Assignment 3: FairMOT

Abdul Bhutta
Electrical and Computer Engineering
Toronto Metropolitan University (TMU)
Toronto, Canada
abdul.bhutta@torontomu.ca

I. Introduction

Multiple Object Tracking (MOT) is a rapidly evolving and exciting research area in computer vision. It is an extension of object detection, addressing the dynamic task of tracking multiple objects in a video sequence over time. The primary goal of MOT is to track various objects in a video frame while maintaining their unique identities. The development of numerous state-of-the-art algorithms has not diminished the challenges in MOT, which remain significant due to occlusion, scale variation, and illumination changes to the objects. However, the continuous introduction and development of several popular MOT algorithms, including DeepSORT, FairMOT, JDE, Tractor, and ByteTrack, reassure us about the progress in the field. This ongoing progress gives us hope and optimism for the future of multiple object tracking. Another excellent use case for MOT is that they are also used for Single Object Tracking (SOT), making such algorithms designed only for SOT obsolete and less of a research area. This assignment covers the state-of-the-art algorithm FairMOT, which aims to address the challenges caused by other tracking algorithms detection and re-ID modules by incorporating anchor-free detection using center points and a simultaneous module to share the extracted features.

II. FairMOT Algorithm

FairMOT algorithm allows tracking multiple objects in a video or real-time application. The detected objects are tracked in the video, while each is assigned a unique ID. The algorithm uses a re-ID module to assign a unique ID to the object, which allows it to preserve the identity of the tracked object. It is a single-shot tracker that simultaneously detects and reidentifies objects using a single network. The two tasks, detection and tracking, are optimized to improve tracking performance, and the algorithm is designed to be highly computationally efficient. The FairMOT algorithm initially takes an image as input and passes it through the DLA34 backbone network to extract features. It contains two main modules: object detection and reidentification, both using the extracted features from the backbone. This unique approach allows the algorithm to detect and reidentify objects simultaneously, setting it apart from traditional tracking methods. The detection head is based on the anchor-free CenterNet and has three parallel heads to estimate the heatmap, box size, and center offset. Each head provides unique information about the object, where the heatmap head gives the object center, and the box size head provides the target box dimensions. The algorithm utilizes downsampling to reduce the height and width of the feature map by a factor of 4. While this may cause the center to be misplaced by up to 4 pixels in each direction, the algorithm's center offset head steps in to provide the offset of the object's center, ensuring precise localization. The Re-ID module generates a unique ID for each detected object and feature vectors to differentiate between them. The algorithm's two-stage matching strategy, which matches objects in the current frame with those in the previous frame, further enhances its precision. In the second stage, the Hungarian algorithm plays a pivotal role. It matches the boxes from the previous frame with the current frame, ensuring that the distance values (Mahalabonis and Cosine) are fused and used effectively. The algorithm compares the overlap or IOU of the bounding boxes of the previous and the current new frame. Based on a specific threshold, the tracks are either kept with the former unique ID or a new unique ID is assigned to the unmatched detection. The model output provides the bounding boxes and vector embeddings for each object in the current frame.

III. Loss Function

The FairMOT algorithm's model incorporates one loss function during training called the $total\_loss$ function. This loss function is encapsulated and constructed from 4 various loss functions: heatmap loss, box loss, offset loss, and re-ID loss. Each loss function is precisely detailed and explained in the original paper while being instrumental in the algorithm's efficiency. The heatmap loss, also known as the Detection loss, is used to pinpoint the object's center in the image accurately. This loss function, eq.1, operates on a pixel-wise logistic regression with focal loss, ensuring the highest level of accuracy in object detection [1].

$$L_{heat} = -\frac{1}{N}\sum_{xy}\begin{cases}(1-\hat{M}_{xy})^{\alpha}\log(\hat{M}_{xy}) & \text{if } M_{xy}=1,\\ (1-M_{xy})^{\beta}(\hat{M}_{xy})^{\alpha} & \\ \log(1-\hat{M}_{xy}) & \text{otherwise}\end{cases} \tag{1}$$

where M_{xy} is the ground truth heatmap, $\hat{M}_{xy}$ is the predicted heatmap, and α and β are the pre-determined/ hyperparameters [1].

The box offset, and size head loss are used to accurately predict the bounding box coordinates and the size of the detected object. The loss function used is the L1 loss [1],

$$L_{box} = \sum_{i} \left(\|o_i - \hat{o}_i\|_1 + \lambda_s \|s_i - \hat{s}_i\|_1 \right) \tag{2}$$

where o_i and s_i are the ground truth offset and size for each object and $\hat{o}_i$ and $\hat{s}_i$ are the predicted offset and size for each object [1].

The re-ID loss function plays a pivotal role in the FairMOT algorithm, enabling the tracking of each unique feature embedding for every object. This function, eq.3, which operates on the cross-entropy loss, allows the algorithm to track objects across different frames.

$$L_{identity} = -\sum_{i} \sum_{k} L_i(k) \log(p(k)) \tag{3}$$

To train the model, the total loss is calculated using the combination of the detection and re-ID loss with the learnable parameters w_1 and w_2 to balance the two losses [1]. The total loss function is calculated as follows,

$$L_{total} = \frac{1}{2} \left(\frac{1}{e^{w_1}} L_{detection} + \frac{1}{e^{w_2}} L_{identity} + w_1 + w_2 \right) \tag{4}$$

where w_1 and w_2 are the learnable parameters to balance the detection and re-ID losses. $L_{detection}$ is the sum of the heatmap loss, box loss, and offset loss,

IV. Training & Testing Dataset

The current FairMOT models were trained using the MOT-17 dataset provided by the MOT challenge, which contained MOT16 video sequences with seven videos for training and seven videos for testing. The new dataset (MOT-17) contains more bounding boxes and is more accurate than its predecessor. Furthermore, each video includes three sets of detection: DPM, FRCNN, and SDP. Each model training used pre-trained weights using the COCO 2017 dataset, a widely used dataset for object detection tasks. These pre-trained weights provide a starting point for the model, allowing it to learn from a broad range of object categories and features before being fine-tuned on the MOT-17 dataset. The training and validation of the COCO dataset were used for 230 epochs, and the learning rate was dropped about ten times during training using the CenterNet algorithm. The MOT17 test set, which contains videos similar to the training set but with different camera angles in the same environment, was used to evaluate the model's performance [2]. This set is crucial as it provides a real-world scenario for the model to demonstrate its effectiveness.

The two models were trained using the algorithm provided by the FairMOT repository [3] and the MOT-17 dataset, which was processed frame by frame and required to detect only the pedestrians (1 class) in each video. The model was trained for 30 epochs with two different datasets. The first model used the entire dataset, containing all the video frames for training and testing. Although there were better approaches than this, it was used to verify the model's results and tracking. The second model used a half-split dataset, where half of the frames for each video sequence were used for training and the other half for testing. The second model is a more realistic approach. It is trained on one dataset and tested using unseen data but in the same environment, which mimics the testing dataset in the MOT challenge. Both datasets allowed the model to be trained, and the test dataset was used with the given ground truth data to provide evaluation metrics for each model.

V. Results

A. Training Dataset Results

Tables 1 and 2 demonstrate the results for the training dataset for the entire dataset (Model 1) and the half-half split (Model 2) with all loss functions for each epoch. Notably, epochs 30 and 29 exhibited the model's efficiency, with the lowest total loss values of -0.2286 and -0.3406, respectively. This indicates the model performs the best with the weights provided at those epochs.

TABLE I
Model 1 Training Results

Epoch	Loss	hm_loss	wh_loss	off_loss	id_loss
1	9.6323	1.0486	2.5409	0.2279	4.6623
5	1.6720	0.4941	0.8555	0.2059	1.1045
10	0.4433	0.3924	0.7270	0.1966	0.4116
15	0.0637	0.3574	0.6855	0.1926	0.2302
20	-0.1601	0.3427	0.6616	0.1914	0.1540
25	-0.3142	0.3206	0.6264	0.1879	0.1061
29	**-0.3406**	**0.3164**	**0.6093**	**0.1869**	**0.1021**
30	-0.3392	0.3189	0.6262	0.1877	0.1021

TABLE II
Model 2 Training Results

Epoch	Loss	hm_loss	wh_loss	off_loss	id_loss
1	10.8061	1.1807	3.1605	0.2302	4.9122
5	2.3370	0.5123	0.8433	0.2066	1.2514
10	0.8026	0.4043	0.7167	0.1962	0.4739
15	0.2496	0.3487	0.6380	0.1903	0.2582
20	-0.0737	0.3178	0.6020	0.1859	0.1633
25	-0.1932	0.3014	0.5673	0.1823	0.1230
30	**-0.2286**	**0.2940**	**0.5624**	**0.1823**	**0.1152**

B. Test Dataset Results

The test dataset results for the two models are shown below. The first model's test results are presented in Table 3, and the second model in Table 4. The first model achieved an overall MOTA of 83.3%, a precision of 97.5%, and a recall of 86.0%. In comparison, the second model's overall MOTA was 67.9%, with a precision of 92.1% and a recall of 75.2%. While these results are less accurate than the first model's, it's important to note that the second model's tracker performed exceptionally well in tracking the objects to unseen data, highlighting its potential in real-world applications.

TABLE III
MODEL 1 TEST DATASET RESULTS

Video	MOTA	IDF1	MT	ML	Prcn	Rcll
MOT17-02-SDP	66.0%	61.9%	23	11	96.9%	69.0%
MOT17-04-SDP	93.9%	90.7%	75	2	98.1%	96.0%
MOT17-05-SDP	75.9%	82.0%	46	16	98.2%	78.0%
MOT17-09-SDP	80.9%	72.1%	20	0	98.7%	82.9%
MOT17-10-SDP	75.1%	75.1%	31	0	96.0%	79.2%
MOT17-11-SDP	87.4%	87.7%	54	4	96.8%	90.8%
MOT17-13-SDP	78.9%	81.8%	73	11	96.8%	82.4%
OVERALL	**83.3%**	**82.1%**	**322**	**44**	**97.5%**	**86.0%**

TABLE IV
MODEL 2 TEST DATASET RESULTS

Video	MOTA	IDF1	MT	ML	Prcn	Rcll
MOT17-02-SDP	42.6%	46.5%	10	15	87.9%	50.9%
MOT17-04-SDP	82.1%	82.6%	52	3	95.6%	86.7%
MOT17-05-SDP	64.9%	70.5%	25	11	93.1%	71.6%
MOT17-09-SDP	74.4%	67.6%	14	0	98.8%	76.1%
MOT17-10-SDP	63.7%	73.0%	12	1	91.2%	71.5%
MOT17-11-SDP	63.1%	64.8%	18	14	87.9%	74.0%
MOT17-13-SDP	50.0%	69.0%	23	6	76.0%	74.6%
OVERALL	**67.9%**	**71.9%**	**154**	**50**	**92.1%**	**75.2%**

Figures 1 and 2 illustrate the tracking results from a video in the test dataset for each model. Both models demonstrated their accuracy and reliability by successfully tracking the objects in the video and assigning a unique ID to each object, providing a strong reassurance of their reliability and instilling a sense of confidence in their performance.

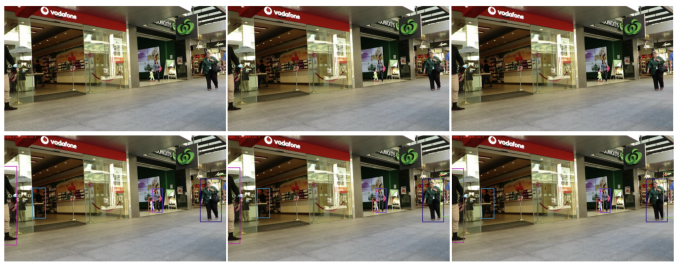

Fig. 1. Model 1 Tracking Results on Test Dataset

Fig. 2. Model 2 Tracking Results on Test Dataset

C. Inference Results

The FairMOT algorithm provides a demo video to run an inference test using the trained model and is part of the MOT-16 Test dataset (MOT16-03). The two trained models were used to track the objects in the video sequence, and each model's results were visualized and analyzed. Although the ground truth for the image was not provided, the results were analyzed based on the tracking results, and it showed both models did an excellent job in tracking the object in the video. The results from both models are provided in Figures 3 and 4.

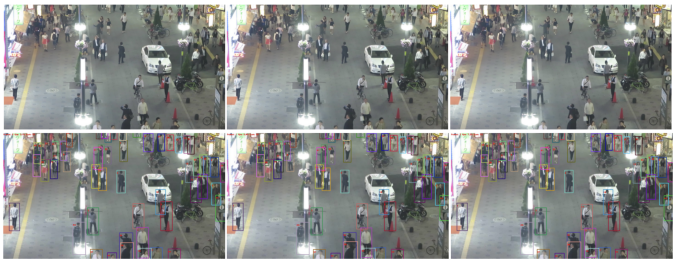

Fig. 3. Inference Results for Model 1

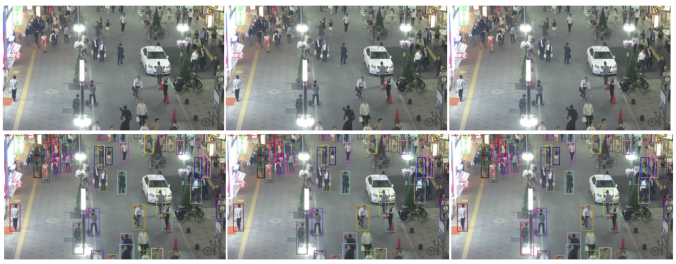

Fig. 4. Inference Results for Model 2

VI. CONCLUSION

As seen by the results, the FairMOT algorithm is a powerful and efficient tracker that can track multiple objects in a video and is considered a state of the art tracker. The use cases for such tracker are vast and can be used in real-time applications such as surveillance, monitoring, and autonomous driving. Although there will be many challenges to overcome such as occlusion, scale variation, and appearance changes, the FairMOT algorithm has shown to be robust and efficient in tracking the objects. The question remains when the solution to such challenges will be solved and how far such algorithms can go in the future with the advancements in AI and computer vision.

REFERENCES

[1] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, 'FairMOT: On the Fairness of Detection and Re-identification in Multiple Object Tracking', International Journal of Computer Vision, vol. 129, pp. 1–19, 11 2021.

[2] P. Dendorfer et al., 'MOTChallenge: A Benchmark for Single-camera Multiple Target Tracking', CoRR, vol. abs/2010.07548, 2020.

[3] I. Zhang, "Ifzhang/Fairmot: [IJCV-2021] Fairmot: On the fairness of detection and re-identification in multi-object tracking," GitHub, https://github.com/ifzhang/FairMOT/tree/master.