2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

CribNet: Enhancing Infant Safety in Cribs through Vision-based Hazard Detection

Shaotong Zhu[1], Amal Mathew[1,2], Elaheh Hatamimajoumerd[1], Michael Wan[3], Briana Taylor[4], Rajagopal Venkatesaramani[2], Sarah Ostadabbas[1*]
[1]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA
[2]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
[3]Institute for Experiential AI, Northeastern University, Portland, ME, USA
[4]Roux Institute, Northeastern University, Portland, ME, USA
*Corresponding Author's Email: ostadabbas@ece.neu.edu

2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG) | 979-8-3503-9494-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/FG59268.2024.10581871

***Abstract*— Recent advancements in object detection and human activity recognition have shown commendable progress, albeit with a predominant focus on adult-centric applications and datasets. This paper proposes a new vision-based, infant-focused hazard detection framework, CribNet, to assess threats to in-crib safety in the form of blanket occlusions and hazardous toys, as a step towards addressing the broad, critical problem of infant sleep safety. CribNet estimates hazards by considering the proximity and characteristics of detected objects around the infants. To evaluate the framework, we created the first publicly available crib hazard detection (CribHD) dataset, consisting of 1,620 images specific to infant-centric environments. These images present a wide range of real-world challenges, including clutter, occlusion, varied lighting conditions, with and without presence of infants in the images. We show that the framework performs with over 80% mean average precision (mAP) in segmenting toys and blankets and accurately assessing hazards, marking a new advancement in infant safety. CribNet and CribHD lay the foundation for future developments in in-crib hazard detection and infant sleep safety[1].**

I. INTRODUCTION

A recent report from the Centers for Disease Control and Prevention (CDC) showed a concerning uptick in the U.S. infant mortality rate for the first time in two decades, with an increase of 3–4% from 2021–2022 [8]. Despite a marked reduction in infant deaths since the 1990s due to large public health campaigns regarding infant sleep safety, a majority of infant deaths still occur during sleep or within the sleep environment [28] and rates of suffocation and strangulation in bed have actually increased in recent years [6]. Sleep is a highly vulnerable state for infants, characterized by reduced arousal and dampened responsiveness to exogenous threats (e.g., suffocation and strangulation hazards). The American Academy of Pediatrics (AAP) recommends avoiding crib accessories such as blankets, crib bumpers, positioners, and toys. However, data show that parents report using these products despite previously held intentions to abide by infant sleep safety recommendations [32], [39]. Moreover, the availability of unsafe crib accessories and commercial depictions of infants sleeping promote the impression that crib accessories are harmless [12], [5]. In fact, data from the 2016 Pregnancy Risk Assessment Monitoring System, a study of over 30,000 women from 29 U.S. states revealed that the use of unsafe crib accessories is not uncommon, with approximately 50% of women reporting the use of blankets and 8–17% reporting the use of crib bumper pads, toys, cushions, or pillows [16]. All of these in-crib accessories have been associated with accidental suffocation or strangulation [10].

Computer vision is well-positioned to advance sleep safety by enabling automatic detection of such safety hazards. Computer vision algorithms can be integrated into many commercially available sleep monitoring systems that are already market-tested for ease of use and acceptability. Furthermore, computer vision predictions like object locations and segmentations are inherently interpretable. This opens up the possibility of safety monitoring systems which could address the aforementioned awareness issue, by just *telling* users about a hazard, but *showing* them *how* it might be hazardous. For instance, rather than simply altering the caregiver to the presense of a blanket in the crib, the algorithm we describe in this work could offer the user a sobering statistic of how often and for how long their infant's face was covered by a blanket.

Despite their promising potential, vision-based detection systems tailored for such environments remain significantly under-explored. One major barrier is the lack of comprehensive datasets capturing in-crib hazards like toys and blankets to support a deep-learning-based approach. Current object detection research predominantly targets adult settings, while the infant domain and infant-centric items are often under-examined, thus underscoring the need for a dataset focused on the infant context.

To address this shortfall, we present CribNet, a novel framework using state-of-the-art computer vision algorithms for the segmentation and analysis of toys and blankets in crib settings (See Fig. 1). CribNet's goal is to identify and assess the potential dangers of these objects. Central to this system is the crib hazard detection (CribHD) dataset, which includes three specific subsets: CribHD-T containing 1000 toy images, CribHD-B containing 500 blanket images in infant environments, and CribHD-C containing 120 challenging images of

This study was supported by the National Science Foundation CAREER Award (NSF-IIS 2143882).

[1]The code and our data are publicly available at https://github.com/ostadabbas/CribNet.

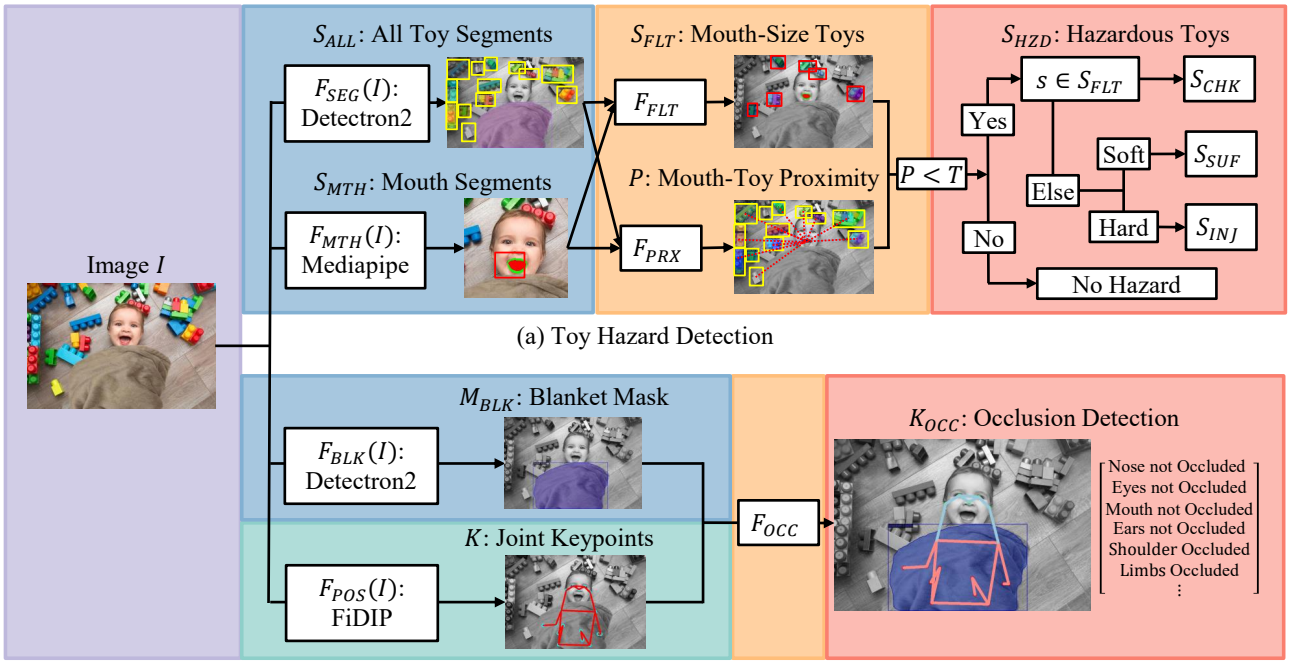

Fig. 1: A schematic workflow of our CribNet framework for detecting hazards to infant safety in cribs, which has two principal components: one system for detecting toy hazards and one for assessing blanket occlusions. We use a color-coded system to delineate components: purple for input, blue for segmentation functions, cyan for infant pose estimation, orange for filtering steps, and red for the final output step. The input image, I, highlighted in purple, is the starting point for both frameworks. In subplot (a), the upper part of the diagram, the method for detecting hazardous toys within I is detailed. This involves choosing toys from the total toy segmentations S_{ALL}, then applying a mouth-size filter F_{FLT} and a proximity threshold T relative to the mouth. A decision tree then assesses the risks, identifying choking hazards S_{CHK} from smaller hard toys, injury hazards S_{INJ} from larger hard toys, and suffocation hazards S_{SUF} from soft toys. Subplot (b), the lower part of the diagram, delineates the blanket occlusion detection framework. This pathway computes occluded infant body parts by overlaying the infant pose estimation data K with the blanket segmentation mask M_{BLK} to identify the occluded keypoints K_{OCC}.

simulated hazard scenes in crib environment, all of which will be accessible to the public. Examples from each category in the CribHD dataset can be found in Fig. 2. This paper's contributions are threefold:

- **Creation of a Hazardous Object Detection Framework (CribNet)**: CribNet integrates advanced detection and segmentation techniques, uniquely combined with an infant-specific pose estimation model, to filter out the hazardous objects in crib environments.
- **Curation of the Crib Hazard Detection Dataset**: We present CribHD, a novel dataset for hazardous object detection in crib environments, with a CribHD-T subset for toys, a CribHD-B subset for blankets, and a CribHD-C for simulated hazard scenes, all richly annotated with bounding boxes and segmentation masks. This dataset fills the existing data gap in crib infant safety research.
- **Comprehensive Evaluation of Detection Methods**: We rigorously evaluate a variety of cutting-edge detection methods from CribNet, validating the robustness and effectiveness of the CribHD dataset and demonstrating its wide applicability and adaptability for different detection models.

Moving forward, we aim to refine these models and expand our dataset to cover a broader spectrum of infant-centric environments and interactions, and beyond toys and blankets to include other hazards like crib bumpers, cushions and pillows, and positioners. Our goal is to evolve these technologies into real-world monitoring systems, providing caregivers with essential insights for proactive infant safety. This work not only advances the state-of-the-art, but also promises substantial real-world impact in enhancing infant well-being.

II. Related Work

A. Object Detection for Hazard Detection Applications

Object detection, a key task in computer vision, plays a pivotal role in hazard detection. It enables the identification and accurate localization of objects in images, thereby significantly contributing to semantic scene interpretation. This automated capability to understand and interpret the content within images or videos marks a crucial step in the evolution of computer vision, as highlighted in recent studies [45]. The object detection problem encompasses two primary steps: object localization for identifying the object's location and object classification for determining the object's category [41]. Object detection frameworks fall into two categories: region proposal-based methods, like Faster R-CNN [11], Mask R-CNN [15], and ClusterNMS [42], focusing on generating and classifying region proposals, and classification-based methods, like SSD [25], POTO [38], and YOLOv7 [37], which treat object detection as a regression problem, aiming to simultaneously classify and locate objects.

(a) (b) (c)

Fig. 2: Examples from the crib hazard detection (CribHD) dataset are presented. Figure (a) shows CribHD-T subset, which is organized by soft and hard toys. Figure (b) displays selected instances from the CribHD-B subset, illustrating the diversity in blanket shapes and colors. Figure (c) depicts scenarios from the CribHD-C subset, portraying meticulously crafted simulations of potential hazard scenes where toys are positioned in proximity to an infant, or blankets partially cover the infant, or a combination of both.

The application of these advanced object recognition technologies in hazard detection has marked a significant leap in this specific area. In the construction industry, Jeelani et al. [21] use Mask R-CNN, fine-tuned with a custom visual vocabulary [9], to detect hazards for construction workers. This illustrates the use of object recognition in industrial contexts. Powers et al. [31] developed a new pipeline for identifying hazardous situations on the battlefield, particularly injuries, and localizing treatment using YOLOv5 [22] and lightweight OpenPose [29]. Meanwhile, Pena-Caballero et al. enhanced YOLOv3 [30] using specialized data to identify road hazards like cracks and potholes. Additionally, Sanjai et al. [34] improved YOLOv5 [22] with a tailored vehicle dataset for the detection of hazardous vehicles on roads, highlighting the effectiveness of these technologies in daily scenarios for hazard detection. These studies exemplify the crucial role of precision and speed in object recognition models for accurately identifying risks and enhancing safety in diverse contexts.

Despite the advancements in object recognition, methods trained on established datasets like MS-COCO [24], Objects365 [35], and open images detection (OID) [23] often overlook hazardous objects pertinent to specific settings. For targeted hazard detection, especially in crib environments, there is a crucial need for a specialized dataset encompassing various hazardous objects. Existing datasets generally do not include specific crib-related hazards, such as small toys that pose choking risks or blankets that might cover an infant's face. As a result, even advanced object detection models specialized for certain hazard detection tasks may be inadequate for accurately identifying such specific in-crib hazards.

B. Vision-based Infant Safety Monitoring

In the field of infant safety monitoring, computer vision technologies have significantly advanced, focusing on applications like pose estimation [17], landmark detection [36], tracking[1], and action recognition [14], [19]. Huang et al. [20] employed an infant-specific fine-tuned domain-adapted 2D pose estimation model called FiDIP (Fine-tuned Domain-adapted Infant Pose), derived from an adult pose estimation model [18] and trained on their SyRIP dataset, coupled with Bayesian methods for precise postural symmetry analysis. Recently, Hatamimajoumerd et al. [14] proposed a pipeline for automatic infant action recognition employing state-of-the-art models such as PoseC3D [7] and InfoGCN [3]. Their pipeline was evaluated on their dataset named InfActPrimitive, containing basic milestones in infant development. Manne et al. [27] introduced AIRFlowNet, a data-efficient framework for non-contact infant respiratory rate estimation using video data, trained on their AIR-125 dataset. Zhu et al. [43] developed a spatial-temporal deep learning method trained on a clinical in-crib dataset for detecting non-nutritive sucking (NNS) in infants. These innovations represent substantial progress in understanding and monitoring infant behavior, offering low-cost, non-invasive methods that could reduce healthcare costs compared to traditional medical examinations.

Despite significant progress in infant monitoring technologies, there remains a notable gap in research focusing on hazard detection in infant environments. Current studies largely center on general infant pose or behavior but do not adequately address the identification of hazardous situations for infants, which is crucial for ensuring their safety [10]. Furthermore, existing datasets like InfAnFace [36], InfActPrimitive [14], NNS clinical in-crib dataset [44], and SyRIP infant pose dataset [18] make valuable contributions but lack specific annotations for hazard detection, missing critical

details about potential risks in an infant's surroundings.

Addressing these identified research gaps, our study introduces CribNet and the CribHD dataset, both aimed at hazard detection in crib environments. Unlike previous research focused on infant behavior and pose, our approach zeroes in on the less-explored domain of in-crib hazard identification. CribNet combines advanced detection, segmentation, and infant-specific pose estimation techniques. Meanwhile, the CribHD dataset is tailored to highlight hazardous objects and scenarios near infants. Together, they offer both a novel dataset and a methodological leap forward in the realm of infant safety, significantly enhancing the capabilities for hazard detection in these critical settings.

III. Methods

In this section, we present our framework, CribNet, for hazard estimation in the crib, evaluated on our novel dataset called Crib Hazard detection (CribHD). We first begin by delving into the detail of CribNet by explaining the entire process from receiving an image as input to final hazard estimation. We then introduce CribHD as data backbone for training and testing our framework.

A. CribNet: A framework for In-Crib Hazard Detection

Detecting hazards related to blankets and toys are two key component of CribNet as shown in Fig. 1. The following sections provide a detailed description of these two components.

*1) **Toy Hazard Detection in CribNet**:* We build an algorithmic framework for toy hazard analysis in crib environments, leveraging the CribHD-T dataset. It is illustrated in Fig. 1(a). The framework aims to detect any toys presenting a suffocation hazard owing to their diminutive size and closeness to the infant's mouth, as well as to pinpoint hard toys situated in proximity that could potentially lead to accidental injuries.

The detailed process is presented in Algorithm 1. In the initial phase, we separate toys from the background with the segmentation model $\mathcal{F}_{\text{SEG}}$, a refined Detectron2 model [40]. The resulting output S_{ALL} identifies the pixel coordinates for each toy's segmentation and corresponding textural characteristics: hard and soft. Concurrently, a Mediapipe model [26] F_{MTH} is employed to identify the facial region in I and to extract the mouth area, using the predicted facial landmark adhering to the established Multi-PIE layout [13]. The segmented mouth area's pixel coordinates are encapsulated in the output S_{MTH}.

With these two classifications, together with the hard-soft classification obtained from the earlier toy segmentation $\mathcal{F}_{\text{SEG}}$, we can apply a simple decision tree to determine the potential hazard class for each toy. Initially, toys are evaluated based on a proximity threshold T, established by specific criteria, to determine their closeness to the mouth using $\mathcal{F}_{\text{PRX}}$. Toys deemed distant from the mouth are labeled as *non-hazards*. Those near the mouth are considered potential hazards S_{HZD} and are further categorized into hard and soft types. Hard toys near the mouth are identified as *injury hazards* (S_{INJ}). The remaining soft toys near the mouth are assessed using $\mathcal{F}_{\text{FLT}}$, with those larger than the mouth size marked as *suffocation hazards* (S_{SUF}), and those softer as *choking hazards* (S_{CHK}). Consequently, each toy is classified as either a non-hazard, an injury hazard, a suffocation hazard, or a choking hazard.

Algorithm 1 Toy Hazard Detection

Require:
I: Input image with toys and infant's face.
T: Min safe distance between toy and infant's mouth.
Ensure:
S_{HZD}: Results of hazardous toy detection.

function $\mathcal{F}_{\text{SEG}}(I)$
  $S_{\text{ALL}} \leftarrow$ Segment toys in I via Detectron2.
  return S_{ALL}
end function

function $\mathcal{F}_{\text{MTH}}(I)$
  $S_{\text{MTH}} \leftarrow$ Segment mouth in I via Mediapipe.
  return S_{MTH}
end function

function $\mathcal{F}_{\text{FLT}}(S_{\text{ALL}}, S_{\text{MTH}})$
  $S_{\text{FLT}} \leftarrow \{s \in S_{\text{ALL}} : \text{size}(s) < \text{size}(S_{\text{MTH}})\}$
  return S_{FLT}
end function

function $\mathcal{F}_{\text{PRX}}(s, S_{\text{MTH}})$
  $P \leftarrow \text{dist}(S_{s,\text{MTH}})$
  return P
end function

$S_{\text{FLT}} \leftarrow \mathcal{F}_{\text{FLT}}(\mathcal{F}_{\text{SEG}}(I), \mathcal{F}_{\text{MTH}}(I))$
for $\{s \| s \in S_{\text{ALL}} : \mathcal{F}_{\text{PRX}}(s, S_{\text{MTH}}) < T\}$ **do**
  if $s \in S_{\text{FLT}}$ **then**
    $S_{\text{CHK}} \leftarrow s$
  else if s[TEX] is Soft **then**
    $S_{\text{SUF}} \leftarrow s$
  else
    $S_{\text{INJ}} \leftarrow s$
  end if
end for
$S_{\text{HZD}} \leftarrow \{S_{\text{CHK}}, S_{\text{SUF}}, S_{\text{INJ}}\}$

Employing the extensive CribHD-T subset, our toy hazard detection framework effectively monitors infant crib spaces. This framework's ability to provide accurate, real-time alerts about toy-related dangers marks a significant leap in proactive infant safety. By harnessing data-driven insights, this system enhances traditional safety protocols, substantially improving infant care.

*2) **Blanket Occlusion Detection in CribNet**:* Next, we turn to the detection of hazards caused by blankets, including the suffocation risk of a blanket covering the infant's face, and limb entanglement. Our approach is straightforward but

Algorithm 2 Blanket Occlusion Detection

Require:
I: Input image with blanket and infant.
Ensure:
K_{OCC}: Occlusion status vector for body joint keypoints.

function $\mathcal{F}_{\text{BLK}}(I)$
  $M_{\text{BLK}} \leftarrow$ Blanket area mask in I via Mask RCNN.
  return M_{BLK}
end function

function $\mathcal{F}_{\text{POS}}(I)$
  $K \leftarrow$ Infant pose estimation via FiDIP.
  return K
end function

function $\mathcal{F}_{\text{OCC}}(k,\, M_{BLK})$
  for $m \in M_{BLK}$ **do**
    if $k \in m$ **then**
      return True
    end if
  end for
  return False
end function

Initialize $K_{OCC} \leftarrow \vec{0}$
for $k \in \mathcal{F}_{\text{POS}}(\text{I})$ **do**
  if $\mathcal{F}_{\text{OCC}}(k, \mathcal{F}_{\text{BLK}}(I))$ **then**
    $K_{OCC}[k] \leftarrow 1$
  end if
end for

addresses an aspect of crib safety often overlooked in infant monitoring.

Our framework, as depicted in Fig. 1(b) and Algorithm 2, consists of two components. First, we develop a blanket segmentation model $\mathcal{F}_{\text{BLK}}$ by fine-tuning a pre-trained Detectron2 segmentation model using images from CribHD-T equipped with blanket segmentation labels. This yields a blanket segmentation mask M_{BLK} for each image. Second, we use a pose estimation model, denoted $\mathcal{F}_{\text{POS}}$, drawn from the Fine-tuned Domain-adapted Infant Pose Model (FiDIP) [18] developed specifically for precise and robust infant pose estimation. This produces a set of keypoint locations $\mathcal{K}$ for each image. Finally, we assess which infant body parts, as represented by their keypoints in $\mathcal{K}$, are and are not occluded by the blanket mask M_{BLK}, and output the classification list with the function $\mathcal{F}_{\text{OCC}}$. This process assumes a top-down camera view and a supine pose, which can be monitored separately with pose estimation.

B. The Crib Hazard Detection (CribHD) Dataset

We present the crib hazard detection (CribHD) dataset, designed to fill a gap in existing infant-centric datasets by focusing on the identification of crib-based hazards, with rich detection and segmentation annotations for toys and blankets. See Fig. 2 for image and annotation samples. CribHD is sourced from Google Images and supplemented with real-world images produced by our lab, to ensure diversity and realism. All data are meticulously annotated by bounding boxes for object detection and object masks for object segmentation, using Roboflow software [4].S The dataset has three subsets: CribHD-T for toy analysis, CribHD-B for blanket examination, and CribHD-C, which features simulated hazardous crib settings.

The CribHD-T subset, as illustrated in Fig. 2(a), consists of 1000 images featuring annotated toys, split evenly between soft and hard varieties. This subset encompasses a diverse range of toy-related scenarios, including 500 images with backgrounds removed and 500 depicting toys against complex real-world backdrops. 700 feature isolated toys, while 300 show toys held by hand. The collection also includes a variety of shapes, with 500 round toys and 500 polygonal ones, and types, such as 500 soft or plushy toys and 500 hard plastic toys. It is important to note that the labels in different aspects are independent.

The CribHD-B subset, shown in Fig. 2(b), comprises 500 images specifically tailored for blanket segmentation, showcasing blankets in a variety of contexts, including 75 images with just only blankets, 380 images covering infants, 45 draped over adults, in folded arrangements, dispersed in room settings, and displaying an array of colors and textures. Additionally, this dataset incorporates images depicting blankets in outdoor environments, crumpled blankets, and blankets adorned with complex patterns.

The Crib Hazard Detection Challenge Dataset (CribHD-C), shown in Fig. 2(c), is meticulously crafted using a realistic doll, toys, and blankets to simulate hazardous crib environments. We arranged various scenarios with the doll in different poses such as lying on its back, chest, or side, amidst a diverse array of toys and blankets. These setups vary in texture, shape, size, and position, with blankets of different colors and patterns, creating scenarios where different body parts are occluded. A total of 120 images were collected for the CribHD-C dataset, specifically to assess the performance of our CribNet hazard detection framework.

IV. Experimental Results

This section presents a detailed evaluation of the CribHD dataset employing the CribNet hazard detection framework, integrating advanced detection and segmentation methods such as YOLOv8 [33], Detectron2 [40], and YOLACT [2]. The effectiveness of both the toy hazard detection and blanket occlusion frameworks are examined individually.

A. Toy Hazard Detection Framework Evaluation

In the toy hazard detection framework, F_{SEG} is key for toy detection and segmentation, adapted for three models. A thorough fine-tuning process, using the CribHD-T subset, involved freezing pre-trained layers and retraining the final layer for 200 epochs to ensure consistent and fair model performance evaluation. The CribHD-T subset was split into 80% training and 20% testing sets to avoid overfitting and

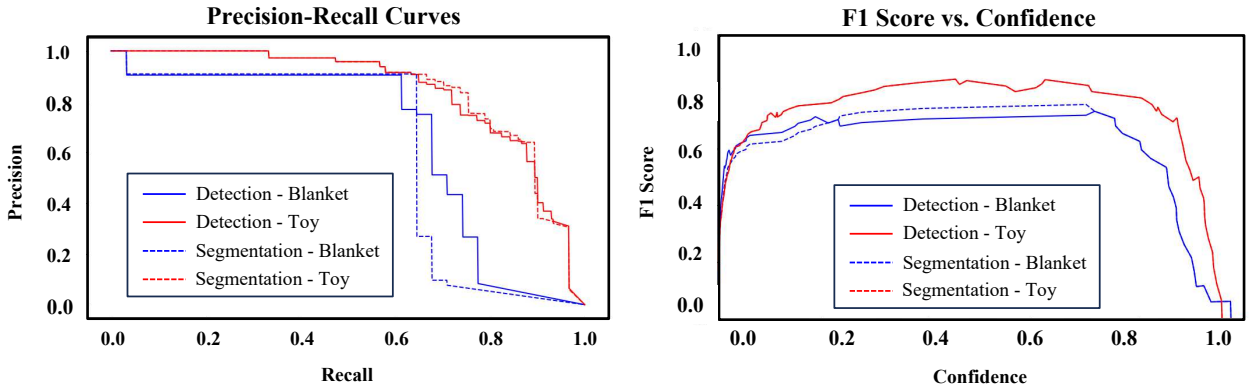

Fig. 3: Precision–recall and F1 Score vs. confidence curves for the YOLOv8 Model. The left graph displays the precision–recall curves, where 'Detection' (solid lines) represents bounding box accuracy and 'Segmentation' (dashed lines) reflects mask-based predictions for blanket and toy objects. The right graph shows the F1 score vs. confidence, illustrating the balance between precision and recall at different thresholds. Blue and red colors distinguish between blanket and toy detections. Please note that the F1 versus confidence curve for toy detection and segmentation shows closely matched performance, resulting in overlapping curves on the graph. This figure collectively highlights the YOLOv8 model's effectiveness in object detection and segmentation tasks.

TABLE I: Performance of object detection models trained on the CribHD dataset in mean average precision (under IoU thresholds of 50%, 70%, and 90%). The bolded numbers represent the highest performance achieved for the specified metric.

	Toy Detection			Blanket Detection		
Model	mAP_{50}	mAP_{70}	mAP_{90}	mAP_{50}	mAP_{70}	mAP_{90}
YOLOv8 [33]	**86.1**	**85.5**	**83.4**	**88.3**	**84.7**	**78.6**
Detectron2 [40]	78.2	77.7	75.0	65.4	58.3	56.5
YOLACT [2]	81.9	78.7	68.5	31.5	27.6	26.8
Avg.	82.1	80.6	75.6	61.7	56.9	54.0

TABLE II: Performance of segmentation models trained on the CribHD dataset in mean average precision (under IoU thresholds of 50%, 70%, and 90%). The bolded numbers represent the highest performance achieved for the specified metric.

	Toy Segmentation			Blanket Segmentation		
Model	mAP_{50}	mAP_{70}	mAP_{90}	mAP_{50}	mAP_{70}	mAP_{90}
YOLOv8 [33]	**88.3**	64.7	68.6	68.4	64.3	61.2
Detectron2 [40]	80.4	**77.5**	**74.7**	**70.4**	**66.2**	**63.1**
YOLACT [2]	77.3	71.4	62.5	30.7	29.8	25.7
Avg.	82.0	71.2	68.6	56.5	53.4	50.0

maintain result reliability. Performance was measured in mean average precision (mAP) across various IoU thresholds (details in Table I and Table II).

YOLOv8 led in toy detection with an 86.1% mAP at 50% IoU, showing high accuracy in toy localization. YOLACT followed with 81.9% mAP, demonstrating robustness. Detectron2 had a strong performance (78.2% mAP), as indicated in Fig. 3, showing YOLOv8's precision across recall levels. What needs to be noted is that the F1-versus-confidence curve for toy detection and segmentation shows very similar performance, resulting in the curves appearing overlapped on the graph. This consistency suggests reliability in different detection scenarios, crucial for infant hazard detection.

In segmentation, YOLOv8 again excelled (88.3% mAP), effectively delineating toy shapes and sizes. Detectron2 and YOLACT, with slightly lower mAPs, still showed balanced segmentation capabilities. The Fig. 3 F1 Score vs Confidence plot revealed YOLOv8's optimal confidence level for peak segmentation, enhancing its application in infant safety.

Overall, the results indicate the models' proficiency in handling toy detection and segmentation in crib environments, a key factor in infant safety monitoring systems.

B. Blanket Occlusion Detection framework Evaluation

For the blanket occlusion detection framework, the fine-tuning approach applied to YOLOv8, Detectron2, and YOLACT using the CribHD-B subset is the same as the toy detection. The evaluation of blanket detection and segmentation are shown in Table I and Table II. YOLOv8 stands out for its high performance, achieving an mAP of 88.3% in detection and 68.4% in segmentation at a 50% IoU threshold, as depicted also in the Precision-Recall segment of Fig. 3. This demonstrates its superior capability to not only detect but also accurately segment blanket positions in various crib scenarios, a crucial aspect for assessing potential suffocation risks. Detectron2 follows closely, showing stable and effective results in both detection and segmentation, indicating its robustness in handling diverse blanket scenarios within crib environments, including different textures and positions. YOLACT shows proficiency in blanket detection (31.5% mAP) but struggles in segmentation (30.7% mAP), indicating difficulties in accurately defining blanket edges, especially in heavily occluded scenes. The accompanying plot underscores these challenges, revealing YOLACT's varying performance across different levels of occlusion and providing a visual comparison of its capability in handling complex scenarios with partial or full blanket coverage.

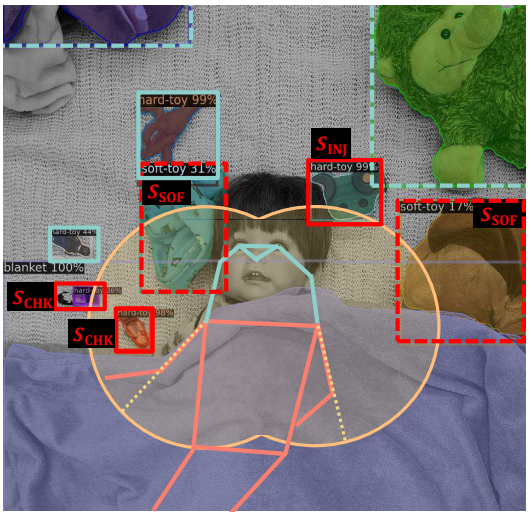

Fig. 4: This figure provides an illustrative depiction of CribNet's analytical capabilities when applied to a simulated hazard scene from the CribHD-C dataset. The blanket occlusion detection framework effectively delineates the blanket coverage, demarcating the occluded infant joints and skeletons in light coral and non-occluded ones in cyan. The arm length, obtained from the predicted infant pose, establishes a proximity threshold T, as illustrated by the light orange lines and region. In the toy hazard detection framework, the threshold is pivotal in pinpointing potentially dangerous toys. Toys are classified as hard (enclosed in solid boxes) or soft (enclosed in dashed boxes). Those intersecting with the specified proximity zone are highlighted in red, indicating potential risks: choking S_{CHK} from small hard toys, injury S_{INJ} from large hard toys, and suffocation S_{SUF} from soft toys. Toys outside this zone are highlighted in cyan, signifying they are safe.

These observations elucidate each model's competencies and potential enhancement points in blanket detection and segmentation, which are crucial for infant safety. The performances of YOLOv8 and Detectron2 suggest their promising applicability in practical crib environments, whereas YOLACT's segmentation struggles highlight the necessity for advanced development to better manage intricate occlusions. Collectively, the analysis accentuates the adaptability and relevance of the CribHD dataset and the CribNet framework in advancing the proactive identification of hazards to ensure infant well-being.

C. Simulated In-Crib Hazard Scene Evaluation

In the evaluation of CribNet on the CribHD-C dataset, which simulates in-crib hazards, the results are promising and indicative of the model's efficacy. This figure in Fig. 4 presents a visual representation of CribNet's analytic proficiency when evaluated against a simulated hazard scenario within the CribHD-C dataset. In this case, we set proximity threshold T equal to the arm length obtained from the predicted infant pose. We tested all images in the CribHD-C dataset and the CribNet demonstrated exceptional precision in toy detection, correctly identifying over 80% of the hazardous toys present across all 120 frames. This high degree of accuracy, with only a marginal fraction of toys being incorrectly predicted or undetected, attests to the model's robustness and the dataset's comprehensive nature. For occlusion prediction, CribNet employs the FiDIP infant pose estimation model to ascertain occluded joints, accomplishing this by superimposing identified joints onto the blanket segmentation mask. Given the exceptional accuracy of our blanket segmentation model, accurate pose estimation typically translates to precise occlusion status determination for each joint. The high robustness of the FiDIP model against occlusion effects allows CribNet to achieve notable accuracy in identifying joint occlusions, with only two images showing exceptions, occurring in cases of extensive body coverage where over 90% of the infant's body and face are nearly indiscernible, resulting in prediction errors. These results highlight the proposed CribNet's effectiveness in most obscured conditions and suggest potential areas for enhancement in handling scenarios with greater occlusion.

Overall, the CribNet's performance, with high accuracy in toy detection and occlusion prediction, underscores the dataset's generalizability and utility in proactive infant hazard detection. These results confirm the capability of CribNet to effectively discern in-crib hazards, validating both the dataset and the method as reliable tools for enhancing infant safety.

V. CONCLUSIONS AND FUTURE WORKS

This study marks a pivotal advancement in in-crib hazard detection, specifically targeting the safety issues posed by toys and blankets. We introduce the innovative CribNet framework for hazard detection in cribs, coupled with the specialized CribHD dataset for identifying potential dangers. Together, these tools demonstrate our commitment to improving infant safety. Looking ahead, the field is set for dynamic progress. Future enhancements in object detection, diversifying the CribHD dataset, and real-world application of these models promise to revolutionize infant care, harnessing advanced computer vision and machine learning for enhanced child safety.

REFERENCES

[1] S. Amraee, B. Galoaa, M. Goodwin, E. Hatamimajoumerd, and S. Ostadabbas. Multiple toddler tracking in indoor videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 11–20, 2024.

[2] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[3] H.-g. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20186–20196, 2022.

[4] F. Ciaglia, F. S. Zuppichini, P. Guerrie, M. McQuade, and J. Solawetz. Roboflow 100: A rich, multi-domain object detection benchmark. *arXiv preprint arXiv:2211.13523*, 2022.

[5] S. de Visme, D. A. Korevaar, C. Gras-Le Guen, A. Flamant, M. Bevacqua, A. Stanzelova, N. T. Trinh, D.-A. Ciobanu, A. A. Carvalho, I. Kyriakoglou, et al. Inconsistency between pictures on baby diaper packaging in europe and safe infant sleep recommendations. *The Journal of Pediatrics*, 264:113763, 2024.

[6] J. Drowos, A. Fils, M. C. Mejia de Grubb, J. L. Salemi, R. J. Zoorob, C. H. Hennekens, and R. S. Levine. Accidental infant suffocation and strangulation in bed: disparities and opportunities. *Maternal and child health journal*, 23:1670–1678, 2019.

[7] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022.

[8] D. M. Ely and A. K. Driscoll. Infant mortality in the united states: Provisional data from the 2022 period linked birth/infant death file. *National Center for Health Statistics. Vital Statistics Rapid Release*, (33), 2023.

[9] D. Gálvez-López and J. D. Tardos. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012.

[10] C. E. Gaw, T. Chounthirath, J. Midgett, K. Quinlan, and G. A. Smith. Types of objects in the sleep environment associated with infant suffocation and strangulation. *Academic pediatrics*, 17(8):893–901, 2017.

[11] R. Girshick. Fast r-cnn. *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[12] M. H. Goodstein, E. Lagon, T. Bell, B. L. Joyner, and R. Y. Moon. Stock photographs do not comply with infant safe sleep guidelines. *Clinical pediatrics*, 57(4):403–409, 2018.

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010.

[14] E. Hatamimajoumerd, P. D. Kakhaki, X. Huang, L. Luan, S. Amraee, and S. Ostadabbas. Challenges in video-based infant action recognition: A critical examination of the state of the art. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 21–30, 2024.

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[16] A. H. Hirai, K. Kortsmit, L. Kaplan, E. Reiney, L. Warner, S. E. Parks, M. Perkins, M. Koso-Thomas, D. V. D'Angelo, and C. K. Shapiro-Mendoza. Prevalence and factors associated with safe infant sleep practices. *Pediatrics*, 144(5), 2019.

[17] X. Huang, N. Fu, S. Liu, and S. Ostadabbas. Invariant representation learning for infant pose estimation with small data. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2021*, December 2021.

[18] X. Huang, N. Fu, S. Liu, and S. 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.

[19] X. Huang, L. Luan, E. Hatamimajoumerd, M. Wan, P. D. Kakhaki, R. Obeid, and S. Ostadabbas. Posture-based infant action recognition in the wild with very limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4911–4920, 2023.

[20] X. Huang, M. Wan, L. Luan, B. Tunik, and S. Ostadabbas. Computer vision to the rescue: Infant postural symmetry estimation from incongruent annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1909–1917, 2023.

[21] I. Jeelani, K. Asadi, H. Ramshankar, K. Han, and A. Albert. Real-time vision-based worker localization & hazard detection for construction. *Automation in Construction*, 121:103448, 2021.

[22] G. Jocher. Ultralytics yolov5, 2020.

[23] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2(3):18, 2017.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll'ar, C. L. Zitnick, et al. Microsoft coco: Common objects in context. *European conference on computer vision*, pages 740–755, 2014.

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[26] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[27] S. K. R. Manne, S. Zhu, S. Ostadabbas, and M. Wan. Automatic infant respiration estimation from video: A deep flow-based algorithm and a novel public benchmark. In *International Workshop on Preterm, Perinatal and Paediatric Image Analysis*, pages 111–120. Springer, 2023.

[28] R. Y. Moon, R. F. Carlin, I. Hand, T. F. on Sudden Infant Death Syndrome, et al. Sleep-related infant deaths: updated 2022 recommendations for reducing infant deaths in the sleep environment. *Pediatrics*, 150(1), 2022.

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

[30] C. Pena-Caballero, D. Kim, A. Gonzalez, O. Castellanos, A. Cantu, and J. Ho. Real-time road hazard information system. *Infrastructures*, 5(9):75, 2020.

[31] T. Powers, E. Hatamimajoumerd, W. Chu, V. Rajendran, R. Shah, F. Diabour, M. Vaillant, R. Fletcher, and S. Ostadabbas. Vision-based treatment localization with limited data: Automated documentation of military emergency medical procedures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1819–1828, 2023.

[32] P. V. Ramos, P. J. Hoogerwerf, P. K. Smith, C. Finley, U. E. Okoro, and C. A. Jennissen. Pre-and postnatal safe sleep knowledge and planned as compared to actual infant sleep practices. *Injury epidemiology*, 10(Suppl 1):55, 2023.

[33] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[34] M. Sanjai Siddharthan, S. Aravind, and S. Sountharrajan. Real-time road hazard classification using object detection with deep learning. In *International Conference on IoT Based Control Networks and Intelligent Systems*, pages 479–492. Springer, 2023.

[35] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019.

[36] M. Wan, S. Zhu, L. Luan, G. Prateek, X. Huang, R. Schwartz-Mette, M. Hayes, E. Zimmerman, and S. Ostadabbas. Infanface: Bridging the infant–adult domain gap in facial landmark estimation in the wild. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4486–4492. IEEE, 2022.

[37] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023.

[38] J. Wang, L. Song, Z. Li, H. Sun, J. Sun, and N. Zheng. End-to-end object detection with fully convolutional network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15849–15858, 2021.

[39] T. C. S. Ward. " things changed very quickly": Maternal intentions and decision-making about infant sleep surface, location, and position. *Birth (Berkeley, Calif.)*.

[40] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[41] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

[42] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, and W. Zuo. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE transactions on cybernetics*, 52(8):8574–8586, 2021.

[43] S. Zhu, M. Wan, E. Hatamimajoumerd, K. Jain, S. Zlota, C. V. Kamath, C. B. Rowan, E. C. Grace, M. S. Goodwin, M. J. Hayes, et al. A video-based end-to-end pipeline for non-nutritive sucking action recognition and segmentation in young infants. *arXiv preprint arXiv:2303.16867*, 2023.

[44] S. Zhu, M. Wan, S. K. R. Manne, E. Zimmerman, and S. Ostadabbas. Subtle signals: Video-based detection of infant non-nutritive sucking as a neurodevelopmental cue. *arXiv preprint arXiv:2310.16138*, 2023.

[45] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 2023.