CCTV-Gun: Benchmarking Handgun Detection in CCTV Images

Zhenghong Li*[1], Srikar Yellapragada*[1], Kevin Bhadresh Doshi[1], Purva Makarand Mhasakar[1], Heng Fan[2], Jie Wei[3], Erik Blasch[4], Bin Zhang[1], and Haibin Ling[1]

[1] Stony Brook University
[2] University of North Texas
[3] City College of New York
[4] Air Force Research Laboratory

Abstract. Gun violence is a critical security problem, and it is imperative to develop effective gun detection algorithms for real-world scenarios, particularly in Closed Circuit Television (CCTV) surveillance data. Despite significant progress in object detection, detecting guns in real-world CCTV images remains a challenging and under-explored task. Firearms, especially handguns, are typically very small, non-salient in appearance, and often severely occluded or indistinguishable from other small objects. Additionally, the lack of principled benchmarks and difficulty collecting relevant datasets further hinder algorithmic development. In this paper, we present a meticulously crafted and annotated benchmark, called **CCTV-Gun**, which addresses the challenges of detecting handguns in real-world CCTV images. Our contribution is three-fold. Firstly, we select and analyze real-world CCTV images from three datasets, manually annotate handguns and their holders, and assign each image with relevant challenge factors such as blur and occlusion. Secondly, we propose a new cross-dataset evaluation protocol in addition to the standard intra-dataset protocol, which is vital for gun detection in practice. Finally, we comprehensively evaluate both classical and state-of-the-art object detection algorithms. The benchmark will facilitate research and development on this topic and ultimately enhance security. Code, annotations, and trained models are available at https://github.com/srikarym/CCTV-Gun.

1 Introduction

Gun violence has been a severe security problem for a long time in many countries, especially the United States [18]. Many gun-related crimes, such as robbery and shootings, occur in public places with surveillance systems. However, reliance on human supervision demands an impractical amount of vigilance. Since most public surveillance systems are Closed Circuit Television (CCTV) cameras, automatic and fast detection of handguns in real-world CCTV imagery has the potential to prevent gun-related violence and/or increase interdiction response. Such a detection algorithm can potentially alert law enforcement agencies when an incident occurs. This study focuses mainly on handguns, the most commonly used type of gun in gun crimes [30].

* Authors make equal contribution to this work

In recent years, many effective deep-learning-based object detectors [25, 16, 33, 21] have been proposed. Handgun detection in real-world crime imagery is much more challenging than general object detection tasks for the following reasons: 1) the size of handguns is very small (e.g., a few pixels) in these images. The frames in crime CCTV videos are only in 320x240 resolution, lower than most object detection datasets [14]. Handguns usually occupy a small area in these images, which means there are no salient texture features, and their features may be obscured in the networks. 2) the holder's hands often occlude the handguns at crime scenes. Only the barrels, which are merely in slender rectangles, can be seen in many images; see Fig. 1 for examples. Therefore, handguns are easily misclassified since there are no salient shape or texture features. 3) the wide variety of camera angles and lighting conditions make detection even more difficult.

Fig. 1: Example surveillance images involving handgun violence from UCF-Crime [26]. CCTV-gun annotations include both handguns and their holders.

Some preliminary studies of gun detection from images/videos have been conducted based on generic object detectors. Most of these studies (e.g., [9, 19, 20]) focus on well-processed gun images, very different from images in real crime scenes. Only minimal efforts [8, 11] pay some attention to real-world scenarios with CCTV images but are restricted in either data size or use of actors for simulation. Moreover, there needs to be more evaluation of state-of-the-art (SOTA) visual detection algorithms for gun detection tasks, let alone a more complicated yet critical study on their generalization capability.

Fig. 2: Example challenging CCTV images for handgun detection from UCF [26] (occlusion), MGD [11] (blur), and USRT [8] (similar objects).

To address the above challenges, this paper presents **CCTV-Gun**, a meticulously crafted and annotated benchmark for real-world handgun detection from CCTV images. Our work tackles handgun detection comprehensively in three aspects: *benchmark construction*, *evaluation protocol*, and *thorough experiments*.

For *benchmark construction*, we first investigate relevant real-world CCTV imagery datasets and judiciously select images from three of them: Monash Gun Dataset (MGD) [11], US Real-time Gun detection dataset (USRT) [8], and UCF Crime Scene dataset (UCF) [26]. MGD and USRT datasets contain images of enacted crime scenes, while UCF is a general-purpose action recognition dataset. We extract frames from these datasets and provide bounding box annotations of person, handgun, and handgun-holder pairs (which person holds each handgun) for all images. Moreover, for each selected image, we label it with challenge factors (e.g., blur), which helps analyze the performance of detection algorithms.

For *evaluation protocol*, we propose two types of experiments: intra-dataset and cross-dataset testing. Intra-dataset is the standard evaluation technique, where a model is trained on the training split and evaluated on the test split of a given dataset. In Cross-

dataset testing, we train a model on two datasets, say MGD and USRT, and test it on the entirety of the third dataset – UCF. Cross-dataset evaluation indicates the generalization capability of the model. We also take the model trained on two datasets (from the previous experiment), fine-tune it on the training split, and evaluate it on the test split of the third dataset. The fine-tuning evaluation signifies if models pre-trained on gun-detection datasets act as a better initialization than the COCO-pre-trained model.

For *thorough evaluation*, we comprehensively test both classical CNN-based object detectors and state-of-the-art (SOTA) transformer-based detectors in all protocols. We also provide in-depth results analysis and insights for future directions.

We believe this benchmark will facilitate further research on this topic and ultimately enhance security. In summary, our main contributions are as follows:

- design the first carefully annotated benchmark, **CCTV-Gun**, for handgun detection in CCTV images,
- develop a new cross-dataset evaluation protocol in addition to the standard intra-dataset protocol, which is vital for gun detection, and
- provide thorough evaluation and analysis of SOTA object detection algorithms for handgun detection.

2 Related Work

2.1 Generic Object detection

Two Stage Methods. R-CNN [7] introduces the first two-stage detection algorithm. It generates region proposals using selective search, computes CNN features, and classifies them using an SVM. Fast R-CNN [6] uses ROI Pooling and jointly learning to detect spatial locations of objects and classify them. Faster R-CNN [25] uses a Region Proposal Network (RPN) to generate region proposals. To further boost the performance, [12] proposes the Feature Pyramid Network (FPN) to capture multi-scale features. Cascade RCNN [2] trains a sequence of RCNNs using the output of one stage to train the next. DetectoRS [21] uses recursive feature pyramids, which provide feedback from the top-down to bottom-up layers of an FPN.

One Stage Methods. One-stage detection models [22–24, 15, 13] skip the region proposal phase and directly make final predictions. YOLOv1-v3 [22–24] are a representative series of one-stage algorithms, regarding detection as a regression task. Focal loss [13] is widely used to solve the mismatch between positive and negative samples.

Transformer-based Methods. Transformers [28] have been the de-facto choice of architecture in natural language processing tasks and have been widely applied for vision tasks [5, 16]. The first representative work for object detection is DETR [3], which uses a Transformer on CNN image features to predict all objects simultaneously directly. Deformable DETR [33] proposes a multi-scale deformable attention module enabling much faster training. More recently, transformer-based detection algorithms [10, 31] keep pushing the front end of detection performance.

2.2 Firearm Detection

There have been some preliminary studies on gun detection, but it remains a seriously underexplored area. A dataset from CCTV recordings with an actor is created in [9] for

Table 1: Details of various subsets. Images from the UCF crime scene are much smaller than the other two, making it the most challenging.

Source	Image size	# Unique backgrounds	# images	# images with handgun	Avg size of handgun in pixels	Avg size of person in pixels
MGD [11]	512×512	41	2857	2852	25	158
USRT [8]	1920×1080	3	3294	1115	47	319
UCF [26]	320×240	76	1616	1597	16	79

gun and knife detection. It selects positive examples for gun detection by annotating the frames in the video with a gun. Internet Movie Firearms Database (IMFDB) [1] is a database of firearms featured in movies, TV shows, and video games. [29] constructs a dataset using images from IMFDB as positive samples and randomly collected internet images as negative samples. [19] presents a dataset of 9,100 images of people holding handguns. The images are obtained from online gun catalogs and advertisements. [20] published a dataset with 51,000 annotated images for gun detection, and most of these images were selected from IMFDB [1], and some from previously published datasets such as [19]. The imagery data from the above datasets are either non-CCTV or not in the real-world surveillance scene, hence inappropriate for our goal.

In [8], the US Real-time Gun Detection dataset (USRT) is constructed from a CCTV during a mock attack and annotated for the presence of handguns. USRT hires multiple people holding guns to walk through rooms with CCTV, and 4,118 images are collected. Synthetic data of people with handguns are also generated using the Unity Game engine. They train a Faster RCNN [25] network on synthetic data and fine-tune it on the mock attack data. [11] constructs the Monash Gun Dataset (MGD) of 2,500 images, enacting crime scenes recorded with a CCTV. They train an M2Det [32] model on a pooled dataset of their images and images from [19]. Despite these efforts, detecting guns from real-world CCTV imagery remains underexplored. The studies in [9, 29] train networks to classify whether an image has a handgun but skip the critical step of gun detection. The datasets in [19, 20] are neither from a CCTV perspective nor a real-world surveillance scene. The dataset in [8] can be used to pre-train a gun detection model, but they still need to evaluate their model on real crime scene data.

Our work is inspired by the above studies but is the first for thorough benchmarking of handgun detection from real-world CCTV images. On the dataset part, we compile a new benchmark by selecting appropriate images from USRT and MGD, together with the real-world UCF Crime Scene dataset (UCF) [26]. Besides, we provide richer annotations, enhanced thorough evaluation protocols, and more comprehensive evaluations.

3 CCTV-Gun Benchmark

3.1 Dataset Construction

There has been some preliminary work on handgun detection datasets, but most need improvement. Our dataset, **CCTV-Gun**, consists of images taken from various CCTV cameras and scenarios. We focus mainly on handguns, the most commonly used type of firearm in gun crimes [30]. Instead of capturing new images, which is a difficult task, we

seek help from three publicly available datasets: Monash Gun Dataset [11], US Real-time Gun detection dataset [8], and UCF Crime scene dataset [26]. More details about images from different subsets can be found in Table 1.

The original MGD [11] dataset had 7,811 images. Each annotated CCTV image in this dataset contains the presence of a handgun. The images are of size 512×512. This dataset had 4,954 stock images obtained from the internet. We discard them as they are not from a CCTV perspective. We take the remaining 2,857 images from 250 recorded CCTV videos in various indoor and outdoor settings.

The USRT [8] dataset consists of 5,149 images from 3 different CCTV cameras, varying lighting conditions, conflicting objects such as fire extinguishers, and often containing multiple people. The photos are of size 1920×1080. This dataset also annotated knives and shotguns, but we ignored them and considered them as background. We discard 650 images with no objects. We take 3294 images from this dataset. MGD and USRT are mock datasets, meaning the creators have acted out the attack scenes.

UCF Crime scene dataset [26] is a large-scale dataset of 128 videos. It contains 1900 untrimmed videos showing 13 anomalies. It is not a gun-detection dataset but a general-purpose anomaly detection dataset. We use the Robbery and Shooting videos (in 320×240 resolution), which are CCTV camera recordings of real-world crime scenes. We select 57 robbery and 17 shooting videos, extract handgun images in 2 frames/second, and obtain 1616 handgun frames.

3.2 Annotation

We annotate two objects in each image: the handgun and the person. MGD and USRT datasets already provide handgun annotations, whereas the UCF dataset has no annotations. We provide handgun holder annotations for the first time, making it different from previous works. It is equally important to detect the holder at a potential crime scene. In total, we obtained 7767 annotated images. Among these images were 5 images from MGD, 19 from UCF, and 2197 from USRT, which didn't have any handguns. We still include them in the dataset, as they serve as negative examples with a person but no handgun. Examples from our dataset can be seen in Fig. 2. We use a graphical image annotation tool *labelimg* [27] to draw bounding box annotations in our dataset.

Table 2: Number of images with challenging attributes in each dataset.

	Occlusion		Blur		Similar objects	
	USRT	UCF	USRT	MGD	USRT	MGD
# of images	17	34	33	17	109	29

We also provide annotations of handgun holder pairs - the person holding each handgun. Although we have not used the pair annotations in training our models, we believe it will benefit future works. Using a human-object interaction model, one can refine the handgun features or find the holder for each handgun in an image.

We annotate the test split of our dataset with the following challenges or attributes:

- **Occlusion**: Handguns are often occluded due to the holder's hands at crime scenes, where only a tiny portion of the handgun is visible.
- **Blur**: Since these images are captured from CCTV, some are blurry due to motion.
- **Similar object**: Other small-sized things, such as mobile phones, can be misclassified since handguns occupy a small area in these images.

Table 4: Intra-dataset performance (average precision at IoU=0.5).

Backbone	Framework	BS	LR	MGD		USRT		UCF	
				handgun	person	handgun	person	handgun	person
ResNet50	Faster RCNN + FPN	12	0.01	86.8	94.8	43.7	80.0	43.4	87.2
ResNet50	Deformable DETR	8	0.0001	89.3	96.8	36.5	80.6	48.4	86.5
ResNet50	DetectoRS	4	0.0002	87.4	95.1	**48.9**	81.6	54.5	89.4
Swin-T	Faster RCNN + FPN	4	0.01	**91.7**	93.5	44.8	86.0	**57.4**	89.4
ConvNeXt-T	Faster RCNN + FPN	6	0.0001	88.2	95.5	48.1	83.1	56.7	89.5

Table 5: Average precision of cross-dataset experiments.

Backbone	Framework	Train : MGD + USRT Test : UCF		Train : USRT + UCF Test : MGD		Train : UCF + MGD Test : USRT	
		handgun	person	handgun	person	handgun	person
ResNet50	Faster RCNN + FPN	3.7	16.0	47.8	89.8	22.1	86.4
ResNet50	Deformable DETR	**11.7**	64.4	**61.2**	95.8	15.9	83.1
ResNet50	DetectoRS	10.3	42.0	60.7	93.7	25.7	88.2
Swin-T	Faster RCNN + FPN	6.8	17.2	48.5	92.1	26.2	87.7
ConvNext-T	Faster RCNN + FPN	7.7	32.5	47.9	93.0	**27.3**	86.4

Table 6: Average precision of COCO pretrained (and then fine-tuned) model on handgun detection. In USRT + UCF pretrained column, models were pretrained on COCO, then USRT + UCF, and finally fine-tuned on the target MGD dataset.

Backbone	Framework	MGD		USRT		UCF	
		COCO pretrained	USRT+UCF pretrained	COCO pretrained	UCF+MGD pretrained	COCO pretrained	MGD+USRT pretrained
ResNet50	Faster RCNN+FPN	86.8	**87.0**	43.7	**47.1**	**43.4**	32.2
ResNet50	Deformable DETR	89.3	**90.5**	36.5	**39.0**	**48.4**	43.2
ResNet50	DetectoRS	87.4	**88.4**	**48.9**	45.9	**54.5**	50.8
Swin-T	Faster RCNN+FPN	91.7	**92.7**	**44.8**	41.0	**57.4**	47.5
ConvNext-T	Faster RCNN+FPN	**88.2**	86.8	48.1	**48.9**	**56.7**	56.7

These three attributes were chosen based on visually examining the images. Handguns in MGD images do not have any occlusions, but there are a lot of similar objects, such as mobile phones. In USRT, we found many blurry photos. Since we consider knives and shotguns in USRT as background, they could confuse the detector. We include images with these objects in the “Similar objects" category. Details can be found in Table 2. Examples of such challenging images are shown in Fig. 2.

3.3 Evaluation Protocols

Table 3: Train-val-test split.

Source	Total	Train	Val	Test
MGD	2857	2164	287	401
USRT	3294	2492	308	494
UCF	1616	1435	0	185

Intra-dataset protocol We train and evaluate the models on each dataset individually. The train-val-test split for each dataset is presented in Table 3.

Cross-dataset protocol Our approach involves training a model on two datasets, $D_1 + D_2$, and testing it on the entirety of D_3. It allows us to assess the generalization ability of SOTA models. We use the same training hyperparameters as Intra-dataset. We then

fine-tune the model trained on two datasets on the third dataset, and evaluate the model's performance on the test split of the third dataset.

4 Experiments and Analysis

4.1 Setup

We perform two kinds of assessment: Intra-dataset and Cross-dataset. We train five object detection methods on our datasets: Faster R-CNN [25], Swin-T [16], Deformable DETR [33], DetectoRS [21], and ConvNeXt-T [17]. Besides, we use the two-stage with refinement variant of Deformable DETR and employ Cascade RCNN head [2] on DetectoRS. We only use two-stage methods as they are generally more accurate [25] and better suited for detecting small objects [22].

The implementations and COCO pre-trained models are based on MMdetection [4]. We train these models for 36 epochs on a 24GB Nvidia A5000 GPU. We decay the learning rate by 0.1 at epochs 27 and 33. Table 4 provides the training and framework details of the models employed. Our analysis uses the average precision (AP) at IoU = 0.5. We compute the average precision value for both handgun and person classes.

4.2 Results and analysis

Intra dataset protocol Table 4 presents the results of Intra-dataset evaluation. All five models perform well on MGD, with Swin-T achieving the highest handgun AP score. Swin-T and DetectoRS were the top-performing models on UCF and USRT. Fig. 3 provides qualitative results. It is noteworthy that despite MGD and USRT being high-resolution images, the models performed considerably worse on USRT. There are several possible reasons. Firstly, the USRT dataset has fewer positive examples of handguns, with only 1115 out of 3294 images featuring handguns. In contrast, handguns are present in almost all the images from MGD, providing detection models with fewer positive examples. Secondly, several images from USRT include similar objects like knives and torch lights. This makes the detection task much more challenging. Finally, we ignored larger guns like Shotguns and Assault rifles in the annotation, treating them as background, which makes it more difficult.

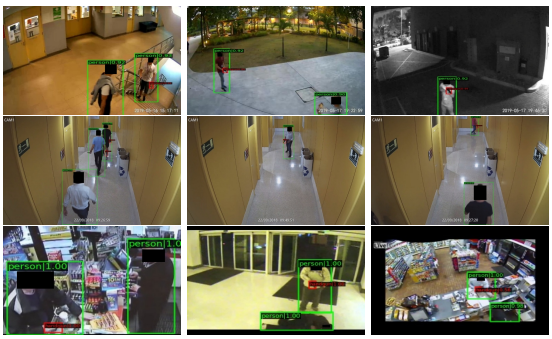

Fig. 3: Qualitative results of the best-performing models. Top: MGD, Center: USRT, Bottom: UCF.

Cross dataset protocol Results for Cross-dataset evaluation (without fine-tuning) can be found in Table 5. We observe that models trained on MGD + USRT perform poorly on the UCF dataset. MGD and USRT are made of enacted crime scenes with clear, high-resolution frames, whereas UCF data comprises real crime scene images taken at

Table 7: Performance of detection models on challenging attributes. We report the average precision of the handgun class computed on the images with selected attributes.

Backbone	Framework	Occlusion		Blur		Similar objects	
		USRT	UCF	MGD	USRT	MGD	USRT
ResNet50	Faster RCNN + FPN	50.9	38.9	76.6	34.4	69.5	45.1
ResNet50	Deformable DETR	16.0	20.6	**77.5**	23.4	70.2	22.1
ResNet50	DetectoRS	26.0	27.7	77.3	29.9	67.8	45.0
Swin-T	Faster RCNN + FPN	45.6	23.3	76.4	40.1	**73.8**	47.7
ConvNext-T	Faster RCNN + FPN	**62.8**	**44.1**	75.5	**42.8**	69.9	**49.6**

low resolution. Models trained on USRT + UCF perform pretty well on MGD since images in MGD are clear images with very few occlusions.

Fine-tuning results can be found in Table 6. Only when fine-tuned on MGD did the Gun-detection (USRT+UCF) trained model perform better than COCO pre-trained one. The effectiveness is inconclusive for USRT, where the Gun-detection (UCF+MGD) trained model performs better in 2 out of 5 cases. In UCF, we observed worse performance when the MGD+USRT model was used for fine-tuning. Since MGD is the least challenging dataset among the three, pre-training may have helped. UCF, with its small images and heavy occlusions, obtains more significant benefits when a COCO pre-trained model is used for fine-tuning. The combined size of MGD + USRT might not have been enough to act as an effective pre-training dataset in this case.

4.3 Challenging attributes

We annotate the test split of our dataset with three challenging attributes: occlusions, blur, and similar objects. We then evaluate models trained on each dataset on these attributes. We report the average handgun precision at IoU = 0.5 for these models in Table 7. Results are similar to Intra-dataset evaluation - models that perform well are more robust towards challenges.

5 Conclusion

To address the challenges in real-world gun detection, we presented **CCTV-Gun**, a meticulously crafted and annotated benchmark for real-world handgun detection from CCTV images. Through detailed bounding box annotations for persons, handguns, and handgun holder pairs, combined with the evaluation protocol and thorough experiments, our benchmark provides a valuable resource for training and evaluating handgun detection algorithms. We hope that the availability of this benchmark will facilitate further research in this area and encourage the development of more effective solutions to address the serious issue of gun violence.

Acknowledgments. We thank Sabbarish Ramana Rajan for his contribution to the early study. We also thank the authors of the UCF Crime Scene dataset [26], US Real-time Gun detection dataset [8], and Monash Gun Dataset [11], which together form the base of CCTV-Gun. The work was supported in part by US NSF Grant 2006665 and AFOSR Grant FA 9550-23-2-0002. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of agencies.

References

1. Internet movie firearms database - guns in movies, tv and video games, http://imfdb.org/
2. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
4. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
7. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Region-based convolutional networks for accurate object detection and segmentation. IEEE transactions on pattern analysis and machine intelligence **38**(1), 142–158 (2015)
8. González, J.L.S., Zaccaro, C., Álvarez-García, J.A., Morillo, L.M.S., Caparrini, F.S.: Real-time gun detection in cctv: An open problem. Neural networks **132**, 297–308 (2020)
9. Grega, M., Matiolański, A., Guzik, P., Leszczuk, M.: Automated detection of firearms and knives in a cctv image. Sensors **16**(1), 47 (2016)
10. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: Accelerate detr training by introducing query denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13619–13627 (2022)
11. Lim, J., Al Jobayer, M.I., Baskaran, V.M., Lim, J.M., See, J., Wong, K.: Deep multi-level feature pyramids: Application for non-canonical firearm detection in video surveillance. Engineering applications of artificial intelligence **97**, 104094 (2021)
12. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
15. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
17. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
18. Lopez, G.: America's unique gun violence problem, explained in 17 maps and charts. Vox. Retrived from https://www. vox. com/policy-and-politics/2017/10/2/16399418/us-gun-violence-statistics-maps-charts (2018)

19. Olmos, R., Tabik, S., Herrera, F.: Automatic handgun detection alarm in videos using deep learning. Neurocomputing **275**, 66–72 (2018)
20. Qi, D., Tan, W., Liu, Z., Yao, Q., Liu, J.: A dataset and system for real-time gun detection in surveillance video using deep learning. In: 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 667–672. IEEE (2021)
21. Qiao, S., Chen, L.C., Yuille, A.: Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10213–10224 (2021)
22. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
23. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7263–7271 (2017)
24. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
25. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
26. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6479–6488 (2018)
27. Tzutalin: Labelimg. https://github.com/heartexlabs/labelImg (2015)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
29. Verma, G.K., Dhillon, A.: A handheld gun detection using faster r-cnn deep learning. In: Proceedings of the 7th international conference on computer and communication technology. pp. 84–88 (2017)
30. Zawitz, M.W.: Guns used in crime. Washington, DC: US Department of Justice: Bureau of Justice Statistics Selected Findings, publication NCJ-148201 (1995)
31. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In: International Conference on Learning Representations (2022)
32. Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., Ling, H.: M2det: A single-shot object detector based on multi-level feature pyramid network. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 9259–9266 (2019)
33. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)