FOOTBALL

RE-IDENTIFICATION

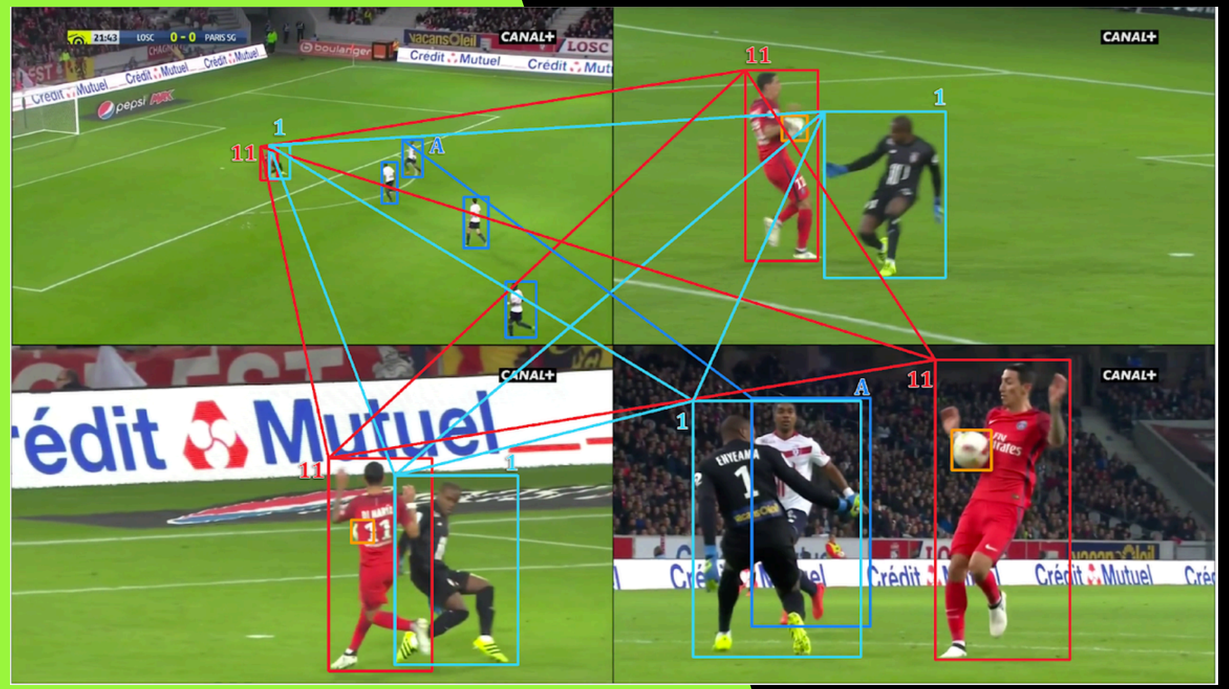
Crea Michelangelo
1993024

Gautieri Alessandro
2041850

Computer vision

WHAT IS RE-ID?

The task of identifying the same player across different angles and cameras in a broadcast video.



Goal

The task of identifying the same player across different angles and cameras in a broadcast video.

WHY IS RE-ID DIFFICULT IN FOOTBALL?

- frequent blockages between players.
- same shirt for many subjects
- Low resolution and motion blur.

CHALLENGE

The challenge is based on the SoccerNet-v3 dataset, one of the most complex challenges in sports computer vision.

Goal

Match player identities despite drastic changes in pose, scale, and viewpoint.

Evaluation Metrics

Rank-1 Accuracy: The probability that the first match found is correct.

mAP (mean Average Precision): The average precision across all queries.

PROJECT GOALS AND CONSTRAINTS

Strategies

Use a Multi-Model Ensemble approach to combine different architectures (CNN and Transformer).

Hardware constraints

The SoccerNet dataset is huge. We had to work with limited resources, making it impossible to use giant models.

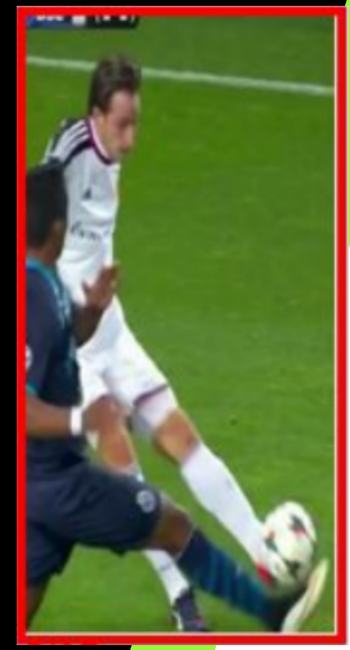
Approach

Balancing computational complexity and accuracy.

DATASET AND PREPROCESSING

- Dataset: SoccerNet-v2 ReID. Provides bounding boxes and IDs across different actions.
- Data Preparation:
  - Data Augmentation: Random erasing, flipping, and color jittering to prevent overfitting.
  - Fixed-resolution scaling.
- Breakdown: Training on dedicated set and validation on 11,638 queries and 34,355 gallery images.



METHODOLOGY

We selected three distinct architectures to capture complementary features

DINOv2

Vision Transformer (ViT) self-supervised.

ResNet-50

Standard CNN, robust baseline.

OsNet-AIN

Specific architecture for ReID.

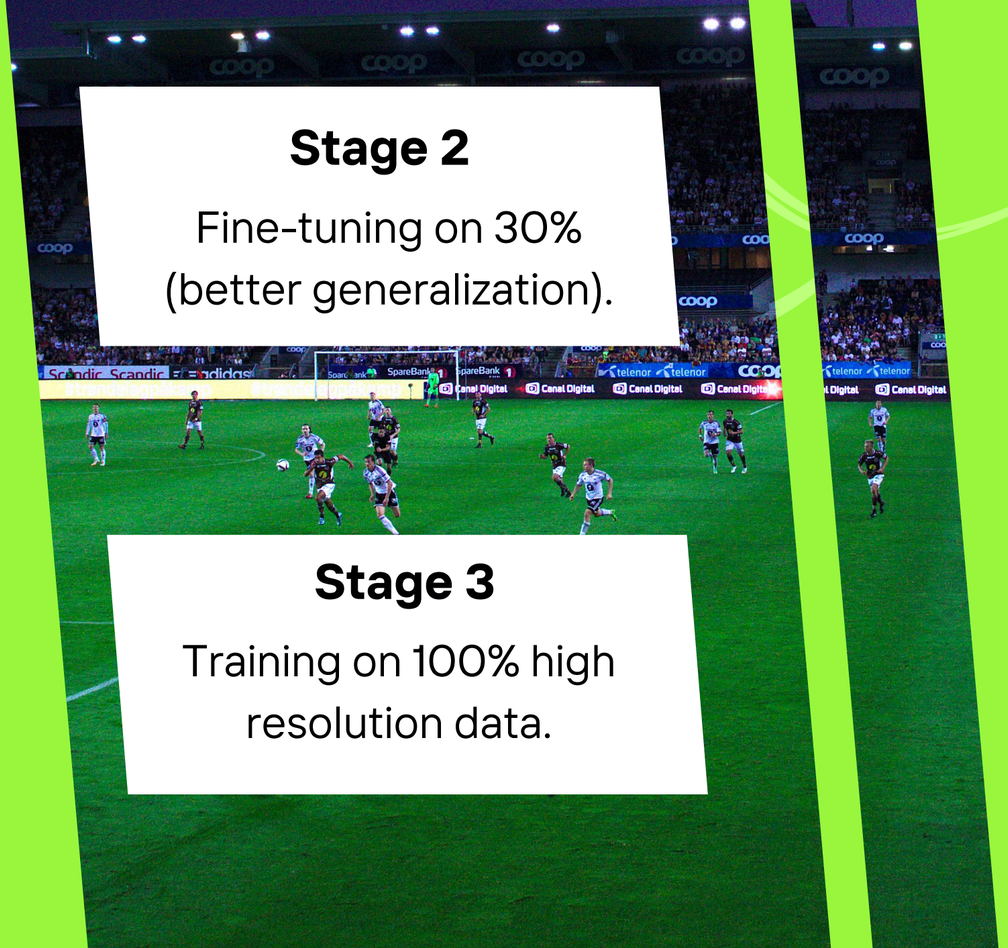
Basic idea

Merge predictions to reduce single model errors.

Training Strategy

Stage1

Fine-tuning on 10% of the data (prototyping).



Stage 2

Fine-tuning on 30% (better generalization).

Stage 3

Training on 100% high resolution data.

DINOV2

Features

Unsupervised trained transformer, great for semantic context.

Partial result

mAP 48.09% - Transformers take a long time to adapt to this domain.

Progressive Resizing technique

Simulates human learning

Low-Res Phase (256 X 128)

Focus on global shapes and dominant colors.

High-Res Phase (384 X 192)

Refining fine details.

RESNET-50

Why ResNet-50

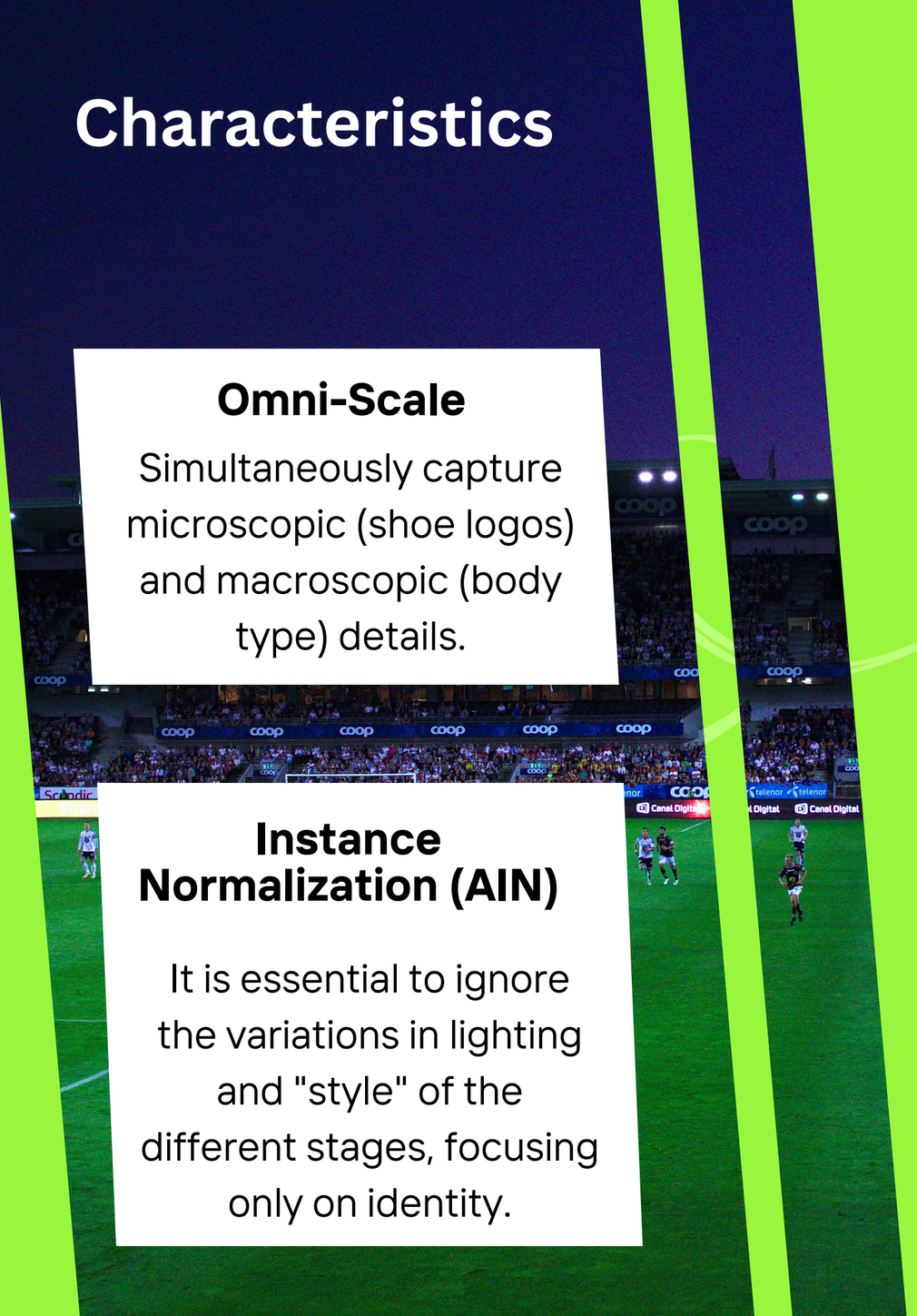
It is the de-facto standard, excellent balance between efficiency and accuracy.

Observation

The low-resolution version generalized better (mAP 46.41%) than the high-resolution one, which suffered from overfitting on unnecessary details.

OSNET-AIN

Characteristics



Omni-Scale

Simultaneously capture microscopic (shoe logos) and macroscopic (body type) details.

Instance Normalization (AIN)

It is essential to ignore the variations in lighting and "style" of the different stages, focusing only on identity.

Architecture

Designed from the ground up for Re-Identification.

Result

It is the best performing model (mAP 56.83%).

ENSEMBLE AND RE-RANKING STRATEGIES

Ensemble attempts

- **Feature Concatenation:** Joining feature vectors.
- **Distance Averaging:** Average of distances (weighted and unweighted).

Re-Ranking (k-reciprocal): Refine results by checking whether two images are "mutual neighbors".

Configurations tested: Standard, Aggressive (small surroundings), Broad (large surroundings).

RESULTS

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 λ=0.5	55.14%	42.60%

ANALYSIS

Ensenble

- **Counterintuitive result:** The ensemble did not outperform the best single model.
- **Ensembles work when the models are diverse and complementary. Here, OsNet was so superior that adding ResNet and DINO only introduced noise, diluting the accuracy.**

Re-Ranking

Even re-ranking didn't help, suggesting that the neighborhoods in the dataset are too noisy.

QUALITATIVE RESULTS

Common Errors:

- False positives due to players from other teams with similar colors
- Partial occlusions

If successful: The model identifies the player despite the change in pose and camera.

CONCLUSIONS AND FUTURE DEVELOPMENTS

Conclusions

- OsNet-AIN is the chosen model: superior in accuracy and 3 times faster in inference than the ensemble.
- Specialized architectures beat generalist ones (ResNet) on this task.

Future Developments

- Train DINOv2 for many more epochs.
- Investigate other lightweight architectures.
- Specific augmentations for the football domain.

OUR TRAINING TIME

Time spent 300 hours of training

You can watch 15 times the
lord of the ring and lo hobbit

You can listen 3000 times
bohemian rhapsody


THANKS FOR YOUR
ATTENTION