CL-MOT: A Contrastive Learning Framework for Multi-Object Tracking

Daniel Silva
danielzgsilva@knights.ucf.edu
Leulseged Tesfaye Alemu
leule.24@gmail.com
Mubarak Shah
shah@crcv.ucf.edu

Center for Research in Computer Vision
University of Central Florida
Orlando, FL

Abstract

The field of Multi-Object Tracking (MOT) has seen tremendous progress with the emergence of joint detection and tracking algorithms. Nevertheless, these approaches still rely on supervised learning schemas with prohibitive annotation demands. To this end, we present CL-MOT, a semi-supervised contrastive learning framework for multi-object tracking. By clustering object embeddings from different views of static frames, CL-MOT learns to discriminate between objects in a scene without relying on identity labels. This simple, yet profound pre-text learning paradigm enables one to transform any object detector into a tracker with no additional annotation labor. We evaluate the proposed approach on the MOT Challenge benchmark, finding that CL-MOT performs on par with supervised trackers and significantly outperforms other self-supervised methods. Interestingly, we even set a new state-of-the-art in re-identification metrics such as IDF1 score, suggesting that CL-MOT learns a more effective feature space than its counterparts. These results demonstrate that our within-frame contrastive learning framework can significantly reduce the annotation demands of tracking while recovering if not surpassing the performance of previous approaches. The code for this work is publicly available at https://github.com/danielzgsilva/CL-MOT

1 Introduction

The ability to track multiple objects in a dynamic environment is a core primitive to perceiving any visual scene. Therefore, it is critical in many down-stream applications such as autonomous navigation, video surveillance, human-computer interaction, and more. Formally known as Multi-Object Tracking (MOT), this task involves detecting object instances while maintaining their identities throughout the frames of a video.

The most common approaches to this problem leverage a paradigm known as tracking-by-detection [3, 5, 16, 21, 33]. These trackers typically utilize off-the-shelf detectors [13, 26, 27] to first localize objects within a video, then link predictions through time with a secondary association step [3, 33]. Recently, however, simultaneous detection and association methods [1, 31, 32, 38, 40] have surpassed the performance of tracking-by-detection approaches. Capable of performing both steps with a single forward pass, these one-shot

Figure 1: CL-MOT requires only bounding box annotations.

Figure 2: We do not require object identity labels.

trackers are significantly less complex, notably faster, and even allow for real-time object tracking. While some of these methods handle association from a local, frame-by-frame perspective [1, 40], the most robust trackers are able to re-identify objects over large windows of time by leveraging appearance information [32, 38].

Despite this success, these methods still rely on large amounts of object identity annotations, posing a costly, time consuming and error prone prerequisite to tracking. Concurrently, recent advancements in contrastive learning [4, 6, 14] have produced power self-supervised frameworks for learning visual representations. However, despite appearance-based tracking having shown such promise [21, 32, 33, 38], there has not been much work in extending these learning principles to tracking.

Inspired by this recent work, we propose CL-MOT, a semi-supervised learning framework for multi-object tracking. Our training approach is unique in that we contrast embeddings retrieved from single images containing multiple objects. Therefore, CL-MOT learns a feature space for object re-identification in a self-supervised manner, allowing us to forgo identity labels and significantly reduce the annotation demands of tracking.

Details regarding our training and online tracking method can be found in section 4, while our network architecture is described in 3. In section 5.4, we compare CL-MOT to the state-of-the-art trackers on the MOT Challenge benchmark [25]. Our results show that CL-MOT performs comparatively to supervised approaches and significantly outperforms other self-supervised methods. Moreover, we set a new top score in re-identification metrics such as IDF1, indicating that CL-MOT learns more robust and functional features than its counterparts. Our ablation studies in section 5.3 also investigate the effects of different contrastive loss functions, including normalized temperature-scaled cross entropy loss [6] and triplet loss with various mining strategies [15]. In summary, our central contribution is a novel learning framework that not only enables one to train a tracker without global identity labels, but even results in improved discriminative and re-identification ability.

2 Related Work

Multi Object Tracking Datasets . In order to craft a multi-object tracking (MOT) dataset, one must manually track an unknown number of objects across the frames of a video. This typically involves annotating the frames with bounding boxes while maintaining each object's global identity throughout the dataset. Even for humans, this is quite a difficult and time-consuming task, especially once considering that objects may leave or enter the scene at any time and must be re-identified over periods of occlusion or appearance change. To illustrate the effort and cost required, Manen *et al*. [24] show that accurately annotating just 6 minutes worth of video from the MOT15 benchmark [20] requires at least 22 hours

of annotation time. Thus, scaling supervised approaches to large datasets is clearly inefficient, expensive and error prone. In spite of this, there have been several smaller-scale MOT datasets created, such as the MOT Challenge [8, 20, 25] and Caltech [10] datasets, which focus on pedestrian tracking.

Joint Detection and Tracking. As mentioned above, tracking typically involves detecting objects and associating each prediction with a previous or new identity, for each subsequent frame in a video. Recently, much work has been done to accomplish both tasks in a single framework, referred to as joint detection and tracking or one-shot tracking. We categorize these one-shot trackers into two high-level groups, the first of which containing those that approach tracking from a local perspective by modeling object motion. For example, Bergmann *et al*. [1] track objects by predicting their future position based off the current frame, and Zhou *et al*. [40] learn a 2D location offset for objects in consecutive frames. While these methods excel in the local regime, they struggle to track objects through periods of occlusion or low frame-rate video. In contrast, the second category of trackers tackle the association step by leveraging appearance information. For instance, approaches such as [31, 32, 38] predict object embeddings, along with their bounding boxes, and link predictions in an online manner based on feature similarity. This allows trajectories to be recovered over large windows of time or occlusion, resulting in more robust performance in the global regime. Despite their differences in association strategy, these methods are all trained in a supervised manner, requiring object identity labels.

Contrastive Learning of Visual Representations. Contrastive learning is a general learning paradigm that aims to maximize agreement between the representations of different views of the same data. Thus, a network is trained to produce consistent and robust feature vectors for class instances, without relying on manual annotations. Typically, this paradigm is used as a pre-training schema for image classification. As such, Dosovitskiy *et al*. [11] treat each image as a distinct class and train a linear classifier on as many classes as images in the dataset. Being that this approach does not scale well to large datasets, Wu *et al*. [34] propose to store and sample pre-computed class representations from a memory bank. To simplify the contrastive learning framework, several other works [9, 17, 36] suggest using samples from within a training batch as contrastive pairs. Most recently, this approach was refined and popularized by Chen *et al*. [6], providing a simple yet effective self-supervised framework for visual representation learning. In CL-MOT, we leverage a similar approach to learn a feature space for object re-identification. However, rather than sampling contrastive pairs from within a batch, we sample them from a single image containing multiple objects.

3 Preliminaries

With the goal of leveraging contrastive learning to train an appearance-based tracker, we adopt the multi-task network architecture presented in [38]. Comprised of an encoder-decoder backbone, along with separate object detection and embedding branches, this one-shot tracking network predicts object bounding boxes and appearance embeddings in a single forward pass. Further information regarding the network architecture is provided in the following sections.

3.1 Backbone Network

We leverage a fully-convolutional ResNet-34 [12, 35] as our backbone with the Deep Layer Aggregation [37] variant proposed in [39]. In addition, we replace all convolutions in the up-sampling layers with deformable convolutions [7], which have been shown to improve performance by dynamically adapting receptive fields according to object scales and poses.

3.2 Object Detection Branch

Following [39], we approach object detection as a center-based keypoint estimation task and regress to other properties such as height and width. Therefore, three parallel regression heads are appended to the backbone network to predict an object heatmap, object center offsets and bounding box sizes, respectively.

Heatmap Head. This head is responsible for predicting an object heatmap $H \in [0,1]^{\frac{H}{R} \times \frac{W}{R} \times 1}$, with a default down-sampling factor R of 4. Local maximums in this heatmap should reflect ground truth object centers, with the response decaying exponentially as the distance between the location in the heatmap and the object center increases.

Center Offset Head. The output stride R used in the heatmap and embedding heads will introduce non-negligible discretization error with respect to object center points. To solve this, [39] introduces a head that estimates the offset $o^i \in \mathbb{R}^2$ between an object's original center point and its location in the down-scaled heatmap. Zhang *et al*. [38] also show that this is critical for ensuring that appearance embeddings are extracted from the accurate down-scaled object centers.

Box Size Head. This head simply estimates the size $s^i \in \mathbb{R}^2$, or height and width, of the bounding box at each predicted object center.

3.3 Object Embedding Branch

The embedding branch aims to generate feature vectors that can be used to re-identify objects over time. Ideally, the similarity between embeddings from the same object instance should be larger than that between different objects. Formally, this head outputs a global feature map $E \in \mathbb{R}^{\frac{H}{R} \times \frac{W}{R} \times C}$, from which an appearance embedding $E^i_{x,y} \in \mathbb{R}^C$ for each object at position (x,y) can be extracted from. Note that C is a hyper-parameter representing the size of the output embeddings. Unless otherwise stated, we use a default C value of 512, which we find results in the strongest performance.

4 Proposed Approach

In this paper, we treat tracking as an online multi-object re-identification task. Therefore, we aim to predict embeddings and bounding boxes for each object in a frame, then match predictions to previous tracklets based on appearance. Drawing motivation from recent contrastive learning work [4, 6, 14], CL-MOT clusters object embeddings retrieved from different views of a static image. By doing so, our network learns a feature space that discriminates between object instances in a single scene. We combine this representation learning framework with an appearance-based association algorithm to achieve online and real-time object tracking. Thus, CL-MOT "learns to track" objects in a self-supervised manner, forgoing the identity annotations typically required for tracking. We do, however, require bounding box

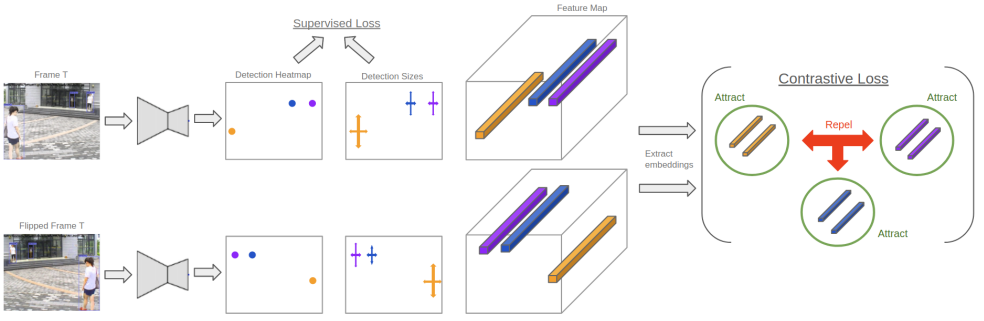

Figure 3: Overview of our semi-supervised contrastive learning approach for joint object detection and association.

annotations in order to constrain the localization capabilities of our network. Our technical approach is simple, yet profound, enabling users to train an appearance-based tracker without the costly demands of labeling object identities across frames. The CL-MOT training pipeline and losses are defined below.

4.1 Training Pipeline

Illustrated in Figure 3, the CL-MOT training pipeline begins with a single image $I_a \in \mathbb{R}^{H\times W\times 3}$ and its corresponding bounding box annotations $b_a^i \ \forall$ objects i. This image is then copied and horizontally flipped to create I_b, a different view of the original scene. Note that I_a and I_b are also augmented with random rotation, translation, scaling and color jittering to increase the contrast between the two views. All geometric augmentations, such as the horizontal flip or rotation, are also applied to b_a, allowing us to maintain a local identity correspondence between the objects in both views.

Both I_a and I_b are then fed through an encoder-decoder network, predicting an object heatmap $\hat{H} \in [0,1]^{\frac{H}{R}\times\frac{W}{R}\times 1}$, bounding box sizes $\hat{s}^i \in \mathbb{R}^2$, offsets $\hat{o}^i \in \mathbb{R}^2$ and a global feature map $E \in \mathbb{R}^{\frac{H}{R}\times\frac{W}{R}\times C}$ for both views. We also use the ground truth bounding boxes $b_{a,b}^i = (x_1^i, y_1^i, x_2^i, y_2^i)$ to estimate a low-resolution center point $(\tilde{c}_x^i, \tilde{c}_y^i) = \left(\left\lfloor \frac{x_1^i + x_2^i}{2R} \right\rfloor, \left\lfloor \frac{y_1^i + y_2^i}{2R} \right\rfloor\right)$ for each object i. Lastly, we use these to extract an embedding $E_{x,y}^i \in \mathbb{R}^C$ from each object's center point in feature maps E_a and E_b. This results in a maximum of $2N$ embeddings, where N is the number of objects in the original image.

4.2 Supervised Learning for Detection

The object detection branch of CL-MOT is trained in a supervised manner, using the ground truth bounding boxes b_a^i for image I_a. Specifically, we apply a focal loss to the estimated object heatmap $\hat{H}_a$, as well as an L1 regression loss to the size $\hat{s}_a^i$ and offset $\hat{o}_a^i$ predictions. Note that these supervised losses are only computed against the predictions for image I_a.

Heatmap Loss Given the ground truth $H \in [0,1]^{\frac{H}{R}\times\frac{W}{R}\times 1}$ and predicted $\hat{H} \in [0,1]^{\frac{H}{R}\times\frac{W}{R}\times 1}$

object heatmaps, we enforce a pixel-wise logistic regression with focal loss [22] as follows:

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

where N is the number of objects, and $\alpha = 2$ and $\beta = 4$ are our default hyper-parameters for the focal loss. The ground truth heatmap H is generated by rendering a Gaussian peak at each object's down-scaled center point $(\tilde{c}^i_x, \tilde{c}^i_y) = (\left\lfloor \frac{x^i_1+x^i_2}{2R} \right\rfloor, \left\lfloor \frac{y^i_1+y^i_2}{2R} \right\rfloor)$. Formally, each response value $H_{x,y}$ within H is computed using Equation 2, where σ_i is a function of object size [19].

$$H_{x,y} = \max_i \left[\exp\left(-\frac{(x-\tilde{c}^i_x)^2 + (y-\tilde{c}^i_y)^2}{2\sigma^2_i} \right) \right] \tag{2}$$

Size and Offset Loss Let $\hat{s}^i$ and $\hat{o}^i$ represent the network's size and offset predictions, respectively. Using the set of bounding boxes $b^i_a = (x^i_1, y^i_1, x^i_2, y^i_2)$, we calculate the ground truth size $s^i = (x^i_2 - x^i_1, y^i_2 - y^i_1)$ and offset $o^i = \left(\frac{x^i_1+x^i_2}{2R}, \frac{y^i_1+y^i_2}{2R}\right) - \left(\left\lfloor \frac{x^i_1+x^i_2}{2R} \right\rfloor, \left\lfloor \frac{y^i_1+y^i_2}{2R} \right\rfloor\right)$ for each object i in image I_a. Our network's size and offset branches are then supervised by the L1 regression loss shown below:

$$L_{bbox} = \sum_{i=1}^{N} \left| s^i - \hat{s}^i \right| + \left| o^i - \hat{o}^i \right| \tag{3}$$

4.3 Self-Supervised Embedding Learning

Recall that the CL-MOT training pipeline produces at most N embeddings for each augmented image I_a and I_b, where N is the number of objects in the original scene. We denote these contrastive sets of embeddings as E^i_a and E^j_b, letting $i \in \mathbb{R}^N$ and $j \in \mathbb{R}^N$ represent an object's local identity. Our training objective, then, is to maximize similarity between embeddings of the same identity, while minimizing agreement between those of different identities. We accomplish this by leveraging a triplet loss [29] with hard mining [15], sampling embedding triplets from the bipartite set $B = E^i_a \cup E^j_b$.

Given that for each embedding $b \in B$, there exists only a single positive pair (E^i_a, E^j_b) where $i = j$, the task of mining hard triplets in CL-MOT reduces to hard negative mining. This is equivalent to searching for objects which look most similar to one another, despite having distinct identities, and only including these pairs in the loss computation. Note that we use Euclidean distance $\mathrm{d}(v,w)$ for measuring the distance in feature space between embeddings:

$$\mathrm{d}(v,w) = \|v - w\| \tag{4}$$

Letting b_p represent the positive pair to each embedding b_i and $m = 0.3$ a default margin term, our embedding loss can then be formulated as:

$$L_{embedding} = \frac{1}{|B|} \sum_{b_i \in B} \max(\|b_i - b_p\| - \min_{\substack{b_k \in B \\ id_i \neq id_k}} \|b_i - b_k\| + m, 0) \tag{5}$$

This loss ensures that the distance in feature space between each object instance and the object most similar to it is greater than that between matching, albeit augmented, instances by

a margin m. As a result, CL-MOT learns to distinguish between the most similar-looking objects in a frame while enforcing consistency across appearance change. Our ablation studies in section 5.3 show that this mining strategy improves discriminative ability when compared to losses that compute over all positive and negative pairs, such as batch-all mining [15] or an N-pair loss [6, 30]. Moreover, our results in section 5 suggest that CL-MOT's pretext learning objective produces more robust and general features than supervised identity classification methods [18, 38].

4.4 Inference

Once trained, CL-MOT runs online and in real time by leveraging an appearance-based association algorithm. Introduced by Wang *et al.* [32], this method describes a tracklet, or object instance, with its appearance embedding E_i and motion state $M_i = (x, y, h, w, \vec{x}, \vec{y}, \vec{h}, \vec{w})$, comprised of the object's center position, height, width and rate of change for each of these. Each tracklet's appearance embedding is initialized with the embedding from its first observation. Furthermore, this method maintains a memory bank of all previous tracklets that current predictions are likely to be associated with.

Then, for each incoming frame, the pair-wise motion affinity matrix A_m and appearance affinity matrix A_e are computed between the current predictions and those in memory. We utilize Mahalanobis distance and cosine similarity for motion and appearance affinity, respectively. To finalize the association step, bipartite matching is performed on the combined cost matrix $C = \lambda A_e + (1-\lambda)A_m$ by the Hungarian algorithm. Lastly, the motion state M_i of all matched tracklets are updated by a Kalman filter and the embeddings E_i are updated by Equation 6 to reflect appearance change over time. Let f_i represent the currently predicted embedding and α a momentum term.

$$E_i = \alpha E_i + (1-\alpha) f_i^t \tag{6}$$

5 Experiments

5.1 Datasets and Metrics

Following previously successful works [32, 38], we form a single large training set from a number of smaller pedestrian detection, tracking and search datasets. In particular, we include the MOT17 [25], Citypersons, CalTech, ETH, PRW and CUHK-SYSU datasets. While the CalTech, MOT17, CUHK-SYSU and PRW datasets do provide object identity annotations, we do not use these during training. Further statistics regarding the full training dataset can be found in Table 1.

The performance of CL-MOT is evaluated on datasets from the MOT Challenge benchmark. Specifically, our main results are gathered from the MOT17 [25] online test server, while our ablation experiments are reported on the training split of MOT15 [20]. We adopt the CLEAR [2] tracking metrics, MOTA and MOTP, as well as IDF1 score [28], for our experimentation and comparison against the state-of-the-art. Multi-object tracking accuracy (MOTA) is defined as: $MOTA = 1 - \frac{\sum_t (GT_t + FN_t + IDSW_t)}{\sum_t GT_t}$, letting $GT_t, FP_t, FN_t, IDSW_t$ be the number of ground truth bounding boxes, false positives, false negatives, and identity switches at time t, respectively. Multi-object tracking precision: $MOTP = \frac{\sum_{i,t} D_t^i}{\sum_i C_t}$, describes the misalignment between ground truth and predicted bounding boxes, where D_t^i is

a trajectories total misalignment and C_t is the total number of matches made. Lastly, IDF1 score is computed as the harmonic mean of a tracker's re-identification precision and recall: $IDF1 = \frac{2IDTP}{2IDTP+IDFP+IDFN}$, where $IDTP$, $IDFP$, and $IDFN$ are the total number of true positive, false positive, and false negative object identifications. We also report the number of mostly tracked (MT) trajectories, those that are tracked for at least 80% of their life span, and mostly lost (ML), those that are covered for at most 20%.

Dataset	MOT17	CalTech	Citypersons	CUHK-SYSU	PRW	ETH	Total
Images	5.3k	26.7	2.5k	11.2k	5.7k	2.1k	53.5k
Bounding Boxes	112k	46k	21k	55k	18k	17k	269k

Table 1: Breakdown of our full training set.

5.2 Implementation Details

Our encoder-decoder is a fully-convolutional DLA-34 variant [37, 39], with additional skip connections and deformable convolutions replacing traditional convolutional up-sampling layers. This model is trained using the Adam optimizer for 40 epochs, an initial learning rate of 1e-4 and a batch size of 8. We also schedule the learning rate to decay by a factor of 10 after epochs 20 and 30. In regards to data augmentation, we use a default input resolution of 1088×608, as well as random rotation, scaling, translation and color jittering for both I_a and I_b.

5.3 Ablation Study

Contrastive Loss Functions As mentioned previously, CL-MOT's default embedding loss is a triplet loss with hard negative mining (triplet-hard). However, many other loss functions have shown success in contrastive learning or re-identification tasks. We compare the triplet-hard loss against other commonly used contrastive loss functions, including normalized temperature-scaled cross entropy loss (NT-Xent) [6] and triplet loss [29] with no mining strategy (triplet-all). Table 2 contains the formula for each loss function being compared. Note that the NT-Xent and triplet-all losses are computed across all possible positive and negative pairs, whereas the triplet-hard loss includes only the hardest triplet for each anchor.

Name	Loss function
NT-Xent	$\frac{1}{\lvert B\rvert}\sum_{b_i\in B} -\log \frac{\exp(\text{sim}(b_i,b_p)/\tau)}{\sum_{\substack{b_n\in B \\ id_i\neq id_n}} \exp(\text{sim}(b_i,b_n)/\tau)}$
Triplet-all	$\frac{1}{\lvert B\rvert}\sum_{\substack{b_i,b_n\in B \\ id_i\neq id_n}} \max(\lVert b_i-b_p\rVert - \lVert b_i-b_n\rVert + m, 0)$
Triplet-hard	$\frac{1}{\lvert B\rvert}\sum_{b_i\in B} \max(\lVert b_i-b_p\rVert - \min_{\substack{b_n\in B \\ id_i\neq id_n}} \lVert b_i-b_n\rVert + m, 0)$

Table 2: Contrastive loss function definitions. B represents the set of all embeddings from a training image, b_p is the positive sample to b_i and b_n is a negative sample. All embeddings are ℓ_2 normalized.

Table 3 compares the chosen contrastive loss functions on the MOT15 [20] training set. We observe that the triplet-hard loss produces the strongest performance, with the NT-Xent loss falling behind both triplet loss variations. These results highlight the advantage of the triplet loss, which is that it does not force positive embedding clusters to collapse to a single point, but to simply be closer to one another than to any negative embedding. Moreover, by mining the hardest negative for each embedding in a scene, the triplet-hard loss focuses on object instances which are maximally difficult to distinguish between. This results in stronger re-identification ability, indicated by a 1.7 increase in IDF1 score over the triplet-all loss.

Loss function	MOTA ↑	MOTP ↑	IDF1 ↑	IDSw ↓	MT ↑	ML ↓	FP ↓	FN ↓
NT-Xent	70.3	78.8	68.9	319	232	127	**2054**	8711
Triplet-all	72.6	80.3	72.5	**269**	268	103	2790	6803
Triplet-hard	**73.8**	**80.5**	**74.2**	283	**277**	**98**	2538	**6600**

Table 3: Results obtained after training with various contrastive loss functions.

Non-linear Projection Head In [6], Chen *et al*. show that casting representations to a latent space prior to applying the contrastive loss improves performance in the context of image classification. This transformation is achieved by introducing a nonlinear projection head that maps representations to a space where the loss is computed. The intuition is that taking the contrastive loss directly on feature vectors will train representations to be invariant to data augmentation, resulting in a loss of information that could be useful in the downstream task. Their results back this hypothesis, indicating that computing the loss in the latent space does allow for more information to be learned and maintained in the representations.

Driven by these findings, we study the effects of including a similar projection head in the CL-MOT architecture. We apply this head as a small two-layer neural network $g(h)$ that outputs latent vectors: $z_i = g(b_i) = W^{(Z)}\sigma(W^{(C)}b_i)$, where σ is a ReLU non-linearity, b_i is a given embedding, C is the size of the embeddings, and Z is the chosen size of the latent vectors. Table 4 contains the results we obtain on the training set of MOT15 [20] after training CL-MOT with various embedding and projection head sizes. These experiments use our default embedding loss (a triplet loss with hard mining) and datasets. It should also be noted that during inference time we use embeddings b_i to link trajectories and that latent representations z_i are only utilized during training.

In contrast to [6], leveraging a projection head does not improve our performance. We observe that CL-MOT's re-identification capabilities are substantially worse when doing so, indicated by lower IDF1 scores. We posit that computing our loss directly on embeddings trains them to be consistent despite the appearance variation of an object. This enables CL-MOT to more accurately re-identify objects across frames, overcoming the lighting, scale, position and orientation changes that objects naturally undergo over time. Ultimately, these findings suggest that appearance-based tracking benefits from general and invariant representations, whereas image classification improves with more informative features.

5.4 Main Results

Table 5 compares CL-MOT to many of the state-of-the-art supervised and unsupervised approaches on the MOT17 test set [25], under the private detection protocol. Methods which do not require object identity labels, such as ours, are marked with an *X* under the *Unsup* column. To achieve the best results, we utilize the training schema described in section

MLP Layer	C	Z	MOTA ↑	MOTP ↑	IDF1 ↑	IDSw ↓	MT ↑	ML ↓
Yes	512	256	73.5	80.3	70.4	**209**	**293**	**90**
Yes	1024	512	73.2	80.2	68.9	210	283	95
None	512	N/A	**73.8**	**80.3**	**74.2**	283	277	98

Table 4: Effects of including a non-linear projection head following the embedding branch. We experiment with several embedding and projection head dimensions, represented by *C* and *Z*, respectively.

5.2 with an additional 10 epochs of fine-tuning on the MOT17 training set. We find that CL-MOT matches the performance of modern supervised trackers and significantly outperforms its self-supervised competitors. Most notably, we improve the state-of-the-art IDF1 score by 1.1. Intuitively, this metric measures a tracker's ability to maintain long consistent tracks without identity switches. Based on this, our improvement indicates that CL-MOT's self-supervised approach to learning object appearance is superior to methods which learn features by classifying objects on their global identity [18, 38]. Moreover, recent work [18, 23, 28] has shown that this measure represents a tracker's performance better than the popular MOTA metric, which is strongly influenced by detection quality.

Overall, then, our results are quite meaningful. Firstly, they show that CL-MOT can effectively sidestep the object identity labels that are typically demanded for tracking. Furthermore, we demonstrate the efficacy of our within-scene contrastive learning method by surpassing the re-identification capabilities of today's trackers.

Method	Unsup	MOTA ↑	MOTP ↑	IDF1 ↑	IDSw ↓	MT ↑	ML ↓
CenterTrack [40]		67.3	78.4	59.9	2898	822	584
FairMOT [38]		**67.5**	80.3	69.8	**2868**	**890**	489
Unsup-MOT [18]	X	61.7	78.3	58.1	1864	640	762
Ours	X	66.0	**80.5**	**70.9**	3225	870	**471**

Table 5: Comparison against the state-of-the-art trackers on the MOT17 [25] benchmark, under the private detection protocol. The column Unsup indicates that a method does not require identity labels.

6 Conclusion

In this work, we present CL-MOT, a novel semi-supervised learning framework for multi-object tracking. The proposed approach enables us to forgo object identify labels by instead leveraging contrastive learning to enforce feature consistency within object tracks. Our experimental results on tracking benchmarks demonstrate that CL-MOT performs comparatively to supervised trackers, while requiring a fraction of the annotation labor. Moreover, we find that our pre-text learning paradigm produces more robust and functional embeddings than other appearance-based approaches, setting a new state-of-the-art in several re-ID metrics. While many of the individual components of CL-MOT have appeared in previous works, we feel that our learning framework provides a valuable contribution by enabling accurate tracking without relying on costly object identity annotations.

References

[1] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles, 2019.

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing. pp. 1-10*, 2008.

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP 2016)*, pages 3464–3468. IEEE, Institute of Electrical and Electronics Engineers, aug 2016. ISBN 9781467399623. doi: 10.1109/ICIP.2016.7533003. URL http://2016.ieeeicip.org/.

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2020.

[5] Long Chen, Haizhou Ai, Zijie Zhuang, and Chong Shang. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. *CoRR*, abs/1809.04427, 2018. URL http://arxiv.org/abs/1809.04427.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. *CoRR*, abs/1703.06211, 2017. URL http://arxiv.org/abs/1703.06211.

[8] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé. MOT20: A benchmark for multi object tracking in crowded scenes. *arXiv:2003.09003*, March 2020. URL http://arxiv.org/abs/1906.04567.

[9] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. doi: 10.1109/iccv.2017.226. URL http://dx.doi.org/10.1109/ICCV.2017.226.

[10] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. pages 304–311, 2009.

[11] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin A. Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *CoRR*, abs/1406.6909, 2014. URL http://arxiv.org/abs/1406.6909.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2017.

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2019.

[15] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017. URL http://arxiv.org/abs/1703.07737.

[16] Hou-Ning Hu, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krähenbühl, Trevor Darrell, and Fisher Yu. Joint monocular 3d vehicle detection and tracking, 2018.

[17] Xu Ji, Andrea Vedaldi, and Joao Henriques. Invariant information clustering for unsupervised image classification and segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. doi: 10.1109/iccv.2019.00996. URL http://dx.doi.org/10.1109/ICCV.2019.00996.

[18] Shyamgopal Karthik, Ameya Prabhu, and Vineet Gandhi. Simple unsupervised multi-object tracking. *arXiv preprint arXiv:2006.02609*, 2020.

[19] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. *CoRR*, abs/1808.01244, 2018. URL http://arxiv.org/abs/1808.01244.

[20] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942*, April 2015. URL http://arxiv.org/abs/1504.01942.

[21] Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese CNN for robust target association. *CoRR*, abs/1604.07866, 2016. URL http://arxiv.org/abs/1604.07866.

[22] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. URL http://arxiv.org/abs/1708.02002.

[23] Andrii Maksai and Pascal Fua. Eliminating exposure bias and metric mismatch in multiple object tracking. June 2019.

[24] Santiago Manen, Michael Gygli, Dengxin Dai, and Luc Van Gool. Pathtrack: Fast trajectory annotation with path supervision, 2017.

[25] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831*, March 2016. URL http://arxiv.org/abs/1603.00831.

[26] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2015.

[28] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. *European Conference on Computer Vision. pp. 17–35*, 2016.

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015. URL http://arxiv.org/abs/1503.03832.

[30] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1857–1865. Curran Associates, Inc., 2016.

[31] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: multi-object tracking and segmentation. *CoRR*, abs/1902.03604, 2019. URL http://arxiv.org/abs/1902.03604.

[32] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. *arXiv preprint arXiv:1909.12605*, 2019.

[33] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. *CoRR*, abs/1703.07402, 2017. URL http://arxiv.org/abs/1703.07402.

[34] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. *CoRR*, abs/1805.01978, 2018. URL http://arxiv.org/abs/1805.01978.

[35] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. *CoRR*, abs/1804.06208, 2018. URL http://arxiv.org/abs/1804.06208.

[36] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019. doi: 10.1109/cvpr.2019.00637. URL http://dx.doi.org/10.1109/CVPR.2019.00637.

[37] Fisher Yu, Dequan Wang, and Trevor Darrell. Deep layer aggregation. *CoRR*, abs/1707.06484, 2017. URL http://arxiv.org/abs/1707.06484.

[38] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. A simple baseline for multi-object tracking. *arXiv preprint arXiv:2004.01888*, 2020.

[39] X. Zhou, D. Wang, Vladlen, and P. Krähenbühl. Objects as points. *arXiv:1904.07850*, 2019.

[40] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. *arXiv:2004.01177*, 2020.