FiDaSS: A Novel Dataset for Firearm Threat Detection in Real-World Scenes

Murilo S. Regio and Isabel H. Manssour
Pontifical Catholic University of Rio Grande do Sul, PUCRS, School of Technology, Porto Alegre, RS, Brazil

Keywords: Surveillance, CCTV, Firearm Detection, Armed Person Detection.

Abstract: For a society to thrive, people must feel safe; otherwise, fear and stress reduce the quality of life. A variety of security measures are used, but as populations grow and firearms become more accessible, societal safety faces new challenges. Existing works on threat detection focus primarily on security cameras but lack common benchmarks, standard datasets, or consistent constraints, making it difficult to assess their real-world performance, especially with low-quality footage. This work introduces a challenging dataset for Firearm Threat Detection, comprising 7450 annotated frames across 291 videos, created under rigorous quality controls. We also developed tools to streamline dataset creation and expansion through semi-automatic annotations. To our knowledge, this is the largest real-world dataset with frame-level annotations in the area. Our dataset is available online alongside the tools developed, including some to facilitate its extension. We evaluated popular detectors and state-of-the-art transformer-based methods on the dataset to validate its difficulty.

1 INTRODUCTION

Security has always been a major concern, and as firearms become increasingly accessible, societal safety grows more fragile (Hurka and Knill, 2020). Firearms allow individuals, even without advanced training, to cause significant harm in public spaces, leading to tragedies such as school and mass shootings (Gius, 2018; Lemieux, 2014). Several measures exist to handle these situations, the most common being monitoring environments using security cameras.

While security cameras offer advantages (Piza et al., 2019), such as recording events for posteriority, they rely heavily on human supervision. Cameras only capture footage, they require operators to actively monitor and respond to incidents. In larger areas or buildings, multiple cameras must be monitored simultaneously, increasing the risk of distractions and human error (Darker et al., 2007).

Effective vigilance requires sustained concentration (Donald and Donald, 2015), yet CCTV operators often struggle to maintain attention over time. Studies show that focus declines significantly after 20 minutes (Velastin et al., 2006), with operators missing 45% of scene elements by 12 minutes and up to 95% after 22 minutes (Ainsworth, 2002). This highlights the limitations of traditional monitoring and the need for solutions to enhance security.

Many studies have tackled firearm detection, often prioritizing model performance over the data used (Gelana and Yadav, 2019; de Azevedo Kanehisa and de Almeida Neto, 2019). Some focus on specific tasks, such as concealed weapon detection (Raturi et al., 2019; Ineneji and Kusaf, 2019), while others address broader security issues, including abandoned luggage (Loganathan et al., 2019), fire (Mehta et al., 2020), or general violence (Pawar et al., 2019).

Despite ongoing concerns about dataset quality, these issues are frequently left for future work (Olmos et al., 2018; Lim et al., 2019). Few authors propose new datasets, and even fewer provide real-world data with detailed object detection annotations. To address this gap, we developed a novel and flexible dataset for firearm threat detection, created methodically from real-world scenes using rigorous selection and annotation processes.

The contributions of this work are threefold:

- A novel challenging dataset called FiDaSS (Firearm Dataset for Smart Surveillance) with 7450 real-world annotated images featuring diverse scenarios, cultural contexts, and detailed annotations for victims, perpetrators, and weapons.
- Tools to streamline dataset creation or expansion by using pre-existing detectors to estimate annotations, which can be manually refined afterward.

Regio, M. S. and Manssour, I. H.
FiDaSS: A Novel Dataset for Firearm Threat Detection in Real-World Scenes.
DOI: 10.5220/0013177800003912
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 683-690
ISBN: 978-989-758-728-3; ISSN: 2184-4321
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

Table 1: Most frequently used datasets in the studied literature, shown in descending order of popularity.

Dataset	Type of Data	Amount of Data	Annotations	Task	Frame Dimension	Year
(Olmos et al., 2018)	Movie	3,000 Frames	Frame Level	Detection	Varied	2018
(IMFDB, 2015)	Movie	396,808 Frames	Frame Level	Classification	Varied	2015
(Sultani et al., 2018)	Real-world	200 Videos	Video Level	Classification	320x240	2018
(Grega et al., 2013)	Acted	7 Videos	Video Level	Classification	640x480	2013
(González et al., 2020)	Synthetic	4000 Frames	Frame Level	Detection	1920x1080	2020
(Gu et al., 2022)	Acted	5000 Frames	Frame Level	Detection	Varied	2022
(Hnoohom et al., 2022)	Acted	8319 Frames	Frame Level	Detection	1920x1080	2022
FiDaSS	Real-world	7450 Frames	Frame Level	Detection	Varied	N/A

- Experiments using state-of-the-art networks to evaluate the quality and difficulty of FiDaSS.

2 RELATED WORK

Through a literature study, we identified 34 datasets used or proposed, and in Table 1 we provide a comparison between the most popular. The datasets found can be roughly categorized based on the data they use. This insight is crucial for quickly filtering undesirable datasets and focusing on those that are adequate to our objectives. The categories identified are as follows:

- **Movie Data:** Datasets based on movies are abundant and offer plenty of data, but have a tendency for lower real-world performance due to cinematic characteristics. E.g., (IMFDB, 2015).
- **Enacted Data:** Simulated real-life scenarios offer better realism but are smaller due to the high effort required for creation. E.g., (Grega et al., 2013).
- **Real Data:** Surveillance footage datasets are rare, small, and often subsets of broader datasets, despite being the most representative of real-world scenarios. E.g., (Sultani et al., 2018).

Although most works focused on CCTV scenarios, only three public datasets are based on real-world data. Movie-based datasets were the most prevalent, with the two most used datasets falling into this category. This mismatch highlights a reliance on inadequate data for real-world applications, likely due to the greater availability and size of movie datasets. Furthermore, when analyzing the datasets shown in Table 1, we notice it is difficult to compare the methods in the area fairly, as works use diverging datasets that focus on different categories. Similarly, there is no standard measure for comparing methods, even considering within the same category.

We identified that the biggest concern in the area is the construction of representative datasets. Some works (Sultani et al., 2018; Lim et al., 2019) stand out for presenting data from actual events captured by security cameras and made available to the public and contain exciting data. However, they lack in amount, diversity, and are composed of a set of videos or contiguous frames marked as containing or not the object of interest instead of precise annotations. Considering this, we created a dataset aiming to address the limitations identified and provide a robust foundation for future research, thus fulfilling the following gaps: **(I)** Real-world data to encourage practical applications; **(II)** High variability in sources, video quality, and cultural representation; **(III)** Frame-level annotations, adaptable for tasks like object detection and scene classification.

3 DATASET DESCRIPTION

We started with a literature review to identify commonly used datasets, their characteristics, and areas for improvement (Section 2). Based on that, we decided that our primary objective for FiDaSS was to portray a diversity of cultures using real-world scenes, thus minimizing regional social biases.

To facilitate FiDaSS creation, we implemented a set of tools to provide useful scripts for manipulating videos, creating and labeling clips, creating bounding box annotations, and generating statistics. We also integrated our annotation pipeline with an open-source video object tracking frameworkto provide suggestions for future annotations.

The following sections describe each step presented in Fig. 1. FiDaSS, the tools developed to create it, and complementary details about it (e.g., dataset splits, geographical diversity, and training configurations) are available online[1].

3.1 Data Collection

To create FiDaSS, we explored a wide range of data sources to assess the existing resources in the literature and identify gaps. We began by analyzing

[1] https://github.com/fidass/fidass_dataset

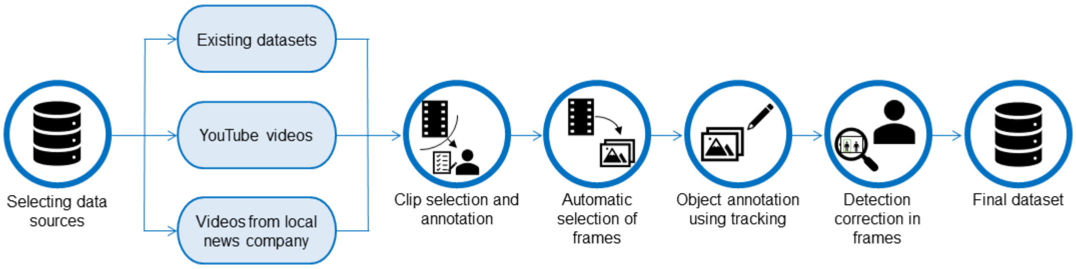
Selecting data sources
Existing datasets
YouTube videos
Videos from local news company
Clip selection and annotation
Automatic selection of frames
Object annotation using tracking
Detection correction in frames
Final dataset

Figure 1: Steps followed to create our dataset.

the well-known datasets listed in Table 1, along with lesser-known ones, to establish a strong foundation for our dataset. Building on this, we sought videos from YouTube and a local news company to enrich FiDaSS with diverse scenarios, situations, and cultures. However, obtaining a substantial amount of varied real-world data remains a significant challenge due to privacy concerns and the limited availability of recordings held by security companies.

From our analysis of existing datasets, we selected UCF Crime as a foundation due to its focus on real-world security camera footage. We selected videos containing moments that clearly displayed a held weapon and included a visible criminal, thus excluding scenarios solely involving law enforcement. Then, to expand beyond the literature, we collected unexplored data from YouTube. We collected an initial pool of videos using a query-based search with the keywords *[surveillance video armed robbery, CCTV assaults, guns in CCTV, assault caught on camera]* in over 20 languages. After filtering based on our criteria (weapon visibility and the presence of a criminal) and removing duplicates, this yielded 139 videos. Furthermore, we used YouTube's recommendation system to discover more content, and by applying the same filtering process, we gathered 162 new videos.

To ensure there were no duplicate videos, we conducted a manual verification to remove all overlapping data from the selection. Thus, ultimately, we selected 301 videos from YouTube depicting crime scenes from different countries and cultures. The playlists with these videos are available online[2].

Finally, to expand our dataset further, we contacted a local news company, requesting access to some videos provided to them depicting recent crime scenes from the region. Upon receiving their approval, we obtained 13 novel videos.

3.2 Dataset Annotation

After collecting the videos, we began annotating clips, while ensuring each was self-contained. For that, each clip had to feature the assailant for at least five seconds, ensuring sufficient relevant information to contribute. Through this process, we reduced 18 hours of video into 2.8 hours of manually selected clips, with each containing either unique footage or a different camera angle. To standardize the dataset, we converted all clips to a frame-rate of 1 fps.

After delimiting each clip, we started annotating each clip on a frame-level for object detection. Our first step was to annotate only the first time each person appeared in each clip. In addition to armed people, we also included their guns and unarmed people in the annotations. This way, models could learn more reliably the difference between armed people and people holding items similar to guns in low-quality videos (such as phones and umbrellas).

Next, we processed all clips and their unique objects using a network designed for object tracking. This approach generated an initial approximation of annotations for every clip with minimal manual effort. To ensure quality, we meticulously reviewed each frame and corrected mismarked instances. This procedure significantly reduced our workload while expediting the creation process of our dataset.

In many recordings, we noticed that the assailant would, for example, stand perfectly still while making demands for $15-30$ seconds, and the people involved were paralyzed listening to their threats during most of this time. These scenarios would cause several clips to be composed of nearly identical images with no substantial changes that would provide new information. Thus, while adjusting the miss-detected bounding boxes, we discarded these long redundant sequences to avoid having inflated results in our experiments, as having the models predict identical frames would make them seem more accurate than they actually were. By doing so, we reduced the total duration of the clips in our dataset to 2.1 hours.

Since we gathered images from several different videos available online, many of them had faces blurred for anonymity. However, as our objective with our dataset is to provide an accurate estimation of a model's performance in real-world scenarios, adding blur to all faces would provide an inaccurate represen-

[2] https://youtube.com/@fidassdataset7285/playlists.

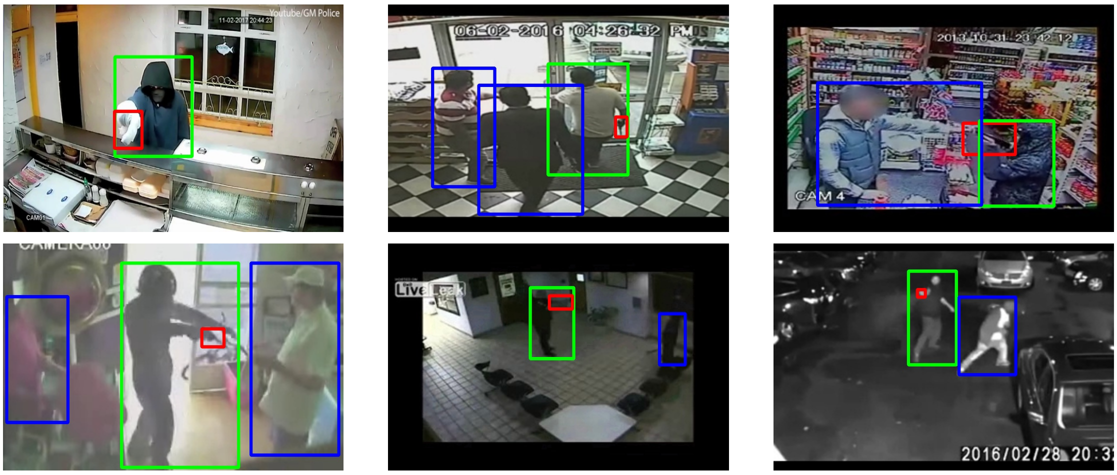

Figure 2: Example of images from our dataset with their corresponding labels, following the color scheme: green for "armed" labels, blue for "unarmed", and red for "firearm".

Table 2: Total data selected from each data source.

Dataset of Origin	Videos	Clips	Frames
Youtube Playlist	301	216	4905
UCF Crime	197	144	2239
News Company	13	19	306
Total	522	379	7450

tation of the data models would find when applied in real scenarios. Therefore, we decided not to add any more blur to our dataset but still use the images we gathered that were already blurred. This way, we hope those images would serve as augmented data during training, hopefully teaching models that the individual's face is not as important as their posture and what they are holding.

The annotation process described involved three contributors, each responsible for annotating a subset of the collected videos. A final group revision was conducted to ensure consistency across all annotations, addressing nuances in low-quality frames, such as determining whether faintly visible individuals in the background should be annotated.

The tools we developed streamlined the methodology described, enabling efficient frame selection, annotations, and subsequent corrections. The resulting dataset comprises a total of 23109 annotated objects across 379 clips (7450 frames), representing approximately two hours of annotated footage. Examples of the dataset's annotations are shown in Fig. 2, with the first row illustrating higher-quality frames and the second row showcasing lower-quality frames that require additional context for proper interpretation.

4 DATASET STATISTICS

This section discusses some properties of FiDaSS while also comparing it to those presented in Table 1 and addressing why the original annotations were insufficient in the datasets we used as a basis for ours.

One crucial characteristic of FiDaSS is that we made the annotations directed toward the task of object detection in real-life scenarios. To the best of our knowledge, considering our literature analysis, there is no dataset presenting those characteristics and containing a substantial number of images. Although the datasets highlighted in Table 1 have a large amount of data, only one presented exclusively real-world data, and none had annotations for the object detection task, only for image or video classification.

Table 2 lists the sources used for FiDaSS, showing the number of videos and selected frames from each. Approximately 40% of FiDaSS derives from existing datasets, but we have rigorously selected and annotated the most relevant frames, which were previously available only as raw videos. The remaining 60% consists of novel data from diverse cultures.

After describing FiDaSS' properties, it is essential to compare it with the datasets identified in Table 1. While Weapons-Detection and IMFDB offer the largest datasets, they primarily consist of movie scenes or context-free images, limiting real-world applicability. The Gun Movies Database provides security camera footage, but comprises only seven laboratory-shot videos. The UCF Crime dataset provides real-world footage, with 150 robbery and 50 shooting videos, but uses clip-level labels instead of frame-level. FiDaSS bridges these gaps with detailed annotations and diverse real-world scenarios.

Table 3: Comparison of the models we explored with our dataset, highlighting the transformer-based models in grayed lines.

Model	Input	Backbone	mAP_{50}	AP_{50}			#Params
				Armed	Unarmed	Firearm	
DAFNe	Frames	ResNet-101	34.00%	50.07%	39.83%	12.11%	5M
Faster-RCNN	Frames	ResNet-50	40.65%	45.69%	48.83%	27.42%	42M
YOLOv10	Frames	CSPDarknet-53	44.30%	56.10%	40.50%	36.10%	24M
DINO	Frames	ResNet-50	60.60%	71.30%	66.09%	44.35%	47M
EVA-02	Frames	FPN-12	**72.07%**	**86.47%**	**75.48%**	**54.27%**	86M
TransVOD	Clips	ResNet-50	45.00%	57.11%	50.62%	27.26%	59M

5 ALGORITHMIC ANALYSIS

FiDaSS introduces a novel, challenging object detection dataset designed for adaptability to various tasks, such as video and frame detection. To evaluate its utility, we conducted experiments using both clip-based context and individual frames. We also performed cross-dataset evaluations to assess model generalization when trained on our dataset. The following sections outline the setup used and discuss the results.

5.1 Experimental Setup

FiDaSS aims to evaluate how accurately models would perform in real surveillance system applications. We tested a range of state-of-the-art architectures to identify their weaknesses in this area.

Our initial experiments employed YOLO (Jocher et al., 2023) and Faster-RCNN (Ren et al., 2015), versatile models effective across diverse tasks. However, these models struggled with our dataset, especially with detecting firearms. To address this, we tested DAFNe (Lang et al., 2021), a specialized architecture for detecting small objects in scenes. We then explored two transformer-based networks, DINO (Zhang et al., 2022) and EVA-02 (Fang et al., 2023), as an alternative to convolutional approaches. Additionally, we evaluated sequence-processing models using TransVOD (Zhou et al., 2022), an enhanced version of DETR (Carion et al., 2020), to analyze performance on short video clips rather than isolated frames. Details of the training configurations are available on the project's GitHub.

5.2 Experimental Results

The results of our experiments are presented on Table 3, including information such as if the experiment focused on individual frames or a clip sequence, the average precision for each class, and the model size. Because of the low quality of the images, when a person has their arms stretched out, the models sometimes detect only the torso, which causes a disparity between the label and the detections. Considering this, we focused our analysis on a 50% IoU threshold, as we infer a person to be correctly detected even if their limbs were not included in the prediction.

From the results gathered on Table 3, we can notice a significant advantage of using transformer models. The first three models barely achieved a mAP of 40%, and a class AP of 50%, while the transformer models achieved a mAP above 60%, with a per-class AP between 70% and 80% for non-firearm classes. Additionally, while the firearm class is considerably lower, peaking at approximately 40%, if we can consistently identify armed individuals, we can infer more easily the presence of firearms on the scene. Thus, we will focus mainly on “Armed” and “Unarmed”.

Recurring errors emerged across all models, with person labeling and firearm detection being the most significant challenges. Models consistently identified people, except in cases of heavy blur or background occlusion. However, they struggled with labeling individuals as “armed” or “unarmed,” often fluctuating between the two across consecutive frames, with a slight bias toward the “armed” class. No model consistently located firearms in the scene due to factors such as image quality, variability in object shape, angle, and distance.

These were the major issues identified that caused the metrics to drop in all models proportionally to how well the model could minimize these mistakes. Because of this, we focus our analysis on the results produced by EVA-02, which had the highest performance, and discuss other issues we identified in more depth while studying its results more thoroughly.

We present correct and incorrect detection samples in Fig. 3, illustrating diverse camera angles from our test set. We can see that even from a distance and with low-quality images, the model consistently located people in the scene, with few recurring exceptions. For instance, individuals with their faces hidden were often mislabeled as “Armed,” leading to errors in cases with innocents with their backs to the camera.

To evaluate how representative our dataset is, we

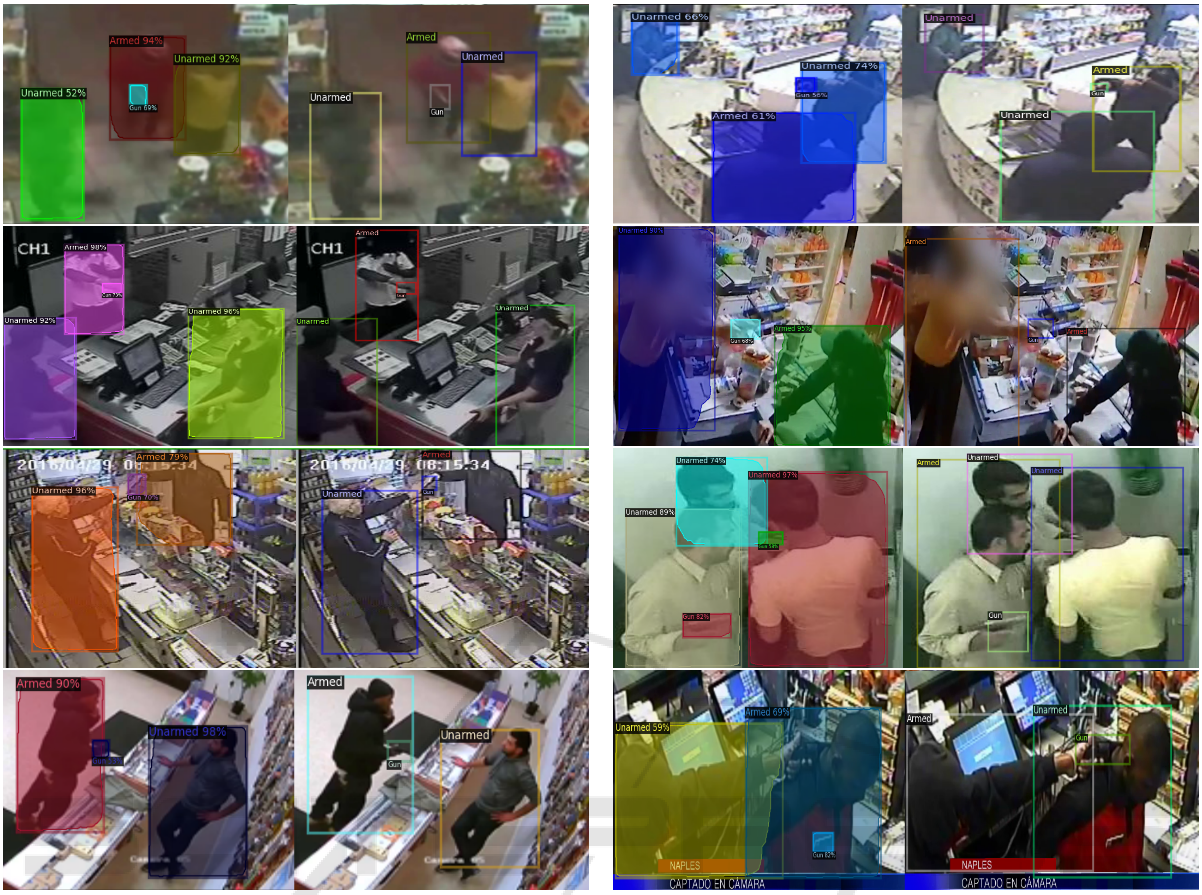

Figure 3: Examples of successful (left) and missed (right) detections by the EVA-02 model. The ground truth for each sample is shown on the right and the model's detection is on the left. The images presented have been zoomed for clarity.

conducted inference of our best-performing model on two datasets from the literature: Sohas-Weapons, which has been frequently used throughout the years, and YouTube-GDD, which is a more recent and very promising dataset. Because of the difference in the datasets' objectives, we decided to focus on a qualitative analysis instead of metrics. In both datasets, we noticed instances of firearms labeled in the scenes but not held by anyone (e.g., in gun stores), which were not detected by the model we trained. We did not consider those as miss-detections, as our objective is to identify firearms being wielded by someone. For further comparisons, examples of our cross-dataset experiments are also available on our GitHub.

Our experiments with Sohas-Weapons showed that, despite its interesting and diverse data (e.g., video game screenshots, stock photos, and selfies), it lacks relevance for real-world applications. Their annotations are also limited, labeling pistols regardless of context (e.g., images of figurines) but excluding people or other types of firearms, which reduces the dataset's applicability and flexibility. EVA-02 struggled most with contextless images or first-person perspectives, which are absent in our training set.

Similarly, experimenting with YouTube-GDD showed that it contains more specialized data, including shooting range videos and demonstrations. Additionally, they provide more complete annotations covering firearms and people, though without distinguishing between armed and unarmed individuals. However, while aligning better with real-world scenarios, it also includes close-ups of firearms or individuals, limiting its practical value. The model we trained performed well in this dataset, mostly just suffering from different firearm annotation policies between their dataset and ours.

Analyzing the results, we can observe that transformer-based networks outperform purely convolutional networks. This is due to attention mechanisms that probably associate global information to classify people in ambiguous scenarios, whereas convolutional networks rely more on the immediate vicinity of the object. Because of this, transformer models can perform more reliably in instances where the presence of a firearm is not clear, either because of its position relative to the camera or because the image quality blurs it out. However, we were surprised that our first experiments with clip sequences performed worse than other transformer strategies. We are unsure why that happened, given that we expected temporal information to help resolve ambiguous scenes even more than spatial information.

6 DISCUSSION, LIMITATIONS, AND FUTURE WORK

FiDaSS is one of the few datasets made entirely with real data from various cultures, to the best of our knowledge, being the only one in this context with annotations for object detection. Moreover, our experiments show that state-of-the-art methods have difficulty with it, primarily with differentiating armed from non-armed people, making it an exciting alternative for future research. Our dataset offers a rich diversity of scenes, capturing different real-world situations from security cameras with varying levels of quality. These diverse scenes, often obfuscated and ambiguous, provide exciting challenges to be tackled by future research. We also observed that only a minority of firearm objects were correctly identified, which we attribute to the high similarity between objects that the camera did not catch well due to their small size and low image quality.

Additionally, by analyzing our results, we identified certain patterns that frequently reappear throughout many videos. The first pattern we identified was cases where the model found the gun but labeled the person holding it as being unarmed. While this is strange for us to observe, we must consider that the model has no “reasoning module” that would associate that those two labels go together. Thus, we believe that this can be addressed in a post-processing module in specialized solutions.

The second important pattern is the fact that the models appeared to associate a person hiding their face with them being armed, which is true in a lot of scenarios. It is not uncommon to find cases of people wearing hoods and masks or bike helmets during robberies, so this association was mostly positive for the model. However, we also identified that, in unclear cases, a person with their back to the camera had the tendency to be labeled as being armed. This introduces several cases of false-positives in our predictions, but taking into consideration the final goal of being usable in real surveillance systems, we consider this a lower priority compared to false-negatives, i.e., cases when an armed person is not detected.

Although we sought diversity, a limitation of FiDaSS is that we still noticed some inherent biases consistently reflected in our results. One issue identified was that, in unclear scenarios where the video quality makes it hard to discern where the firearm is, the models tend to mark a man as “armed”, even if it was a woman who was holding a gun. This characteristic shows us that, even after including comprehensive data from different cultures, our dataset still contains an overwhelming amount of examples of a man holding guns compared to a minority of instances with an armed woman.

For future work, we want to expand FiDaSS further with new and unexplored data from more countries and cultures to introduce even more diversity of images. Moreover, we want to provide more substantial and representative data to avoid social biases. However, we expect that by using the proposed tools, the task of enhancing the dataset becomes easier and more efficient. Finally, we are also interested in exploring more specialized approaches that may achieve better detection results on our task since the results we achieved with general-purpose models were so low.

7 CONCLUSIONS

This work introduced a novel and challenging dataset for firearm threat detection, focusing on object detection in real-world scenarios. Our experiments, both individual frames and video sequences, resulted in very low AP scores, highlighting the dataset’s difficulty. The dataset contains 7450 annotated frames from diverse cultures, environments, and situations and is easily extendable using the provided tools.

We hope our work will help stimulate the research area and provide a challenging dataset that could assist in comparing the performance of works. Finally, besides promoting research in the area, we hope to contribute to security in our everyday lives.

ACKNOWLEDGEMENTS

This research is partially funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. Research supported by HP Brasil Indústria e Comércio de Equipamentos Eletrônicos Ltda. using financial incentives of IPI refund regarding the Law (Law nº 8.248 of 1991). Manssour would like to thank the financial support of the CNPq Scholarship - Brazil (303208/2023-6).

While preparing and revising this manuscript, we used ChatGPT and Grammarly to ensure clarity and

grammatical precision, as English is our second language. The authors are responsible for creating the entire content and ensuring technical accuracy.

REFERENCES

Ainsworth, T. (2002). Buyer beware. *Security Oz*, 19:18–26.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Darker, I., Gale, A., Ward, L., and Blechko, A. (2007). Can cctv reliably detect gun crime? In *2007 41st Annual IEEE International Carnahan Conference on Security Technology*, pages 264–271. IEEE.

de Azevedo Kanehisa, R. F. and de Almeida Neto, A. (2019). Firearm detection using convolutional neural networks. In *ICAART (2)*, pages 707–714.

Donald, F. M. and Donald, C. H. (2015). Task disengagement and implications for vigilance performance in cctv surveillance. *Cognition, Technology & Work*, 17(1):121–130.

Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., and Cao, Y. (2023). Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*.

Gelana, F. and Yadav, A. (2019). Firearm detection from surveillance cameras using image processing and machine learning techniques. In *Smart innovations in communication and computational sciences*, pages 25–34. Springer.

Gius, M. (2018). The effects of state and federal gun control laws on school shootings. *Applied economics letters*, 25(5):317–320.

González, J. L. S., Zaccaro, C., Álvarez-García, J. A., Morillo, L. M. S., and Caparrini, F. S. (2020). Real-time gun detection in cctv: an open problem. *Neural networks*.

Grega, M., Lach, S., and Sieradzki, R. (2013). Automated recognition of firearms in surveillance video. In *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2013 IEEE International Multi-Disciplinary Conference on*, pages 45–50.

Gu, Y., Liao, X., and Qin, X. (2022). Youtube-gdd: A challenging gun detection dataset with rich contextual information. *arXiv preprint arXiv:2203.04129*.

Hnoohom, N., Chotivatunyu, P., and Jitpattanakul, A. (2022). Acf: an armed cctv footage dataset for enhancing weapon detection. *Sensors*, 22(19):7158.

Hurka, S. and Knill, C. (2020). Does regulation matter? a cross-national analysis of the impact of gun policies on homicide and suicide rates. *Regulation & Governance*.

IMFDB (2015). Internet movie firearms database. http://www.imfdb.org/index.php?title=Main_Page&oldid=911151. Last accessed in 18/01/2021.

Ineneji, C. and Kusaf, M. (2019). Hybrid weapon detection algorithm, using material test and fuzzy logic system. *Computers & Electrical Engineering*, 78:437–448.

Jocher, G., Chaurasia, A., and Qiu, J. (2023). YOLO by Ultralytics.

Lang, S., Ventola, F., and Kersting, K. (2021). Dafne: A one-stage anchor-free approach for oriented object detection. *arXiv e-prints*, pages arXiv–2109.

Lemieux, F. (2014). Effect of gun culture and firearm laws on gun violence and mass shootings in the united states: A multi-level quantitative analysis. *International Journal of Criminal Justice Sciences*, 9(1):74.

Lim, J., Al Jobayer, M. I., Baskaran, V. M., Lim, J. M., Wong, K., and See, J. (2019). Gun detection in surveillance videos using deep neural networks. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1998–2002. IEEE.

Loganathan, S., Kariyawasam, G., and Sumathipala, P. (2019). Suspicious activity detection in surveillance footage. In *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, pages 1–4. IEEE.

Mehta, P., Kumar, A., and Bhattacharjee, S. (2020). Fire and gun violence based anomaly detection system using deep neural networks. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 199–204. IEEE.

Olmos, R., Tabik, S., and Herrera, F. (2018). Automatic handgun detection alarm in videos using deep learning. *Neurocomputing*, 275:66–72.

Pawar, M., Dhanki, M., Parkar, S., Dandekar, C., and Gupta, B. (2019). A novel approach to detect crimes and assist law enforcement agency using deep learning with cctvs and drones.

Piza, E. L., Welsh, B. C., Farrington, D. P., and Thomas, A. L. (2019). Cctv surveillance for crime prevention: A 40-year systematic review with meta-analysis. *Criminology & Public Policy*, 18(1):135–159.

Raturi, G., Rani, P., Madan, S., and Dosanjh, S. (2019). Adocw: An automated method for detection of concealed weapon. In *2019 Fifth International Conference on Image Information Processing (ICIIP)*, pages 181–186. IEEE.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Sultani, W., Chen, C., and Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488.

Velastin, S. A., Boghossian, B. A., and Vicencio-Silva, M. A. (2006). A motion-based image processing system for detecting potentially dangerous situations in underground railway stations. *Transportation Research Part C: Emerging Technologies*, 14(2):96–113.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., and Shum, H.-Y. (2022). Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhou, Q., Li, X., He, L., Yang, Y., Cheng, G., Tong, Y., Ma, L., and Tao, D. (2022). Transvod: end-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.