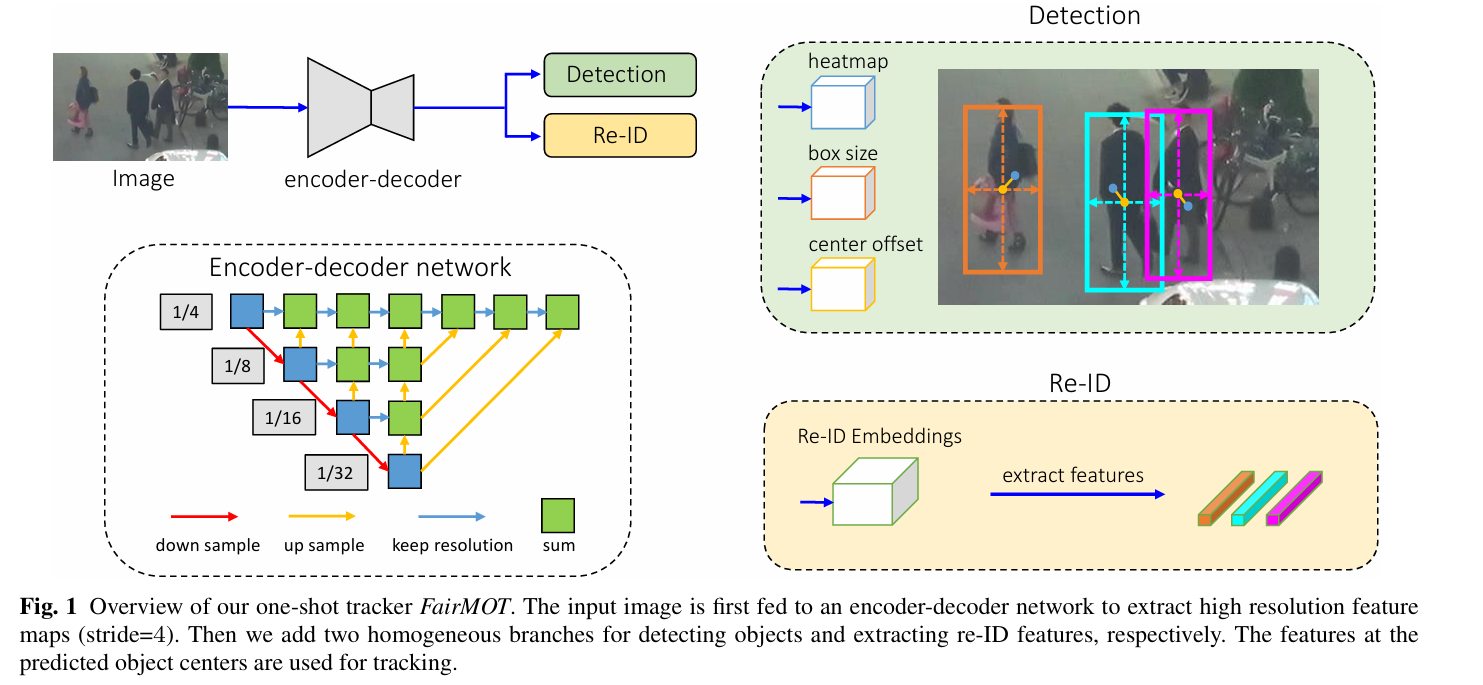
FairMOT : state-of-the-art Multi-Object Tracking model

# My Understanding –

* FairMOT is a *state-of-the-art* multi-object tracker model that integrates detection and re-identification into the same network, hence striking a balance between accuracy and efficiency.
* The advantage this model provides over others is that it **treats the detection and re-identification tasks equally**, unlike previously introduced models that make detection the primary task and re-id a secondary one, significantly affecting the accuracy and computation of the same.
* FairMOT is an **anchor-free** object detection model based on the CenterNet architecture.



* Above depicts a general architecture of the FairMOT model. Note that detection and re-id branches are homogeneous. Detection is done based on the CenterNet architecture, removing the need for anchors.

## Disadvantage of the two-step method

* The two step method, that is first detection, followed by re-id suffers from scalability issues as it is fails to perform optimally in real-time when the number of objects in the given environment is large.
* The reason being, the two task (and their models) do not share any features and therefore, for every bounding box, the re-id model is applied repetitively.

## Issues caused by the one-shot trackers

* Due to the reliance on **anchors** for initial object detection, the training process might become biased towards detection accuracy, neglecting the refinement of Re-ID features. This means that the model could be very good at detecting objects but poor at re-identifying them across different frames, resulting in higher identity switches and track fragmentation in MOT scenarios.
* **Feature Sharing –** Many one-shot trackers work in such a manner that re-id features are learned from detection features itself extracted by the same network. This becomes an issue as the authors suggest that both tasks need their own independent feature embeddings to avoid any conflicts in features.
* **Feature Dimension –** The dimensions of the re-id features are much higher (512 or 1024) than that of object detection. This size difference between the two embeddings comes at the cost of performance of the both the tasks.

# FairMOT architecture details

* **ResNet34** and **DLA (Deep Layer Aggregation)** is used as the backbone structure of this model for image feature extraction. This is chosen as a viable option for its simultaneous accuracy and speed.
* DLA is used to enhance the feature extraction capabilities of the backbone network by effectively aggregating features from different layers. It aggregates features through iterative and hierarchical merging, improving the semantic richness of the features at every scale. It integrates low-level details with high-level semantic information, ensuring that the extracted features are rich and multi-scale.
* **CenterNet:** CenterNet is an anchor-free detection framework that predicts object centers without relying on predefined anchor boxes, which simplifies the network design and reduces the hyperparameters involved in defining anchors.
* **Heatmap Prediction**: It uses a fully convolutional network to predict a heatmap where each peak corresponds to the center of an object. This method directly maps the input image to output spatial dimensions, making it highly efficient.
* **Size and Offset Prediction**: Alongside the heatmap, CenterNet predicts the size of each object’s bounding box and an offset value for each detected center. The size predictor gives the width and height of the box, while the offset predictor compensates for the discretization error caused by the output stride of the network.
* **Detection Branch:** The detection branch of FairMOT, based on CenterNet, identifies the center of each object, predicts the size of bounding boxes, and computes center offsets which are crucial for precise localization.
* **Re-ID Branch:** This branch uses the high-resolution features generated by the backbone and aggregated through DLA to compute re-identification features at each object center. These features are then used to match objects across different frames, maintaining their identities.