SoccerNet Re-Identification Challenge

Crea Michelangelo 1993024 Gautieri Alessandro 2041850

Abstract

This report outlines the various phases of the project dedicated to Player Re-Identification (Re-ID) across different matches. First, we explain the project's goal by introducing the *SoccerNet Re-Identification Challenge*: this competition served as our starting point, providing both the necessary datasets and a baseline to benchmark our training.

In the second step, we detail the dataset and how the data was prepared. In this section, we describe the cleaning and *preprocessing* procedures applied to the images to make them suitable for processing by our models.

Next, the document describes the techniques and architectures we selected. We explain how the individual models work and, most importantly, how we combined their predictions into a final **Ensemble** strategy to achieve better results than using a single model alone.

Following this, we present the concrete results. We report the performance achieved by each individual model and the final ensemble, supported by test values and metrics calculated to measure the system's accuracy.

Finally, we analyze the conclusions drawn from the work and look ahead, proposing potential improvements and future developments to make the project even more effective.

1. Introduction

The SoccerNet Re-Identification (ReID) Challenge presents a complex computer vision problem: re-identifying soccer players across multiple camera views in broadcast video footage. This task is complicated by factors such as low resolution, motion blur, occlusion, and similar player appearances (team kits).

Figure 1 illustrates the core challenge: given a player detected in one camera view (query), the system must identify all instances of the same player in other camera views (gallery). The colored lines represent identity associations across views, with different colors indicating different players. This multi-view matching must account for drastic appearance changes due to viewpoint, scale, and occlusion.

To address these challenges and maximize performance, we adopted a multi-model ensemble strategy. Leveraging distinct architectures allows us to capture complementary features, improving the robustness of the re-identification process.

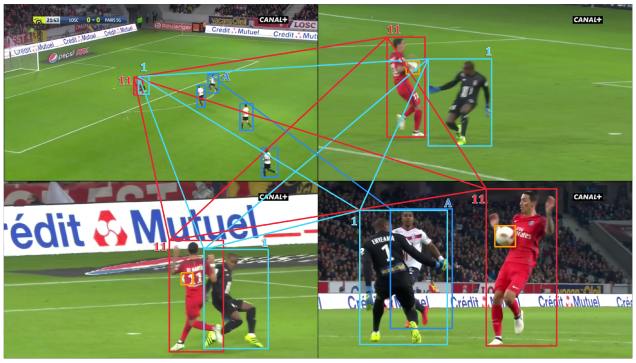

Figure 1. Multi-view person re-identification in broadcast soccer footage. Players detected in different camera views (top and bottom rows) are linked by identity. Each color represents a different player identity, with lines connecting the same person across views. The challenge involves matching players despite significant variations in pose, scale, and viewpoint.

Our approach was also shaped by significant computational constraints. We operated in a heterogeneous hardware environment due to the enormous size of the SoccerNetV3 dataset, training large-scale models on a single consumer-grade machine was infeasible within reasonable timeframes.

2. Dataset and Preprocessing

We utilized the SoccerNet-v2 dataset [1, 2], a large-scale benchmark for holistic understanding of broadcast soccer videos. For the ReID task, the dataset provides bounding box annotations and identity labels across distinct actions and camera angles.

Given the challenge's scale, effective preprocessing was crucial. We adhered to standard ReID practices, including resizing images to a fixed resolution and applying data augmentation techniques such as random erasing, flipping, and color jittering to prevent overfitting.

3. Methodology

Our methodology centers on an ensemble of three distinct deep learning models. By combining the strengths of different architectures, we aim to achieve a higher mean Average Precision (mAP) and Rank-1 accuracy than any single

model could achieve in isolation. The core idea is that different models learn different feature representations; fusing them can smooth out individual prediction errors.

The three models selection was driven by the need for diversity:

1. **DINOv2**: A Vision Transformer (ViT) based on self-supervised learning, excellent for capturing semantic context.
2. **ResNet50-IBN**: A Convolutional Neural Network (CNN) variant optimized for domain generalization.
3. **OsNet-AIN**: is a lightweight architecture specifically for Re-Identification, capable of simultaneously capturing fine details and global characteristics through multi-scale features.

3.1. Model 1: DINOv2 and Transfer Learning

For our first model, we selected DINOv2 [3], a state-of-the-art Vision Transformer trained with self-supervised learning. DINOv2 is particularly effective at generating robust visual features without requiring vast amounts of labeled data for initialization.

To adapt DINOv2 for the specific domain of soccer player re-identification, we implemented a progressive transfer learning strategy, a technique widely recognized for improving model performance in target domains with limited data [4]. This approach allowed us to incrementally adapt the model weights while managing computational resources:

1. **Stage 1 (10% Dataset)**: We initially fine-tuned the model on a small, representative 10% subset of the training data. This allowed for rapid prototyping and hyperparameter tuning.
2. **Stage 2 (30% Dataset)**: The best checkpoint from Stage 1 was then fine-tuned on a larger 30% subset. This stage significantly improved the model's generalization capabilities.
3. **Stage 3 (100% Dataset)**: Finally, utilizing the weights from Stage 2, we executed the final training run on the complete 100% dataset. This final run employed a higher input resolution (matching the 256x128 resolution of the ResNet model) to maximize fine-grained feature extraction.

This staged approach ensured stable convergence and effectively leveraged the extensive dataset, overcoming local hardware limitations during the initial phases.

3.2. Modello 2: ResNet-50

For our project, we selected the **ResNet-50** [8] model. This Convolutional Neural Network (CNN) was chosen because it offers the best balance between computational efficiency, training stability, and accuracy of results. ResNet-50 is widely recognized in the field of Computer Vision thanks to its use of **residual connections** (or *shortcut connections*). These are "shortcuts" that allow the network to skip certain layers, making the learning process smoother and preventing the model from getting stuck as it becomes deeper.

We chose this model for the Re-Identification (Re-ID) task for three main reasons:

1. **De-Facto Standard:** ResNet-50 is the reference model (*baseline*) in scientific research. Since almost all Re-ID studies start here, using this model allows us to easily compare our results with existing literature.
2. **Robust Feature Extraction:** Thanks to its deep structure, the network learns incrementally: it first recognizes simple details (such as edges and colors) and then understands complex high-level characteristics (such as posture or body structure). This is crucial for recognizing players even under varying positions or lighting conditions.
3. **Transfer Learning:** Instead of training the network from scratch, we leveraged pre-trained "weights" from the ImageNet dataset. This means the model started with a basic "visual knowledge," allowing us to achieve excellent results on the SoccerNet dataset in significantly less time.

To train the model, we used a technique called **Progressive Resizing**. The concept is similar to how humans learn: first, general shapes are understood, then the focus shifts to details. The training was divided into two phases based on image resolution:

1. **Low Resolution Phase (256×128):** In the first 60 epochs, we used smaller images. At this resolution, fine details disappear, forcing the model to focus only on the most important global features, such as body shape and dominant kit colors. This phase provided several technical advantages:
   - It allowed us to use a larger **Batch size**, making initial learning more stable.
   - The *loss function* (the model's error rate) decreased much faster.
   - The lower resolution reduced noise, helping the model avoid **overfitting** (i.e., learning useless or incorrect details).
2. **High Resolution Phase (384×192):** In the second part, we increased the image size. Since the model had already learned solid foundations in the previous phase, it only had to "refine" its knowledge by adding finer details in this step. This method proved much more effective than starting immediately with high resolution.

3.3. Modello 3: OsNet-AIN

The third model selected is **OsNet** [6], an architecture designed from scratch specifically for the Person Re-Identification task. The distinctive feature of OsNet is the use of **Omni-Scale features**. The main challenge in Re-ID lies in the heterogeneous nature of the details: distinguish-

ing between two people requires capturing both "microscopic" features (e.g., shoe logos, facial hair) and "macroscopic" features (e.g., body build, kit color).

OsNet addresses this issue by introducing a **multi-stream residual block**. At each layer, the network processes images using different **receptive fields** simultaneously, capturing details of various scales at the same time. Thanks to a mechanism called **Unified Aggregation Gate**, the model dynamically learns how to combine these local and global details, acquiring the full range of information needed to distinguish between similar players.

Specifically, we utilized the **OsNet-AIN** [7] variant. This version integrates **Instance Normalization**, a technique that makes the model significantly more robust to changes in style and lighting. In the context of SoccerNet, where footage comes from different stadiums with varying lighting and angles, this normalization helps the network ignore environmental variations (the "style") and focus exclusively on the player's identity (the "content").

3.4. Ensemble and Re-Ranking

3.4.1. Ensemble Techniques

To potentially improve upon individual model performance, we experimented with several ensemble strategies that combine predictions from multiple models. The core idea behind ensemble learning is that different architectures learn complementary feature representations; by fusing them, individual prediction errors can be smoothed out [9].

We evaluated the following techniques:

1. **Feature Concatenation**: The feature vectors extracted by each model are concatenated into a single, higher-dimensional representation. After concatenation, L2 normalization is applied to ensure all features contribute equally. This approach preserves the full information from each model but increases the dimensionality proportionally.
2. **Distance Averaging (Equal Weights)**: Instead of combining features, we compute separate distance matrices from each model and average them. This method is robust to different feature dimensions and treats all models as equally important.
3. **Weighted Distance Fusion**: An extension of distance averaging where each model's contribution is weighted differently. We performed a grid search over weight combinations to find the optimal balance, testing configurations such as $(0.33, 0.33, 0.34)$, $(0.25, 0.25, 0.50)$, $(0.20, 0.20, 0.60)$, and others.

3.4.2. Re-Ranking

Re-ranking is a post-processing technique that refines the initial ranking by exploiting the *k-reciprocal neighbors* relationship [10]. The intuition is simple: if two images are mutual nearest neighbors (i.e., each appears in the other's top-k list), they are more likely to belong to the same identity.

The algorithm works as follows:

1. Compute the initial distance matrix between query and gallery images.
2. For each query, find its k_1 nearest neighbors in the gallery.
3. Expand this set by including neighbors that share reciprocal relationships.
4. Compute a Jaccard distance based on the overlap of these neighbor sets.
5. Combine the original distance with the Jaccard distance using a weighting factor λ.

We tested several configurations:

- **Standard**: $k_1 = 20$, $k_2 = 6$, $\lambda = 0.3$
- **Aggressive**: $k_1 = 10$, $k_2 = 3$, $\lambda = 0.3$ (smaller neighborhoods)
- **Broad**: $k_1 = 60$, $k_2 = 10$, $\lambda = 0.3$ (larger neighborhoods)
- **Lambda Variant**: $k_1 = 20$, $k_2 = 6$, $\lambda = 0.5$ (more weight on Jaccard distance)

4. Final Results

All experiments were conducted on the **SoccerNet Re-ID v3 Validation Set**, which contains 11,638 query images and 34,355 gallery images. We report mean Average Precision (mAP) and Rank-1 accuracy as primary metrics.

4.1. Results DINOv2

The DINOv2 model, trained with LoRA on the full dataset (100%), achieved the following results:

- Rank-1 accuracy: **35.66%**
- mAP: **48.09%**

This suggests that Vision Transformers struggle to precisely localize discriminative features on SoccerNet's challenging domain, likely due to the need for significantly longer training compared to CNNs to fully adapt weights pre-trained on generic natural images.

4.2. Results ResNet-50

4.2.1. ResNet-50 Low Resolution Phase (256×128)

The ResNet-50 model trained with the low-resolution setting (256×128) achieved:

- Rank-1 accuracy: **33.34%**
- mAP: **46.41%**

While this model performed reasonably, it was outperformed by both OsNet-AIN and DINOv2, indicating that the standard ResNet architecture may not be optimally suited for the specific challenges of soccer player re-identification.

4.2.2. ResNet-50 High Resolution Phase (384×192)

The ResNet-50 model trained with the low-resolution setting (384×192) achieved:

- Rank-1 accuracy: **38.4%**
- mAP: **52.6%**

Despite the additional visual information, the model performed worse in both Rank-1 and mAP. Training stabilized earlier, suggesting that the model may have reached a plateau or begun slightly overfitting to the superfluous details present in the high-resolution data

4.2.3. Conclusions on Res-Net50

The low-resolution model demonstrated better generalization capabilities. This result suggests that, for the ResNet-50 architecture on this specific dataset, global structural features are more reliable predictors than fine, high-frequency details, which may instead introduce noise and make optimization more complex.

4.3. Results OsNet-AIN

The OsNet-AIN model achieved the **best individual performance**:

- Rank-1 accuracy: **43.64%**
- mAP: **56.83%**

This represents a significant improvement over the other models (+8.74% mAP over DINOv2 and +10.42% over ResNet). The multi-scale feature extraction and instance normalization proved highly effective for handling the domain variability present in broadcast soccer footage.

4.4. Results Ensemble

Surprisingly, ensemble methods did not surpass the best individual model:

- **Feature Concatenation**: mAP 53.36%, Rank-1 41.70%
- **Distance Averaging (Equal)**: mAP 53.36%, Rank-1 41.70%
- **Best Weighted Fusion** $(0.20, 0.20, 0.60)$: mAP 55.47%, Rank-1 43.50%

Even the best ensemble configuration, which heavily weights OsNet (60%), achieved lower performance than OsNet alone. This counter-intuitive result is explained by ensemble learning theory: ensembles are most effective when the constituent models exhibit **diversity** and **complementarity** [9]. When models share similar error patterns or when one model significantly outperforms the others, adding weaker models can actually dilute the ensemble's predictions.

4.5. Results Re-Ranking

Re-ranking was applied to the best ensemble configuration (weighted fusion with $w = 0.20, 0.20, 0.60$):

- **Standard** ($k_1 = 20, k_2 = 6$): mAP 54.83%, Rank-1 42.03%
- **Aggressive** ($k_1 = 10, k_2 = 3$): mAP 55.38%, Rank-1 43.28%
- **Broad** ($k_1 = 60, k_2 = 10$): mAP 52.29%, Rank-1 38.13%
- **Lambda=0.5**: mAP 55.14%, Rank-1 42.60%

None of the re-ranking configurations improved upon the base ensemble or OsNet alone. The broad configuration significantly degraded performance, suggesting that larger neighborhoods introduce too much noise in this dataset.

4.6. Performance Analysis

Figure 2 provides a visual comparison of all evaluated methods, clearly showing that OsNet-AIN outperforms all other approaches including ensembles and re-ranking strategies.

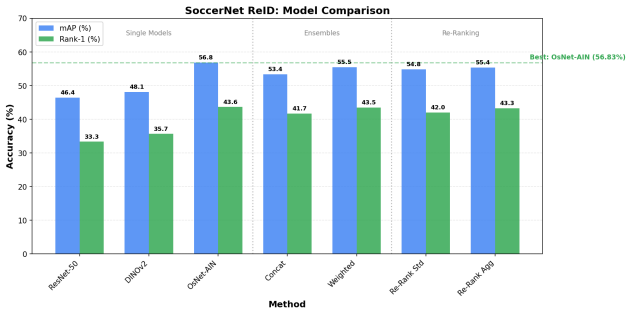

Figure 2. Comparison of mAP and Rank-1 accuracy across all evaluated methods. OsNet-AIN achieves the best performance (mAP: 56.83%, Rank-1: 43.64%), surpassing even the weighted ensemble and re-ranking approaches. The dashed line indicates OsNet-AIN's performance as the upper bound.

Figure 3 displays the Cumulative Matching Characteristic (CMC) curves for the three single models. The CMC curve shows the probability of finding a correct match within the top-k ranked gallery images. OsNet-AIN consistently outperforms the other models at all ranks, with a particularly notable advantage at Rank-1 (43.64% vs 35.66% for DINOv2).

Key observations from the performance analysis:

- **Single Model Superiority**: OsNet-AIN alone achieves better results than any ensemble or re-ranking strategy, demonstrating the importance of architectural design for this domain.
- **Gap at Higher Ranks**: The performance gap narrows at higher ranks (Rank-10, Rank-20), indicating that all models capture relevant features but differ in their ability to rank the correct match first.
- **Ensemble Degradation**: Despite conventional wisdom, ensemble methods underperform when one model is significantly stronger than others.

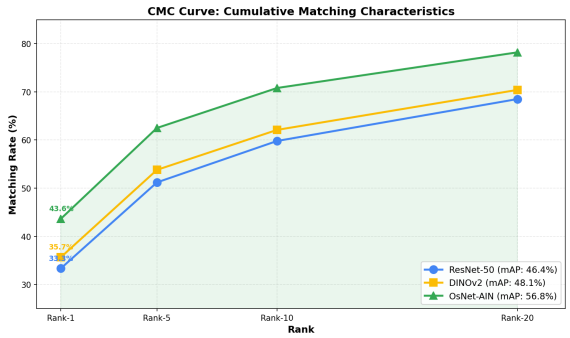

Figure 3. CMC curves for the three single models. OsNet-AIN (green) shows superior performance across all ranks, achieving 78.2% accuracy at Rank-20. The shaded area under OsNet-AIN highlights its consistent advantage over ResNet-50 and DINOv2.

4.7. Summary of Results

Table 1 presents a comprehensive comparison of all evaluated methods.

Table 1. Complete Results Comparison on SoccerNet ReID v3 Validation Set

Method	mAP	Rank-1
Single Models		
ResNet-50	46.41%	33.34%
DINOv2 (LoRA)	48.09%	35.66%
OsNet-AIN	**56.83%**	**43.64%**
Ensemble Methods		
Feature Concatenation	53.36%	41.70%
Distance Avg (Equal)	53.36%	41.70%
Weighted (0.2, 0.2, 0.6)	55.47%	43.50%
Re-Ranking (on Best Ensemble)		
Re-Rank Standard	54.83%	42.03%
Re-Rank Aggressive	55.38%	43.28%
Re-Rank Broad	52.29%	38.13%
Re-Rank $\lambda = 0.5$	55.14%	42.60%

4.8. Qualitative Results

Figure 4.8 shows representative examples of re-identification results obtained with OsNet-AIN. Each row displays a query image (leftmost, blue border) followed by the top-10 retrieved gallery images. Green borders indicate correct matches (same person identity), while red borders indicate incorrect matches.

The visualizations highlight key challenges in broadcast soccer re-identification:

- **Uniform similarity**: Players from different teams may wear similar colors, leading to false positives
- **Pose variation**: The same player appears in vastly different poses (standing, running, kicking)

Figure 4. Player #37 in burgundy jersey: 5 correct matches in top-10 (R1, R2, R4, R6, R7). The model successfully identifies the player across different camera angles and poses, demonstrating robustness to viewpoint changes and motion blur.

Figure 5. Player in blue jersey: 3 correct matches in top-10 (R6, R7, R10). This harder case shows challenges when similar jerseys appear across different teams and when significant occlusion occurs.

- **Occlusion and motion blur**: Broadcast footage often contains partial views and motion artifacts
- **Camera angle diversity**: Multi-camera coverage results in drastically different viewpoints

4.9. Final Model Selection

Based on our experimental results, we selected **OsNet-AIN as our final model** without additional ensemble or re-ranking support. This decision is grounded in both empirical evidence and theoretical considerations:

1. **Superior Performance**: OsNet-AIN achieved the highest mAP (56.83%) and Rank-1 (43.64%), outperforming all ensemble and re-ranking combinations.
2. **Ensemble Ineffectiveness**: Research on ensemble learning shows that combining models is most beneficial when they exhibit high diversity and complementary error patterns [9]. In our case, OsNet-AIN significantly outperforms the other models, meaning the ensemble is dominated by OsNet's predictions. Adding ResNet and DINOv2 introduces noise rather than complementary information, diluting the overall performance.
3. **Computational Efficiency**: Using a single model reduces inference time by 3x compared to an ensemble of three models, making it more practical for real-world deployment.

5. Conclusions and Future Works

Our experiments revealed several important insights about person re-identification in the challenging domain of broadcast soccer footage:

The experiment with DINOv2 showed moderate performance (mAP 48.09%), below CNN-based alternatives. While LoRA enabled memory-efficient training, Vision

Transformers appear to require significantly more training epochs to adapt their generic pre-trained weights to this specific domain.

ResNet-50 achieved lower performance (mAP 46.41%) than expected, suggesting that while it remains a solid baseline, more specialized architectures offer significant advantages for this task.

OsNet-AIN emerged as the clear winner with an mAP of 56.83%. Its omni-scale feature extraction combined with instance normalization proved highly effective for handling the variability in broadcast footage from different stadiums, lighting conditions, and camera angles.

A key finding was that ensemble methods and re-ranking did not improve upon the best single model. This aligns with theoretical understanding that ensembles require diverse, complementary models to be effective. When one model significantly outperforms the others, combining them can actually hurt performance by introducing inferior predictions.

The primary challenge throughout our project was computational resource limitations. We had to carefully balance model complexity with training feasibility, ultimately selecting architectures and strategies that could deliver strong results within our hardware constraints.

Among evaluated but discarded approaches, we considered **CLIP-ReID** with a **ViT-Huge** backbone. While this approach offers potentially superior accuracy and robustness, its prohibitive computational cost (billions of parameters, requiring enterprise-grade GPUs) made it infeasible for our setup.

Future Work could explore:

- (1) training DINOv2 for significantly more epochs to allow full adaptation
- (2) investigating other lightweight architectures designed for Re-ID
- (3) applying domain-specific augmentations tailored to broadcast soccer footage.

References

[1] Deliege, A., Cioppa, A., Giancola, S., Seikavandi, M. J., Duehard, J. V., Ghanem, B., Van Droogenbroeck, M., & Moeslund, T. B. (2021). SoccerNet-v2: A Dataset and Benchmarks for Holistic Understanding of Broadcast Soccer Videos. *CVPR*. 1

[2] Cioppa, A., Deliege, A., Giancola, S., Ghanem, B., Van Droogenbroeck, M., Gade, R., & Moeslund, T. B. (2022). Scaling up SoccerNet with multi-view spatial localization and re-identification. *Scientific Data*, 9(1), 356. 1

[3] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*. 2

[4] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345-1359. 2

[5] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., & Hoi, S. C. (2021). Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*.

[6] Zhou, K., Yang, Y., Cavallaro, A., & Xiang, T. (2019). Omni-scale feature learning for person re-identification. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2

[7] Zhou, K., Yang, Y., Cavallaro, A., & Xiang, T. (2021). Learning generalisable omni-scale representations for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 3

[8] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2

[9] Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2), 181-207. 3, 4, 5

[10] Zhong, Z., Zheng, L., Cao, D., & Li, S. (2017). Re-ranking person re-identification with k-reciprocal encoding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3