MADA – NET: Multi-frame Affinity-Based Deep Association for Multi-Object Tracking

BMVC 2020 Submission # 869

Abstract

Multi-object tracking (MOT) is one of the most important computer vision problems, due to its applications in video surveillance, autonomous driving, sports analysis, etc. In this paper, we propose a multi-frame object tracking technique as an end-to-end deep clustering called Multi-frame Affinity-Based Deep Association (*MADA*-NET). The proposed method is based on a data clustering algorithm known as Constraint Dominant Sets Clustering (CDSC). For the first time, we are able to integrate the maximization of a differentiable parametrized quadratic function in a deep-model as an objective function for solving a multi-frame tracking problem. *Unlike the predominant two-frame affinity-based MOT schemes employing bi-partite graph, we formulate the underlying association process as finding a highly coherent cluster from a k-partite graph, where the clusters correspond to the object tracks.* Therefore, our method does not require special measures to handle newly entering, exiting, occluded, and miss-detected objects. Extensive experiments on benchmark datasets show that the proposed method does not only achieve competitive results but also opens up a new research direction by formulating the MOT problem as a cluster optimization problem into an end-to-end deep learning fashion.

1 Introduction

Due to its numerous applications ranging from video surveillance, human computer interface, autonomous driving to sports analysis etc, multi-object tracking (MOT) has long been an indispensable computer vision problem. Given a video, the goal of MOT model is to predict the trajectories or tracks of each object in the video, while maintaining their identities across multiple video frames. The predominant MOT methods decompose multi-object tracking into two separate modules, such as object detection and tracking, also known as tracking-by-detection paradigm [6, 7, 26]. The detection module detects bounding boxes around objects of interest independently in each frame. Subsequently, the association module links the bounding boxes across time to form tracks. Given the detected objects, classical methods [18, 21, 44] associate the detections across frames by computing *pairwise* affinity. To that end, these methods employ different representation models such as appearance model [2, 21, 51], motion model [6, 15] and composite models [3, 43] to compute affinity. Appearance models have long been exploited to discriminate objects, represented by bounding boxes, employing their appearances or local features [22, 45]. However, these methods suffer from misleading appearance similarities among objects across frames that is caused by illumination changes, noisy background pixels, and very similar cloths of pedestrians. On the other hand, the motion model has been utilized to discriminate among

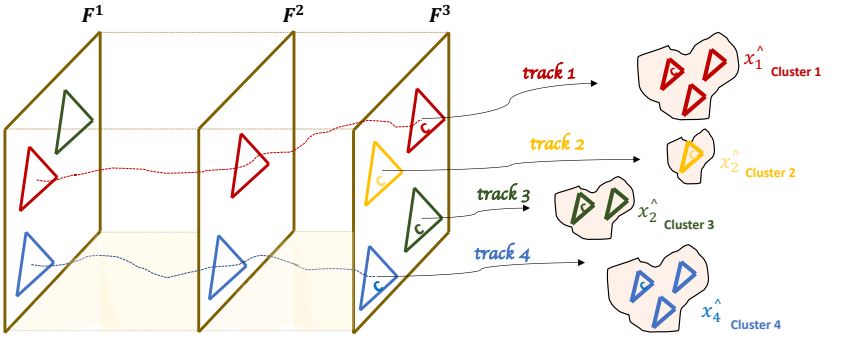

Figure 1: Toy-example of the proposed method, where the triangles in each frame denote objects to be tracked and the colors of the triangle define their identities. F^1, F^2 and F^3 refer to video-frames at $time_1, time_2$, and $time_3$ respectively. c denotes the constraints or objects in the current frame, and each cluster correspond to object track, represented by $x^\wedge$.

pedestrians, as customary, these models assume a constant velocity assumption(CVA) [6]. Nevertheless, the pedestrian motion gets more complex in crowded scenarios. In addition to the linear models, recent methods also attempt to overcome the real-world motion scenario, by employing complex motion models, in terms of non-linear models [46]. However, both the linear and non-linear models fail to tackle object occlusion. To deal with this, a composite model attempts to combine appearance and motion models, though, fail to succeed in a real-world scenario. A majority of above method use off the shelf hand crafted appearance features. To overcome some of the limitations of the aforementioned methods, Deep Affinity Network (DAN) [33] proposed an affinity-based deep neural network model. DAN jointly *learns* target object appearance and their affinities in a pair of video frames in an end-to-end manner. However, instead of exploiting the global information across video frames, DAN only considers objects in *a pair of video-frames*. Thus, it is not sufficiently effective to tackle the problem of occlusion involving multiple frames. In most two-frame MOT methods data association is formulated as a bi-partite graph matching problem and solved employing Hungarian Algorithm.

In this work, we propose a novel deep learning method called Multi-frames Affinity-Based Deep Association ($MADA-NET$). We formulate a multi-object tracking problem as a differentiable parametrized quadratic program, implying an end-to-end clustering. As a result, our method is able to exploit the global relationship among objects across *multiple* frames to determine optimal object trajectories. To this end, we adopt the well-known graph and game theoretic clustering algorithm called dominant sets clustering. In our formulation, we cast the concept of a track as a cluster, and the object in the current frame as a constraint; subsequently, we constrain the clustering process to search for objects across the previous frames with the same identities as of the constraint object. Unlike the predominant deep association methods that only consider a pair of video frames to perform online object association; our clustering formulation enables us to exploit objects across multiple previous frames to determine the association between a sequence of consecutive frames. Thereby, we effectively tackle longstanding multi-object tracking challenges such as object occlusion, object entry and exit and miss detection. Unlike DAN [33], our formulation operates on a symmetrized k-partite graph, thus, it naturally handles unidentified objects without the need for further modifications of the affinity matrix. Furthermore, to the best

of our knowledge, this is the very first method that incorporates a clustering scheme in an end-to-end manner to tackle the multi-frame tracking problem. The conceptual diagram of the proposed method is depicted in the toy-example, Figure 1. As can be seen from the toy-example Figure 1, due to the extended search space, involving multiple frames, the clustering process can effectively retrieve the occluded object, green triangle, which is occluded in Frame 2. The rest of the paper is structured as follows. In the next section 2, we briefly review the literature, which is followed by the proposed method in section 3; subsequently extensive experiments are presented in section 4, finally the paper is concluded in section 5.

2 Related Works

In this section we briefly discuss the related works, we refer the readers to [10] for in-depth literature review of recent MOT methods. Due to the effectiveness of recently proposed object detection techniques tracking-by-detection paradigm [6, 7, 26] has been dominantly used for multi-object tracking problem. The most common multi-object tracking schemes are broadly divided into two major categories such as online and offline methods. Online MOT methods [28, 39, 47] determine the association between objects based on the current and the previous frames, while offline methods [11, 40, 41, 48] leverage the all video frames to form trajectories. Thus, online methods are preferred for real-time applications. Data association in tracking can be performed in either local or global association techniques. Local data association [35, 37, 38] techniques leverage a pair of video frames to perform data association. Nevertheless, such multi-object tracking algorithms have critical limitations in handling occlusions or noisy detections and consequently tend to produce short fragmented trajectories. On the other hand, global techniques use a larger number of video frames to construct object trajectories. The network flow formulation [60] is an example of the global association techniques, that exploit multiple video frames to perform data association in both online and offline scenarios.

Classical global MOT methods (CGM). Among the CGM, we will review the most related methods to the proposed scheme that cast tracking problems in a graph formulation, such as GMMCP[11], CDSMOT [41], and DSMOT [40]. GMMCP[11] solves tracking as a generalized maximum clique problem. However, this method has a limitation of introducing dummy nodes, which is computationally expensive to handle outliers and associate weak-detections. In DSMOT [40], Tesfaye et. al. cast a tracking problem as finding a dominant set cluster from a fully connected edge-weighted graph. Nevertheless, this method adopts a pill-off strategy to enumerate all possible clusters, as a result, it alters the original size of the graph. CDSMOT method proposed in [41] does not require extra nodes to handle occlusion nor change in scale of the problem. However, all the aforementioned graph-based global methods do not employ any learning mechanism in order to exploit the advantage of deep learning. Also, these methods perform feature extraction and hence object affinities separately from the data association therefore are not end-to-end methods.

Deep MOT methods. Recently, there have been several methods ranging from contrastive loss based [16, 33], recurrent neural network [13], to graph-based methods [8, 34, 38] that attempt to solve the data association problem in an end-to-end manner. The authors in DAN [38] jointly model the appearance of objects in a pair of frames employing *bi-partite* graph and their association in an end-to-end manner. Thereby, they are able to model the

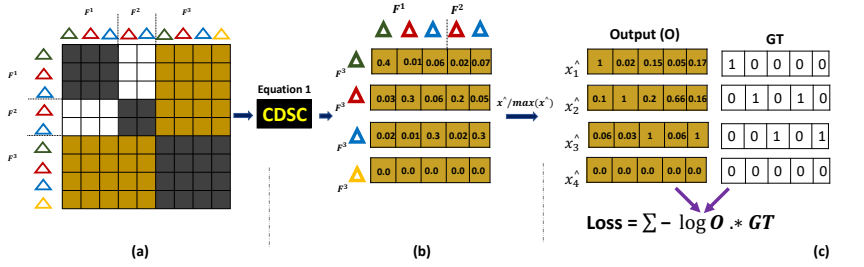

Figure 2: We show the structure of our model output and the cross-entropy loss computation. (a) demonstrate the multi-frame affinity matrix representation of our toy-example in figure 1; the matrix entries corresponding to objects in the same frame is set to 0, marked by black; and, the similarities among previous frames object is represented by white, while the similarity between objects in the current frame (F^3) and previous frames (F^1&F^2) is marked as golden. (b) Each row shows the characteristics vector obtained from CDSC, taking F^3 objects as a constraint. Followed by a normalization operation. (c) Shows cross-entropy loss computation, where, $.^*$ defines elementwise multiplication.

appearance and leverage it to estimate object affinities in a pair of frames. Nonetheless, DAN [32] takes extra measures to handle missing, or entering new objects and exiting objects in the affinity construction step. For instance, DAN appends a dummy variable to manage the missing objects.

In order to overcome aforementioned limitations, in this paper we propose a flexible affinity-based approach that can encode the relationship among objects across *multiple frames* with a symmetric *k-partite* graph and extract the underlying track in a form of constrained cluster. Moreover, the proposed work naturally handles entering, exiting, and occluded objects without the need for further modification of the affinity matrix by adding extra rows and columns.

3 MOT as a Deep Clustering problem

In the following, we elaborate on how we cast the complex multi-object tracking problem into an end-to-end multi-frame based data association. We first define some of the notations we use in this paper; we then provide a brief introduction to constrained dominant sets clustering, followed by section 3.2 which deals with the formulation of tracking problems as a differentiable parametrized quadratic program, implying an end-to-end clustering.

3.1 Problem formulation and Notations

Let us first define some of the notations we use in this paper. Given a video sequence S, we denote the video-frame at time t as F^t. The features corresponding to each objects, bounding boxes, in F^t are denoted by B^t_u, where $u \in 1, 2, ...L$ and L is the total number of bounding boxes in the corresponding frame F^t.

3.2 Constrained Dominant Sets Clustering

Constrained dominant sets clustering (CDSC) [49] is an extension of dominant sets clustering [29] that allows one to enforce a constraint on the clustering process such that the

resulting cluster contains the specified constraint. Dominant sets clustering (DSC) [29] is a graph-theoretic notion of a cluster, which generalizes that of a maximal complete subgraph to edge-weighted graphs. Unlike some clustering methods, for instance, K-means [19], DSC does not need to know the number of clusters in the given data in advance. First, the data to be clustered are represented as an undirected edge-weighted graph with no self-loops $G = (V, E, w)$, where $V = \{1, ..., N\}$ is the vertex set, $E \subseteq V \times V$ is the edge set, and $w : E \rightarrow R^*_+$ is the (positive) weight function. Vertices in G define data points, edges correspond to neighborhood relationships, and edge-weights reflect the similarity between pairs of linked vertices. We then represent the graph G with the corresponding weighted adjacency (or similarity) matrix, which is the $N \times N$ nonnegative, symmetric matrix $A = (a_{ij})$, defined as $a_{ij} = w(i, j)$, if $(i, j) \in E$, and $a_{ij} = 0$ otherwise. The diagonal of A corresponding to each node is always set to zero indicating that there are no self-loops in graph G. Zemene et. al [49] extends DSC [29] into a parametrized quadratic program, and show that by properly controlling a regularization parameter, which determines the structure and the scale of the underlying problem, one can extract groups of DSC that are constrained to contain predefined constraints. When the constraint is zero the local solutions are known to be in one-to-one correspondence with the dominant sets. A highly coherent constrained cluster can easily be extracted by defining the following parametrized quadratic equation as

$$\begin{array}{ll} \text{maximize} & f_c^{\alpha}(x) = \mathbf{x}'(A - \alpha \hat{I}_c)\mathbf{x}, \\ \text{subject to} & \mathbf{x} \in \Delta \end{array} \tag{1}$$

where, Δ is the standard simplex of $\mathbb{R}^N$. $\hat{I}_c$ represents $N \times N$ diagonal matrix whose diagonal elements are set to 1 in correspondence to the vertices contained in $V \setminus c$ and to zero otherwise. According to [49], we have a theoretical gurantee that if $\alpha > \lambda_{max}(A_{V \setminus c})$, where $\lambda_{max}(A_{V \setminus c})$ is the largest eigenvalue of the principal submatrix of A indexed by the elements of $V \setminus c$. Hence, if $\mathbf{x}$ is a local maximizer of $f_c^{\alpha}(\mathbf{x})$ in Δ, then $\delta(\mathbf{x}) \cap c \neq \emptyset$, where, $\delta(\mathbf{x}) = i \in V : \mathbf{x}_i > 0$. We refer the reader to [49] for the proof. Equation 1 can be simply solved with a straightforward continuous optimization technique from evolutionary game theory called replicator dynamics [42].

3.3 Deep CDSC

In this section, we discuss how we contextualize the constrained dominant sets clustering in a multi-object tracking problem. We cast a multi-object tracking problem as finding a constrained cluster from a complete k-partite graph G that is constructed accounting for multiple frames in the given video sequences, S. To this end, we first represent the given video sequence, S, by a k-partite graph G; we then apply the graph and game theoretic clustering scheme, CDSC, in a differentiable manner. The output, *constrained* cluster, is treated as a track. The constraint here is that an object in the current frame, represented by B_u^t, and other members of the cluster collected from the previous frames, represent the same object. The solution cluster membership-scores ($x^{\wedge}$) define the degree of confidence about how well the objects in previous frames, B_u^{t-n}, match the constraint-object, B_u^t in the current frame. In other terms, we cast the data association problem as a clustering problem.

K-partite graph construction. Recent methods that exploit affinity matrix for data association, DAN [38], take extra measures to handle missing objects in the affinity construction step. For instance, appending a dummy variable to handle entering or exiting and missing

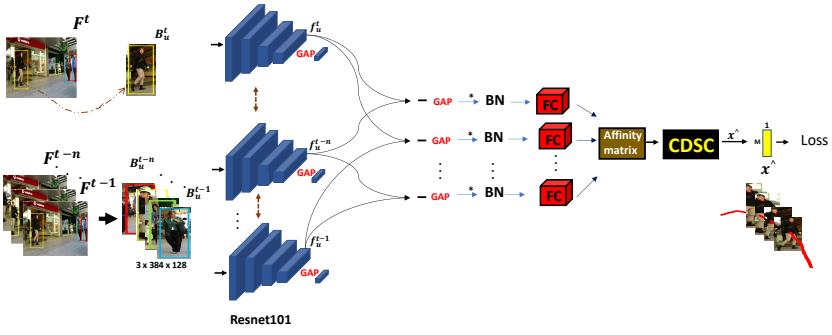

Figure 3: The pipeline of our mode. The upper frame, F^t, shows the current frame while the one below, $F^{t-1}...F^{t-n}$, show the previous frames. And, B_u^t and $B_u^{t-n}...B_u^{t-1}$ depicts the objects from F^1, to be searched, and objects collected from the previous frames, respectively. For each object, we extract the feature, f_u^t, right before the GAP layer of ResNet101; that is followed by the computation of objects difference $(-)$, *GAP*, elementwise square $(*)$ and Batch Normalization (BN). Subsequently, we fed the difference to the *FC* layer to find a scalar value. After constructing N x N affinity matrix, we run CDCS to find the characteristics vector $x^\wedge$; we finally apply max normalization and compute the cross-entropy loss.

objects. However, these measures are unnatural and need additional manually fixed parameters. In this work, we naturally handle exiting and newly entering objects through our formulation. As can be seen from Figure 1, one of the objects, green triangle, leaves the sequence at frame F^2, since we are encoding these relationships among objects in a symmetric multi-frame affinity matrix (see Figure 2), we do not need extra row and column in our model. In addition, proposed method can also deal with a newly appearing object in F^3 shown by yellow-triangle, which results in a singleton cluster; the membership score corresponding to the remaining vertices in this cluster is zero (see Figure 2 (b)). Moreover, we organize the ground truth, Figure 2(c), in a way that can seamlessly suit the loss computation. Intriguingly, besides the aforementioned perks of our model, leveraging k-partite graph allows us to enforce $can-not-link$ constraints among objects that belong to the same frame. As a result, we assure that objects from the same frame are not be grouped in the same cluster.

A pipeline of the proposed model is shown in Figure 3. We adopt a Siamese like network as in [1, 36] to compute object features as follows. Given the current video frame, F^t, and previous video frames, $F^{t-1},...,F^{t-n}$, along with the object coordinates obtained from the object detection module; we first crop image patches of size 384 x 124, bounding boxes B_u^{t-n}, of current and previous video-frames. We then extract the object features f^t from right before the global average pooling layer (GAP) of the ResNet101 network. Next, we compute the difference using elementwise subtraction and apply Global Average Pooling (GAP), batch normalization (BN), and finally a fully connected layer (FC) to find the similarity between a pair of objects. Using these pairwise similarities among the objects in the current frame, B_u^t, and objects in the previous frames, B_u^{t-n}, we construct the complete k-partite graph G. Finally, we apply the *CDSC* to obtain the characteristics vector, $x^\wedge$ that defines the cluster membership probability of each node in the graph.

Building k-partite Graph In addition to the appearance-similarity, we leverage the IOU

	MOTA ↑	MOTP ↑	IDF1 ↑	MT ↑	ML ↓	FP ↓	FN ↓
FHFD [18]	51.3	-	47.6	21.4	35.2	24101	247921
DMAN [52]	48.2	75.5	**55.7**	19.3	38.3	26218	263,608
Tracktor++ [4]	53.5	78.0	52.3	19.5	36.6	12201	248047
FAMNet [8]	52.0	76.5	48.7	19.1	33.4	14138	253616
DAN [38]	52.4	76.9	49.5	21.4	**30.7**	25423	234592
Ours	**57.6**	**78.5**	52.9	**22.8**	33.7	**10,823**	**223,951**

Table 1: A comparison with state-of-the-art methods on the MOT17 benchmark.

between a pair of objects to learn their spatial overlap. Therefore, we build the complete k-partite graph (section 3.2, equation 1), employing affinity matrix which consists of feature similarities, A_m, and intersection-over-union between two corresponding bounding boxes, A_{IOU}, as follows: $\gamma A_m + (1-\gamma)A_{IOU}$, where, γ is a fusing parameter that is used to tune the similarity fusion, empirically set to 0.8 in our experiments.

Loss computation. We employ cross-entropy loss to train our model. To do so, we normalize the output (O) (see Figure 2), the characteristics vector, $x^{\wedge}$, excluding the membership score of the constraint node, by dividing it with the maximum value.

4 Experiments

We demonstrate the validity of the proposed end-to-end deep clustering-based tracking approach by conducting several experiments using benchmark datasets and performing ablation studies to verify the contribution of different components of the proposed model.

Datasets: we use different benchmark pedestrian datasets from MOTChallenge [27], namely 2D MOT2015 [23] and MOT17[27]. These datasets provide crowded pedestrian video sequences captured in real-world outdoor and indoor scenarios. The 2D MOT15 dataset contains 11 video sequences and also provides ACF [12] detection. The MOT17 dataset comprises 14 video sequences, divided into 7 sequences for training and 7 for testing.

Metrics: As the problem of tracking is complex, it needs to be assessed from different aspects. Thus, in addition to the widely used CLEAR MOT metrics [5], such as MOTA, MOTP, false negatives (FN), false positives (FP), IDF1 score [31]; we also adopt metrics like percentage of mostly lost targets (ML) and mostly tracked targets (MT) from [24]. MOTA and IDF1 score shows the object coverage and Identity; MOTP defines the misalignment between the annotated and the predicted bounding boxes. MT and ML show the ratio of ground-truth trajectories that are covered by a track hypothesis for at least 80 % and at most 20 % of their respective life span, respectively.

4.1 Implementation Details

Following recent multi-object tracking methods we asses our model using publicly available detections provided by the MOTChallenge, such as DPM [14], FRCNN [30], and SDP [12]. Since the public detections in MOTChallenge [27] are prone to error, we introduce bounding boxes filtering mechanism whereby we filter out noisy bounding boxes. To detect bounding

	MOTA ↑	MOTP ↑	IDF1 ↑	MT ↑	ML ↓	FP ↓	FN ↓
KFC [9]	38.9	70.6	44.5	16.6	31.5	7321	29501
DAN [33]	38.30	71.10	45.60	17.60	41.20	1290.25	**2700**
FAMNet [8]	40.6	71.1	-	12.5	34.4	4678	31018
Tracktor++ [4]	44.1	75.0	46.7	18.0	**26.2**	6477	26577
Ours	**47.0**	**76.0**	**47.9**	**121**	186	**4524**	28691

Table 2: A comparison with state-of-the-art methods on the 2D MOT15 benchmark.

Figure 4: Shows the qualitative results of our model, from the MOT17 dataset video sequence 12, 03 and 08 on top, middle, and bottom, respectively.

boxes, we adapt Faster R-CNN [30] with ResNet101 [17] and Feature Pyramid Network (FPN) [25], that is trained using MOT17Det [27] detection dataset. As a result, objects which pass the pruning stage are included in the subsequent graph construction process, otherwise, they are put into an inactive track-list. Our $MADA-NET$ model is built upon ResNet101 [17] network, that is pre-trained on Imagenet dataset.

4.2 Results on MOTChallenge Benchmarks

We evaluate the effectiveness of the proposed scheme experimenting on the MOT17 and MOT15 benchmarks. Table 1 shows the performance of our model on the MOT17 dataset with a comparison to published online state-of-the-art methods for different metrics. As can be observed, our model surpasses most of the recent online multi-object tracking methods. Furthermore, the proposed scheme outperforms the affinity-based MOT method, DAN [33], which exploits a pair of video-frames through the Hungarian association. Our method also surpasses the performance of the recent multi-frame affinity-based method called FamNet [8]. In particular, the proposed approach achieves a smaller number of false positives and negatives compared to its counter methods, DAN [33], and FamNet [8]. Table 2 shows the performance of our model on the 2D MOT15 [23], as can be seen our model achieve state-of-the-art performance; furthermore, it outperforms recent affinity-based tracking methods [34], [33], [8] in different metrics. Thereby, summing up the improvements in different metrics we can draw our conclusion that the proposed end-to-end clustering-based MOT method has a clear benefit on improving the quality of multi-object tracking models.

k	MOTA ↑	MOTP ↑	IDF1 ↑	MT ↑	ML ↓	FP ↓	FN ↓
2	57.10	80.50	60.40	22	35	125	22581
4	58.50	82.30	67.10	17	24	83	18453
6	62.40	85.20	66.40	43	36	87	15433
10	65.5	85.9	68.7	34	22	74	16472

Table 3: Impact of numbers of video frames, parameter k, on the MOT17 training set sequence 04.

		MOTA↑	MOTP↑	IDF1↑	MT↑	ML↓	FP↓	FN↓
Seq-04	Without CDSC	61.2	85.1	66.8	23	34	77	16612
	With CDSC	65.5	85.9	68.7	34	22	74	16472
Seq-13	Without Out CDSC	70.4	86.2	68.3	63	15	400	2814
	With CDSC	73.1	88.1	70.0	64	9	349	2722

Table 4: Ablation study on the MOT17 dataset, training set, taking the average of the results obtained using three detectors such as DPM [14], FRCNN [30], and SDP [12].

4.3 Ablation Study

In order to asses the impact of the proposed deep clustering module in our tracking pipeline, we also conduct ablation studies by removing different components of the proposed model. We utilized MOT17 training set sequences to perform the ablation study. We consider MOT17-04 and MOT17-13 sequences as a test set while using remaining sequences as a training set. We first experimented excluding the clustering module both at the training and testing phase. Consequently, we conducted the test without leveraging the CDSC module, Without CDSC. As can be seen from Table 4, the results we obtained using the clustering module surpass the result without utilizing the clustering module. Thereby, from the results in Table 4, we can witness the benefits of exploiting the global relationship among objects across multiple frames, through the deep clustering, to determine their association with the newly detected object in the current frame. Thus, our model effectively leveraged clustering in an end-to-end manner and demonstrate its effectiveness in the reported experimental analysis.

Moreover, to analyze the impact of leveraging global information through exploiting multiple frames, we conduct ablation study by varying the number of frames, k, used to do the association. As can be observed from Table 3, setting k to a larger value has shown its advantages over using only a pair of images to decide the track-id of the current frame objects. In addition to that, in the qualitative result (Figure 4), it can be seen that *MADA* − *NET* has effectively handle occlusions.

5 Conclusion

In this work, we presented a novel multi-object tracking scheme that cast a multi-object tracking problem into an end-to-end clustering problem. Furthermore, unlike existing methods, our scheme can easily be able to leverage multiple frames, more than two, to perform the data association in an end-to-end learning process. Extensive experiments demonstrate the benefit of treating track as a constrained cluster and optimize it in an end-to-end fashion. Moreover, the reported result achieves the state-of-the-art result.

References

[1] Leulseged Tesfaye Alemu, Mubarak Shah, and Marcello Pelillo. Deep constrained dominant sets for person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9854–9863, 2019.

[2] Seung Hwan Bae and Kuk-Jin Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1218–1225. IEEE Computer Society, 2014.

[3] Seung Hwan Bae and Kuk-Jin Yoon. Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(3):595–610, 2018.

[4] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 941–951, 2019.

[5] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP J. Image and Video Processing*, 2008, 2008.

[6] Michael D. Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 1515–1522. IEEE Computer Society, 2009.

[7] Wongun Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3029–3037. IEEE Computer Society, 2015.

[8] Peng Chu and Haibin Ling. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6171–6180. IEEE, 2019.

[9] Peng Chu, Heng Fan, Chiu Chiang Tan, and Haibin Ling. Online multi-object tracking with instance-aware tracker and dynamic model refreshment. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pages 161–170. IEEE, 2019.

[10] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, 2020.

[11] Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah. GMMCP tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4091–4099. IEEE Computer Society, 2015.

[12] Piotr Dollár, Ron Appel, Serge J. Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(8):1532–1545, 2014.

[13] Kuan Fang, Yu Xiang, Xiaocheng Li, and Silvio Savarese. Recurrent autoregressive networks for online multi-object tracking. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*, pages 466–475, 2018.

[14] Pedro F. Felzenszwalb, David A. McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. IEEE Computer Society, 2008. doi: 10.1109/CVPR.2008.4587597. URL https://doi.org/10.1109/CVPR.2008.4587597.

[15] François Fleuret, Jérôme Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):267–282, 2008. doi: 10.1109/TPAMI.2007.1174. URL https://doi.org/10.1109/TPAMI.2007.1174.

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 1735–1742, 2006.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[18] Roberto Henschel, Laura Leal-Taixé, Daniel Cremers, and Bodo Rosenhahn. Fusion of head and full-body detectors for multi-object tracking. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1428–1437. IEEE Computer Society, 2018.

[19] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[20] Margret Keuper, Siyu Tang, Bjoern Andres, Thomas Brox, and Bernt Schiele. Motion segmentation & multiple object tracking by correlation co-clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(1):140–153, 2020.

[21] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M. Rehg. Multiple hypothesis tracking revisited. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4696–4704. IEEE Computer Society, 2015.

[22] Cheng-Hao Kuo, Chang Huang, and Ram Nevatia. Multi-target tracking by on-line learned discriminative appearance models. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 685–692. IEEE Computer Society, 2010.

[23] Laura Leal-Taixé, Anton Milan, Ian D. Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *CoRR*, abs/1504.01942, 2015.

[24] Yuan Li, Chang Huang, and Ram Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 2953–2960. IEEE Computer Society, 2009.

[25] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 936–944. IEEE Computer Society, 2017.

[26] Anton Milan, Stefan Roth, and Konrad Schindler. Continuous energy minimization for multitarget tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(1):58–72, 2014.

[27] Anton Milan, Laura Leal-Taixé, Ian D. Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *CoRR*, abs/1603.00831, 2016.

[28] Anton Milan, Seyed Hamid Rezatofighi, Anthony R. Dick, Ian D. Reid, and Konrad Schindler. Online multi-target tracking using recurrent neural networks. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, 2017.

[29] Massimiliano Pavan and Marcello Pelillo. Dominant sets and pairwise clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(1):167–172, 2007.

[30] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015.

[31] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. 2016.

[32] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 300–311, 2017.

[33] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823, 2015.

[34] Samuel Schulter, Paul Vernaza, Wongun Choi, and Manmohan Chandraker. Deep network flow for multi-object tracking. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2730–2739, 2017.

[35] Khurram Shafique and Mubarak Shah. A noniterative greedy algorithm for multiframe point correspondence. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(1):51–65, 2005.

[36] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2265–2274, 2018.

[37] Guang Shu, Afshin Dehghan, Omar Oreifej, Emily M. Hand, and Mubarak Shah. Part-based multiple-person tracking with partial occlusion handling. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 1815–1821, 2012.

[38] ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal Mian, and Mubarak Shah. Deep affinity network for multiple object tracking. *CoRR*, abs/1810.11780, 2018.

[39] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3701–3710. IEEE Computer Society, 2017.

[40] Yonatan Tariku Tesfaye, Eyasu Zemene, Marcello Pelillo, and Andrea Prati. Multi-object tracking using dominant sets. *IET Comput. Vis.*, 10(4):289–297, 2016.

[41] Yonatan Tariku Tesfaye, Eyasu Zemene, Andrea Prati, Marcello Pelillo, and Mubarak Shah. Multi-target tracking in multiple non-overlapping cameras using fast-constrained dominant sets. *Int. J. Comput. Vis.*, 127(9):1303–1320, 2019.

[42] Jörgen W Weibull. *Evolutionary Game Theory*. MIT press, 1995.

[43] Longyin Wen, Wenbo Li, Junjie Yan, Zhen Lei, Dong Yi, and Stan Z. Li. Multiple target tracking based on undirected hierarchical relation hypergraph. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1282–1289. IEEE Computer Society, 2014.

[44] Yu Xiang, Alexandre Alahi, and Silvio Savarese. Learning to track: Online multi-object tracking by decision making. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4705–4713. IEEE Computer Society, 2015.

[45] Kota Yamaguchi, Alexander C. Berg, Luis E. Ortiz, and Tamara L. Berg. Who are you with and where are you going? In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 1345–1352. IEEE Computer Society, 2011.

[46] Bo Yang and Ram Nevatia. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 1918–1925. IEEE Computer Society, 2012.

[47] Ju Hong Yoon, Chang-Ryeol Lee, Ming-Hsuan Yang, and Kuk-Jin Yoon. Online multi-object tracking via structural constraint event aggregation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1392–1400. IEEE Computer Society, 2016.

[48] Amir Roshan Zamir, Afshin Dehghan, and Mubarak Shah. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II*.

[49] Eyasu Zemene and Marcello Pelillo. Interactive image segmentation using constrained dominant sets. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, volume 9912 of *Lecture Notes in Computer Science*, pages 278–294. Springer, 2016.

[50] Li Zhang, Yuan Li, and Ramakant Nevatia. Global data association for multi-object tracking using network flows. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*, 2008.

[51] Shengping Zhang, Xiangyuan Lan, Hongxun Yao, Huiyu Zhou, Dacheng Tao, and Xuelong Li. A biologically inspired appearance model for robust visual tracking. *IEEE Trans. Neural Networks Learn. Syst.*, 28(10):2357–2370, 2017.

[52] Ji Zhu, Hua Yang, Nian Liu, Minyoung Kim, Wenjun Zhang, and Ming-Hsuan Yang. Online multi-object tracking with dual matching attention networks. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*.