Focusing on What Matters: Fine-grained Medical Activity Recognition for Trauma Resuscitation via Actor Tracking

Wenjin Zhang[1], Keyi Li[1], Sen Yang[2], Sifan Yuan[1], Ivan Marsic[1],
Genevieve J. Sippel[3], Mary S. Kim[3], and Randall S. Burd[3]
[1]Rutgers University. [2]Waymo. [3]Children's National Hospital.
{wz315, kl734, sy358, sy609}@scarletmail.rutgers.edu, marsic@rutgers.edu
{gsippel, mskim, rburd}@childrensnational.org

Abstract

Trauma is a leading cause of mortality worldwide, with about 20% of these deaths being preventable. Most of these preventable deaths result from errors during the initial resuscitation of injured patients. Decision support has been evaluated as an approach to support teams during this phase to reduce errors. Existing systems require manual data entry and monitoring, which makes tasks challenging to accomplish in a time-critical setting. This paper identified the specific challenges of achieving effective decision support in trauma resuscitation based on computer vision techniques, including complex backgrounds, crowded scenes, fine-grained activities, and a scarcity of labeled data. To address the first three challenges, the proposed system involved an actor tracker that identifies individuals, allowing the system to focus on actor-specific features. Video Masked Autoencoder (Video-MAE) was used to overcome the issue of insufficient labeled data. This approach enables self-supervised learning using unlabeled video content, improving feature representation for medical activities. For more reliable performance, an ensemble fusion method was introduced. This technique combines predictions from consecutive video clips and different actors. Our method outperformed existing approaches in identifying fine-grained activities, providing a solution for activity recognition in trauma resuscitation and similar complex domains.

1. Introduction

Trauma is one of the leading causes of death worldwide [27]. Despite individual and team training, errors persist during the initial resuscitation phase of trauma that can contribute to these deaths [35, 40]. Failures to adhere to established protocols (rule-based errors) and inadequate or incorrect knowledge (knowledge-base) account for 85% of human errors in trauma resuscitation resulting in preventable deaths [16]. While computerized decision support systems offer potential solutions to mitigate these errors, their practical implementation during active resuscitation is limited by the need for manual data entry by a dedicated team member [15]. This challenge led to the development of an automated system that uses computer vision to monitor patient status and activities during trauma resuscitation. Although computer vision can effectively recognize resuscitation activities that are easily visible, previous approaches for identifying relevant but more fine-grained activities (e.g., Intravenous (IV) placement and temperature measurement) have had poor performance.

Several previous studies have used deep learning to identify activities through ceiling-mounted cameras during trauma resuscitation procedures [10, 47]. One system [10] extracted activity-related features (e.g., patient location and medical devices) and applied Markov Logic Networks to differentiate activities with these features. Because of its reliance on manually crafted features, this system may not perform well in the complex and chaotic environments typical of trauma resuscitation due to its inability to generalize [47]. Another system [47] used a 3D Convolutional Neural Network (ConvNet) to recognize activities. This system performed well for easily visualized activities but performed poorly for fine-grain activities like intravenous (IV) catheter placement, likely because of an inability of the model to identify small regions of interest.

Recognizing activities during trauma resuscitation has several challenges, including a complex background, a crowded setting leading to partial and transient visual occlusion of activities, and a need to recognize fine-grained activities that are represented in a small area (i.e., $< 5\%$ out of the whole scene). Obtaining and annotating video data is an additional challenge in this domain. In contrast to existing public video datasets (e.g., Kinetics [26], Something-Something [17], and AVA [18]), patient and provider confidentiality limits the available video samples available for

training models for trauma resuscitation. Domain knowledge is also required for ground truth coding, increasing the cost of video annotation.

This research introduced an approach to address the challenges of identifying fine-grained trauma resuscitation activities, which first identifies the actor and then recognizes the activities performed by that actor. Using this approach, the model focused on the movements of individuals rather than the entire scene, reducing background distractions from a noisy background and attention to irrelevant activities. The system begins with the design of a tracker that detects and tracks actors. After extracting the feature map from the original video stream, the system isolates actor-specific features from the feature map using tracked coordinates and actor identifiers. To address the problem of limited labeled training data, a Video Masked Autoencoder (VideoMAE) pre-trained the feature extractor, i.e., Vision Transformer(ViT), with unlabeled videos. This pretraining process can enhance the ViT model's ability to extract meaningful features for medical activities. To enhance reliability, an ensemble fusion method merged predictions from consecutive video clips and various actors. In summary, the main contributions of this work are as follows:

- The development of fine-grained medical activity system achieves excellent performance, with a mean average precision of 0.78.
- The introduction of an actor tracker isolates actor-specific features to enhance model's focus and recognition performance.
- Ensemble fusion method increases reliability by integrating predictions from consecutive video clips and multiple actors.
- Evaluations indicate that the developed method surpasses existing methods, especially in recognizing fine-grained activities, e.g., IV placement and measuring temperature.

2. Related work

2.1. Vision-based activity recognition

Activity recognition is an active area of research in many domains. This work is focused on identifying specific actions within video segments. Several approaches have been used in this field, including two-stream 2D ConvNets [25, 37], 2D ConvNets combined with LSTMs [12,46], 3D ConvNets [9, 23, 45], and Transformers [3, 36, 39]. These methods have been successfully applied to several activity recognition datasets, including Kinetics-400 [26], UCF-101 [38], Something-Something [17], and Jester [33]. Although these model frameworks can effectively learn temporal and spatial features to recognize actions in short, segmented clips, these approaches face limitations when applied to longer videos typically found in real-world applications. Other investigators integrated object detection with activity recognition for spatiotemporal action detection. Faster-RCNN was used to generate proposals, followed by SlowFast networks to perform multi-label predictions [14]. Subsequent work showed improved results using more robust backbone networks [39]. TubeR was proposed as a solution [49], providing an end-to-end solution capable of simultaneously locating and recognizing activities. The successes in activity recognition [14, 37, 39, 45, 49] are partly due to the availability of high-quality, labeled datasets. These techniques, however, have limited applicability to complex settings with limited labeled data, such as trauma resuscitation.

2.2. Medical activity recognition

Activity recognition systems have been evaluated for application in healthcare, using various sensing techniques, including radio frequency identification (RFID) [5], accelerometers [2, 34], microphones [19], and cameras [47]. Body-attached RFID tags have been used to track the movements of objects and medical personnel in surgical settings [5]. The use of body-attached sensors presents challenges, such as potential interference with medical procedures, maintenance demands, and the time needed for sensor placement, limiting the effectiveness in time-critical settings. To overcome the limitations of wearable sensors, some systems have adopted fixed sensors. RFID tags on medical tools and the use of the received signal strength indication (RSSI) feature are examples of this approach [28, 29]. Fixed microphones have been used to capture speech for recognizing activities [19]. These fixed sensor systems avoid the need for provider interaction but have performance constraints. These limitations include a requirement for pre-tagging tools and a lack of applicability to activities not involving taggable tools. Speech recognition can assist in activities but is limited by the costs and effort of creating transcripts and the challenge of noisy environments with overlapping speech. Vision-based technologies address many of these challenges, including ease of maintenance and the lack of reliance on physical sensors. A 3D ConvNet was introduced for medical activity recognition, achieving high performance for many medical activities [47]. This system recognized medical activities from the entire video view without focusing on the region of interest. Consequently, it exhibits limited performance in identifying fine-grained activities, such as IV placement. This limitation highlights the need for a fine-grained activity recognition method using computer vision.

3. Method

Our activity recognition system consisted of an actor tracker, a feature extractor, an MLP classifier head and an ensemble fusion(Figure 1).

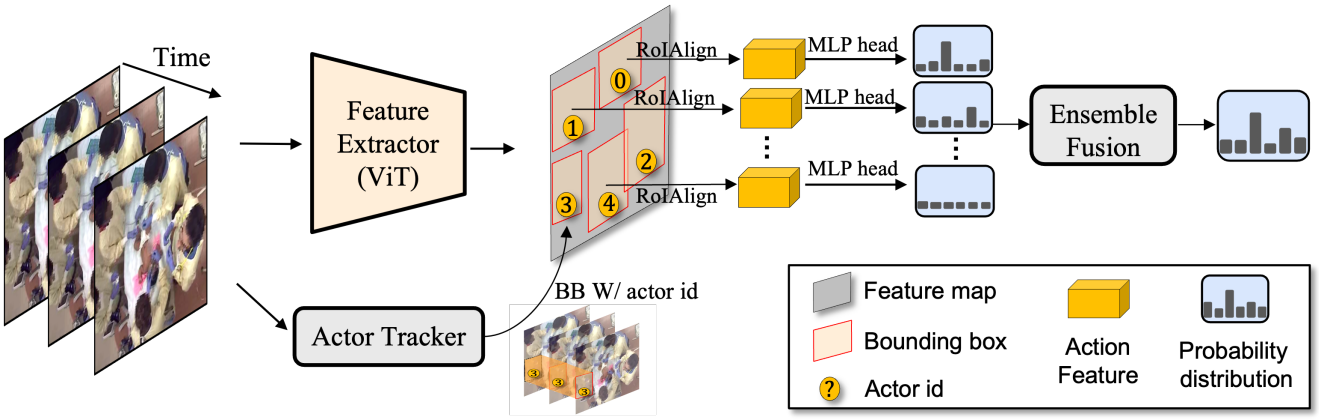

Figure 1. System flow: the actor tracker first tracked actors (i.e., bounding boxes (BB) and actor identification) in the frame sequences. A vision transformer (ViT) extracted spatial-temporal features from the video stream. With the help of bounding boxes and actor identification, the MLP classifier head focused on the action features of each actor to recognize medical activities. Region of interest align (RoIAlign) [21] aligned the dimension of actor features with the input of MLP classifier. Ensemble fusion smoothed the prediction across time and aggregated prediction scores of actors to enhance performance.

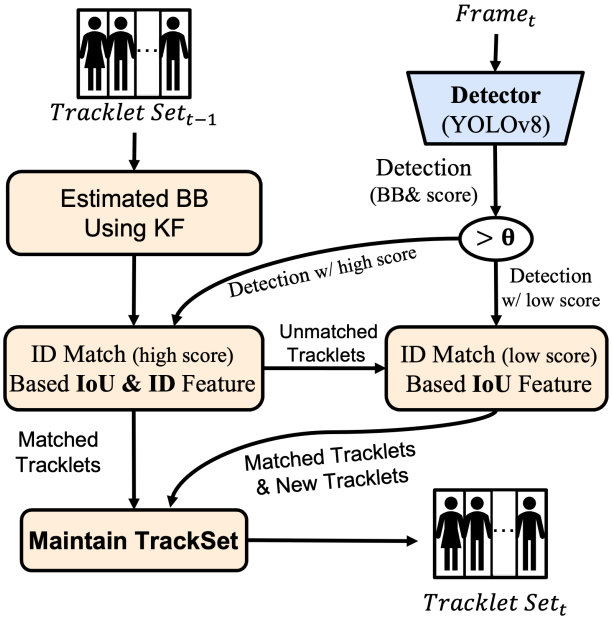

Figure 2. Matching pipeline of actor tracker. The use of a Kalman filter estimated the bounding boxes of Tracklet at the next time frame. The system then matched Tracklet with new detections based on similarity features and updated Tracklet with the matching results.

3.1. Actor Tracker

Multiple actors appear simultaneously in the videos. For this reason, activity recognition relies on the recognition of consecutive features belonging to the same actor. This led us to develop an actor tracker that assigns a unique identifier to the first detected actor and maintains tracking of that actor using the same identifier. The most influential paradigm, tracking-by-detection, was used for this purpose [41,43,48]. Tracklet set was defined as $\mathcal{T}_{t-1}$, which stores tracklet (i.e., bounding box, actor ID, and actor's id-related features) at frame $t-1$. After actor detector (i.e. YOLOv8) predicting the new detections in frame, t, the tracking procedure (Figure 2) was as follows:

- Model the movement of each actor and predict their bounding boxes for the next time frame t to update $\mathcal{T}_t$.
- Associate new detections with current $\mathcal{T}_t$ and maintain $\mathcal{T}_t$:
  - If a new detection matches an existing tracklet, update the tracklet's bounding box with that of the new detection.
  - If a new detection does not match any existing tracklets, create a new tracklet for it in $\mathcal{T}_t$.
  - If an existing tracklet does not find a matching new detection, remove that tracklet from $\mathcal{T}_t$.

Given tracklet $\mathcal{T}_{t-1}$, the Kalman filter (KF) [8] predicted the location of each actor in frame t and update $\mathcal{T}_t$. This estimation process was applied to every actor from the previous frame to refresh the tracklet set $\mathcal{T}_t$. Tracking of multiple actors relied on ID Match process, which calculates the similarity between tracklets and new detections, and then pairs them based on this similarity.

Similarity metric. Location and appearance features play a significant role in matching tracklets [7]. Advances in representation learning have led to the adoption of the Re-Identification (Re-ID) feature model, which extracts deep appearance features and facilitates long-term tracking [1, 42, 48]. In medical rooms, actors may be crowded, and occlusions of medical activities by actors can occur. To

address this challenge, both location and ID features are incorporated into our similarity metric. For location similarity, intersection over union (IoU) distance metric was used:

$$d_{i,j}^{IoU} = 1 - IoU(BB_i^{det}, BB_j^{tracklet}), \quad (1)$$

where BB_i^{det} and $BB_j^{tracklet}$ are bounding box of i^{th} detection and bounding box of j^{th} tracklet, respectively. The IoU distance measure can be ineffective for long-term tracking because it becomes unreliable when an actor is occluded for a prolonged time, leading to inaccurate bounding box estimations for similarity checks. To address this challenge, our approach involved Re-ID feature matching. The advanced BoT(SBS) model [22] extracted Re-ID features, with ResNeSt50 [44] serving as the backbone. Given the potential unreliability of Re-ID features during occlusion, the exponential moving average (EMA) method updated the average Re-ID feature for the i^{th} tracklet using the formula $emb_i^t = \beta * emb_i^{t-1} + (1-\beta) * e_i^t$, where e_i^t is the Re-ID feature embedding of latest matched detection, emb_i^{t-1} represents the previous average feature and β (with a value of 0.9 [1]) acts as the momentum term. The similarity of the Re-ID features was then measured using cosine distance.

$$d_{i,j}^{cos} = 1 - \frac{emb_i^t * e_j^t}{\left\|emb_i^t\right\|_2 \left\|e_j^t\right\|_2}, \quad (2)$$

where e_j^t is the Re-ID feature of j^{th} new detection and emb_i^t is the RE-ID feature of i^{th} tracklet.

Matching pipeline. The matching pipeline was designed to ensure comprehensive tracking of activities in the trauma room, necessitating the tracking of all potential actors while accommodating some false positives, such as background elements. To consider all potential actors, a low detection threshold of 0.2 was applied to the detector. Our pipeline used different strategies for high-score and low-score detections divided at a threshold $\theta = 0.6$. For high-score detections, the pipeline matched them with tracklets based on the lowest IoU and Re-ID distances. For low-score detections, IoU distance was used to match unmatched tracklets, given the potential unreliability of Re-ID features at low detection scores. After associating, tracklet set was maintained by creating new tracklets, updating the tracklets' bounding box with new detection, removing inactive tracklets, and updating Re-ID feature.

3.2. Feature extractor

The feature extractor processes video frames or sequences to obtain high-level features necessary for task recognition. Vision Transformer (ViT) served as our primary feature extraction method. ViT was chosen because of ViT's performance in other applications [13]. VideoMAE-based self-supervised learning further refined the feature extraction process using unlabeled videos.

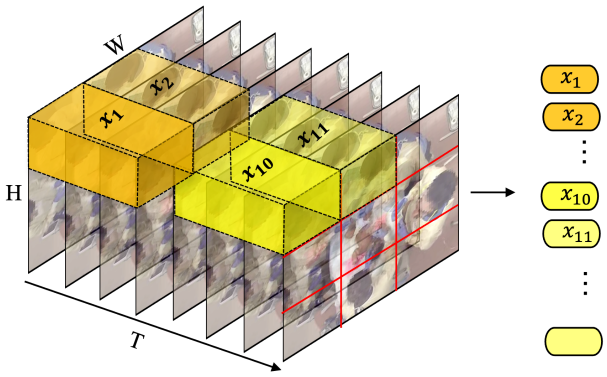

Figure 3. Cube embedding. To tokenize the video as input of ViT, it extracted and linearly embedded discrete tubes that cover the spatiotemporal input volume without overlapping

Video stream pre-processing. In addition to standard pre-processing techniques, including random cropping, flipping, and rescaling, stride temporal downsampling and cube embedding were applied on video data. This is because videos often have high frame rates (e.g., 30 FPS) but slow semantic changes. This method involved initially selecting a clip of N consecutive frames from the video, which was then reduced to $T = \frac{N}{4}$ frames through temporal sampling, with a stride of 4. Each frame had $H * W * 3$ pixels, where H and W denoted the frame's height and width, respectively. In the cube embedding process (Figure 3), video frames were divided into non-overlap cubes, each containing $2*16*16$ pixels, to create individual token embeddings. This division resulted in a total of $\frac{T}{2} * \frac{W}{16} * \frac{H}{16}$ tokens, each of which was then converted into a D-dimensional ($D = 1024$) token embedding. These token embeddings served as the input for the ViT model. This approach reduced the spatial and temporal dimensions of the input, effectively minimizing the spatiotemporal redundancy present in the video data.

Video-MAE-based labeled-free pre-training. Labeling large volumes of video data is challenging. With only limited labeled videos available, training a ViT model was difficult due to the lack of inductive biases in ViTs [13]. A self-supervised learning method, Video-MAE [39]pre-trained the ViT model. Video-MAE introduced a video reconstruction task by randomly masking out video tubes (i.e., video tokens) and reconstructing them through an encoder (i.e., ViT) and a decoder. The decoder was an auxiliary network (a shallow transformer) that assisted with ViT training. For a video $V \in R^{3*H*W*T}$, the training involved a mean squared error (MSE) loss calculation between the normalized masked tokens and their reconstructed counterparts in the pixel space. The loss function was defined as:

$$\mathcal{L}_{mae} = \frac{1}{\Omega} \sum_{idx \in \Omega} \left| V(idx) - \widehat{V}(idx) \right|^2, \quad (3)$$

where idx is the index of a masked tube, Ω represents the set of all masked tubes, and $\widehat{V}$ is the reconstructed video. Following pre-training, the encoder model was used to initialize the ViT, which was then fine-tuned using our labeled video dataset.

3.3. Ensemble fusion

Visual occlusions can lead to noisy predictions from the model. An ensemble fusion technique leveraged medical knowledge to enhance the system's robustness.

Smooth filter across time. Domain knowledge and statistical analyses showed that the minimum duration for activities that we sought to recognize was over 30 seconds [47]. Given that this duration significantly exceeds our model's inference interval of 1 second, a moving average method smoothed the prediction scores over time using the formula:

$$P^{'}(t, id_{actor}) = \frac{1}{N_w} \sum_{t-\lfloor \frac{N}{2} \rfloor}^{t+\lfloor \frac{N}{2} \rfloor} P(t, id_{actor}), \qquad (4)$$

where $P(t, id_{actor}) \in R^C$ is the prediction score at time t, C is number of activity classes, id_{actor} is the actor id and N_w is sliding window size.

Fusion across multiple actors. In resuscitation, patient care activities often involve collaboration among multiple actors. To improve recognition of these coordinated activities, our fusion method integrated predictions from actors using a maximum fusion method. The predicted probabilities of all actors at time t were denoted by $P^{'}(t) \in R^{C*N_{actor}}$, where C is number of classes and N_{actor} is number of actors. The fused predictions were then calculated as $P^{''}(t) = max(P^{'}(t), dim = 1)$. $P^{''}(t) \in R^C$ represents the highest probability across the actors for each class at time t.

4. Experiment

Video recording. Our study was conducted at Children's National Medical Center with approval from the hospital's Institutional Review Board. The emergency department resuscitation room is equipped with a ceiling-mounted camera that records trauma resuscitation for performance improvement. The recording system is activated upon patient entry into the room and terminated upon departure. Parental or guardian informed consent is obtained for the use of videos for research. Over two years, video recordings included 230 trauma resuscitation cases. These videos were captured at a frame rate of 30 fps, with a resolution of 640x480 pixels. The average length of the recordings was 25 minutes, ranging from 16 to 35 minutes.

Labeling strategy. 230 trauma resuscitation videos were divided into three subsets: 170 for unlabeled training, 40 for labeled training, and 20 for testing. The labeling process involved two steps: drawing bounding boxes around each actor and labeling each actor's bounding box with their activity. The labeled training dataset videos were prepared for annotation by extracting frames at one frame per second. Due to the labor-intensive process of manually drawing bounding boxes, **AI-assisted bounding box labeling** was introduced. This method randomly sampled 200 frames from training videos and manually outlined each actor with bounding boxes using the computer vision annotation tool (CVAT) [11]. These manually annotated frames were then used to fine-tune a detection model using YOLOv8 [24], enabling automated bounding box annotations for the rest of the frames in our training dataset. Following annotations, domain experts labeled each actor's bounding box with the medical activities being performed. Using AI-assisted bounding box labeling, annotating a video takes about 1 hour, compared to 7.25 hours for manual labeling. The use of AI-assisted bounding box annotations reduced annotation time by 86%.

Medical activities. This research concentrated on five critical and frequently executed medical activities during trauma resuscitation: cervical collar (c-collar) placement, IV catheter placement, rolling the patient (log roll), placement of a warm blanket (covering warm blanket), and temperature measurement. C-collar placement ensures spinal stability during patient log roll for injury assessment. IV placement facilitates diagnostic and therapeutic interventions through blood draws and medication administration. Temperature measurement is needed to monitor the need for warming measures while covering a patient with a warm blanket helps maintain core body temperature. Among these activities, IV placement and temperature measurement present challenges in computer vision due to the small size of the essential tools. For example, the IV needle and thermometer probe constitute much less than 5% of the total camera view.

Baseline. Due to its specific design for medical activities, our method is compared exclusively with our earlier work in medical activity recognition [47]. Variants of general video understanding models within the medical activity recognition domain are evaluated in the ablation study, focusing on the effects of different backbones.

4.1. Training pipeline.

Actor detection and activity recognition models were trained separately. For both models, a two-step training process was adopted involving pre-training and fine-tuning. Initially, the YOLOv8 model was pre-trained on the COCO dataset [30] and subsequently fine-tuned using our own dataset with labeled bounding boxes, running this process for 100 epochs. This approach aligned with established practices [3, 6] for training video transformers from scratch

on video datasets with limited available samples. For the activity recognition model, a two-stage pre-training process was also conducted. Initially, the ViT-Large model was pre-trained on the ImageNet-1K dataset for 1600 epochs to learn spatial features. The model was then adapted for temporal feature learning by inflating the 2D patch embedding layer into a cube embedding layer. Pre-training continued on the Kinetics datasets [26]. This step helps the model to better understand temporal dynamics by randomly masking out cube embeddings and reconstructing them. The two-stage pre-training process is resource-intensive and time-consuming. For this reason, we opted to download the pre-trained weights. To address the unique characteristics of medical activity motion, the ViT-Large model was further trained using VideoMAE on the unlabeled training set. This process aided in domain adaptation learning. Finally, both the ViT and MLP classifier were fine-tuned using the labeled dataset in a supervised manner for 30 epochs, ensuring that the model was adapted to specific features of targeted activities.

4.2. Implementation detail

Our experiments were conducted with four RTX 3090 GPUs. Our pre-training setup followed a previously outlined configuration on VideoMAE [39]. The ViT-Large and MLP classifier were fine-tuned through several steps. Initially, frames were randomly re-scaled to a range of 256 to 320 pixels, cropped to 256x256 pixels, and subjected to random horizontal flipping with a 50% probability. The AdamW optimizer [32] was used, setting the learning rate to 2.5e-4 and the weight decay to 0.05. A layer-wise learning rate decay [4] of 0.75 was implemented. The batch size was 128 (32 per GPU). Due to the insufficient VRAM capacity of the 3090 GPUs for a batch size of 128, gradient accumulation was employed during training. The training process spanned 30 epochs, including an initial 5-epoch warm-up phase, and utilized a cosine decay schedule [31] for the learning rate.

Table 1. Average Precision. Our approach outperforms the existing vision-based approach.

Activity	ConvNet [47]	Our method
C-collar placement	0.57	0.94
IV placement	0.02	0.45
Log roll	0.87	0.98
Warm blanket	0.65	0.84
Temperature measurement	0.02	0.69
mAP	0.42	0.78

4.3. Results

The effectiveness of our approach was measured using the average precision (AP) metric. The results (Table 1) showed that our method outperformed an existing medical activity recognition method that uses i3D [47], particularly in detecting small-scale activities like IV placement and temperature measurement. Despite the substantial improvements in two small-scale activities, a performance gap remains when compared to other large-scale activities. This gap occurred because essential features in IV placement and temperature measurement, such as tourniquets and thermometer probes, are frequently hidden for extended periods. To address this issue, we planned to enhance accuracy by incorporating multiple viewing angles in future work.

4.4. Ablation study

We conducted several ablation experiments to gain insights into the characteristics of our approach and its influence on the performance of medical activity analysis. We varied actor tracker, ensemble fusion, and video-MAE in Table 2 and studied backbones in Table 3.

Ensemble fusion. Our method showed improvements in performance across all activities compared to the approach without ensemble fusion. A notable improvement in recognition of c-collar placement activity was observed. In this task, an actor's hand often covers the C-collar for periods as brief as 1 second. Ensemble fusion helped mitigate the effects of these short-term occlusions. We observed no difference in the recognition of activities like log roll and IV placement after applying ensemble fusion. The log roll activity was less likely to be occluded by actors. IV placement, however, was often long-term obscured by a warm blanket. Based on these observations, the ensemble fusion proved particularly beneficial for activities prone to brief occlusions.

Actor tracker. To assess the impact of using an actor-tracking-based focus mechanism, we developed a baseline method, i.e., "W/ actor tracker". This method took the entire feature map as input for an MLP classifier. Performance significantly dropped without the actor tracker, particularly for small-scale activities like IV placement and temperature measurement, which have small critical regions and are susceptible to noise. Our improved focus explained the success of our method on small-scale activities. Insufficient training data prevented model learning from focusing on relevant regions. We introduced an actor-tracking-based focus to reduce the difficulty of learning focus and the requirement of training data.

Video-MAE based label-free pre-training. We studied the impact of pre-training with our own unlabeled video dataset, as the benefits of pre-training with a large volume of public videos have already been examined in previous research [39]. We conducted an ablation experiment where we did not pre-train the model using the unlabeled trauma video dataset ("W/O VideoMA"), still pre-training on publicly available data [26]. VideoMAE pre-training

Table 2. Ablation study to system flow. The average precision of our method was reported by removing ensemble fusion, actor tracking, or Video-MAE pre-training

Activity	Our method	W/O ensemble fusion	W/O actor tracker	W/O VideoMAE
C-collar placement	**0.94**	0.75	0.55	0.89
IV placement	**0.45**	0.41	0.02	0.46
Log roll	**0.98**	0.96	0.99	0.97
Covering warm blanket	**0.84**	0.79	0.85	0.82
Temperature measurement	**0.69**	0.56	0.03	0.64
mAP	**0.78**	0.69	0.49	0.75

only marginally enhanced model performance in recognizing the C-collar placement and temperature measurement activities. This minor improvement may be due to the close association of these activities with specific medical equipment (e.g., C-collars and thermometers). Pre-training using VideoMAE likely aided the model's adaptation to the medical domain by learning the spatial features of medical equipment from unlabeled videos. Given that learning these spatial features appears sufficient for domain adaptation, we planned to use imageMAE, which focuses on spatial features only, for domain adaptation to reduce training expenses.

Pre-training with VideoMAE may have helped the model adapt to the medical domain by learning spatial features of medical equipment from unlabeled videos. Due to learning spatial features being enough for domain adaptation, We planned for the future evaluation of performing domain adaptation using imageMAE [20] to save training costs.

Backbone selection. Different backbone architectures were tested, including ViT-Large [13], ViT-Base [13], i3D [9], and SlowFast [14], to assess their impact on recognition performance. Among these, ViT-Large achieved the highest performance, likely due to its greater model capacity and complexity. ViT-Large models had more parameters and layers than smaller models, likely leading to an enhanced ability to discern more detailed features in video data. When comparing ViT-based backbones with conventional convolutional backbones, it is observed that ViT models derive greater benefits from transfer learning and pre-training on extensive datasets. This advantage allowed them to acquire more generalizable features during pre-training, a function allowing models to be efficiently fine-tuned for activity recognition.

5. Conclusion and future work

A fine-grained medical activity recognition system was developed. This system addressed several challenges, including complex environments, crowded scenes, subtle movements, and a lack of labeled data. An actor tracker was used to better focus the model on relevant actions. Its performance was enhanced by implementing video-MAE-based self-supervised learning for domain adaptation and by using ensemble fusion to improve prediction reliability. Testing confirmed the effectiveness of our approach, especially in identifying fine-grained activities. Our system has limitations. First, our method does not entirely resolve long-term occlusion problems, as it mainly reduces the effects of short-term obstructions. Second, deploying our ViT model in clinical settings might be challenging due to resource constraints (e.g. without GPU).

Table 3. Ablation study to backbone. The average precision of our method was reported by varying backbone.

Activity	ViT-L	ViT-B	i3D	SlowFast
C-collar placement	**0.94**	0.89	0.57	0.59
IV placement	**0.45**	0.40	0.32	0.27
Log roll	**0.98**	0.91	0.87	0.68
Covering warm blanket	**0.84**	0.78	0.66	0.77
Temperature measurement	**0.69**	0.65	0.43	0.45
mAP	**0.78**	0.73	0.57	0.56

To address the issue of occlusion, we intend to set up multiple cameras to provide different perspectives (e.g., top, side, and front). This multi-view approach, however, also has challenges. Activities may look different from each angle, complicating the task of consistently recognizing the same activity across various views. Our system must be scalable and capable of accommodating more cameras or adapting to environmental changes without losing performance. Given the limited availability of labeled videos, self-supervised learning, such as constructive learning, is a promising strategy for training models for this purpose. This technique uses the inherent structure and relationships in multi-view data to build strong representations, reducing the dependency on extensive labeled datasets.

Our goal is to create an efficient recognition model that any hospital can use because the computation resource (e.g., GPU VRAM) is limited in hospitals. Privacy concerns prevent us from using cloud-based services (like AWS or Azure) to host our recognition servers. To overcome the challenges of computation resources, we aim to develop a more compact model that retains its effectiveness while requiring less computational power and memory. We plan to

compress our model through several techniques, including model distillation, pruning, quantization, and the use of efficient transformer variants. These methods will help reduce the model's size and resource needs, making it more feasible for widespread deployment.

6. Acknowledgment

This research has been funded under NIH/NLM grant 2R01LM011834-05 and NSF grant IIS-1763827.

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking, 2022. 3, 4

[2] SA Ahmadi, N Padoy, SM Heining, H Feussner, M Daumer, and N Navab. Introducing wearable accelerometers in the surgery room for activity detection. *Computer-und Roboter-Assistierte Chirurgie (CURAC)*, 2008. 2

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer, 2021. 2, 5

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022. 6

[5] Jakob E Bardram, Afsaneh Doryab, Rune M Jensen, Poul M Lange, Kristian LG Nielsen, and Søren T Petersen. Phase recognition during surgical procedures using embedded and body-worn sensors. In *2011 IEEE international conference on pervasive computing and communications (PerCom)*, pages 45–53. IEEE, 2011. 2

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?, 2021. 5

[7] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, Sept. 2016. 3

[8] Robert Grover Brown and Patrick Y. C. Hwang. Introduction to random signals and applied kalman filtering: with matlab exercises and solutions. (3), 1997. 3

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018. 2, 7

[10] Ishani Chakraborty, Ahmed Elgammal, and Randall S. Burd. Video based activity recognition in trauma resuscitation. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013. 1

[11] CVAT.ai Corporation. Computer Vision Annotation Tool (CVAT), Nov. 2023. 5

[12] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description, 2016. 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 4, 7

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition, 2019. 2, 7

[15] Mark Fitzgerald, Peter Cameron, Colin Mackenzie, Nathan Farrow, Pamela Scicluna, Robert Gocentas, Adam Bystrzycki, Geraldine Lee, Gerard O'Reilly, Nick Andrianopoulos, Linas Dziukas, D Cooper, Andrew Silvers, Alfredo Mori, Angela Murray, Susan Smith, Yan Xiao, Dion Stub, Frank McDermott, and Jeffrey Rosenfeld. Trauma resuscitation errors and computer-assisted decision support. *Archives of surgery (Chicago, Ill. : 1960)*, 146:218–25, 02 2011. 1

[16] E. Girard, Q. Jegousso, B. Boussat, P. François, F.-X. Ageron, C. Letoublon, and P. Bouzat. Preventable deaths in a french regional trauma system: A six-year analysis of severe trauma mortality. *Journal of Visceral Surgery*, 156(1):10–16, 2019. 1

[17] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter 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, 2017. 1, 2

[18] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions, 2018. 1

[19] Yue Gu, Ruiyu Zhang, Xinwei Zhao, Shuhong Chen, Jalal Abdulbaqi, Ivan Marsic, Megan Cheng, and Randall S Burd. Multimodal attention network for trauma activity recognition from spoken language and environmental sound. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–6. IEEE, 2019. 2

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 7

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018. 3

[22] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification, 2020. 4

[23] Ying Ji, Yu Wang, and Jien Kato. Spatial-temporal concept based explanation of 3d convnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15444–15453, 2023. 2

[24] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, Jan. 2023. 5

[25] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 2

[26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 1, 2, 6

[27] Blanca Larraga-García, Manuel Quintana-Díaz, and Álvaro Gutiérrez. The need for trauma management training and evaluation on a prehospital setting. *International Journal of Environmental Research and Public Health*, 19(20), 2022. 1

[28] Xinyu Li, Dongyang Yao, Xuechao Pan, Jonathan Johannaman, JaeWon Yang, Rachel Webman, Aleksandra Sarcevic, Ivan Marsic, and Randall S Burd. Activity recognition for medical teamwork based on passive rfid. In *2016 IEEE international conference on RFID (RFID)*, pages 1–9. IEEE, 2016. 2

[29] Xinyu Li, Yanyi Zhang, Ivan Marsic, Aleksandra Sarcevic, and Randall S Burd. Deep learning for rfid-based activity recognition. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*, pages 164–175, 2016. 2

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 5

[31] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. 6

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 6

[33] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2874–2882, 2019. 2

[34] Christian Meißner, Jürgen Meixensberger, Andreas Pretschner, and Thomas Neumuth. Sensor-based surgical activity recognition in unconstrained environments. *Minimally Invasive Therapy & Allied Technologies*, 23(4):198–205, 2014. 2

[35] T.P. Saltzherr, K.W. Wendt, P. Nieboer, M.W.N. Nijsten, J.P. Valk, J.S.K. Luitse, K.J. Ponsen, and J.C. Goslings. Preventability of trauma deaths in a dutch level-1 trauma centre. *Injury*, 42(9):870–873, 2011. 1

[36] Javier Selva, Anders S. Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B. Moeslund, and Albert Clapes. Video transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–20, 2023. 2

[37] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos, 2014. 2

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

[39] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022. 2, 4, 6

[40] Sandra M. Vioque, Patrick K. Kim, Janet McMaster, John Gallagher, Steven R. Allen, Daniel N. Holena, Patrick M. Reilly, and Jose L. Pascual. Classifying errors in preventable and potentially preventable trauma deaths: a 9-year review using the joint commission's standardized methodology. *The American Journal of Surgery*, 208(2):187–194, 2014. 1

[41] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric, 2017. 3

[42] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric, 2017. 3

[43] Fengwei Yu, Wenbo Li, Quanquan Li, Yu Liu, Xiaohua Shi, and Junjie Yan. Poi: Multiple object tracking with high performance detection and appearance feature, 2016. 3

[44] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks, 2020. 4

[45] Wenjin Zhang and Jiacun Wang. Dynamic hand gesture recognition based on 3d convolutional neural network models. In *2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC)*, pages 224–229, 2019. 2

[46] Wenjin Zhang, Jiacun Wang, and Fangping Lan. Dynamic hand gesture recognition based on short-term sampling neural networks. *IEEE/CAA Journal of Automatica Sinica*, 8(1):110–120, 2021. 2

[47] Yanyi Zhang, Yue Gu, Ivan Marsic, Yinan Zheng, and Randall Burd. Video-based concurrent activity recognition for trauma resuscitation. volume 2020, pages 1–6, 11 2020. 1, 2, 5, 6

[48] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box, 2022. 3

[49] Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Shuai Bing, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, Ivan Marsic, Cees G. M. Snoek, and Joseph Tighe. Tuber: Tubelet transformer for video action detection, 2022. 2