-Scale Benchmark for Crowd Counting and Localization. *TPAMI*, 2020. 1

[68] Ethan Weber, Nuria Marzo, Dim P. Papadopoulos, Aritro Biswas, Agata Lapedriza, Ferda Ofli, Muhammad Imran, and Antonio Torralba. Detecting natural disasters, damage, and incidents in the wild. In *ECCV*, August 2020. 1, 2, 3

[69] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning Actor Relation Graphs for Group Activity Recognition. In *CVPR*, 2019. 3, 4

[70] Peng Wu, Jing Liu, Yujiao Shi, Yujia Sun, Fang Shao, Zhaoyang Wu, and Zhiwei Yang. Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision. In *ECCV*, 2020. 1, 3

[71] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. TubeDETR: Spatio-Temporal Video Grounding with Transformers. In *CVPR*, pages 16442–16453, 2022. 7, 8

[72] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 4

[73] Guijuan Zhang, Dianjie Lu, and Hong Liu. IoT-based positive emotional contagion for crowd evacuation. *IEEE Internet of Things Journal*, 8(2):1057–1070, 2020. 1

[74] Lin Zhuo, Zhen Liu, Tingting Liu, Chih-Chieh Hung, and Yanjie Chai. Modeling crowd emotion from emergent event video. *Computer Animation and Virtual Worlds*, 32(5):e1988, 2021. 1