https://doi.org/10.2352/EI.2023.35.3.MOBMU-363

Improvement of Vehicles Accident Detection using Object Tracking with U-Net

Kirsnaragavan Arudpiragasam[1], ***Taraka Rama Krishna Kanth Kannuri***[1], ***Klaus Schwarz***[1,3], ***Michael Hartmann***[1], ***Reiner Creutzburg***[1,2]

[1] ***SRH Berlin University of Applied Sciences, Berlin School of Technology, Ernst-Reuter-Platz 10, D-10587 Berlin, Germany***
Email: klaus.schwarz@srh.de, reiner.creutzburg@srh.de

[2] ***Technische Hochschule Brandenburg, Department of Informatics and Media, IT- and Media Forensics Lab, Magdeburger Str. 50, D-14770 Brandenburg, Germany***
Email: creutzburg@th-brandenburg.de

[3] ***University of Granada, Faculty of Economics and Business, P.° de Cartuja, 7, ES-18011 Granada, Spain***

Keywords: Anomaly Detection; Confusion Metric; Residual Connection; Computer Vision; GAN; u-net; Hyperparameters; YOLOv4; Object Detection; Object Tracking

Abstract

The number of surveillance footage has dramatically expanded due to the increasing use of CCTV cameras in the public domain, like roads, supermarkets, and other public places, to increase the safety of the people. According to the "Road Safety Annual Report, 2021", between 20 and 25 million accidents occurred in 2021; this proves that detecting accident in surveillance videos is crucial for safety and maintenance. However, detecting anomalies is a significant challenge due to the various anomaly events, occlusions, and objects in the frame at different times. Further, researchers have proposed various approaches over the last decade to detect anomalies, including multimodal techniques, image reconstruction with optical flow, object detection (OD), and GAN. However, none of the studies has applied frame reconstruction with object tracking (OT) to detect anomalies. Therefore, this study focuses on road accident detection using a combination of OT and image reconstruction using a u-net with multiple variants such as u-net with skip -, residual-, and attention-connection. Firstly, background removal (BR) techniques were applied to remove and reduce the background variant in the frame. After that, the u-net algorithm was developed for reconstructing the images. Besides, YOLOv4 was used to detect objects, followed by DeepSort to track objects in the frame. Finally, the Mahalanobis distance and the reconstruction error (RCE) were determined using the Kalman filter and the u-net model. As mentioned in the proposed model, combining the u-net with skip plus residual connection and the OT algorithm achieved the highest accuracy.

INTRODUCTION

Detecting anomalies in video scenes is one of the biggest challenges because of variants of anomalies events, context information of the frames, heterogeneous class of anomalies, occlusions, the number of objects in the frame, and the moments of these objects. The invention of emerging computing power, computer vision (CV), deep learning (DL), and machine learning(ML) techniques have attracted the attention of researchers. Anomalies refer to that [1] unexpected behaviors or irregular patterns of data that deviate from the normal state of the data. Moreover, Anomaly detection (AD) can be applied to different areas, including network security intrusion detection, bank fraud detection, out-layer detection [2], bank fraud detection[3], production line defect detection[4], and incident detection[5].

The emerging improvement in closed-circuit television (CCTV) camera technologies and their underlying infrastructure components, such as networking, storage, and processing hardware, has led to the deployment of large numbers of surveillance cameras worldwide to improve security and public safety. The surveillance system automatically finds anomalies to improve people's safety problems, such as road safety, analyzing traffic accidents, seat belts, traffic signs, speed limits, license plates, and motorcycle helmets. Furthermore, successfully improving the CV approaches can more accurately detect objects such as vehicles, people, animals, etc. Thus, researchers focused on multi-modal techniques to detect anomalies in live streaming.

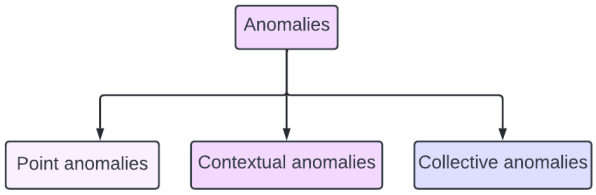

Figure 1. *Types of the anomalies that occurred based on the input data*

Figure 1 represents the anomaly types based on the input data.

- **Point anomalies:** In this type of anomaly, the observation is too far away from other data.
- **Contextual anomalies:** If the event is abnormal in a particular context, employing context-based techniques for iden-

tifying anomalies become crucial as data complexity increases. This type of anomaly often occurs with time-series data.

- **Collective anomalies:** A collective anomaly represents a set of abnormalities in the dataset but not in individual objects.

LITERATURE REVIEW

In recent years, there has been a rise in the number of studies dedicated to developing new approaches for detecting anomalies. Researchers initially started with a handcrafted features technique for detecting anomalies in a video. This approach involves a basic three-stage process: i) feature extraction of the image involves the features defined using the cascade model; ii) describing the behavior under normal conditions or regular patterns of features using trained models; iii) clustering the output features as anomalies or non-anomalies based on the similarity [6]. Conversely, in recent years, researchers have suggested several cutting-edge CV techniques to detect anomalies because CV algorithms can comprehend scene behavior and predict whether a frame is an anomaly.

Handcrafted Features Based Anomaly Detection

Handcrafted features refer to concepts developed by humans about the object's essential features and feed into the ML model to identify whether the same characteristics are present in the images [7]. Developing handcrafted features involves finding the suitable trade-off between accuracy and computational efficiency. The trajectory-based detection approach, statistical model learning [8], and dictionary learning [9] are prime examples of handcrafted techniques. Here, trajectory-based handmade processes are popular techniques to encode the non-anomalies scene. The study discussed a method based on the low-level features of multiple monitors in a set of fixed spatial positions. Finally, anomalies were detected by comparing each local monitor value [8]. Another study discussed the space-time Markov random field (MRF) model to detect abnormalities in the video. The nodes in the MRF map correspond to the phase of the images' local areas and connect to neighboring node links in space and time. The distribution of optical flow is determined using a mixture of probabilistic principal component analyzers. Eventually, the MRF model could detect abnormal activity in a local, and a global sense [10].

Accordingly, the study started with dictionary learning techniques, the concept of spatial RC cost over the dictionary learning approach to recognize the anomalies in the event; also, and the sparsity consistency constraint defined for the dictionary size [9]. The following study proposed the dictionary learning approach to the sparse RC methods in the following. Sliding windows with spatial and temporal axes define events or points of interest, which they describe using the histogram of gradients (HoG) and optical flow histogram. Spatial encoding updates these points in the dictionary. Finally, regular events in the video are reconstructed from the event dictionary, while unusual circumstances cannot be reconstructed [11].

Breakthrough, many traditional methods of trajectory-based detection of abnormal behavior depend on low-level features such as flow points, velocity vectors, and control points [10]. Also, this trajectory was generated by a human to detect specific abnormalities in the videos. The study presented an approach for trajectory-based detection of high-level feature information and contained three stages. In the first stage, three separate Gaussian mixture models (GMMs) build to understand spatial science, and each GMM was responsible for learning all entry, turn, and exit points. The second stage was responsible for learning region-to-region segments of travel patterns between access, turn, and exit regions. In the last step, probabilistic and particle filtering algorithms were used to efficiently obtain multiple hypotheses about the agents' destinations in the scene. This framework helped to classify trajectories in motion paths so that abnormal behavior is detected as soon as it occurs rather than after it is completed [12]. Another study discussed foreground segmentation techniques to improve the handcrafted feature model. First, the input images' initial foreground segmentation confines the foreground objects' analysis. Then, foreground images divide into non-overlapping spatial cells. Afterward, features are extracted based on the foreground cells' motion, size, and texture and analyzed to confirm the presence of anomalies in each cell using a speed check classifier. It evaluates the probability for motion size of foreground objects cell by the size-texture check, which first estimates the likelihood of the size of a foreground object. Furthermore, the classifier's output is divided into abnormal and non-anomalies cells [13].

However, these methods are not robust and can fail quickly due to (i) the regularity of human-defined features, (ii) the fact that model learning is not a self-learning process, (iii) occlusions and shadows in images, (iv) the statistical limitations of parametric models, and (v) the instability of the definitions of normality and abnormality.

Deep Learning Based Anomaly Detection

DL approaches have proven successful in many CVs, such as image classification, object recognition, OT, semantic segmentation, instance segmentation, image RC, and image capturing.

In recent years, the design methods of EDs, AEs, and GANs have attracted the attention of researchers. The authors proposed the AD network for appearance and motion [AMDN]. The proposed model contains three alternative ED architectures. Furthermore, the first ED is responsible for learning discriminative appearance in a random and unsupervised manner. Similarly, the second ED is accountable for learning motion's features in an arbitrary and unsupervised way. The third ED is responsible for learning the joint representation of motion and appearance features by performing initial fusion techniques to establish a pixel-by-pixel relationship between them. Finally, the post-fusion method combines the results of several single-class SVM classifiers [14]. The study developed the encoder part using a 3D convolutional neural network (CNN) to extract 3D information on the spatial and temporal. The decoder builds using a 3D CNN for image RC and another decoder to predict future images. Moreover, a weight-decreasing prediction loss introduces to evaluate the prediction of future images. A regularity value determines the RC of the image composed of the RC and the prediction loss. The model achieved an AUC value of 92.3% for the Ped 1 dataset [15].

Another study gave the anomaly technique a new dimension by introducing disparities between the ground truth and future frame predictions to discover anomalies. The intensity gradient loss was estimated as the difference between the future and predicted images after using a skip-connected Generator (u-net) to predict the future image. On the other hand, i) optical flow-1 calculates the difference between the actual frame and the predicted

image, and ii) optical flow-2 determines the difference between the ground truth frame and the expected frame. The difference between "optical flow-1" and "optical flow-2" is then used to calculate the final optical flow loss, and adversarial learning aids in determining if the prediction of the future picture is accurate. Finally, anomalies were detected in the video using all error values. The proposed method achieves 83.1% and 95.4% accuracy on the UCSD Ped1 and UCSD Ped2 datasets, respectively [16]. The above-mentioned study predicts future images and objects' optical flow in frames. However, in these instances, the temporal information of the image is lost. In addition, to fill the gap, another study suggested a variational AE in combination with a convolutional LSTM to predict future frames, where the convolutional LSTM aids in retrieving the temporal information of the previous frame. With suggested method achieves an AUC of 86.26% and 96.06% for the UCSD Ped1 and UCSD Ped2 datasets, respectively [17].

Subsequently, the study presented a theoretical evaluation of a DL-based solution for AD in the video. The proposed method consists of two phases. In the first step GAN model is used to predict future images, and a YOLOv3 model is used to detect objects. At the end of this phase, i) the MSE value calculates using the predicted and the actual image's difference, ii) the predicted object class, and iii) the object position calculates. The second step is a statistical step where MSE and adverse losses are analyzed, for which a constant false alarm rate threshold [CFAR] is to detect abnormal events [18]. The following study added a new concept involving images using a delay space vector as input to an encoder and a U-shaped network to predict future images. The optical flow loss of the SpyNet calculates by predicting the difference between the visual loss 1 of the future image and the future image and the optical loss 2 of the future image and the reconstructed image. Finally, a defined threshold for the error and optical flow loss aids in detecting abnormal events and evaluating the presence of non-anomalies behavior in the surveillance video. The proposed method achieves 86.9% and 96.1% of AUC for the UCSD Ped1 and UCSD Ped2 datasets, respectively [19]. Above mentioned techniques show excellent future frame prediction and RC performance using u-net with ED, GAN, and an AE.

Further, previous researchers used these aspects to predict future high-quality images. However, there was a lack of information between pixels that were spatially and temporally distant from each other. Then, a non-local u-net Generator introduces to predict the subsequent frames. Moreover, the non-local blocks define by considering the contraction and expansion paths. The systolic way of the non-local u-net uses convolutional layers and subsampling to extract features. While the expansion path helps to locate and retrieve the information as accurately as possible, the non-local block can enhance temporal and spatial characteristics and create video images with long-range dependencies [20]. Subsequently, previous researchers have used these aspects to predict future high-quality frames. However, there was a lack of information between pixels that were spatially and temporally distant from each other. Moreover, the study proposed model can capture anomalies and dark events. The new framework for AD, including improved prediction of future frames using a GAN-based model, facilitates anomalies events in the video. While previous GANs based on u-networks use 2D convolutional networks to account for spatial and temporal information, 3D u-net introduces to predict future images using 3D convolution combined with spatial and temporal features. Finally, anomalies are detected based on the RC score [21].

Most studies address AD but do not consider the computational time for AD in the video, which is critical for live-streaming surveillance videos. The study proposed a crossed u-net model to improve AD's accuracy and computational time in surveillance video. The crossed u-net model uses two u-net-based subnets and allows the output of each third layer of the systolic channel to combine with the result of the corresponding layer of the second u-net. The yield of the second sub-networks corresponding layer applies as the next layer's input. In addition, the cascading sliding window method determines the difference between each patch in the image. These patches are selected depending on which patch has the change to determine the anomaly score [22].

Research Gap

The researchers have developed multimodal techniques to deal with anomalies. However, none of the studies performed AD by removing the image background and combining OT and image reconstruction methods. Therefore, this study proposed a combination of OD, OT, and image reconstruction, using YOLOv4, DeepSort, and u-net to bridge the gap.

METHODOLOGY

Overview of AD System

Figure 2 shows the overview of the proposed model for detecting anomalies. Firstly, only the vehicle accident videos were separated from the UCF-Crime and Shanghai Tech datasets. The next phase is converting videos into irregularities and non-anomalies by manual inception. After that, all frames' background was removed using a u^2-net model. The third phase is responsible for developing the model, and the u-net model was developed with various variants such as:

- u-net model with SC
- u-net model with attention connection
- u-net model with skip and attention connection

On the other hand, to detect the objects in the frame, the YOLOv4 model was prepared; later, a DeepSort algorithm was applied to track the object noticed by the YOLOv4 model. The final phase is responsible for comparing the error metric received by each model variant.

Data Preparation

The data quality is the crucial input factor in validating any DL or ML models. However, after going through the research papers, we came across various datasets used by previous researchers.

Previously, many researchers have developed AD datasets, but few are publicly available. Table 1 shows some publicly available datasets used for the AD and the types of anomalies considered in the dataset. Among them, i) USCD – Ped1 and Ped2 are - one-channel (black and white) dataset and contains very few accident videos; therefore, it is not suitable for our research study, ii) the street scene dataset contains vehicle accident videos and other types of anomalies, and iii) UCF-Crime contains around twelve various anomaly events including the accidents videos.

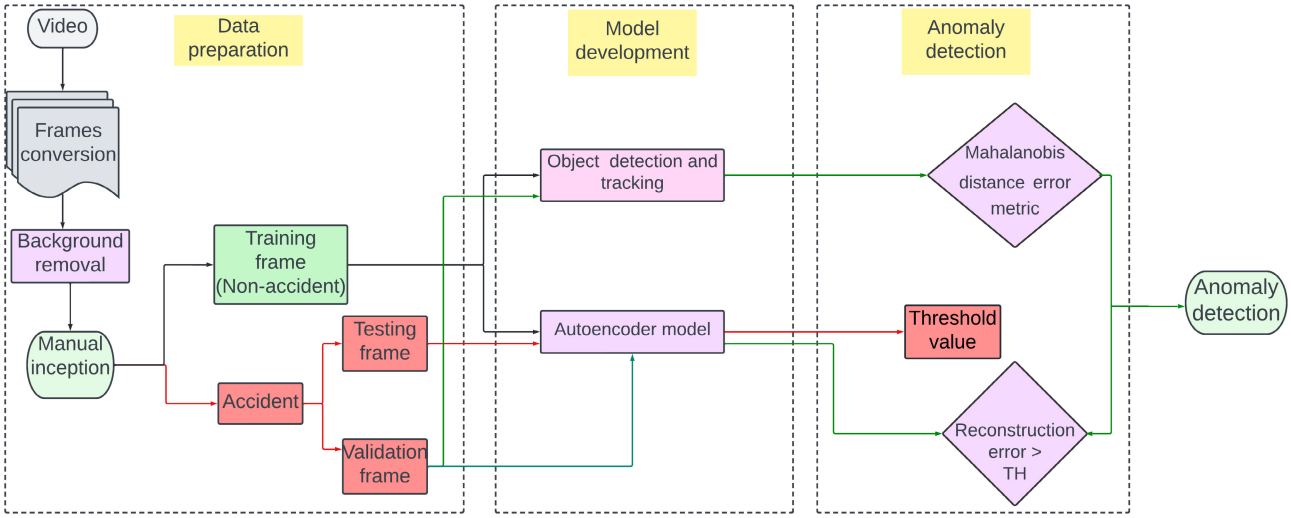

Figure 2. *Overview of the accident detection algorithm*

AD dataset with anomaly events

Data set	Abnormal events	Frames	Examples of anomalies types
UCSD - Ped1 [23]	40	14000	biker, cart, and cycle
UCSD - Ped2 [23]	12	4,560	biker, cart, and cycle
UCF-Crime [24]	13.8M	-	abuse, arrest, arson assault, burglary, explosion, fighting, and accident
Street Scene [25]	205	203,257	jaywalking, a biker on the sidewalk, skateboarder in a bike lane, biker outside the lane - pedestrian reversing direction, and a person sitting on a bench
Shanghai Tech [16]	130	317,398	chasing, brawling in a sudden motion, and wrong detection

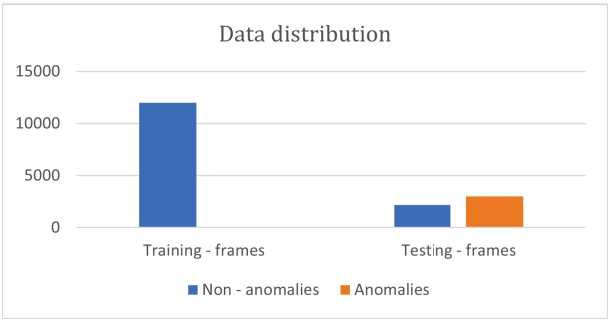

Figure 3. *Distribution of the classes of training and testing dataset*

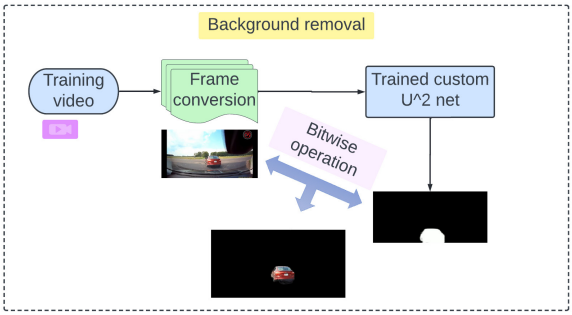

Figure 4. *Frame's background removal algorithm*

Accordingly, in the final verdict, accident videos are collected from the above datasets table 1 and prepared the anomalies dataset for our study.

Background Removal

Figure 4 shows removing the frame's background methods. The first video has been converted to frames, and the u^2 model has been trained using those frames. Afterward, the bitwise operation is performed between the output of the u^2 network and the input image to obtain the same channels as the input image.

State-of-the-Art

u-net with Skip Connection

Encoder: The u-net with SC connection architecture shows in figure 5. The model evolves without the features or data trans-

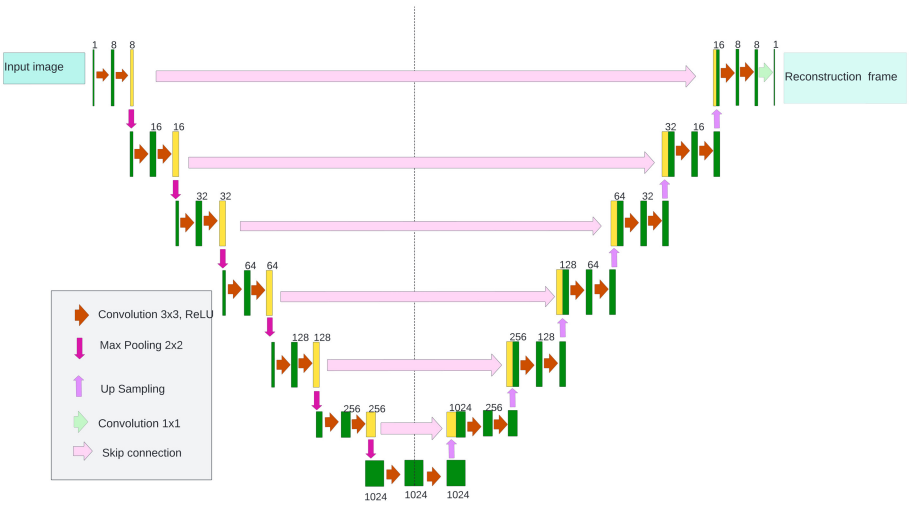

Figure 5. *u-net with skip connection*

action between the contraction and expansion phases. First, the $256 \times 256 \times 3$ size of the image considers the standard size for developing the model. For the encoder stage, six blocks are used as stacked. Further, each block also includes the CL, which begins by capturing the smaller feature maps using convolution techniques with a kernel size of 3×3. After that, the output sends it through batch normalization (BN), AF, dropout layer, and max pooling layer. Afterward, the output from the first layer transfer to the next CL, which has the same feature map size as the previous CL. Finally, the output of the first block sends it through and max pooling layer to the next block, where it starts from the CL and increasingly features maps and follows the previous structure.

Decoder: The last max pooling's outputs of the encoder block send it through the first block of the decoder, which is precisely the same as the final block of the encoder except for the pooling layer. Further, each block also includes the CL, which begins by capturing the larger feature maps using the convolution techniques with the kernel size of 3×3. After that, the output sends it through BN, AF, and dropout, and then the result is transferred to the next CL, which has the same feature map size as the previous CL. Finally, the output of the first block sends through upsampling layers to the next block, where it starts from the CL and decreasingly features maps and follows the previous structure. Features are transferred between the encoder and the decoder, as shown in the model in figure 5. The output of each max pooling of the encoder is concatenated with the CL of the decoder block. SC ensures maximum information flow between network layers, which is achieved by directly connecting the encoder's max-pooling layer with an upsampling layer of the decoder. In addition, SC shifts features from the encoder path to the decoder path to recover the spatial information lost during max sampling.

u-net with Residual Connection

Figure 6 shows the u-net skip connections with the residual connections. As shown in the figure, the residual connections are added to the encoder-decoder block between the input and CNN. Only a fraction of the input is passed via the first CNN; the remainder is added to the 2nd CNN's output of the 1st encoder block. The residual connection has the following advantages: i) allows the creation of a deeper network without the problem of gradient disappearance or explosion; ii) facilitates the training of the network; iii) allows a better flow of information between the different layers, which helps to flow of information between different layers during back propagation.

Results

The proposed AD algorithm is trained using Google-Colab GPU with Tesla T4 power with NVIDIA-SMI 460.32.03 in Tensorflow API.

Background Removal

After categorizing the image into anomalies and non-anomalies, the u^2 model was trained using a custom image to eliminate the background. The masked image is obtained as output from the u^2 model. Finally, The bitwise-AND operator was applied between the input frame and the masked image to obtain the RGB image without background, as shown in figure 7.

Object Detection and Tracking

Figure 8 shows the OT output from DeepSort. In some images, DeepSort cannot track the objects correctly. For example, even if two objects are visible in the frame, it sometimes detects both as one object. Also, DeepSort occasionally has problems accurately tracking objects in an image. For instance, DeepSort has a problem tracking the initial object as a new object, regardless of the event before or after an accident. However, this error contributes to determining anomalies.

u-net with Skip Connection

Figure 5 shows the final architecture of the u-net with an SC. While training the model, a few hyperparameters, such as learning rate (LR), feature maps, and optimizer, are essential to define. LR-scheduler techniques were performed to find out the optimal LR. Traverse a set of LR values starting from 1e-4, increasing by $10^{\frac{epoch}{20}}$ every epoch, as shown in figure 9, and LR exponentially increases as the number of epochs increases.

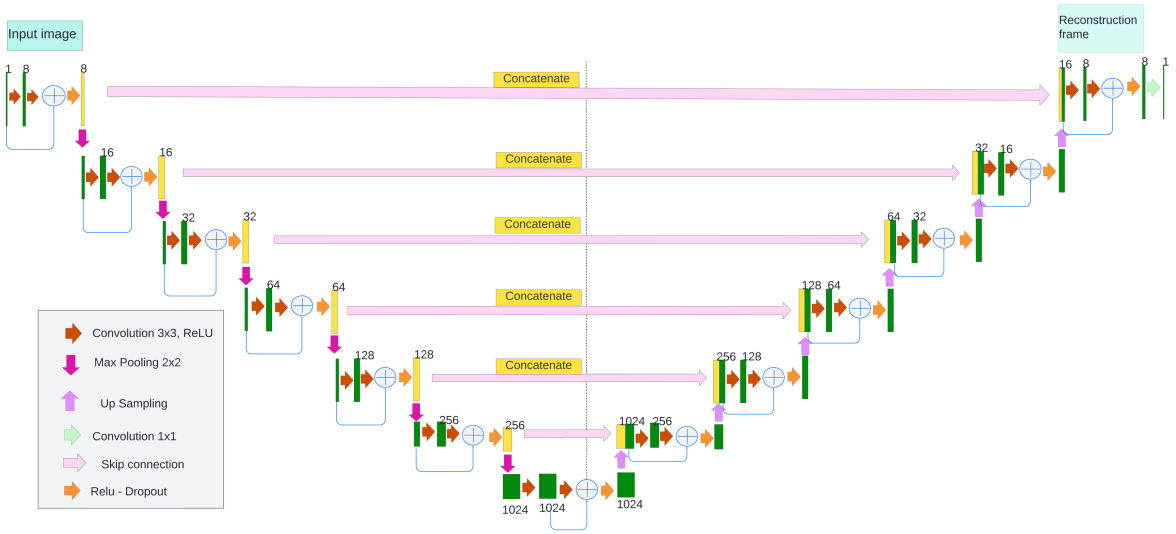

Figure 6. *u-net with skip connection and residual connection*

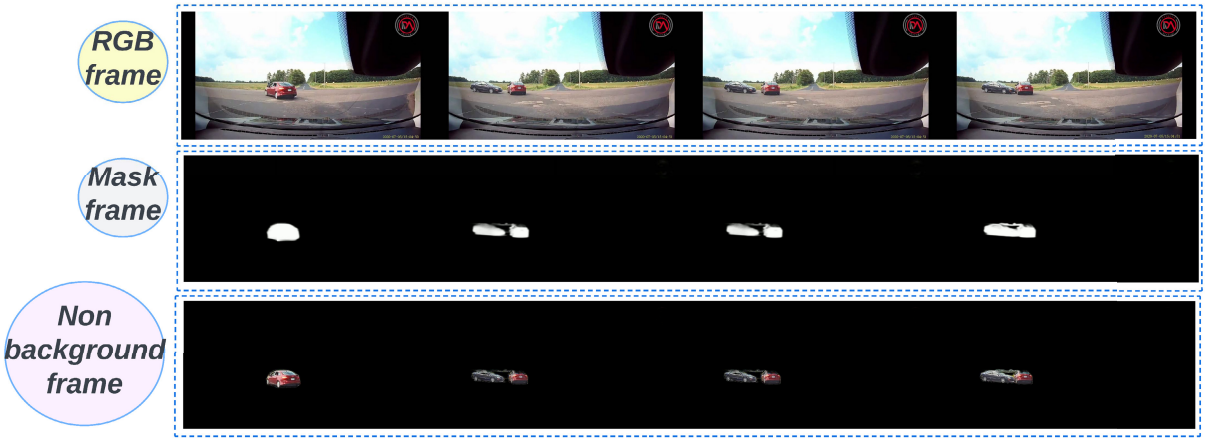

Figure 7. *Background removal output after applying the bitwise operation*

Figure 10 illustrates the LR behavior based on the accuracy and loss. The model's accuracy determined the LR's lowest and highest threshold values during the learning phase. Finally, optimist LR was calculated by averaging the range of the LR. After training the u-net model with the SC model, the new non-anomalies images were reconstructed using the model. Each frame's RCE value was plotted as normal distribution to find the threshold value. Further, The OT results are combined with the u-net model to improve the model's performance and resolve missing behaviors. Figure 11 shows the Mahalanobis distance error metric and the RCE. The RCE value exceeds the threshold several times to ensure that the Mahalanobis distance error metric is combined with the RCE. In addition, If both errors exceed the threshold, the frame is considered an accident.

Figure 12 depicts the confusion metric for the u-net with the SC model, which was combined with the Mahalanobis distance error metric. The analysis revealed that the model accurately predicted 71% of actual accidents framed as accident frames but misclassified 2% of non-accident frames as accident frames. The model's precision value was approximately 93.74%, and its recall value was 71.43%. Overall, the model's accuracy was 81.32%.

u-net skip and Residual Connection

To address the misbehavior of the prior model, a new variation called residual connection was added to the existing architecture, as shown in figure 6. The residual connection in the model addresses the prior model's lack of temporal characteristics while reconstructing the frame.

Figure 13 shows the error analysis of the u-net with the SC and residual connection model; this model falsely predicts a few accident frames as non-accidents. However, this model is performing better than the previous model. Additionally, the OT output is combined with the u-net model to enhance the performance and address the model's missing behavior.

Figure 15 shows the confusion matrix for the u-net with the SC model after combining it with the Mahalanobis distance error metric. The model correctly predicts:

- 81.25% of actual accidents framed as accident frames,
- 3% of actual non-accident frames as an accident frame,
- model precision value is around 86.64%

Figure 8. *OT output from the DeepSort*

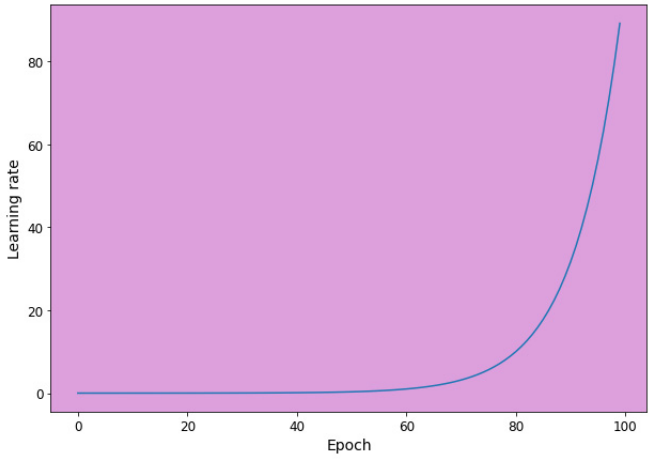

Figure 9. *Range of the LR to find the optimal value*

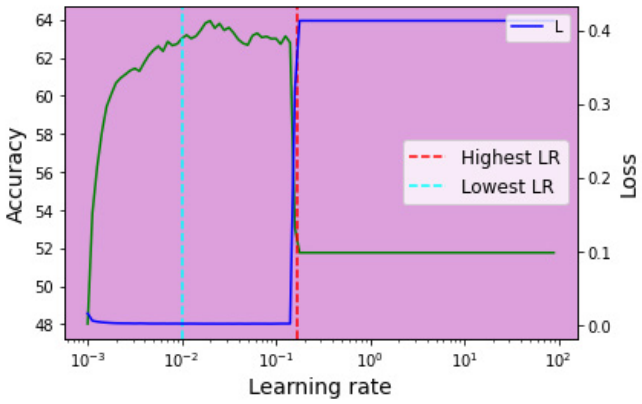

Figure 10. *Finding the optimal LR using the LR scheduler method*

- recall value is 81.43%
- model F1 score is 83.32%

Overall model performance has increased by 3%, and precision and recall values are increased significantly. However, the model still predicted 18% of actual accident frames as non-accident frames. The attention connection will be added to the model to improve the performance.

Comparison of the u-net Results using Confusion Metric

The final results are similar for the u-net combined with SC plus attention plus OT algorithms and the u-net combined with SC plus residual attention connection plus OT algorithms. However, the u-net with the SC plus attention - residual plus OT algorithms obtained higher F1 scores and accuracy values. Furthermore, this model still predicts incident frames as non-incident frames because defining an incident's start and end frames is difficult.

Conclusion and Outlook

AD is crucial in surveillance videos based on anomaly events, such as accidents, abuse, and kidnaps. Likewise, many researchers have studied AD using ML and CV methods for decades. Still, it remains one of the biggest challenges due to the different variants of anomalies events, occlusions, the number of objects in the frame, and the moment of these objects. Surveillance video is essential in various domains, such as law enforce-

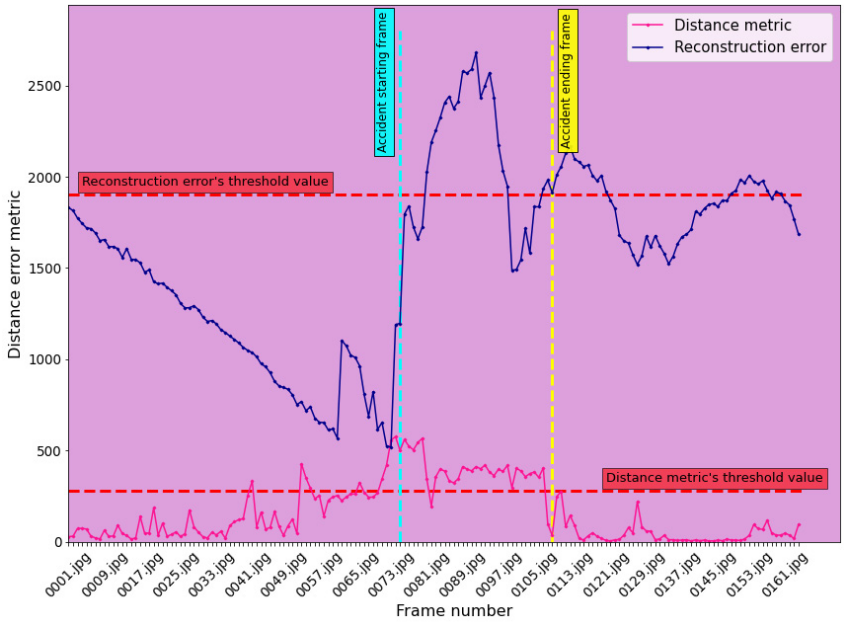

Figure 11. *Representation of the reconstruction error with Mahalanobis distance error metric*

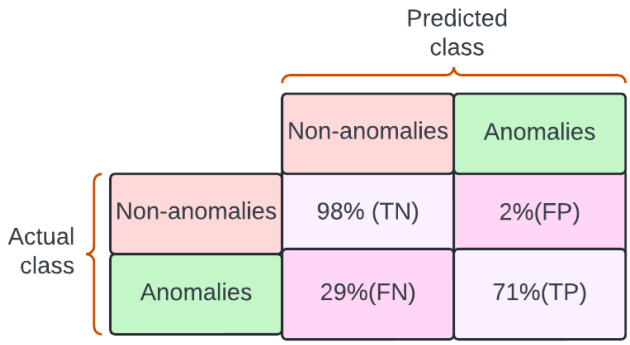

Figure 12. *Confusion metric of the u-net with skip connection algorithm after being combined with Mahalanobis distance error metric*

ment, transportation, and environmental monitoring, to improve security and public safety. Therefore, the road surveillance system will find anomalies to improve road safety, analyzing traffic accidents, traffic signs, speed limits, and license plates. The proposed algorithms in this study can be used to detect road accidents in various scenarios. The proposed method takes into account the following ideas: i) background removal - avoiding background variants and unwanted objects in the frame, ii) object detection and tracking - identifying which accident occurred and tracking the object, for example, the car with a human or car with motorbike or car with human, iii) image reconstruction – the u-net used as the main backbone of the algorithm with different variants such as i) skip connections - transmitting information about temporal features while reconstructing the frame, ii) attention connection - focusing on the main temporal features when transmitting from the encoder block to the decoder block, and iii) residual connections - avoiding vanishing/exploding gradient problems.

This study shows that removing the background allows images to be reconstructed with fewer errors because images contain the exact variants after removing the background. In addition, DeepSort can track objects such as cars, bikes, humans, trucks, and motorbikes at 30 frames per second. Besides, u-net reconstructs frames with fewer errors after adding functionality such as skip, residual, and attention connections. Moreover, DeepSort and u-net with residual and attention connection achieved the highest accuracy.

This study contributes to the existing body of knowledge in the area of AD by providing i) the effectiveness of BR methods for frame reconstruction and OT, ii) the identification of accident scenes with vehicle classes, iii) the effect of the without SC, with SC, attention connection, and residual connection with attention connection when reconstructing frames using u-net and iv) the use of DeepSort's Mahalanobis error metric and u-net's RCE with different variants to detect anomalies. In that respect, the study provides a systematic approach for designing the automation of the Accident detection algorithm, which detects accident vehicle classes.

Furthermore, our solution could be improved by: i) adding recurrent functionality to the u-net model, which will enhance the relationship between each layer of the input and output, ii) replacing DeepSort with recurrent techniques that can be added to OT algorithms, iii) training the model under different weather and daylight conditions.

REFERENCES

[1] Chandola, V., Banerjee, A., and Kumar, V., "Anomaly detection: A survey," *ACM computing surveys (CSUR)* **41**(3), 1–58 (2009).

[2] Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., and Ahmad, F., "Network intrusion detection system: A systematic study of machine learning and deep learning ap-

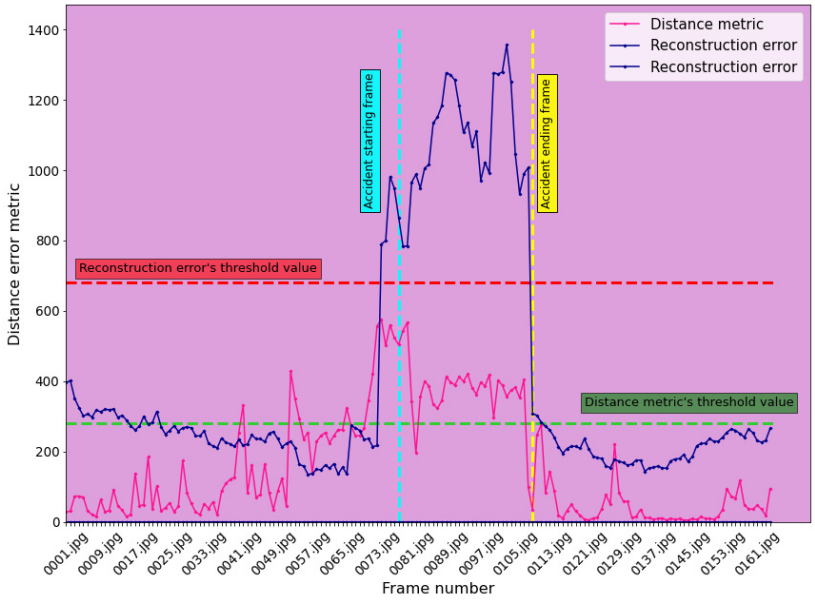

Figure 13. *Reconstruction error analysis for the AD*

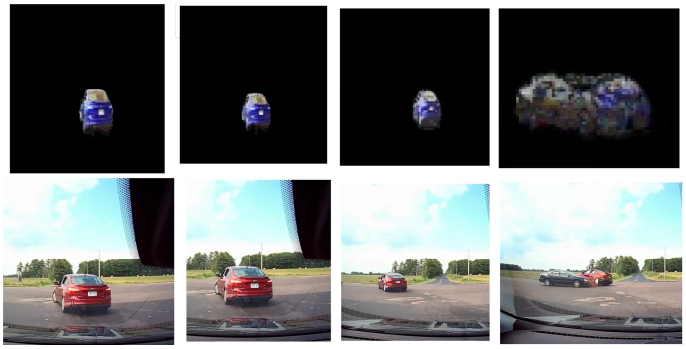

Figure 14. *Example of the reconstruction frame by the model*

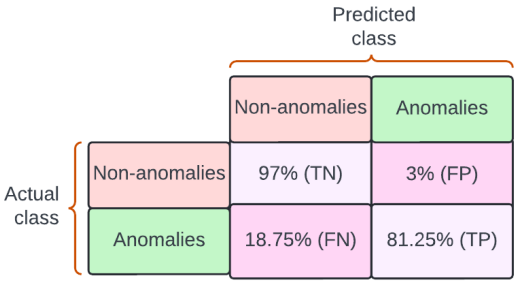

Figure 15. *Confusion metric of the u-net with skip connection SC algorithm after being combined with Mahalanobis distance error metric*

proaches," *Transactions on Emerging Telecommunications Technologies* **32**(1), e4150 (2021).

[3] Maniraj, S., Saini, A., Ahmed, S., and Sarkar, S., "Credit card fraud detection using machine learning and data science," *International Journal of Engineering Research* **8**(9), 110–115 (2019).

[4] Fernandes, M., Corchado, J. M., and Marreiros, G., "Machine learning techniques applied to mechanical fault diagnosis and fault prognosis in the context of real industrial manufacturing use-cases: a systematic literature review," *Applied Intelligence* , 1–35 (2022).

[5] Wang, X., Ding, H., Gu, X., Yuan, J., and Shen, Q., "Study of traffic incident detection with machine learning methods," in [*Education and Awareness of Sustainability: Proceedings of the 3rd Eurasian Conference on Educational Innovation 2020 (ECEI 2020)*], 725–728, World Scientific (2020).

[6] Hospedales, T., Gong, S., and Xiang, T., "Video behaviour mining using a dynamic topic model," *International journal of computer vision* **98**(3), 303–323 (2012).

[7] Nanni, L., Ghidoni, S., and Brahnam, S., "Handcrafted vs. non-handcrafted features for computer vision classification," *Pattern Recognition* **71**, 158–172 (2017).

[8] Adam, A., Rivlin, E., Shimshoni, I., and Reinitz, D., "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE transactions on pattern analysis and machine intelligence* **30**(3), 555–560 (2008).

[9] Cong, Y., Yuan, J., and Liu, J., "Sparse reconstruction cost for abnormal event detection," in [*CVPR 2011*], 3449–3456,

IEEE (2011).
[10] Kim, J. and Grauman, K., "Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates," in [*2009 IEEE conference on computer vision and pattern recognition*], 2921–2928, IEEE (2009).
[11] Zhao, B., Fei-Fei, L., and Xing, E. P., "Online detection of unusual events in videos via dynamic sparse coding," in [*CVPR 2011*], 3313–3320, IEEE (2011).
[12] Tung, F., Zelek, J. S., and Clausi, D. A., "Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance," *Image and Vision Computing* **29**(4), 230–240 (2011).
[13] Reddy, V., Sanderson, C., and Lovell, B. C., "Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture," in [*CVPR 2011 WORKSHOPS*], 55–61, IEEE (2011).
[14] Siemenn, A. E., Shaulsky, E., Beveridge, M., Buonassisi, T., Hashmi, S. M., and Drori, I., "A machine learning and computer vision approach to rapidly optimize multiscale droplet generation," *ACS Applied Materials & Interfaces* **14**(3), 4668–4679 (2022).
[15] Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., and Hua, X.-S., "Spatio-temporal autoencoder for video anomaly detection," in [*Proceedings of the 25th ACM international conference on Multimedia*], 1933–1941 (2017).
[16] Liu, W., Luo, W., Lian, D., and Gao, S., "Future frame prediction for anomaly detection–a new baseline," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 6536–6545 (2018).
[17] Lu, Y., Kumar, K. M., shahabeddin Nabavi, S., and Wang, Y., "Future frame prediction using convolutional vrnn for anomaly detection," in [*2019 16Th IEEE international conference on advanced video and signal based surveillance (AVSS)*], 1–8, IEEE (2019).
[18] Doshi, K. and Yilmaz, Y., "Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate," *Pattern Recognition* **114**, 107865 (2021).
[19] Qiang, Y., Fei, S., Jiao, Y., and Li, L., "Anomaly detection of predicted frames based on u-net feature vector reconstruction," in [*Journal of Physics: Conference Series*], **1627**(1), 012014, IOP Publishing (2020).
[20] Di Mattia, F., Galeone, P., De Simoni, M., and Ghelfi, E., "A survey on gans for anomaly detection," *arXiv preprint arXiv:1906.11632* (2019).
[21] Yang, J., Cai, Y., Liu, D., and Xie, J., "3d u-net for video anomaly detection," in [*Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering*], 1640–1645 (2021).
[22] Kim, Y., Yu, J.-Y., Lee, E., and Kim, Y.-G., "Video anomaly detection using cross u-net and cascade sliding window," *Journal of King Saud University-Computer and Information Sciences* (2022).
[23] Mahadevan, V., Li, W., Bhalodia, V., and Vasconcelos, N., "Anomaly detection in crowded scenes," in [*2010 IEEE computer society conference on computer vision and pattern recognition*], 1975–1981, IEEE (2010).
[24] Sultani, W., Chen, C., and Shah, M., "Real-world anomaly detection in surveillance videos," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 6479–6488 (2018).
[25] Ramachandra, B. and Jones, M., "Street scene: A new dataset and evaluation protocol for video anomaly detection," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*], (March 2020).

Author Biography

Kirsnaragavan Arudpiragasam holds a Master's in Engineering and Sustainable Technology Management from SRH Berlin University of Applied Sciences, where he focused on the Mobility and Automotive Industry. He previously earned his Bachelor's degree in Manufacturing and Industrial Engineering in 2019. His research interests are computer vision, deep learning, multimodal learning, and ethical considerations related to autonomous driving.

Taraka Rama Krishna Kanth Kannuri is pursuing a Master's in Engineering and Sustainable Technology Management, focusing on the Mobility and Automotive industry at SRH Berlin University of Applied Sciences. He received his Bachelor's in Mechanical Engineering from Koneru Lakshmaiah University, Vijayawada, India, in 2018. His research interests include autonomous driving, computer vision, ethical decision, and deep learning.

Klaus Schwarz received his B.Sc. and M.Sc. in Computer Science from Brandenburg University of Applied Sciences (Germany) in 2017 and 2020, respectively. He is currently a Ph.D. student at the University of Granada, Spain. His research interests include IoT and smart home security, OSINT, mechatronics, additive manufacturing, embedded systems, artificial intelligence, and cloud security. As a faculty member, he is developing a graduate program in Applied Mechatronic Systems focusing on Embedded Systems at SRH Berlin University of Applied Sciences.

Reiner Creutzburg is a Retired Professor for Applied Computer Science at the Technische Hochschule Brandenburg in Brandenburg, Germany. Since 2019 he has been a Professor of IT Security at the SRH Berlin University of Applied Sciences, Berlin School of Technology. He has been a member of the IEEE and SPIE and chairman of the Multimedia on Mobile Devices (MOBMU) Conference at the Electronic Imaging conferences since 2005. In 2019, he was elected a member of the Leibniz Society of Sciences to Berlin e.V. His research interest is focused on Cybersecurity, Digital Forensics, Open Source Intelligence (OSINT), Multimedia Signal Processing, eLearning, Parallel Memory Architectures, and Modern Digital Media and Imaging Applications.