

---

# Sports Re-ID: Improving Re-Identification Of Players In Broadcast Videos Of Team Sports

---

**Bharath Comandur**

Apple Inc.

cjrbharath@gmail.com

## Abstract

Re-identification (re-id) of people in images is a well-studied problem in computer vision for many applications. This work focuses on player re-identification in broadcast videos of team sports. Specifically, we focus on identifying the same player in images captured from different camera viewpoints during any given moment of a match. This task differs from traditional applications of person re-id in a few important ways. Firstly, players from the same team wear highly similar clothes, such as a team jersey/uniform, thereby making it harder to tell them apart. Secondly, there are only a few number of samples for each identity, which makes it harder to train a re-id system. Thirdly, the resolutions of the images are often quite low and vary a lot. This combined with heavy occlusions and fast movements of players greatly increase the challenges for re-id. In this paper, we propose a simple but effective hierarchical data sampling procedure and a centroid loss function that, when used together, increase the mean average precision (mAP) by 7 - 11.5 and the rank-1 (R1) by 8.8 - 14.9 without any change in the network or hyper-parameters used. Our data sampling procedure improves the similarity of the training and test distributions, and thereby aids in creating better estimates of the centroids of the embeddings (or feature vectors). Surprisingly, our study shows that in the presence of severely limited data, as is the case for our application, a simple centroid loss function based on euclidean distances significantly outperforms the popular triplet-centroid loss function. Our proposals provide comparable improvements for both convolutional networks and vision transformers. Our approach is currently ranked #1 on the ongoing SoccerNet Re-Identification Challenge 2022 leaderboard (test-split) with a mAP of 86.0 and a R1 of 81.5. On the sequestered challenge split, we achieve an mAP of 84.9 and a R1 of 80.1. While we demonstrate results on soccer matches, our proposals naturally extend to any team sport. Research on re-id for sports-related applications is very limited and our work presents one of the first discussions in the literature on this. Code will be available in the supplementary material.

## 1 Introduction

Person re-identification (re-id) in images and videos is quite useful for many applications that require tracking people across multiple cameras. Examples of such applications include surveillance in airports and public spaces, tracking customers in automated grocery stores such as Amazon Go, etc. While there exists a lot of prior art on person re-id for surveillance, in this paper, we focus on an interesting application of re-id that is under-represented in the literature. Our application involves re-identifying players in broadcast videos of team sports such as soccer, basketball, etc. Specifically, we want to develop a system that can quickly re-identify players across different camera viewpoints at any given moment in a match. Such a system can be used for many applications including tracking players across multiple cameras, building automatic highlight videos that focus on a single player,



Figure 1: Some examples of the power of our approach in challenging scenarios. The results in the column on the left are obtained by training an OSNet [33] using random data sampling and triplet loss [10]. The results in the column on the right are obtained by training the same network with our proposed hierarchical sampling and centroid loss. The query image is highlighted with a **red** border and the true matches in the corresponding gallery images are marked with a **blue** border. The gallery images are ranked in increasing order of their predicted distances from the query. Comparing Figs. 1a and 1b, we see that our procedure helps the same network learn powerful features that can correctly re-identify a player even if the gallery image contains only a small part of the player in the query image. Comparing Fig. 1c with Fig. 1d shows that our approach correctly identifies a player from different viewpoints even in the presence of occlusions.

developing better tools to assist referees, etc. In fact, this is the main objective of the first edition of the SoccerNet Re-Identification Challenge 2022 [1; 4].

This re-id task differs from traditional re-id applications, such as for surveillance, in the following important ways – 1) players that belong to the same team often look very similar because they wear similar uniforms or team jerseys; 2) the resolutions of the images are often quite low and vary a lot. This combined with significant occlusions, and fast movements of players greatly increase the challenges for re-id. Figs. 1 and 2 illustrate these issues; 3) there are only a few number of samples for each identity, which makes it harder to train machine-learning models.

Through extensive evaluation in Section 6, our study shows that due to these differences, the performance of re-id systems that have been developed for surveillance applications, degrades when applied to the task of player re-id without any modifications.

Our contributions are summarized as follows:

1. We present one of the first research studies that is specifically focused on re-identifying players in broadcast videos of team sports.
2. Via extensive evaluation, we show that conventional data sampling and batching strategies used for surveillance re-id are not suitable for our application, as they create dissimilar data distributions during training and testing.
3. Motivated by the above observation, we propose a simple but effective hierarchical data sampling and batching procedure to reduce the gap between the training and test data distributions.
4. Many SOTA approaches for re-id such as [24; 5; 27; 29; 20; 12; 2] use a triplet-centroid loss to improve upon the popular triplet loss [10]. However, our study shows that in the presence of severely limited data, as is the case for our application, a triplet-centroid loss hardly

offers any improvement over a triplet loss. Surprisingly, we show that a simple centroid loss function based on euclidean distances significantly outperforms triplet-centroid loss functions for our application.

5. Our proposals yield an increase in the range of 7 - 11.5 for the mean average precision (mAP) and an increase in the range of 8.8 - 14.9 for the rank-1 (R1), without changing the network or hyper-parameters. We demonstrate comparable increases in mAP and R1 for both convolutional neural networks and vision transformers [6; 19].

## 2 Related work

An exhaustive survey of prior art in re-id is beyond the scope of this work, since person re-id is an extremely popular research area. The studies in [25; 21; 18] provide an in-depth survey of existing re-id techniques. On the one hand, the algorithms in [27; 5; 27; 29; 20; 12; 24; 2] represent a whole category of approaches that focus on developing suitable loss functions to improve re-id. On the other hand, the work in [9; 33; 34] represent a category of approaches that concentrate on developing suitable networks to extract multi-scale, multi-view invariant features. The authors of [9] apply transformers to re-id while those of [33; 34] develop ‘omni-scale’ features for finegrained re-id.

Other techniques including the ones in [3; 16; 26; 28; 22; 31] propose hierarchical clustering and grouping strategies for re-id. However, they use the term ‘hierarchical’ in the context of feature-based grouping or clustering, and subsequently learning a mapping function to assign each person to a position within the group structure. This is different from our context, where we use a simple, effective and deterministic rule-based approach for hierarchically sampling data. Our approach is suitable for many team sports.

Some popular public datasets for re-id include CUHK03[13], Market-1501 [30], MSMT17 [23], Street2Shop[7], etc. However, almost all of them are for surveillance or fashion related applications. To the best of our knowledge, the most relevant public sports-focused re-id contribution that we found is the work in [17] that creates a dataset for re-id in long-distance running. Research on re-id in sports is very limited. Infact, this year is the first edition of the SoccerNet [1; 4] Re-ID Challenge. There is a lot of avenues for new research in re-id for sports applications.

## 3 Some basic concepts and definitions for person re-identification

Consider an image that contains a person of interest. We refer to this image as the query image. We want to identify this person in another collection of images that we refer to as the gallery images. Broadly speaking, many SOTA approaches [24; 9; 34] that employ neural networks for re-id, use some variant of the following procedure.

Each unique person in the training data is assigned a unique id. An object detector is used to detect bounding boxes of each person in an image. The image is then cropped using these bounding boxes to yield a set of detections per image. The class label for each such detection is the id of the person that it contains. Let  $Q$  denote a detection produced from a query image that contains the person of interest. We refer to this as the query detection. Let  $G = \cup G_i$  denote the set of all detections obtained from the gallery images. We refer to  $G$  as gallery detections.  $G_i$  is the  $i^{th}$  gallery detection. These detections are then fed to a classifier such as a neural network. This network is trained using a combination of loss functions such as a classification (cross-entropy) loss and a triplet loss [10]. At inference, the final layer of the network that outputs the class scores is discarded. For each input detection, the output of the penultimate layer of the network is used as its corresponding feature vector or embedding. Let  $F_Q$  denote the embedding of  $Q$ , and  $F_{G_i}$  denote the embedding of  $G_i$ .  $F_G = \cup F_{G_i}$ . The distance between  $F_Q$  and every element  $F_{G_i} \in F_G$  is computed. This distance can be an euclidean distance or cosine distance, etc. For some applications, the  $G_i$  that has the smallest distance to  $Q$  is selected as the match. For other applications, the  $G_i$ ’s are ranked in increasing order of their distances from  $Q$ .

### 3.1 Random sampling, triplet loss, and BATCH HARD

The novel study in [10] shows that the procedure used to create training batches is key when using a triplet loss for re-id. They propose the following approach that they refer to as BATCH HARD.

For each batch, we collect  $K$  samples for a specific id. We do this for  $M$  randomly sampled unique ids, resulting in a batch size of  $K \cdot M$ . We apply the network on this batch to predict the class label and the embedding for each sample. We then compute all pairwise distances between the embeddings in this batch. Let  $D_{I,J}$  denote the distance between two detections I and J. Since the  $M$  ids in this batch are randomly sampled, we refer to this procedure as random sampling. We treat each sample in this batch as a query detection, and treat the other samples in the batch as the corresponding gallery detections. Thus for each sample  $A$ , we find the hardest positive  $P$  within the batch, i.e., the sample that is farthest from  $A$  while having the same id as  $A$ . We also find the hardest negative  $N$  within the batch, i.e., the sample that is nearest to  $A$  but has a different id than  $A$ . Nearest and farthest are calculated using the pairwise distances.

The triplet loss for this sample is then defined as

$$L_T(A) = [m + D_{A,P} - D_{A,N}]_+ \quad (1)$$

where  $[W]_+$  denotes  $\max(0, W)$ .  $P$  and  $A$  have the same id.  $N$  and  $A$  have different ids.

The triplet loss for a batch  $B$  is simply the sum over the samples.

$$L_T = \sum_{A, A \in B} [m + D_{A,P} - D_{A,N}]_+ \quad (2)$$

### 3.2 Player re-id in broadcast videos of team sports

Consider a broadcast video of a basketball game or soccer match. We refer to each moment of interest in the match as an "action". We call the first frame in which we see an action in the broadcast video as the "action frame" for that action. A broadcast video often contains replays captured from other camera viewpoints. This means that the same action can appear in the broadcast video at later frames. We call the other frames that depict the same action, and that appear later in the broadcast video, as "replay frames". Thus while an action frame and its corresponding replay frames appear at different times in the broadcast video, they represent roughly the same timestamp with respect to the match itself. For example, an action occurring at the 5 minute mark in a match can have the action frame at frame number 100 and a replay frame at frame number 1000.

Our objective is to re-id a player in the action frame, across the different replay frames for that action. This is highly useful for tracking players across multiple cameras, creating highlight videos, virtual assistance for referees, etc. Note that person detection is beyond the scope of this paper. We assume that we already have the person detections for the action and replay frames. Note that due to the nature of team sports, a detection can contain multiple players with occlusions as shown in Figs. 1 and 2. This makes re-id challenging. In our context, each detection from a given action frame is a query and the detections from the corresponding replay frames are the gallery detections for that query. Our re-id system ranks these replay frames based on their distance to the query.

## 4 Our proposals to improve player re-id in broadcast videos

### 4.1 Motivation

At inference, the people in the gallery and query detections in one batch will belong to one of two teams, or in some cases be a referee. Now consider a single batch during training time, that is created using the BATCH HARD procedure [10] described in Section 3.1. This batch consists of  $K$  samples for each of  $M$  different ids. Hence the batch size is  $K \cdot M$ . We now make the following hypothesis. If the  $M$  different ids for this single batch are *randomly sampled* from the entire dataset consisting of samples from many matches (which is the conventional approach), then the distribution of the pairwise differences between the samples of this batch will be different from the distribution of the pairwise differences between the samples of the batch that the network sees during inference. For instance, with random sampling of the ids, each training batch contains images of players from multiple teams, and the network potentially learns to extract image features that exploit the differences in the clothing. However, such features won't have sufficient discriminative power if we are comparing a player against other players from the same team, *which is precisely the case at inference time*.

Table 1: Hierarchical grouping for data sampling and batching

Level	Description
O	A random sample with its action, match, year and two participating teams.
I	All samples with the same action. See Section 3.2 for the definition of an "action".
II	All samples from the same match
III	All samples from any match between the same two teams in the same year
IV	All samples from any match between the same two teams in any year
V	All samples from any match that involves at least one of the two teams in the same year
VI	All samples from any match that involves at least one of the two teams in any year
VII	All samples

## 4.2 Hierarchical data sampling

Based on the hypothesis described above, we propose an alternate hierarchical data sampling procedure that we denote as hierarchical sampling. To create this hierarchy, we make use of the metadata that comes with the images. For each sample at training, we know its "action" (Section 3.2), which match it belongs to, the year in which the match was played, and also the names of the two teams participating in that match. Note that we do not know which of the two teams the player belongs to. Using this information, we hierarchically group the data into multiple levels as shown in Table 1:

For a single training batch, we randomly select a sample (Level O in Table 1). We then proceed to create a training batch by first sampling from Level I in Table 1. If we do not have sufficient samples for a batch, we proceed to Level II, while excluding samples that have already been added to the batch. In this fashion, we proceed from level to level until we have enough samples for a batch. All the samples selected in this batch are then discarded from the pool of available samples (for this epoch). For the next epoch, we re-initialize the pool of available samples, and again apply the same procedure. In this way, we ensure that the batches are randomized. So a single sample is compared against different samples in different epochs. This hierarchical procedure increases the possibility of having similar samples in each batch, which is the case at inference.

One potential drawback of our proposed sampling procedure is that it does reduce the total number of possible batches that can be created for training, when compared to pure random sampling. In addition, as mentioned before, there are only a few number of samples for each identity, which makes it harder to train a network for our application. We address this via the use of a centroid loss.

## 4.3 Centroid loss vs triplet-centroid loss

### 4.3.1 Triple-centroid loss

The triplet loss in equation 2 operates on a per-sample level, meaning that we compare a sample embedding  $A$  with two other embeddings,  $P$  and  $N$ , where  $P$  is the hardest positive for  $A$  within a batch and  $N$  is the hardest negative for  $A$  within a batch. Hence it can be sensitive to outliers. To handle this, triplet loss functions using centroids have been proposed in prior work such as [5; 27; 29; 20; 12; 24; 2]. However, it is important to note that these studies employ a *triplet (or n-tuplet) loss that uses centroids*. For example, the study in [24], that is currently among the SOTA approaches on many public re-id benchmarks, defines a triplet-centroid loss for a sample  $A$  as follows:

$$L_{TC}(A) = [m + D_{A,\text{Centroid}_P} - D_{A,\text{Centroid}_N}]_+ \quad (3)$$

where  $[W]_+$  denotes  $\max(0, W)$ .  $\text{Centroid}_P$  denotes the centroid of the cluster with the same id as  $A$  and  $\text{Centroid}_N$  is the centroid of the cluster with ids different from the id of  $A$ .

However, in our experiments, we observe that adding this triplet-centroid loss  $L_{TC}$  does not improve the mAP or R1 for our application. We show this in Table 3.

We hypothesize that this is because our re-id application has much less data than the studies in [5; 27; 29; 20; 12; 24; 2], and because of the differences in the data distributions. Even though it uses centroids, the triplet-centroid loss is still calculated on a per-sample level with one triplet for each sample  $A$ . When there is sufficient amount of labelled data, this is not an issue since gradients

are repeatedly calculated over multiple batches, thereby reducing any negative effect of individual outliers, and the network learns to average information over the samples.

For our limited data case, we obtain much better results by calculating a simple L2 loss using the centroids, instead of a triplet-centroid loss. Specifically, for each unique id  $M_i$  in the batch, we partition the samples into two clusters. Cluster-I contains only the samples with the same id  $M_i$ , and cluster-II contains the remaining samples. We then calculate the embedding of the centroid of cluster-I and cluster-II. The loss function is simply the euclidean distance between these two centroids. We sum this loss for each unique id in the batch. This teaches the network to create embeddings that don't just push individual samples, but rather the clusters of different ids away from each other. Formally, for a batch with  $K$  samples for each of  $M$  ids, let  $y(J)$  denote the id of the sample  $J$  and  $F_J$  the feature embedding for this sample  $J$ . Cluster-I for an id  $M_i$  will have  $K$  samples and cluster-II for this id will have  $K \cdot M - K$  samples. The centroids of cluster-I and cluster-II for this id  $M_i$  are calculated as follows:

$$C_{\text{cluster-I}}(M_i) = \frac{1}{K} \cdot \sum_{J, y(J)=M_i} (F_J) \quad (4)$$

$$C_{\text{cluster-II}}(M_i) = \frac{1}{K \cdot M - K} \cdot \sum_{J, y(J) \neq M_i} (F_J) \quad (5)$$

The centroid loss for id  $M_i$  is calculated as follows:

$$L_{\text{Centroid}}(M_i) = \|C_{\text{cluster-I}}(M_i) - C_{\text{cluster-II}}(M_i)\|^2 \quad (6)$$

The centroid loss for a batch is calculated as the sum over all ids in that batch.

$$L_{\text{Centroid}} = \sum_{M_i} L_{\text{Centroid}}(M_i) \quad (7)$$

While  $L_{\text{Centroid}}$  is conceptually much simpler than the triplet-centroid loss  $L_{TC}$ , we show in Table 3, that it does much better than  $L_{TC}$  for our application.

The final loss function that we use is given by

$$L = \alpha \cdot L_T + \beta \cdot L_C + \gamma \cdot L_{\text{Centroid}} \quad (8)$$

where  $\alpha$ ,  $\beta$  and  $\gamma$  are weights that are set empirically.  $L_T$  and  $L_C$  are the triplet and classification loss defined in Section 3.1.

## 5 Experimental evaluation

### 5.1 Datasets

We use the ongoing SoccerNet Re-Identification Challenge 2022 dataset [1; 4] to evaluate our experiments. This is the first edition of this re-id challenge. This dataset is composed of 340993 players thumbnails extracted from the SoccerNet videos at different events, and images from their replays. The data is divided into train, validation, test and challenge splits. The training data has 248234 samples in total.

The test split has 11777 query images and 34989 gallery images. However, the challenge website [1] states that "player identity labels are derived from links between bounding boxes within an action and are therefore only valid within the given action. Consequently, player identity labels do not hold across actions and a given player has a different identity for each action he has been spotted in. For that reason, during the evaluation process, only samples within the same action are matched against each other." Therefore, for each query sample, we only need to compare against the gallery samples that have the same action. For evaluating on the test split, we train our networks on the train split only. The test split leaderboard is public and hence we can compare our performance with other methods. The challenge split is composed of separate player thumbnails from different games and is sequestered, meaning that the ground-truth labels are unknown. There are 9021 query samples and 26082 gallery samples. Since the challenge is currently ongoing, the challenge split leaderboard is

not public. Hence we do not know our position on the challenge split leaderboard. Therefore, we only report the performance of our network on the challenge split. We evaluate the same networks on the test and challenge splits. Only the train split is used for training in this case as well.

## 5.2 Network

We conduct our evaluation using five different network architectures. Two of them are convolutional networks and the remaining three are vision transformers.

**Convolutional networks:** - The first is a ResNet50-fc512 network which is a ResNet50 [8] with an extra fully connected layer of 512 output channels and a batchnorm layer as the penultimate layers, followed by a classification layer. We chose this network because the SoccerNet challenge provides it as a baseline. However, note that their baseline ResNet50-fc512 produces an mAP of 57.40 and a R1 of 45.89. We observed that by just adjusting the batchsize and a few hyperparameters, the same network produces an mAP of 70.3 and a R1 of 61.2 on the test split. So, we use the latter as our baseline. This network has 24.6M trainable parameters. The second convolutional network that we use is the OSNet\_x1\_0 [33; 34]. This network has produced compelling results on prior re-id literature and is very lightweight with 2.2M trainable parameters.

**Vision transformers:** The third network is a data-efficient image transformer DeiT-Tiny/16 [19] which is relatively lightweight with only 5.5M parameters but does well on image classification tasks. The fourth is a larger variant of this transformer called DeiT-S/16 [19] with 21.7M trainable parameters. The fifth network is the popular vision transformer ViT-B/16 [6] which produces SOTA results on image classification tasks. This network is quite huge with 57.7M trainable parameters. For all three transformer networks, we replace the final MLP classification layer with a fully connected layer with 512 output channels, a batchnorm layer followed by a final layer for classification. Hence all the five networks produce an embedding of length 512, for consistency. We choose these three transformers since the study in [9] shows good results for re-id applications with them.

## 5.3 Training and hyperparameters

We train the convolutional networks with ADAM [11] optimizer and a linearly decaying learning rate scheduler. Transformers on the other hand are relatively trickier to train and therefore we use an ADAMW [15] optimizer, with a linear warmup for the learning rate for the first 3000 iterations, followed by cosine annealing [14]. We use the loss in equation 8 with  $\alpha = 0.9$ ,  $\beta = 0.5$  and  $\gamma = 0.5$  chosen empirically. We also carry out ablation studies where we disable the centroid loss, and use the triplet-centroid loss (equation 3). The SoccerNet challenge provides the Torchreid [32] library to develop our algorithms. Training is done on a single NVIDIA Ampere A100 GPU. To prevent selective/biased reporting of our improvements, each network is trained for 40 epochs with periodic checkpointing and we report the best mAP and R1 for each network. For fair comparison, the hyperparameters such as learning rate, weight decay etc. of each network are fixed across all experiments for that network.

# 6 Results

## 6.1 Evaluation on test split

Table 2 summarizes our results on the test split. We observe that for all five networks, hierarchical sampling when used with the additional centroid-loss (equation ??) increases the mAP by 7 - 11.5 and R1 by 8.8 - 14.9, when compared to random sampling and triplet loss.

## 6.2 Evaluation on challenge split

Since the challenge is currently ongoing, the sequestered challenge split leaderboard is not public. Hence we do not know our position on the challenge split leaderboard. Also, since there is a hard limit on the number of submissions we can make to the challenge leaderboard, we only evaluate our best performing ViT-B/16 network trained with hierarchical sampling and the additional centroid loss on this split. It yields a mAP of 84.9 and a R1 of 80.1 on the challenge split.

Table 2: Evaluation on the SoccerNet Re-Identification Challenge 2022 test split. Best numbers for each network are marked in bold font. The ViT-B/16 network with a mAP of 86.0 and a R1 of 81.5 is currently ranked #1 on the test split leaderboard. Classification loss weight  $\beta = 0.5$  for all cases.

Network	Sampling	Triplet loss	Centroid loss	mAP	R1
ResNet50-fc512	random	✓	✗	70.3	61.2
	random	✓	✓	75.4	68.7
	hierarchical	✓	✗	66.7	57.7
	hierarchical	✓	✓	<b>81.8</b>	<b>76.1</b>
OSNet_x1_0	random	✓	✗	76.4	69.2
	random	✓	✓	78.5	72.6
	hierarchical	✓	✗	75.8	69.1
	hierarchical	✓	✓	<b>83.4</b>	<b>78.0</b>
DeiT-Tiny/16	random	✓	✗	73.2	65.0
	random	✓	✓	74.9	67.1
	hierarchical	✓	✗	80.5	74.4
	hierarchical	✓	✓	<b>82.2</b>	<b>76.2</b>
DeiT-S/16	random	✓	✗	75.3	67.8
	random	✓	✓	78.9	72.3
	hierarchical	✓	✗	81.0	74.7
	hierarchical	✓	✓	<b>84.3</b>	<b>79.4</b>
ViT-B/16	random	✓	✗	75.7	68.2
	random	✓	✓	78.4	72.0
	hierarchical	✓	✗	81.4	75.4
	hierarchical	✓	✓	<b>86.0</b>	<b>81.5</b>

### 6.3 Ablation studies

#### 6.3.1 Centroid loss vs triplet-centroid loss

SOTA approaches such as the studies in [5; 27; 29; 20; 12; 24; 2] use triplet-centroid losses which is defined in equation 3. As mentioned in Section 6.3.1, due to the limited training data per id, and the difference in data distributions, we observe that the triplet-centroid loss hardly improve mAP and R1. This is shown in Table 3. Note that we only show a subset of results in this table due to space considerations. The full table is present in the supplementary material and our observations are similar across networks. The simple euclidean centroid loss is much more powerful for our task.

#### 6.3.2 Using hierarchical sampling without the centroid loss

Table 2 also shows the mAP and R1 when the networks are trained without the centroid loss term. The results are quite interesting. For all three transformer networks, DeiT-Tiny/16, DeiT-S/16 and ViT-B/16, we see an increase of 5.7 - 7.3 in the mAP and an increase of 6.9 - 9.4 in the R1 by just using hierarchical sampling, without the centroid loss. However, for the two convolutional networks, we see a decrease. The mAP and R1 decrease by  $\sim 3.6$  for the ResNet50-fc512 which is, broadly speaking, the weakest of the five networks in terms of image classification accuracy. For the relatively more powerful OSNet\_x1\_0, the mAP and R1 decrease slightly by 0.6 and 0.1 respectively. We are quite surprised by this clear separation in the behavior of convolutional networks and transformers. More analysis is needed before we can draw any definitive conclusion from this. But we think it is interesting enough to point out.

Table 3: Comparison of centroid loss vs triplet-centroid loss. We only show a subset of results in this table due to space considerations. The full table is present in the supplementary material and our observations are similar across networks. Best numbers for each network are marked in bold font.

Network	Sampling	Triplet loss	Centroid loss	Triplet-centroid loss	mAP	R1
DeiT-Tiny/16	hierarchical	✓	✗	✗	80.5	74.4
		✓	✗	✓	80.5	74.2
		✓	✓	✗	82.2	76.2
ViT-B/16	hierarchical	✓	✓	✓	<b>82.3</b>	<b>76.4</b>
		✓	✗	✗	81.4	75.4
		✓	✗	✓	81.7	75.9
		✓	✓	✗	<b>86.0</b>	<b>81.5</b>
		✓	✓	✓	85.8	81.1

#### 6.4 Additional results



Figure 2: Additional examples of successful re-id in challenging scenarios. The query is highlighted with a red border and the true matches in the corresponding gallery images are marked with a blue border. Fig. 2a shows that our procedure teaches the network to recognize the same number on the jersey in the query and in the shorts in the gallery. Fig. 2b shows that the network learns to distinguish the subtle differences in the poses of the legs, despite all the players wearing the same white jersey.

Figs. 1 and 2 show examples of challenging cases where our procedure successfully re-identifies the player. The supplementary material contains examples of failure cases and a discussion of the limitations of our approach.

## 7 Conclusions

Re-identifying players in broadcast videos has a number of significant differences from surveillance re-id applications. These differences merit the needs for a specialized data sampling strategy and loss function. While we focus on soccer in this work due to availability of data, the ideas discussed in this work apply to many team sports. Hierarchical sampling can be easily extended to sport-specific cases, such as grouping by leagues, grouping by country etc. An unexpected result of our study is that a simple euclidean centroid loss is much more suited for our task with limited number of samples per id, when compared to SOTA triplet-centroid losses. Additional insights are included in the supplementary material.

## 8 References

### References

- [1] Soccernet player re-identification challenge 2022. <https://www.soccer-net.org/tasks/re-identification>. Accessed: 2022-05-19.
- [2] A. Alnissany and Y. Dayoub. Modified centroid triplet loss for person re-identification. 2022.
- [3] A. Bialkowski, P. Lucey, X. Wei, and S. Sridharan. Person re-identification using group information. In *2013 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6, 2013. doi: 10.1109/DICTA.2013.6691512.
- [4] A. Deliége, A. Cioppa, S. Giancola, M. J. Seikavandi, J. V. Dueholm, K. Nasrollahi, B. Ghanem, T. B. Moeslund, and M. V. Droogenbroeck. Soccernet-v2 : A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021.
- [5] T.-T. Do, T. Tran, I. Reid, V. Kumar, T. Hoang, and G. Carneiro. A theoretically sound upper bound on the triplet loss for improving the efficiency of deep distance metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10404–10413, 2019.
- [6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [7] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the IEEE international conference on computer vision*, pages 3343–3351, 2015.
- [8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [9] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15013–15022, 2021.
- [10] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.
- [11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [12] M. Lagunes-Fortiz, D. Damen, and W. Mayol-Cuevas. Centroids triplet network and temporally-consistent embeddings for in-situ object recognition. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10796–10802. IEEE, 2020.
- [13] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.
- [14] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- [15] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [16] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1363–1372, 2016.

- [17] A. Penate-Sanchez, D. Freire-Obregon, A. Lorenzo-Melian, J. Lorenzo-Navarro, and M. Castrillon-Santana. Tgc20reid: A dataset for sport event re-identification in the wild. *Pattern Recognition Letters*, 138:355–361, 2020.
- [18] V. Radh and S. Suresh. A literature survey on person re-identification.
- [19] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.
- [20] J. Wang, K.-C. Wang, M. T. Law, F. Rudzicz, and M. Brudno. Centroid-based deep metric learning for speaker recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3652–3656. IEEE, 2019.
- [21] Z. Wang, Z. Wang, Y. Zheng, Y. Wu, W. Zeng, and S. Satoh. Beyond intra-modality: A survey of heterogeneous person re-identification. *arXiv preprint arXiv:1905.10048*, 2019.
- [22] Z. Wang, L. He, X. Tu, J. Zhao, X. Gao, S. Shen, and J. Feng. Robust video-based person re-identification by hierarchical mining. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.
- [23] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018.
- [24] M. Wieczorek, B. Rychalska, and J. Dąbrowski. On the unreasonable effectiveness of centroids in image retrieval. In *International Conference on Neural Information Processing*, pages 212–223. Springer, 2021.
- [25] D. Wu, S.-J. Zheng, X.-P. Zhang, C.-A. Yuan, F. Cheng, Y. Zhao, Y.-J. Lin, Z.-Q. Zhao, Y.-L. Jiang, and D.-S. Huang. Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing*, 337:354–371, 2019.
- [26] M. Ye, X. Lan, J. Li, and P. Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [27] Y. Yuan, W. Chen, Y. Yang, and Z. Wang. In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 354–355, 2020.
- [28] K. Zeng, M. Ning, Y. Wang, and Y. Guo. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13657–13665, 2020.
- [29] Z. Zhang, C. Lan, W. Zeng, Z. Chen, and S.-F. Chang. Beyond triplet loss: Meta prototypical n-tuple loss for person re-identification. *arXiv preprint arXiv:2006.04991*, 2020.
- [30] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.
- [31] Y. Zheng, S. Tang, G. Teng, Y. Ge, K. Liu, J. Qin, D. Qi, and D. Chen. Online pseudo label generation by hierarchical cluster dynamics for adaptive person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8371–8381, 2021.
- [32] K. Zhou and T. Xiang. Torchreid: A library for deep learning person re-identification in pytorch. *arXiv preprint arXiv:1910.10093*, 2019.
- [33] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3702–3712, 2019.
- [34] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Learning generalisable omni-scale representations for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.