The Imposter on the Pitch: Learning Team-Agnostic Signatures for Counterfactual Evaluation

Davide Danesi

1. Introduction

A fundamental challenge in football analytics is the problem of disentanglement. A player's observable metrics, such as their pass completion rates or expected goals (xG) contribution [1], are not a pure function of their individual ability. These metrics are heavily influenced by confounding variables, including tactical systems, teammate quality, and opposition strength. This makes it difficult to answer critical questions for scouting and analysis: Is a striker's high goal tally a product of elite finishing, or of playing in a dominant system? Would a creative midfielder's output diminish in a different tactical structure? To answer these questions, methods are needed to isolate a player's context-neutral 'signature'.

A robust method for achieving this isolation is counterfactual analysis, which is often used in causal inference to estimate hypothetical "what if" scenarios. This approach moves beyond simple correlations by comparing observed outcomes with unobserved, simulated alternatives. In a football context, this would allow an analyst to ask, "What was the likely outcome had Player B taken that shot instead of Player A?". To conduct such simulations, however, a context-independent representation of "Player A" and "Player B" is a prerequisite. This representation must capture their intrinsic skill, separate from the situation itself.

Learning such representations from complex data is a challenge. In recent years, Self-Supervised Learning (SSL) has emerged as a successful approach in fields like Natural Language Processing. SSL models, such as Word2Vec [2] and BERT [3], have proven effective at learning dense vector representations, or embeddings, from large unlabeled datasets. Football, with its large spatiotemporal and event datasets, is a clear candidate for these techniques.

In this paper, we introduce a dual-objective framework designed to learn these very representations, and then we try to apply it to counterfactual evaluation. This framework, which we propose as the main contribution, has two components:

1. Contextual Swap Detection: A self-supervised, contrastive task. The model is presented with a game situation and must learn to distinguish the authentic player's signature from a set of 'imposter' signatures. To succeed, the model must learn the stylistic patterns unique to an individual, beyond their positional role.
2. Adversarial Purification: A simultaneous objective, inspired by domain-adversarial networks [4], to remove team-specific information. A secondary network is trained to

predict the player's team from their signature. The main model is, in turn, trained to produce embeddings that fool this adversary.

To conduct a foundational experiment given significant temporal and computational constraints, we simplified the problem as a first step in three key areas. First, we narrowed the scope to a single domain: learning a 'Passing Signature' from pass events. Second, we simplified the architecture, opting for a feature-engineered MLP in place of the more complex, and likely more promising, spatio-temporal models that we will detail in the methodology section. Third, we simplified the temporal scope, analyzing isolated events rather than full sequences. This non-sequential choice simplified the model's implementation and computational requirements, though we acknowledge that the evolution of an action likely represents a player's style more effectively than a static snapshot.

We present an analysis of this preliminary study. The model's outputs, upon qualitative inspection, yielded results that did not fully align with established, real-world player assessments. This suggests the model, in its simplified form, did not capture the full, nuanced stylistic signatures we aimed for. This paper, therefore, serves a dual purpose: to formally introduce the complete learning framework, and to document the findings of this foundational experiment. By analyzing the model's shortcomings, we identify several lessons learned and propose a clear path for future research.

2. Related works

Our proposed framework intersects several established areas of research. This section reviews relevant work in football performance analysis, representation learning, and the self-supervised and adversarial methods from machine learning that inform our approach and from which we drew inspiration.

Efforts to quantify player contributions in football have evolved from simple event counts to more sophisticated probabilistic models. Metrics such as Expected Goals (xG) [1] and Expected Assists (xA) [5] measure the quality of a shot or a pass by accounting for the situational context in which it occurred. Frameworks like Valuing Actions by Estimating Probabilities (VAEP) [6] go further, attempting to assign a value to every on-ball action by measuring its impact on the probability of scoring or conceding. While these models are powerful descriptive tools, their outputs remain fundamentally context dependent. They are effective at valuing a player's past production within a specific system, as their calculations are based on the aggregated outcomes of similar situations. However, they are not designed to disentangle the player's intrinsic ability from that system, which limits their predictive power for counterfactual analysis.

The core mechanism of our "Contextual Swap Detection" task is inspired by self-supervised learning in Natural Language Processing. Specifically, it draws from the principles popularized by models like Word2Vec [2]. In this paradigm, a model learns a word's meaning, its vector embedding, based on the "distributional hypothesis", which posits that a word's meaning is defined by the contexts in which it appears. The model is trained on a simple, self-supervised task, such as predicting a target word from its surrounding context words (Continuous Bag-of-Words) or, more relevant to our work, distinguishing a true context-word pair from randomly generated "negative" or "imposter" pairs. This contrastive objective, known as negative sampling, forces the model to learn a semantically rich representation from unlabeled text. We propose to adapt this principle to the football domain: we treat the player as the "word" and the game situation (pass characteristics and 360-degree player locations) as the "context." By training a model to distinguish the authentic player for a given situation from a set of 'imposter' players, we aim to force the model to learn a "signature" that captures that player's unique and appropriate fit within that specific game context.

Our proposal for "Adversarial Purification" builds on principles from domain-adversarial training. This technique is often structured as a two-player game, most famously in Generative Adversarial Networks (GANs). In a Domain-Adversarial Neural Network (DANN) [7], a model is trained on two tasks simultaneously: a main task (e.g., classification) and a secondary "adversarial" task (e.g., predicting a "nuisance" attribute, such as the data's domain). The model's feature extractor is trained to succeed on the main task while failing at the adversarial task, often by maximizing the adversary's loss via a gradient reversal layer or an equivalent loss-subtraction method. This forces the model to learn representations that are "invariant" to the nuisance attribute. We adapt this concept directly, treating a player's "team" as the nuisance variable we wish to remove. By training our primary model to produce player signatures that an adversary cannot use to predict the player's team, we aim to "purify" the embedding, removing team-specific tactical biases and isolating the individual's context-free style.

3. A Framework for Learning Player Signatures

This chapter details the complete methodological framework designed to learn disentangled player signatures. The architecture is predicated on the hypothesis that a player's intrinsic style can be learned by a model trained to distinguish that player's signature from 'imposter' signatures within a given game context.

We first formalize the comprehensive, spatio-temporal architecture that we hypothesize is necessary to fully capture the nuances of player style. This represents the 'full vision' of the project. We then describe the foundational, simplified model that was subsequently implemented as the basis for the experimental results presented in this paper.

This project utilizes data provided by Hudl, covering five full seasons of the Italian Serie A (2020/21 to 2024/25). The framework and subsequent experiments detailed in this paper are designed to operate on two rich, interconnected datasets: StatsBomb Event Data , which provides timestamped information on all on-ball actions (e.g., passes, shots), and StatsBomb 360 Data, which provides 'freeze frames' detailing the locations of players on the pitch corresponding to these events.

3.1 The Full Vision: A Spatiotemporal Graph Model

A player's "style" is unlikely to be fully encoded in a single, isolated action. It is more likely a function of sequential decision-making, off-ball movement, and reactions to complex spatial configurations. An isolated pass event, for example, lacks the context of the player's prior movements or the subtle shifts in team shape that preceded it. Therefore, we posit that the ideal unit of analysis is the full possession sequence, modeled using a spatiotemporal architecture.

This architecture would consist of two primary components: a spatial encoder and a temporal encoder.

The spatial component would be designed to interpret the complex relationships between all players on the pitch at a given moment. The StatsBomb 360 data provides a freeze frame for an event, detailing the location of players. This data is a natural fit for a Graph Neural Network (GNN). In this model, each player in the 360 frames would be treated as a node. Node features would include their 2D coordinates, velocity (derived from sequential frames), and their role in the event, such as teammate, keeper, or actor. The GNN's role would be to automatically learn the high-order relationships between these nodes—such as defensive structures, passing lanes, and areas of spatial control—replacing the need for extensive manual feature engineering.

The temporal component would then be required to understand how these spatial snapshots evolve. A possession is a sequence of events, each with its own corresponding graph representation from the GNN. A Recurrent Neural Network (RNN), such as an LSTM, would process this sequence of graph embeddings. The final hidden state of the RNN would, in theory, represent a rich, spatiotemporal embedding of the entire possession. This complete possession embedding would then serve as the definitive "context" for the learning tasks detailed in Section 3.3.

3.2 Core Learning Objectives

The architecture described in Section 3.1, regardless of its specific implementation, is designed to serve two simultaneous training objectives. These objectives are the core of our proposed framework: a primary task to learn a player's style, and a secondary adversarial task to ensure that style is disentangled from team-level context.

3.2.1 Contextual Swap Detection: Learning Stylistic Signatures

The primary learning objective is a self-supervised, contrastive task we refer to as "Contextual Swap Detection." The central hypothesis is that a player's individual style can be learned by a model trained to recognize the "fit" between a player and the specific game situation in which they are acting.

The task is structured as a binary classification problem. For a given event, such as a pass, the model is presented with a "positive pair" and several "negative pairs."

1. Positive Pair: This pair consists of the authentic situation embedding (derived from the GNN/RNN or a feature-engineered MLP) and the player embedding (from the main embedding table) of the player who was involved in that event. This pair is assigned a label of 1 (true).
2. Negative Pairs: These pairs consist of the same situation embedding but are paired with "imposter" player embeddings-signatures of players who were not involved in the event. These pairs are assigned a label of 0 (false).

A classifier, termed the Swap-Detector, takes a concatenated situation and player embedding and outputs a single logit score. The model is trained to minimize the binary cross-entropy loss on these pairs, learning to output a high score for the positive pair and low scores for all negative pairs.

The efficacy of this approach is highly dependent on the "negative sampling" strategy. A naive random sampling of imposters would present the model with a trivial task. For example, if the situation is a defensive tackle in the penalty box, and the imposter is an attacking striker, the model can easily learn to reject the pair. This, however, would only teach the model to distinguish roles, not individual style.

To force the model to learn a more granular signature, we propose a sampling curriculum with increasing difficulty:

- Easy Samples (Different Role): A small fraction of imposters are drawn from players with a different positional role. This teaches the model the basic, high-level context of different positions.
- Medium Samples (Same Role, Different Team): A larger fraction of imposters are drawn from players who share the same role as the authentic player but play for a different team. This forces the model to move beyond simple role-detection and begin to distinguish, for example, the characteristics of a possession-based midfielder from a counter-attacking midfielder.
- Hard Samples (Same Role, Same Team): The largest fraction of imposters are drawn from the authentic player's own teammates who share the same role. This is the most important step. By forcing the model to distinguish between two players who operate in the same position and the same tactical system (e.g., two starting center-backs), the model can only succeed by learning the subtle, unique, and repeatable patterns of individual behavior that constitute "style."

The loss for this task is the sum of the binary cross-entropy for the positive example and the mean of the binary cross-entropy for all negative examples.

3.2.2 Adversarial Purification: Disentangling Team Context

The Swap Detection task alone is insufficient to guarantee a context-neutral signature. The situation embedding is derived from a team's collective action, and the player embedding will therefore learn to absorb this team-wide information. The model may learn that a player who fits a "short, patient build-up" context is likely a "Player X" type, but it might also learn that this "Player X" type is characteristic of their specific team. The resulting signature would be entangled with team identity.

To solve this, we introduce a secondary, simultaneous objective: adversarial purification. This task, inspired by domain-adversarial networks, treats the team as a "nuisance" variable that we want the player signature to be invariant to.

The mechanism involves a second classifier, the Team Adversary, which is trained only to predict a player's team from their player embedding. The training process becomes a "game" with two steps:

1. Adversary Step: The Team Adversary network is trained to get better at its job. It minimizes a standard cross-entropy loss (L_adv_ce) based on its prediction of the team from the (detached) player embedding.
2. Main Model Step: The main model's loss function is modified. It is trained to minimize the Swap Detection task's loss (L_swap) while simultaneously maximizing the Team Adversary's loss. This is achieved by subtracting the adversary's loss from the main loss function: $L_{main} = L_{swap} - \lambda \cdot L\_adv\_ce.$

By subtracting the adversary's loss (a technique equivalent to a gradient reversal layer), we create a gradient update that penalizes the model's signature-learning process for any information that helps the adversary. The player signatures themselves are thus trained to become maximally useful for the swap detection task (stylistically accurate) while being maximally useless for the team prediction task (context-neutral).

3.2.3. The Combined Training Objective

The framework's training process requires two separate optimizers. The first is dedicated solely to the adversarial network, updating its weights to minimize the team-prediction loss. The second optimizer is responsible for all main model components, updating the core player representations and the modules related to the swap-detection task. This optimizer minimizes the combined loss function, which simultaneously rewards stylistic accuracy and penalizes any information that aids the adversary. The hyperparameter λ controls this trade-off, weighting how strongly the model prioritizes purification versus accuracy.

3.2.4. Application: Counterfactual Evaluation

The learned player signatures are not an end in themselves. They are a foundational asset designed for use in downstream supervised tasks. The primary application we envisioned for these purified embeddings is counterfactual analysis, the method by which we aim to isolate a player's individual contribution.

The process would involve training a separate, supervised "outcome model." This model would learn to predict the probability of a successful outcome, such as pass completion, by taking two inputs: the encoded situation and a player signature.

Once this outcome model is trained, it can be used to run simulations. We can take a large, static set of real-world situations (e.g., thousands of passes or shots attempts from the dataset) and iterate through our database of players. For each situation, we would predict the outcome by pairing the encoded situation with each player's signature. By aggregating these predicted probabilities, we can estimate how many passes or shots Player A would have been expected to complete in those situations versus Player B. This aggregated difference would represent a context-neutral measure of their individual skill, fully isolated from the specific situations they happened to face.

3.3 A Foundational Experiment: The Isolated Pass Model

To conduct an initial test of the core training objectives, we designed a practical experiment that adheres to the principles from Section 3.2. This experiment necessitates significant simplifications of the "Full Vision" architecture. These design choices were made to reduce the computational and temporal costs associated with training spatiotemporal GNNs on sequential data, allowing for a focused, iterative validation of the learning logic itself.

The first simplification was to limit the scope of the analysis. Rather than modeling all event types, this experiment focuses exclusively on passes. Passes are the most frequent on-ball action in a match, are rich in stylistic variation, and provide a well-defined domain to test the "Contextual Swap Detection" hypothesis. This modular approach suggests that if a "Passing Signature" can be successfully learned, the same framework could be independently applied to other domains, such as learning "Shooting Signatures" from shot events, with these specialized embeddings potentially being combined to offer a complete, multifaceted representation of a player.

The second simplification was to remove the temporal component. We moved the unit of analysis from a full possession sequence to an isolated event. This decision removes the need for a recurrent (RNN) architecture. The cost of this simplification is the loss of all sequential information, such as a player's off-ball movement leading up to the pass or the chain of events that created the situation. The "context" is therefore limited to the single snapshot in time when the pass occurs.

The third simplification was to replace the spatial GNN encoder with a feature-engineered Multi-Layer Perceptron (MLP). Instead of tasking a GNN with learning spatial relationships from raw 360 data, we manually curated a flat feature vector for each pass. This vector combines information from both the event data (pass geometry, type) and the 360 data (player locations, distances). This situation encoder (an MLP) thus relies on our ability to manually identify and extract the most salient features, rather than learning them automatically.

The resulting model is a three-headed architecture trained on this feature-engineered pass data. Its inputs are processed as follows:

- Categorical Features: pass outcome and pass height are converted to integer indices.
- Continuous & Boolean Features: A large vector of continuous features is compiled, including pass geometry (e.g., pass length, pass angle), spatial context (e.g., the distance to the nearest opponent), and event-level metrics (e.g., StatsBomb's on ball value [8]). This vector is normalized using z-score standardization.
- Final Input: The outputs of the categorical embedding layers are concatenated with the normalized continuous feature vector to form the final input to the situation encoder.
- The model architecture consists of a central, learnable stylistic embedding table of 128 dimensions for each player, which is the primary output of this experiment. This table is trained by three distinct MLP "heads":

- The situation encoder: An MLP that processes the final input vector (described above) and outputs a 256-dimensional situation embedding.
- The Swap Detector: An MLP that takes a concatenated situation embedding and a 128-dimensional player embedding as input. It outputs a single logit score, representing the "correctness" of that player in that situation.
- The Team Adversary: An MLP that takes only the 128-dimensional player embedding as input and outputs a logit score for each team.
- This model was trained using the dual-objective loss function $L_{main} = L_{swap} - \lambda \cdot L\_adv\_ce$, as defined in Section 3.2.3, with two separate Adam optimizers. The negative sampling strategy for L_swap was explicitly defined with a 10/20/70 ratio of (different-role, same-role/different-team, same-role/same-team) imposters to force the model to learn fine-grained stylistic differences.

The intended outcome of this experiment was to produce a validated, frozen stylistic embedding table. This table would then serve as a key component for downstream supervised models, as described in Section 3.2.4.

4. Results

This section details the results of the experiment described in Section 3.3. We first provide the specific architectures of the models, then present the quantitative training metrics, and finally, we conduct a qualitative and quantitative analysis of the learned player embeddings.

4.1. Model Architectures

The experiment consisted of two models: a primary signature-learning model and a downstream counterfactual model.

4.1.1. Signature Learning Model

This model is trained to produce the player embeddings and is composed of three main components:

- Player Embedding Table: A learnable lookup table size (1377 players, 128 dimensions), initialized from a normal distribution $\mathrm{N}(0, 0.01^2)$.
- Pass Situation Encoder: An MLP that processes the input features. The input vector has 61 dimensions (37 continuous features + 24 from categorical embeddings for pass outcome and pass height). The architecture consists of three layers:
    1. Linear(61 -> 256) -> BatchNorm1d(256) -> ReLU -> Dropout(0.1)
    2. Linear(256 -> 256) -> BatchNorm1d(256) -> ReLU -> Dropout(0.1)
    3. Linear(256 -> 256) -> ReLU

    This outputs a 256-dimensional embedding representing a situation (event + associated 360 data).

- Swap Detector Head: An MLP classifier that takes the 384-dimensional concatenated vector (situation embedding + player embedding) as input.

    1. Linear(384 -> 256) -> ReLU -> Dropout(0.1)
    2. Linear(256 -> 256) -> ReLU -> Dropout(0.1)
    3. Linear(256 -> 1) (Outputting a single logit).

- Team Adversary Head: An MLP classifier that takes the 128-dimensional player embedding as input.

    1. Linear(128 -> 256) -> ReLU -> Dropout(0.1)
    2. Linear(256 -> 256) -> ReLU -> Dropout(0.1)
    3. Linear(256 -> 28) (Outputting 28 team logits).

4.1.2. Counterfactual Pass-Outcome Model

This is a separate, supervised model trained to predict pass completion.

- Context Encoder: An MLP that processes a 157-dimensional input (84 continuous features + 73 from categorical embeddings for pass height, body part, etc.). It consists of two layers (Linear(157 -> 256) and Linear(256 -> 256)) and outputs a 256-D context vector.
- Player Embedding Table: An embedding table of (1377, 128).
- Predictor Head: An MLP that takes the 384-D concatenated context and player embedding to predict a single logit for pass success.

4.2. Model Training & Convergence

The signature-learning model was trained for 12 epochs using 64 negative samples per pass and an adversarial weighting $\lambda = 0.2$. To evaluate the model's learning process on its dual objectives, we will focus our analysis on two key metrics: the accuracy of the swap-detection task and the accuracy of the adversarial task.

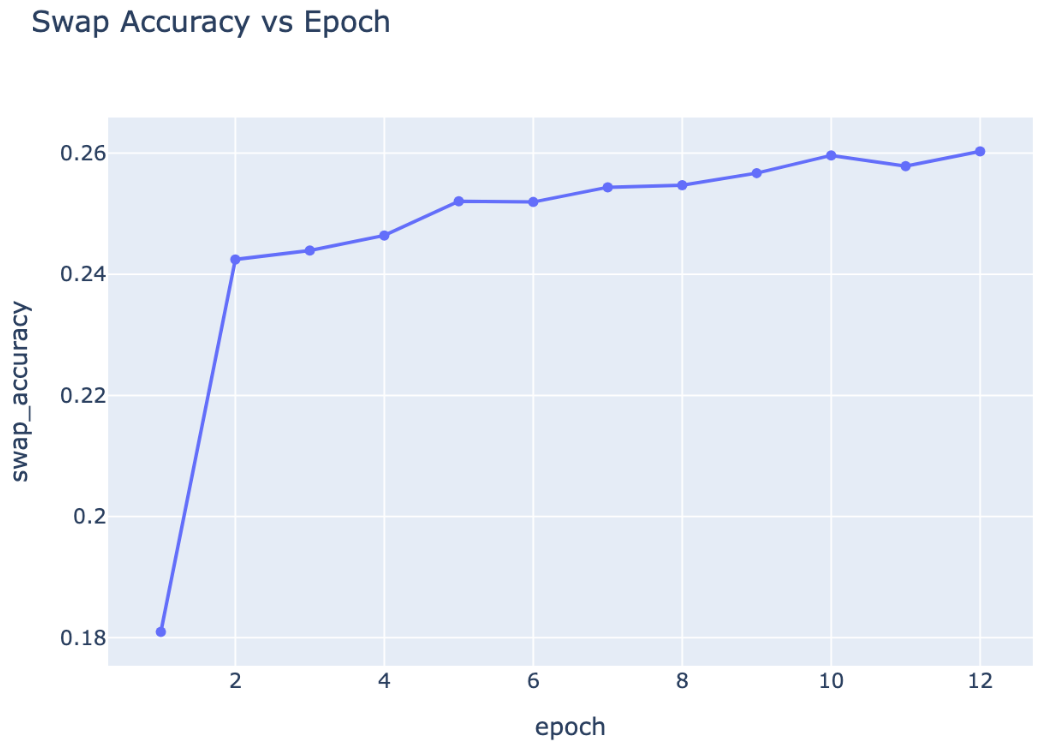

Figure 4.1: Validation swap accuracy per epoch. This metric measures the "ranking accuracy," or the model's success rate at scoring the authentic player higher than all 64 imposter players.

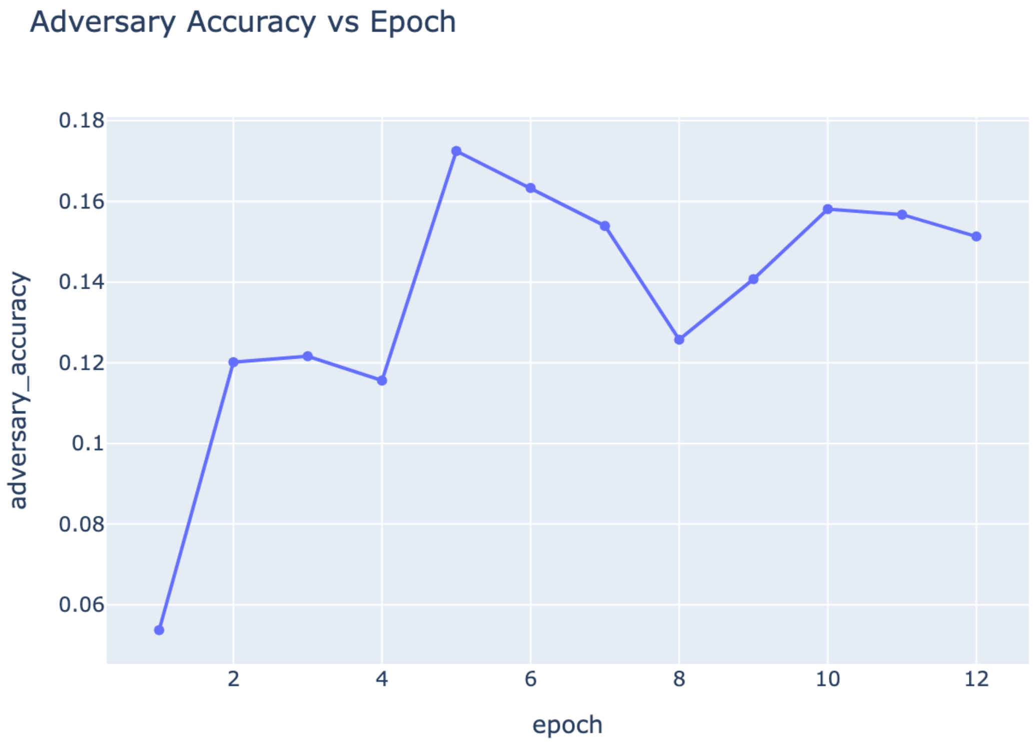

Figure 4.2: Validation adversary accuracy per epoch. This metric measures the success rate of the adversarial network in predicting a player's team from their signature.

Figure 4.1 shows the evolution of the swap-detection accuracy. It is important to note this metric is not a simple binary accuracy but a "ranking accuracy." It measures the percentage of time the model correctly scores the one positive pair higher than all 64 negative pairs. With 65 total options (1 positive, 64 negative), the baseline for random chance is approximately 1.5%. The model's final validation accuracy of ~26% is therefore significantly better than random, indicating that the model successfully learned a meaningful signal from the data.

Figure 4.2 shows the validation accuracy of the team-adversary network. With 28 teams in the dataset, a random guess would yield an accuracy of ~3.6%. The adversary's accuracy, while fluctuating, remains low (peaking at ~17% and ending at ~15%). This suggests that while the adversary can find some team-related signal, the main model's purification objective is largely effective in suppressing a strong, team-based clustering.

4.3. Qualitative Analysis of the Learned Signature Space

To understand the structure of the learned 128-dimensional player signatures, we used Uniform Manifold Approximation and Projection (UMAP). UMAP is a non-linear dimensionality reduction technique that projects high-dimensional data into a lower-dimensional space, in this case, two dimensions for visualization. It is effective at preserving the topological structure of the data, revealing underlying clusters and relationships.

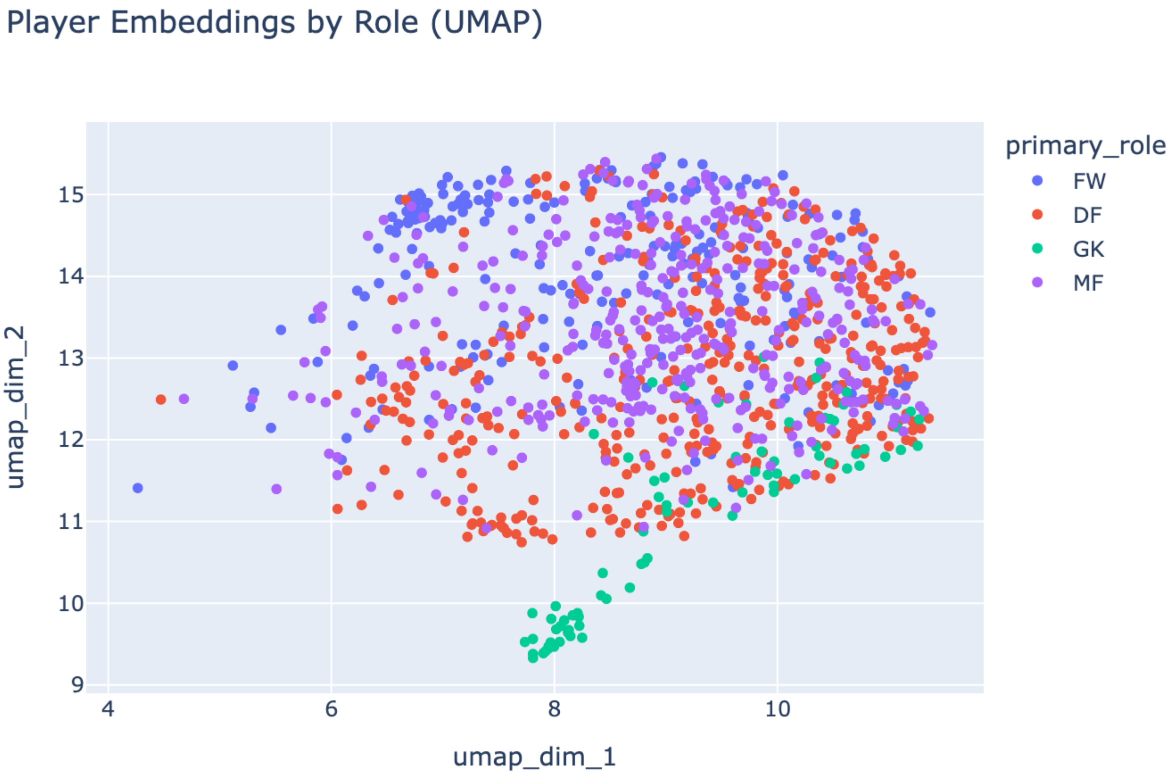

Figure 4.3: A UMAP visualization of the learned 128-dimensional player signatures, colored by primary role.

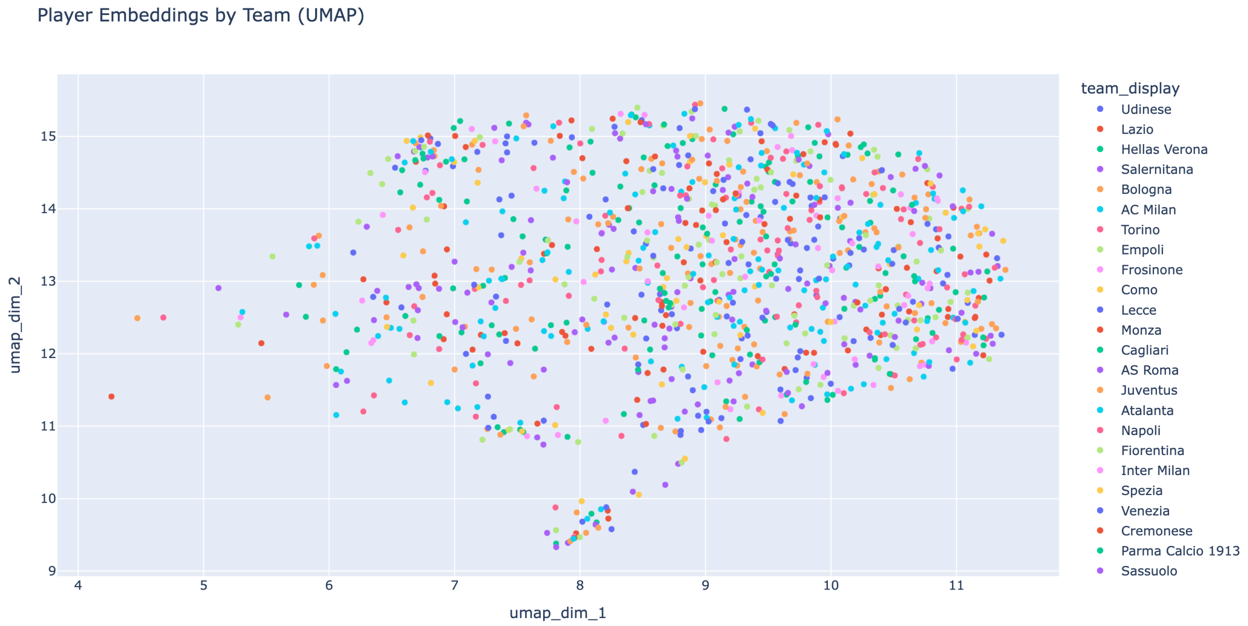
Player Embeddings by Team (UMAP)

Figure 4.4: The same UMAP visualization, colored by the player's team to assess the effectiveness of the adversarial purification. Legend is capped at 24 teams.

Figure 4.3 shows the resulting embedding space, with each player colored by their primary positional role (FW, MF, DF, GK). The visualization shows clear, distinct regions. Goalkeepers (GK) form a tight, isolated cluster, which is expected given their unique passing profile. Defenders (DF) and Forwards (FW) also form relatively distinct regions, with Midfielders (MF) occupying a central space between them. This result confirms that the model successfully learned the high-level passing characteristics that define a player's role on the pitch.

Figure 4.4 shows the same embedding space, but colored by the players' teams. In stark contrast to the role-based clusters, there are no discernible team-based groupings. The team colors are intermixed in a seemingly random distribution. This provides strong visual evidence that the adversarial purification objective was effective, since the model did not organize the embedding space along team-specific lines.

4.4. Quantitative Analysis of Embedding Similarity

To quantitatively substantiate the clustering patterns observed in the UMAP visualizations, we computed the mean cosine similarity between the learned player embeddings. Cosine similarity measures the orientation of two vectors; a high score (near 1.0) indicates that two players' signatures are very similar, while a score near 0.0 indicates dissimilarity. This analysis allows us to move beyond a purely visual interpretation and apply a rigorous metric to our findings.

We performed two separate analyses: one based on role and one based on team. For each analysis, we calculated the average similarity for all unique pairs of players within the same group (e.g., 'Intra-Role' for all pairs of midfielders, or 'Intra-Team' for all pairs of teammates). We then compared this value to the average similarity of pairs from different groups (e.g., 'Cross-Role' or 'Cross-Team'). A high intra-group similarity, relative to the cross-group baseline, would provide quantitative evidence of distinct clustering.

Grouping	Comparison	Mean Cosine similarity
Role-Based	Intra-Role (Goalkeepers)	0.224
	Intra-Role (Avg. Defenders)	0.014
	Intra-Role (Avg. Midfielders)	0.238
	Intra-Role (Avg. Forwards)	0.079
	Cross-Role	0.002

Table 4.1 (Role): A comparison of mean cosine similarity for player pairs within the same positional role (Intra-Role) versus different roles (Cross-Role).

Grouping	Comparison	Mean Cosine similarity
Team-Based	Intra-Team (Teammates)	0.010
	Cross-Team (Non-Teammates)	0.010

Table 4.2 (Team): A comparison of mean cosine similarity for player pairs on the same team (Intra-Team) versus different teams (Cross-Team).

The results in Table 4.1 confirm the visual analysis. The role similarity for Goalkeepers (0.224) is an order of magnitude higher than for other roles, quantitatively matching the tight cluster seen in the UMAP plot. Crucially, the team similarity (0.0097) is functionally identical to the cross-team similarity (0.0104). This result quantitatively supports the visual finding that the learned embeddings are team-agnostic.

4.5. Plausibility Check of Learned Signatures

A key qualitative test for the learned signatures is a real-world plausibility check: do the 'stylistically similar' players identified by the model make sense? We used a nearest-neighbors analysis to find the five most similar players (by cosine similarity) for a sample of players.

Target Player	1st Nearest Neighbor (similarity)	2nd Nearest Neighbor (similarity)
G. Di Lorenzo	S. Chukwueze (0.425)	W. Kechrida (0.368)
M. Locatelli	S. De Maio (0.361)	M. Pašalić (0.329)
M. de Roon	N. Bajrami (0.477)	N. Cambiaghi (0.381)
A. Rrahmani	F. Ravaglia (0.399)	C. Terzi (0.389)
A. Bastoni	D. Criscito (0.495)	Gervinho (0.394)

Table 4.3: Sample of nearest-neighbor analysis. This table shows the top two most similar players (by cosine similarity) for a selection of target players.

The results from this analysis, sampled in Table 4.3, reveal some limitations of the learned embeddings. While some pairings show intuitive similarity (e.g., Alessandro Bastoni's closest neighbor is Domenico Criscito, another left-sided, ball-playing defender), many of the other recommendations are counter-intuitive.

While a player's passing style is not strictly defined by their positional role, some of the model's pairings are difficult to interpret. For example, the model finds Amir Rrahmani (DF) to be most similar to Federico Ravaglia (GK), and Manuel Locatelli (MF) to be most similar to Sebastian De Maio (DF). This suggests that while the model successfully learned a high-level role representation and discarded team context, the remaining signal captured in the embedding is not yet sufficiently informative to align with granular, real-world assessments of individual passing style.

4.6. Counterfactual Model Plausibility Check

The final test was to use the downstream Counterfactual Pass-Outcome Model to assess the real-world plausibility of the learned embeddings. To do this, we first trained the model (described in Section 4.1.2) to predict pass completion.

We then created a challenging validation dataset by filtering for all pass events where the Statsbomb-provided success probability was less than 0.7, ensuring we did not only evaluate trivial passes. This resulted in a set of 1,622,086 pass situations. For each of these situations, we ran a counterfactual simulation: we paired the encoded situation with every player's learned signature (for all 1,377 players) and had the model predict the success probability.

Players were then ranked by their mean predicted probability of completing these passes, with the results for the top 20 and bottom 10, filtered to players with at least 750 total passes in the dataset, shown in Table 4.4.

Top 15 Players: Mean Predicted Pass Success		
1. A. Vogliacco (0.8212)	6. D. Kamada (0.8174)	11. M. Demiral (0.8162)
2. D. De Gea (0.8193)	7. T. Reijnders (0.8172)	12. G. Medel (0.8160)
3. S. Handanovic (0.8191)	8. M. Adopo (0.8171)	13. N. Gyömbér (0.8159)
4. A. Terzič (0.8179)	9. K. Thuram (0.8167)	14. H. Kamara (0.8155)
5. M. Kempf (0.8176)	10. M. Lopez (0.8164)	15. Y. Paredes (0.8154)

Table 4.4: Counterfactual Model Rankings for Predicted Pass Success

The rankings in Table 4.4 show a partial alignment with intuition. For example, the top 15 contains several players who are known for their passing ability (e.g., Daichi Kamada, Tijjani Reijnders, Paredes). Conversely, the bottom 10 includes players like Stefano Sturaro and Perparim Hetemaj, who are primarily "interdicting midfielders", not known for passing precision.

However, these details are overshadowed by several counter-intuitive results. The top of the list features unexpected players, and many of the league’s most elite playmakers are absent. A significant anomaly is the presence of four goalkeepers (De Gea, Handanovič, Sepe, Patrício) in the top 20. This suggests that the passing patterns of goalkeepers represent a fundamentally distinct domain from outfield players, and it may be more appropriate to model them separately.

The most critical finding, however, is the narrow range of the mean predicted probability scores. The difference between the top-ranked player (0.8212) and the bottom-ranked player (0.7948) is less than 2.7 percentage points. This extremely small variance indicates that the player signature's influence on the counterfactual model's prediction is limited. The model appears to be relying heavily on the pass situation features, with the signature itself acting as a less impactful component. This reinforces the findings from the nearest-neighbor analysis: the learned signature, in its current form, may not yet be sufficiently informative to capture the full, nuanced range of individual passing skill.

5. Discussion

The results of our experiment present a nuanced picture. The model's performance on its dual objectives requires careful interpretation. The findings from Section 4.2 and 4.3 suggest the training process was successful in two key areas. First, the model's swap accuracy, which reached 26% against a random-chance baseline of ~1.5%, confirms that the model learned a meaningful signal. Second, the adversarial objective appears to have been largely effective. The UMAP plot by team (Figure 4.4) shows no discernible team-based clusters, and the quantitative similarity metrics (Table 4.1) confirm that the intra-team similarity is functionally identical to the cross-team similarity.

However, these positive findings are met with challenges in the qualitative validation. The nearest-neighbor analysis (Section 4.5) produced pairings that are often counter-intuitive from a footballing perspective. This suggests that while the model successfully learned a team-agnostic representation of a player's role, the remaining information captured in the embedding, in this form, does not appear sufficiently rich to fully align with observable, real-world assessments of individual passing style.

We do not believe these findings are contradictory. Rather, they suggest a coherent hypothesis: the architectural simplifications made for this foundational experiment may have resulted in an input signal that was not sufficiently complex for the model's most difficult task. The model was given a "hard" objective in the swap-detection task: to distinguish between two players in the same role and on the same team. To solve this, the model would need access to a rich, stylistic passing signal. Our adversarial network, however, was simultaneously (and successfully) removing one of the strongest available signals: team-wide tactical identity. Once both role (the "easy" signal) and team (the "adversarial" signal) were accounted for, it is plausible that the remaining stylistic signal in the static, isolated pass data was subtle and perhaps insufficient for the model to learn a deeply nuanced representation.

This points to the two primary simplifications as areas for further investigation. First, the reduction of the problem from sequences to isolated events removes temporal context. This simplification is analogous to providing a scout with a set of static photographs of a pass, rather than a video of the entire play. A player's passing style is likely not encapsulated in a single snapshot, but is rather a dynamic property revealed by their off-ball movement, their decision to pass instead of carry, or their patterns of play across an entire possession. By design, our model did not have access to this sequential information.

Second, the substitution of a GNN with a feature-engineered MLP required us to manually define what constitutes the "situation." Our 61-dimension feature vector was an attempt to describe the game state, but it is possible that this description did not capture all the subtle interactions a model would need to identify fine-grained passing style. The framework, therefore, may not have been provided with the richest possible signal, and the resulting embeddings, while successfully purified, are not yet as semantically meaningful as intended.

This leads to a final, positive interpretation. The fact that the two core components of the framework functioned, the adversarial network successfully purified the embeddings, and the swap-detection task learned a structured, role-based signal, is a promising outcome. It suggests that the framework itself is a viable approach for this problem. The limitations identified in this experiment appear to stem not from a flaw in the training logic, but from the simplified nature of the input signal. It is hypothesized that by applying this proven framework to a richer, spatiotemporal data representation, such as one derived from GNNs and full possession sequences, it holds considerable potential for learning the nuanced, individual signatures we set out to find.

6. Conclusion

This paper addressed the problem of disentangling player skill from situational context, a key challenge in football analytics. We introduced a novel learning framework designed to learn context-neutral player signatures. These signatures are designed to be a versatile, foundational asset for various downstream tasks. For instance, one application we envisioned is their use in counterfactual analysis, a method which isolates an individual's contribution by simulating the outcomes of events under the hypothetical condition that a different player was involved. Our framework is defined by a training process with two simultaneous, competing objectives: a self-supervised 'Contextual Swap Detection' task to learn individual player style, and an 'Adversarial Purification' objective to remove confounding, team-specific information.

The primary contribution of this work is the formalization of this framework. We then presented a foundational experiment designed to test its core viability. This experiment was intentionally simplified, focusing on a pass-only model that analyzed isolated events using a feature-engineered Multi-Layer Perceptron, rather than the more complex spatiotemporal architecture we initially hypothesized would be necessary.

As detailed in Section 4, the analysis of this foundational study's embeddings presented a nuanced picture. The qualitative and counterfactual results indicated that while the model successfully learned role-based patterns, the resulting signatures did not yet fully align with real-world assessments of individual style. One interpretation of this finding is that the architectural oversimplification, rather than an issue with the training logic, may be a primary factor. The result suggests that a player's "style" might be a complex signal, perhaps not strongly present in isolated event data. It is possibly encoded in temporal dependencies and complex spatial interactions, which were the elements simplified in this model. The value of this paper, therefore, is in providing a formalization of the proposed framework and presenting a critical analysis of a foundational experiment, which we hope can inform subsequent research.

This result suggests several potential paths for future work. A logical next step would be to replace the feature-engineered situation encoder with a Graph Neural Network (GNN) to automatically learn spatial relationships from the 360 data. Following this, a temporal component, such as a recurrent network or transformer, could be implemented to model full possession sequences. Further investigation is also warranted into the stability of the dual-objective loss and the optimization of the negative sampling curriculum. Once validated, this modular framework could then be extended to other domains, such as learning 'Shooting Signatures' or 'Defensive Signatures', with such embeddings potentially being combined to offer a complete, multifaceted representation of a player.

References

[1] Eggels, H., van Elk, R., & Pechenizkiy, M. (2016). Expected goals in soccer: Explaining match results using predictive analytics. Proceedings of the Machine Learning and Data Mining for Sports Analytics Workshop (MLSA 2016).

[2] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems 26 (NIPS 2013).

[3] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019.

[4] Ganin, Y., Ustinova, E., Ajakan, H., et al. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(1), 2096–2030.

[5] Stats Perform. (2021). What are expected assists (xA)? The Analyst (Stats Perform explainer article).

[6] Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD 2019).

[7] Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems 27 (NIPS 2014).

[8] StatsBomb. (2021). Introducing On-Ball Value (OBV). StatsBomb News/Blog (September 16, 2021).