Vision-Based Treatment Localization with Limited Data: Automated Documentation of Military Emergency Medical Procedures

Trevor Powers[1,2], Elaheh Hatamimajoumerd[2], William Chu[1], Vishakk Rajendran[3], Rishi Shah[1,3], Frank Diabour[4], Marc Vaillant[1†], Richard Fletcher[1,3†], Sarah Ostadabbas[2†*]
[1]MIT Lincoln Laboratory, [2]Northeastern University Augmented Cognition Lab
[3]Massachusetts Institute of Technology, [4]Texas Tech University
[*]Corresponding Author's Email: ostadabbas@ece.neu.edu

Abstract

In response to the challenges faced in documenting medical procedures in military settings, where time constraints and cognitive load limit the completion of life-saving Tactical Combat Casualty Care (TCCC) Cards, we present a novel end-to-end computer vision pipeline for autonomous detection and documentation of common military emergency medical treatments. Our pipeline is specifically designed to handle limited and challenging data encountered in military scenarios. To support the development of this pipeline, we introduce SimTrI, a labeled dataset comprising 116 twenty-second videos capturing patients undergoing four prevalent treatment procedures. Our pipeline incorporates training and fine-tuning of object detection and human pose estimation models, complemented by a proprietary pose-enhancement algorithm and a range of unique filtering and post-processing techniques. Through comprehensive development and optimization, our pipeline achieves exceptional performance, demonstrating 100% precision and 62% recall on our dedicated 23-video test set. Furthermore, the pipeline automates the generation of TCCC-relevant information, significantly improving the efficiency of TCCC documentation. Comparative analysis against previous state-of-the-art techniques in emergency medical autonomous documentation demonstrates that our pipeline performs exceptionally‡

1. Introduction

In 2009, Secretary of Defense Robert Gates' Golden Hour policy mandated that all critically injured military personnel, known as "battlefield casualties," that are at risk of losing life, limb, or eyesight, would receive a medical evacuation from the point of injury to surgical care within sixty minutes or less [30]. The Golden Hour originates from renowned military surgeon R. Adam Cowley, who identified the urgency for treatment in the hour following an injury, stating "There is a golden hour between life and death. If you are critically injured, you have less than 60 minutes to survive. You might not die right then; it may be three days or two weeks later – but something has happened in your body that is irreparable" [27]. Studies later determined that this time was much lower, between 19 and 23 minutes [3].

Currently, combat medics are required to use a portion of this valuable, limited time to document interventions on the casualty via a tactical combat casualty care (TCCC) Card, which is essential for informing higher echelons of care (flight medics and hospital surgeons) of the casualty's status. An estimated time for a combat medic filling out a TCCC Card and conducting a patient hand-off is about 3 minutes [23]. However, numerous studies indicate this TCCC documentation leads to an increased survival rate among casualties [24, 5]. By doctrine, this card should be attached to the casualty [6]. Unfortunately, two senior US Army combat medics interviewed by this project estimated that only $10-15\%$ of TCCC Cards reach the surgical team receiving the casualty [23, 22]. Because of the inherent time constraints and the unreliability of the TCCC Card, casualty status is often communicated only verbally at the patient

† Senior Author

‡ Our code and the manually annotated dataset can be found at https://github.com/ostadabbas/Vision-Based-Treatment-Localization-with-Limited-Data

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

This material is based upon work supported by the Combatant Commands under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Combatant Commands.

hand-off. This communication is conducted between the combat medic and flight medic in a noisy, high-stress environment via the much shorter and less informative MIST Report - a report summarizing Mechanism of injury, Injuries, Symptoms, and Treatments for a patient [28].

In this paper, we introduce a novel pipeline utilizing computer vision for autonomous TCCC documentation, shown in Fig. 1. To do so, we introduce the first-of-its-kind Simulated Trauma Interventions (SimTrI) dataset, train computer vision models with limited and challenging data, propose a variety of filtering methods, develop our own algorithm to support pose algorithms facing challenging partial body data, and design unique evaluation metrics specific to our use-case. If implemented as a fielded prototype, this software would significantly decrease the time and cognitive load combat medics currently face when documenting casualty status. As a result, medics could solely concentrate on delivering life-saving interventions. Moreover, with automated generation and digital formatting, such a system would guarantee that every TCCC Card (as shown in Fig. 2) reaches all levels of medical care *before* the arrival of the casualty, ensuring comprehensive coverage and enabling preparation for specific procedures at higher echelons of care.

2. Related Works

In recent years, the field of emergency medicine, both in military and civilian contexts, has faced challenges in efficiently documenting and transmitting casualty treatment information across different levels of care. Several studies have addressed this issue, employing various techniques with differing levels of automation [26, 20, 31, 25, 10]. One notable manual approach was the US Air Force's BATDOK system, which provided combat medics with a user interface for manual data entry [14]. However, feedback from military combat medics revealed that this manual data entry was impractical in high-stress combat environments [25, 22, 23]. To enhance autonomy, researchers explored two primary avenues: wearable biosensors and machine learning (ML) methods.

Considering wearable biosensors, several studies have successfully achieved autonomous detection of critical biometrics for emergency treatment, including the BATDOK system. However, these technologies primarily focused on capturing physiological data and did not provide actual treatment information [2, 26]. The application of machine learning (ML) has been primarily limited to the use of automatic speech recognition (ASR) for documenting medical treatments. Woo et al. utilized noise-resilient ASR, multi-style training, customized lexicon, and speech enhancement to predict medically relevant treatment speech at a word error rate of 33.3% [31] to fill out a TCCC Card. This was done by using the Switchboard and Common Voice datasets to train a base ASR model; subsequent modular model improvements were made by generating battlefield noise with a generative adversarial network and domain-specific medical and military data from the Carnegie Mellon University Sphinx Knowledge Base Tool. Similarly, McGeorge et al. heavily relied on ASR and systemic functional grammar models to detect and parse medically relevant text [20]. This group also introduced a small computer vision component, but this was limited to optical character recognition for implementing patient identification. While these advancements reduced the need for manual documentation, they still required medical teams to provide speech input, adding to the cognitive load of combat medics in high-stress military environments.

The application of computer vision in the medical field has seen significant progress [7]; however, its utilization in emergency treatment documentation remains limited. A notable study by Heard et al. introduced a pipeline that utilized Myo devices on a medic's hands to extract arm movements and electromyography data, enabling the detection of various emergency room treatments [10]. Despite this effort, the study's performance fell short of achieving more than 50% accuracy for all treatments.

One of the primary reasons for the scarcity of computer vision-based research in this domain is the inherent challenges associated with the data. Firstly, the availability of visual data within medical treatment spaces is constrained due to legal and ethical medical privacy concerns, resulting in limited datasets [11]. Secondly, egocentric data captured from medics treating patients often exhibit partial body views and rare poses. Patients are frequently in a lying-down position, which presents challenges for pose algorithms, as important information such as facial features may be occluded in many frames. Several studies have endeavored to address these challenges [18, 19, 17, 8]. For instance, Vyas et al. developed a 3D synthetic model generation pipeline to augment body pose data, mitigating the issues posed by limited data in critical applications like healthcare [29]. Liu et al. introduced the Simultaneously-collected multimodal Lying Pose dataset to specifically tackle the challenges associated with lying-down, partially occluded poses [16].

To overcome the hurdles associated with autonomously documenting emergency military medical treatments, we have devised an innovative computer vision-based end-to-end pipeline. This pipeline has been designed to operate effectively, even when faced with limited and challenging data, allowing for real-time identification of treatments administered to military casualties.

3. Methods

In this section, we present our end-to-end pipeline for the detection and documentation of casualty status, supported

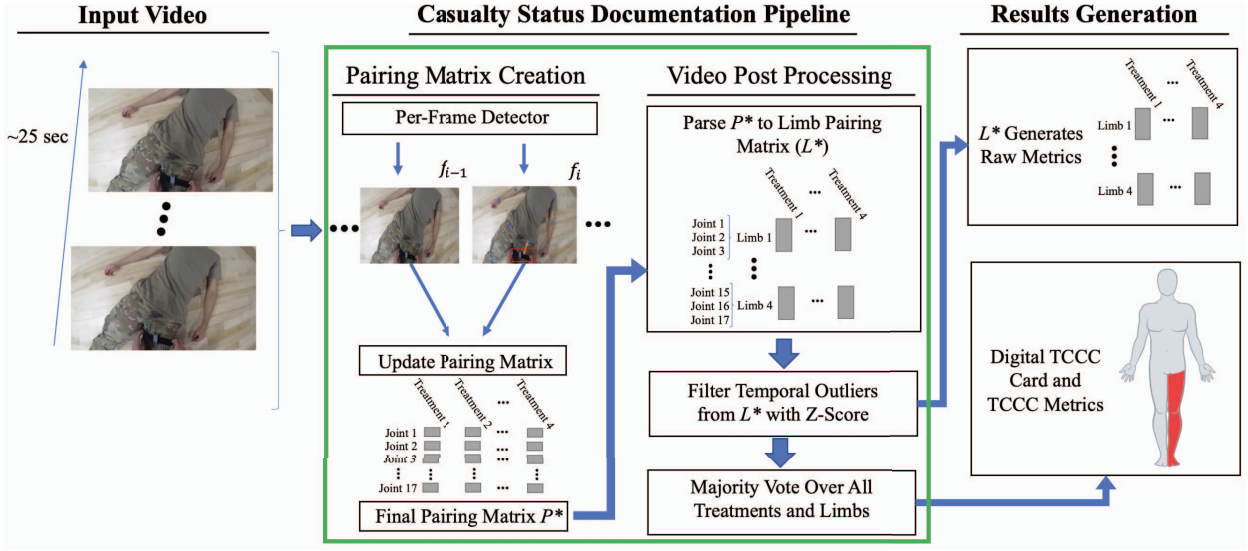

Figure 1. Illustration of our comprehensive pipeline for casualty status documentation. The pipeline consists of two main stages, shown from left to right. In the first stage (Pairing Matrix Creation), the input video is processed frame by frame, and relevant detections are analyzed and summarized to generate a pairing matrix. In the second stage (Video Post Processing), the summarized detections undergo post-processing to extract TCCC-relevant information. Subsequently, in the Results Generation stage, the pipeline generates a digital TCCC card formatted with the extracted information, and its metrics are reported based on the ground truth data.

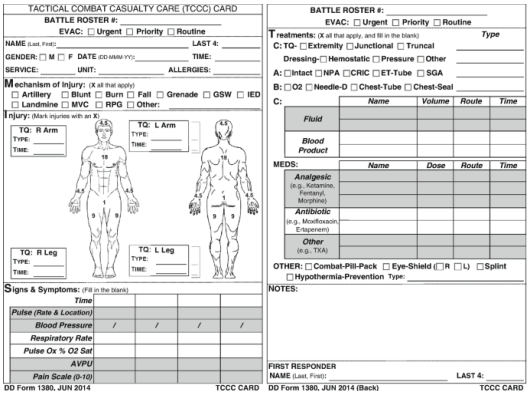

Figure 2. An illustration of the tactical combat casualty care (TCCC) Card.

Figure 3. Example screenshots from a variety of videos in the Simulated Trauma Interventions (SimTrI) dataset. Faces are blurred to protect the identities of the personnel used in these example images.

by our novel dataset called Simulated Trauma Interventions (SimTrI). We begin by introducing the SimTrI dataset, which serves as the foundation for training and evaluating our pipeline. Subsequently, we delve into the details of the pipeline, which encompasses the entire process from receiving a video as input to generating treatment information in a matrix format, mirroring a portion of the content typically seen on a TCCC Card.

3.1. SimTrI Dataset

Due to operational security concerns and the unavailability of public datasets, we collaborated with US Army Special Forces combat medics to generate a unique dataset specifically designed for our study. This dataset, named Simulated Trauma Interventions (SimTrI), consists of 116 egocentric videos that were carefully recorded and approved by the school and Army Institutional Review Boards (IRB). SimTrI features simulated casualties represented by a mannequin and three human subjects with diverse racial backgrounds and body types. The videos capture the performance of four standard military trauma care interventions: tourniquet application, pressure dressing, hemostatic dressing, and chest seal placement. These interventions were selected because they allow for a range of anatomical placement options, necessitating the incorporation of a localization component into our research pipeline. Furthermore, these treatments are commonly taught to all military personnel as part of Combat Life Saver (CLS) training [9]. It is essential to emphasize that the individuals responsible for video documentation in SimTrI are CLS certified military

personnel, ensuring the accuracy and adherence to established medical protocols. The videos were recorded using a helmet-mounted camera positioned approximately three inches above the forehead. The camera was set to record at a rate of 30 frames per second and aligned with the user's line of sight.

In Fig. 3, we provide sample screenshots from various videos within the SimTrI dataset, illustrating the diversity of scenarios and interventions captured in the dataset. By creating the SimTrI dataset in collaboration with US Army Special Forces combat medics, we have obtained a valuable resource for training and evaluating our research pipeline. This dataset enables us to explore new solutions for the detection and documentation of trauma interventions in military settings, ultimately enhancing the care provided to casualties in the field.

Nevertheless, the process of capturing the SimTrI dataset presented certain difficulties. Due to the unique poses and movements of the casualties, as well as the close proximity of the medics to the patient, there were instances where the recorded frames focused primarily on a specific body part, without capturing the face and neck regions. This particular challenge poses a significant obstacle for current human pose estimation (HPE) algorithms, as they heavily rely on visible face and neck features to accurately estimate the pose of an individual.

3.2. Casualty Status Documentation Pipeline

Treatment detection, pose estimation, and the pairing process for treatment localization are the key components of our pipeline, as shown in Fig. 1. To achieve this, we customized and incorporated YOLOv5 [12] and Lightweight OpenPose (LOP) [21] as the backbone for treatment detection and pose estimation, respectively. These frames are then processed through the pipeline's components, enabling treatment detection, pose estimation, and ultimately visualizing the treatments on the TCCC card.

Problem Formulation– This pipeline takes a set of frames, denoted as F, as input. Let f denote an individual frame in F, such that $F = \{f_1, \ldots, f_g\}$, where $g = |F|$. T_i is defined as the array of all treatments detected in frame f_i. Additionally, K_i is defined as the pose key point array corresponding to the patient in frame f_i. If there are multiple pose arrays detected in f_i, the largest pose (by bounding box area which encloses key points) is selected as K_i. Finally, it is important to note the treatment detector is trained on m classes, and the pose detector is trained to recognize n key points to form a human skeleton.

It is critical to pair any detected treatment with a pose key point, in order to localize the treatment to part of the body. To do this, a pairing matrix is used to count the number of pairings made between any detected treatment and pose key point. Upon initiating the pipeline and receiving F as input, this pairing matrix, denoted as P, is initialized with shape (n x m). All values in P are initially set to 0. P is shown below.

$$P = \begin{vmatrix} p_{00} & \cdots & p_{0n} \\ \vdots & \ddots & \vdots \\ p_{m0} & \cdots & p_{mn} \end{vmatrix}$$

Per-Frame Detector– After P is initialized, the per-frame detector (PFD) is sequentially run on all $f \in F$, updating P with the treatment-key point pairings made in each frame. PFD is introduced in Fig. 1 and expounded upon in Fig. 4. PFD receives a frame f_i as input and runs it through the treatment detector and the human pose estimation (HPE) model to output T_i and K_i. If either T_i or K_i is empty, f_i is ignored, and PFD moves on to the next frame. Otherwise, PFD optionally employees a pose enhancement algorithm to enhance K_i, and then maps any detected treatment to the nearest key point while also considering various restriction criteria to filter erroneous detections, as described below.

Pose-Enhancement Algorithm– Given that the data is challenging as many frames only show part of the patient and the patient is often taking on a rare pose, the pipeline may optionally utilize our custom pose-improvement algorithm introduced by this paper, called Pose-Enhancing Transformation Algorithm (PeTA). PeTA is used to enhance any K_i which is incomplete, by adding missing key points to K_i based on poses from previous frames.

PeTA continually estimates a mean casualty pose matrix B of shape $(n \times 2)$ which contains all key points of the standard LOP pose representation. B may then be used to estimate missing joints in frames for which the detected pose is incomplete. To iteratively build B for any frame f_i, a running Procrustes average [13] is calculated from the set of all frames prior to f_i with complete key point arrays. The first frame's key point array of this set is used as the initial estimate of B. Each subsequent frame's key point array in the set is mapped to B by finding the optimal similitude transformation (rotation, translation, and scale) that minimizes the median of squares between corresponding key points. Least median of squares (LMedS) is applied to optimize the transformation and enhance resilience against outliers and noise. The mapped key points are averaged. B is then updated to this new average, and the process is repeated for all remaining frames in the set.

B becomes useful when a K_i does not contain all n points, but only detects a subset of points, as often happens on the challenging data in this topic – that is, $3 < |K_i| < n$. In this case, B may be used to estimate the missing joints in K_i, denoted as $K_{i-missing}$. To do this, it is first necessary to estimate an optimal similitude transformation between B and K_i, to produce a transformation matrix, denoted as τ. LMedS is once again utilized to increase robustness. τ is then applied to B to obtain an estimate of the current key

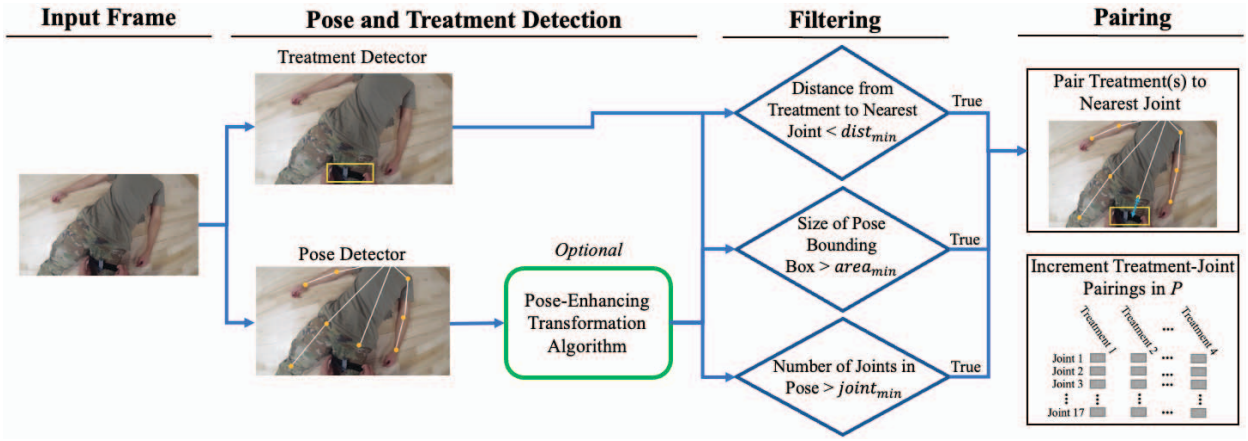

Figure 4. Illustration of our per-frame detector (PFD) pipeline. From left to right, PFD receives a single frame as input. It independently searches for a treatments and a patient pose, optionally employing a pose enhancement algorithm on the detected pose. It considers these detections against three restriction criteria. If all criteria are passed, each detected treatment is mapped to the nearest respective key point of the patient pose.

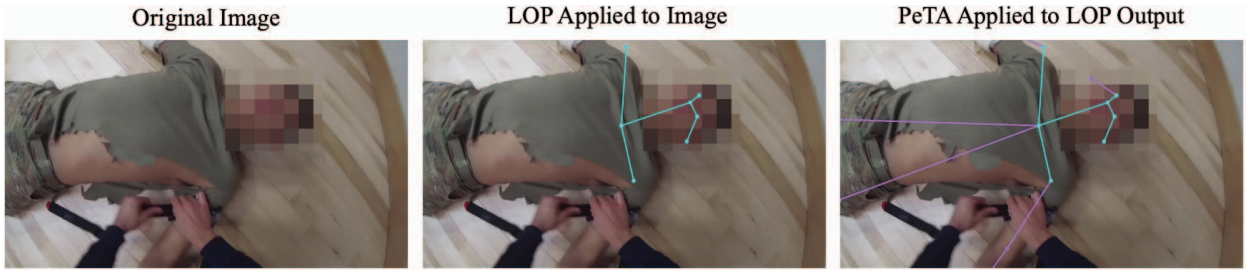

Figure 5. An example of the application of our Pose-Enhancing Transformation Algorithm (PeTA). From left to right, this figure shows the input image, the image following the application of the Lightweight OpenPose (LOP) [21] algorithm, and finally the image after using PeTA to infer the missing joints. Note that the face is blurred for the purposes of this paper, but is not blurred for LOP or PeTA algorithms.

points $\widehat{K} = \tau(B)$. The missing points are then taken from this estimate so that $K_{i-missing} = \widehat{K}_{i-missing}$.

Finally, the estimated missing key points are concatenated to K_i, to form a complete pose key point array. An example of this process is shown in Fig. 5. It is important to note that some of the key points inferred from PeTA are often outside the frame, due to the partial-body nature of the data. Ultimately, PeTA enables estimation of pose key points for frames which fail to detect a full pose but still detect a partial pose. However, for frames that detect fewer than four pose key points ($|K_i| \leq 3$), PeTA is not utilized and the frame is not considered in treatment-key point mapping.

Treatment Key Point Mapping– For any treatment $t_i \in T_i$ for some frame f_i, it is assumed this treatment is on the casualty and must be mapped to some key point $k_i \in K_i$. To do this, for all $t_i \in T_i$, a binary mapping is performed to the nearest pose key point in K_i and P is updated accordingly for $\forall\ t \in T_i$ as:

$$P_{t,k} = P_{t,k} + \begin{cases} 0 & \text{if any } r \in R \text{ is False} \\ 1 & \text{else } \underset{k \in K_i}{\operatorname{argmin}} \textbf{dist}(k,t) \end{cases} \quad (1)$$

A set of restriction criteria R, are also introduced in Eq. (1) to remove pairings that are unlikely to be correct. Any criterion may be tuned to achieve a more or less restrictive pipeline. R is summarized as $R = \{dist_{\min}, area_{\min}, joints_{\min}\}$. Here, $dist_{\min}$ indicates the normalized minimum pixel distance requirement between a potential key point-treatment pairing. Next, $area_{\min}$ indicates the minimum normalized pixel area requirement of a K_i bounding box for the pose to be valid and used for key point-treatment pairing. Lastly, $joints_{\min}$ indicates the minimum number of joints in K_i for the pose to be valid. If any criterion in R is false, the pairing is not counted.

Video Post-Processing– After all $f \in F$ have been analyzed, P is now denoted as P^*. Going forward, each column in P^*, denoted as $P^*_{t \in T}$ is considered independently, as it forms a histogram indicating the various body locations a given treatment class has been mapped to throughout the entirety of F. Now, the casualty may be analyzed at limb-treatment level pairings by summing the entries from $P^*_{t \in T}$ that correspond to the same limb to produce $L^*_{t \in T}$. Then, for each $L^*_{t \in T}$, any limbs that are invalid for a given treatment are dropped (for example, a tourniquet may not be

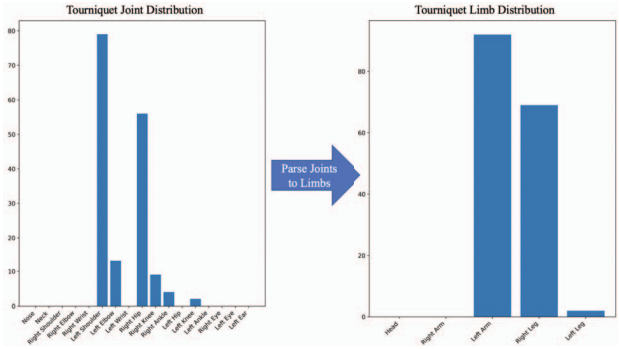

Figure 6. Example of P_t^* being converted to L_t^*, where t is tourniquet.

placed on head, so these erroneous entries are replaced with a 0). The process of converting a treatment column (or histogram) from the joint level to the limb level is depicted in Fig. 6.

Next, temporal outlier detections for $L^*_{t\in T}$ may be filtered. This filtering process is optional. To determine if any pairing in some L_t^* is a temporal outlier, its z-score is computed, denoted as z, with respect to other pairings within the same L_t^*. For any frame f_i, it is assigned a value of 1 if f_i contributed to L_t^*, and a 0 if it did not contribute. A window of w frames preceding f_i is then considered, where all frames also have a value of either 1 or 0. The average of the window μ_w and the standard deviation of the window σ_w are calculated.

The z-score z is computed using the formula: $z = \frac{f_i - \mu_w}{\sigma_w}$. It measures the number of standard deviations by which the pairing value of f_i deviates from the average within the window. If the calculated z-score z exceeds a predefined threshold t, we consider the pairing at frame f_i as an outlier and remove it from further consideration in L_t^*. The purpose of this filtering step is to identify and exclude pairings that are temporally distant from other pairings, and therefore likely erroneous. This process of z-score filtering is conducted for all $L^*_{t\in T}$. Fig. 7 demonstrates an example of this process.

Generating Metrics and Results– Finally, majority voting over $L^*_{t\in T}$ is performed to output a TCCC Card prediction for the m treatments the treatment detector is trained on. To model this digital TCCC Card, a new binary array denoted as H is created with shape (m x $|L|$), where 1 indicates a treatment on a limb is present at the respective index, and 0 indicates the opposite. To predict H from L^*, the following is used:

$$H_{t,\ell} = \begin{cases} 1 & L^*_{t,\ell} \geq \max_{\hat{\ell}\in L} L^*_{t,\hat{\ell}} \times c \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where, for any $t \in T$ and $\ell \in L$, $H_{t,\ell}$ will be 1 if the corresponding $L^*_{t,\ell}$ is greater than the maximum value across all limbs for that t times some constant c, such that $c < 1$.

Using L^* and H, this paper reports two types of metrics - TCCC metrics to determine the accuracy of the outputted TCCC Card and raw metrics to determine the accuracy of L^*. Per-video ground truth labels denoted as Y are utilized to determine these metrics. Similar to H, Y is a binary array with shape (m x $|L|$), where 1 indicates a treatment on a limb is present at the respective index, and 0 indicates the opposite.

For TCCC metrics, true positives TP_H, false positives FP_H, and false negatives FN_H are provided. For raw metrics, only true positives TP_L and false positives FP_L are provided since per-frame truth labels are unavailable to determine the accuracy of negatively predicted frames.

The calculation for TCCC metrics are shown in Eq. (3) through Eq. (5) and the calculation for raw metrics are shown in Eq. (6) through Eq. (7), where $\hat{Y} = 1 - Y$ and $\hat{H} = 1 - H$. From these equations, this paper reports precision P_T and P_R for TCCC and Raw metrics respectively, and recall R_T for TCCC metrics.

$$TP_H = \sum_a^m \sum_b^{|L|} H_{ab} \times Y_{ab} \tag{3}$$

$$FP_H = \sum_a^m \sum_b^{|L|} H_{ab} \times \hat{Y}_{ab} \tag{4}$$

$$FN_H = \sum_a^m \sum_b^{|L|} \hat{H}_{ab} \times Y_{ab} \tag{5}$$

$$TP_L = \sum_a^m \sum_b^{|L|} L^*_{ab} \times Y_{ab} \tag{6}$$

$$FP_L = \sum_a^m \sum_b^{|L|} L^*_{ab} \times \hat{Y}_{ab} \tag{7}$$

4. Experimental Results

This section presents the experimental results for our pipeline. We begin by discussing the performance of our treatment detector and pose estimation models. Subsequently, we evaluate the overall performance of our pipeline using different pipeline configurations. All the results are obtained using our dedicated 23-video test set, which is a subset of the SimTrI dataset. All videos in the test set are of one human subject in uniform. This human subject was not present in any training or fine-tuning sets.

4.1. Treatment Detection Results

We trained our modified YOLOv5 model with three datasets: (1) the BBN PTG-MAGIC dataset (soon to be publicly available) [1], (2) an open-source RoboFlow dataset [4], and (3) our SimTrI dataset. We utilized an

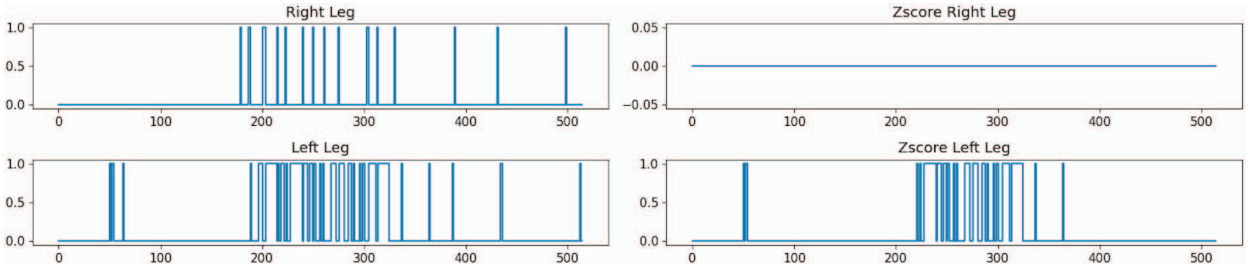

Figure 7. Example of Z-Score Filtering for Leg Detections. The x-axis indicates the frame number within the test video, while the y-axis indicates the presence of a pairing from the given treatment to the respective limb. This example graph demonstrates the utilization of Z-Score filtering to effectively eliminate erroneous detections on the right leg while preserving accurate detections on the left leg.

Table 1. Training data and results for the highest-performing YOLOv5 model.

Treatment	Annotations	Precision	Recall
Tourniquet	13459	87%	54%
Pressure Dressing	9652	94%	80%
Hemostatic Dressing	982	80%	96%
Chest Seal	11211	94%	92%
Average	8826	89%	81%

Table 2. An overview of the pose estimation models, including their names, the number of annotations used in fine-tuning - human or mannequin - and their performance measured as Percentage of Correct Keypoints (PCK). The Base model is the out-of-the-box LOP model. Mann and HuMann are fine-tuned on mannequin data and both mannequin and human data respectively.

Model Name	Human	Mannequin	PCK
Base	0	0	38%
Mann	0	553	51%
HuMann	540	553	57%

80%-20% train-test split, initialized with the YOLOv5 pre-trained weights and freezing no layers. For the purpose of labelling, we defined four classes each corresponding to the various treatments we seek to predict in SimTrI. Tourniquets presented a unique challenge as they are small and often blend in with the body. In addition, the appearance of a tourniquet significantly changes once it is applied to the body. This often resulted in a large number of false negatives. The training set annotations and model performance are summarized in Table 1.

4.2. Pose Estimation Results

For developing our HPE model, we utilized the publicly provided LOP training weights, trained initially on the COCO dataset [15] for 370,000 training iterations. We fine-tuned these weights with different strategies using labeled images from both mannequin and human videos in SimTrI and evaluated these models on 152 randomly selected images of humans in uniform from SimTrI. The 2nd and 3rd columns of Table 2 show the number of images in the training set for various fine-tuned pose models for human and mannequin data respectively with roughly the same number of images with and without uniform clothing. The Percentage of Correct Keypoints (PCK) scores are reported in the last column of Table 2. The best results are obtained when we use both mannequin and human images in training. Apart from having relatively small training and test datasets, a particular challenge to this data is the lack of face or neck in many images as pose estimation relies on detecting and associating key features on the body with one another. Additionally, the uniform worn by the subject is unlike the clothing found in most HPE training datasets. However, our model learns well, and the mannequin-only model clearly shows generalizability to human test data. This indicates strong potential for scalability, as human data is more difficult to obtain within the medical domain, but if mannequin data may achieve similar results, this pipeline may be easily scaled for broader future use.

4.3. End to End Pipeline Results

We evaluated the overall performance of our end-to-end pipeline with SimTrI. We analyzed our pipeline's precision and recall based on Eq. (3) through Eq. (7). The results are reported in Table 3.

In these results, we evaluated the impact of various parameters. We considered all HPE models (see Table 2), the z-score filtering window size w (if z-score is applied), the majority vote c value in Eq. (2), as well as the use of PeTA. We also considered the restriction parameters R in Eq. (1).

Our results indicate our pipeline could provide the software backbone of a promising solution to emergency medical documentation. While our baseline model alone struggles to achieve high-level results, we demonstrate in Table 3 that the iterative improvements our pipeline implemented throughout Section 3 enable favorable results.

First, we show our base pipeline with no parameters in use, providing a baseline score. Next, we show that post-video filtering via majority voting and z-score filtering lead

Table 3. Results of our pipeline with different parameters. In the table, *T*, *F*, and *NA* represent True, False, and Not Applied, respectively. P_T and R_T denote TCCC precision and recall for the respective pipelines, while P_R indicates raw precision for the respective pipeline.

HPE	w	$d_{\min}$	$\alpha_{\min}$	$j_{\min}$	c	$PeTA$	$\mathbf{P_R}$	$\mathbf{P_T}$	$\mathbf{R_T}$
Base	*NA*	1	0	1	0	*F*	67%	37%	48%
Base	60	1	0	1	.5	*F*	73%	68%	48%
Base	60	1	0	1	.5	*T*	85%	90%	59%
Base	60	.25	.1	10	.5	*F*	91%	96%	52%
Mann	60	.1	.1	5	.5	*F*	94%	**100%**	57%
Mann	60	.25	.1	5	.5	*F*	94%	96%	**62%**
HuMann	60	.1	.1	1	.5	*F*	**99%**	**100%**	**62%**
HuMann	60	.25	.3	1	.5	*T*	96%	96%	**62%**
Mann	60	.5	.1	1	.5	*T*	93%	**100%**	**62%**

to improvements in all metrics. Then we demonstrate that the addition of either filtering with R parameters or utilizing PeTA can improve results. However, we found that utilizing both at the same time provided sub-optimal results on the Base LOP model.

We next consider LOP fine-tuned models. We show with relatively little data, fine-tuning our LOP model with mannequin-only data leads to significant improvement for our overall pipeline, indicating promising results for the generalizability of mannequin data to our problem set. However, utilizing both mannequin and human data leads to slightly higher metrics for our pipeline, as expected given the test data is solely human data.

Consequently, we add PeTA for our optimal Mann and HuMann pipeline configurations. Here, it is important to note that the most important metrics are the TCCC-metrics, as this predicts the accuracy of a TCCC card, the end goal output. When we compare our optimal Mann pipeline configuration with PeTA, against the overall optimal configuration (which uses HuMann, but no PeTA), we find these pipelines produce the exact same TCCC-metrics. The implications of this must not be understated. This implies that by fine-tuning only on mannequin data and tuning our pipeline parameters, we may achieve the same TCCC results that we would achieve with human data. Given significant ethical and legal considerations often slow progress for the collection of human data, these results indicate this research may scale more rapidly than it would if it were dependent on human data for optimal results.

It is worth noting that in the context of combat medics, high precision is considered more important than high recall [23]. This is because false positives can be more detrimental than false negatives, as they would require medics to verify the accuracy of the entered data, which could be time-consuming and hinder their workflow. Conversely, false negatives, where certain areas are left blank, do not undermine the medics' trust in the software or discourage its usage, as they can easily fill in the missed areas while still benefiting from the detections in other areas. Therefore, our ability to achieve high precision is a promising outcome for future research aiming to develop a fielded prototype based on this software pipeline.

5. Conclusion and Future Work

In response to the critical need for automated TCCC Card documentation within the US military, we have introduced a comprehensive end-to-end pipeline for military treatment documentation. In addition to this processing pipeline, we have created and curated the SimTrI dataset, which represents a significant contribution to this application domain and enables researchers to develop new computer vision solutions.

Our processing pipeline leverages state-of-the-art techniques, utilizing human pose estimation and object detection as its foundation, while incorporating various filtering and post-processing methods to enhance the accuracy of the results. Despite encountering several challenges during the development process, including the limited number of labeled frames and the partial visibility of facial features or the full body of casualties in the SimTrI dataset, we have achieved highly favorable results. Our pipeline attained an excellent precision rate of 100% in accurately predicting TCCC-relevant information, with a recall rate of 62%, indicating a substantial level of accuracy in identifying and localizing treatments administered.

Looking ahead, there are several exciting avenues for further advancements. Future work should focus on expanding the capabilities of our pipeline to encompass a wider range of treatments and incorporate a more diverse set of patients, thereby enhancing its applicability and versatility in various scenarios. Furthermore, optimizing the recall rate will be a key objective, aiming to increase the percentage of accurately predicted TCCC Cards. Additionally, ongoing efforts will be dedicated to refining and improving the underlying machine learning models that form the foundation of our pipeline.

References

[1] Bbn ptg-magic. www.bbn.com/ptg-magic, 2023. 6

[2] Shireen Bedi. BATDOK improves, tailors to deployed medical Airmen. *Air Force Medical Service*, May 2019. 2

[3] Brian C. Beldowicz, Michael Bellamy, and Robert Modlin. Death ignores the golden hour. *Military Review*, 100(2):25–33, 2020. 1

[4] Bio1SysCombined. piie combined ds test 3 dataset. https://universe.roboflow.com/bio1syscombined/piie-combined-ds-test-3, jan 2023. visited on 2023-01-16. 6

[5] Frank K. Butler. Two decades of saving lives on the battlefield: tactical combat casualty care turns 20. *Mil Med*, 182(3):e1563–e1568, Mar 2017. 1

[6] Combat Studies Institute Press. *Combat Studies Institute Press Publications*. US Army Command and General Staff College, 2018. 1

[7] Andre Esteva, Kelvin Chou, Serena Yeung, and et al. Deep learning-enabled medical computer vision. *npj Digital Medicine*, 4(1):5, 2021. 2

[8] Masoud Farshbaf, Rasoul Yousefi, M Baran Pouyan, Sarah Ostadabbas, Mehrdad Nourani, and Matthew Pompeo. Detecting high-risk regions for pressure ulcer risk assessment. In *2013 IEEE International Conference on Bioinformatics and Biomedicine*, pages 255–260. IEEE, 2013. 2

[9] Maranda Flynn. Combat lifesaver course trains soldiers to save lives on, off battlefield. *Army.mil*, July 2014. 3

[10] Jamison Heard, Richard A. Paris, Deirdre Scully, Candace McNaughton, Jesse M. Ehrenfeld, Joseph Coco, Daniel Fabbri, Bobby Bodenheimer, and Julie A. Adams. Automatic clinical procedure detection for emergency services. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 337–340, 2019. 2

[11] Xiaofei Huang, Nihang Fu, Shuangjun Liu, and Sarah Ostadabbas. Invariant representation learning for infant pose estimation with small data. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021. 2

[12] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, Laughing, tkianai, Adam Hogan, lorenzomammana, yxNONG, AlexWang1900, Laurentiu Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Francisco Ingham, Frederik, Guilhen, Hatovix, Jake Poznanski, Jiacong Fang, Lijun Yu, changyu98, Mingyu Wang, Naman Gupta, Osama Akhtar, PetrDvoracek, and Prashant Rai. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements, Oct. 2020. 4

[13] David G. Kendall. A survey of the statistical theory of shape. *Statistical Science*, 4(2):87–99, 1989. 4

[14] Air Force Research Labs. Battlefield assisted trauma distributed observation kit (batdok) software tools, 2022. 2

[15] 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. 7

[16] Shuangjun Liu, Xiaofei Huang, Nihang Fu, Cheng Li, Zhongnan Su, and Sarah Ostadabbas. Simultaneously-collected multimodal lying pose dataset: Enabling in-bed human pose monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1106–1118, 2023. 2

[17] Shuangjun Liu and Sarah Ostadabbas. A vision-based system for in-bed posture tracking. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1373–1382, 2017. 2

[18] Shuangjun Liu and Sarah Ostadabbas. Seeing under the cover: A physics guided learning approach for in-bed pose estimation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 236–245. Springer, 2019. 2

[19] Shuangjun Liu, Yu Yin, and Sarah Ostadabbas. In-bed pose estimation: Deep learning with shallow dataset. *IEEE journal of translational engineering in health and medicine*, 7:1–12, 2019. 2

[20] Nicolette M McGeorge, Susan Latiff, Christopher Muller, Lucas Dong, Ceara Chewning, Daniela Friedson-Trujillo, and Stephanie Kane. Design and development of a prototype heads-up display: Supporting context-aware, semi-automated, hands-free medical documentation. In *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, volume 10, pages 18–22. SAGE Publications Sage CA: Los Angeles, CA, 2021. 2

[21] Daniil Osokin. Real-time 2d multi-person pose estimation on CPU: lightweight openpose. *CoRR*, abs/1811.12004, 2018. 4, 5

[22] Trevor Powers. Interview with Chris Macnamara. Personal Interview, July 2021. 1, 2

[23] Trevor Powers. Interview with Kyle Johnson. Personal interview, July 2021. 1, 2, 8

[24] JB Robinson, MP Smith, KR Gross, SW Sauer, JJ Geracci, CD Day, et al. Battlefield documentation of tactical combat casualty care in afghanistan. *US Army Med Dep J*, 2016(2-16):87–94, 2016. 1

[25] Rishi Shah. *An Autonomous Casualty Status Communication Tool*. PhD thesis, Massachusetts Institute of Technology, 2021. 2

[26] Alex Sorkin, Avishai M Tsur, Roy Nadler, Ariel Hirschhorn, Ezri Tarazi, Jacob Chen, Noam Fink, Guy Avital, Shaul Gelikas, and Avi Benov. Bladeshield 101: A novel prehospital digital wearable combat casualty card. *The Israel Medical Association Journal: IMAJ*, 24(9):602–605, 2022. 2

[27] University of Maryland Medical System. Shock trauma - about - history. https://www.umms.org/ummc/health-services/shock-trauma/about/history, Accessed: 2023. 1

[28] U.S. Army. *Tactical Combat Casualty Care Handbook*. Center for Army Lessons Learned, 2017. 2

[29] Kathan Vyas, Le Jiang, Shuangjun Liu, and Sarah Ostadabbas. An efficient 3d synthetic model generation pipeline for human pose data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1542–1552, June 2021. 2

[30] Leslie Waghorn. New research shows ‘golden hour’ trauma care saves lives on the battlefield. *Vital Record*, 2014. 1

[31] MinJae Woo, Prabodh Mishra, Ju Lin, Snigdhaswin Kar, Nicholas Deas, Caleb Linduff, Sufeng Niu, Yuzhe Yang, Jerome McClendon, D Hudson Smith, Stephen L Shelton, Christopher E Gainey, William C Gerard, Melissa C Smith, Sarah F Griffin, Ronald W Gimbel, and Kuang-Ching Wang. Complete and resilient documentation for operational medical environments leveraging mobile hands-free technology in a systems approach: Experimental study. *Journal of Medical Internet Research*, Volume(Number):Pages, Oct 2021. 2