2023 IEEE International Conference on Image Processing (ICIP) | 978-1-7281-9835-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIP49359.2023.10222473

OBJECT DETECTION AND COUNTING CHALLENGES IN REAL STREET MONITORING: CASE STUDY OF HOMELESS ENCAMPMENTS

Abdullah Alfarrarjeh[§], *Seon Ho Kim*[‡], *Utkarsh Baranwal*[‡], *Yash Bitla*[‡]

[‡] *Integrated Media Systems Center, University of Southern California, Los Angeles, CA 90089, USA*
[§] *Department of Computer Science, German Jordanian University, Amman 11180, Jordan*
abdullah.alfarrarjeh@gju.edu.jo, ⟨ seonkim, baranwal, bitla ⟩@usc.edu

ABSTRACT

Wide area urban street monitoring is highly demanding in various smart city applications. Manual monitoring is both laborious and time-consuming, hence automatic vision-based monitoring is a more feasible alternative. An essential part of vision-based street monitoring is detecting and counting objects of interest. However, these tasks are not straightforward due to various challenges, i.e., noisy conditions in a real environment, such as occlusion and high illumination. This study investigates the impact of these challenges on object detection and counting accuracy, then provides an empirical study to address the challenges with respect to video-based street monitoring. The selected case study demonstrates detecting and counting of homeless encampments in Los Angeles streets using street-level videos collected from a moving vehicle.

Index Terms— Street monitoring, object detection and counting, homeless encampment, noisy conditions

1. INTRODUCTION

Urban streets constitute the nervous system of a city and the issues related to streets may impact various aspects of city life such as transportation, public health, pollution, economy, and tourism. Given that an efficient monitoring of wide area urban streets is in need, manual monitoring is impractical; hence automatic street monitoring might be more feasible. Towards this end, automatic street monitoring solutions using visual data have been proposed, such as traffic flow analysis [1, 2], road damage detection [3, 4, 5, 6], street cleanliness [7, 8], graffiti detection [9, 10], air pollution detection [11, 12], and water leakage detection [13]. Most of these solutions require a visual dataset along with a machine-learning algorithm to train a model capable of detecting specific objects.

Vision-based street monitoring aims at understanding street scenes. Conventionally, understanding a scene implies detecting objects of interest in an image that convey comprehensive information about streets. In some cases, counting the number of object instances is additionally required for a better understanding of scenes. A fundamental approach is based on video streams due to the ubiquity of surveillance cameras, dash cameras, and smartphone cameras providing plentiful visual data around streets. In a straightforward way, video streams can be converted into a sequence of frames (i.e., still images) and a conventional object detection algorithm (e.g., Faster R-CNN [14] or YOLOv5 [15]) can be employed to detect objects of interest (e.g., vehicles, illegal trash dumping, and homeless encampments). However, such an image-based object detection method may result in duplicate counting across a series of frames in a video. Therefore, a video-based analysis method is needed to detect and correctly count objects. One of the state-of-the-art solutions is an object tracking algorithm in which objects of interest are detected and assigned unique identifiers, then tracked in the following frames in a video. The result of an object tracking algorithm is straightforwardly a list of uniquely tracked objects, which becomes implicitly the counting result. There are well-known object tracking algorithms including OpenCV Trackers [16], Sort [17], and DeepSort [18].

In a real environment, however, visual street monitoring may become a more challenging task than a straightforward object tracking because of the existence of some factors that complicate the task. Examples include occlusion, high illumination, and capturing settings (e.g., the distance from an object), to name a few. As a result of the surrounding environment or data acquisition setup, we refer to these factors as *noisy conditions*. This study is to investigate and identify which noisy conditions would affect (or not) the overall accuracy of object detection and counting in real street monitoring. The noisy conditions make training a model difficult (i.e., more detection failures) since they cause an ambiguity for learning objects of interest in images. However, this might not affect the overall object counting result in a video as the missing objects can be detected in the following video frames with different noisy conditions.

For a focused study, we chose homeless encampment detection and counting as the case study since homelessness is one of the most serious social issues in cities, e.g., Los Angeles's homeless population reached 66,433 in 2020 [19]. Furthermore, the locations and shapes of homeless encampments in urban streets provide a challenging complexity with various noisy conditions. Our experiments demonstrated that learning with selected noisy conditions outperformed learning with all noisy conditions or no noisy conditions.

2. RELATED WORK

Monitoring under Noisy Conditions: Certain research works have tackled the problem of object detection under noisy conditions [20, 21]. Also, another group of work investigated the problem of object tracking under noisy conditions [22, 23]. To minimize the impact of noisy conditions, researchers have attempted several strategies including attention using transformers and interrelation of tracked objects using graph convolutional neural networks. It is worth noting that the impact of noisy conditions on object counting is different from that on both object detection and tracking especially on video-based monitoring. For example, if an object is not detected at a certain video frame or fails to be tracked in a specific video segment but that object is detected and tracked in the following frames, that object is implicitly counted in the final result. Hence, solutions that minimize the effects of noisy conditions on detection and tracking are not necessarily sufficient for object counting.

Counting: The work for counting objects in images can be classified into three categories: counting by detection, counting by glance, and counting by subitizing. In counting by detection, an object detection algorithm (e.g., Faster R-CNN [14] or YOLOv5 [15]) can be used to localize objects and classify them; hence counting can be conducted implicitly. Glancing estimates object count in one shot without the need to detect and localize objects within an image. Counting by glance was investigated in two ways; one that trains a CNN model to estimate a global count in one step [24, 25], and the other that uses regression methods to estimate the count [26, 27]. In subitizing, images are divided into regions and a conventional method (e.g., counting by glance) is used for each region [28]. Regarding counting objects in videos, a class of proposed approaches targets to analyze a virtual region of interest to count objects (e.g., [29]) and such approaches are suitable when the capturing camera is stationary while objects are movable. Another approach depends on object track algorithms to count trajectories of objects [30].

Homeless Encampment Detection. A few research works have investigated homeless encampment detection using visual data, such as refugee camp detection in satellite imagery using various methods [31, 32, 33]. In particular, Giada et. al. [31] proposed approaches using pixel-based supervised classification, unsupervised classification, mathematical morphology, and object-based segmentation and classification. Land et al. [33] studied on object-based segmentation and classification while Wang et. al. [33] used mathematical morphology analysis. Recent research work is the one by Fisher et. al. [34] which proposed a deep learning framework named TentNet based on ResNet2 [35], InceptionV3 [36], and MobileNet [37] for detecting homeless camps in satellite images. To the best of our knowledge, our work is the first one which aims at detecting and counting homeless encampments using street-level video data.

3. CHALLENGES IN STREET MONITORING

In real street monitoring, image scenes may be associated with noisy conditions. In our study, we focus on the following six types of noisy conditions (note that we explain them with regard to the homeless encampment objects), see Fig. 1:

- *An Occluded Encampment*: This occurs when the view of an encampment is blocked partially by other nearby objects (e.g., pedestrians, poles, and trash cans).
- *A Truncated Encampment:* This occurs when a part of the encampment is not visible because the camera field of view does not capture the whole encampment.
- *A Hand-made Encampment*: This occurs when an encampment is made by a person, instead of being purchased from a store. Such an encampment appears in a non-standardized irregular shape.
- *A Tiny-appeared Encampment*: This occurs when an encampment appears small in an image since it is located far from the capturing camera.
- *A Blurry Encampment*: This occurs when an encampment is not captured clearly, e.g., due to a sudden movement in the capturing camera position.
- *A High-illuminated Encampment*: This occurs when an encampment is captured in a high-brightness lighting condition.

4. OBJECT DETECTION & COUNTING USING VIDEO STREAMS

4.1. Impacts of Noisy Conditions

In our study, we adopted a video-based street monitoring solution in which we integrated YOLOv5 as an object detector into DeepSort as an object tracker[1]. This solution enables the detection and counting of homeless encampments.

Real street monitoring for homeless encampments is inevitably associated with several noisy conditions (as discussed in Section 3), which makes accurate detection and counting challenging. There are two main points to address for a more accurate result. First, since video-based street monitoring is composed of three sequential stages (namely, detection, tracking, and counting), the effect of a detection error in the first stage is propagated to the next ones. In particular, incorrect object detection (e.g., identifying a car as a homeless encampment) results in tracking the wrong object and reporting it in the final counting result (see Fig. 2). Therefore, it is critical to increase the detection precision by decreasing the number of false positives in the context of our problem. Second, the impact of a failure to detect an object in a certain video segment is limited when that object is detected in a subsequent segment and then counted in the result. The dismissal of an object is rectifiable when it is due to a sudden or temporary noise on the object (e.g., view truncation or tiny appearance, see Fig. 3). However, when an object is associated with a permanent noisy condition, such an object may not be detected and counted (see Fig. 4). Therefore, the recall of a video-based monitoring method may not be critical based on the nature of noise conditions (i.e., temporary or permanent).

4.2. Our Approaches

Based on the discussion in the previous subsection, we present the following three approaches to address the issue of video-based street monitoring for homeless encampments.

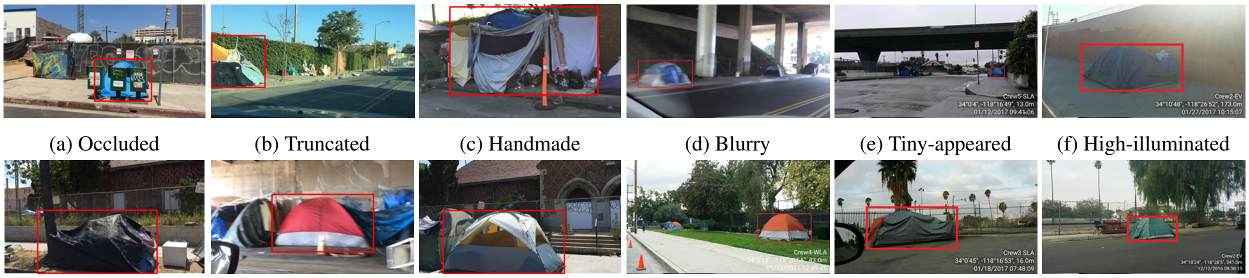
1. *Real Environment Learning Approach* ($RELA$): this involves training a model using a dataset that includes all examples of homeless encampments, regardless of whether they are noise-free or noisy. It is important to note that the specific types of noisy conditions are not considered during the training process, and they are not taken into account in the approach's results. The sole focus of this approach is to detect homeless encampments without identifying the specific types of noisy conditions associated with them. In other words, the training phase does not involve any metadata related to the noisy conditions.
2. *Optimal Environment Learning Approach* ($OELA$): this simulates an ideal environment by training a model solely on a dataset that contains noise-free examples of homeless encampments.
3. *Semi-Real-Environment Learning Approach* ($SELA$): this aims to examine the impact of each noisy condition on the counting results, as well as the false detection or false dismissal rates (precision and recall). To achieve this, we propose $SELA$ where multiple models are trained. Each model is trained using all examples of homeless encampments except those affected by a specific noisy condition. This allows us to investigate the influence of each noisy condition on the results.

5. DATASET AND EXPERIMENTS

5.1. Dataset Collection

To the best of our knowledge, there is no public dataset for homeless encampments. Therefore, to investigate our problem of street

[1] YOLOv5 and DeepSort are one of the state-of-the-art solutions for object detection and tracking, respectively. Selecting the best object detection and tracking algorithms is out of the scope of this study.



(a) Occluded (b) Truncated (c) Handmade (d) Blurry (e) Tiny-appeared (f) High-illuminated

(g) Non-occluded (h) Non-truncated (i) Standard-shape (j) Non-blurry (k) Non-Tiny-appeared (l) Non-high-illuminated

Fig. 1: Image Examples for Homeless Encampment with and without Noisy Conditions

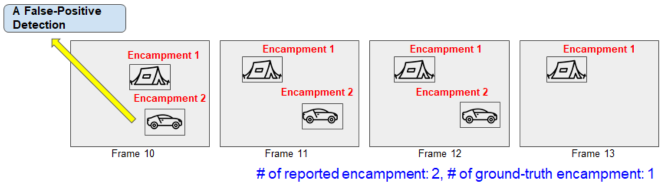


Fig. 2: Potential Impact of Noisy Conditions on the Precision of Video-based Street Monitoring

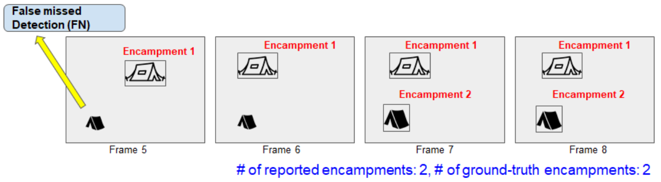


Fig. 3: Potential Impact of Temporary Noisy Conditions on the Recall of Video-based Street Monitoring

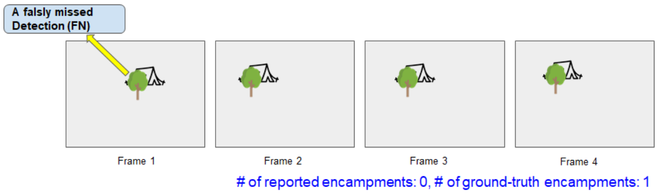

A falsly missed Detection (FN)
Frame 1
Frame 2
Frame 3
Frame 4
# of reported encampments: 0, # of ground-truth encampments: 1

Fig. 4: Potential Impact of Permanent Noisy Conditions on the Recall of Video-based Street Monitoring

monitoring for the purpose of homeless encampment detection and counting, we collected a set of videos and images in urban streets. For the video subset, we captured a set of videos from a moving vehicle on urban streets in Los Angeles. This subset includes 21 videos in MP4 format. On average, each video is 17 seconds long, 29 MB in size, and contains 509 frames. The total number of frames extracted from all videos is 10,680. Regarding the subset of images, the Los Angeles Sanitation and Environment (LSAN) department provided us with 817 images and we crawled 135 images from the Internet. In total, the image subset comprises 952 images.

5.2. Data Annotation

For a supervised learning, the collected images and videos were manually annotated by graduate students using an open-source annotation tool named Visual Object Tagging Tool (VoTT) [38]. To enhance the annotation consistency and accuracy, annotation tasks were repeated and alternated among the students. As a result, we obtained 1,927 instances of homeless encampments in the collected images and 12,560 instances in the video frames.

When annotating a video frame or an image, a bounding box was created surrounding a homeless encampment. Each encampment was associated with one label composed of six flags. These flags refer to the status of the six noisy conditions discussed in Section 3. Annotators manually set flags for noisy conditions including occlusion, truncation, illumination, and non-standardized shape. The remaining two flags were set automatically using a program. Specifically, if the area of a bounding box occupies less than 2% of the image/frame area, it is considered as a tiny-appeared encampment. To annotate blurry encampment, we followed a technique based on wavelet transform (proposed by Tong et al [39]) and used an edge detection threshold of 25 and a MinZero threshold of 0.001. Table 1 shows the distribution of examples of the annotated homeless encampments among the six noisy conditions.

5.3. Experimental Settings and Evaluation Metrics

For the object detection stage, we trained a YOLOv5m model with a batch size of 64 images, an image resolution of 704x704, and the SGD optimizer. At the training phase, the YOLOv5 model was trained using all collected images and frames extracted from a selected group of five videos. The remaining sixteen videos were used for testing the trained model. At the testing phase, we set the intersection over union (IoU) threshold to 0.6 and the confidence threshold to 0.6. For the object tracking stage, we used DeepSORT model with a min_confidence of 0.5, a nms_max_overlap of 0.5, max_age of 30, and n_init of 3.

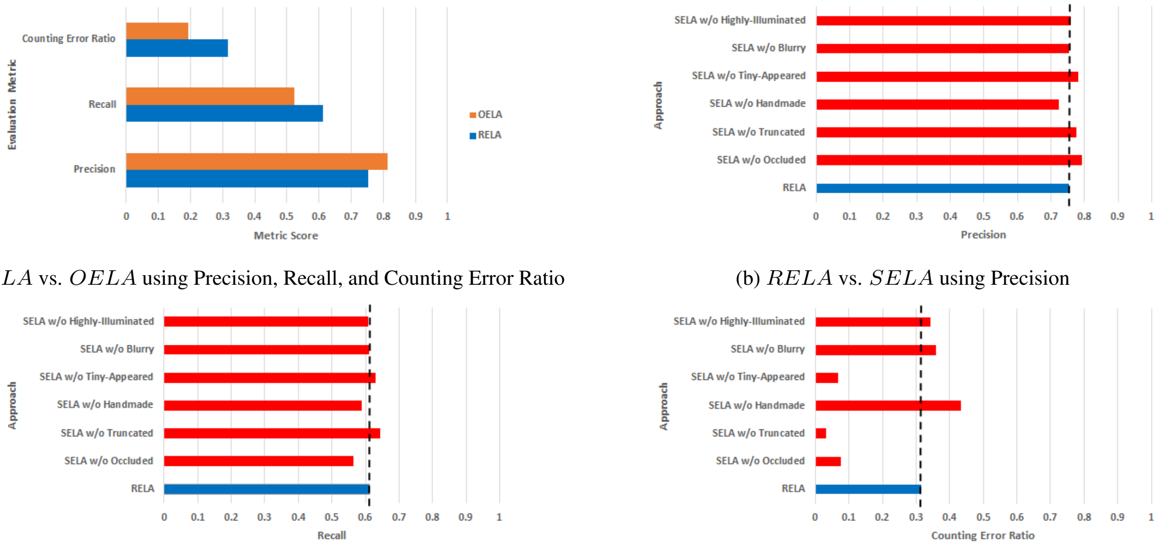
In terms of evaluation metrics, we used precision and recall to evaluate the result accuracy of the object detection stage. We also reported the evaluation of the counting stage[2] using counting error ratio as illustrated in Eq. 1.

$$Counting\ Error\ Ratio\ (CER) = \frac{|Reported\ Count - Ground\ Truth\ Count|}{|Ground\ Truth\ Count|} \tag{1}$$

[2] We omitted the evaluation of the middle stage due to the space limitation.

Table 1: Statistics of the Number of Homeless Encampment Objects with Noisy Conditions

	# of All Homeless Examples	# of Homeless Encampment Examples with Certain Noisy Conditions					
		Occluded	Truncated	Hand-made	Tiny-appeared	Blurry	High-illuminated
Training Phase	3979	2639	2425	2367	872	1777	851
Testing Phase	8581	6185	6148	5174	2422	2255	1120



(a) $RELA$ vs. $OELA$ using Precision, Recall, and Counting Error Ratio

(b) $RELA$ vs. $SELA$ using Precision

(c) $RELA$ vs. $SELA$ using Recall

(d) $REAL$ vs. $SELA$ using Counting Error Ratio

Fig. 5: Evaluation Results

5.4. Results

Real vs. Optimal Environments: As shown in Fig. 5a, the baseline approach trained using a real environment dataset (i.e., $RELA$) achieved a precision of 0.75, a recall of 0.61, and a counting error ratio of 0.31. Meanwhile, the second baseline trained using an optimal environment dataset (i.e., $OELA$) achieved a better precision and counting error ratio of 0.81 and 0.19, respectively, while it obtained a lower recall of 0.52. $OELA$ achieved a better precision and less counting error compared to that of $RELA$ because training on an optimal environment dataset reduces confusion; hence decreasing false positives and potentially counting encampments with a high certainty. However, $OELA$ obtained a lower recall because the approach was not exposed to noisy examples of homeless encampments that inevitably exist in real environments. Therefore, such an approach could be enhanced by achieving a better recall along with a better precision and counting error ratio.

Real vs. Semi-real Environments: The objective of a semi-real environment learning approach is to discover the type of noisy conditions that can be excluded given that the recall of street monitoring may not be always affected by all types of noisy conditions. Therefore, This experiment demonstrates the accuracy of this approach using several variants, each trained using a dataset excluding a specific type of noisy conditions. As shown in Figs. 5b-5d, the worst-performing variant of $SELA$ was when handmade encampments were excluded, resulting in a lower precision and recall and a higher counting error ratio. This indicates that such a noisy condition should be included during training to achieve better results. Such noise is categorized as a permanent noise and its impact on the recall is not rectifiable. Meanwhile, the best variants of $SELA$ are when the approach is trained without either tiny-appeared or truncated encampments. In particular, with these variants, $SELA$ achieved better in both precision and recall along with less counting error ratio. Such noisy conditions have a temporary effect on the detection and counting of encampments accurately. Therefore, such noisy examples can be excluded in the training phase to produce an enhanced approach. The other variants did not enhance all metrics (i.e., precision, recall, and counting error ratio); thus, the corresponding noisy conditions should be excluded in the training phase. For example, the variants when excluding either highly-illuminated or blurry encampments, their corresponding variants achieved almost similar precision and recall values compared with $RELA$; however, these variants obtained higher counting error ratios.

6. CONCLUSION

This study investigated the impact of several challenges (termed as *noisy conditions*) existing on object detection and counting in video-based street monitoring. The study focused on identifying which noisy conditions affect (or not) the overall accuracy of counting objects of interest, specifically in the case of homeless encampment detection and counting. The identification of noisy conditions with minimal impact on accuracy can be effectively overlooked in vision-based machine learning development concerning data collection and training.

7. ACKNOWLEDGEMENTS

This research has been funded in part by the USC Integrated Media Systems Center.

8. REFERENCES

[1] Ruimin Ke et al., "Real-time traffic flow parameter estimation from uav video based on ensemble classifier and optical flow," *IEEE trans Intell Transp Syst*, vol. 20, no. 1, pp. 54–64, 2018.

[2] Seon Ho Kim et al., "Real-time traffic video analysis using intel viewmont coprocessor," in *DNIS 2013*. Springer, 2013, pp. 150–160.

[3] Hiroya Maeda et al., "Road damage detection and classification using deep neural networks with smartphone images," *COMPUT-AIDED CIV INF*, vol. 33, no. 12, pp. 1127–1141, 2018.

[4] Abdullah Alfarrarjeh et al., "A deep learning approach for road damage detection from smartphone images," in *Big Data*. IEEE, 2018, pp. 5201–5204.

[5] Vinuta Hegde et al., "Yet another deep learning approach for road damage detection using ensemble learning," in *Big Data*. IEEE, 2020, pp. 5553–5558.

[6] Maitry Bhavsar et al., "Country-specific ensemble learning: A deep learning approach for road damage detection," in *Big Data*. IEEE, 2022, pp. 6387–6394.

[7] Hema Begur et al., "An edge-based smart mobile service system for illegal dumping detection and monitoring in san jose," in *UIC*. IEEE, 2017, pp. 1–6.

[8] Abdullah Alfarrarjeh et al., "Image classification to determine the level of street cleanliness: A case study," in *BigMM*. IEEE, 2018, pp. 1–5.

[9] Albert Parra et al., "Automatic gang graffiti recognition and interpretation," *JEI*, vol. 26, no. 5, pp. 051409–051409, 2017.

[10] Abdullah Alfarrarjeh et al., "Recognizing material of a covered object: A case study with graffiti," in *ICIP*. IEEE, 2019, pp. 2491–2495.

[11] Chao Zhang et al., "On estimating air pollution from photos using convolutional neural network," in *ACM MM*, 2016, pp. 297–301.

[12] Yuncheng Li et al., "Using user generated online photos to estimate and monitor air pollution in major cities," in *ICIMCS*, 2015, pp. 1–5.

[13] Jiawei Chen et al., "Augmenting a deep-learning algorithm with canal inspection knowledge for reliable water leak detection from multispectral satellite images," *Adv. Eng. Inform.*, vol. 46, pp. 101161, 2020.

[14] Shaoqing Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[15] "YOLOv5," `https://github.com/ultralytics/yolov5`, 2020.

[16] "OpenCV Tracker API," `https://docs.opencv.org/3.4/d9/df8/group__tracking.html`, 2023.

[17] Alex Bewley et al., "Simple online and realtime tracking," in *ICIP*. IEEE, 2016, pp. 3464–3468.

[18] Nicolai Wojke et al., "Simple online and realtime tracking with a deep association metric," in *ICIP*. IEEE, 2017, pp. 3645–3649.

[19] "LA county homeless count resumes in january after covid-19 shut it down in 2021," `https://www.dailynews.com/2021/12/13/la-county-homeless-count-resumes-in-january-after-covid-19-shut-it-down-in-2021/`, 2021.

[20] Hsiu-Ming Yang et al., "Traffic sign recognition in disturbing environments," in *ISMIS*. Springer, 2003, pp. 252–261.

[21] PP Halkarnikar et al., "Object detection under noisy condition," in *AIP*. American Institute of Physics, 2010, vol. 1324, pp. 288–290.

[22] Mustansar Fiaz et al., "Tracking noisy targets: A review of recent object tracking approaches," *arXiv preprint arXiv:1802.03098*, 2018.

[23] Mk Bashar et al., "Multiple object tracking in recent times: A literature review," *arXiv preprint arXiv:2209.04796*, 2022.

[24] Cong Zhang et al., "Cross-scene crowd counting via deep convolutional neural networks," in *CVPR*, 2015, pp. 833–841.

[25] Santi Seguí et al., "Learning to count with deep object features," in *CVPR Workshops*, 2015, pp. 90–96.

[26] Antoni B Chan and Nuno Vasconcelos, "Bayesian poisson regression for crowd counting," in *ICCV*. IEEE, 2009, pp. 545–551.

[27] Victor Lempitsky and Andrew Zisserman, "Learning to count objects in images," *NIPS*, vol. 23, 2010.

[28] Prithvijit Chattopadhyay et al., "Counting everyday objects in everyday scenes," in *CVPR*, 2017, pp. 1135–1144.

[29] Jae-Won Kim et al., "Real-time vision-based people counting system for the security door," in *IEEK*. The Institute of Electronics and Information Engineers, 2002, pp. 1416–1419.

[30] Antoni B Chan et al., "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *CVPR*. IEEE, 2008, pp. 1–7.

[31] Silvia Giada et al., "Information extraction from very high resolution satellite imagery over lukole refugee camp, tanzania," *IJRS*, vol. 24, no. 22, pp. 4251–4266, 2003.

[32] Stefan Lang et al., "Earth observation (eo)-based ex post assessment of internally displaced person (idp) camp evolution and population dynamics in zam zam, darfur," *IJRS*, vol. 31, no. 21, pp. 5709–5731, 2010.

[33] Shifeng Wang et al., "Detecting tents to estimate the displaced populations for post-disaster relief using high resolution satellite imagery," *Int J Appl Earth Obs Geoinf*, vol. 36, pp. 87–93, 2015.

[34] Andrew Fisher et al., "Tentnet: Deep learning tent detection algorithm using a synthetic training approach," in *IEEE SMC*. IEEE, 2020, pp. 860–867.

[35] Kaiming He et al., "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[36] Christian Szegedy et al., "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.

[37] Andrew G Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[38] "Visual Object Tagging Tool VoTT," `https://github.com/microsoft/VoTT`, 2023.

[39] Hanghang Tong et al., "Blur detection for digital images using wavelet transform," in *ICME*. IEEE, 2004, vol. 1, pp. 17–20.