applied sciences

MDPI

Article

Kalman Filtering and Bipartite Matching Based Super-Chained Tracker Model for Online Multi Object Tracking in Video Sequences

Shahzad Ahmad Qureshi [1], Lal Hussain [2,3], Qurat-ul-ain Chaudhary [1], Syed Rahat Abbas [1], Raja Junaid Khan [4], Amjad Ali [5] and Ala Al-Fuqaha [5,*]

[1] Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad 45650, Pakistan
[2] Department of Computer Science and Information Technology, King Abdullah Campus Chatter Kalas, University of Azad Jammu and Kashmir, Muzaffarabad 13100, Pakistan
[3] Department of Computer Science and Information Technology, Neelum Campus, University of Azad Jammu and Kashmir, Athmuqam 13230, Pakistan
[4] Department of Computing, Islamabad Campus, Iqra University, Islamabad 44000, Pakistan
[5] Information and Computing Technology (ICT) Division, College of Science and Engineering (CSE), Hamad Bin Khalifa University (HBKU), Doha P.O. Box 34110, Qatar
* Correspondence: aalfuqaha@hbku.edu.qa

check for updates

Citation: Qureshi, S.A.; Hussain, L.; Chaudhary, Q.-u.-a.; Abbas, S.R.; Khan, R.J.; Ali, A.; Al-Fuqaha, A. Kalman Filtering and Bipartite Matching Based Super-Chained Tracker Model for Online Multi Object Tracking in Video Sequences. *Appl. Sci.* **2022**, *12*, 9538. https://doi.org/10.3390/app12199538

Academic Editor: Dariusz Frejlichowski

Received: 11 August 2022
Accepted: 9 September 2022
Published: 23 September 2022

Abstract: Object tracking has gained importance in various applications especially in traffic monitoring, surveillance and security, people tracking, etc. Previous methods of multiobject tracking (MOT) carry out detections and perform object tracking. Although not optimal, these frameworks perform the detection and association of objects with feature extraction separately. In this article, we have proposed a Super Chained Tracker (SCT) model, which is convenient and online and provides better results when compared with existing MOT methods. The proposed model comprises subtasks, object detection, feature manipulation, and using representation learning into one end-to-end solution. It takes adjacent frames as input, converting each frame into bounding boxes' pairs and chaining them up with Intersection over Union (IoU), Kalman filtering, and bipartite matching. Attention is made by object attention, which is in paired box regression branch, caused by the module of object detection, and a module of ID verification creates identity attention. The detections from these branches are linked together by IoU matching, Kalman filtering, and bipartite matching. This makes our SCT speedy, simple, and effective enough to achieve a Multiobject Tracking Accuracy (MOTA) of 68.4% and Identity F1 (IDF1) of 64.3% on the MOT16 dataset. We have studied existing tracking techniques and analyzed their performance in this work. We have achieved more qualitative and quantitative tracking results than other existing techniques with relatively improved margins.

Keywords: object tracking; object detection; MOTA; real-time multiobject tracking

1. Introduction

Object video tracking is important for many applications for cohorts working in the computer vision arena, including video and security surveillance, traffic control, video communication, robotics and animation, video sensor targeting, robotics, human computing, etc. Several object-tracking methods are used in video sequencing to identify moving objects. Multiple Object Tracking (MOT), based on computer vision, has many important implications related to object recognition, belonging to many categories, like chairs, glasses, cars, and pedestrians, and its tracking without its posterior details about the shape characteristics, including the objects' count [1]. MOT is a very complicated task that is difficult to solve; therefore, we need to develop robust methods for object detection [2]. A gap exists to accommodate smooth but reasonable tracklets by reducing ID switches, assigning IDs to wrong objects, and using the tracked trajectories. Keeping only the neighboring frames and consequently raising the speed of object tracking ensures robustness by evaluating the unknown position of the target. The minimum mapping criterion should be selected to raise the speed of OT by avoiding misleading tracklets followed by an object recognition system.

Convolutional Neural Networks (CNN) are state-of-the-art in spatial pattern extraction and useful in various tasks like image classification [3–5] and its positional details. Object recognition aims to describe a series of related vision tasks involving activities such as the identification of digital photographic objects. Classifying images requires tasks such as predicting an object's class in an image. Object position refers to the location of an image and the drawing of a box around the extent of one or more objects, and it works by integrating these two tasks by finding one or more objects in an image. The goal of tracking objects in consecutive video frames is the association of target objects. The object's shape, location, and position in the video frames are needed for object tracking.

Object detection and classification in computer vision systems are the most important phases in object tracking [6–8]. Object detection is the first and most important tracking step to identify or locate the moving object in a picture. Subsequently, objects detected could be categorized in terms of cars, persons, swaying trees, birds, and others. Image processing techniques make monitoring objects in consecutive frames difficult or challenging. Different problems may occur due to complex object motion, irregular object form, object-to-object occlusion, and processing requirements in real time. Tracking video objects is an important computer vision research subject. Many algorithms for object tracking have been proposed. However, there are still various problems with tracking objects in the images, such as a change in illumination, occlusion, movement blurring, change in size, change in scale, appearance changes, camera movement, confusing scenes, and so on. Similarly, the interval between two adjacent frames affects target tracking by increasing the inherent frame rate at which the video sequence is acquired and provides extra time for more information in the form of a frame, leading to the loss of information in case the lower frame rate is maintained.

In this article, the following aspects are discussed:

1. The proposed method, the SCT model, based on Kalman filtering and bipartite matching, is presented. It is an online MOT model to optimize feature extraction, object detection, and data association. It performs data association to object detection (pairwise).
2. The informative regions are improved using box pair regression of SCT with the help of a joint attention unit to enhance performance.
3. The proposed SCT reaches remarkable performance with datasets of MOT16 and MOT17.

The paper is organized as follows; Section 2: Literature Review, Section 3: Materials and Methods, and Section 4: Results and Discussion, followed by Conclusions. The nomenclature of the terminology used in this article has been given in Table A1.

2. Literature Review

Considering the subject's importance, many cohorts across the globe have worked on object tracking algorithms and used them in different applications. This section covers the prominent findings of research activities carried out in the relevant field. Tangirala and Ramesh [9] used a particle filtering approach for monitoring objects in a video sequence. They used six parameters model for modeling motion edge. They determined the velocity of the moving edge using spatio-temporal filtering techniques. The multi-dimensional posterior density was used to distribute the filtering of particles over time. They proposed using sampling and sample impoverishment techniques to avoid particles' degeneracy leading to particle-filter convergence failure. Their study has a limitation that the current implementation is manual and involves user interaction relying on some video sequence information, including edge orientation, to collect the correct motion cube. Similarly, Rehman et al. [10] proposed an online surveillance algorithm using particle filtering. Their methodology was that a particle filtering-based framework developed for online anomaly detection finds video frames with anomalous activities from the posterior probability of activities present in a video sequence. They proposed that the anomaly detection algorithm was better than other contemporary algorithms in terms of reduced equal error rate (EER) and processing time. Tikala and Pietikainen [11] have proposed a real-time tracker system. The operational mechanism of the tracker system was based on the integrated use of color, texture, and movement information. They exploited RGB color histogram and correlogram as color cues using local binary patterns (LBP) to represent texture properties. After extraction of all features, including texture-based background subtraction, they built a unified distance measure. Their unified system worked well on indoor and outdoor control videos compared to a single feature-based system. The proposed system performed satisfactorily under the low lighting and low frame rates conditions, typically in large-scale monitoring systems. The use of a flexible set of indicators makes it less color sensitive. Their proposed system showed robust performance. Yang et al. [12] presented a 3D multi-object-tracking algorithm by establishing the track-to-detection correspondence. The findings of their study compared with challenging datasets reported that their proposed approach accomplished outclass performance. Shu [13] introduced three video control activities: firstly, he implemented a method of human detection to boost an overall human detector for a video using an unsupervised learning framework to train a particular video classifier; secondly, he defined a robust semi-crowded scene tracking by detection system to improve the overall tracking performance using a partial model to manage the appearance changes, and partial occlusions; finally, he implemented a real-time tracking device in the high-resolution video for single target tracking. He also introduced a new form of multiple human segmentation in videos that used partial detection in addition to the video's space-time detail. Gruosso et al. [14] introduced a model for segmentation of humans in videos using a deep convolutional neural network for human activity monitoring by generating a high-quality mask for human segmentation. Their study reported a better possibility of results in terms of automatically recognizing and segmenting objects in videos. Gao et al. [15] proposed a robust hybrid tracker system that combined deep features with color features forming a hybrid of features. The notion of this technique was de-noising the image using Gaussian smoothing and cropping the irrelevant areas. Zhong et al. [16] proposed a novel deblurring approach by adding residual dense blocks into recurrent neural network (RNN) cells to efficiently extract the current frame's spatial features. They also proposed a spatio-temporal module to count the past and future features. They claimed better deblurring performance with less computational cost than the competing methods. A multi-stage system based on deep CNN detection and tracking system in videos for object detection was proposed by Kang et al. [17] that constituted of two principal modules: firstly, a proposal module for tubelets that provided object detection, and the object tracking proposals for tubelets; secondly, a classification and scoring module for tubelets that provided max spatial pooling with robust box scoring, and time convolution to integrate temporal consistency. In this manner, the identification and monitoring of artifacts worked closely. On one hand, object detection provided high trust anchored to trigger tracking while simultaneously decreasing spatial tracking (max-pooling) failure. The monitoring also produced new object detection proposals, and the tracked boxes served as anchors for combining existing detections.

A spatially supervised RNN was developed by Ning et al. [18] for visual object tracking. They proposed the Recurrent you Only Look Once (ROLO) approach extended neural network research to the spatiotemporal domain. Their proposed tracker was both deep in space and time to tackle serious occlusive problems effectively with extreme motion blur. Based on a comparison with benchmark tracking datasets, the detailed experimental results showed that the proposed tracker was more robust, low computational cost, and accurate against competing methods. Similarly, Held et al. [19] proposed an offline neural network method to track new objects at 100 fps, an enormous frame rate, in tested time slots. Ma et al. [20] used a hierarchical convolutional for tracking purposes using the visual features and found improvement in the system's robustness by representation learning for object recognition datasets. Wang, et al. [21] studied object tracking using full CNN and found that different levels of convolutional layers had different properties. To capture the semantic information of the target, and distinguish the target from background distractors, they opted for an integrated approach and considered the properties jointly. Shen et al. [22] studied moving object detection in aerial video based on spatio-temporal saliency. They suggested a new method of spatiotemporal savings using the saliency technique for hierarchical movement of goal detection. The experimental results demonstrated that this method recognized high-precision and efficient moving substances in the airborne film. Furthermore, compared to the Motion History Image (MHI) technique, this method accompanied no time delay effect. Still, this approach measured the positions of the objects in any video frame as self-sufficient and unavoidable false alarms. Zhang et al. [23] introduced computation on unclear areas rejecting useless information following computational efficiency, namely coarse-to-fine object localization with object identification for Unmanned Armed Vehicles (UAV) imagery using lightweight CNN. This frame selection philosophy detects the objects coarsely, increasing detection speed and accuracy. They generated a deep motion saliency map for UAVs using LiteFlowNet, a lightweight CNN model, for refining the detection results [24] and extracted deep features through PeleeNets [25]. Guo et al. [26] proposed a method for object detection for video frame tracking. The simulation result showed that this technique was efficient, precise, and robust for good performance of generic object class detection. Ayed et al. [27] suggested a method for text data detection based on big data analytics texturing in video frames. The video frames were decomposed into blocks (fixed size) and analyzed by Haar wavelet transformations followed by training of a neural network to distinguish blocks of text from non-text. Viswanath et al. [28] proposed a non-panoramic background modeling technique that modeled each pixel with a single spatio-temporal Gaussian model. The result of the simulations showed that moving the substances was detected with fewer false alarms. This method suffered a serious drawback of insufficient features from the scene. Soundrapandiyan et al. [29] proposed an adaptive pedestrian detection system using image pixel intensities where the foreground objects were isolated from the background. They used a high boost filter to improve the front edges. Similarly, YOLO algorithm park [30] is a real time object tracking method that is known for its speed and accuracy. It is subjected to low recall with high localization error due to insufficient positioning of bounding boxes, and it struggles to detect close or small objects. This problem can be attributed to instability to detect multiple objects that are too close.

Tracktor [31] carried out the temporal realignment of the bounding boxes of the object for multiobject tracking manipulating its regression head for the detector. Dual Matching Attention Networks (DMAN) [32] are based on two attention mechanisms: namely spatial- and temporal types, where the former, creating dual attention maps, uses network focus for image-pair (derived through input) by the matching patterns, while the later deemphasizes the noisy details by allocating multi-level attention to samples in the tracklet. Some cohorts worked on the correlation among targets, and occlusion-based drift between them was alleviated using the Spatial-Temporal Attention Mechanism (STAM) [33]. The resultant target map was parameterized, leading to the attention map on a spatial basis achieving MOTA as 46% on the challenging dataset of MOT16. Choi [34] introduced Near-Online Multi-target Tracking (NOMT) framework by computing data covariance sequentially between objects detected in a time window and target frames. The detection due to representation-based learning with high performance resulted in efficient MOT in offline and online systems [35]. Sanchez-Matilla et al. [36] introduced a Person of Interest (POI) tracker for an online solution, while the K-Dense Neighbors Tracker (KDNT) for its offline counterpart. The targets and labels are based on the relationship between object localization-path after its prediction. The Early Association Multi Target Tracker (EAMTT) highlights the weak features (only for the existing tracks where detections have lower confidence levels) as well as strong detections (with higher confidence levels) for tracking as well as its initialization. Similarly, Wojke et al. [37] introduced Simple Online and Real-time Tracking (SORT) as an extension to the previous work. They used transfer learning-based association merging it as appearance-features information, thereby allowing stranded occlusion bursts reducing identity switch cardinality. Mahmoudi et al. [38] pointed out multi-target tracking based on CNN (CNNMTT) used frames at (t-1 interval) group-wise for describing the useful cross-correlation between discriminative appearance-features working as a good grouping technique for online $\mathbb{R}^2$-environments without any prior camera calibration information.

Recently, a unified tracking graph representation was proposed that combined detections and tracks in one graph and improved tracking performance by providing an online method of tracking objects in 3D [39]. Another method, Motion-Aware Tracker (MAT), was introduced by Han et al. [40] as a plug-and-play way out with the sole objective of motion-based prediction, including association and reconnection. Tracking-by-detection (TBD) [41] was proposed, which localized the pedestrian in each frame and connected object hypotheses into the trajectories without initial labeling. These methods are mostly dependent on the accuracy and data association of objects. Another framework, namely CenterPoint, was proposed by Yin et al. [42], that used a keypoint detector in the initial stage to locate centers of objects with attributes like 3-D size, 3-D orientation, and velocity. Secondly, they added some additional point features to improve estimates.

The proposed model, SCT, can take two adjacent frames as an input at a time, also known as a chain node. This online MOT system is different and more efficient than the other MOT systems, which only take one frame at a time as input. In this method, we have applied ResNet50 for feature extraction, which extracted all the semantic features. Feature Pyramid Networks (FPN) [43] have been used to generate features' multiscale representation for subsequent predictions. Furthermore, IoU matching, Kalman filter, and bipartite matching are used for sequencing two nodes on a shared common frame.

3. Materials and Methods

Video scene comprehension and analysis of human conduct are critical. Computer vision has high-level functions with enormous real-scene applications depending on many other tasks, which are critical for multiobject tracking. However, MOT remains a major challenge, particularly for crowded scenes, because of occlusions, including overlapping items with difficult backgrounds.

3.1. Problems

Despite great efforts made in the last few years and supporting improvement, the MOT solutions have two big issues. One is that these systems are based on plausible but sub-optimal tracking models [44] that result in the infeasibility of overall optimization in an end-to-end manner. It typically involves three main divisions: the detection of objects, the extraction of features, and the association of data. However, dividing the entire task into separate subtasks resulted in getting local optimum and was helpful in performing calculations. Furthermore, the association of data was dependent on the consistency of the objects detected that couldn't deliver accurate and consistent results in different frames as it rejected the adjacent frames having temporal relationships. The other problem is that issues are arising for boosting efficiencies of multiobject tracking; therefore, two important points; reidentification and attention, are found to enhance MOT efficiency. Reidentity extracts robust data association features (or ID verification). Attention tends to concentrate on the model more, avoiding interruption by irrelevant details, like the background. Regardless of productivity, the presence of existing solutions causes an increase in the complexity of the model and machine costs considerably. We have proposed a new method for online tracking, namely the Super Chained-Tracker (SCT), that defines an end-to-end model incorporating feature extraction with data association based on object detection. Our proposed model is convenient, simpler, and more efficient than existing MOT detection techniques. In a single regression model, the joint detection with tracking is carried out simultaneously using the adjacent pairs of frames that show matched bounding boxes as the goal. We have added a joint attention module to increase the efficiency of our SCT using expected confidence maps. It directs the regression branch for paired cases to concentrate on insightful spatial areas with two other branches. Firstly, the branch of object classification that provides the appropriate values for the box in pairs appears first in the detected cases. Scores from this are then used to direct the regression industry to concentrate on the region of interest. The ID verification branch focuses on forecasting the regression branch to highlight the regions that meet the target. In the last stage, the bounding boxes are classified, and the association of generated box pairs takes place using methods such as Intersection over Union (IoU) [45], Kalman filter, and Bipartite Matching from the adjacent framework pairs with their boxes in the common framework. Thus, chaining of all neighboring frame pairs may be used to ensure the tracking operation.

3.2. Problem Formulation

With an image sequence of frames "N," MOT aims to decipher all the bounding boxes with identity labels for all the interesting objects in all frames, including their corresponding locations. For the input image sequence $\{F_t\}_{i=1}^{N}$, here $F_t \in \mathbb{R}^{c\times w\times h}$ representing the tth frame, and the output would be all bounding boxes represented as $\{B_t\}_{i=1}^{N}$, whereas labels are represented as $\{L_t^{GT}\}_{i=1}^{N}$, for all the relevant objects present in a frame. Moreover, B_t is defined as a subset of $\mathbb{R}^4$ indicating the ground truth bounding boxes of the number of targets (K_t) in the tth frame and L_t^{GT} is defined as the subset of $\mathbb{Z}$, which indicates their identities/labels. The MOT is divided into multiple processes, namely object detection using feature extraction process with data association. However, the researchers revealed that association efficiency relies on detection performance [46], thereby indicating its importance among the three procedures. To better exploit their correlation, the proposed Super Chained-Tracker (SCT) uses a single network mechanism.

3.3. Chained-Tracker Pipeline

3.3.1. Framework

The proposed SCT model, which is capable of taking two adjacent frames as an input at a time, also known as a chain node, is more efficient than the other MOT systems that only take one frame at a time as an input. In our system, the initial chain node is represented by (F_1, F_2), whereas the last one is represented by the (F_N, F_{N+1}). We are taking the copy of frame F_N as $F_{N+1,}$ as we know F_N is the last frame. The input node is given as (F_{t-1}, F_t); SCT can generate pair of bounding boxes [9] provided that the targets are the same in the initial and last frames where it generates bounding box pair $[((B_{t-1}^{i}, \hat{B}_t^{i}))]_{i=1}^{n_{t-1}}$, where n_{t-1} is represented as the pairs' count, the two bounding boxes $[B_{t-1}^{i} \in B_{t-1} \subset \mathbb{R}^4$ and $\hat{B}_t^{i} \in B_t \subset \mathbb{R}^4]$ are representing the same target. Therefore, we can get the bounding boxes defined as $[((B_{t-1}^{i}, \hat{B}_t^{i}))]_{i=1}^{n_{t-1}}$ in the next node (F_t, F_{t+1}). In Figure 1, it is illustrated that the detected boxes $\hat{B}_t^{i}$ and B_t^{j} are defined as the same target in a common frame of adjacent nodes, with a slight difference of two boxes. The existing MOT techniques having appearance features of complex nature can be avoided by involving simpler and more efficient matching strategies to perform chaining of the two bounding boxes. With the help of chaining nodes in a sequential manner, we can get all the detected targets by using long paths [4].

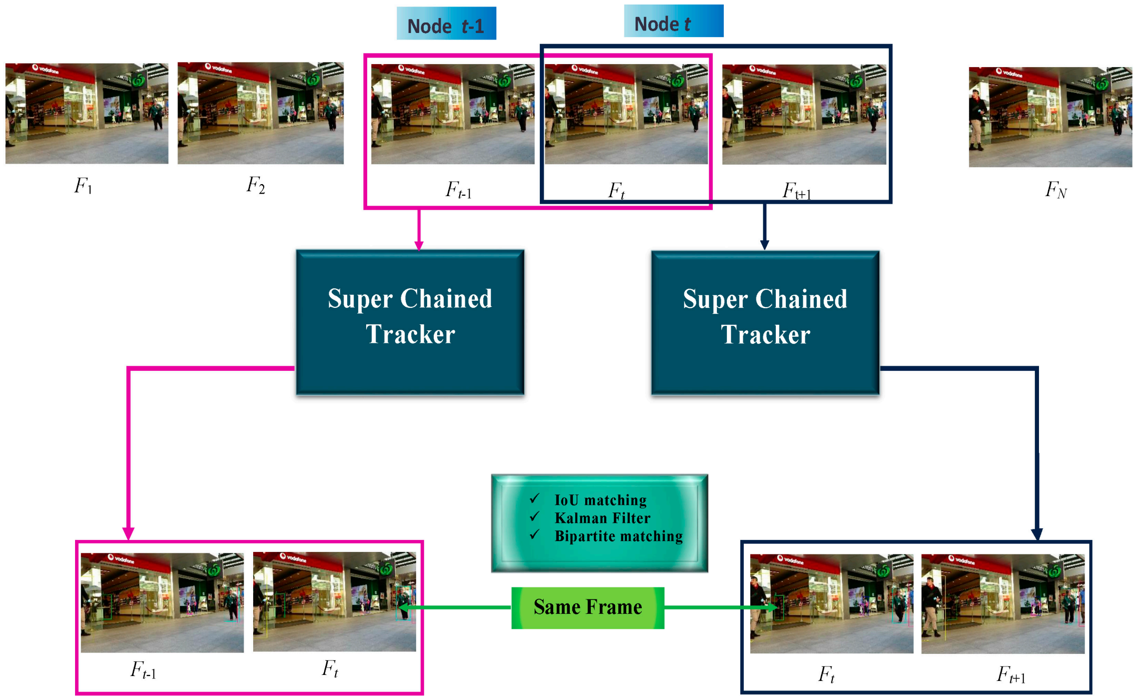


Figure 1. Schematic details of the chaining node: the pair of bounding boxes denoted by $(B_{t-1},\ \hat{B}_t)$ generated, and after that, SCT exploits two random adjacent nodes represented by (F_{t-1}, F_t) and (F_t, F_{t+1}). Subsequently, IoU matching, Kalman filtering, and Bipartite matching are used on the common frame for sequencing the two nodes. The long trajectories are generated for the video sequence by using adjacent nodes, and chaining is performed sequentially.

3.3.2. Node Chaining

For convenience, we have used (B_{t-1}) to represent $[((B^i_{t-1},\ \hat{B}^i_t))\ ]^{n_{t-1}}_{i=1}$. Two important steps for chaining can be highlighted. Firstly, we have nodes with identity tracks randomly assigned to each of them, and hence, the bounding boxes $B^i_1 \in \mathrm{B}_1$ are detected during the initialization phase. Secondly, the two nodes $(F_{t\ -1}, F_t)$ and $(F_t, F_{t\ +\ 1})$ are chained for any node t provided the condition of adjacency is preserved. The IoU, Kalman filter, and Bipartite matching between the bounding boxes in B_t and $\hat{B}_t$ are calculated as illustrated in Figure 1 where $\hat{B}_t$ is the last box set of $(B_{t-1},\ B_t)$, and B_t are the earlier box set of $(B_t,\ \hat{B}_{t+1})$. If the IoU affinity is obtained, the detected $\hat{B}_t$ and B_t boxes are matched by applying the Kuhn-Munkres (KM) algorithm [47]. The tracklet, to which $\hat{B}^i_t$ belongs for each of the matching box pair $\hat{B}^i_t$ and B^j_t, is updated by adding B^j_t. The unparalleled box B^k_t is initialized with an identity as a tracklet. Therefore, long paths are generated for the video sequence by using adjacent nodes, and chaining is performed sequentially.

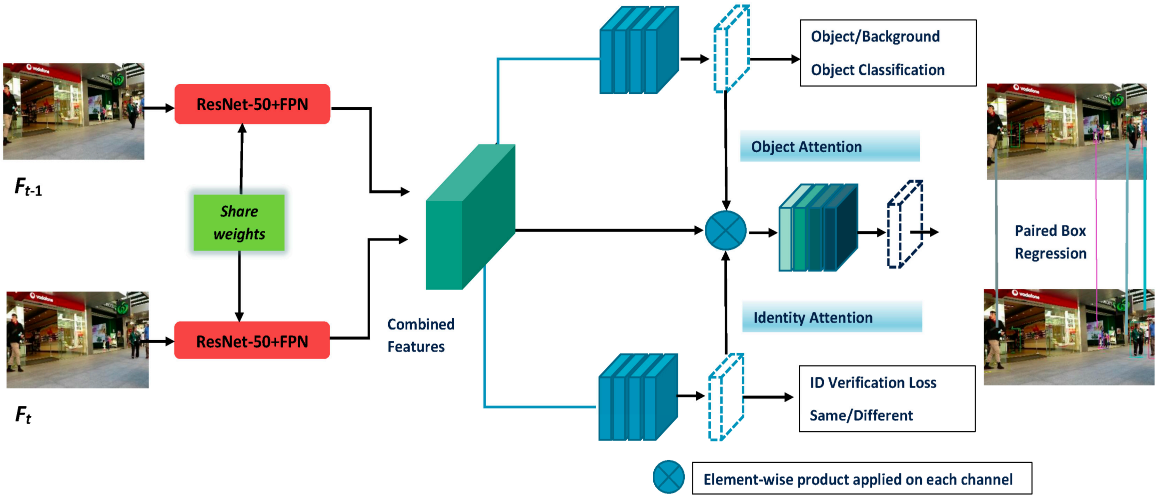
Two nodes were chained by just IoU matching, which is insufficient to chain two nodes and thus resulted in a greater number of IDs (Identity Switch) than previous methods (Section 4.2). The use of box IoU is the reason behind a lot of ID switch present. These are mostly used for fast camera motion in a crowd. IoU [48] is the prevalent evaluation metric for object detection. We have linked two nodes with IoU using strategies as given by:

1. Adding Kalman Filter [49] (predicts object location in the next frame) helps to get smooth and reasonable tracklets which causes a decrease in the number of ID switches. We can keep only those detections for tracking purposes whose predicted location is near to previous frame detection. The mathematical equations of the Kalman filter provide an efficient computational means for evaluating the states of a process by helping evaluate the present, past, and future states. It can also help evaluate when the modeled system is unknown.

2. Using Bipartite Matching [50] (one-to-one mapping), we can assign one identity to one person in the next frame. It ensures one person maps to only one person in the next frame and thus reduces the ID switch. The bipartite matching framework [51] allows us to factorize node similarity in the search for a one-to-one correspondence between nodes in two graphs.

3.4. SCT Framework

The proposed model for object detection takes two frames adjacent to one another as input and a common target for the bounding box pair. We have applied ResNet-50 [8] for feature extraction, which extracted all the semantic features. The Feature Pyramid Networks (FPN) is used to generate the subsequent prediction on a multiscale basis. The scale-level feature maps are collected from each frame and concatenated together, and then these frames are used for the regressed prediction of the bounding box pairs. It is illustrated in Figure 2 that two important branches are: the classification and paired box (regression) branches. Object classification is used for predicting the confidence level of the foreground, whereas the paired box is used for each target. The classification branch avoids inappropriate information, and an ID verification branch is added for attention guidance.



(**a**)

$$\{F \in \mathbb{R}^{c\times w\times h}\}_{\substack{ResNet50_{F_{t-1}}\\+\\FPN}} \oplus \{F \in \mathbb{R}^{c\times w\times h}\}_{\substack{ResNet50_{F_t}\\+\\FPN}}$$

(**b**)

Figure 2. Feature extraction frame-wise with two adjacent frames using two backbone branches consisting of ResNet-50 with FPN by parametric sharing for classification and regression branches, (**a**) SCT architecture, and (**b**) operation expression for combined features.

3.4.1. Paired Boxes Regression

A chained anchor from the regression branch of the paired boxes is used to regress two boxes simultaneously. The chained anchors, mostly used in object detection distributed on a less sparse grid using a spatial basis, are used to predict an instance with bounding boxes for subsequent frames of objects. The k-means clustering algorithm has been adopted for real scenes (on large scales) for ground truth boxes to obtain the chained anchors' scale. Each of the clusters is mapped to a particular level of FPN for performing prediction on later scales. The soft-NMS is used to process the bounding boxes [52], and subsequently, the filtering is carried out based on the confidence level score in the corresponding classification branch. In the end, the IoU is used to chain the box pairs for the entire tracking path. For simplicity of our model, the stack of four consecutive 3 × 3 convolutional layers is used so that ReLU is interleaved after the third convolution layer.

3.4.2. Joint Attention Module

The regression industry introduced the Joint Attention Module (JAM) to extract the combined features using local informative regions [53]. The ID verification branch provides accurate results indicating whether the two boxes of the pair identified belong to the same goal, as illustrated in Figure 2. The forecast of the confidence map of the ID verification branch is used as the attention map and the object classification branch. The trust maps in the classification branch concentrate on frontal regions, while the forecast from the ID verification branch is used to illustrate characteristics having the same target. Note that the advice of the branches is complementary.

3.4.3. Feature Reuse

The common frame is used again while tracking subsequent frames as input for the SCT framework. The MSM saves the extracted features related to the current frame until reused by the next node, as illustrated in Figure 3. The copy of frame N is created to be hypothetical frame $N + 1$, highlighting the suggestion as to the last node, and consequently, the MSM reduces the computational complexity.

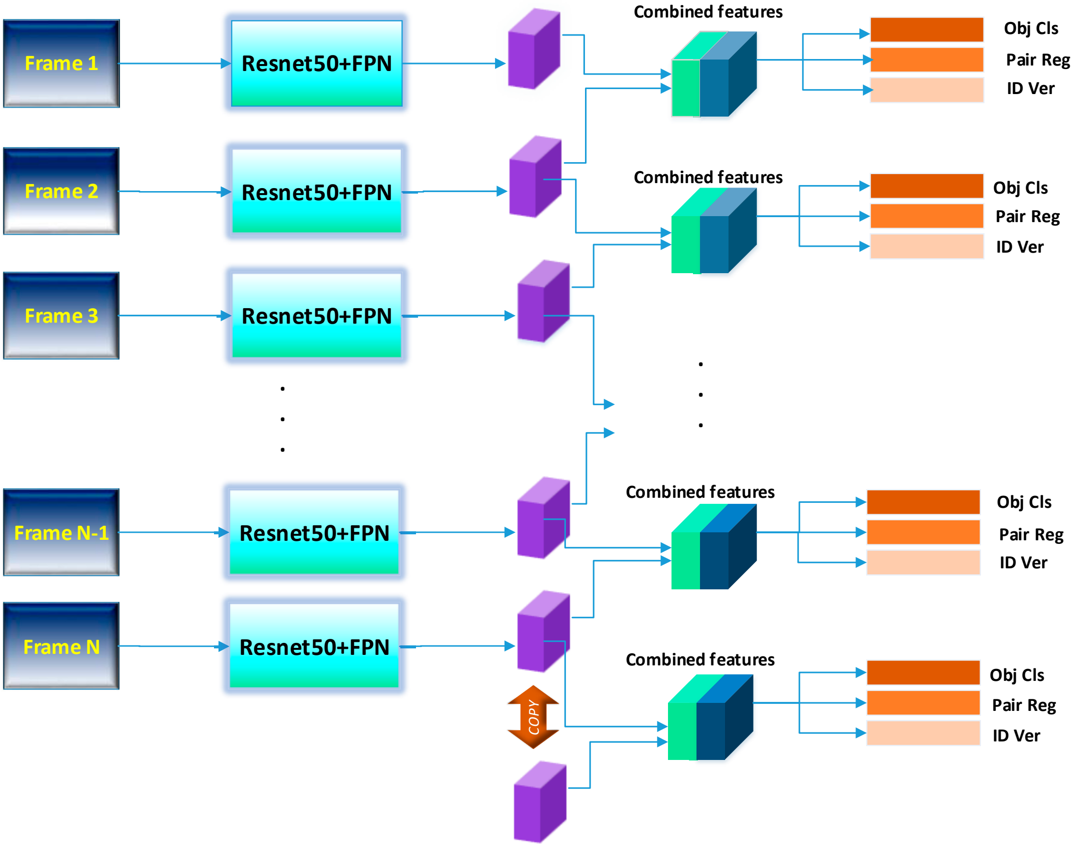


Figure 3. The Memory Sharing Mechanism (MSM) shows the features extracted from each frame (not from the first one) using Resnet-50 with FPN. These features are used for current and even future nodes. The computation cost is alleviated by using the last frame N to avoid its duplicate copy in the case of $N + 1$.

3.4.4. Loss Function and Assigning Labels

Learning the label assignment techniques adjust the network parameters leading iteratively to an optimum position on the error surface. The loss function influences the detection results during the object detection phase. The Single Shot-multibox Detector (SSD) matching technique is used for the ground truth bounding box. The matching result is stored in matrix "S". If G_t represents the bounding box from ground-truth on F_t for ith chained anchor A_i, checked by: (IoU ratio > threshold t_p), then we have $S_{ij} = 1$. If the smaller threshold is represented by t_n, so that: (IoU ratio < t_n), then $S_{ij} = 0$. Depending on the value of S, the ground-truth label (c^i_{cls}) is assigned to the SCT classification branch for A_i as given by:

$$c^i_{cls} = \begin{cases} 1,\ if\ \sum_{j=1}^{k_t} S_{ij} = 1, \\ 0,\ \ if\ \sum_{j=1}^{k_t} S_{ij} = 0, \end{cases} \tag{1}$$

where k_t represents the cardinality of the bounding boxes for the ground-truth frame. The ID verification branch is assigned the ground-truth label for A_i, the predicted pair and the corresponding ground-truth (bounding boxes) are (B^i_t, $\hat{B}^i_t$) and $\left(GB^j_t,\ GB^k_{t+1}\right)$ respectively, as given by the kernel function:

$$c^i_{ID} = \begin{cases} 1,\ if\ c^i_{cls} = 1 \text{ and } I\left(GB^j_t\right) = I(\ GB^k_{t+1}), \\ 0, \qquad\qquad\qquad\qquad\qquad\quad \text{otherwise}, \end{cases} \tag{2}$$

where $I[\cdot]$ denotes the target identity in the bounding box. The loss function is given by:

$$L_{all} = \sum_{t,i}[L_{reg}\left(\Delta^{t,i}_d,\ \Delta^{t+1,i}_{\hat{d}},\ \Delta^{t,j}_g,\ \Delta^{t+1,k}_g\right) + \alpha\mathcal{F}\left(\mathcal{P}^i_{cls},\ c^i_{cls}\right) + \beta\mathcal{F}(\mathcal{P}^i_{id},\ c^i_{id})], \tag{3}$$

where the focal losses, given by $\mathcal{F}\left(\mathcal{P}^i_{id},\ c^i_{id}\right)]$ and $\mathcal{F}\left(\mathcal{P}^i_{cls},\ c^i_{cls}\right)$ [54], are used for both branches of classification and ID verification (particularly for alleviating class imbalance problems), respectively. Moreover, $\mathcal{P}^i_{id}$ and $\mathcal{P}^i_{cls}$ denote the prediction score or confidence level; where α and β are the weighting factors, $\Delta^{t,i}_d$, $\Delta^{t+1,i}_{\hat{d}}$ denote the bounding box offset generated by Faster R-CNN [6], and $\Delta^{t,j}_g$, $\Delta^{t+1,k}_g$ denote the offsets for the ground-truths.

3.5. Datasets and Evaluation Metrics

MOT16 [55,56]: The goal of this publicly available benchmark dataset is to perform the detection of multiple objects present in a frame. MOT16 consists of 14 challenging video sequences with equal distribution of 7 training and 7 testing sequences, where video sequences are carefully annotated. Each video sequence is different from the others according to the camera's static view or movement, the camera's viewpoint, and conditions like weather, humidity, temperature, etc. MOT16 includes only detector Deformable Part Models (DPM), which provide pictorial structures of an elegant framework for object detection [57].

MOT17 [56]: The objective of this benchmark dataset is multiobject tracking. It consists of equally distributed training and testing video sequences, but it provides more accurate ground Truth than MOT16. It consists of 7 sequences, 5316 frames for training and 7 sequences, 5919 frames for testing. The major difference between MOT17 with MOT16 is that it has three types of detectors used in video sequences; Faster R-CNN, Deformable part model (DPM), and Scale-Dependent pooling (SDP). Faster R-CNN [58] has three major components; bottom convolutional layers, a region proposal network, and a bounding box regressor or region-of-interest (ROI) based classifier. Due to its flexibility, Faster R-CNN can carry out instance segmentation [59]. Scale Dependent Pooling (SDP) [60] technique takes care of scale variation in object detection problems. SDP tackles the fixed input size requirement by selecting a proper feature layer depending on the size of the input image to describe an object proposal.

The comparison with other existing methods was carried out by training two models separately using the training data of MOT16 and MOT17. These models were checked for generalization using MOT16 and MOT17 test data. The most widely used CLEAR MOT Metrics [61] include Multiobject Tracking Accuracy (*MOTA*), Multi-Object Tracking Precision (MOTP), Identity Switch (*IDSw*), the Mostly Tracked Trajectory (MT), False Positives (*FP*), False Negatives (*FN*), and Mostly Lost Trajectories (MLT), have been used to determine tracking performance.

3.5.1. Identification-F1 (*IDF*1)

The fraction of ground truth detections (computed) which are the corrected identifiable detections, is known as Identification *F*1 (Precision-Recall). *ID*-*F*1 [62] is defined as the correct ratio of detections over the total number of detections based on the ground truth.

$$IDF1 = 1 - \frac{(IDTP)}{2IDTP + IDFN + IDFP} \tag{4}$$

The *IDF*1 ranks all the trackers on a single scale, which balances precision and recall identifications with their harmonic means. The high rising field of sensors' development has led to its improved accuracy resulting in the enhancement of the tracking accuracy; thereby, consistency of the object trajectory remains intact, i.e., *IDF*1.

3.5.2. Number of Identity Switch (*IDSw*)

ID switches are the count of tracked trajectory changes matching ground-truth identity. The smaller value of the ID switch represents better results [63]. This factor increases when the ID of a person is assigned to another person.

3.5.3. Multiple Object Tracking Accuracy (MOTA)

The MOTA [61] interprets alignment and structural errors in frames due to misses, tracking, false positives, etc., including overall object detection. The field of computer vision uses this performance metric for multiobject tracking and is useful for intuitively measuring the tracker's evaluation by keeping the paths/trajectories. It also estimates the object's location by making the precision independent [61]. The relationship for MOTA is given by:

$$MOTA = 1 - \frac{(FN + FP + IDSw)}{GT} \tag{5}$$

The bases of error are combined in this relation as *GT* (number of ground truths), *FN* (false negatives), *FP* (false positives), and *IDSw* (Number of Identity Switches).

4. Results and Discussion

This section is related to the working of SCT using the datasets, as explained in Section 3.5. The two aspects that have been used to illustrate the results are qualitative and quantitative perspectives.

4.1. Qualitative Results

The data for which we cannot quantize needs qualitative analysis to get a deep perception of the trends hidden inside the structure of the problem and its association with the prediction algorithms related to its parametric variations leading to conclusive results. Below are the results of applying our SCT on the test set of the MOT 16 dataset. These are the images taken from video sequences of the test set. Each person is assigned a bounding box for tracking with an identity number (ID-Verification). The bounding box moves along with the person showing the same ID number. The qualitative results have been shown in Figures 4–9. From these, we can conclude that in two adjacent frames of Figures 4 and 5 from the test set of the MOT16 dataset, we can see the movement of bounding boxes along with two ladies. Rectangular bounding boxes are drawn around every person, and a numeric identity is assigned to every person. It has been illustrated in Figure 4 that two ladies have been detected where bounding boxes were contributed by our SCT (the bounding boxes assigned: ID 7 and 43 to the ladies). As a token of performance, it can be verified that the same IDs remain intact in the next frame, as illustrated in Figure 5. The identities assigned to the pedestrians in Figure 6 are *ID* (1, 3, 4, 5 and 6)^{frame,} and in the contiguous frame in Figure 7, IDs remain the same for the pedestrians with identity numbers (4, 5 and 6)^{frame}. It can be concluded that our tracker correctly tracks persons in the videos, and it also shows robustness in the qualitative results. By comparing Figures 8 and 9, we can conclude that identities and bounding boxes are assigned to pedestrians passing through a camera with ID (4, 6 and 14)^{frame} are assigned to the pedestrians by SCT, and after some frames, the same IDs (4, 6 and 14)^{frame} remain assigned to same pedestrians in Figure 9. Our tracker correctly tracks persons present in the videos with different frames.

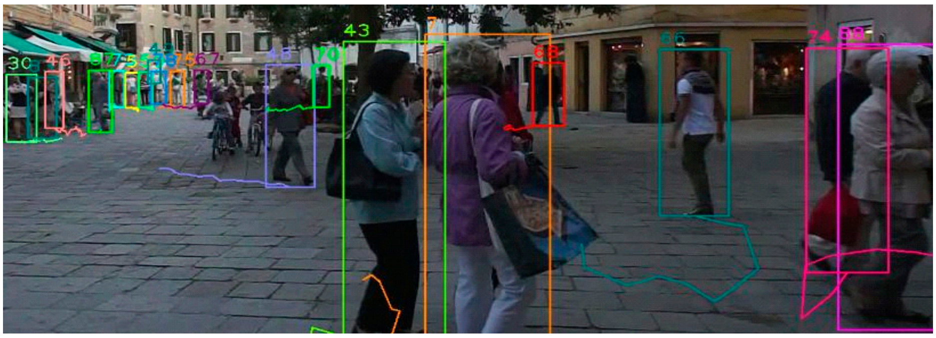


Figure 4. The proposed model (SCT) results using MOT16-02.

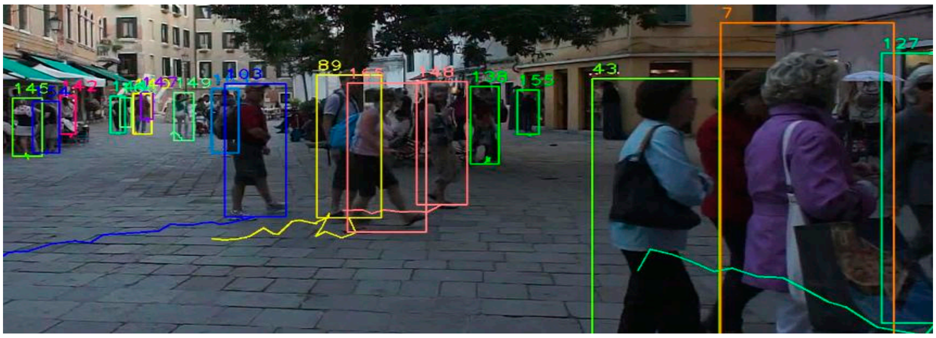


Figure 5. The proposed model (SCT) results using MOT16-02.

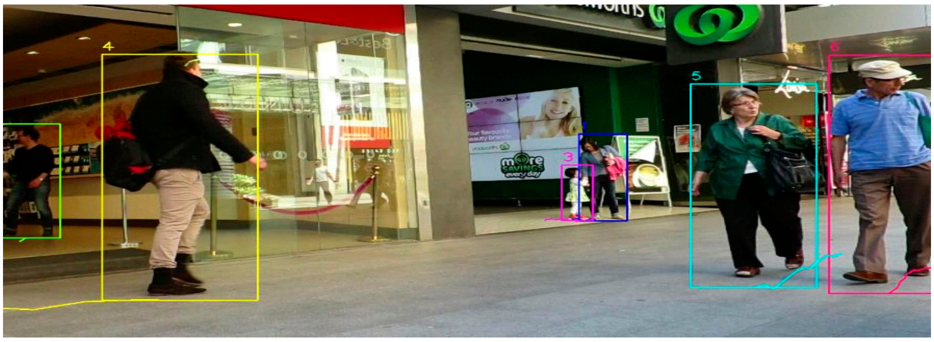


Figure 6. The proposed model (SCT) results using MOT 16-09.

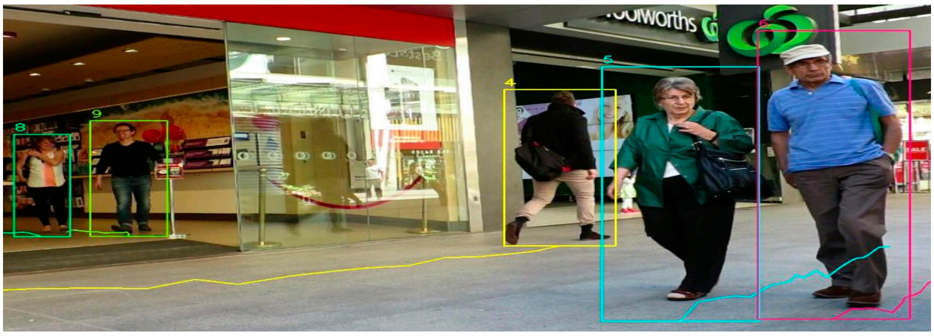


Figure 7. The proposed model (SCT) results using MOT 16-09.

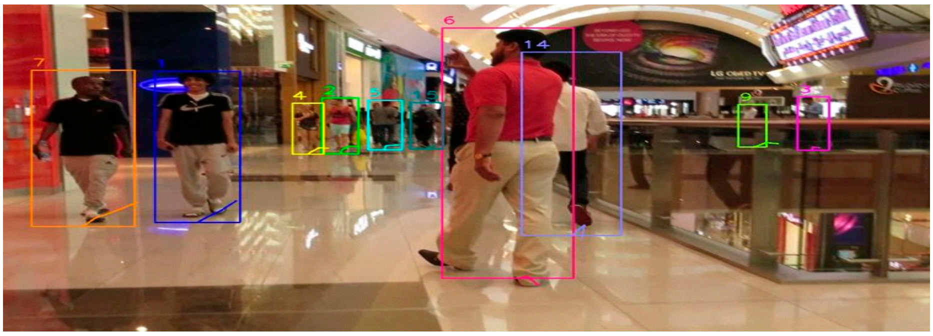


Figure 8. The result of applying the proposed tracker on MOT 16-11.

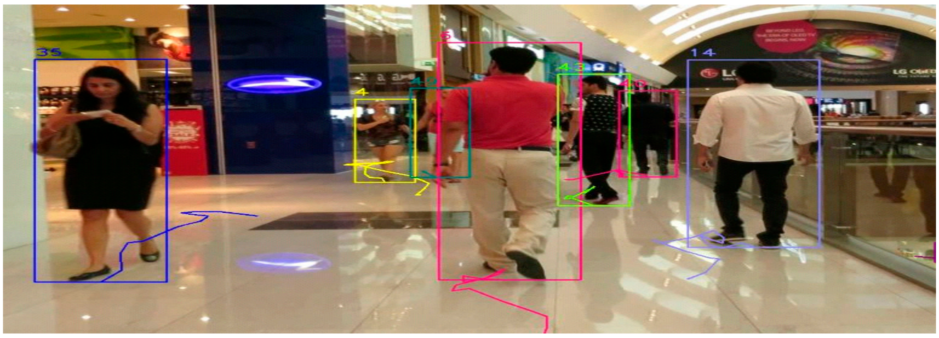


Figure 9. The proposed model (SCT) results using MOT 16-11.

4.2. Quantitative Results

Quantitative data are defined as that can either be counted or compared on a numeric scale. In contrast, its analysis has important implications for data sparsity, generally encountered in representation learning-based environments, for studying scarce and random tracking processes. The quantitative results for both offline and online methods have been illustrated in Tables 1 and 2 for public detectors, respectively. These detectors generate frames based on DPM, FRCNN, and SDP. In Table 1, results with offline methods are listed where the up arrow shows that the higher the value, the better it is, and the down arrow shows that the lesser the value, the better it is. The optimization technique used for MOT is of two types: offline and online versions. The offline MOT, based on training before tracking using past and future frames simultaneously forming a tracking path, performs object detection in the spatial domain by a pre-trained hypothesis in still images, and another pre-trained classifier (offline) is used for object association in the time domain. Similarly, the online version of MOT, based on training while tracking, uses contiguous frame information [64].

Table 1. Results of public detectors on MOT16 dataset (offline methods).

Method	*MOTA*↑	*IDF1*↑	*FP*↓	*FN*↓	*IDSw*↓
LMPR [65]	48.8	51.3	6654	86,245	481
Quad-CNN [66]	44.1	38.3	6388	94,775	745
MHT-bLSTM [67]	42.1	47.8	11,637	93,172	753
EDMT [68]	45.3	47.9	11,122	87,890	639

Table 2. Results of public detectors on MOT16 dataset (online methods).

Method	*MOTA*↑	*IDF1*↑	*FP*↓	*FN*↓	*IDSw*↓
Tracktor [31]	54.4	52.5	3280	79,149	682
DMAN [32]	46.1	54.8	7909	89,874	532
STAM [33]	46.0	50.0	6895	91,117	473
CDA-DDAL [69]	43.9	45.1	6450	95,175	676

In Table 2, results with online methods are listed where the up arrow shows that the higher the value, the better it is, and the down arrow shows that the lesser the value, the better it is. A synthetic image set is generated from the first frame using the object of interest and is further used for the training of the online trackers. In the next frame, the online trackers detect the object of interest in the subsequent frames. The final frames in the video sequence are used to update the detector online. The bold text represents the best results found in the same performance measuring category. In offline methods, Lifted Multicut and Person Re-identification LMPR [65] for tracking multiple people performed outclass compared to other offline methods in *MOTA*, *IDF*1, *FN*, and *IDSw*. Quadruplet Convolution Neural Network (Quad-CNN) [66] for multiobject tracking achieves lower *FP* than the other methods. Tracktor [31] performed better in online methods than other competing methods in *MOTA*, *FP*, and *FN*. Dual Matching Attention Networks (DMAN) [32] for detection of the multiobjects has the highest *IDF*1 54.8, and Spatial-Temporal Attention Mechanism (STAM) [33] achieves minimum IDSs.

Similarly, Tables 3 and 4 show the results of private detectors with offline and online methods that use any technique of choice for tracking purposes, respectively. In Table 3, offline methods are listed that assess the trajectory based on past and future observations. Table 4 lists online methods that assess the trajectory based on past to current frames' information. The optimum results are illustrated in Tables 3 and 4 for offline and online methods.

Table 3. Results of private detectors on MOT16 dataset (Offline Methods).

Method	*IDF1*↑	*MOTA*↑	*FP*↓	*FN*↓	*IDSw*↓
NOMT [34]	62.6	62.2	5119	63,352	406
KDNT [35]	60.0	68.2	11,479	45,605	933
MCMOT-HDM [70]	51.6	62.4	9855	57,257	1394

Table 4. Results of private detectors on MOT16 dataset (Online Methods).

Method	*IDF1*↑	*MOTA*↑	*FP*↓	*FN*↓	*IDSw*↓
EAMTT [36]	53.3	52.5	4407	81,223	910
DeepSORT [37]	62.2	61.4	12,852	56,668	781
CNNMTT [38]	62.2	65.2	6578	55,896	946
CTracker [46]	57.2	67.6	8934	48,305	1897
Proposed system	64.3	68.4	2517	31,405	909

Near-Online Multi-target Tracking (NOMT) [34], as an offline entity, competed with other offline methods in *IDF*1, *FP*, and *IDSw*. The KDNT [35] achieved the best *MOTA* and *FN* compared to the other offline methods. In online methods, our proposed tracker outperformed the other methods like Early Association Multi-Target Tracking EAMTT [36], Deep Simple Online and Realtime Tracking (DeepSORT) [37], CNN Multi-target tracking (CNNMTT) [38], and CTracker [46] in *IDF*1, *MOTA*, *FP*, and *FN*.

It has been concluded from the previous results that our tracker generally outperforms all the other detectors in *MOTA*, *FP*, and *FN*, where *IDF*1 increments approximately 7%, *FP* reduces from 8934 to 2517, *FN* reduces from 48,305 to 31,405 and *IDSw* diminishes to half as compared to the CTracker. This concludes the enhancement in results when compared to the previous methods.

Initially, the absence of a target object in the previous frame, although appearing in the subsequent frame, then no paired bounding boxes are generated in the *t*th frame. When the target object remains present in the subsequent frame, the object is easily identifiable in the future ($t + 1$)th chain node. Furthermore, if the target object is present in the previous ($t - 1$)th frame and it is not present in *t*th frame, then the bounding box pair will not be generated in the future frame, ultimately blocking either of the previous nodes $(t - 1)$ or ($t - 2$) resulting in the disappearance of the frame. Therefore, the chaining operation used in the proposed method is challenging to parameterize, and consequently, regressions do not guarantee optimization in frame label congruency.

5. Conclusions

We proposed a robust Super Chained Tracker to detect and monitor multiobjects because we have used reidentification to improve performance and focus and avoid irrelevant information. This online tracker detects objects, extracts features and combines data into an end-to-end solution. Unlike existing techniques, chain nodes as two subsequent frames are used as input to the network that generates bounding boxes paired by the Joint Attention Module in both adjacent frames. The adjacent nodes that overlap are chained together in a common frame using IoU, Kalman filtering, and Bipartite matching. Tracking paths are generated by applying box regression and node chaining alternatively. Extensive experimentations on MOT16 and MOT17 datasets have shown that our approach is efficient and performs relatively better than the existing state-of-the-art algorithms. Our SCT achieves an Identity *F*1 (*IDF*1) 64.3% and Multi-Object Tracking Accuracy (*MOTA*) 68.4%. Tracking object anchors are more complex and require a lot of calculation, making the system slower. Our tracker gains about a 7% increase in *IDF1*, False Positives (*FP*) decreased from 8934 to 2517, False Negatives (*FN*) decreased from 48,305 to 31,405, and Identity Switch (*IDSw*) reduced to half as compared to the previous methods.

Author Contributions: Conceptualized the study, analyzed data, wrote, revised and supervised the paper, S.A.Q.; conceptualized the study, analyzed data, wrote the paper, revised and supervised the paper, L.H.; implemented, analyzed and edited the paper, Q.-u.-a.C.; conceptualized the study and edited the paper, S.R.A.; analyzed the data and edited the paper, R.J.K.; analyzed the data and edited the paper, A.A.; edited and supervised the paper, A.A.-F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable as the dataset is publically available.

Data Availability Statement: The datasets are publically available.

Acknowledgments: We are thankful to Qatar National Library (QNL) for supporting the publication charges of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The nomenclature of the terminology used in this article is illustrated in Table A1.

Table A1. Nomenclature used throughout the article.

Symbol	Use
CNN	Convolution Neural Network
CNNMTT	Multi-target tracking based on CNN
DeepSORT	Deep Simple Online and Realtime Tracking
DPM	Deformable part model
DMAN	Dual Matching Attention Networks
EER	Equal error rate
EAMTT	Early Association Multi-Target Tracker
FPN	Feature Pyramid Networks
ID-F1	Identification-F1
IDSw	Identity Switch
IoU	Intersection over Union
JAM	Joint Attention Module
LBP	Local Binary Patterns
LMPR	Lifted Multicut and Person Re-identification
MAT	Motion-Aware Tracker
MOT	Multiobject Tracking
MOTA	Multiobject Tracking Accuracy
MOTP	Multiobject Tracking Precision
MHI	Motion History Image
MSM	Memory Sharing Mechanism
MT	Mostly Tracked Trajectory
MLT	Mostly Lost Trajectories
NOMT	Near-Online Multi-target Tracking
POI	Person of Interest
Quad-CNN	Quadruplet Convolution Neural Network
RNN	Recurrent Neural Network
R-CNN	Region-based Convolutional Neural Network
RPN	Region Proposal Network
ROI	Region of Interest
SDP	Scale-Dependent pooling
SCT	Super Chained Tracker
SSD	Single Shot-multibox Detector
STAM	Spatial-Temporal Attention Mechanism
SORT	Simple Online and Real-time Tracking
TBD	Tracking-by-detection
UAV	Unmanned Armed Vehicles
YOLO	You Only Look Once

References

1. Ciaparrone, G.; Sánchez, F.L.; Tabik, S.; Troiano, L.; Tagliaferri, R.; Herrera, F. Deep learning in video multi-object tracking: A survey. *Neurocomputing* **2020**, *381*, 61–88. [CrossRef]
2. Xu, B.; Liang, D.; Li, L.; Quan, R.; Zhang, M. An Effectively Finite-Tailed Updating for Multiple Object Tracking in Crowd Scenes. *Appl. Sci.* **2022**, *12*, 1061.
3. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
4. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
5. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
6. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Processing Syst.* **2015**, *28*, 91–99. [CrossRef]
7. 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*; Springer: Berlin/Heidelberg, Germany, 2016.

8. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
9. Tangirala, K.V.; Namuduri, K.R. Object tracking in video using particle filtering. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05), Philadelphia, PA, USA, 23 March 2005.
10. Tariq, S.; Farooq, H.; Jaleel, A.; Wasif, S.M. Anomaly detection with particle filtering for online video surveillance. *IEEE Access* **2021**, *9*, 19457–19468.
11. Takala, V.; Pietikainen, M. Multi-object tracking using color, texture and motion. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007.
12. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987.
13. Shu, G. Human Detection, Tracking and Segmentation in Surveillance Video. Ph.D. Thesis, University of Central Florida, Orlando, FL, USA, 2014.
14. Wang, L.; Liu, T.; Wang, G.; Chan, K.L.; Yang, Q. Video tracking using learned hierarchical features. *IEEE Trans. Image Processing* **2015**, *24*, 1424–1435.
15. Gao, T.; Wang, N.; Cai, J.; Lin, W.; Yu, X.; Qiu, J.; Gao, H. Explicitly exploiting hierarchical features in visual object tracking. *Neurocomputing* **2020**, *397*, 203–211.
16. Zhong, Z.; Gao, Y.; Zheng, Y.; Zheng, B. Efficient spatio-temporal recurrent neural network for video deblurring. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020.
17. Kang, K.; Ouyang, W.; Li, H.; Wang, X. Object detection from video tubelets with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
18. Ning, G.; Zhang, Z.; Huang, C.; Ren, X.; Wang, H.; Cai, C.; He, Z. Spatially supervised recurrent convolutional neural networks for visual object tracking. In Proceedings of the 2017 IEEE International Symposium on Circuits and Systems (ISCAS), Baltimore, MD, USA, 28–31 May 2017.
19. Held, D.; Thrun, S.; Savarese, S. Learning to track at 100 fps with deep regression networks. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016.
20. Ma, C.; Huang, J.-B.; Yang, X.; Yang, M.-H. Hierarchical convolutional features for visual tracking. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
21. Wang, L.; Ouyang, W.; Wang, X.; Lu, H. Visual tracking with fully convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
22. Shen, H.; Li, S.; Zhu, C.; Chang, H.; Zhang, J. Moving object detection in aerial video based on spatiotemporal saliency. *Chin. J. Aeronaut.* **2013**, *26*, 1211–1217. [CrossRef]
23. Zhang, J.; Liang, X.; Wang, M.; Yang, L.; Zhuo, L. Coarse-to-fine object detection in unmanned aerial vehicle imagery using lightweight convolutional neural network and deep motion saliency. *Neurocomputing* **2020**, *398*, 555–565.
24. Hui, T.-W.; Tang, X.; Loy, C.C. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
25. Wang, R.J.; Li, X.; Ling, C.X. Pelee: A real-time object detection system on mobile devices. *arXiv* **2018**, arXiv:1804.06882.
26. Guo, L.; Liao, Y.; Luo, D.; Liao, H. Generic object detection using improved gentleboost classifier. *Phys. Procedia* **2012**, *25*, 1528–1535. [CrossRef]
27. Ayed, A.B.; Halima, M.B.; Alimi, A.M. MapReduce based text detection in big data natural scene videos. *Procedia Comput. Sci.* **2015**, *53*, 216–223. [CrossRef]
28. Viswanath, A.; Behera, R.K.; Senthamilarasu, V.; Kutty, K. Background modelling from a moving camera. *Procedia Comput. Sci.* **2015**, *58*, 289–296. [CrossRef]
29. Soundrapandiyan, R.; Mouli, P.C. Adaptive pedestrian detection in infrared images using background subtraction and local thresholding. *Procedia Comput. Sci.* **2015**, *58*, 706–713. [CrossRef]
30. Park, Y.; Dang, L.M.; Lee, S.; Han, D.; Moon, H. Multiple object tracking in deep learning approaches: A survey. *Electronics* **2021**, *10*, 2406. [CrossRef]
31. Bergmann, P.; Meinhardt, T.; Leal-Taixe, L. Tracking without bells and whistles. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 2 November 2019.
32. Zhu, J.; Yang, H.; Liu, N.; Kim, M.; Zhang, W.; Yang, M.-H. Online multi-object tracking with dual matching attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
33. Chu, Q.; Ouyang, W.; Li, H.; Wang, X.; Liu, B.; Yu, N. Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
34. Choi, W. Near-online multi-target tracking with aggregated local flow descriptor. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
35. Yu, F.; Li, W.; Li, Q.; Liu, Y.; Shi, X.; Yan, J. Poi: Multiple object tracking with high performance detection and appearance feature. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016.
36. Sanchez-Matilla, R.; Poiesi, F.; Cavallaro, A. Online multi-target tracking with strong and weak detections. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016.

37. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 7 February 2017.
38. Mahmoudi, N.; Ahadi, S.M.; Rahmati, M. Multi-target tracking using CNN-based features: CNNMTT. *Multimed. Tools Appl.* **2019**, *78*, 7077–7096. [CrossRef]
39. Zaech, J.-N.; Liniger, A.; Dai, D.; Danelljan, M.; Van Gool, L. Learnable online graph representations for 3d multi-object tracking. In Proceedings of the IEEE Robotics and Automation Letters, Philadelphia, PA, USA, 23–27 May 2022.
40. Han, S.; Huang, P.; Wang, H.; Yu, E.; Liu, D.; Pan, X. Mat: Motion-aware multi-object tracking. *Neurocomputing* **2022**, *476*, 75–86. [CrossRef]
41. Sun, Z.; Chen, J.; Mukherjee, M.; Liang, C.; Ruan, W.; Pan, Z. Online multiple object tracking based on fusing global and partial features. *Neurocomputing* **2022**, *470*, 190–203. [CrossRef]
42. Yin, T.; Zhou, X.; Krahenbuhl, P. Center-based 3d object detection and tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
43. Tsai, C.-Y.; Su, Y.-K. MobileNet-JDE: A lightweight multi-object tracking model for embedded systems. *Multimed. Tools Appl.* **2022**, *81*, 9915–9937. [CrossRef]
44. Breitenstein, M.D.; Reichlin, F.; Leibe, B.; Koller-Meier, E.; Van Gool, L. Robust tracking-by-detection using a detector confidence particle filter. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009.
45. Bochinski, E.; Eiselein, V.; Sikora, T. High-speed tracking-by-detection without using image information. In Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017.
46. Peng, J.; Wang, C.; Wan, F.; Wu, Y.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; Fu, Y. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020.
47. Qin, C.; Zhang, Y.; Liu, Y.; Lv, G. Semantic loop closure detection based on graph matching in multi-objects scenes. *J. Vis. Commun. Image Represent.* **2021**, *76*, 103072. [CrossRef]
48. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
49. Patel, H.A.; Thakore, D.G. Moving object tracking using kalman filter. *Int. J. Comput. Sci. Mob. Comput.* **2013**, *2*, 326–332.
50. Meinhardt, T.; Kirillov, A.; Leal-Taixe, L.; Feichtenhofer, C. Trackformer: Multi-object tracking with transformers. *arXiv* **2021**, arXiv:2101.02702,.
51. Shokoufandeh, A.; Dickinson, S. Applications of bipartite matching to problems in object recognition. In Proceedings of the ICCV Workshop on Graph Algorithms and Computer Vision, Corfu, Greece, 21 September 1999.
52. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS-improving object detection with one line of code. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
53. Ravindran, V.; Osgood, M.; Sazawal, V.; Solorzano, R.; Turnacioglu, S. Virtual reality support for joint attention using the Floreo Joint Attention Module: Usability and feasibility pilot study. *Jmir Pediatrics Parent.* **2019**, *2*, e14429. [CrossRef] [PubMed]
54. 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, Venice, Italy, 22–29 October 2017.
55. Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; Schindler, K. MOT16: A benchmark for multi-object tracking. *arXiv* **2016**, arXiv:1603.00831.
56. Yu, E.; Li, Z.; Han, S.; Wang, H. Relationtrack: Relation-aware multiple object tracking with decoupled representation. *IEEE Trans. Multimed.* 2022. [CrossRef]
57. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1627–1645. [CrossRef]
58. Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; Van Gool, L. Domain adaptive faster r-cnn for object detection in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
59. Dai, J.; He, K.; Sun, J. Instance-aware semantic segmentation via multi-task network cascades. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
60. Yang, F.; Choi, W.; Lin, Y. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
61. Bernardin, K.; Stiefelhagen, R. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP J. Image Video Processing* **2008**, *2008*, 246309. [CrossRef]
62. Liu, Y.; Bai, T.; Tian, Y.; Wang, Y.; Wang, J.; Wang, X.; Wang, F.-Y.I. SegDQ: Segmentation Assisted Multi-Object Tracking with Dynamic Query-based Transformers. *Neurocomputing* **2022**, *481*, 91–101. [CrossRef]
63. Li, Y.; Huang, C.; Nevatia, R. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.

64. Lee, J.; Kim, S.; Ko, B.C. Online multiple object tracking using rule distillated siamese random forest. *IEEE Access* **2020**, *8*, 182828–182841. [CrossRef]
65. Tang, S.; Andriluka, M.; Andres, B.; Schiele, B. Multiple people tracking by lifted multicut and person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
66. Son, J.; Baek, M.; Cho, M.; Han, B. Multi-object tracking with quadruplet convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
67. Kim, C.; Li, F.; Rehg, J.M. Multi-object tracking with neural gating using bilinear lstm. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
68. Chen, J.; Sheng, H.; Zhang, Y.; Xiong, Z. Enhancing detection model for multiple hypothesis tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017.
69. Bae, S.-H.; Yoon, K.-J. Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 595–610. [CrossRef] [PubMed]
70. Lee, B.; Erdenee, E.; Jin, S.; Nam, M.Y.; Jung, Y.G.; Rhee, P.K. Multi-class multi-object tracking using changing point detection. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016.