UNIVERSITÀ
di **VERONA**
Dipartimento
di **INFORMATICA**

From Prediction to Interaction: Exploring Deep Learning Forecasting Models for Human-Centered Industrial Applications

Andrea Avogaro

Supervisor:
Prof. Marco Cristani

Department of Computer Science
University of Verona

Doctoral Dissertation
Doctoral Program in Computer Science
37^{th} *cycle*

2025

La borsa di dottorato è stata cofinanziata con risorse del
Programma Operativo Nazionale Ricerca e Innovazione 2014-2020, risorse FSE REACT-EU
Azione IV.4 “Dottorati e contratti di ricerca su tematiche dell’innovazione”
e Azione IV.5 “Dottorati su tematiche Green”

Declaration

I hereby declare that the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Andrea Avogaro
2025

Abstract

The integration of Artificial Intelligence (AI), particularly Deep Learning (DL), is driving a paradigm shift in industrial practices, moving beyond traditional automation and production management. In the era of Industry 5.0, the focus is shifting toward human-centered systems that not only improve efficiency but also promote safety, adaptability, and sustainability. Deep Learning plays a key role in this transition, offering the ability to process complex, multimodal data and provide real-time insights in line with the dynamic nature of modern industrial environments. By enabling systems to anticipate, adapt, and interact, Deep Learning transforms prediction from a passive tool to an active tool of Human Robot Collaboration (HRC).

Traditional forecasting methods, while effective in structured and predictable scenarios, often struggle with the complexity and variability inherent in dynamic industrial applications. This thesis investigates how Deep Learning models can bridge this gap by progressively scaling capabilities across different applications of Time-Series Forecasting: from production planning to adaptive systems, and ultimately to real-time interaction between robots and humans.

The contributions of this thesis are demonstrated through three key industrial tasks, each exemplifying the main challenges of the design of a more sustainable, safe, and human-centered industry. The first tackled task is called New Fashion Product Performance Forecasting (NFPPF), addressing the complexity of predicting demand for garments with no historical sales data. This task highlights how can support sustainable production planning, reduce waste, and align supply with market needs. The second task explores human trajectory forecasting in dynamic indoor environments, such as warehouses and factories, where understanding and predicting human motion is essential for operational safety and efficiency. By enabling systems to adapt to unpredictable movements, this task lays the foundation for safer Human Robot Collaboration. Finally, this thesis investigates Human Pose Forecasting (HPF), a critical element in enabling seamless interaction between humans and robots. In

scenarios such as assembly lines or shared workspaces, accurate predictions of human movement in real-time allow robots to respond proactively, ensuring safety and enhancing cooperative workflows.

Lastly, this thesis highlights the potential of Deep Learning in aligning technological advances with human-centered principles, offering a roadmap for industries to embrace the transformative vision of Industry 5.0-where Artificial Intelligence and humans work together to redefine the future of work and manufacturing.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 The Deep Learning Role in the Industrial Landscape

The rapid advancement of Artificial Intelligence (AI) is profoundly reshaping industrial realities, giving rise to a new era of automation, intelligence, and connectivity. Central to this transformation is the emergence of Deep Learning (DL), a subset of AI that enables machines to analyze large amounts of complex data, learn intricate patterns, and make accurate predictions. While traditional industrial systems have largely relied on deterministic and rule-based approaches [1], the growing demand for flexibility, adaptability, and efficiency has necessitated a shift to dynamic and intelligent systems.

DL has already made significant inroads in industrial environments, particularly in the automation of simple, repetitive tasks. An important example is quality control, where Convolutional Neural Networks (CNNs) are used to detect product defects [2]. These systems excel at image recognition and detection, identifying inconsistencies such as cracks, warping, or discoloration in manufactured products. By automating these tasks, DL not only improves accuracy and consistency but also frees human workers to focus on more complex responsibilities. Another DL solution widely adopted in factories involves process optimization through robotic automation systems. For example, DL-driven robotic arms are increasingly used for assembly tasks, leveraging machine vision to identify components, accurately position them, and assemble products with remarkable precision [3]. In this sense, all these innovations relate to a fourth-generation industry, where various production

processes are being automated not only for performance issues, but to marginalize the need for human intervention as much as possible.

However, when the goal is to maximize production, more and more critical issues have emerged in recent years regarding the sustainability and ethical aspects of this industry concept [4], making room for a further refined fifth-generation of industry, where machines do not replace humans but become a useful collaborative and informative tool to achieve a more responsible, ethical and sustainable level of production [5].

For this reason, Industry 5.0 represents a paradigm shift where Human Robot Collaboration (HRC) becomes the cornerstone of technological progress [6]. Collaboration between humans and machines, however, should not be restricted only at a production level: the main goal is to use machines as informative tools, especially for making decisions about product planning and management, ensuring efficient production, improving planning processes, reducing waste, and enabling safe human-robot collaboration.

Time-Series Forecasting addresses these challenges by enabling predictive insights that shift industrial processes from reactive to proactive. By analyzing historical data and identifying patterns, forecasting models can anticipate future demands, optimize supply chains, and enhance production scheduling [7]. In human-robot collaboration, trajectory and behavior forecasting ensure safer and more adaptive interactions, reducing risks and improving workflow efficiency. These capabilities make Time-Series Forecasting an essential tool in shaping Industry 5.0, where intelligent systems support human workers in creating more efficient, sustainable, and safe industrial environments.

Definition. A Time-Series is a sequence of data points recorded over time, typically at regular intervals, such as hourly temperature readings, daily stock prices, or monthly sales figures. For this reason, Time-Series data is inherently temporal, meaning that the order of observations carries essential information about trends, patterns, and dependencies over time.

Time-Series Forecasting is the task of predicting future values based on past observations. Forecasting methods range from statistical approaches, like autoregressive models and ARIMA, to advanced DL techniques that capture complex dependencies [8, 9]. Accurate Time-Series Forecasting is crucial in various appli-

cations, including demand planning, predictive maintenance, and human behavior modeling, enabling industries to make data-driven decisions and optimize processes.

Improving Industry through Time-Series Forecasting. Many modern factories already generate large amounts of data from sensors, machines, and order management systems. These data sets often remain unused or untapped, representing an important resource of useful information [10]. Time-Series Forecasting models offer a powerful means of analyzing these data, uncovering patterns and trends that can inform decision-making. For example, predictive maintenance systems can process sensor data to predict equipment failures, enabling timely repairs and minimizing downtime [11, 12, 13, 14]. Similarly, by analyzing production and inventory data, these models can optimize schedules, align production with demand, and reduce material waste. Time-Series Forecasting models can also improve supply chain efficiency by predicting delays, managing inventory levels more effectively, and reducing costs. In addition, this knowledge can help companies adapt to market trends, improve customer satisfaction, and achieve a higher level of operational intelligence [15].

In the context of safety, Time-Series Forecasting can play a key role. For example, predicting human trajectories in shared workspaces can help prevent accidents by enabling real-time adjustment of robot movements [16, 17]. Similarly, predicting environmental conditions, such as temperature fluctuations or air quality, can inform decisions that protect workers' health and safety. This thesis explores how forecasting systems, powered by deep learning, can address some of these challenges and contribute to safer, more efficient, and more sustainable industrial practices.

1.2 Time-Series Forecasting for Planning, Adaptation and Interaction

This thesis addresses practical challenges in diverse applicative scenarios, focusing on three interconnected themes: planning, adaptation, and interaction. Each theme represents a critical step in leveraging predictive systems to address real-world problems. A visual overview can be seen in Figure 1.1.

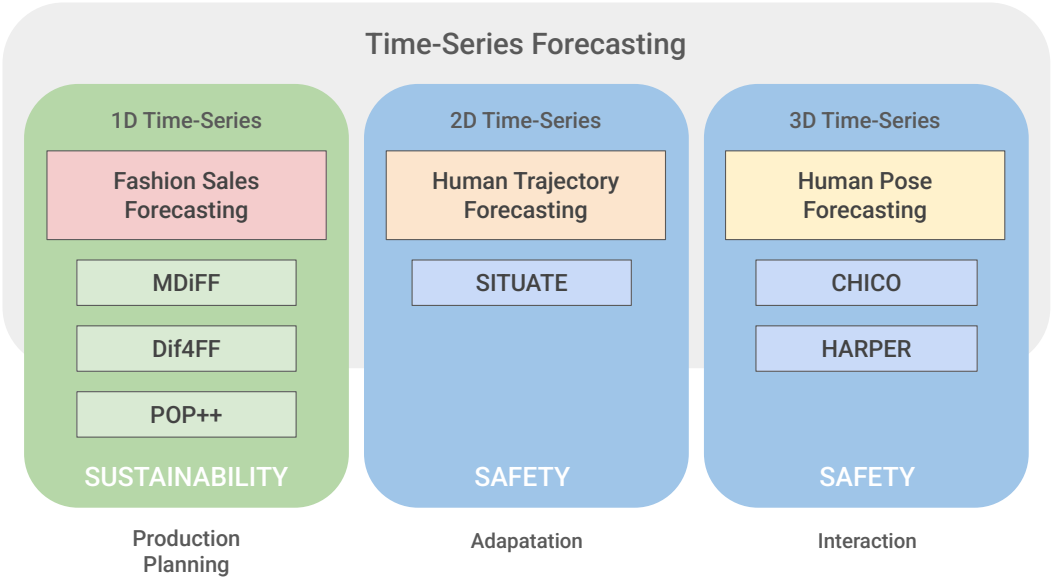

Fig. 1.1 An illustrated overview of this thesis, organized into three macro-areas. From a methodological point of view, this thesis explores how different architectures can scale and adapt to different problems, explained in Section 1.3. In the Figure, the contributions of this thesis are shown, together with the related practical scenario studied. All the three main topics and tasks explored in this thesis are described in Section 1.2

1.2.1 Planning: New Fashion Product Performance Forecasting

One-dimensional forecasting typically revolves around single-variable time models, such as forecasting future sales or demand based on historical data. These problems are fundamental and critical to decision-making in many industries, but they often involve relatively simple linear relationships over time. Planning places the foundation of industrial optimization, enabling businesses to allocate resources efficiently, reduce waste, and meet market demand. In the context of this thesis, planning is exemplified by the task of New Fashion Product Performance Forecasting (NFPPF). This task involves predicting consumer preferences for garments with no historical sales data, a problem complicated by trends, seasonality, and the multimodal nature of the data [18, 19, 20, 21].

The contributions in this area leverage advanced generative AI techniques, such as diffusion models, to address the inherent uncertainty of the task. The approach captures complex dependencies that drive consumer behavior by conditioning these models on multimodal information, such as product attributes, visual features, and market trends. Additionally, the integration of graph-based methodologies, such

as Spatial-Temporal Graph Convolutional Networks (GCNs) [22, 23], enhances the interpretability and refinement of predictions. These innovations highlight the potential of forecasting systems to support sustainable production practices, aligning supply with demand while minimizing overproduction and environmental impact.

1.2.2 Adaptation: Indoor Human Trajectory Forecasting

Adaptation represents the next stage in the progression of forecasting systems, where models must adjust dynamically to changing conditions and environments [24]. This concept is embodied in the task of indoor human trajectory forecasting, a critical challenge in shared spaces such as warehouses, factories, and public environments. Predicting the paths of individuals in these settings requires models to understand spatial relationships, environmental constraints, and the interactions between agents [25, 26].

The research in this domain introduces a novel approach using Equivariant and Invariant GCNs [27], augmented with a visual representation of the environment. By incorporating environmental context into trajectory predictions, the models provide a more holistic understanding of movement patterns, enabling adaptive responses to dynamic scenarios. For example, these systems can support autonomous robots in avoiding collisions with humans or optimizing navigation paths in real-time, demonstrating the transformative potential of adaptive forecasting in enhancing safety and operational efficiency.

1.2.3 Interaction: Human Pose Forecasting

The third scenario explored in this thesis is Human Pose Forecasting (HPF). Standard collision avoidance systems are usually based on proxemics and distance heuristics, systems that do not allow for any fine-grained collaboration between humans and robots in close HRC environments, where anticipating human movements can prevent accidents and ensure seamless cooperation.

This domain requires the integration of spatiotemporal dynamics to model how human joints move and interact in real-time environments. These models are expected not only to predict human behavior accurately but also to anticipate and respond to it dynamically [16, 28].

This thesis contributes to this field by introducing novel datasets, such as CHICO and HARPER, designed specifically for human pose forecasting in safety-critical and

collaborative environments. These datasets capture diverse interactions from strict HRC scenarios where humans and robots must cooperate in solving a specific task, enabling robust model training and evaluation. The methodological contribution includes the development of advanced GCN architectures [29] tailored to capture the intricate dependencies between human joints and their interactions with robotic systems. By enabling robots to predict and adapt to human movements, these models represent a significant step toward intelligent, interactive systems that prioritize safety and collaboration.

1.3 Scaling Complexity in Time-Series Forecasting

The field of Time-Series Forecasting has long been a cornerstone of industrial optimization, but its evolution has reflected the increasing complexity and dynamism of the environments in which it is applied. Traditional Time-Series Forecasting methods, while effective in stable and predictable environments, often fail to capture the complexities of dynamic industrial systems. Such methods typically rely on historical data to extrapolate future trends, offering little flexibility in the face of changing conditions or novel scenarios. In contrast, DL models excel in handling diverse and complex datasets, enabling predictions that adapt to real-time changes and account for contextual factors.

1.3.1 From 1D to 3D Time-Series Analysis

From a methodological point of view, this thesis reflects on the progression of Time-Series Forecasting techniques, presenting a unifying narrative that links research contributions spanning the one-dimensional (1D), two-dimensional (2D), and three-dimensional (3D) forecasting domains. The progression from one-dimensional (1D) to three-dimensional (3D) prediction reflects an important step in addressing increasing complexity, not only due to the growing interdependencies among elements as the feature space expands but also in terms of computational demands.

As discussed in Section 1.2, each forecasting task presents distinct requirements regarding computational efficiency, inference time, and memory consumption. For instance, in NFPPF, inference time is not a critical factor, as it does not affect the model's practical utility. On the other hand, in HPF, inference time must remain a fraction of the prediction horizon to ensure real-time applicability. While advances in

methodology and the increasing availability of computational resources have made it feasible to run more complex models, careful architectural design remains essential to balance performance and efficiency.

Throughout this thesis, GCN models have been extensively employed across all tasks due to their adaptability, lower memory footprint, and efficient inference time, making them particularly well-suited for the challenges posed by high-dimensional forecasting problems.

1.3.2 Time-Series Forecasting Through Common Principles and Modern Tools

Despite the distinct requirements of each domain, the basic principles of forecasting remain consistent: identifying patterns across different time frames, modeling dependencies, and making accurate predictions. By applying similar fundamental techniques in different contexts, this work emphasizes the universality of Time-Series Forecasting methodologies. At the same time, it shows the need for customization to address domain-specific requirements, such as incorporating multimodal data in sales forecasting or using spatial embeddings in trajectory forecasting. This versatility not only broadens the impact of forecasting but also reveals new opportunities for interdisciplinary innovation.

Graph Convolutional Networks. GCNs are a class of neural networks designed to process data structured as graphs. Unlike traditional neural networks that operate on regular grid-like data (e.g., images or sequences), GCNs can model complex relationships between entities by leveraging graph structures. In a graph, nodes represent entities (such as sensors, people, or locations), while edges define connections between them, capturing spatial or relational dependencies.

GCNs extend the concept of convolution from grid-based data to graph domains by aggregating and propagating information from neighboring nodes, during the so-called message-passing operation. This allows them to learn representations that incorporate both local and global graph structures. GCNs [30, 31] are widely used in applications where understanding relationships between entities is essential for accurate prediction, *e.g.* encoding spatial relationships between body joints for HPF [29], mapping interactions between agents in trajectory prediction [27], or capturing multimodal dependencies in sales forecasting [32], GCNs provide a

flexible and powerful framework. Their ability to generalize to different tasks makes them an indispensable component of the contributions of this thesis.

Denoising Diffusion Probabilistic Models. This research explores also the use of generative AI, such as Denoising Diffusion Probabilistic Models (DDPMs) [33], to improve traditional Time-Series Forecasting approaches. DDPMs are a class of generative models that learn to create high-quality data samples by progressively removing Gaussian noise from a noisy input. Inspired by diffusion processes in physics, DDPMs work by gradually corrupting data with noise in a forward process and then learning to reverse this process to reconstruct the original data.

Unlike traditional generative models, such as Generative Adversarial Networks (GANs) or Variational AutoEncoders (VAEs), DDPMs generate samples step by step, allowing for high-quality and diverse outputs. Their iterative refinement process makes them particularly effective in applications like image synthesis and data augmentation. Diffusion models have been particularly effective in addressing tasks such as Time-Series Forecasting [34, 4, 35] and NFPPF [22, 36], where historical data may be sparse or nonexistent.

Together, the progression in dimensional scaling, the exploration of diverse contexts, and the application of common methodologies highlight a cohesive narrative: Time-Series Forecasting is not only a powerful analytical tool but also a critical enabler of intelligent, adaptive systems. By scaling forecasting techniques across dimensions and domains, this thesis provides a comprehensive foundation for advancing the field, with applications that extend from production planning to dynamic and real-time interaction.

1.4 Contributions

This section will present the main contributions of the thesis, divided into the three different use cases presented in the previous section, namely ***Prediction for Planning***, ***Prediction for Adaptation***, and ***Prediction for Interaction***. As reported in Section 1.3.1, these scenarios and use cases are the practical application of different methodologies related to the forecasting of time series with different dimensionalities.

Prediction for Planning

- **MDiFF [36]:** Accurate forecasting of sales volumes for new, unreleased products is crucial for improving efficiency and reducing waste. However, this task is particularly difficult due to the lack of historical data and rapidly changing trends. Traditional deterministic models often struggle with domain shifts when dealing with products outside their training data distribution. To address these challenges, this method proposes a novel approach using a two-step multimodal pipeline based on diffusion models. The method first employs a score-based diffusion process to generate multiple future sales predictions for new fashion items. These predictions are then refined using a lightweight Multi-Layer Perceptron (MLP) to produce the final forecast. This combination leverages the adaptability of diffusion models and the efficiency of neural networks, resulting in an accurate and scalable forecasting system that advances sustainability and resource optimization in the fast fashion industry.

- **Dif4FF [22]:** An enhanced version of the MDiFF pipeline that incorporates more complex data structures and introduces an upgraded model, replacing the MLP with a GCN. This upgrade improves the system's adaptability for predicting diverse product lines by better capturing relationships within the data. The extension is designed to generalize the methodology, increasing its versatility and making it applicable to a wider range of scenarios within the fast fashion industry.

- **POP++ [37]:** This work introduces a data-centric approach to improve the forecasting of new fashion products' market performance before their release. Unlike traditional methods that struggle with the absence of historical data, this approach constructs a historical context by leveraging multimodal information. The pipeline starts by extracting textual tags from the garment's image to retrieve similar fashion items from web sources. These retrieved images, representing both successful and unsuccessful styles from the past year, undergo a series of filtering techniques—such as confident learning, pose filtering, and mask filtering—to remove noisy or irrelevant samples. The refined dataset is then used to generate an exogenous predictive signal, termed POtential Performance++ (POP++), which simulates how the item would have performed had it been available earlier.

Prediction for Adaptation

- **SITUATE [25]:** This work proposes a novel approach for indoor human trajectory prediction. Indoor motion patterns, characterized by self-loops and rapid direction changes, differ significantly from outdoor trajectories. SITUATE addresses these challenges by combining geometric learning, which captures intrinsic symmetries and human movement dynamics, with a self-supervised vision representation that extracts spatial-semantic information from the environment. Evaluated on the THÖR and Supermarket datasets, SITUATE achieves state-of-the-art results in indoor trajectory forecasting while also demonstrating strong generalization to outdoor scenarios. This highlights the robustness of indoor-oriented models in diverse applications.

Prediction for Interaction

- **CHICO Dataset and Ses-GCN [16]:** The CHICO dataset is a benchmark for studying HRC in industrial settings. It provides multi-view videos and 3D joint trajectories of human operators collaborating with a robotic arm in realistic assembly-line tasks. CHICO focuses on enabling cobots to forecast human motion and detect collisions, ensuring safer and more productive interactions in shared workspaces. This work introduces a novel model, SeS-GCN, for efficient and accurate human pose forecasting in industrial HRC. By leveraging depthwise-separable graph convolutions, space-time separable adjacency matrices, and sparse graph structures, the model achieves state-of-the-art performance with reduced computational complexity. Validated on the CHICO dataset, SeS-GCN enhances cobot awareness by predicting human movements and enabling proactive collision detection.

- **HARPER Dataset [28]:** is a dataset designed for studying human behavior in collaboration with a mobile quadruped robot, Boston Dynamics' Spot. It includes synchronized data from Spot's sensors and a high-precision motion capture system, capturing interactions like pose forecasting, partial-body pose estimation and collision prediction. HARPER emphasizes scenarios where the robot has a limited field of view, advancing research on mobile cobots in diverse environments.

1.5 Publications

The publications carried out during the author's PhD program are listed below in chronological order. Please note that some of these articles are not part of this thesis. Those included are highlighted in ***bold***.

- Sampieri, A., D'Amely di Melendugno, G. M., Avogaro, A., Cunico, F., Setti, F., Skenderi, G., Cristani, M., and Galasso, F. (2022).
  Pose Forecasting in Industrial Human-Robot Collaboration.
  In Computer Vision – ECCV 2022.

- Cunico, F., Girella, F., Avogaro, A., Emporio, M., Giachetti, A., and Cristani, M. (2023).
  Oo-dmvmt: A deep multi-view multi-task classification framework for real-time 3d hand gesture classification and segmentation.
  In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop.

- Avogaro, A., Cunico, F., Rosenhahn, B., and Setti, F. (2023).
  Markerless human pose estimation for biomedical applications: a survey.
  In Frontiers in Computer Science.

- Emporio, M., Caputo, A., Pintani, D., Cunico, F., Girella, F., Avogaro, A., Cristani, M., and Giachetti, A. (2023).
  Gesture Based Interaction with the Hololens 2.
  In Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter

- Avogaro, A., Capogrosso, L., Fummi, F., and Cristani, M. (2024).
  MDiFF: Exploiting Multimodal Score-based Diffusion Models for New Fashion Product Performance Forecasting.
  In Computer Vision – ECCV 2024 Workshops.

- Avogaro, A., Toaiari, A., Cunico, F., Xu, X., Dafas, H., Vinciarelli, A., Li, E., and Cristani, M. (2024).
  Exploring 3D Human Pose Estimation and Forecasting from the Robot's Perspective: The HARPER Dataset.
  In 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

- Capogrosso, L., Toaiari, A., Avogaro, A., Khan, U., Jivoji, A., Fummi, F., and Cristani, M. (2024).
  SITUATE: Indoor Human Trajectory Prediction through Geometric Features and Self-Supervised Vision Representation.
  In International Conference on Pattern Recognition.

- Avogaro, A., Capogrosso, L., Fummi, F., and Cristani, M. (2024).
  Dif4FF: Leveraging Multimodal Diffusion Models and Graph Neural Networks for Accurate New Fashion Product Performance Forecasting.
  In International Conference on Pattern Recognition.

- Cunico, F., Aldegheri, S., Avogaro, A., Boldo, M., ... (2024).
  Enhancing Safety and Privacy in Industry 4.0: The ICE Laboratory Case Study.
  In IEEE Access.

- Avogaro, A., Capogrosso, L., Toaiari, A., Fummi, F., and Cristani, M. (2025).
  New Fashion Products Performance Forecasting: A Survey on Evolution, Models, and Emerging Trends.
  In Springer Nature on Computer Science.

- Avogaro, A., Girella, F., Capogrosso, L., Fummi, F., Skenderi, G., and Cristani, M. (2025).
  POP++: Refining POtential Performance of New Fashion Products through Human Proxemics
  In IEEE Access **[Under Revision]**.

Chapter 2

Sustainable New Fashion Product Performance Forecasting

The fast fashion industry represents the second most polluting industry in the world, responsible for 79 trillion liters of water consumed and 92 million tonnes of waste produced per year [38], contributing 8% of all carbon emissions and 20% of all global wastewater [39]. Being able to predict sales volumes for an unreleased product more precisely could represent a significant step toward making this market more efficient, reducing the use of resources for production, and especially minimizing leftover unsold inventory [18]. Although forecasting time series with a known historical past has been extensively analyzed [40, 8], very little attention has been paid to a much more practical and challenging scenario: forecasting new products that the market has not seen before.

Specifically, this problem, known as New Fashion Product Performance Forecasting (NFPPF) [20], is far from trivial, as unreleased products do not have sales data available. Thus, it is necessary to retrieve valuable information from the available data, which may be technical specifications of the product (such as color, type, material), release period, or interest shown for similar products in the past [20]. An example of the task is shown in Figure 2.1.

Due to the rapidly changing nature of fashion trends, determining what is considered fashionable or outdated can be challenging. This makes it difficult to accurately predict the market performance of a specific item and identify the key factors that influence its popularity. Traditional deterministic forecasting models have shown reasonable performance in specific situations. However, they are limited by their assumption that the characteristics of past-season products are directly applicable to

new items, which often exhibit distinct features. This leads to inaccurate predictions due to the shift in the characteristics of the input data.

Recently, Denoising Diffusion Probabilistic Models (DDPMs) [33], or Diffusion Models in general, have impressed with their realistic image and video generation capabilities. Moreover, DDPMs have shown promising results even in the context of time series analysis [41]. Specifically, DDPMs implicitly learn the probability distributions of data, such as images [42, 43], or fashion sales, as we demonstrate in this research. In applications like image generation [44, 45], DDPMs generative process does not start from the input but with Gaussian noise, and it's iteratively denoised to create an output that conforms to the learned data distribution. An input, even a complete outlier, only acts as a conditioning signal to guide this denoising process, not as a direct source. The model's core training forces it to produce a sample that lies on its learned manifold, effectively pulling the final result back towards an "in-distribution" appearance. This inherent constraint means that even when guided by anomalous data, the output remains plausible and resistant to being distorted by the outlier [46].

In contrast, deterministic approaches are less reliable because they directly link input features to predicted sales, which can result in incorrect predictions for new feature combinations. This characteristic is critically important, making DDPMs an ideal tool for NFPPF.

MDiFF [36] (Section 2.2). In this chapter, we present **MDiFF**, a two-stage architecture specifically crafted to tackle NFPPF. We first train a multimodal score-based diffusion model that learns to generate samples from the true sales distribution. A second refinement model, based on an Multi-Layer Perceptron (MLP), is then used to refine the prediction. DDPMs, by nature, generates output from Gaussian noise, guaranteeing a fundamental property, which is the generation of non-deterministic samples. In order to better control this behavior and have more stable results, what we do is generate multiple sales signals for the same object. Specifically, 50 samples are generated and given as input to the refinement model, which subsequently generates the final sales prediction. This strategy ensures that MLP receives data in the distribution that matches more closely the actual sales data distribution, leading to improved pipeline reliability.

Fig. 2.1 A practical overview of the importance of NFPPF. In the left part of the image, one can see how, for classic or "evergreen" garments, there is the possibility of using certain historical data to apply forecasting algorithms, analyzing features such as trend and seasonality. Instead, in the second case on the right, NFPPF applies, where exogenous data must be used for products that have never been released before.

Dif4FF [22] (Section 2.3). The previous work was later extended in **Dif4FF**, in which we conditioned the multimodal score-based diffusion model on more data, and the refining module was replaced by a Graph Convolutional Network (GCN). Specifically, our GCN builds two types of graphs from the input. One graph focuses on the time dimension, highlighting important connections among weak sales. Then, we create another graph based on prediction space, pinpointing strong connections among model-predicted samples. Finally, we use three Conv1D layers to compress the graph network's output and generate the prediction vector. This allowed us to process the expanded multimodal input more effectively.

Visual Search-Based Models in NFPPF. Another emerging field of research in the literature related to NFPPF is to use external knowledge to allow models to better understand the concept of *popularity*, key factor and very valuable information for any multimodal NFPPF model. Several approaches have been proposed in the literature to automate the creation of high-quality training data that can be used to improve any forecasting model that accommodates multivariate time-series forecasting [47, 20]. Of particular importance in this research field is the work of [47], where the authors propose a fully automatic pipeline that generates a signal related to a product's popularity, dubbed POP. The pipeline starts by extracting

textual tags related to the visual attributes of the probe, automatically or directly (*e.g.*, by an available technical sheet). The set of tags is expanded with *positive* and *negative* terms that are used to perform a *time-dependent* query online, *i.e.*, collecting images of "fashionable" and "unfashionable" items related to the tags, which have been uploaded during some specified K_{past} intervals in the past. These images are used to *confidently learn* [48] a binary classifier that captures what is fashionable VS unfashionable in that interval. This learning procedure prunes noisy images from both the positive and negative classes, resulting in a robust database to build the exogenous popularity (POP) signal. The aforementioned signal is then generated as the average cosine similarity between the probe image and the fashionable images.

POP++ [22] (Section 2.4). Existing pipelines for generating exogenous signals like [47] rely on the knowledge contained on a large platform that can reflect fashion trends, like Google Images or Google Trends. These pipelines have shown remarkable results in terms of performance, extracting visual features from a large pool of "fashionable" images of garments. Despite the effectiveness of those Data-Centric pipelines, they often fail to filter out irrelevant elements from images retrieved on the web, such as backgrounds, overlapping objects, or unrelated garments, which can introduce noise and bias into the data. These issues arise from the lack of mechanisms to isolate the garment of interest, leading to less precise feature extraction and signal generation.

To address these limitations, we propose a novel pipeline, POP++, which incorporates Human Pose Estimation (HPE) and garment-specific segmentation. By leveraging pose information, our approach filters out irrelevant images and isolates the garment of interest, ensuring that the extracted signals are focused exclusively on the relevant features of the product. This refinement makes the pipeline more robust and capable of handling the variability of non-standardized datasets, such as those collected from the Web.

Our proposed pipeline represents a significant step toward creating cleaner, higher-quality data for forecasting applications, enabling more accurate and reliable insights into NFPPF architectures.

2.1 Related Works

The existing literature on NFPPF with deep learning tools is limited but growing. One of the first articles to investigate the sales forecasting problem is [49]. Following, [50] attempts to emulate an expert's decision-making process, training a ConvNet-based model to extract visual features from the image and then, through the k-Nearest Neighbors (k-NN) algorithm, confront the features with other elements already seen to produce the final prediction of the sales.

With [51, 52], there was a first attempt at tackling this task by exploring various algorithms. In particular, [51] compares several machine learning algorithms, such as gradient boosting and random forest, discussing which is better for NFPPF. Moreover, the authors also tried two deep learning approaches, *i.e.*, Feed-Forward Networks (FFNs) and Long Short-Term Memorys (LSTMs), both fed with multimodal signals. Those signals were obtained from static attributes of the items, such as category, color, and fabric, and variable information, such as discounts and promotions. Similarly, [52] uses an architecture based on Recurrent Neural Networks (RNNs), including more signals such as past sales, images, textual embeddings, and discounts. The model also operates a soft-attention mechanism to understand which information is more relevant to produce the predicted sales signal. However, their autoregressive model produced the same prediction between products of different seasons. Unfortunately, the authors' code and dataset are proprietary and not publicly available.

Building on the value of exogenous signals in fashion forecasting, [20] propose an encoder-decoder Transformer-based architecture incorporating as input of the model all the multimodal data offered by the dataset [19]. The encoder is fed with Google Trend signals, while the decoder receives an ensemble of features extracted from images, textual descriptions, and the item's temporal information (release date). This approach effectively extends previous work by using a more powerful architecture to extract insights from exogenous signals, showing the effectiveness of Google Trends information in relation to NFPPF.

In contrast to previous approaches, this research introduces the first diffusion model-based implementation for solving the NFPPF task. In this way, we solve the problem common to all the previous methods: unrealistic predictions due to the shift in the input feature domain.

DCAI [53] shifts the focus from models to the data used to train and evaluate them. It is a topic whose importance is growing continuously in many AI communities [54,

55], with important effects on Computer Vision (CV) & Machine Learning (ML). In general, DCAI investigates methodologies to accelerate open-source dataset creation from lower-quality resources. Consequently, it is tightly coupled with learning from noisy data, which aims to produce consistent low-noise data samples or to remove labeling noise and inconsistencies from existing data [56, 57, 48]. The application of such methodologies to the fashion industry has been limited, despite showing potential [47, 20]. In particular, the POP signal [47] was the first to propose the idea of generating a web-learned popularity signal for later use in forecasting.

Datasets for NFPPF. Publicly available datasets for fashion forecasting, such as [58], take into account applications that are different from NFPPF. They have usually been used to forecast fashion styles, which are aggregates of products of multiple brands in terms of popularity based on social networks, such as, for example, Instagram. In our case, the task is different since we focus only on single products and not on groups of products, so we have less data to reason on. In addition, we are considering genuine sales data and not popularity trends. As a result, in our research, we use the VISUELLE dataset [19, 20], the only dataset available in the literature for NFPPF, and thus the de facto standard for this task. Due to its nature, our research is also impactful from an industrial level.

2.2 MDiFF

In this section, we introduce **MDiFF**, a two-stage architecture shown in Figure 2.2 specifically crafted to tackle NFPPF.

Section Organization

In the following, we introduce the problem formalization (Section 2.2.1), our score-based diffusion models (Section 2.2.2), the methodology used to guide sample generation (Section 2.2.3), the MLP-based refinement stage (Section 2.2.4), and the experimental results (Section 2.2.5).

2.2.1 Problem Formalization

Given a new product j, we want to predict $y \in \mathbb{R}^W$ expressed as the performance vector in terms of sales in an interval of W weeks since its release date.

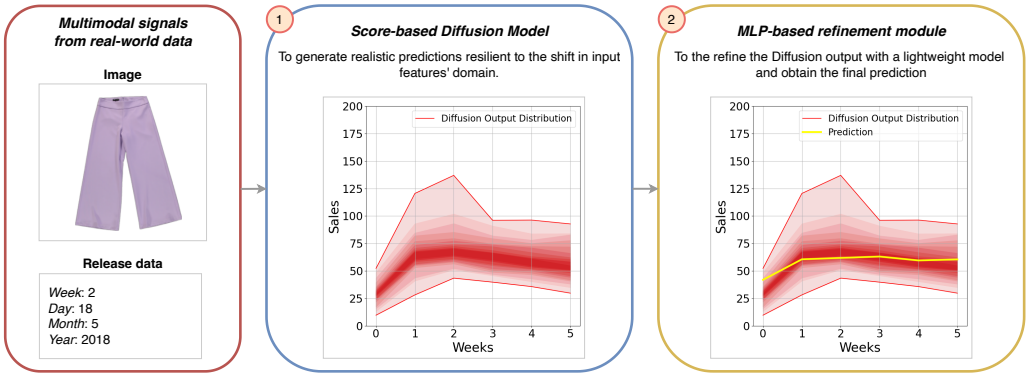

Fig. 2.2 MDiFF: a two-stage pipeline for NFPPF. Starting from multiple signals of a single fashion product, we build a multimodal score-based diffusion model to generate an initial prediction of the sales, addressing potential objects with features beyond the training distribution. Then, we refine the Diffusion output using a lightweight MLP to obtain the final prediction.

For every j, a set of 2 attributes is given: an image of the product $i_j \in \mathbb{R}^{w\times h\times 3}$ with $w = h = 256$ and the release date $t_j \in \mathbb{R}^4$ composed of four digits representing the day, week, month and year of release.

2.2.2 Our Score-Based Diffusion Model

Score-based diffusion models [59] generalize DDPMs [33] generative models trained to reverse a discrete-time diffusion process. A Gaussian noise diffusion process, also known as the *forward process*, can be summarized as a chain of steps in which Gaussian noise is progressively added to the initial distribution, as described by the following equations:

$$q(x^1,\ldots,x^T|x^0 = y) = \prod_{t=1}^{T} q(x^t|x^{t-1}) \,, \tag{2.1}$$

$$q(x^t|x^{t-1}) := \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t I) \,, \tag{2.2}$$

where $q(x^T) \approx \mathcal{N}(0,1)$, $y = q(x^0)$ is the true data distribution, β_t is the variance of the additive noise, and $t \in [0,T]$ represents the number of noising steps.

A model p_θ is then trained to reverse the diffusion process by gradually removing noise, also known as the *backward process*, to restore the initial distribution.

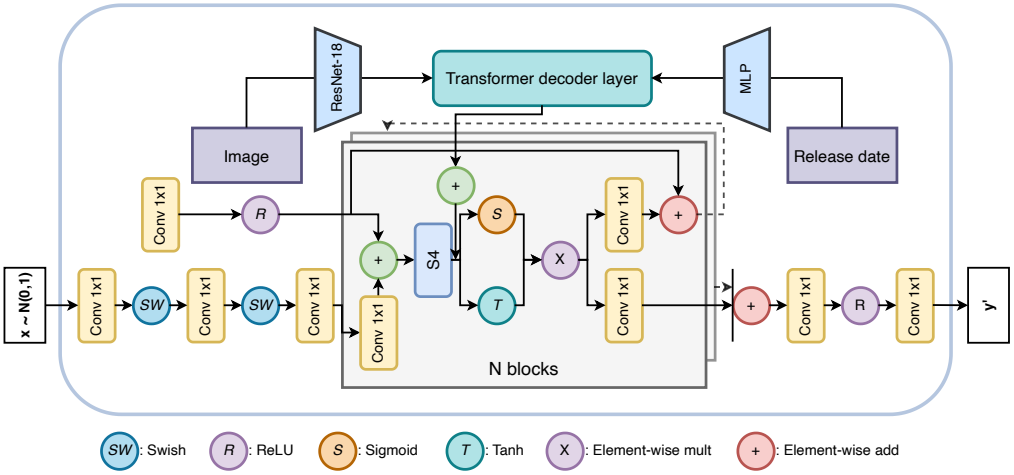

Fig. 2.3 An overview of our multimodal score-based diffusion model. The diffusion basic block is taken from TS-Diff [34] (grey square), modified to be injected with the output of the transformer decoder layer, a module responsible for producing an embedding representing the two modalities of input related to the item. Each block contains two outputs: one for the subsequent block and another for a skip connection. The summation of all skip connections forms the model's final output. The primary component of each block is typically an S4 block [60], chosen by the authors of [34] for its efficiency when it comes to time series and structured data. The input of the **MDiFF** is noisy data, and the output is the denoised sample.

Specifically, the backward process is formalized as follows:

$$p_\theta(x^{t-1}|x^t,c) = \mathcal{N}(x^{t-1};\mu_\theta(x^t,t) \\ +s\sigma_t^2\nabla_{x^t}\log p(x^t|x^0),\sigma_t^2 I)\ , \tag{2.3}$$

where σ is the variance for each timestep, and s is the parameter that controls the strength of the conditioning.

Specifically, Figure 2.3 shows the architecture of our multimodal score-based diffusion model. The network is a stack of multiple N blocks. Every block has two outputs, one for the next block and one for a skip connection. The summation of all the skip connections represents the actual output of the model. Every block is mainly made up of an S4 block [60]. The *multimodal conditioning* c_j is directly added to the output of the S4 block, and the cross-attention is implemented using a Transformer decoder layer [61].

2.2.3 Multimodal Conditioning

To guide the generation of the future sales of the j-th product by the diffusion model, we use multimodal data composed of images i_j of the product and release date t_j, as described in Section 2.2.1.

We train two different encoders I_θ, T_θ to extract features from the images and release dates respectively. Then, these features are used to produce the conditional embedding through a cross-attention mechanism defined as:

$$c_j = \text{Softmax}\left(\frac{Q_j K_j^T}{\sqrt{d_k}}\right) V_j \,, \tag{2.4}$$

where $K_j = V_j = I_\theta(i_j)$, $Q_j = T_\theta(t_j)$ and $\sqrt{d_k}$ is the dimensionality of the number of features of the embeddings.

We designed the conditioning module of **MDiFF** to weight, through the attention mechanism, the image embedding $I_\theta(i_j)$ with the features $T_\theta(t_j)$ obtained from the temporal information given by the release date. The idea behind this architectural choice came from the fact that every visual feature of the item has to be considered with respect to the fashionable concept of the current season to effectively guide the reverse diffusion process. To effectively and efficiently use the cross-attention mechanism, we used a Transformer decoder layer to serve this scope. More details on the implementation of the various encoders are reported in Section 2.2.5.

2.2.4 MLP-Based Diffusion Outputs Refinement

We approach the refinement stage as a regression task to predict a continuous output value. The model is designed to reduce the dimensionality of all 50 predictions of the diffusion model, analyzing both the feature and the temporal dimensions. Given $x \in \mathbb{R}^{WxN}$ the set of predictions of the diffusion model, y the true sales signal and $\hat{y}$ the output of the model with $y, \hat{y} \in \mathbb{R}^{1xW}$, the model is defined as follows:

$$\begin{aligned} x_t &= \phi_t(x) \,, \\ \hat{y} &= \phi_n(x_t') \,, \end{aligned} \tag{2.5}$$

where $N = 50$ is the number of generated samples of the diffusion model, $W = 6$ is the number of weeks of prediction, $\phi_t(x)$ the temporal MLP defined as $\phi_t(x) = W_t x + B_t$ and $\phi_n(x) = W_n x + B_n$ the MLP that regress the final sales signal.

Specifically, for $\phi_t(x)$, $W_t \in \mathbb{R}^{WxW}$ and $B_t \in \mathbb{R}^{WxN}$; on the other hand, for $\phi_n(x)$ we have $W_n \in \mathbb{R}^{1xN}$ and $B_n \in \mathbb{R}^{1xW}$. In this case, the dimensionality W is the number of weeks to predict, while W_t, W_n are the weight matrices of the two MLPs. Specifically, we utilized an MLP network trained with a Mean Squared Error (MSE) loss function denoted as $\mathscr{L}$MSE. The mean squared error loss measures the dissimilarity between the predicted output and the ground truth. Mathematically, it is defined as:

$$\mathscr{L}_{\text{MSE}}(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \,, \tag{2.6}$$

In detail, y_i denotes the true output value for sample i, while $\hat{y}_i$ signifies the predicted output value for sample i. More details about the structure of the refinement module are explained in Section 2.2.5.

2.2.5 Experimental Results

This section describes the experimental trials that have been carried out to validate our claims, along with their implementation details and results.

Implementation Details. The backbone of our model is a TS-Diff [34] architecture, a multi-purpose score-based diffusion model developed for predicting, reconstructing, and refining *univariate* time series data. TS-Diff consists of a series of $M = 4$ S4 Blocks [60] layers, connected via skip connections. We have extended the model to be conditioned as described in Section 2.2.3, using a Transformer decoder layer [61] to implement the multi-head cross-attention used to ensembling the two embeddings.
Image Encoder. We used as Image Encoder I_θ a ResNet-18 [62] pre-trained on ImageNet-1K [63]. We substituted the last two layers of the model with a Conv1D and a Linear to reduce the dimensionality of the features extracted, obtaining a tensor $I_\theta(i_j) \in \mathbb{R}^{C\times W}$, with C channels equal to 64 and W forecasting horizon of six weeks.
Temporal Encoder. The Temporal Encoder T_θ comprises four different MLPs that expand the dimension from 1 to C. The output of this model is then concatenated along the channel dimension and fed into another MLP that reduces the feature number from $4C$ to C, resulting in $T_\theta(t_j) \in \mathbb{R}^C$.
MLP-Based Refinement. The refinement network is based on two MLPs working on different dimensions of the input tensor. The first part is a stack of five Linear layers working on the temporal dimension of the input (*i.e.*, the six weeks of prediction),

expanding and compressing the feature space to match the same dimensionality of the input. The second part comprises three other linear layers, operating on the sample's dimensionality. These layers gradually reduce and compress the $N = 50$ predictions to achieve the actual forecasting. We used the ReLU as an activation function, with a skip connection between the input tensor and the output of the first MLP.

Dataset Description. We used the VISUELLE fast-fashion dataset [19, 20] to test our proposal. The dataset provides a comprehensive collection of fashion products and consumer behavior data. It encompasses three primary components: product information, customer data, and market trends.

Product information includes detailed descriptions of individual items. This involves visual representations in the form of high-resolution images showcasing the product on a plain background. In addition, textual attributes such as product category, color, fabric, and release date are provided.

Customer data offers meaningful insights into consumer preferences and purchasing habits. It contains anonymized information about a large number of customers, including their purchase history, specific items purchased, purchase dates, and the stores where purchases were made.

Finally, market trend data is incorporated into the Google Trends time series. This information tracks the popularity of product attributes, such as color, category, and fabric, over time, providing valuable information on consumer interest and demand fluctuations.

Specifically, the VISUELLE dataset provides purchase information from 667K users, containing data on 5,577 products exposed in 100 stores of Nunalie, an Italian fast-fashion company. The dataset contains 5,080 samples for the training set and 497 samples for the testing set, and unfortunately does not serve as a suitable validation set to evaluate during the training procedure model. Following the protocol given in [19, 20], we then used the test set as the validation set during the training of the model. Therefore, at the end of every training epoch, the model has been evaluated directly on the test set. At the end of the training procedure, the best-performing model has been kept.

Since we aim to put ourselves in a more challenging scenario, we rely only on the image and the release date as multimodal conditioning for each product, skipping the Google Trends and description ground truth available. Specifically, the fact that

MDiFF is conditioned only on image and release date is definitely a pro since these data types are extremely easy to find, minimizing the need for annotations by object operators and much more applicable on a large scale automatically in the fast fashion market.

Paired with every item, the dataset gives the sum of the sales across all 100 shops of the specific product in the 12 weeks after the release date. Following the evaluation protocol of [47, 20], we only predict the first 6 values of the available interval.

Evaluation Metrics. The Mean Average Error (MAE) and Weighted Absolute Percentage Error (WAPE), *i.e.*, the two main metrics representing the quality of the forecasting, are used to evaluate **MDiFF**. Formally, they are defined as:

$$\text{MAE} = \frac{\sum_{t=0}^{T} |y_t - \hat{y}_t|}{T}, \tag{2.7}$$

$$\text{WAPE} = \frac{\sum_{t=0}^{T} |y_t - \hat{y}_t|}{\sum_{t=0}^{T} y_t}, \tag{2.8}$$

where y represents the actual values of the time series, $\hat{y}$ represents the forecasted values, and T represents the total number of observations in the time series. Within the scope of NFPPF, MAE and WAPE are two complementary metrics: while MAE offers an indispensable, direct measure of the average error magnitude, it is insufficient on its own for evaluating forecasts across garments with different sales volumes. To overcome this problem, WAPE has been introduced as a complementary metric because it provides a scale-independent and robust measure of forecast accuracy.

Training Details. All the code is implemented in PyTorch [64]. For the multimodal score-based diffusion model, we train the network for 500 epochs, with a learning rate of 1×10^{-3}, a weight-decay of 5×10^{-4}, using AdamW [65] as an optimizer, on a NVIDIA RTX 4090. On the other hand, for the MLP networks, a Bayesian algorithm was used to search for the best training hyperparameters of the refinement MLP network.

Table 2.1 Quantitative results of **MDiFF** expressed in terms of WAPE and MAE, described in Equation (2.8) and Equation (2.7), respectively. In **bold**, the best results. Underlined, the second best.

Model	**Image**	**Temporal Condition**	**Description**	**Google Trends**	**WAPE** ↓	**MAE** ↓
Attribute k-NN [52]			✓		59.8	32.7
Image k-NN [52]	✓				62.2	34
Attr + Image k-NN [52]	✓		✓		61.3	33.5
GBoosting [66]	✓	✓			64.1	35
GBoosting+G [66]	✓	✓		✓	63.5	34.7
Cat-MM-RNN [52]		✓	✓	✓	63.3	34
X-Att-RNN [52]	✓	✓	✓		59.5	32.3
GTM-Transformer [20]	✓	✓	✓	✓	55.2	30.2
MDiFF (ours)	✓	✓			**54.7**	**30.1**

MDiFF Quantitative Results

This section discusses the quantitative results obtained with **MDiFF**. As we can see from Table 2.1, **MDiFF** outperforms all other state-of-the-art methods without using the information from Google Trends and the textual description of the various samples. The reason behind this choice is mostly practical: Google Trends are signals quite complex to obtain. Following the procedure described in [19], the practical adaptation into a real industrial scenario would be difficult. Therefore, we just chose to use images and release date, two modalities that do not require any specific pipeline or supervision to be acquired. Results shown in Table 2.1 clearly show that the choice of using **MDiFF** can drastically improve results when it comes to prediction error, using less information compared to other methods like [20].

In particular, we report comparisons between eight other existing models. [20] is the most similar in terms of performance to **MDiFF**. The most noticeable improvement is in WAPE, which dropped from 55.2 to 54.7, with a slight improvement in MAE as well. These enhancements in performance offer significant benefits that might not be immediately apparent. Primarily, as already introduced in Section 2.2, a diffusion-based model is always preferable for this task since it should better maintain performance with out-of-distribution objects, ensuring greater stability in practice when used in real-world usage contexts. Secondly, given our specific architecture, we only need less information to achieve better results.

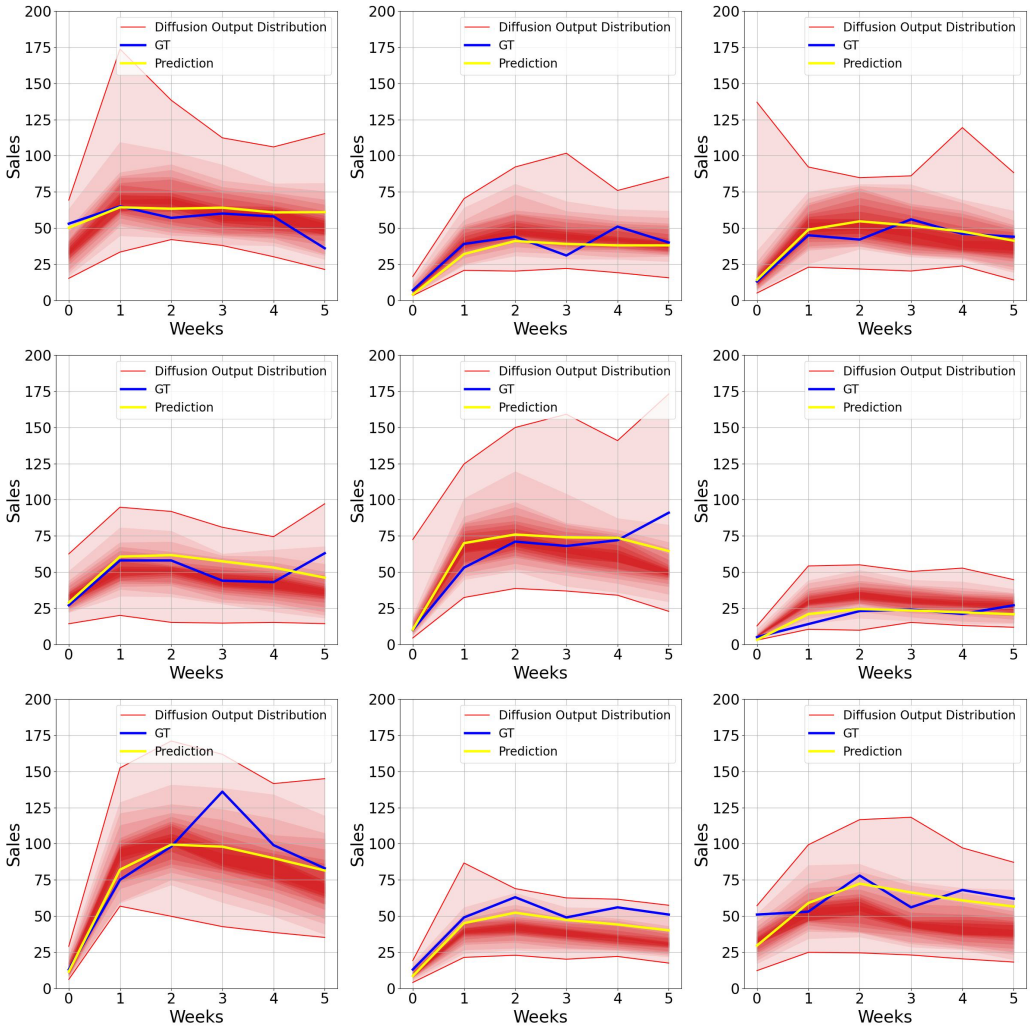

Fig. 2.4 In the figures above are presented some visual representations of the multimodal score-based diffusion model output. In particular, the red region represents the output distribution of the diffusion model given a certain sample. The red area is obtained by computing the weekly quantiles among the 50 outputs. The Prediction line, on the other hand, is the output of the refinement MLP, *i.e.*, the final prediction. The forecasting period is for 6 weeks from the date of release. The y-axis shows the number of units sold of a specific garment in the chain's various shops.

MDiFF Qualitative Results

In this section, we explain the role of the MLP in refining the multimodal score-based diffusion model output forecasting sales. Starting with the visualization of a few samples in Figure 2.4, it can be seen that the model itself gives as output a distribution of predictions that follows the ground truth very closely. Therefore, the role of the

Table 2.2 Table representing the different tests made with the same multimodal score-based diffusion model. We tested our model first without the temporal condition and then without images.

Model	WAPE ↓	MAE ↓
MDiFF (ours) without the temporal condition	56.4	31.1
MDiFF (ours) without the images	56.8	31.6
MDiFF (ours)	**54.7**	**30.1**

refinement MLP may seem obsolete, but there are several factors to consider, which instead make it crucial.

Firstly, the model output consists of N predictions; it is then necessary to use some technique to obtain a single final prediction. Examples might be simply taking the prediction mean or median. However, this would result in performance degradation, as the ground truth often differs from the distribution's mean or median.

As a result, we implemented a lightweight MLP network trained to refine the diffusion model output, as described in Section 2.2.5. As we can see from Figure 2.4, specifically in the second and final image, the ground truth does not reside in the densest region of the distribution, and the refinement network very effectively follows its movement away from the median of the diffusion output.

In order to understand how much the refinement model actually helps to improve the performance (and not just predict the mean/median) of the diffusion distribution output, we conducted additional ablative studies that show that the use of the refinement module is a winning strategy, improving performance considerably without excessively increasing the model complexity. We didn't explore how performance might improve with more complex models since our simple MLP already yielded good results. We plan to delve deeper into this matter and conduct a more thorough analysis in the future.

Ablation Studies

We conducted ablative studies to test how well the model performs using different conditioning setups. This helps us understand which configuration best achieves optimal performance with the diffusion model. It should be noted that the error values of each test done were obtained by running the entire **MDiFF** pipeline and not just the diffusion model. The results are reported in Table 2.2.

Looking at the results, it is clear that conditioning the model with just the images is insufficient. This is because the features of the dress without information on the season and period in which it is sold is insufficient to predict an accurate sales value. Indeed, it is not difficult to think that, since the fashion market is a sector strongly influenced by trends, a certain garment may be very fashionable in one season but remain completely unsold in the next.

On the other hand, it is quite straightforward to understand why just the temporal information without any further detail on the item's color, fabric, or shape is insufficient to determine an accurate prediction of the sales.

It is important to note that for other models such as [20], the importance of the various multimodal data types may differ. In **MDiFF**, unlike [20], conditioning is not processed directly to obtain a prediction but is only used to guide the process of reverse diffusion. The impact that one type of conditioning can have on different architectures is, therefore, very different from another.

2.3 Dif4FF

Section Organization

This section first discusses how we guide the generation process using multimodal information (Section 2.3.1). Then, we explore the proposed neural network-based graph refinement (Section 2.3.2) and show the experimental results (Section 2.3.3).

Instead, for a review of the theoretical background of multimodal score-based diffusion models, look at Section 2.2.2.

2.3.1 Our Improved Multimodal Conditioning

We leverage a combination of data sources to predict future sales for each product. This data will include images of the products (represented as i_j), the Google Trend signal g_j, and their release dates t_j.

Specifically, Figure 2.5 shows the architecture of our multimodal score-based diffusion model. The network is a stack of multiple N blocks. Every block has two outputs, one for the next block and one for skip connection. The summation of all the skip connections represents the actual output of the model. Every block is mainly composed of an S4 block [60]. With respect to MDiFF [36], our multimodal conditioning also includes Google Trend signals, so it is different in its composition.

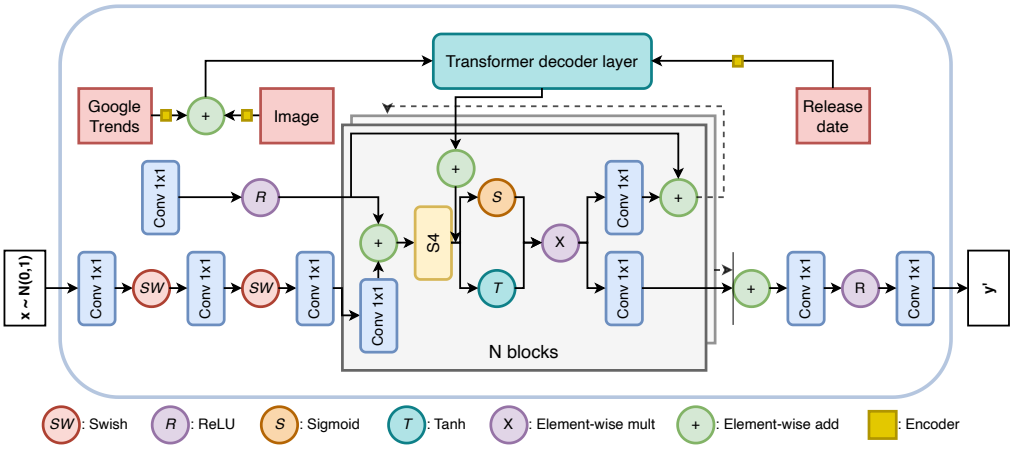

Fig. 2.5 An overview of our improved multimodal score-based diffusion model. Each block contains two outputs: one for the subsequent block and another for a skip connection. The summation of all skip connections forms the model's final output. The primary component of each block is typically an S4 block [60]. With respect to MDiFF [36], our multimodal conditioning also includes Google Trend signals, so it is different in its composition.

We train three different encoders I_θ, T_θ and G_θ to extract features from the input information. Then, these features are used to produce the conditional embedding through a cross-attention mechanism defined as:

$$c_j = \text{Softmax}\left(\frac{Q_j K_j^T}{\sqrt{d_k}}\right) V_j \,, \tag{2.9}$$

where $K_j = V_j = G_\theta(g_j) + T_\theta(i_j)$, $Q_j = I_\theta(i_j)$ and $\sqrt{d_k}$ is the dimensionality of the number of features of the embeddings.

The idea behind this choice is to start from the Google trend signal and use the release date embedding as positional encoding. Lastly, the Google Trend signal is weighted on the visual features (*i.e.*, shape, color, fabric, etc.) extracted from the image. The *multimodal conditioning* c_j is added directly to the output of the S4 block, and the cross-attention is implemented using a Transformer decoder layer [61].

2.3.2 GCN-based Diffusion Outputs Refinement

The refinement module comprises two main modules: the first is based on ST-GCN [23]. The GCN block operates on two different dimensions by constructing two graphs.

The first graph operates on the dimension of the space of the predictions made by the diffusion model. The mathematical formulation is the following:

$$X_s = \phi_s(A_s X) \,, \tag{2.10}$$

where $A_s \in \mathbb{R}^{S\times S}$ is the adjacency matrix, X output of the diffusion model and ϕ_s an MLP.

The second block works on the time dimension, learning a graph that weighs the connections between the various prediction weeks. Formally, it is defined as:

$$X_t = \phi_t(A_t X_s) \,, \tag{2.11}$$

where $A_t \in \mathbb{R}^{W\times W}$ is the adjacency matrix, X_t output of the first GCN block and ϕ_t an MLP.

These blocks are designed to compress the N predicted samples and regress the final prediction. The network is then trained with a Mean Squared Error (MAE) loss function denoted as $\mathscr{L}_{\text{MAE}}$, mathematically defined as:

$$\mathscr{L}_{\text{MAE}}(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|;, \tag{2.12}$$

where $\hat{y}$ represent the ground truth value and $\hat{y}_i$ is the value predicted by the model.

2.3.3 Experimental Results

This section describes the experimental trials that have been carried out to validate our claims, along with their implementation details and results. See Section 2.2.5 for dataset and evaluation metric details.

Implementation details. Our model is based on TS-Diff [34]. This architecture is a score-based diffusion model that is particularly useful for predicting, reconstructing, and refining time series tasks. TS-Diff uses a series of building blocks called "S4 Blocks" [60], stacked together four times with connections that allow information to flow directly (skip connections).

We've made some modifications to TS-Diff to incorporate the additional information from images and release dates. As described in the subsection on multimodal conditioning (Section 2.3.2), we've added a layer inspired by Transformers [61].

This layer helps to combine the features extracted from the images i_j, the release dates t_j, and the Google Trends signals g_j to properly guide the diffusion model.

Image encoder. We used as Image Encoder I_θ a ResNet-18 [62] pre-trained on ImageNet-1K [63]. We substituted the last two layers of the model with a Conv1D and a Linear to reduce the dimensionality of the features extracted, obtaining a tensor $I_\theta(i_j) \in \mathbb{R}^{C\times W}$, with C channels equal to 64 and W forecasting horizon of six weeks.

Temporal encoder. The Temporal Encoder T_θ comprises four different MLPs that expand the dimension from 1 to C. The output of this model is then concatenated along the channel dimension and fed into another MLP that reduces the number of features from $4C$ to C, resulting in $T_\theta(t_j) \in \mathbb{R}^C$.

Google Trends encoder. Lastly, as G_θ, we adopted a Transformer Encoder layer [61] that performs a self-attention operation on Google Trends. The encoder layer also reduces the dimensionality of the encoding, resulting in $G_\theta(g_j) \in \mathbb{R}^{C\times W}$.

GCN-based refinement. As described in Section 2.3.3, the architecture used for the refinement network is based on ST-GCN [23]. The first GCN-based module comprises two ST-GCN blocks, with an expansion and subsequent reduction of the channels to create a first embedding. On the other hand, the second module is composed of 1D convolutions, reducing the sample dimension from 50 to 1 (to obtain an actual final prediction). Specifically, this compression is done by three layers of 1D convolutions with PReLUs as activation functions.

Training details. All the code is implemented in PyTorch [64]. For the multimodal Score-based diffusion model, we train the network for 500 epochs, with a learning rate of 1×10^{-3}, a weight-decay of 5×10^{-4}, using AdamW [65] as an optimizer, on a NVIDIA RTX 4090.

Given the nature of diffusion models, different seeds produce different predictions. As a result, we ensured that our score-based diffusion model was executed in the most deterministic setup possible by setting the seed value to 32 across all libraries used. A Bayesian algorithm was used for the GCN networks to search for the best training hyperparameters.

Dif4FF Quantitative Results

Here, we discuss the quantitative results obtained with **Dif4FF**. As we can see from Table 2.3, **Dif4FF** outperforms all other state-of-the-art methods. To verify the statistical significance of the results, we ran 10 instances of the **Dif4FF** pipeline. The

Table 2.3 Quantitative results of **Dif4FF** expressed in terms of WAPE and MAE on VISUELLE, described in Equation (2.8) and Equation (2.7), respectively. In **bold**, the best results. Underlined, the second best.

Method	IMAGE	RELEASE	DESCR.	GOOGLE T.	WAPE ↓	MAE ↓
Mean predictor					60.1	32.8
Median predictor					50.3	31.8
Attribute k-NN [52]			✓		59.8	32.7
Image k-NN [52]	✓				62.2	34.0
Attr+Image k-NN [52]	✓		✓		61.3	33.5
GBoosting [66]	✓	✓			64.1	35.0
GBoosting+G [66]	✓	✓		✓	63.5	34.7
Cat-MM-RNN [52]		✓	✓	✓	63.3	34.4
X-Att-RNN [52]	✓	✓	✓		59.5	32.3
GTM-Transformer [20]	✓	✓	✓	✓	55.2	30.2
MDiFF [36]	✓	✓			54.7	30.1
Dif4FF (ours)	✓	✓		✓	**54.6**	**30.0**

mean MAE was 30.2 with a variance of 0.04, and the mean WAPE was 54.8 with a variance of 0.08.

In **Dif4FF**, we do not use the information from the textual description of the various samples since these could worsen the model performance in our diffusion-based architecture. Furthermore, the diffusion model probably retrieves the color, fabric, and color information directly from the image features. Furthermore, an objective product description is notoriously difficult to obtain in a real-world scenario since the description of fashion garments involves a complex interplay of abstract concepts and stylistic elements, making objective descriptions for each garment in the dataset challenging [67]. For more information on that, see Section 2.3.3.

In particular, we report the comparisons among eight other existing models. [20] is the most similar in terms of performance to **Dif4FF**. The most noticeable improvement can be seen in WAPE, which has a sharp drop from 55.2 to 54.7, with a slight improvement in MAE as well. These enhancements in performance offer significant benefits that might not be immediately apparent. Primarily, as already introduced in Section 2.3, a diffusion-based model is always preferable for this task since it should better maintain performance with out-of-distribution objects, ensuring greater stability in practice when used in real-world contexts. Secondly, given our specific architecture, we need less information to achieve better results with respect to the current state-of-the-art.

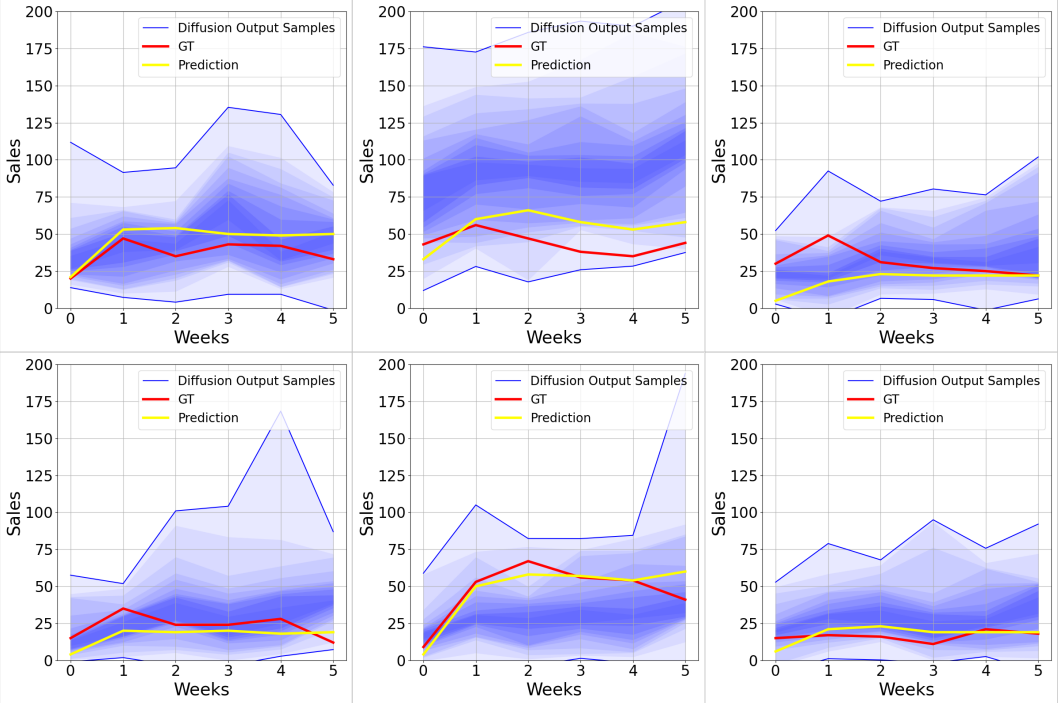

Fig. 2.6 In the figures above are presented some visual representations of the multimodal score-based diffusion model outputs. In particular, the blue region represents the output distribution of the diffusion model given a certain sample. Specifically, the blue area is obtained by computing the weekly quantiles among the 50 outputs. The Prediction line, on the other hand, is the output of the refinement GCN, *i.e.*, the final prediction. The forecasting period is six weeks from the release date, depicted on the x-axis. On the y-axis, the number of units sold of a specific garment in the various shops is shown.

Dif4FF Qualitative Results

In this section, we explain the role of the GCN in refining the multimodal score-based diffusion model output forecasting sales. Starting with the visualization of a few samples in Figure 2.6, it can be seen that the model itself gives as output a distribution of predictions that follows the ground truth very closely. Therefore, the role of the refinement GCN may seem obsolete, but there are several factors to consider that make it crucial.

Firstly, the model output consists of N predictions; then, it is necessary to use some technique to obtain a single final prediction. Examples might be simply taking the prediction mean or median. However, this would result in performance degradation, as the ground truth often differs from the distribution's mean or median.

Table 2.4 Table representing the different tests made with the same multimodal score-based diffusion model. We tested our model first without the temporal condition and then without images.

Model	WAPE ↓	MAE ↓
Dif4FF (ours) without the temporal condition	56.6	31.5
Dif4FF (ours) without the images	56.2	31.4
Dif4FF (ours) without the Google Trends	55.6	30.9
Dif4FF (ours)	**54.6**	**30.0**

As a result, we implemented a powerful GCN network trained to refine the diffusion model output, as described in Section 2.3.3. As we can see from Figure 2.6, specifically in the second and fifth images, the ground truth does not reside in the densest region of the distribution, and the refinement network very effectively follows its movement away from the median of the diffusion model outputs.

In order to understand how much the refinement model actually helps to improve the performance (and not just predict the mean/median) of the diffusion distribution output, the errors in terms of MAE and WAPE of both static measurements were calculated. Using the median to extract a final prediction, we obtained a result of 34.8 of MAE and 59.3 of WAPE. On the other hand, for the mean, an MAE error of 34.5 and WAPE of 58.7 were obtained. Thus, adopting a refinement network is a winning strategy that significantly improves performance without excessively increasing the complexity of the model.

Ablation Studies

The first ablation study tests the model performance using different conditioning setups. It should be noted that the error values of each test were obtained by running the entire **Dif4FF** pipeline. As we can see from Table 2.4 it is clear that images and temporal information are crucial for the model to predict sales accurately. This is because the item's features without information on the season and the release period are insufficient to predict an accurate sales value. Since the fashion market is a sector strongly influenced by trends, a certain garment may be very fashionable in one season but remain completely unsold in the next. However, it is quite straightforward to understand why temporal information and Google Trends, without any further details on the item's color, fabric, or shape, are insufficient to determine an accurate

Table 2.5 Table representing the results obtained in the domain-shift example. It is clear that our diffusion model is more resilient to the domain shift due to the different years it has been used compared to the second-best performing method.

Model	WAPE ↓	MAE ↓
GTM-Transformer [20]	56.9	32.8
Dif4FF (ours)	**55.9**	**31.4**

prediction of sales. Google Trends represents information on the public appreciation of a certain item, and the impact on performance is lower when removed.

The second ablation study is related to the domain shift, a well-known issue in the fashion domain. To check the resiliency of the models to this phenomenon, we trained both GTM-Transformer and **Dif4FF**, removing from the train set every garment related to 2018, leaving all of them just in the test set. The VISUELLE test set is already composed of only 2018 garments. As shown in Table 2.5, our method significantly reduces the error when tested on garments completely outside the training distribution. This highlights that while existing methods may perform adequately on training data, their real-world performance suffers significantly under domain change. As a result, **Dif4FF** becomes a killer application for real-world scenarios where domain shift is a common challenge.

Finally, since it is well recognized that a common issue with diffusion models is their tendency to converge on the mean of the data distribution, we investigated the performance of **Dif4FF** compared to a simple predictor that uses only the mean and variance of the training data for the test set. The results are reported in Table 2.3 and reveal that based on the mean and variance, this naive predictor significantly underperforms compared to our **Dif4FF**. This finding reinforces the effectiveness of our proposed approach for the NFPPF task.

2.4 POP++

In this Section, we introduce **POP++**, a Data-Centric pipeline to retrieve a **POP**ularity embedding to enrich the information available to models related to NFPPF.

Section Organization

In the following, we introduce the problem formalization (Section 2.4.1), the data collection and the confident learning pipeline (Section 2.4.2), the human pose estimation (Section 2.4.3) and the segmentation (Section 2.4.4) filtering mechanism. Lastly, there will be the mathematical formulation of the POP++ signal (Section 2.4.5) and the experimental results (Section 2.4.6).

2.4.1 Problem Formalization

The goal of our data-centric method is to produce POP++, an exogenous signal that can help a forecasting model predict a product's future sales. POP++ is intimately related to the POtential Performance (POP) signal presented in [47], an estimation of the performance that a product might have had in the past, which was shown to help prediction models in sales and popularity predictions. We improve the quality and quantity of the information contained in the POP signal through two additional operations, detailed in the following sections, resulting in the POP++ signal.

Following [47], our approach takes as input a probe image $\mathbf{z}$ and an associated *observation time* t, which is the date from when we began looking into the past. The output is the POP++ signal $S_{\mathbf{z}}^{(t)}$, indicating the performance S of the item z in the time steps preceding t. Formally, POP++ is defined as:

$$S_{\mathbf{z}}^{(t)} = s_{\mathbf{z}}^{(t-K_{\text{past}})}, \ldots, s_{\mathbf{z}}^{(t-k)}, \ldots, s_{\mathbf{z}}^{(t-1)}, \tag{2.13}$$

where K_{past} is a fixed number of time steps (or observations) preceding t, $k = 1, \ldots, K_{\text{past}}$, and $s_{\mathbf{z}}^{(t-k)} \in [-1, 1]$.

The observation times are discretized by weeks ($K_{\text{past}} = 52$). The next sections will sequentially detail the pipeline of our approach, which is visually depicted in Figure 2.7.

2.4.2 Data Collection and Confident Learning

As a data-centric approach, we rely on information accessible on the Web to compute the exogenous signal POP++. Following the procedure presented in [47], we start by extracting tags pertaining to the class of the item (*e.g.*, long sleeves), fabric (*e.g.*, cotton) and color (*e.g.*, gray) from the probe image $\mathbf{z}$. In the fashion industry, these tags are often already paired to the probe image, but their extraction is easily

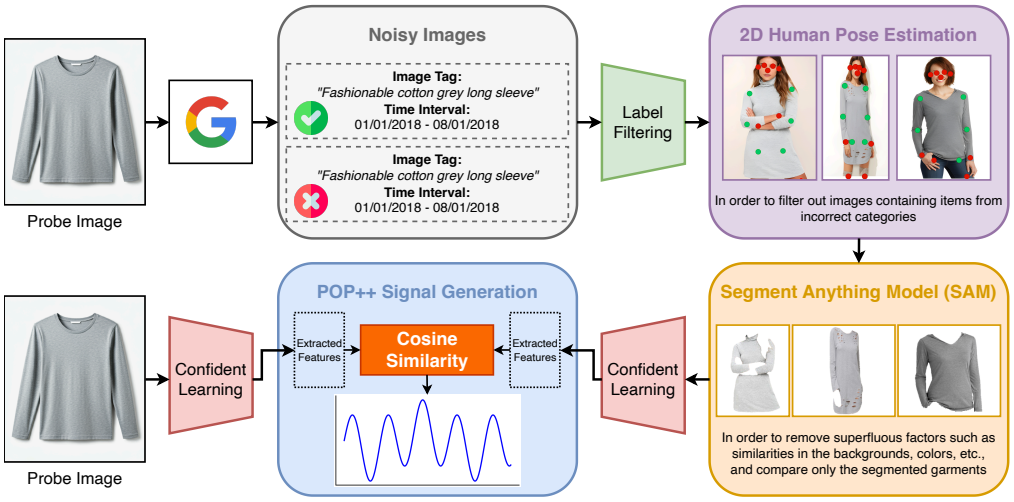

Fig. 2.7 A schematic overview of our proposed data-centric approach. We start with a probe image and obtain a specific POP++ signal for it at the end. The first step is mining for related, time-dependent images through a time-dependent query expansion on the web, followed by a confident learning procedure to eliminate noisy labels, as proposed in [47]. Afterward, we extract Human Pose (details in Section 2.4.3) and Segmentation (details in Section 2.4.4) features, which are used alongside the mined image features to calculate a similarity score with the starting probe image. This gives rise to the POP++ signal. Notably, the generated signals are highly multi-modal, containing information about past popularity based on temporal, image, pose, and segmentation features. We empirically show that such an elaborate data-centric and multi-modal approach leads to state-of-the-art forecasting results in Section 2.4.6.

automatable with off-the-shelf classifiers and attribute taggers. A textual query q is created by concatenating the tags:

$$q = \text{“<fabric> <color> <class>”}\,, \tag{2.14}$$

resulting in queries like “cotton gray long sleeve”. The goal of this query is to prompt a web search engine to collect images depicting the item, which we will use to build our exogenous signal. The intuition behind this operation is that popular items (*i.e.*, high-selling, trendy) would be among the top retrieved items when querying a search engine. As such, comparing the probe item $\mathbf{z}$ with the most popular similar items should provide us with valuable information on the potential performance of $\mathbf{z}$. However, the work in [47], noted that simply collecting images using such queries resulted in noisy data. This is because trending web images do not always depict *fashionable* items: it is common for *unfashionable* and controversial items to be featured in trending discussions. To automatically filter fashionable and

unfashionable items, a confident learning [48] approach is adopted to train classifiers that can discern true fashionable and unfashionable items. This approach can be summarized in the following three steps.

i) Query Expansion. This step augments the query q with the tags "fashionable" and "unfashionable", resulting in a positive (*i.e.*, fashionable) and negative (*i.e.*, unfashionable) query of the probe $\mathbf{z}$, as shown in Figure 2.7.

ii) Image Retrieval. This step uses positive and negative queries to prompt a web search API. To introduce temporal localization, for each of the k weeks prior to t (our observation starting point), the top a images that trended in the weeks spanning $k-W$ and k are collected. The hyperparameter W defines a sliding window of W weeks in which the items retrieved are grouped. After empirical evaluations, $a = 25$ and $W = 4$ were selected. The downloaded data are associated with a noisy binary label $\tilde{y} \in \{\text{"fashionable"}, \text{"unfashionable"}\}$, depending on which query was used to retrieve the images.

iii) Confident Learning. In this final step, classifiers are trained on the noisy labels [48] and automatically filter incorrectly labeled data. In short, the data downloaded in the previous step is used to train a binary classifier θ on the noisy labels $\tilde{y}$. Afterward, θ is used to classify the positive data downloaded (*i.e.*, data downloaded with the positive query) and record its confidence score in the classifications. The wrongly classified images (*e.g.*, "fashionable" labeled data classified as "unfashionable") are pruned, as well as all images for which confidence is lower than a threshold. Since θ has been trained on the same data, low classification confidence (or misclassification) indicates that the item was originally mislabeled or, in our case, the image downloaded with the "fashionable" tag depicted an item similar to "unfashionable" ones, and is thus removed from the final set of images. The downloaded data are then divided into *cleaned positive* (*i.e.*, "fashionable") images $\mathscr{I}^{(+)}$ that are correctly (and confidently) classified by θ and *negative* images $\mathscr{I}^{(-)}$, including the original "unfashionable" images and all images removed from the "fashionable" pool. Finally, a new binary classifier θ' is trained on the new data, resulting in a feature extractor that is capable of confidently distinguishing "fashionable" images. The implementation details for θ are described in Section 2.4.6.

2.4.3 Human Pose Estimation (HPE) Filtering

The confident learning pipeline described in the previous step successfully removes images that are clearly unfashionable or irrelevant to the desired criteria. However, it encounters significant challenges when filtering out images containing items from incorrect categories. To address this issue, we propose a novel data filtering pipeline that uses HPE techniques. This advanced approach uses pose detection algorithms to identify where key body joints, such as shoulders, elbows, hips, or knees, would logically be located in the items displayed in the images. By analyzing the spatial relationship between these joints and the surrounding objects, we can systematically remove images that do not contain relevant clothing items.

For example, when searching for pants, our method can check for the presence of leg joints within the image. The image probably does not depict pants or related lower-body clothing if these joints are absent. Consequently, such images can be excluded from the dataset of positive pants examples. We employ a bottom-up HPE model to avoid discarding relevant images that do not depict a person wearing the item of interest, unlike top-down models, which detect the person before extracting joint positions. For this reason, we are able to employ them to locate possible joint locations even when no person is present, thanks to their bias toward clothed people. This process ensures a more refined and accurate selection of images, effectively addressing the shortcomings of the initial pipeline and improving the quality of the data used in subsequent stages.

We start with a bottom-up, pre-trained HPE model HPE, the set of categories available in the dataset C_{dataset}, and the cleaned positive images $\mathscr{I}^{(+)}$. For each garment category $c \in C_{\text{dataset}}$, we define a hand-made binary mask $m_c = [m'_0, m'_1, ..., m'_{b_j}]$ with $m'_i \in \{0,1\}$, with b_j number of body joints. This mask defines whether the garment is expected to be worn over the joint represented by that specific index in terms of HPE in the image plane [68, 69].

The positive images are then fed to the model HPE($\mathscr{I}^{(+)}$), providing as a result a set of poses $P_{x_i} \in \mathbb{R}^{3\times b_j}$, where x_i is the i-th image in HPE($\mathscr{I}^{(+)}$). The third dimension of each P_{x_i} is a score vector σ_{x_i} assigned by the model that indicates how confident the model is about the result of the j-th joint. If this confidence score is close to 0, the model is not confident that that specific joint is present in the product within the image. Naturally, occlusions can also result in lower scores. We would like to stress that the model can estimate joint positions even in the absence of humans, given its bottom-up nature. We then apply a binarization of the scores

based on a threshold $\beta \in [0,1]$ in order to define which joint is actually present in the image, resulting in another visibility mask $m_c^{x_i}$ of the same form as m_c that contains these scores. Then, if $\langle m_c^{x_i}, m_c \rangle \geq \alpha \sum_{b_j} m_c$, with $\langle \cdot, \cdot \rangle$ being the dot product, the image is considered relevant and is kept. Otherwise, it is discarded because it is considered to not confidently contain the correct garment's category. $\alpha \in [0,1]$, in this case, is a threshold that regulates the tolerance between the overlapping part of the visibility mask $m_c^{x_i}$ w.r.t. m_c. In simpler words, we check if the HPE model confidently estimates a fraction greater than or equal to α of m_c. At the end of this process, the resulting data $\mathscr{I}_{\text{HPE}}^{(+)}$ is a filtered version of $\mathscr{I}^{(+)}$ where all images that do not contain the queried class are removed.

2.4.4 Segmentation

The ultimate goal of our pipeline is to generate a data-centric exogenous signal by comparing the similarity of a probe element with fashionable images from the past. A crucial aspect that we wish to consider here is that the similarity between images can be quite spurious due to superfluous factors such as similarities in the backgrounds, colors, etc. To bypass this issue and obtain a final result that is both as automatic and as clean as possible, we utilize a segmentation module to later compare only the segmented garments.

To do this, we use a recent and powerful segmentation model SAM [70]. Starting with cleaned positive images $\mathscr{I}_{\text{HPE}}^{(+)}$, pose P_{x_i}, and C_{dataset}, it is possible to inform SAM which element on the image plane to segment using the pose and mask information. Specifically, based on the mask m_c, we define $P_{x_i}^+ \subseteq P_{x_i}$ as the subset of positive points found in the binary mask $m_c^{x_i}$, $P_{x_i}^+ = \{p_{x_i}{}^{(j)} \in P_{x_i} \,|\, m_c^{x_i}[j] = 1\}$, where $p_{x_i}{}^{(j)}$ is the j-th column vector of P_{x_i}. On the other hand, we can also define the subset of points outside the garment w.r.t. $m_c^{x_i}$ as $P_{x_i}^- = P_{x_i} \setminus P_{x_i}^+$, where $\setminus$ is the set difference operation. Letting these two masks over the whole image set be denoted as P^+ and P^- respectively, we can then use them as additional information to inform the SAM segmentation model about which part of the image it should focus on, thereby defining the segmented images as:

$$\mathscr{I}_{seg} = \text{SAM}(\mathscr{I}_{\text{HPE}}^{(+)}, P^+, P^-) \,. \tag{2.15}$$

2.4.5 Signal Formation

The POP++ signal $S_{\mathbf{z}}^{(t)} = s_{\mathbf{z}}^{(t-K_{past})}, \ldots, s_{\mathbf{z}}^{(t-k)}, \ldots, s_{\mathbf{z}}^{(t-1)}$, is computed by considering at each timestep k the cleaned and segmented fashionable images $\mathscr{I}_{seg} = \{\mathbf{x}_i\}_{i=1,\ldots,M^{(t-k)}}^{(t-k)}$, the robust model θ', and the image $\mathbf{z}$, as follows:

$$s_{\mathbf{z}}^{(t-k)} = \frac{1}{M^{(t-k)}} \sum_{i=1}^{M^{(t-k)}} \frac{\left\langle \theta'\left(\mathbf{x}_i^{(t-k)}\right), \theta'(\mathbf{z}) \right\rangle}{\| \theta'\left(\mathbf{x}_i^{(t-k)}\right) \| \| \theta'(\mathbf{z}) \|}, \tag{2.16}$$

where $M^{(t-k)}$ indicates the number of images present after the cleaning procedure at timestep k, $\theta'(\cdot)$ indicates the extracted features of the input image, $\langle \cdot \rangle$ indicates the dot product, and $|| \cdot ||$ the Euclidean norm.

The POP++ signal value $s_{\mathbf{z}}^{(t-k)}$ is, therefore, the average cosine similarity between the embedding of the probe image $\mathbf{z}$ and each fashionable image $\mathbf{x}_{i^{(t-k)}}$ from the $M^{(t-k)}$ downloaded images.

2.4.6 Experimental Results

In this section, we empirically validate our proposed approach by showcasing how a forecasting model can gain an informative context on the past by using POP++ time-series as exogenous information, allowing it to achieve state-of-the-art performance. Specifically, we evaluate on the NFPPF task [51, 52, 20], where the goal is to predict a time-series indicating the future sales of a probe fashion product that has not been placed on the market before. By relying on extensive experiments, statistical analysis, and ablation studies, we demonstrate the effectiveness of POP++.

Implementation Details

The binary classifier θ for learning on noisy data (Section 2.4.2) is implemented through a ResNet-50 [62] pre-trained on ImageNet [63], with two additional fully connected layers. During the confident learning procedure, we fine-tune its last convolutional block and fully connected layers for 50 epochs with a batch size of 64, using Cross-Entropy loss, following a 5-fold cross-validation protocol. AdamW [71] is used as an optimizer, with a learning rate of $1e-4$.

The forecasting neural network models are all trained for 200 epochs with a batch size of 128 and L2 loss, using the AdaFactor [72] optimizer.

The *HPE* model defined in Section 2.4.3 is implemented through OpenPose [73] pretrained on COCO [74], while for the *SAM* model we use the official implementation of [70] and the corresponding pretrained weights. *HPE* and *SAM* are only used in inference mode and are not fine-tuned. All experiments were performed on a single NVIDIA 4090 RTX GPU.

Quantitative Results

New Fashion Product Performance Forecasting (NFPPF). Given their popularity in the literature, we use the VISUELLE dataset [20] and protocols for all forecasting experiments. The forecasting model output for a probe clothing item $\mathbf{z}^{(t)}$ is a time-series $\hat{Y}_{\mathbf{z}} = \hat{y}_{\mathbf{z}}^{(t+1)}, \ldots, \hat{y}_{\mathbf{z}}^{(t+T)}$, indicates how many pieces of z will be sold starting from the observation time t, which coincides with the planned release date, for the next T time steps. For the sake of clarity, we have overloaded some notations for the predicates y and T, which have been seen before. For evaluation, we apply the same metrics used in the original experimental protocol of [20], namely:

$$\text{MAE}(Y_{\mathbf{z}}, \hat{Y}_{\mathbf{z}}) = \frac{\sum_{k=t}^{t+T} |y_{\mathbf{z}}^{(k)} - \hat{y}_{\mathbf{z}}^{(k)}|}{T}, \tag{2.17}$$

$$\text{WAPE}(Y_{\mathbf{z}}, \hat{Y}_{\mathbf{z}}) = \frac{\sum_{k=t}^{t+T} |y_{\mathbf{z}}^{(k)} - \hat{y}_{\mathbf{z}}^{(k)}|}{\sum_{k=t}^{t+T} y_{\mathbf{z}}^{(k)}}, \tag{2.18}$$

where T is the number of timesteps to forecast into the future, $y_{\mathbf{z}}^{(k)}$ is the ground truth signal value at timestep k, with $y_{\mathbf{z}}^{(k)} \geq 0, \forall k \in \{t, \ldots, t+T\}$), and $\hat{y}_{\mathbf{z}}^{(k)}$ is the model's forecast. In order to facilitate its interpretation as a percentage metric, we report $100 * \text{WAPE}(Y_{\mathbf{z}}, \hat{Y}_{\mathbf{z}})$ in all of our experiments. For more information on the evaluation protocol and the NFPPF task, we refer the reader to [20].

Table 2.6 displays the results for different approaches available in the literature, alongside their data modalities used to compute the forecasts, on the VISUELLE *release setup* task. At the top of the table, we also report, as an indicative baseline, the error obtained using the mean and median of the training set sales as predictions for the testing set. All models are trained to predict 12-week sales signals, while the evaluation is performed on the first 6-week horizon. This is argued to give the best predictions while simulating the politics of real fashion companies [47], a topic we will dive into in more detail in the following subsection.

Table 2.6 Quantitative results of POP++ expressed in terms of WAPE and MAE on VISUELLE for the NFPPF task. The *Mean* and *Median* are two naive predictors based on the corresponding statistics over the whole training set. Additionally, we report the exogenous data modalities that each model uses in order to compute the forecasts. In **bold**, the best results. Underlined, the second best. (*) indicates that the model uses description as a further modality retrieved from other image-to-text models.

Method	IMAGE	RELEASE	DESCRIPTION	GOOGLE TRENDS	WAPE ↓	MAE ↓
Mean predictor	✗	✗	✗	✗	60.1	32.8
Median predictor	✗	✗	✗	✗	58.3	31.8
Attribute k-NN [52]	✗	✗	✓	✗	59.8	32.7
Image k-NN [52]	✓	✗	✗	✗	62.2	34.0
Attr+Image k-NN [52]	✓	✗	✓	✗	61.3	33.5
GBoosting [66]	✓	✓	✗	✗	64.1	35.0
GBoosting+G [66]	✓	✓	✗	✓	63.5	34.7
Cat-MM-RNN [52]	✗	✓	✓	✓	63.3	34.4
X-Att-RNN [52]	✓	✓	✓	✗	59.5	32.3
GTM-Transformer [20]	✓	✓	✓	✓	55.2	30.2
MDiFF [36]	✓	✓	✗	✗	54.7	30.1
Dif4FF [22]	✓	✓	✗	✓	54.6	30.0
GTM-Transformer + POP [47]	✓	✓	✓	✓	<u>52.3</u>	<u>28.6</u>
MuQAR [21]*	✓	✓	✓	✓	52.6	28.7
GTM-Transformer + Ours (ours)	✓	✓	✓	✓	**51.4**	**28.1**

As seen in Table 2.6, POP++ allows GTM-Transformer [20] to obtain the best performance on the NFPPF task, setting the new state-of-the-art. A notable result from our experiments is that deep learning models that can effectively use all available modalities perform the best, suggesting multi-modal deep learning as a promising research direction for forecasting applications.

In Figure 2.8, we compare some forecast signals when using POP and POP++, using GTM [20] as our forecasting model. The green squares indicate the ground-true 6-week-long sales signal to forecast; the red and the blue circles report the predicted sales using POP and POP++ as exogenous signals, respectively. Although GTM still struggles to predict sudden changes in signals with both POP++ and POP (*e.g.*, the drop in sales in Figure 2.8-d), it is clear how POP++ provides an overall improvement: both trend prediction (*i.e.*, increase or drop) and the magnitude of predicted sales qualitatively improve when using POP++ over POP.

POP++ is Model Agnostic. To further validate the quality of our POP++ signal, we perform additional experiments on the *first-order setup* and the *release setup* of VISUELLE, using POP++ as an exogenous signal with different models. The goal of these experiments is two-fold: *i)* proving that the use of POP++ provides a

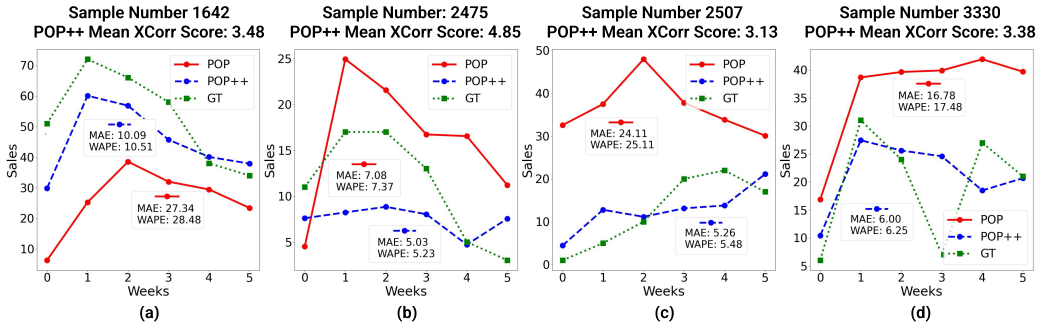

Fig. 2.8 Comparisons between the predictions made by GTM [20] trained using POP (red) and POP++ (blue) on four sale signals (green) from VISUELLE. The four examples have high average Cross-Correlation with POP++ (Section 2.4.6), as seen in the title of each subsplot). The MAE and WAPE for both exogenous signals are reported within the white boxes. These qualitative results clearly depict how using POP++ signals that have higher average cross-correlation with the ground truth sales leads to improved forecasts.

performance improvement to all NFPPF models; *ii)* prove that POP++ outperforms the previous POP in all scenarios presented in VISUELLE [20].

The *first-order setup* and the *release setup* introduced in VISUELLE consist of simulating how fast-fashion companies deal with new products in two different moments during production planning. In the *first-order setup*, the company decides which products and how many pieces to order by looking at the predicted sales of the probe items during the season. This scenario is particularly interesting for seasonal products, which are expected to perform well only during their main season (*e.g.*, heavy sweaters during autumn-winter). The only relevant data for forecasting the sales volume of these products is their previous years' seasonal performance (*e.g.*, for heavy sweaters released in autumn-winter, the previous years' autumn-winter performance). As such, in this scenario, fashion companies (and, in our case, NFPPF models) only have access to data regarding the 28 weeks that coincide with the release season of the previous year, formally from $t - K_{\text{past}}$ to $t - K_{\text{past}} + 28$.

On the other hand, the *release setup* is more suitable for products expected to sell throughout the whole year (*e.g.* jeans pants). This scenario occurs right before the start of a season, and the fashion company's goal is to obtain an accurate forecast of probe items' sales to plan any eventual stock replenishment. To do so, they have access to data from the previous 52 weeks (*i.e.*, previous year), from $t - K_{\text{past}}$ to $t - 1$.

Following the original protocol [20], both scenarios must forecast a 6-week signal, from t to $t + 5$. To quantitatively measure the performance of the model using

Table 2.7 Results on VISUELLE with the *first order setup*; "W" stands for WAPE, "M" for MAE, and "ERP" for the Edit distance with Real Penalty. Lower is better for all metrics.

First Order Setup (K_{best} = 28 weeks)															
Exogenous Signal	**Gradient Boosting** [75]			**Concat MM RNN** [52]			**Residual MM RNN** [52]			**X-Attention RNN** [52]			**GTM Transformer** [20]		
	W ↓	M ↓	ERP ↓	W ↓	M ↓	ERP ↓	W ↓	M ↓	↓	W ↓	M ↓	ERP ↓	W ↓	M ↓	ERP ↓
No Signal	64.1	35.0	0.43	63.3	34.4	0.42	64.2	34.9	0.44	59.5	32.3	0.38	56.6	30.9	0.37
Google Trends	64.3	35.1	0.43	64.1	34.8	0.43	68.1	37.0	0.47	58.7	31.9	0.38	56.8	31.1	0.35
POP Signal	63.8	34.8	0.42	58.1	31.7	0.39	58.9	32.2	0.39	57.8	31.6	0.38	53.4	29.2	0.32
Ours Signal (ours)	**58.3**	**31.9**	**0.30**	**56.2**	**30.7**	**0.33**	**56.6**	**30.9**	**0.35**	**54.3**	**29.7**	**0.28**	**52.6**	**28.7**	**0.29**

Table 2.8 Results on VISUELLE with the *release setup*; 'W" stands for WAPE, "M" for MAE, and "ERP" for the Edit distance with Real Penalty. Lower is better for all metrics.

Release Setup (K_{best} = 52 weeks)															
Exogenous Signal	**Gradient Boosting** [75]			**Concat MM RNN** [52]			**Residual MM RNN** [52]			**X-Attention RNN** [52]			**GTM Transformer** [20]		
	W ↓	M ↓	ERP ↓	W ↓	M ↓	ERP↓	W ↓	M ↓	ERP ↓	W ↓	M ↓	ERP ↓	W ↓	M ↓	ERP ↓
No Signal	64.1	35.0	0.43	63.3	34.4	0.42	64.3	34.9	0.44	59.5	32.3	0.38	56.6	30.9	0.37
Google Trends	63.5	34.7	0.42	65.9	35.8	0.44	68.5	37.2	0.48	59.0	32.1	0.38	55.2	30.2	0.33
POP Signal	63.4	34.6	0.42	57.4	31.4	0.36	58.4	31.9	0.39	57.4	31.3	0.36	52.4	28.6	0.29
Ours Signal (ours)	**58.3**	**31.9**	**0.31**	**55.7**	**30.4**	**0.33**	**56.1**	**30.6**	**0.34**	**53.3**	**29.1**	**0.30**	**51.4**	**28.1**	**0.27**

POP++, in addition to the WAPE and MAE metrics presented previously, we report the similarity between the slope of the predicted curve and the ground truth using the Edit Distance with Real Penalty (ERP). This metric quantifies the number of edit operations—insertions, deletions, and replacements—required to transform one time-series into another. To account for continuous values, we use a threshold of $\varepsilon = 0.03$ to determine whether two values are sufficiently different to require editing.

In both the *first order setup* (Table 2.7) and the *release setup* (Table 2.8), POP++ consistently improves the performance of all models, achieving better results compared to all other exogenous signals in time-series, indicating its quality and informative properties. An interesting outcome of our data-centric approach is the significant performance boost observed in simpler models like Gradient Boosting (GB). When using POP++ as exogenous input, GB achieves performance levels comparable to the more complex X-Attention RNN with POP. This finding has important implications for resource-constrained scenarios, where simpler models can deliver competitive results with a lower computational cost, and it further highlights the strength of the predictive priors embedded in POP++.

Correlation Analysis. To provide additional quantitative evidence on the differences between POP and POP++, in this section, we present two different correlation

analyzes between the aforementioned approaches. The first calculates the Pearson correlation coefficient between the first-order differences of the two signals. Intuitively, this allows us to compare the linear correlation on absolute week-over-week change in the two signals. In this way, we can avoid capturing correlations simply because of temporal trends. Let the POP and POP++ signals for a product z be defined as $\Gamma_{\mathbf{z}}^{(t)}$ and $S_{\mathbf{z}}^{(t)}$, respectively. Then, the Pearson correlation coefficient between these two signals is calculated as:

$$\Gamma_{\mathbf{z}}^{\prime(t)} = \Gamma_{\mathbf{z}}^{(t)} - \Gamma_{\mathbf{z}}^{(t-1)} \,, \tag{2.19}$$

$$S_{\mathbf{z}}^{\prime(t)} = S_{\mathbf{z}}^{(t)} - S_{\mathbf{z}}^{(t-1)} \,, \tag{2.20}$$

$$r = \frac{\sum_{k=1}^{K_{\text{past}}-1}(s_{\mathbf{z}}^{\prime(t-k)} - S_{\mathbf{z}}^{\prime\bar{(t)}})(\gamma_{\mathbf{z}}^{\prime(t-k)} - \Gamma_{\mathbf{z}}^{\prime\bar{(t)}})}{\sqrt{\sum_{k=1}^{K_{\text{past}}-1}(s_{\mathbf{z}}^{\prime(t-k)} - S^{\prime(\bar{t})})^2}\sqrt{\sum_{k=1}^{K_{\text{past}}-1}(\gamma_{\mathbf{z}}^{\prime(t-k)} - \Gamma^{\prime(\bar{t})})^2}} \,, \tag{2.21}$$

where r is the correlation coefficient between the two signals, $\Gamma^{\prime\bar{(t)}}, S^{\prime\bar{(t)}}$ are the average values of the signals over $K_{\text{past}} - 1$ time steps (one value is lost due to the first-order difference: the rate of change at time step $t - K_{\text{past}}$ is 0). The ultimate goal behind this procedure is to understand if the signals behave similarly when stationary. Figure 2.10 displays the results, showing that the distribution of the correlation coefficients is symmetric, bell-shaped, and centered at 0, indicating that these coefficients are essentially uncorrelated on average. This fact shows that the additional filtering steps we propose in this thesis give rise to an exogenous signal different from the original POP.

Knowing this, we proceed to analyze the correlation of these exogenous signals with the sales in VISUELLE, something that directly affects the NFPPF task. Given the results presented in Table 2.6, one would expect POP++ to report a better correlation with the product sales compared to POP. To quantitatively validate this intuitive conclusion, we calculate the cross-correlation (*i.e.*, the sliding dot product) between exogenous signals and the corresponding product sales. Given a lag parameter n and two real-valued discrete signals f and g, the cross-correlation between them is another signal:

$$C_{fg}[n] = \sum_{k=0}^{|g|-1} f[n+k]g[k] \,, \tag{2.22}$$

Fig. 2.9 Cross-correlation signals of POP++ (blue) and POP (red) [47] and the product sales on VISUELLE. On the left side, we present a density plot that represents every quantile of the respective distributions by different color intensities. The right-hand side reports all the cross-correlation signals. It is immediately clear that POP++ has much higher alignment over time with the product sales compared to POP, with some cases reporting very high cross-correlation.

Fig. 2.10 Histogram of the Pearson correlation coefficient between the first-order differences of the POP signal [47] and POP++ signals of each product in VISUELLE. The resulting distribution closely resembles a standard Gaussian, indicating that the rates of change in the two signals are not correlated on average.

Table 2.9 Ablation on the different components of the POP++ signal creation pipeline, using GTM[20] as the forecasting model. The last row is the proposed state-of-the-art configuration.

HPE Filtering	SAM Segmentation	WAPE ↓	MAE ↓
✗	✗	52.3	28.6
✓	✗	52.2	28.5
✓	✓	**51.4**	**28.1**

where $|g|$ is the cardinality of signal g and $0 \leq n \leq |f| - (|g| - 1)$. We use the POP/POP++ signals as f, the sales as g, and compute the values for all available n, resulting in cross-correlation signals. The results are reported in Figure 2.9, where it is immediately visible that POP++ has a much higher affinity for sales compared to POP, complementing the results showing improved forecasting ability. The density and line plots in Figure 2.9 show that the average cross-correlation over different lags is higher for POP++, with some exception cases reporting values > 3. We show some qualitative results for these cases in Figure 2.8, where we pick four samples that have an average cross-correlation value > 3 and then compare the forecast sales curves using the corresponding POP++ and POP signals. A notable result is that the forecast values of the POP++ signals with high sales cross-correlation are located close to the ground truth and, most importantly, are almost never overestimated, unlike POP. This aspect contributes to less waste and better stock management. Furthermore, as expected from the quantitative results in Tables 2.6, 2.7, and 2.8, it is easy to see that the MAE when using POP++ is lower compared to POP. These qualitative results clearly reveal that having higher cross-correlation leads to improved forecasts, which is advantageous for our approach given the results in Figure 2.9.

Ablation Studies

In this section, we ablate two different aspects of our proposed pipeline. The first regards only the different components of the POP++ signal creation pipeline, with the associated NFPPF performance changes. The second is a comparison with the ablations presented in [47], where the focus is on showing how the newly proposed POP++ performs better compared to alternative modifications of the query expansion and confident learning steps of the pipeline. In all these experiments, we only consider the release setup, seeing how POP++ proves very effective in the first-order setup (Table 2.7).

The different configurations and their results are presented in Table 2.9. The previous POP signal did not use HPE filtering or SAM background segmentation, and we report its performance in the first row. By filtering the web images based on the joints found by the HPE model, we see a slight performance improvement compared to the baseline POP. This is expected as confident learning classifiers cannot distinguish between garments in the image. When extracting features from downloaded images, the classifiers cannot remove images containing garments that do not match the probe class (*e.g.*, removing images of skirts when the probe image is a jacket). This introduces noise in the signal when some attributes are fashionable for one class but unfashionable for another (*e.g.*, scratched jeans might be fashionable, but scratched shirts might not).

Furthermore, the classifier's inability to distinguish between the subject of interest and background information can degrade the quality of the signal. As explained in Section 2.4.3, when extracting features from the downloaded gallery, the classifiers might inject additional information about the background and, more importantly, other garments present in the image that do not match the probe category. By removing unwanted information from the image through segmentation (SAM), we see an additional improvement in the downstream forecasting task. These results highlight the effectiveness of all the additional filtering processes introduced, resulting in our rich and informative exogenous signal POP++.

2.5 Discussion

In this chapter, we propose MDiFF and Dif4FF, two new two-stage multimodal pipelines based on diffusion models, exploiting generative Artificial Intelligence (AI) models for NFPPF for the first time. The two works investigate two different case histories, exploiting different types of modalities as conditioning for generation, to provide different architectures to solve the problem depending on the data that are available to the user. Despite the effectiveness of diffusion models, they can sometimes produce inconsistent predictions for the same object. To address this problem, we generate multiple forecasts for each sample using the diffusion model and then use them as input for forecasting an MLP-based model in the case of MDiFF, then extend to a Graph Convolutional Network (GCN) model for Dif4FF.

Lastly, POP++ introduces a new data-centric solution to the problem of past POtential Performance of fashion products. By building a historical context from

textual tags and retrieving visually and thematically similar web images, we simulate the trajectory of a garment's potential performance. The integration of noisy label learning methods, combined with novel human pose filtering and garment segmentation steps, leads to the creation of high-quality, noise-free exogenous data that can be used for reliable predictions. POP++ provides a unique lens for fashion trend analysis, effectively simulating a fashion expert analyzing past market performance and using this information as a strong exogenous indicator of future success.

Chapter 3

2D Time-Series Forecasting: Human Trajectory Forecasting

Human trajectory prediction is the task of predicting the likely path that a subject will take to reach its designated endpoint [76]. This predictive process finds its applicability and utility in a multitude of domains [77]. For example, in the context of robotics, it serves as a tool for facilitating the predictions on potential future robot trajectories, useful for intelligent planning considering human responses [78]. In industry, human trajectory prediction becomes critical for optimizing automated systems and ensuring seamless interactions with other occupants and components of a production line [16].

Despite the significant volume of research over the past decade devoted to outdoor trajectory prediction [79, 80, 78, 81, 82, 83, 84], there has been a notable scarcity of studies that exploited user trajectory data in indoor settings [85, 86, 87, 88, 89, 90, 91], also considering the crucial role these predictions play nowadays in the development of location-based services within indoor spaces. This gap in research inspired this work, which investigates a learning framework designed explicitly for indoor trajectory prediction.

Motivations for this Method. In Figure 3.1, we can note the distinctive nature of indoor settings, where users can encounter numerous choices and potential pathways. This factor implies that the dynamic of the motion can be strongly influenced by the environment setup [85, 93]. Users can navigate through different interconnected rooms, corridors, doors, and elevators, often having the freedom to deviate from straightforward paths and choose alternative routes. Indoor spaces also have a higher

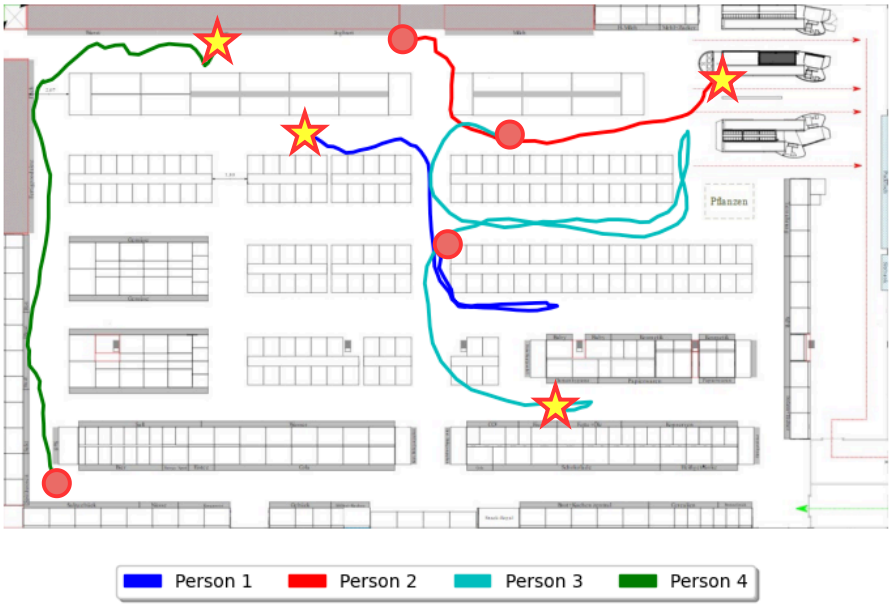

Fig. 3.1 Examples of different trajectories from the Supermarket [92] dataset to show the difficulty of the indoor trajectory prediction task. In particular, the dataset showcases long trajectories (Person 4), self-loops (Person 1 and Person 3), and confusing movements (Person 2) performed in an environment that strongly affects the people's paths. Specifically, the red circle represents the starting point of a trajectory, and the yellow star represents its final point.

density of structural elements and potential obstacles, such as furniture, walls, and partitions, as shown in Figure 3.1, related to the Supermarket [92] dataset. Outdoor environments provide more open spaces, where visibility is less restricted, and the impact of physical barriers is typically reduced [94].

Consequently, indoor trajectory prediction requires a deeper understanding of the context and semantics of the indoor space, as users may have specific goals, like finding a particular room, reaching a specific point of interest, or accessing various facilities [95]. This contextual richness adds a layer of complexity to the prediction process since it also makes it necessary to consider the space's physical layout. Considering the omnipresence of indoor environments in human lives, it is imperative to address trajectory forecasting in these situations. Indeed, recent studies show that humans spend most of their time in indoor environments such as homes, supermarkets, airports, conference facilities, and train stations [96]. These considerations form the basis of the research conducted in this work.

Innovations in this Method. While outdoor trajectory forecasting has received significant attention, indoor forecasting is still an underexplored research area. As a result, we present SITUATE, the first model designed specifically to cope with indoor trajectory forecasting by leveraging equivariant and invariant geometric feature learning and a self-supervised vision representation. Equivariance refers to the model's ability to adapt its output consistently with transformations applied to the input. In the case of trajectory forecasting, this means that if the entire scene or coordinate system undergoes a rigid transformation (e.g., translation or rotation), the predicted trajectory should transform in the same way. This is particularly important indoors, where spatial layouts vary but movement patterns (e.g., navigating around obstacles or toward goals) remain structurally similar. Invariance, on the contrary, requires the output to remain unchanged under certain transformations of the input. For instance, in tasks like predicting the next action or classifying the type of movement, the result should be invariant to global shifts or rotations of the environment. In such cases, the identity of the action or its classification does not depend on where it happens, only on how the movement evolves. On the other hand, the self-supervised vision representation module enabled us to acquire spatial-semantic information about the environment, using the scene or space layout images when available, to predict users' future locations meaningfully and accurately.

In summary, the main contributions of this chapter are:

- We present SITUATE, a novel approach for indoor human trajectory forecasting based on equivariant and invariant geometric feature learning modules and a self-supervised vision representation;
- The equivariant and invariant modules are used to cope with the problem related to the more complicated movements inherent in indoor spaces;
- The vision representation module is used to acquire spatial-semantic information about the environment to predict users' future locations more accurately;
- SITUATE also achieves competitive results in outdoor scenarios, showing that indoor forecasting models generalize better than outdoor-oriented ones.

3.1 Related Works

Indoor Human Trajectory Prediction. Predicting the evolution of a pedestrian trajectory in the future is a long-standing task whose interest is constantly renewed by the emergence of new scenarios that can benefit from it, *e.g.*, autonomous driving [76]. When proposing a methodology to tackle trajectory forecasting, one should take care of several aspects, from the environment's geometry [85] to the presence of obstacles [97] and the possible interactions between multiple agents [98]. Some traditional methods to approach this task involved force models [99], Markov models [100], and RNNs [101]. Notably, considering common sense rules and conventions that humans observe in social spaces helps to manage simultaneous predictions in crowded scenes [79].

Multiple deep learning-based models have been applied successfully to forecast pedestrian trajectories, such as GNNs [102], Trasformers [81] and Conditional Variational AutoEncoders (CVAEs) [103]. More recently, diffusion models have also been applied to solve this problem [82]. However, most proposed methods are tested only on datasets representing outdoor scenarios. This is due to a lack of comprehensive indoor datasets and the fact that indoor trajectories can be considered more "difficult" or non-linear [87]. When traversing indoors, our immediate movement decision is influenced by the objects in our path and the surrounding walls [85]. In indoor settings, people navigate in loosely constrained but cluttered spaces with multiple goal points that can be reached in many ways [86]. Moreover, people in indoor scenarios tend to focus on their surroundings, fixating on the most interesting parts of the scene, alternating movement and stationary phases [91]. At the same time, outside, the movement area can be much larger, and the subjects can move further apart.

Some works have been proposed to address the specific problem of indoor trajectory forecasting [85, 86, 87, 89, 90], highlighting the differences between indoor and outdoor trajectory forecasting. In [87], the authors address the problem of generalizability, proposing a novel indoor dataset and new metrics to normalize common biases. They tackle the problem of aleatoric multimodality with the GAN-Tri model, which uses a heuristic to produce samples corresponding to different behaviors. [85] examine trajectories, modeled as a Markov chain, within 3D environments, introducing the concept of an occupancy map to represent the relative accessibility of each point on the map with respect to its geometry. The study emphasizes the importance of proximity from each point to the destination and the occupancy frequency in

constructing a probability transition matrix for trajectory prediction. Unlike them, our approach considers indoor spaces' detailed scene layouts and non-trivial human movements.

Equivariant and Invariant Graph Neural Networks. Inspired by the research on rotation-equivariant convolutional neural networks within the 2D image domain [104], the advent of Graph Neural Network (GNN) architectures opened doors to investigating symmetries beyond rotations [105]. For example, in [106], the authors proposed partial equivariance by focusing on translation equivariance. Meanwhile, [107] constructed filters using spherical harmonics, enabling equivariance to rotations and translations and facilitating transformations between higher-order representations.

In [108], a new model for learning equivariant graph neural networks, dubbed EGNNs, is proposed. Differently from the previous works, this formulation maintains the flexibility of GNNs while remaining E(n) equivariant (translation, rotation, and reflection equivariant) without the need to compute expensive higher-order operations. [109] further extended this concept by incorporating geometrical constraints implicitly encoded in the forward kinematics when tackling molecular dynamics prediction and human motion capture. However, a significant limitation of current methods is their focus solely on state prediction, preventing models from effectively using sequence information.

Recently, EqMotion [27] extended on these ideas to propose an equivariant motion prediction parametric network with an invariant interaction reasoning module, able to tackle distinct problems such as particle and molecule dynamics, human pose forecasting, and outdoor pedestrian trajectory prediction. Interaction invariance is fundamental in ensuring the agents' interactions remain constant under input transformation.

In our research, we adapt some of the concepts presented in [27] to propose an equivariant model for the human trajectory prediction task, which, in combination with a module to extract semantic information about the scenes, unlock more precise forecasting capabilities in indoor settings.

Self-supervised Vision Representation. One way to get image representations without heavily relying on annotated data is to perform Self-Supervised Learning (SSL). In a nutshell, SSL learns deep feature representations invariant to sensible

transformations of the input data. Then, the learned representations could be used in supervised downstream tasks.

The self-supervised vision representation state-of-the-art rapidly evolved, with Transformer-based architectures emerging as leading models. The Vision Transformer (ViT) [110], and its variants like DeiT [111], have demonstrated impressive performance in learning powerful visual representations from unlabeled data. Specifically, these models leverage self-attention mechanisms to capture global context and long-range dependencies within images, enabling them to encode rich semantic information efficiently.

In this method, to extract semantic information from scenes represented in a 2D map, we use the pre-trained BEiT [112], the state-of-the-art self-supervised vision representation model. This offers a powerful framework for learning visual representations without explicit supervision, effectively capturing high-level semantics and intricate features inherent in visual data.

3.2 SITUATE

In this section, SITUATE will be introduced, a novel architecture tackling Indoor Human Trajectory Forecasting. The model is based on a GCN specifically designed to ensure invariance and equivariance properties to the input graph. Moreover, the model is informed about the scene through a representation of the environment. In Section 3.2.1 there is a short background on the geometrical properties of equivariance and invariance, in Section 3.2.2 the formalization of the problem, and in Section 3.2.3 the description of the architecture.

3.2.1 Mathematical Background

Given a set of transformations $T_x : X \to X$, a function $F : X \to Y$ is called Equivariant if exists a transformation $T_y : Y \to Y$ equivalent to T_x, on the Euclidean space such that:

$$F(T_x(X)) = T_y(F(X)) \ . \tag{3.1}$$

Moreover, we also want the model to have the invariance property. Given the same set of transformations, a function $F : X \to Y$ is called Invariant on the Euclidean

space if it is a transformation $T_y : Y \to Y$ such that:

$$F(X) = F(T_x(X)) \ . \tag{3.2}$$

Specifically, this work addresses the problem of multi-person trajectory forecasting by considering the input trajectories as a graph. As proven by [108], during the message passing of a GNN, the property of equivariance can be ensured by enriching the features of the neighbor nodes with the $L2$ distance between nodes. Let $G = \{V, E\}$ be an input graph representing the input trajectory with nodes $v_i \in V$ and edges $e_{ij} \in E$. For every node v_i, a feature vector $h \in \mathbf{R}^h$ and an absolute position $x_i \in \mathbf{R}^3$ are given. To preserve equivariance among different layers of the model, we update the position as follows:

$$m_{ij} = \phi_e \left(h_i^l, h_j^l, \left\| x_i^l - x_j^l \right\|^2 \right) \ , \tag{3.3}$$

$$x_i^{l+1} = x_i^l + C \sum_{j \neq i} \left(x_i^l - x_j^l \right) \phi_x \left(m_{ij} \right) \ , \tag{3.4}$$

where C is equal to $1/(M-1)$ with M number of nodes, ϕ_e and ϕ_x are learnable Multi-layer Perceptrons (MLPs), defined as $\phi_e(\cdot) = W_e \cdot + B_e$, l indicates the layer and m_{ij} represents the information passed between two nodes during the message passing.

As reported in [108], ϕ_x has to be a scoring function $\phi_x : X \to S$, with $S \in \mathbf{R}^1$. With this procedure, the update of V is consistent, allowing the model to learn without being affected by SO(2) transformations, with SO(2) being the group of all rotations in the plane around the origin that preserve the Euclidean norm, mathematically described by 2×2 matrices. Furthermore, the features learned across layers must be consistent and invariant to graph transformations. To do so, the following procedure governs the final message-passing operations and the update of the features carried out by the $i-th$ layer:

$$m_i = \sum_{j \neq i} m_{ij} \ , \tag{3.5}$$

$$h_i^{l+1} = \phi_h \left(h_i^l, m_i \right) \ , \tag{3.6}$$

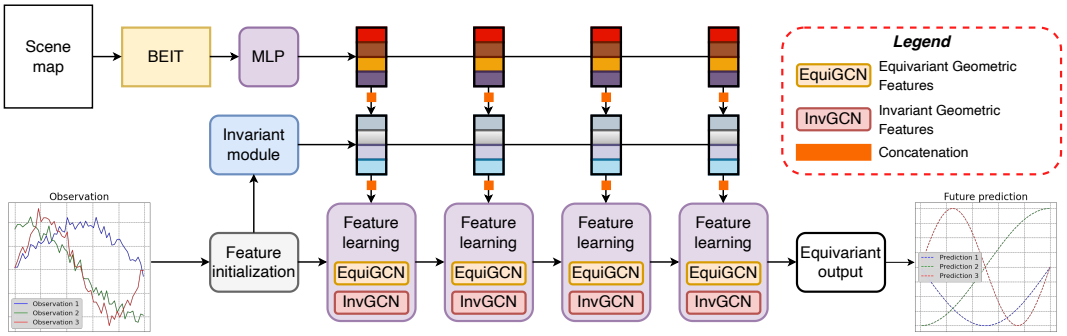

Fig. 3.2 In SITUATE, we first produce a feature vector regarding the scene using the self-supervised vision representation module. Then, a feature initialization layer is used to initialize geometric and pattern features. We then successively update the geometric and pattern features by the equivariant geometric feature learning and invariant pattern feature learning layers, obtaining expressive feature representation. We further use an invariant reasoning module to infer an interaction graph used in equivariant geometric feature learning. Finally, we use an equivariant output layer to obtain the final prediction.

with ϕ_h, an MLP also designed as $\phi_h(X) = W_h X + B_h$, responsible for the invariant feature learning. Mixing these two components allows us to build an Equivariant and Invariant GNN using Euclidean SO(2) transformations.

3.2.2 Problem Formalization

Here, we introduce the general problem formulation of motion prediction. We have a multi-agent system with m agents. Each agent is represented as A_i, where $i = 1, 2, \ldots, m$. The goal is to predict the future motions of these agents based on their historical observations. For each agent A_i, we can denote the historical observations as X_i. These observations typically include positions and can be represented as $X_i = \{x_0^i, x_1^i, \ldots, x_t^i\}$, where x_t^i represents the position of agent A_i at time step t. We also add the velocity $S_i = \{s_{t+1}^i, s_{t+2}^i, \ldots, s_{t+f}^i\}$ as input information of the model. The velocity of an agent is a natural invariant feature because it is not affected by any translation or SO(2) transformation. We use the velocity to compute the initial feature vector of a specific agent A_i. More details in Section 3.2.4. Specifically, for each agent A_i we aim to predict its future f positions $Y_i = \{y_{t+1}^i, y_{t+2}^i, \ldots, y_{t+f}^i\}$.

3.2.3 The SITUATE Prediction Network

In this section, we present SITUATE, our motion prediction network that explicitly uses equivariant and invariant geometric features and a self-supervised scene rep-

resentation module to tackle the indoor trajectory prediction problem. The model architecture is shown in Figure 3.2.

The first module we present is in charge of producing the scene-representation encoding. As anticipated, the subjects' motion characteristics differ greatly from those of the outdoor case when considering indoor trajectory forecasting. Indeed, the motion is strongly characterized and limited by the objects and obstacles in the scene. Knowing the available space that limits the viable paths in the scene can, for every X_i, strongly reduce the cardinality of all the possible outcomes of the model. Starting from the assumption that all the scene objects and structure are available in the form of a scene layout or a camera image, BEiT [112] is first used to output visual tokens T_s, the so-called scene-representation encodings. These tokens T_s are fed into a learnable MLP defined as $\phi_t : T_s \rightarrow T_e$ and then concatenated into the input.

The input concatenated with T_e is then fed into two modules: Equivariant block (*EquiGCN*) and Invariant block (*InvGCN*). Following [27], these two blocks are both based on the implementation of the message passing described in Equation 3.3, modified to accept also T_e:

$$m_{ij} = \phi_e \left(h_i^l, h_j^l, \left\| x_i^l - x_j^l \right\|^2, T_e, a_{ij} \right) , \tag{3.7}$$

where a_{ij} is the edge attribute (or weight), which can be derived from the adjacency matrix.

Specifically, the *EquiGCN* block is responsible for the update of the node's coordinates x, and it represents the implementation of the update function described in Equation 3.4. On the other hand, *InvGCN* is the implementation of the update function of the node's features h in Equation 3.6.

The possible pathways are learned by the module ϕ_t, starting from the token produced by the pre-trained BeIT model. The outputs of *EquiGCN* and *InvGCN* are computed as reported respectively in Equation 3.4 and Equation 3.6, updating h_i^{l+1} and x_i^{l+1}.

To understand the contribution of these two modules in less formal and more practical terms, imagine a navigation system. It can get from point A to point B but might struggle with tricky situations. In SITUATE, EquiGCN injects a sense of direction like a compass you wear on your hat. Specifically, it ensures the network understands the environment's layout, regardless of where it starts "looking". InvGCN, on the other hand, acts like a map you hold – it helps the network account

for different starting points and body orientations, making the predictions more robust.

3.2.4 Feature Initialization

The input given to our model is a set of trajectories of different agents. The first step is to define a node for every position x_t for every agent A_i. Every node x_t is connected with x_{t-1} and x_{t+1} if they are related to the same agent A_i. Since only trajectories (and thus positions x_t) are given as starting data, it is necessary to define for each trajectory a vector of initial features h_i^0 to be used as input together with the positions x_i^0.

As stated in [109], having an invariant feature vector h_i^0 is necessary to guarantee equivariance. Given that as input data we only have position X_i, we followed the procedure in [27] to use velocities in order to create h_i^0 as follows:

$$\hat{x}_i = \phi_X(X_i + \overline{\mathbb{H}}) + \overline{\mathbb{H}}\,, \tag{3.8}$$

$$\rho_i^t = \|v_i^t\|_2\,, \tag{3.9}$$

$$\theta_i^t = \text{angle}(v_i^t, v_i^{t-1})\,, \tag{3.10}$$

$$h_i^0 = \phi_{h_0}(\rho_i, \theta_i)\,, \tag{3.11}$$

where h_i^0 is the initial features vector of the $i-th$ agent. v_i^t represent the velocity of the agent and is defined as $\triangle\hat{x}_i^t$, where $\triangle$ is the finite difference operator, $\overline{\mathbb{H}}$ is the centroid of the observed trajectories of all agents in the scene. ϕ_X and ϕ_{h_0} are two fully connected layers responsible for encoding and producing the initial graph and the initial features of the trajectory.

To compute h_i^0, two different types of velocities are needed (thus, information invariant to rotation and translation): $\triangle x_i^t$ effectively represents the Euclidean velocity of the agent and θ_i^t represents the angular velocity on a certain time step t. Note that both ϕ_{x_0} and ϕ_{h_0}, and in general all the operations described, are linear transformations: this is necessary to preserve both equivariance and invariance properties of the remaining part of the model.

3.2.5 Experimental Results

Our experimental evaluation is tailored toward two objectives. Firstly, in Section 3.2.5, we show the superiority of SITUATE in the two most well-known indoor

datasets, defining the new state-of-the-art in indoor scenarios. Secondly, in Section 3.2.5, we prove that SITUATE can also achieve comparable results concerning other competitors on outdoor datasets. Finally, we report some ablation studies in Section 3.2.5.

Evaluation Setup

Datasets. We evaluate SITUATE on the state-of-the-art indoor datasets and the most well-known outdoor human trajectory prediction dataset.
THÖR. The THÖR dataset [86] includes human motion trajectory and gaze data collected in an indoor environment with accurate ground truth for the participants' position. It comprises 395K frames at 100 Hz, 2531K people detections, and over 600 individual and group trajectories between multiple resting points. The map was taken from the dataset's official website.
Supermarket. The Supermarket dataset [92] comprises 4 different scenario: German1, German2, German3, and German4, *i.e.*, four different supermarket. The dataset collection involved attaching devices on shopping carts/baskets and recording their movements during customer usage. Each subset includes a file with a map of the supermarket.
ETH-UCY. The ETH [113] and UCY [114] dataset group consists of five different scenes: ETH & HOTEL (from ETH) and UNIV, ZARA1, & ZARA2 (from UCY). The scenes are captured in unconstrained outdoor environments with few objects blocking the pedestrian paths. In this case, images of the scene were used.

Evaluation metrics. We use standard metrics for the trajectory prediction task, *i.e.*, minimum Average Displacement Error (ADE), and minimum Final Displacement Error (FDE). In particular, ADE measures the average $L2$ difference between the prediction at all time steps and the ground truth. On the other hand, FDE measures the difference between the predicted endpoint and the ground truth.

Prediction mode. Following the evaluation protocol of [27], SITUATE is employed in two prediction modes: deterministic and multi-prediction. Deterministic means the model only outputs a single prediction for each input motion observation, while multi-prediction means the model has 20 predictions for each input motion observation. Under multi-prediction, ADE and FDE will be calculated using the best-predicted trajectory. To adapt to multi-prediction, we modify SITUATE to

Table 3.1 Deterministic prediction performance (ADE (m)/FDE (m)) on the THÖR and the Supermarket datasets. The **bold**/underlined font denotes the **best**/second-best result.

	Performance (ADE (m) ↓ / FDE (m) ↓)		
Deterministic Evaluation	**THÖR**	**Supermarket**	**Average**
TransF [81]	2.62/4.81	2.56/2.90	2.59/3.85
MemoNet [83]	0.78/5.05	1.79/2.94	1.28/3.99
EqMotion [27]	0.56/0.94	1.71/2.65	1.13/1.79
SITUATE (ours)	**0.45/0.93**	**1.21/1.84**	**0.83/1.38**

repeat the last feature updating layer and the output layer 20 times in parallel to have a multi-head prediction.

Implementation details. As a backbone for our model, we used the structure of [27]. The model architecture has four layers of geometric feature learning. We use the SiLU activation function and dropout with a 0.5 probability to regularise within all MLPs. The visual embeddings of the image, *i.e.*, the floor plans (look at Figure 3.1) for context information are derived from the last layer of the BEiT model. The model is provided with past trajectory information spanning eight discrete time steps, and the model's task is to predict 12 steps into the future. In addition to the dropout mentioned above, we apply the Discrete Cosine Transform (DCT) to the input data as a regularisation technique. Specifically, by representing the data in the frequency domain, it becomes easier to distinguish between signal and noise components, resulting in a cleaner signal. The impact of these regularization approaches is discussed in Section 3.2.5. We train our models with a batch size of 64 for 60 epochs, using AdamW [65] as an optimizer within the PyTorch Lightning framework on an NVIDIA RTX 3090.

Indoor Human Trajectory Prediction Results

We conducted comparative experiments to assess the soundness of our approach against existing trajectory prediction methods. The methods include deterministic evaluation models (TransF [81], MemoNet [83]), as well as multi-prediction evaluation models (PECNet [103], GP-Graph [84]). Our evaluation for both prediction modes also encompasses EqMotion [27], the state-of-the-art method with invariant end equivariant interaction reasoning. Table 3.1 and Table 3.2 show the results.

Table 3.2 Multi-prediction performance (ADE (m)/FDE (m)) on the THÖR and the Supermarket datasets. The **bold**/underlined font denotes the **best**/second-best result.

	Performance (ADE (m) ↓ / FDE (m) ↓)		
Multi-prediction Evaluation	**THÖR**	**Supermarket**	**Average**
PECNet [103]	–	1.57/3.45	–
GP-Graph [84]	2.80/3.92	3.19/4.57	2.99/4.24
EqMotion [27]	1.32/1.03	1.29/1.77	2.61/1.40
SITUATE (ours)	**0.50/0.86**	**0.53/0.65**	**0.51/0.75**

The results show that the proposed SITUATE consistently outperforms all baseline methods in all cases. On the THÖR dataset, SITUATE achieves an ADE of 0.45 and an FDE of 0.93, showcasing its superiority over other models. Notably, compared to EqMotion, the second-best model, SITUATE exhibits a substantial 22% reduction in ADE and a 1% reduction in FDE.

Similarly, on the Supermarket dataset, SITUATE continues demonstrating its effectiveness with an ADE of 1.21 and an FDE of 1.84. Compared to EqMotion, again the closest competitor, SITUATE achieves a 29% reduction in ADE and a 31% in FDE, reinforcing its dominance.

The consistently good performance of SITUATE across both datasets underscores its robustness and efficacy in trajectory prediction tasks. The results further suggest that SITUATE is accurate and that using scene information when tackling indoor prediction scenarios offers a key advantage compared to the available approaches.

Outdoor Human Trajectory Prediction Results

We also evaluate the performance of SITUATE with the deterministic and multi-prediction modalities with outdoor scenarios. Here, we show that SITUATE achieves competitive results, showing that indoor-oriented forecasting models tend to generalize better than outdoor-oriented ones. Table 3.3 and Table 3.4 present the quantitative results.

Specifically, when considering the deterministic prediction case, SITUATE demonstrates a performance improvement by obtaining state-of-the-art results in both ADE and FDE across the ETH (0.94/1.90), HOTEL(0.30/0.57), and UNIV (0.50/1.10) scenes. It places second in the ZARA1 and ZARA2 scenes while performing on par with EqMotion [27] when considering average performance. In the

Table 3.3 Deterministic prediction performance (ADE (m)/FDE (m)) on the ETH-UCY dataset. The **bold**/underlined font denotes the best/second-best result.

	Performance (ADE (m) ↓ / FDE (m) ↓)					
Deterministic	**ETH**	**HOTEL**	**UNIV**	**ZARA1**	**ZARA2**	**Average**
S-LSTM [79]	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17	0.72/1.54
SGAN-ind [80]	1.13/2.21	1.01/2.18	0.60/1.28	0.42/0.91	0.52/1.11	0.74/1.54
Traj++ [78]	1.02/2.00	0.33/0.62	0.53/1.19	0.44/0.99	0.32/0.73	0.53/1.11
TransF [81]	1.03/2.10	0.36/0.71	0.53/1.32	0.44/1.00	0.34/0.76	0.54/1.17
MemoNet [83]	1.00/2.08	0.35/0.67	0.55/1.19	0.46/1.00	0.37/0.82	0.55/1.15
EqMotion [27]	0.96/1.92	**0.30**/0.58	**0.50/1.10**	**0.39/0.86**	**0.30/0.68**	**0.49/1.03**
SITUATE (ours)	**0.94/1.90**	**0.30/0.57**	**0.50/1.10**	0.41/0.89	0.32/0.70	**0.49/1.03**

Table 3.4 Multi-prediction performance (ADE (m)/FDE (m)) on the ETH-UCY dataset. The **bold**/underlined font denotes the best/second-best result.

	Performance (ADE (m) ↓ / FDE (m) ↓)					
Multi-prediction	**ETH**	**HOTEL**	**UNIV**	**ZARA1**	**ZARA2**	**Average**
SGAN [80]	0.87/1.62	0.67/1.37	0.76/0.52	0.35/0.68	0.42/0.84	0.61/1.21
STGAT [115]	0.65/1.12	0.35/0.66	0.34/0.69	0.29/0.60	0.52/1.10	0.43/0.83
STAR [102]	0.36/0.65	0.17/0.36	0.31/0.62	0.29/0.52	0.22/0.46	0.26/0.53
NMMP [116]	0.61/1.08	0.33/0.63	0.52/1.11	0.32/0.66	0.43/0.85	0.41/0.82
Traj++ [78]	0.61/1.02	0.19/0.28	0.30/0.54	0.24/0.42	0.18/0.31	0.30/0.51
PECNet [103]	0.54/0.87	0.18/0.24	0.35/0.60	0.22/0.39	0.17/0.30	0.29/0.48
Agentformer [117]	0.45/0.75	0.14/0.22	0.25/0.45	0.18/**0.30**	0.14/0.24	0.23/0.39
GroupNet [118]	0.46/0.73	0.15/0.25	0.26/0.49	0.21/0.39	0.17/0.33	0.25/0.44
MID [82]	**0.39**/0.66	0.13/0.22	**0.22**/0.45	**0.17/0.30**	**0.13**/0.27	**0.21**/0.38
GP-Graph [84]	0.43/0.63	0.18/0.30	0.24/**0.42**	**0.17**/0.31	0.15/0.29	0.23/0.39
EqMotion [27]	0.40/**0.61**	**0.12/0.18**	0.23/0.43	0.18/0.32	**0.13/0.23**	**0.21/0.35**
SITUATE (ours)	0.41/0.64	0.13/0.20	0.23/0.43	0.22/0.35	0.14/0.26	0.22/0.37

context of multi-prediction modality, SITUATE secures the second rank in terms of ADE and FDE across nearly 75% of the scenes within the ETH-UCY dataset while maintaining an overall second place in average performance.

Designed for indoor scenarios and their peculiar conformations, SITUATE remarkably demonstrates robust capabilities, even when tested on outdoor datasets. In contrast, this is not always true for architectures tailored for outdoor instances, as we can observe in Table 3.1 and Table 3.2, that often struggle when confronted with scenes that differ from those for which they were designed.

Table 3.5 Ablation results (ADE (m)/FDE (m)) of SITUATE. We assess the contribution of the scene representation module and regularization methods in the deterministic prediction case.

		Performance (ADE (m) ↓ / FDE (m) ↓)		
Scene Representation	**Regularization**	**THÖR**	**Supermarket**	**Average**
✗	✗	0.50/1.02	1.92/1.55	1.21/1.29
✗	✓	0.56/0.74	1.79/2.94	1.18/1.84
✓	✗	0.57/0.96	1.29/1.89	0.93/1.43
✓	✓	**0.45/0.93**	**1.21/1.84**	**0.83/1.38**

Ablation Study

We quantitatively evaluate the impact of the scene representation module and regularization methods by considering the deterministic indoor prediction scenario. Results are summarized in Table 3.5. It is observed that both contributions play a crucial role in enhancing the overall performance of SITUATE. In particular, the scene representation module effectively encodes semantic information from the visual scene maps, facilitating an accurate understanding of the environment. On the other hand, the regularization methods ensure robustness and generalization of the model by mitigating overfitting and improving its ability to generalize to unseen data. Therefore, we assert that both contributions are indispensable for achieving the desired outcomes and validating the efficacy of our approach.

3.3 Discussion

This chapter presents SITUATE, a graph neural network-based model designed specifically to cope with indoor human trajectory prediction. SITUATE, using geometric features and self-supervised vision representations, models the intricate human movements inherent in indoor spaces and accurately predicts users' future locations. The scene vision representation module provides insights about the environment, particularly helping in those indoor scenes that are more constrained and full of obstacles. Thus, the model reaches results comparable to the state-of-the-art, even in conditions with limited access to the environment information, as shown in Table 3.5.

We evaluate our method on two well-known indoor trajectory prediction datasets, *i.e.*, THÖR and Supermarket, and achieve state-of-the-art prediction performance. Furthermore, we also achieve competitive results in outdoor scenarios, showing that indoor-oriented forecasting models generalize better than outdoor-oriented ones.

Chapter 4

3D Time Series Forecasting: Human Pose Forecasting

Collaborative robots, or cobots, are reshaping modern industrial environments by fostering close and seamless Human-Robot Collaboration (HRC). Unlike traditional industrial robots, which operate in segregated workspaces and follow rigidly preprogrammed instructions, cobots share environments with human workers and adapt dynamically to their behavior. This shift, a defining feature of the transition from Industry 4.0 to Industry 5.0, is driven by the need for flexible, intelligent systems capable of real-time decision-making and contextual understanding. As cobots operate in shared workspaces, they must move beyond simple task execution to interpret and predict human actions, ensuring not only efficient collaboration but also safety in environments characterized by physical proximity and direct interaction.

At the heart of this transition is the development of advanced robotic capabilities, such as visual perception, action recognition, intent prediction, and safe motion planning. These capabilities enable cobots to interact more naturally with humans, adapting their behavior in real-time to changing scenarios. However, such advancements come with significant challenges. Cobots must not only detect human actions as they happen but also anticipate future movements, foresee potential collisions, and dynamically adjust their trajectories to maintain safety and efficiency. Addressing these challenges is essential to unlock the full potential of cobots in a wide range of applications.

This thesis contributes to this evolving field by presenting two complementary innovations. The first is a novel method for efficient human pose forecasting designed to improve cobots' predictive capabilities in industrial environments. The second is

a groundbreaking dataset that captures human behavior from a quadruped robot's perspective, enabling new research avenues in collaborative robotics. Together, these contributions tackle critical aspects of HRC, focusing on enabling cobots to predict and respond to human actions with a level of foresight and adaptability previously unattainable.

Industrial Human-Robot Collaboration

Collaborative robots (cobots) and modern Human-Robot Collaboration (HRC) depart from the traditional separation of functions of industrial robots [119], because of the shared workspace [120]. Additionally, cobots and humans perform actions concurrently and they will therefore physically engage in contact. While there is a clear advantage in increased productivity [121] (improved by as much as 85% [122]) due to the minimization of idle times, there are challenges in the workplace safety [123]: it is not about whether there will be contact, but rather about understanding its consequences [124].

The pioneering work of Shah et al. [122] has already shown that, in order to seamlessly and efficiently interact with human co-workers, cobots need to abide by two collaborative principles: (1) Making decisions on the fly and (2) Considering the consequences of their decision on their teammates. The first calls for promptly and accurately detecting human motion in the workspace. The second principle implies that cobots need to anticipate the pose trajectories of their human co-workers and predict future collisions.

Motivated by these problems, the first contribution of our work is a novel Separable-Sparse Graph Convolutional Neural Network (SeS-GCN) for human pose forecasting. Pose forecasting requires an understanding of the complex spatiotemporal joint dynamics of the human body and recent trends have highlighted the promises of modeling body kinematics within a single GCN framework [125, 126, 127, 128, 129, 130, 131]. We have designed SeS-GCN with performance and efficiency in mind, by bringing together, for the first time, three main modeling principles: depthwise-separable graph convolutions [132], space-time separable graph adjacency matrices [29], and sparse graph adjacency matrices [133].

In SeS-GCN, *separable* stands for limiting the interplay of joints with others (space), at different frames (time) and per channel (depth-wise). Within the GCN, different channels, frames, and joints still interact by means of multi-hop messages.

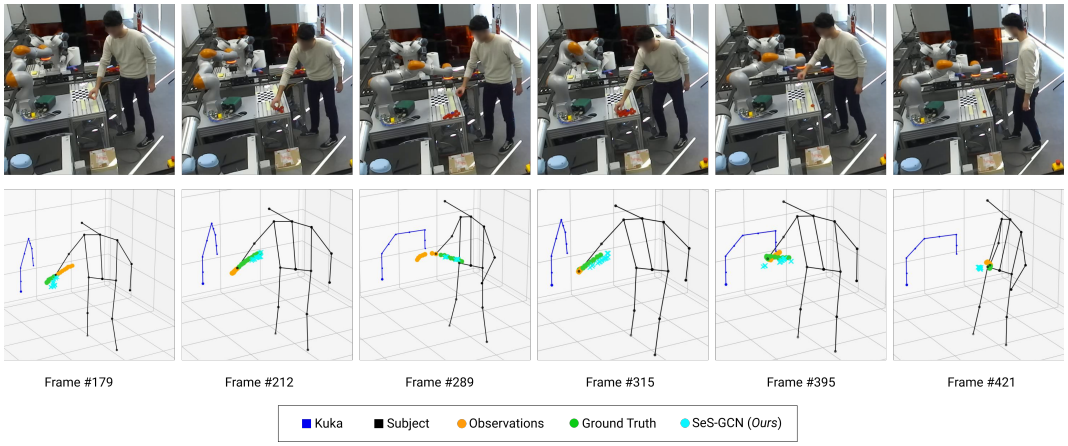

Fig. 4.1 A collision example from our CHICO dataset. On the top row some frames of the *Lightweight pick and place* action captured by one of the three cameras. On the bottom row, operator + robot skeletons. The forecasting takes an observation sequence (in yellow, here pictured for the right wrist only), and performs a prediction (cyan) which is compared with the ground truth (green). On frame 395 is it easy to see the robot hitting the operator, which is retracting, as it is evident in frame 421. See how the predictions by SeS-GCN follow closely the GT, except during the collision. At collision time, due to the impact, the abrupt change of the arm motion produces uncertain predictions, as it shown by the very irregular predicted trajectory.

For the first time, sparsity is achieved by a teacher-student framework. The reduced interaction and sparsity results in comparable or fewer parameters than all GCN-based baselines [132, 29, 133], while improving performance by at least 2.7%.

Compared to the state-of-the-art (SoA) [134], SeS-GCN is lightweight, only using 1.72% of its parameters, it is $\sim$4 times faster while remaining competitive with just 1.5% larger error on Human 3.6M [135] when predicting 1 sec in the future.

The model is described in detail in Section 4.2.1, and experiments and ablation studies are illustrated in Section 4.2.3. We also introduce the very first benchmark of Cobots and Humans in Industrial COllaboration (CHICO, an excerpt in Fig. 4.1). CHICO includes multi-view videos, 3D poses, and trajectories of the joints of 20 human operators, in close collaboration with a robotic arm *KUKA LBR iiwa* within a shared workspace. The dataset features 7 realistic industrial actions, taken at a real industrial assembly line with a marker-less setup.

The goal of CHICO is to endow robots with perceptive awareness to enable human-cobot collaboration with contact. Towards this frontier, CHICO proposes to benchmark two key tasks: human pose forecasting and collision detection. Cobots currently detect collisions by mechanical-only events (transmission of contact

wrenches, control torques, sensitive skins). This ensures safety but it harms the human-cobot interaction because collisions break the motion of both, which reduces productivity and may be annoying to the human operator.

CHICO features 240 1-minute video recordings, from which two separate sets of test sequences are selected: one for estimating the accuracy in pose forecasting, so cobots may be aware of the next future (1.0 sec); and one with 226 genuine collisions, so cobots may foresee them and possibly re-plan. The dataset is detailed in Section 4.2.2, and experiments are illustrated in Section 4.2.4.

When tested on CHICO, the proposed SeS-GCN outperforms all baselines and reaches an error of 85.3 mm (MPJPE) at 1.00 sec, with a negligible run time of 2.3 msec (as reported in Table 4.6). Additionally, the forecast human motion is used to detect human-cobot collision, by checking whether the predicted trajectory of the human body intersects that of the cobot. This is also encouraging, as SeS-GCN allows to reach an F1-score of 0.64. Both aspects contribute to a cobot awareness of the future, which is instrumental for HRC in industrial applications.

Human-Robot Interaction from the Robot's Perspective

One of the main changes characterizing the transition from Industry 4.0 to Industry 5.0 is the shift from Human-Robot *Interaction* to Human-Robot *Collaboration* [136]. This shift requires robots to evolve into cobots—intelligent platforms equipped with capabilities like visual perception, action recognition, intent prediction, and safe online motion planning. These technologies empower cobots with awareness of their surroundings, enabling them to adapt their behavior in real-time, a stark contrast to the rigid, pre-programmed routines of traditional robots [137]. In other words, understanding human behavior is a crucial requirement for a robot to become a cobot and, correspondingly, to be capable of adaptive and seamless interaction with its users [138].

Thus motivated, we propose *Human from an Articulated Robot Perspective* (HARPER), a new, publicly available dataset revolving around the interaction between human users and Spot, the quadruped robot manufactured by Boston Dynamics. Such a platform attracts increasingly more attention, and, not surprisingly, it was recently included in *Habitat 3.0* [139], one of the most advanced environments for simulating Human-Robot interactions. Moreover, Spot is an ideal cobot candidate for at least three reasons: 1) the four-legged design and the biologically inspired locomotion enable it to operate on diverse and challenging terrains (the robot can

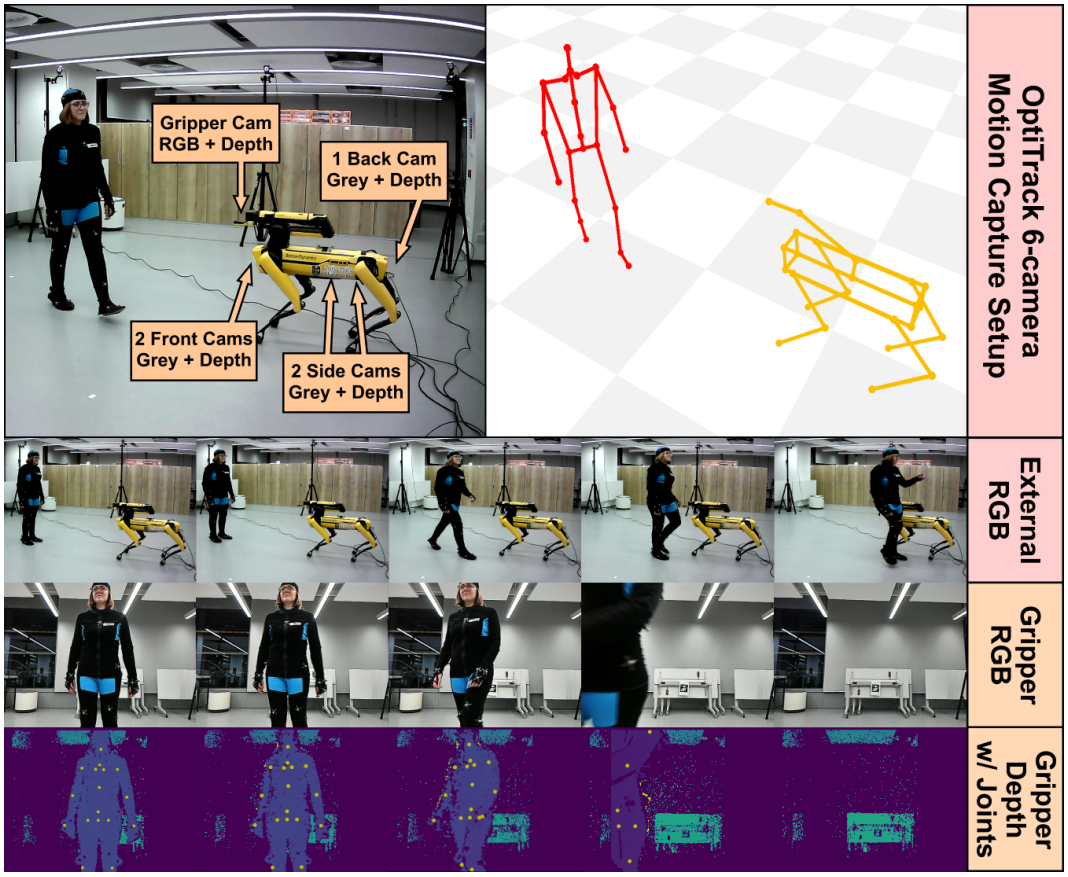

Fig. 4.2 HARPER Showcase. (Top-left) We exploit the Spot on-board equipment to let the robot perceive people. (Top-right) Thanks to a 6-camera Optitrack setup we capture 3D human poses represented with 21-joints and 0.035 mm of error. (Second row) An additional external RGB camera shows the actions performed. (Third row) The gripper RGB camera Point of View. (Fourth row) The gripper depth camera Point of View, with the ground truth joints in yellow.

even climb stairs [140]), thus making Spot a potential companion in a wide range of settings [141, 142, 143, 144]; 2) Spot is equipped with one of the most advanced self-balancing systems available on the market, significantly limits the risk of accidents during close physical interaction with users; and 3) Spot is equipped with a total of 5 greyscale + depth sensors mounted on its body and an RGB-D camera on its grasper arm (see Fig. 4.2). This is important because such a sensing apparatus makes Spot particularly suitable for analysis and understanding of human behavior, a critical step in the evolution from robot to cobot.

HARPER includes dyadic interactions between Spot and 17 human users, 5 females and 12 males, each performing 15 actions that require different degrees of

collaboration with the robot (see Section 4.1 for more details). The data captured with the Spot sensors were enriched with the recordings of a 6-camera Optitrack Motion Capture (MoCap) system capable of extracting users' skeletal models. The joints were localized with a precision of less than one millimeter (see Fig. 4.2), thus providing highly accurate ground-truth information about the pose and position of the users. This is a significant advantage because Spot's sensors and MoCap cameras are synchronized. Therefore, skeletal models can be used to reliably validate approaches for human behavior analysis and understanding based solely on Spot sensors.

In addition to the above, skeletal representations enable one of the key novelties of HARPER, namely the possibility to train approaches capable of recognizing 3D body pose and movement when the Spot, due to its limited height, can "see" its users only partially, something that happens whenever the distance between them is small. To the best of our knowledge, this is one of the first datasets that allows the investigation of such a problem in 3D.

We asked the 17 HARPER participants to stage two types of physical contact with the robot, namely *unintentional* and *intentional*. The first type includes (staged) collisions, while the second includes punches, kicks and soft touches. We paid special attention to the first type because of the major role collisions play in scenarios based on co-located interactions. We enriched HARPER with benchmarks, *i.e.*, reproducible experimental protocols, and baseline approaches designed to address three tasks relevant to the analysis of physical contact, namely 3D Human Pose Estimation (especially when Spot can "see" its users only partially), 3D Human Pose Forecasting and Collision Prediction. This allows researchers interested in HARPER to rigorously compare their results with those presented in this article (see Section 4.3.2).

Overall, the main contributions of this method can be summarized as follows:

- We propose the first dataset that includes not only the "point of view" of the robot (the data captured with the Spot's sensors) but also a panoptic point of view (the data captured with the MoCap system) that provides accurate ground-truth information for position and pose of both users and robot;
- To the best of our knowledge, HARPER is the first dataset enabling the reconstruction of the human users' pose with the data captured with a quadruped robot, a problem which is challenging because Spot is short and the subject is usually close to it (hence, the cameras cannot capture the whole body of the user);

- HARPER allows, for the first time, the visual prediction of collisions between a mobile robotic platform and users.

4.1 Related Works

Human Pose Forecasting

Human pose forecasting is a recent field that has some intersection with human action anticipation in computer vision [127] and HRC [145]. Previous studies exploited Temporal Convolutional Networks (TCNs) [146, 147, 148, 149, 150, 151] and Recurrent Neural Networks (RNNs) [152, 153, 154]. Both architectures are naturally suited to model the temporal dimension. Recent works have expanded the range of available methods by using Variational Auto-Encoders [155], specific and model-agnostic layers that implicitly model the spatial structure of the human skeleton [156], or Transformer Networks [157].

Human Pose Forecasting using Graph Convolutional Networks (GCN)

Most recent research uses GCNs [126, 128, 134, 29, 131]. In [134], the authors have mixed GCN for modeling the joint-joint interaction with Transformer Networks for the temporal patterns. Others [128, 29, 131] have adopted GCNs to model the space-time body kinematics, devising, in the case of [126], hierarchical architectures to model coarse-to-fine body dynamics.
We identify three main research directions for improving efficiency in GCNs: **i.** space-time separable GCNs [29], which factorizes the spatial joint-joint and temporal patterns of the adjacency matrix; **ii.** depth-wise separable graph convolutions [158], which has been explored by [159] in the spectral domain; and **iii.** sparse GCNs [133], which iteratively prunes the terms of the adjacency matrix of a GCN. Notably, all three techniques also yield better performance than the plain GCN. Here, for the first, we bring together these three aspects into an end-to-end space-time-depthwise-separable and sparse GCN. The three techniques are complementary to improve both efficiency and performance, but their integration requires some structural changes (*e.g.*, adopting teacher-student architectures for sparsifying), as we describe in Section 4.2.1.

Table 4.1 Comparison between the state-of-the-art datasets and the proposed CHICO; *unk* stands for "unknown".

	Quantitative Details							Rec. Scene	Actions Type		Tasks			Markerless
	# Classes	# Subj.	Avg Rec. Time	# Joints	FPS	Aspect Ratio	# Sensors		Industr.	HRC	Action Recog.	Pose Forec.	Coll. Det.	
Human3.6M [135]	15	11	100.49 s	32	25	normalized	15	mo-cap studio				✓		
AMASS [160]	11265	344	12.89 s	variable	variable	original	variable	mo-cap studio				✓		
3DPW [161]	47	7	28.33 s	18	60	original	18	outdoor locations				✓		
ExPI [162]	16	4	*unk*	18	25	original	88	mo-cap studio				✓		
CHI3D [163]	8	6	*unk*	*unk*	*unk*	original	14	mo-cap studio				✓		
InHARD [164]	14	16	< 8 s	17	120	original	20	assembly line	✓	✓	✓			
CHICO (ours)	7	20	55 s	15	25	original	3	assembly line	✓	✓		✓	✓	✓

Datasets for Human Pose Forecasting

Human pose forecasting datasets cover a wide spectrum of scenarios, see Table 4.1 for a comparative analysis. Human3.6M [135] considers everyday actions such as conversing, eating, greeting, and smoking. Data were acquired using a 3D marker-based motion capture system, composed of 10 high-speed infrared cameras. AMASS [160] is a collection of 15 datasets where daily actions were captured by an optical marker-based motion capture. Human3.6M and AMASS are standard benchmarks for human pose forecasting, with some overlap in the type of actions they deal with. The 3DPW dataset [161] focuses on outdoor actions, captured with a moving camera and 17 Inertial Measurement Units (IMU), embedded on a special suit for motion capturing [165]. The recent ExPI dataset [162] contains 16 different dance actions performed by professional dancers, for a total of 115 sequences, and it is aimed at motion prediction. ExPI has been acquired with 68 synchronized and calibrated color cameras and a motion capture system with 20 mocap cameras. Finally, the CHI3D dataset [163] reports 3D data taken from MOCAP systems to study human interactions.

None of these datasets answer our research needs, i.e., a benchmark taken by a sparing, energy-efficient markerless system, focused on the industrial HRC scenario, where forecasting may be really useful for anticipating collisions between humans and robots. In fact, the only dataset relating to industrial applications is InHARD [164]. Therein, humans are asked to perform an assembly task while wearing inertial sensors on each limb. The dataset is designed for human action recognition, and it involves 16 individuals performing 13 different actions each, for a total of 4800 action samples over more than 2 million frames. Despite showcasing

Table 4.2 Main HRI datasets revolving around human movement and its analysis. Values in the Participants column indicated with the asterisk (*) refer to datasets captured in uncontrolled scenarios.

Dataset	Participants	Actions	Mobile Robot	Robot POV	Human Skeleton	Human Joints	Marker-Based MoCap	Robot Skeleton	Collisions / Intended Contacts
THÖR [181]	9	13	✓	✗	✗	✗	✓	✗	✗
THÖR-Magni [182]	40	48	✓	✓	✗	✗	✓	✗	✗
JRDB [183]	3.5K*	N/A	✓	✓	✗	✗	✗	✗	✗
L-CAS Multisensor [184]	N/A*	N/A	✓	✓	✗	✗	✗	✗	✗
FROG [185]	1M*	N/A	✓	✓	✗	✗	✗	✗	✗
CODa [186]	N/A*	N/A	✓	✓	✗	✗	✗	✗	✗
PTUA [187]	N/A	N/A	✓	✓	✗	✗	✓	✗	✗
InHARD [188]	16	14	✗	✗	3D	17	✗	✗	✗
JRDB-Pose [189]	5K*	N/A	✓	✓	2D	17	✗	✗	✗
HuRoN [190]	N/A* (5/17 for exp)	N/A	✓	✓	✗	✗	✗	✗	✓
NatSGD [191]	18	11	✗	✓	estim. 2D	25	✗	Arm	✗
CHICO [16]	20	7	✗	✗	2D, 3D	15	✗	Arm	✓
SCAND [192]	N/A* (14 for exp)	12	✓	✓	✗	✗	✗	Quadruped, Wheeled	✗
UF-Retail-HRI [193]	8	2	✓	✓	3D	23	✗	Arm	✗
HARPER	17	15	✓	✓	2D, 3D	21	✓	Quadruped	✓

a collaborative robot, in this dataset the robot is mostly static, making it unsuitable for collision forecasting.

Human-Robot Collaboration (HRC)

HRC is the study of collaborative processes where human and robot agents work together to achieve shared goals [166, 167]. Computer vision studies on HRC are mostly related to pose estimation [73, 168, 169] to locate the articulated human body in the scene.

In [170, 171, 172], methodologies for robot motion planning and collision avoidance are proposed; their study perspective is opposite to ours since we focus on the human operator. In this regard, the works of [173, 174, 175, 176] model the operators' whereabouts through detection algorithms that approximate human shapes using simple bounding boxes. Approaches that predict human motion during collaborative tasks are in [177, 178] using RNNs and in [179] using Gaussian processes. Other work [180] models the upper body and the human right hand (which they call the Human End Effector) by considering the robot-human handover phase. As a motion prediction engine, DCT-RNN-GCN [134] is considered, against which we compare in the experiments.

Dataset for Human-Robot Collaboration

Table 4.2 shows the main differences between HARPER and existing datasets of similar scope. Most available corpora are based on the analysis of people's trajectories. The THÖR dataset [181], a well-known example, contains the 2D trajectories of 9 human users moving together with a robot. Besides this, the data includes 6D head positions, LiDAR data from a stationary sensor, and the participants' orientations and eye gaze directions. THÖR-Magni [182], a significantly more extensive dataset from the authors of THÖR, introduces onboard sensors on the mobile robot and semantic attributes describing the roles and activities of detected people. Similarly, the JRDB [183] dataset aims to enable mobile robots to detect and track humans in both indoor and outdoor settings. The data includes stereo cylindrical RGB videos and LiDAR point clouds annotated with 2D and 3D bounding boxes. In addition, the dataset includes benchmarks for both 2D and 3D detection and tracking. A more recent version includes 2D human-pose skeletal annotations [189].

Other datasets provide information about objects that the robots can encounter while moving. For example, CODa [186] aims at both object detection and semantic segmentation. It was acquired with a wheeled robot, and it features sequences in indoor and outdoor settings of a university campus, as well as 3D semantic segmentation and 3D object detection benchmarks. In the case of FROG [185], based on LiDAR sensors placed on a wheeled robot at roughly the height of human knees, the problem is the detection of people in possibly crowded sites where humans can be confused with static and dynamic obstacles. A similar issue is at the core of the dataset proposed in [187], where the material is collected with an RGB-D camera mounted on a small mobile robot. The annotations include attributes such as *e.g.*, the presence of static obstacles, illumination, and humans' poses. An Optitrack MoCap system provides information about the position of both the robot and its users. The problem of navigating through an environment, possibly shared with humans, is the focus of HuRon [190]. The data was collected using a Roomba bot with LiDAR, bumper collision detectors, video, and odometry sensors. However, no pose annotation is provided about the people sharing the space with the robot.

HARPER shows major novelties with respect to the datasets above. The availability of 3D skeletons for both the robot and users provides unprecedentedly detailed information about the interaction between the two, especially when considering that the joints are captured with a precision of less than one millimeter. A similar acquisition precision is achieved with InHARD [188], an industrial HRI dataset featuring

both RGB images and MoCap data of a person performing multiple manual tasks, captured with wearable devices. A robotic arm, mostly stationary, is the platform used for the experiments, and it never collides with the user, offering a looser type of interaction. This is not the case in HARPER, which includes physical contacts of different types. In [193], a mobile wheeled robot is employed to capture an HRI dataset in a retail environment. Multiple people navigate the room and perform picking and sorting actions while the robot moves along with them. Egocentric and scene videos, eye gaze directions, point clouds, and other data are collected. The human poses are collected through an IMU-based MoCap device, which requires careful setup and calibration for every person. However, the Spot used in HARPER is a more advanced robotic platform, and its movement is significantly less constrained.

Skeleton representations were also used in other corpora. In [16], the scenario is a collaboration between a user and a robotic arm in an industrial setting. A MoCap system captures the user's skeleton from an external point of view, missing the robot's perspective. Furthermore, the acquisition is markerless, making the joint localisation less precise. In another dataset, the multimodal NatSGD [191], the goal is imitation learning, and the data includes human commands, such as speech and gestures, with a focus on robot behavior in the form of synchronized robot trajectories. However, the joint localization is, once again, less precise than in HARPER because it is performed by applying Openpose [73] to videos.

Finally, to the best of our knowledge, the only other dataset in which the robot Spot was actually involved is SCAND [192], where two robots, a wheeled one and the Spot, are teleoperated in human-populated environments. A large variety of data is acquired thanks to an additional LiDAR sensor mounted on the two robots. However, no skeletal models are considered for humans, a major difference with respect to HARPER. The dataset we propose appears to have distinctive characteristics with respect to those currently available in the literature.

4.2 CHICO

4.2.1 Problem Formalization

We build an accurate, memory efficient and fast GCN by bridging three diverse research directions: **i.** Space-time separable adjacency matrices; **ii.** Depth-wise separable graph convolutions; **iii.** Sparse adjacency matrices. This results in an all-

separable and Sparse GCN encoder for the human body kinematics, which we dub SeS-GCN, from which the future frames are forecast by a Temporal Convolutional Network (TCN).

Problem Formalization

Pose forecasting is formulated as observing the 3D coordinates $x_{v,t}$ of V joints across T frames and predicting their location in the K future frames. For convenience of notation, we gather the coordinates from all joints at frame t into the matrix $X_t = [x_{v,t}]_{v=1}^{V} \in \mathbb{R}^{3\times V}$. Then we define the tensors $\mathscr{X}_{in} = [X_1, X_2 ..., X_T]$ and $\mathscr{X}_{out} = [X_{T+1}, X_{T+2} ..., X_{T+K}]$ that contain all observed input and target frames, respectively.

We consider a graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ to encode the body kinematics, with all joints at all observed frames as the node set $\mathscr{V} = \{v_{i,t}\}_{i=1,t=1}^{V,T}$, and edges $(v_{i,t}, v_{j,s}) \in \mathscr{E}$ that connect joints i, j at frames t, s.

Graph Convolutional Networks (GCN)

A GCN is a layered architecture:

$$\mathscr{X}^{(l+1)} = \sigma\left(A^{(l)}\mathscr{X}^{(l)}W^{(l)}\right), \tag{4.1}$$

The input to a GCN layer l is the tensor $\mathscr{X}^{(l)} \in \mathbb{R}^{C^{(l)}\times V\times T}$ which maintains the correspondence to the V body joints and the T observed frames, but increases the depth of features to $C^{(l)}$ channels. $\mathscr{X}^{(1)} = \mathscr{X}_{in}$ is the input tensor at the first layer, with $C^{(1)} = 3$ channels given by the 3D coordinates. $A^{(l)} \in \mathbb{R}^{VT\times VT}$ is the adjacency matrix relating pairs of VT joints from all frames. Following most recent literature [126, 134, 133, 29], $A^{(l)}$ is learnt. $W^{(l)} \in \mathbb{R}^{C^{(l)}\times 1\times 1}$ are the learnable weights of the graph convolutions. σ is the non-linear PReLU activation function.

Separable & Sparse Graph Convolutional Networks (SeS-GCN)

We build SeS-GCN by integrating the three mentioned modeling dimensions: **i.** separating spatial and temporal interaction terms in the adjacency matrix of a GCN; **ii.** separating the graph convolutions depth-wise; **iii.** sparsifying the adjacency matrices of the GCN.

STS-GCN [29] has factored the adjacency matrix $A^{(l)}$ of the GCN, at each layer l, into the product of two terms $A_s^{(l)} \in \mathbb{R}^{V\times V\times T}$ and $A_t^{(l)} \in \mathbb{R}^{T\times T\times V}$, respectively

responsible for the temporal-temporal and joint-joint relations. The GCN formulation becomes:

$$\mathscr{X}^{(l+1)} = \sigma\left(A_s^{(l)} A_t^{(l)} \mathscr{X}^{(l)} W^{(l)}\right) , \tag{4.2}$$

Eq. (4.2) bottlenecks the interplay of joints across different frames, implicitly placing more emphasis on the interaction of joints on the same frame ($A_s^{(l)}$) and on the temporal pattern of each joint ($A_t^{(l)}$). This reduces the memory footprint of a GCN by approx. 4x while improving its performance. Note that this differs from alternating spatial and temporal modules, as it is done in [194] and [195], respectively for trajectory forecasting and action recognition.

Inspired by depth-wise convolutions [196, 158], the approach in [132] has introduced depth-wise graph convolutions for image classification, followed by [159] which resorted to a spectral formulation of depth-wise graph convolutions for graph classification. Here we consider depth-wise graph convolutions for pose forecasting. The depth-wise formulation bottlenecks the interplay of space and time (operated by the adjacency matrix $A^{(l)}$) with the channels of the graph convolution $W^{(l)}$. The resulting all-separable model which we dub STS-DW-GCN is formulated as such:

$$\mathscr{H}^{(l)} = \gamma\left(A_s^{(l)} A_t^{(l)} \mathscr{X}^{(l)} W_{DW}^{(l)}\right) , \tag{4.3a}$$

$$\mathscr{X}^{(l+1)} = \sigma\left(\mathscr{H}^{(l)} W_{\text{MLP}}^{(l)}\right) , \tag{4.3b}$$

Adding the depth-wise graph convolution splits the GCN of layer l into two terms. The first, Eq. (4.3a), focuses on space-time interaction and limits the channel cross-talk by the use of $W_{DW}^{(l)} \in \mathbb{R}^{\frac{C^{(l)}}{\alpha} \times 1 \times 1}$, with $1 \leq \alpha \leq C^{(l)}$ setting the number of convolutional groups ($\alpha = C^{(l)}$ is the plain single-group depth-wise convolution). The second, Eq. (4.3b), models the intra-channel communication just. This may be understood as a plain (MLP) 1D-convolution with $W_{\text{MLP}}^{(l)} \in \mathbb{R}^{C^{(l)} \times 1 \times 1}$ which re-maps features from $C^{(l)}$ to $C^{(l+1)}$. γ is the ReLU6 non-linear activation function. Overall, this does not significantly reduce the number of parameters, but it deepens the GCN without over-smoothing [197], which improves performance.

Sparsifying the GCN.

Sparsification has been used to improve the efficiency (memory and, in some cases, runtime) of neural networks since the seminal pruning work of [198]. [133] has sparsified GCNs for trajectory forecasting. This consists of learning masks $\mathscr{M}$ which

selectively erase certain parameters in the adjacency matrix of the GCN. Here we integrate sparsification with the all-separable GCN design, which yields our proposed SeS-GCN for human pose forecasting:

$$\mathscr{H}^{(l)} = \gamma\Big((\mathscr{M}_s^{(l)} \odot A_s^{(l)})(\mathscr{M}_t^{(l)} \odot A_t^{(l)})\mathscr{X}^{(l)}W_{DW}^{(l)}\Big)\ , \tag{4.4a}$$

$$\mathscr{X}^{(l+1)} = \sigma\left(\mathscr{H}^{(l)}W_{\text{MLP}}^{(l)}\right)\ , \tag{4.4b}$$

$\odot$ is the element-wise product and $\mathscr{M}_{\{s,t\}}^{(l)}$ are binary masks. Both at training and inference, [133] generates masks, it uses those to zero certain coefficients of the adjacency matrix A, and it adopts the resulting GCN for trajectory forecasting. By contrast, we adopt a teacher-student framework during training. The teacher learns the masks, and the student only considers the spared coefficients in A. At inference, our proposed SeS-GCN only consists of the student, which simply adopts the learned sparse A_s and A_t. Compared to [133], the approach of SeS-GCN is more robust at training, it yields fewer model parameters at inference ($\sim$30% less for both A_s and A_t), and it reaches a better performance.

Decoder Forecasting

Given the space-time representation, as encoded by the SeS-GCN, the future frames are then decoded by using a temporal convolutional network (TCN) [146, 147, 148, 29]. The TCN remaps the temporal dimension to match the sought output number of predicted frames. This part of the model is not considered for improvement because it is already efficient and performs satisfactorily.

4.2.2 The CHICO dataset

In this section, the CHICO dataset is detailed by describing the acquisition scenario and devices, the cobot, and the performed actions. We release RGB videos, skeletons, and calibration parameters.

The Scenario

We are in a smart-factory environment, with a single human operator standing in front of a $0.9\,\text{m} \times 0.6\,\text{m}$ workbench and a cobot at its end (see Fig.4.1). The human operator has some free space to turn towards some equipment and carry out certain

assembly, loading and unloading actions [199]. In particular, light plastic pieces and heavy tiles, a hammer, abrasive sponges are available. The detailed setups for each action are reported graphically in the additional material. A total of 20 human operators have been hired for this study. They attended a course on how to operate with the cobot and signed an informed consent form prior to the recordings.

The Collaborative Robot

A 7 degrees-of-freedom Kuka LBR iiwa 14 R820 collaborates with the human operator during the data acquisition process. Weighing in at 29.5 kg and with the ability to handle a payload up to 14 kg, it is widely used in modern production lines.

The Acquisition Setup

The acquisition system is based on three RGB HD cameras providing three different viewpoints of the same workplace: two frontal-lateral, and one rear view. The frame rate is 25Hz. The videos were first checked for erroneous or spurious frames, then we used Voxelpose [200] to extract 3D human pose for each frame. Extrinsic parameters of each camera are estimated w.r.t. the robot's reference frame by means of a calibration chessboard of 1×1m, and temporal alignment is guaranteed by synchronization of all the components with an Internet Time Server. In our environment, Voxelpose estimates a joint positioning accuracy in terms of Mean Per Joint Position Error (MPJPE) of 24.99mm using three cameras, which is enough for our purposes, as an ideal compromise between the portability of the system and accuracy. We confirm these numbers in two ways: the first is by checking that human-cobot collisions were detected with a 100% F1 score (we have a collision when the minimum distance between the human limbs and the robotic links is below a predefined threshold). Secondly, we show that the new CHICO dataset does not suffer from a trivial zero velocity solution [201], *i.e.* results achieved by a zero velocity model underperform the current SoA in equal proportion as for the large-scale established Human3.6M.

Actions

The 7 types of actions of CHICO are inspired by ordinary work sessions in an HRC disassembly line as described in the review work of [202]. Each action is repeated over a time interval of $\sim$1 minute on average. Each action is associated with a goal

that the human operator has to achieve by a given time limit, which requires them to move with a certain velocity. Each action consists of repeated interactions with the robot (*e.g.*, robot place, human picks) which, due to the limited space, lead to some *unconstrained collisions*[1] which we label accordingly. Globally, from the 7 actions $\times$ 20 operators, we collect 226 different collisions. In the following, each action is briefly described.

- ***Lightweight pick and place*** (*Light P&P*). The human operator is required to move small objects of approximately 50 grams from a loading bay to a delivery location within a given time slot. The bay and the delivery location are at the opposite sides of the workbench. Meanwhile, the robot loads on of this bay so that the human operator has to pass close to the robotic arm. In many cases, the distance between the limbs and the robotic arm is a few centimeters.

- ***Heavyweight pick and place*** (*Heavy P&P*). The setup of this action is the same as before, but the objects to be moved are floor tiles weighing 0.75 kg. This means that the actions have to be carried out with two hands.

- ***Surface polishing*** (*Polishing*). This action was inspired by [204], where the human operator polishes the border of a 40 by 60cm tile with some abrasive sponge, and the robot mimics a visual quality inspection.

- ***Precision pick and place*** (*Prec. P&P*). The robot places four plastic pieces in the four corners of a 30$\times$30cm table in the center of the workbench, and the human has to remove them and put them on a bay before the robot repeats the same unloading.

- ***Random pick and place*** (*Rnd. P&P*). Same as the previous action, except for the plastic pieces which were continuously placed by the robot randomly on the central 30$\times$30cm table, and the human operator had to remove them.

- ***High shelf lifting*** (*High lift*). The goal was to pick light plastic pieces (50 grams each) on a sideway bay filled by the robot, putting them on a shelf located at 1.70m, on the opposite side of the workbench. Due to the geometry of the workspace, the arms of the human operator were required to pass above or below the moving robotic arm. In this way, close distances between the human arm and forearm and the robotic links were realized.

[1] Unconstrained collisions is a term coming from [203], indicating a situation in which only the robot and human are directly involved in the collision.

Table 4.3 MPJPE error (millimeters) for long-range predictions (25 frames) on Human3.6M [135] and numbers of parameters. Best figures overall are reported in bold, while underlined figures represent the best in each block. The proposed model has comparable or less parameters than the GCN-based baselines [158, 133, 29] and it outperforms the best of them [29] by 2.6%.

	Depth	MPJPE	Parameters (K)	DW-Separable	ST-Separable	Sparse	w/ MLP layers	Teacher-Student
GCN	4	123.2	222.7					
DW-GCN [132]	4+4	119.8	223.2	✓			✓	
STS-GCN^2 [29]	4	<u>117.0</u>	**<u>57.6</u>**		✓			
Sparse-GCN [133]	4	122.7	257.9			✓		
STS-GCN	5	115.9	<u>68.6</u>		✓			
STS-GCN	6	116.1	79.9		✓			
STS-GCN w/ MLP	5+5	125.2	101.4		✓		✓	
STS-DW-GCN	5+5	<u>114.8</u>	70.0	✓	✓		✓	
STS-DW-Sparse-GCN	5+5	115.7	122.4	✓	✓	✓	✓	
SeS-GCN (proposed)	5+5	**<u>113.9</u>**	<u>58.6</u>	✓	✓	✓	✓	✓

- ***Hammering*** (*Hammer*). The operator hits with a hammer a metallic tie held by the robot. In this case, the interest was to check how much the collision detection is robust to an action where the human arm is colliding close to the robotic arm (that is, on the metallic tile) without properly colliding *with the robotic arm*.

4.2.3 Experimental Results on Human3.6M

We benchmark the proposed SeS-GCN model on the large and established Human3.6M [135]. In this section, we analyze the design choices corresponding to the models discussed in Section 4.2.1, then we compare with the state-of-the-art in Section 4.2.3.

Human3.6M

[135] is an established dataset for pose forecasting, consisting of 15 daily life actions (e.g. Walking, Eating, Sitting Down). From the original skeleton of 32 joints, 22 are sampled as the task, representing the body kinematics. A total of 3.6 million poses are captured at 25 fps. In line with the literature [134, 201, 126], subjects 1, 6, 7, 8, 9 are used for training, subject 11 for validation, and subject 5 for testing.

Metric

The prediction error is quantified via the MPJPE error metric [135, 129], which considers the displacement of the predicted 3D coordinates w.r.t. the ground truth, in millimeters, at a specific future frame t:

$$L_{\text{MPJPE}} = \frac{1}{V}\sum_{v=1}^{V} ||\hat{x}_{vt} - x_{vt}||_2 \,. \tag{4.5}$$

Efficient GCN Baselines

In Table 4.3, we first validate the three different modeling approaches to efficient GCNs, namely space-time separable STS-GCN [29], depth-wise separable graph convolutions DW-GCN [132], and Sparse-GCN [133]. STS-GCN yields the lowest MPJPE error of 117.0 mm at a 1-sec forecasting horizon (2.4% better than DW-GCN, 4.8% better than Sparse-GCN) with the fewest parameters, 57.6k (ca. x4 less). We build therefore on this approach.

Deeper GCNs

It is a long-standing belief that Deep Neural Networks (DNN) owe their performance to depth [62, 205, 206, 207]. However, deeper models require more parameters and have a longer processing time. Additionally, deeper GCNs may suffer from over-smoothing [197]. Seeking both better accuracy and efficiency, we consider three pathways for improvement: (1) add GCN layers; (2) add MLP layers between layers of GCNs; (3) adopt depth-wise graph convolutions, which also add MLP layers between GCN ones (cf. Section 4.2.1).

As shown in Table 4.3, there is a slight improvement in performance with 5 STS-GCN layers (MPJPE of 115.9 mm), but deeper models underperform. Adding MLP layers between the GCN ones (depth of 5+5) also decreases performance (MPJPE of 125.2). By contrast, adding depth by depth-wise separable graph convolutions (STS-DW-GCN of depth 5+5) reduces the error to 114.8 mm. This may be explained by the virtues of the increased depth in combination with the limiting cross-talk of joint-time channels, which existing literature confirms [196, 132, 29]. We note that space-time and depth-wise channel separability are complementary. Altogether, this performance is beyond the STS-GCN performance (114.8 Vs. 117.0 mm), at a slight increase of the parameter count (70k Vs. 57.6k).

Sparsifying GCNs and the Proposed SeS-GCN

Finally, we aim to improve efficiency by model compression. Trends have reduced the size of models by reducing the parameter precision [208], by pruning and sparsifying some of the parameters [209], or by constructing teacher-student frameworks,

whereby a smaller student model is paired with a larger teacher to reach its same performance [210, 211]. Note that the last technique is the current go-to choice in deploying very large networks such as Transformers [212].

We start off by compressing the model with sparse adjacency matrices following the approach of Sparse-GCN [133]. We combined this process with a teacher-student approach, pruning the teacher graph $(\mathscr{M}_t^{(l)} \odot A_t^{(l)})$ during the student training process, distilling the result of $(\mathscr{M}_t^{(l)} \odot A_t^{(l)})$ directly into $A_s^{(l)}$, as described in Section 4.2.1. As illustrated in Table 4.3, the approach of [133] does not make a viable direction (STS-DW-Sparse-GCN), since the error increases to 115.7 mm and the parameter count to 122.4k.

Reminiscent of teacher-student models, in the proposed SeS-GCN we first train a teacher STS-DW-GCN, then use its learned parameters to sparsify the affinity matrices of a student STS-DW-GCN, which is then trained from scratch. SeS-GCN achieves a competitive parameter count and the lowest MPJPE error of 113.9 mm, being comparable with the current SoA [134] and using only 1.72% of its parameters (58.6k Vs. 3.4M).

Comparison with the State-of-the-art (SOTA)

In Table 4.4, we evaluate the proposed SeS-GCN against the three most recent techniques, over a short time horizon (10 frames, 400 msec) and a long time horizon (25 frames, 1000 msec). The first, DCT-RNN-GCN [134], the current SoA, uses DCT encoding, motion attention and RNNs and, differently from other models, demands more frames as input (50 vs. 10). The other two, MSR-GCN [126] and STS-GCN [29] adopt GCN-only frameworks, the former adopts a multi-scale approach, the latter acts a separation between spatial and temporal encoding.

Both on Short- and long-term predictions, at the 400 and 1000 msec horizons, the proposed SeS-GCN outperforms other techniques [29, 134] and it is within a 1.5% error w.r.t. the current SoA [134], while only using 1.72% parameters and being $\sim$4 times faster than [134].

4.2.4 Experimental Results on CHICO

We benchmark on CHICO the SoA and the proposed SeS-GCN model. The two HRC tasks of human pose forecasting and collision detection are discussed in Section 4.2.4 and Section 4.2.4 respectively.

Table 4.4 MPJPE error in mm for short-term (400 msec, 10 frames) and long-term (1000 msec, 25 frames) predictions of 3D joint positions on Human 3.6M. The proposed model achieves competitive performance with the SoA [134] while adopting 1.72% of its parameters and running ∼4 times faster, cf. Table 4.6. Results are discussed in Section 4.2.3.

	Walking		Eating		Smoking		Discussion		Directions		Greeting		Phoning		Posing	
Time Horizon (msec)	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*
DCT-RNN-GCN [134]	**39.8**	**58.1**	**36.2**	**75.5**	**36.4**	**69.5**	**65.4**	119.8	56.5	106.5	**78.1**	138.8	**49.2**	105.0	**75.8**	178.2
MSR-GCN [126]	45.2	63.0	40.4	77.1	38.1	71.6	69.7	117.5	**53.8**	**100.5**	93.3	147.2	51.2	**104.3**	85.0	174.3
STS-GCN^2 [29]	51.0	70.2	43.3	82.6	42.3	76.1	71.9	118.9	63.2	109.6	86.4	**136.1**	53.8	108.3	84.7	178.4
SeS-GCN (proposed)	48.8	67.3	41.7	78.1	40.8	73.7	70.6	**116.7**	60.3	106.9	83.8	137.2	52.6	106.7	82.6	**173.5**

	Purchases		Sitting		Sitting Down		Taking Photo		Waiting		Walking Dog		Walking Together		Average	
Time Horizon (msec)	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*
DCT-RNN-GCN [134]	**73.9**	**134.2**	**56.0**	**115.9**	**72.0**	**143.6**	**51.5**	**115.9**	**54.9**	108.2	**86.3**	**146.9**	**41.9**	**64.9**	**58.3**	**112.1**
MSR-GCN [126]	79.6	139.1	57.8	120.0	76.8	155.4	56.3	121.8	59.2	**106.2**	93.3	148.2	43.8	65.9	62.9	114.1
STS-GCN^2 [29]	83.1	141.0	60.8	121.4	79.4	148.4	59.4	126.3	62.0	113.6	97.3	151.5	49.1	72.5	65.8	117.0
SeS-GCN (proposed)	82.2	139.1	59.9	117.5	78.1	146.0	57.7	121.2	58.5	107.5	94.0	147.7	48.3	70.8	64.0	113.9

Table 4.5 MPJPE error in mm for short-term (400 msec, 10 frames) and long-term (1000 msec, 25 frames) prediction of 3D joint positions on CHICO dataset. The average error is 7.9% lower than the other models in the short-term and 2.4% lower in the long-term prediction.

	Hammer		High Lift		Prec. P&P		Rnd. P&P		Polishing		Heavy P&P		Light P&P		Average	
Time Horizon (msec)	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*
DCT-RNN-GCN [134]	41.1	**39.0**	69.4	128.8	50.6	83.3	52.7	88.2	42.1	76.0	64.1	121.5	62.1	104.2	54.6	91.6
MSR-GCN [126]	41.6	39.7	67.8	130.2	50.2	81.3	53.4	90.3	41.1	73.2	62.7	118.2	61.5	101.9	54.1	90.7
STS-GCN [29]	46.6	52.1	64.2	116.4	48.3	79.5	52.0	87.9	42.1	73.9	60.6	106.5	57.2	95.2	53.0	87.4
SeS-GCN (proposed)	**40.9**	49.3	**62.1**	**116.3**	**46.0**	**77.4**	**48.4**	**84.8**	**38.8**	**72.4**	**56.1**	**104.4**	**56.2**	**92.2**	**48.8**	**85.3**

Pose Forecasting Benchmark

Here we describe the evaluation protocol proposed for CHICO and report a comparative evaluation of pose forecasting techniques. We create the train/validation/test split by assigning 2 subjects to the validation (subjects 0 and 4), 4 to the test set (subjects 2, 3, 18, and 19), and the remaining 14 to the training set.

For short-range prediction experiments, abiding by the setup of Human3.6M [135], we consider 10 frames as observation time and 10 or 25 frames as forecasting horizon. Different from all reported techniques, DCT-RNN-GCN requires 50 input frames. We adopt the same Mean Per Joint Position Error (MPJPE) [135] as Human3.6M, in Eq. (4.5), which also defines the training loss for the evaluated techniques.

None of the motion sequences for pose forecasting contain collisions. In fact, the objective is to train and test the "correct" collaborative human behavior, and not the human retractions and the pauses due to the collisions[7].

[7]After the collisions, the robot stops for 1 seconds, during which the human operator usually stands still, waiting for the robot to resume operations.

Comparative Evaluation

In Table 4.5, we compare pose forecasting techniques from the SoA and the proposed SeS-GCN. On the short-term predictions, the best performance is that of SeS-GCN, reaching an MPJPE error of 48.8 mm, which is 7.9% better than the second-best STS-GCN [29].

On the longer-term predictions, the best performance (MPJPE error of 85.3 mm) is also detained by SeS-GCN, which is 2.4% better than the second best STS-GCN [29]. The proposed model outperforms all techniques on all actions except *Hammer*, a briefly repeating action that may differ for single hits. We argue that DCT-RNN-GCN [134] may get an advantage from using 50 input frames (all other methods use 10 frames)

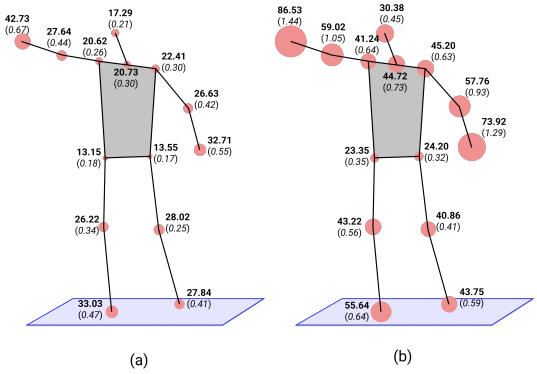

Fig. 4.3 Average MPJPE distribution for all actions in CHICO on different joints for (a) short-term (0.40 s) and (b) long-term (1.00 s) predictions. The radius of the blob gives the spatial error with the same scale as the skeleton.

For a graphical illustration, Fig. 4.3 shows a distribution of the error per joint calculated over all the actions, for the horizons 400 (*left*) and 1000 msec (*right*). In both cases, the error gets larger as we get closer to the extrema of the kinematic skeleton since those joints move the most. The slightly larger error at the right hand (70.03 and 125.76 mm, respectively) matches that subjects are right-handed (but some actions are operated with both hands).

For a sanity check of results, we have also evaluated the performance of a trivial zero velocity model. [201] has found that keeping the last observed positions may be a surprisingly strong (trivial) baseline. For CHICO, the zero velocity model scores an MPJPE of 110.6 at 25 frames, worse than the 85.3 mm score of SeS-GCN. This is in line with the large-scale dataset Human3.6M [135], where the performance of the trivial model is 153.3 mm.

Table 4.6 Evaluation of collision detection performance achieved by competing pose forecasting techniques, with the indication of inference run time. See discussion in Section 4.2.4.

Time Horizon (msec)	***1000***			
Metrics	*Prec*	*Recall*	F_1	*Inference Time (sec)*
DCT-RNN-GCN [134]	0.63	0.58	0.56	9.1×10^{-3}
MSR-GCN [126]	0.63	0.30	0.31	25.2×10^{-3}
STS-GCN [29]	0.68	0.61	0.63	$\mathbf{2.3 \times 10^{-3}}$
SeS-GCN (proposed)	0.84	0.54	**0.64**	$\mathbf{2.3 \times 10^{-3}}$

Collision detection experiments

We consider a collision to occur when any body limb of the subject gets too close to any part of the cobot, i.e. within a distance threshold, for at least one frame. In particular, a collision refers to the proximity between the cobot and the human in the forecast portion of the trajectory. The (Euclidean) distance threshold is set to 13 cm.

The motion of the cobot is scripted beforehand, thus known. The motion of the human subjects in the next 1000 msec needs to be forecast, starting from the observation of 400 msec. The train/validation/test sets sample sequences of 10+25 frames with stride of 10.

Evaluation of Collision Detection

For the evaluation of collision, following [213], both the cobot arm parts and the human body limbs are approximated by cylinders. The diameters for the cobot are fixed to 8cm. Those of the body limbs are taken from a human atlas.

In Table 4.6, we report precision, recall and F_1 scores for the detection of collisions on the motion of 2 test subjects, which contains 21 collisions. The top performer in pose forecasting, our proposed SeS-GCN, also yields the largest F_1 score of 0.64. The lower performing MSR-GCN [126] yields poor collision detection capabilities, with an F_1 score of 0.31.

4.3 HARPER

4.3.1 The HARPER Dataset

The main motivation behind the design of HARPER is to expand the research opportunities enabled by previous HRI datasets (see Section 4.1), especially considering

the transition from robots to cobots. The collection of the corpus involved 17 participants who were asked to perform 15 actions (the same for all participants). The data was captured with the sensors equipped on Spot: 5 greyscale + depth sensors and one RGB-D camera mounted on the gripper. Moreover, we used 6 MoCap sensors (Optitrack) and one RGB camera capturing the full setting (see Fig. 4.4).

Overall, HARPER contains 607 sequences for a total of over 60000 RGB images, grayscale images, depth frames, and 3D data from multi-sensor recordings. In the following, we discuss the acquisition setup (Section 4.3.1), we describe the actions we captured and their annotations (Section 4.3.1), and, finally, we provide key statistics about the data (Section 4.3.1).

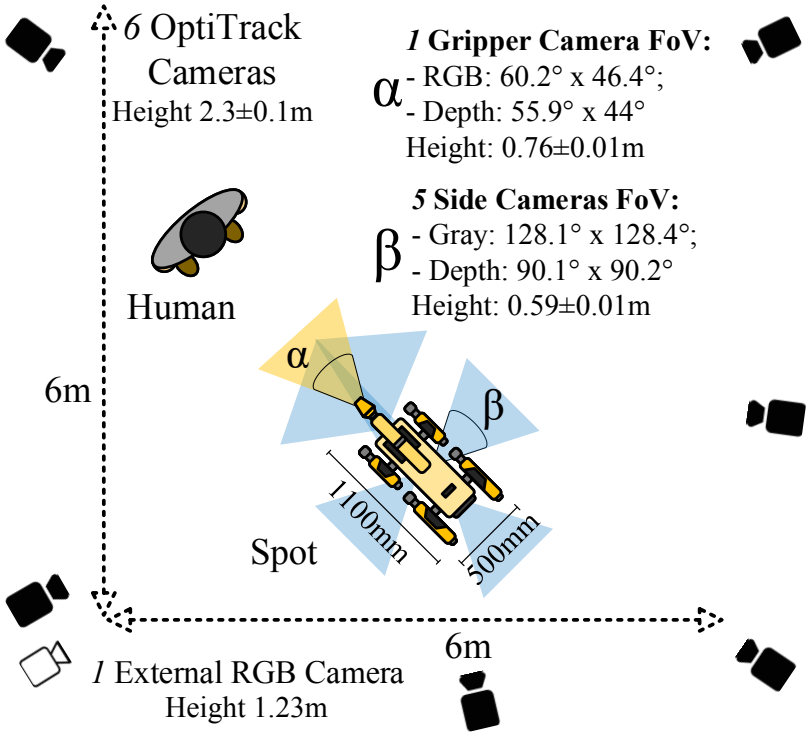

Fig. 4.4 A 6-camera Optitrack system covers a 6×6 squared meters area where users and Spot can freely move. The external RGB camera's Field of View covers the setting. The 5 Spot-on-body greyscale + depth cameras and the RGB-D frontal camera (gripper) cover the environment surrounding the robot.

Acquisition Setup

We collected all data in a laboratory (the layout is in Fig. 4.4). The 6 cameras of the Optitrack MoCap system were arranged to cover a 6×6 m^2 area, free of

obstacles, in which the participants performed the 15 actions of the HARPER scenario. All participants wore a motion capture suit with 37 reflective markers distributed according to the Optitrack *Baseline Marker Set* configuration. After calibration, the Optitrack tracks the markers with a 0.035 mm error at a sampling frequency of 120Hz. Furthermore, thanks to the configuration above, the Optitrack software (Motive) automatically extracts a 21-joint skeleton representation based on the marker positions.

The robot involved in the experiment is the Boston Dynamics Spot, a 12-DoF (3 per leg) quadruped robot equipped with 5 stereo cameras (greyscale with depth) around its body and one RGB-D camera on the gripper. The Spot acts within the Optitrack area described above, and its skeleton is obtained by applying forward kinematics to its internal motors state, acquired through the API provided by Boston Dynamics. The Spot skeleton is positioned in the same 3D scene as the participants' skeletons using a 4-marker rigid body mounted on its back and tracked by the Optitrack.

The Spot cameras operate at about 10 FPS and their data is synchronized with the Optitrack data. This ensures one of the most distinctive features of HARPER, i.e., the availability of two viewpoints, one from the robot and a panoptic view that covers the whole scene. The synchronization is achieved by aligning the timestamps of the data, with a temporal alignment error of less than 2 milliseconds. It is worth noticing that the overlap between Spot cameras is limited to the 3 frontal cameras with a very partial overlap. As a reference, we added an external RGB camera, positioned outside the Optitrack delimited area, to capture the whole scene (see bottom left part of Fig. 4.4). All the videos recorded with such a camera are provided with the dataset.

Actions and Annotations

We invited 17 university students to participate in the data collection (5 females and 12 males). They all signed an informed consent letter, and the information they provided, including the data collected during their participation, was handled according to the ethical regulations of the university in which the data was collected. Each participant interacted with the Spot individually in a session that followed a consistent sequence of steps. First, participants were assisted in wearing the suit necessary for marker tracking (see above), and then they were asked to display a T-pose for calibrating the skeleton extraction.

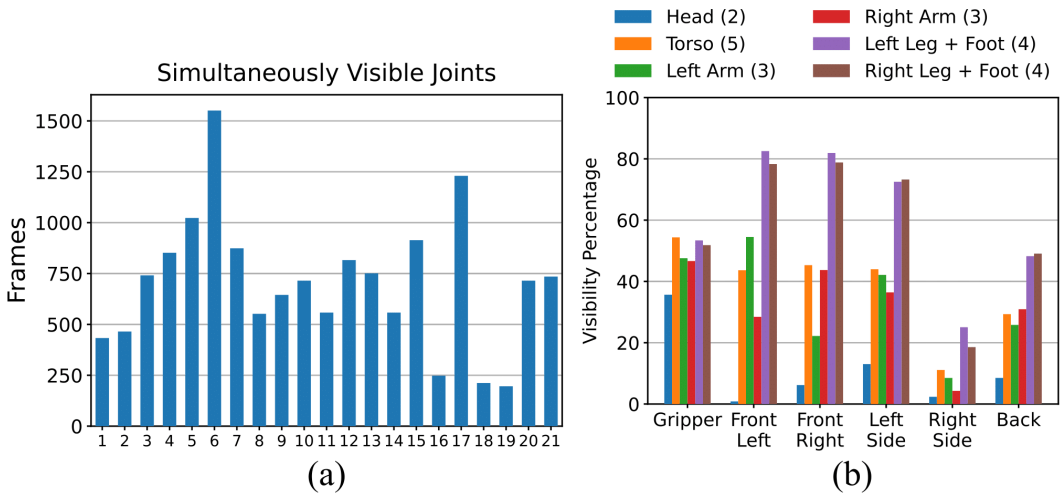

Fig. 4.5 Joints visibility from the robot's perspective. The left chart shows how many frames contain exactly n joints for $n = 1, \ldots, 21$. The right plot shows the percentage of frames in which the different parts of the skeleton are visible.

After calibrating the Optitrack, we asked the participants to perform 15 actions designed to reproduce different situations (Table 4.7), with the robot standing still for 8 actions and moving for 7 actions. In particular, the participants were instructed to act collisions as realistically as possible, *i.e.*, as if they were accidentally and unintentionally bumping into the Spot. The area covered by the Optitrack is sufficiently broad to perform the actions comfortably. However, some participants moved inadvertently out of it, leading to missed markers in a few frames. Similarly, some occlusions prevented the Optitrack from working properly for a few moments. These issues concerned no more than 3% of the total frames, and missing markers were effectively replaced through linear interpolation, ensuring that the skeleton representation was acquired with continuity and with the same precision at all times.

Optitrack and Spot share the same reference system, which enables the projection of 3D skeletons onto the videos captured with the robot's cameras (greyscale and RGB). This allows for accurate annotation of the joint positions in the video. In addition, given that the robot's leg motor state is known, forward kinematics was applied to compute the position of the robot's joints in the 3D space, starting from the rigid body mounted on the Spot back. This allowed us to obtain a 21-joint representation of both the participants' skeletons and the robot's skeleton.

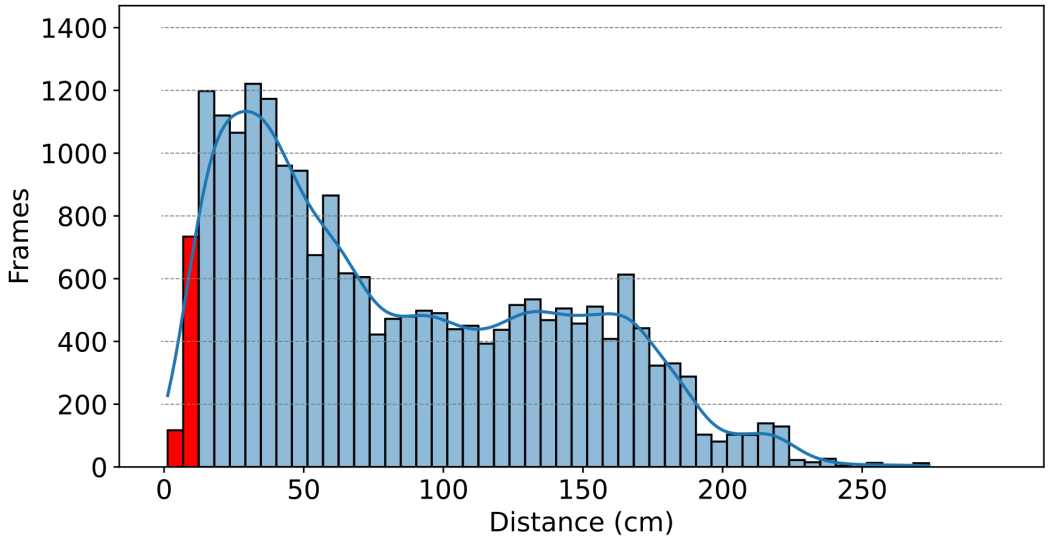

Fig. 4.6 Distribution of distances between Spot and users (the distance considers the two closest joints of human and robot). Red columns correspond to distances lower than 10 cm, considered as cases of physical contact.

Dataset Statistics

Fig. 4.5a shows, for all possible values of n, the number of frames in which exactly n human skeleton joints are visible to the robot. This information is important for understanding the difficulty level involved in addressing one of the new tasks that HARPER enables, namely the analysis and understanding of human poses when they are only partially visible. Similar information is shown in Fig. 4.5b, where human joints are grouped according to five body parts, *i.e.* head (2 joints), torso (5 joints), left/right arm (3 joints), and left/right leg (4 joints). The figure reports the percentage of times these body parts are visible (one joint is sufficient for the part to be considered visible) to each camera. One of the main patterns is that the gripper camera is more likely to capture the upper part of the body and legs, but not the feet (*i.e.* the spike on 17 visible joints caused by the gripper camera field of view).

Regarding the interaction between Spot and the participants, Fig. 4.6 shows the histograms of the distances between their closest joints. Two modes appear: below and above a distance of 1.3 meters. Distances corresponding to physical contact are highlighted in red. A threshold distance of 10 cm was used to determine whether physical contact was happening (see details in Section 4.3.2).

4.3.2 Experimental Results

HARPER provides three benchmarks: *3D Human Pose Estimation* (3D-HPE), *3D Human Pose Forecasting* (3D-HPF), and *Collision Prediction* (CP). All benchmarks are from the *robot's perspective*, *i.e.*, they are based on data captured with the robot's sensors, which is one of the key novelties of HARPER. Participants S1-S12 were used for training (15984 frames), while participants S13-S17 were used for testing (5542 frames). For 3D-HPF, we sampled 10990 sequences (of 20 frames each) for the training set and 3502 for the test set, keeping the same distribution of participants. The sequences were sampled by using a rolling window with a 1-frame step. Sequences without visible joints were excluded from the test set.

3D Human Pose Estimation

3D-HPE from the robot's perspective involves finding the 3D coordinates of the visible human joints using greyscale images and synchronized depth maps captured by the robot's sensors. The main challenge is that humans may not be fully visible (Section 4.3.1). The proposed baseline approach first uses a 2D pose estimator to find the position of the visible joints and then computes their 3D positions by exploiting the depth values, as shown in [215] (Fig. 4.7).

The 2D pose estimator is HRNet [214], trained on HARPER training data after resizing the images to 256×256 (no augmentation is applied). The depth sensors' Field of View (FoV) is narrower than that of the video cameras. Therefore, if an estimated joint is out of the depth FoV, it is considered non-visible.

The positions of the joints, along with their corresponding depth values, can then be mapped into the 3D Optitrack coordinate system. Once this task is completed, the 3D points inferred by the approach can be compared with those of the MoCap ground-truth skeleton.

We evaluated 2D pose estimation performance with the Percentage of Correct Keypoints (PCK) [216], *i.e.*, the fraction of correct predictions within a distance threshold τ (set to 0.5 on the predicted heatmaps). For the 3D joints estimation, we used the Mean Per Joint Position Error (MPJPE) [217], *i.e.*, the mean Euclidean distance between the visible estimated joints and the ground-truth Optitrack ones. We obtained a PCK of 86.8% and an average MPJPE (computed only on the visible joints) of 151 mm on 2D and 3D poses, respectively (Fig. 4.7). The 2D baseline

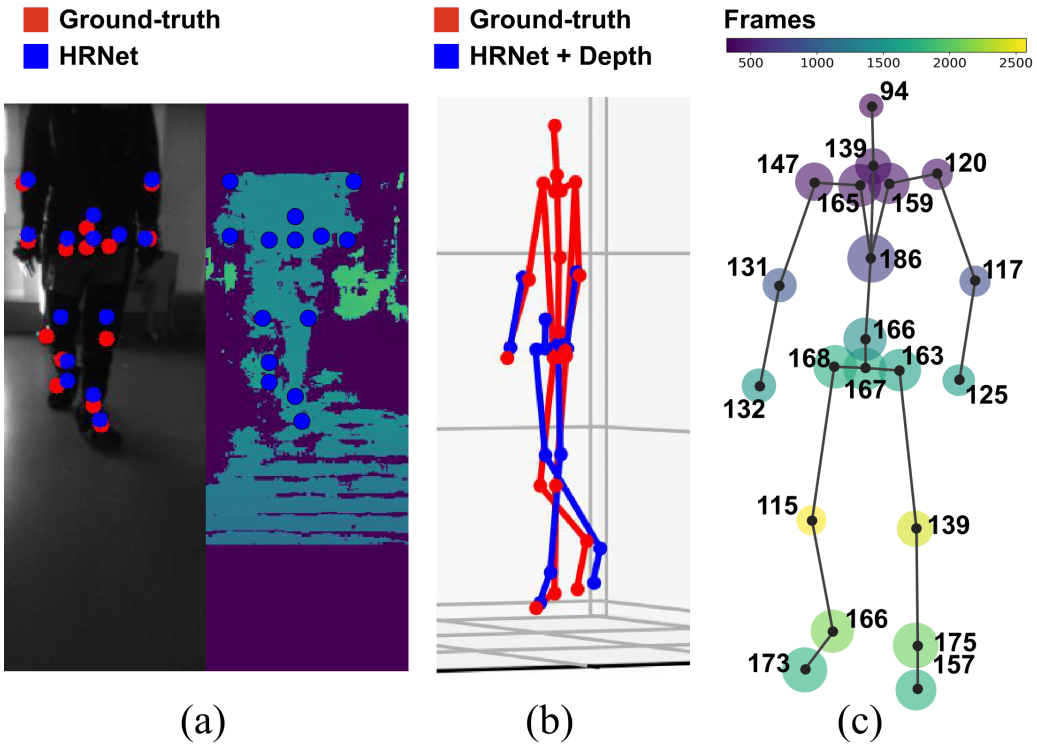

Fig. 4.7 Results of 3D human pose estimation from the robot perspective. (a) On the left, the predicted 2D joints (in blue) by HRNet [214] and the corresponding ground-truth joints (in red). On the right is the depth image with the same 2D detections. The depth will serve to do the lifting. (b) The lifted 3D poses alongside the complete Optitrack skeletons. (c) MPJPE (in mm) for every visible joint (inside the depth Field of View) on the test set. The size of the blobs is proportional to the errors, while colors are related to the number of times a joint is visible from the robot's perspective.

performs well, especially when considering that, in many cases, only one limb is visible or the participant is very close to the Spot.

The 3D lifting shows some limitations due to the noise in the depth maps, especially when the participants are far from the Spot. However, the performance was sufficient to address 3D-HPF and CP, both from the robot's perspective.

3D Human Pose Forecasting

3D-HPF from the robot's perspective involves predicting the future poses of human users with the robot's sensors. The pose at time t can be denoted as $X_t \in \mathbb{R}^{D \times J_h}$, where $D = 3$ is the dimension of the space and $J_h = 21$ is the total number of joints in the human skeleton (X_t is the set of all joint positions in 3D). Correspondingly,

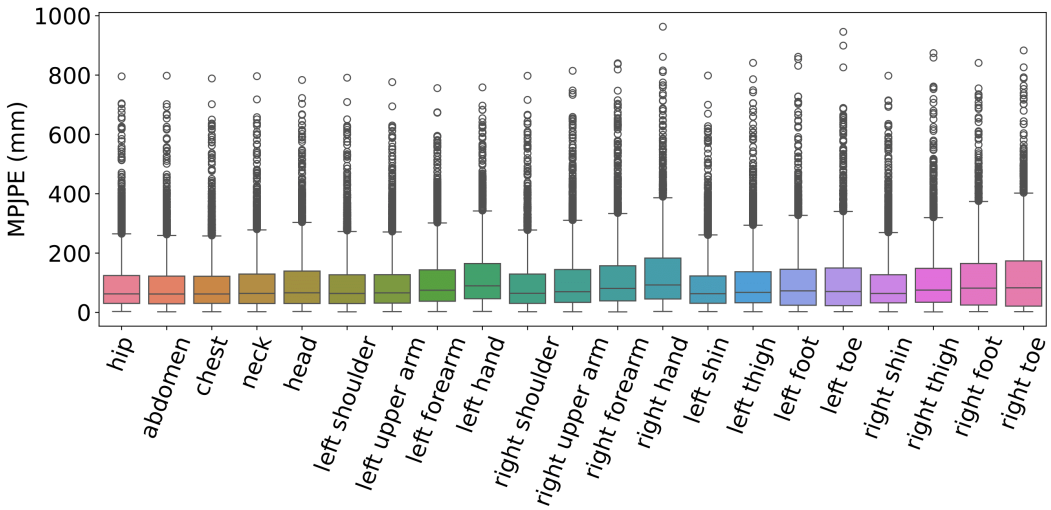

Fig. 4.8 MPJPE for each joint using EqMotion with GT as input and a forecasting horizon of 1000 ms.

3D-HPF means predicting $X_{t+1:t+K}$ based on $X_{t-T+1:t}$, where $X_{i:j} = X_i, X_{i+1}, \ldots, X_j$, and K is the *horizon*. In line with widely-used experimental protocols [135, 218], we set $T = 10$ and $K = 4$ (roughly 400 ms) or $K = 10$ (roughly 1000 ms), two cases referred to as *short-term* and *long-term* forecasting, respectively. We used average MPJPE over the K predicted frames (average MPJPE) or MPJPE over the K^{th} predicted frame (final MPJPE) as performance metrics. The pose forecasting baselines we applied are STS-GCN [218], SiMLPe [219], and EqMotion [27]. All three were trained using MPJPE as a loss function without applying augmentation. The training was performed using the 21-joint poses obtained with the Optitrack sensor as a ground truth.

Each baseline has three variants based on different assumptions about the input data: *GT* assumes that the robot can access all ground-truth joints in the human skeleton, (*GT+R*) assumes that the robot can access only the joints visible to its sensors, and (*HRNet+D+R*) represents the 3D pose as described in Section 4.3.2. The GT+R and HRNet+D+R baselines deal with input sequences of incomplete poses. These cannot be processed with the forecasting baselines above and, in general, with any of the approaches in the literature. Therefore, we used a diffusion-based time series imputation model, CSDI [220], built on a cascade of transformer blocks with skip connections. This model takes a sequence of incomplete poses as input and uses them to condition the generation of a complete pose, reconstructing the position of missing joints.

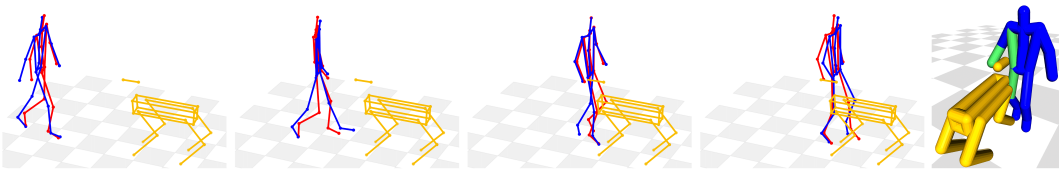

Fig. 4.9 Qualitative results for the pose forecasting with the 1000 ms horizon. (a) shows the human pose forecasted in blue and the ground truth in red. At the end of the sequence, an accidental collision occurs. In (b), the collision (highlighted in green) is detected as explained in Section 4.3.2. The forecasting approach used is EqMotion [27] on the GT data.

Table 4.8 shows the results for the three variants of every baseline. GT achieves the best performance, while HRNet+D+R, corresponding to the most challenging task, performs the worst. EqMotion [27] delivers the best absolute results *when in the presence of GT data*: 43 mm, on average, over the 400 ms horizon, and 70 mm over the 1000 ms horizon. However, STS-GCN [218] bridges the performance gap with EqMotion when the data is noisier like *e.g.*, in the HRNet+D+R case: the best average MPJPE is 313 mm over the 400 ms horizon and 332 mm over the 1000 ms horizon, while EqMotion achieves an MPJPE of 309 mm over the 400 ms horizon, and of 333 mm over the 1000 ms horizon.

Finally, we computed the MPJPE for each joint using the baseline with the smallest average error, *i.e.*, EqMotion [27], with GT as input (Fig. 4.8). We also estimated the correlation r between these errors and the average velocity of each human joint with the Pearson coefficient (r=0.79, p=1.79e-05), noticing that the faster a joint moves, the harder it is to predict its trajectory in the future.

Collision Prediction

CP from robot's perspective is the task of predicting whether the user and the robot will have physical contact, regardless of intent. Our focus is on the contacts or collisions caused by humans with the robot, due to the intricate challenges associated with predicting human movements, especially when the body is partially visible. Table 4.7 shows that HARPER includes four types of physical contact, all acted out to the best of the participants' abilities. Since they differ significantly in terms of energy and limbs involved, we addressed them as different cases in the experiments. The CP process takes as input a sequence of human poses $X_{t+1:t+K}$ (see above for the notation) and a sequence of robot's poses $Y_{t+1:t+K} = (Y_{t+1}, \ldots, Y_{t+K})$, where $Y_t \in \mathbb{R}^{D \times J_r}$ ($D = 3$ and J_r is the number of joints of the robot). The sequence $Y_{t+1:t+K}$

is assumed to be known because the robot plans its actions in advance. The goal is to determine whether $X_{t+1:t+K}$ and $Y_{t+1:t+K}$ indicate a *physical contact*, defined as two cylinders of radius $r = 5.0$ cm centered around the skeletal links of Spot and user are closer than a threshold $d_h = 10.0$ cm (Fig. 4.9b). Performance metrics used are accuracy, sensitivity, and specificity [221]. Sensitivity is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$ (measuring how effectively the system avoids False Positives), while specificity is $\mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$ (measuring how effectively the system predicts True Negatives).

We started the CP experiments by feeding the methods of Section 4.3.2 with the Optitrack ground-truth data. This gave us an upper bound of the performance and showed that punches and kicks are the most difficult to predict (Table 4.7), due to the speed and energy involved. In contrast, touch, with the lowest speed and energy showed the best performance. Then, we replaced the ground-truth data with the pose forecasts output by EqMotion [27] in its HRNet+D+R variant, completed by the CSDI [220] diffusion process. Table 4.7 shows a performance decrease, but not significantly.

Finally, we evaluated a straightforward baseline called *Depth-Based* in Table 4.9, showing that CP requires more sophisticated approaches. The baseline is a linear regression over the future K depth frames given T previous frames. This allowed us to test whether any points are predicted to get closer than d_h. We set $T = 10$, $K = 10$, and $d_h = 10$ cm, as used in the pose forecasting baselines. As expected, the Depth-Based method performed worse than the other methods, except for kicks, where accuracy and specificity were higher, possibly because the robot's cameras capture users' legs more easily than other body parts.

4.4 Discussion

This thesis chapter introduces two distinct datasets, HARPER and CHICO, designed for training and evaluating Human Pose Forecasting (HPF) systems. Together, these datasets aim to comprehensively cover the various scenarios encountered in a production line, particularly in terms of the different viewpoints from which HPF data can be collected.

The CHICO dataset consists of data captured by a multi-view system of RGB cameras, offering diverse perspectives on human-robot interactions. In contrast, HARPER provides a broader range of viewpoints, including a panoptic view from an OptiTrack system and a first-person perspective from RGB-D cameras mounted

on the robot's chassis. Additionally, we considered two complementary industrial use cases. CHICO focuses on a fixed robotic arm operating at an assembly station, while HARPER explores a more dynamic setting, where a quadruped robot navigates around the workspace.

Methodologically, both benchmarks leverage Graph Convolutional Network (GCN)-based networks. Our implementation, inspired by [29], achieves competitive performance while maintaining high computational efficiency, with fewer than 100K parameters. This lightweight architecture ensures that the model can be deployed on a wide range of hardware devices without requiring substantial computational resources.

Beyond providing datasets, we also establish a structured methodology to facilitate the adaptation and deployment of collision prediction systems in industrial environments. This integrated approach ensures that the proposed models are not only theoretically sound but also practically applicable in real-world human-machine collaboration scenarios.

Table 4.7 HARPER Actions. The expression *Contact* means that the distance between Spot and the user is lower than 10 cm.

Action	Action Description	Robot Moving	Contact
A1 Walk+Crash Frontal	Human walks towards Spot oriented frontally then collides;		✓
A2 Walk+Crash 45°	Human walks towards Spot oriented at 45° then collides;		✓
A3 Walk+Crash Sideway	Human walks towards Spot oriented at 90° then collides;		✓
A4 Walk+Crash Backwards	Human walks towards Spot oriented backwards then collides;		✓
A5 Walk+Stop	Human walks towards Spot, then stops right before colliding;	✓	
A6 & A7 Walk+Avoid	Human and Spot walk towards each other avoiding collision at last second on the right (A6) / left (A7).	✓	
A8 Walk+Touch	Human walks towards Spot, then physically touch it;		✓
A9 Walk+Kick	Human walks towards Spot, then kicks it;		✓
A10 & A11 Walk+Punch	Human walks towards Spot oriented at 0° (A10) / 90° (A11), then punches it		✓
A12 Circular Walk	Human and Spot walk together in a circular path	✓	
A13 Circular Follow +Touch	Human follows Spot in a circle, then touches it with the hand	✓	✓
A14 Circular Follow + Avoid	Spot follows the human in a circle, then avoids contact	✓	
A15 Circular Follow + Crash	Spot follows the human in a circle, then a collision happen	✓	✓

Table 4.8 Pose forecasting errors. We provide the MPJPE expressed in mm with a prediction horizon of 400 and 1000 ms. The errors are computed for the particular frame for each action (first nine columns) as well as the average over all frames (*Average*) and the average over the last frame of each action instance (*Last frame average*).

Actions		**A1-4**		**A5**		**A6-7**		**A8**		**A9**		**A10-11**		**A12**		**A13**		**A14**		**A15**		***Average***		***Last frame average***	
Time (ms)		*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*
STS-GCN [2]	GT	115	169	104	150	99	134	116	129	143	218	144	213	148	229	179	268	112	148	150	224	120	171	147	260
	GT+R	182	229	121	153	281	340	126	145	148	213	154	224	156	234	193	288	262	302	306	338	190	242	212	326
	HRNet+D+R	414	469	188	249	559	674	276	306	197	283	257	340	204	294	337	430	646	684	560	606	**382**	453	**400**	**538**
SiMLPe [3]	GT	63	150	60	140	42	99	31	72	66	145	77	183	89	208	99	221	46	108	90	216	60	141	98	264
	GT+R	158	231	100	174	261	351	87	121	82	163	114	199	112	226	137	260	231	302	280	353	158	237	193	359
	HRNet+D+R	436	528	213	275	674	917	290	368	232	349	310	451	205	307	412	610	681	820	785	1170	441	595	511	790
EqMotion [4]	GT	42	110	37	91	25	61	23	60	40	104	60	139	66	174	73	164	35	96	68	168	**41**	**104**	**69**	**197**
	GT+R	147	200	81	125	255	326	79	108	61	124	102	169	92	188	112	185	221	286	267	328	**146**	**205**	**171**	**299**
	HRNet+D+R	419	488	194	234	569	660	274	299	187	237	262	319	187	285	340	407	632	678	613	628	386	**446**	422	552

Table 4.9 Performance of the different collision prediction methods with a 1000 ms horizon in terms of accuracy, sensitivity, and specificity score. The evaluation is divided into the four categories of contacts represented in the HARPER dataset.

Method	**Input Type**	**Unintended**			**Touch**			**Kick**			**Punch**		
		Acc.↑	**Sen.↑**	**Spec.↑**	**Acc.↑**	**Sen.↑**	**Spec.↑**	**Acc.↑**	**Sen.↑**	**Spec.↑**	**Acc.↑**	**Sen.↑**	**Spec.↑**
STS-GCN [218]	GT	0.93	0.92	**0.94**	0.92	0.91	0.93	0.85	0.78	0.90	0.78	0.83	0.71
SiMLPe [219]	GT	0.93	0.93	0.92	0.94	0.92	**0.97**	0.81	0.83	0.79	0.78	**0.89**	0.63
EqMotion [27]	GT	**0.95**	**0.96**	**0.94**	**0.95**	**0.96**	0.94	**0.93**	**0.88**	**0.96**	**0.82**	0.83	**0.82**
Depth-based	D	0.49	0.25	**0.90**	0.53	0.33	0.87	0.71	0.39	**0.94**	0.60	0.61	0.59
EqMotion [27]	HRNet+D+R	**0.74**	**0.61**	**0.90**	**0.89**	**0.88**	**0.91**	**0.80**	**0.73**	0.86	**0.67**	**0.66**	**0.69**

Chapter 5

Conclusion

This thesis has investigated the application of Deep Learning (DL) models for Time-Series Forecasting in diverse industrial settings, focusing particularly on human-machine interaction and sustainability. Through three main case studies, this research highlights the importance of Time-Series Forecasting techniques, scaled from 1D to 3D analysis, in improving efficiency, safety, and sustainability within the industry.

Prediction for Planning. Firstly, the thesis addressed the complex challenge of New Fashion Product Performance Forecasting (NFPPF), which is difficult due to the absence of historical sales data and rapidly evolving trends. Two innovative diffusion-based models were presented: MDiFF and Dif4FF. MDiFF utilizes a two-stage approach, employing score-based diffusion models to generate multiple predictions, refined using a Multi-Layer Perceptron (MLP). Dif4FF, on the other hand, uses a diffusion model with Graph Convolutional Network (GCN) for refinement, incorporating Google Trends data. Both models demonstrated superior performance compared to state-of-the-art methods, highlighting the effectiveness of diffusion models for this forecasting task. Specifically, Dif4FF exhibited greater resilience to domain shifts, a common issue in the fashion industry. Additionally, POP++, a data-centric method, was introduced to improve the quality of data used by forecasting models. POP++ enriches the POtential Performance (POP) signal with information from human poses, segmentations, and image features, leading to more accurate predictions. It was shown that the POP++ signal aligns more closely with actual sales than the original POP signal.

Prediction for Adaptation. Next, the thesis focused on human trajectory forecasting, which is critical for ensuring safety in human-robot interactions. The SITUATE model, based on GCN, was introduced for human trajectory forecasting in indoor environments. SITUATE uses equivariant and invariant geometric features, along with a self-supervised visual representation of the environment, achieving state-of-the-art results across various indoor datasets. This highlights the importance of understanding the environmental context and movement patterns for accurate trajectory prediction in dynamic indoor settings.

Prediction for Interaction. Finally, the thesis explored Human Pose Forecasting (HPF), which is essential for seamless and safe Human Robot Collaboration (HRC). Two new datasets were introduced: CHICO and HARPER. The CHICO dataset captures human-robot interactions from various perspectives using a multi-camera system, providing valuable data for the development of robust models. HARPER provides a panoptic view of human behavior using data from a quadruped robot, making it unique. The SeS-GCN model was proposed for human pose forecasting and achieved competitive results on the Human3.6M and CHICO datasets. This model uses separable space-time adjacency matrices, depth-wise separable graph convolutions, and sparse adjacency matrices, contributing to the development of more efficient and accurate pose forecasting systems.

In summary, this thesis has contributed to the advancement of Time-Series Forecasting through the application of DL models in diverse industrial scenarios. The models and datasets introduced to provide practical solutions for improving production planning, safety, and HRC. The research aligns with the vision of Industry 5.0, where artificial intelligence works synergistically with humans to achieve more sustainable, ethical, and efficient production. The use of diffusion models for sales forecasting, the analysis of indoor human trajectory forecasting, and the proposed datasets for HPF from multiple viewpoints demonstrate the importance of multi-dimensional and multi-modal approaches to complex industrial problems.

5.1 Future Works

The research presented in this thesis has explored three key forecasting challenges in industrial applications: planning, adaptation, and interaction. While significant

advancements have been made, several avenues remain open for further exploration. This section outlines potential directions for future research in each of these domains.

Prediction for Planning. The NFPPF task remains a highly complex problem due to the inherent uncertainty of market trends and the lack of historical sales data for new garments. While diffusion-based models such as MDiFF and Dif4FF have demonstrated strong predictive capabilities, future research should investigate the integration of POP++ inside the Denoising Diffusion Probabilistic Model (DDPM) model. The knowledge given by the POP++ signal could drastically decrease the model error and guide the conditioning towards better sample generation. Another extension of DDPMs based methods is the use of the uncertainty and variance given by the multiple predictions, to train uncertainty-aware models. Finally, incorporating real-time social media trends, user sentiment analysis, and e-commerce behavioral data could enhance forecasting accuracy. Additionally, developing more robust domain adaptation techniques would help mitigate the impact of seasonality and regional market differences. Another promising avenue is the exploration of reinforcement learning strategies to dynamically adjust production plans based on forecasted demand, thereby further optimizing supply chain efficiency and sustainability.

Prediction for Adaptation. Human Trajectory Forecasting in indoor industrial environments is crucial for ensuring safe and efficient HRC. While the SITUATE model has made strides in leveraging geometric features and self-supervised vision representations, future work should focus on improving the generalizability of trajectory forecasting models across different environments. Incorporating unsupervised domain adaptation techniques could enable models to transfer knowledge from one setting to another with minimal retraining. Additionally, the integration of transformer-based architectures with graph neural networks may improve long-term trajectory prediction by capturing both local and global motion patterns. Future efforts should also explore how the model would react when both human actions and the scene are changing rapidly. So far, the model has only been trained on static scenes and environments, and new methodologies could be explored to overcome this issue.

Prediction for Interaction. The ability to accurately predict human poses is a key enabler for seamless and intuitive HRC. While this thesis introduced the CHICO and HARPER datasets to improve pose forecasting in collaborative settings, future research should explore the fusion of multimodal sensor data, including wearable sensors and LiDAR, to enhance prediction robustness. Additionally, further development of lightweight HRC models could enable efficient, low-latency inference suitable for real-time applications. Another critical direction involves integrating human intent recognition into forecasting models, allowing robots to anticipate not only physical movements but also underlying task intentions. Finally, extending these methods to diverse and unstructured environments beyond industrial settings—such as healthcare and assistive robotics—could broaden the impact of human pose forecasting research.

By addressing these challenges, future research can continue to push the boundaries of time-series forecasting in industrial applications, further advancing the principles of Industry 5.0 by fostering greater synergy between artificial intelligence and human-centric processes.

References

[1] D A Zakoldaev, A V Shukalov, and I O Zharinov. "From Industry 3.0 to Industry 4.0: production modernization and creation of innovative digital companies". In: *IOP Conference Series: Materials Science and Engineering* 560.1 (June 2019), p. 012206.

[2] Javier Villalba-Diez, Daniel Schmidt, Roman Gevers, Joaquín Ordieres-Meré, Martin Buchwitz, and Wanja Wellbrock. "Deep Learning for Industrial Computer Vision Quality Control in the Printing Industry 4.0". In: *Sensors* (2019).

[3] Jorge Ribeiro, Rui Lima, Tiago Eckhardt, and Sara Paiva. "Robotic Process Automation and Artificial Intelligence in Industry 4.0 – A Literature review". In: *Procedia Computer Science* 181 (2021). CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020, pp. 51–58.

[4] Lidia Alexa, Marius Pıslaru, and Silvia Avasilcăi. "From Industry 4.0 to Industry 5.0—an overview of European Union enterprises". In: *Sustainability and Innovation in Manufacturing Enterprises: Indicators, Models and Assessment for Industry 5.0* (2022), pp. 221–231.

[5] Saeid Nahavandi. "Industry 5.0—A Human-Centric Solution". In: *Sustainability* 11.16 (Aug. 2019), p. 4371.

[6] Jiewu Leng, Weinan Sha, Baicun Wang, Pai Zheng, Cunbo Zhuang, Qiang Liu, Thorsten Wuest, Dimitris Mourtzis, and Lihui Wang. "Industry 5.0: Prospect and retrospect". In: *Journal of Manufacturing Systems* 65 (2022), pp. 279–295.

[7] LI Xiao-rui, BAN Xiao-juan, YUAN Zhao-lin, and QIAO Hao-ran. "Review on deep learning models for time series forecasting in industry". In: *Chinese Journal of Engineering* (2022).

[8] Bryan Lim and Stefan Zohren. "Time-series forecasting with deep learning: a survey". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379.2194 (Feb. 2021), p. 20200209.

[9] José F Torres, Dalil Hadjout, Abderrazak Sebaa, Francisco Martínez-Álvarez, and Alicia Troncoso. "Deep learning for time series forecasting: a survey". In: *Big Data* 9.1 (2021), pp. 3–21.

[10] Soujanya Mantravadi and Charles Møller. “An overview of next-generation manufacturing execution systems: How important is MES for industry 4.0?” In: *Procedia manufacturing* 30 (2019), pp. 588–595.

[11] Federico Cunico, Stefano Aldegheri, Andrea Avogaro, Michele Boldo, Nicola Bombieri, Luigi Capogrosso, Ariel Caputo, Damiano Carra, Stefano Centomo, Dong Seon Cheng, Ettore Cinquetti, Marco Cristani, Mirco De Marchi, Florenc Demrozi, Marco Emporio, Franco Fummi, Luca Geretti, Samuele Germiniani, Andrea Giachetti, Federico Girella, Enrico Martini, Gloria Menegaz, Niek Muijs, Federica Paci, Marco Panato, Graziano Pravadelli, Elisa Quintarelli, Ilaria Siviero, Silvia Francesca Storti, Carlo Tadiello, Cristian Turetta, Tiziano Villa, Nicola Zannone, and Davide Quaglia. “Enhancing Safety and Privacy in Industry 4.0: The ICE Laboratory Case Study”. In: *IEEE Access* 12 (2024), pp. 154570–154599.

[12] Tiago Zonta, Cristiano André Da Costa, Rodrigo da Rosa Righi, Miromar Jose de Lima, Eduardo Silveira da Trindade, and Guann Pyng Li. “Predictive maintenance in the Industry 4.0: A systematic literature review”. In: *Computers & Industrial Engineering* 150 (2020), p. 106889.

[13] Georgios Makridis, Dimosthenis Kyriazis, and Stathis Plitsos. “Predictive maintenance leveraging machine learning for time-series forecasting in the maritime industry”. In: *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. 2020.

[14] Uzair Khan, Dong Cheng, Francesco Setti, Franco Fummi, Marco Cristani, and Luigi Capogrosso. “A Comprehensive Survey on Deep Learning-based Predictive Maintenance”. In: *ACM Transactions on Embedded Computing Systems* (2025).

[15] Nataliia Kashpruk, Cezary Piskor-Ignatowicz, and Jerzy Baranowski. “Time Series Prediction in Industry 4.0: A Comprehensive Review and Prospects for Future Advancements”. In: *Applied Sciences* 13.22 (2023), p. 12374.

[16] Alessio Sampieri, Guido Maria D’Amely di Melendugno, Andrea Avogaro, Federico Cunico, Francesco Setti, Geri Skenderi, Marco Cristani, and Fabio Galasso. “Pose forecasting in industrial human-robot collaboration”. In: *European Conference on Computer Vision*. Springer. 2022.

[17] David Fridovich-Keil, Andrea Bajcsy, Jaime F Fisac, Sylvia L Herbert, Steven Wang, Anca D Dragan, and Claire J Tomlin. “Confidence-aware motion prediction for real-time collision avoidance1”. In: *The International Journal of Robotics Research* 39.2-3 (2020), pp. 250–265.

[18] Andrea Avogaro, Luigi Capogrosso, Andrea Toaiari, Franco Fummi, and Marco Cristani. *New Fashion Products Performance Forecasting: A Survey on Evolutions, Models and Emerging Trends*. 2025.

[19] Geri Skenderi, Christian Joppi, Matteo Denitto, Berniero Scarpa, and Marco Cristani. “The multi-modal universe of fast-fashion: the Visuelle 2.0 benchmark”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2022, pp. 2240–2245.

[20] Geri Skenderi, Christian Joppi, Matteo Denitto, and Marco Cristani. "Well Googled is Half Done: Multimodal Forecasting of New Fashion Product Sales with Image-based Google Trends". In: *Journal of Forecasting* 43.6 (Mar. 2024), pp. 1982–1997.

[21] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Ioannis Kompatsiaris. "Multimodal Quasi-AutoRegression: forecasting the visual popularity of new fashion products". In: *International Journal of Multimedia Information Retrieval* 11.4 (2022), pp. 717–729.

[22] Andrea Avogaro, Luigi Capogrosso, Franco Fummi, and Marco Cristani. "Dif4FF: Leveraging Multimodal Diffusion Models and Graph Neural Networks for Accurate New Fashion Product Performance Forecasting". In: *International Conference on Pattern Recognition (ICPR)*. 2024.

[23] Sijie Yan, Yuanjun Xiong, and Dahua Lin. "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 2018).

[24] Timm Faulwasser, Tobias Weber, Pablo Zometa, and Rolf Findeisen. "Implementation of nonlinear model predictive path-following control for an industrial robot". In: *IEEE Transactions on Control Systems Technology* 25.4 (2016), pp. 1505–1511.

[25] Luigi Capogrosso, Andrea Toaiari, Andrea Avogaro, Uzair Khan, Aditya Jivoji, Franco Fummi, and Marco Cristani. "SITUATE: Indoor Human Trajectory Prediction Through Geometric Features and Self-supervised Vision Representation". In: *27th International Conference on Pattern Recognition (ICPR)*. 2024.

[26] Chaoyang Zhu. "Intelligent robot path planning and navigation based on reinforcement learning and adaptive control". In: *J. Logist. Inform. Serv. Sci* 10.3 (2023), pp. 235–248.

[27] Chenxin Xu, Robby T. Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. "EqMotion: Equivariant Multi-Agent Motion Prediction with Invariant Interaction Reasoning". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023, pp. 1410–1420.

[28] Andrea Avogaro, Andrea Toaiari, Federico Cunico, Xiangmin Xu, Haralambos Dafas, Alessandro Vinciarelli, Emma Li, and Marco Cristani. "Exploring 3D Human Pose Estimation and Forecasting from the Robot's Perspective: The HARPER Dataset". In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2024, pp. 5828–5835.

[29] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. "Space-Time-Separable Graph Convolutional Network for Pose Forecasting". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021.

[30] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. "Graph convolutional networks: a comprehensive review". In: *Computational Social Networks* 6.1 (2019), pp. 1–23.

[31] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. "Simple and deep graph convolutional networks". In: *International conference on machine learning*. PMLR. 2020, pp. 1725–1735.

[32] Yuzhou Chen, Ignacio Segovia, and Yulia R Gel. "Z-GCNETs: Time zigzags at graph convolutional networks for time series forecasting". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 1684–1694.

[33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), pp. 6840–6851.

[34] Marcel Kollovieh, Abdul Fatir Ansari, Michael Bohlke-Schneider, Jasper Zschiegner, Hao Wang, and Yuyang Bernie Wang. "Predict, Refine, Synthesize: Self-Guiding Diffusion Models for Probabilistic Time Series Forecasting". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2024).

[35] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. "Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8857–8868.

[36] Andrea Avogaro, Luigi Capogrosso, Franco Fummi, and Marco Cristani. "MDiFF: Exploiting Multimodal Score-based Diffusion Models for New Fashion Product Performance Forecasting". In: *European Conference on Computer Vision (ECCV)*. 2025.

[37] Andrea Avogaro, Federico Girella, Luigi Capogrosso, Franco Fummi, Geri Skenderi, and Marco Cristani. "POP++: Refining POtential Performance of New Fashion Products through Human Proxemics". In: *IEEE Access [currently under review]* (2025).

[38] Kirsi Niinimäki, Greg Peters, Helena Dahlbo, Patsy Perry, Timo Rissanen, and Alison Gwilt. "The environmental price of fast fashion". In: *Nature Reviews Earth & Environment* 1.4 (Apr. 2020), pp. 189–200.

[39] Kerrice Bailey, Aman Basu, and Sapna Sharma. "The Environmental Impacts of Fast Fashion on Water Quality: A Systematic Review". In: *Water* 14.7 (Mar. 2022), p. 1073.

[40] Nesreen K Ahmed, Amir F Atiya, Neamat El Gayar, and Hisham El-Shishiny. "An Empirical Comparison of Machine Learning Models for Time Series Forecasting". In: *Econometric Reviews* 29.5–6 (Aug. 2010), pp. 594–621.

[41] Lequan Lin, Zhengkun Li, Ruikun Li, Xuliang Li, and Junbin Gao. "Diffusion models for time-series applications: a survey". In: *Frontiers of Information Technology & Electronic Engineering* 25.1 (Dec. 2023), pp. 19–41.

[42] Luigi Capogrosso, Federico Girella, Francesco Taioli, Michele Chiara, Muhammad Aqeel, Franco Fummi, Francesco Setti, and Marco Cristani. "Diffusion-Based Image Generation for In-Distribution Data Augmentation in Surface Defect Detection". In: *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and Technology Publications, 2024.

[43] Federico Girella, Ziyue Liu, Franco Fummi, Francesco Setti, Marco Cristani, and Luigi Capogrosso. "Leveraging Latent Diffusion Models for Training-Free In-Distribution Data Augmentation for Surface Defect Detection". In: *arXiv preprint arXiv:2407.03961* (2024).

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 10674–10685.

[45] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. *Diffusers: State-of-the-art diffusion models*. https://github.com/huggingface/diffusers. Accessed: 2024-11-03.

[46] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. "Diffusion Models: A Comprehensive Survey of Methods and Applications". In: *ACM Computing Surveys* 56.4 (Nov. 2023), pp. 1–39.

[47] Christian Joppi, Geri Skenderi, and Marco Cristani. "POP: Mining POtential Performance of New Fashion Products via Webly Cross-modal Query Expansion". In: *Computer Vision – ECCV 2022*. Springer Nature Switzerland, 2022, pp. 34–50.

[48] Curtis Northcutt, Lu Jiang, and Isaac Chuang. "Confident Learning: Estimating Uncertainty in Dataset Labels". In: *Journal of Artificial Intelligence Research* 70 (2021), pp. 1373–1411.

[49] Shuyun Ren, Hau-Ling Chan, and Pratibha Ram. "A Comparative Study on Fashion Demand Forecasting Models with Multiple Sources of Uncertainty". In: *Annals of Operations Research* 257.1–2 (Apr. 2016), pp. 335–355.

[50] Giuseppe Craparotta, Sébastien Thomassey, and Amedeo Biolatti. "A Siamese Neural Network Application for Sales Forecasting of New Fashion Products Using Heterogeneous Data". In: *International Journal of Computational Intelligence Systems* 12.2 (2019), p. 1537.

[51] Pawan Kumar Singh, Yadunath Gupta, Nilpa Jha, and Aruna Rajan. "Fashion Retail: Forecasting Demand for New Items". In: *arXiv preprint arXiv:1907.01960* (2019).

[52] Vijay Ekambaram, Kushagra Manglik, Sumanta Mukherjee, Surya Shravan Kumar Sajja, Satyam Dwivedi, and Vikas Raykar. "Attention based Multi-Modal New Product Sales Time-series Forecasting". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '20. ACM, Aug. 2020.

[53] Mohammad Motamedi, Nikolay Sakharnykh, and Tim Kaldewey. “A data-centric approach for training deep neural networks with less data”. In: *arXiv preprint arXiv:2110.03613* (2021).

[54] Ariful Islam Anik and Andrea Bunt. “Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency”. In: *Conference on Human Factors in Computing Systems*. 2021.

[55] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. “Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2021.

[56] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. “SELFIE: Refurbishing Unclean Samples for Robust Deep Learning”. In: *36th International Conference on Machine Learning (ICML)*. 2019.

[57] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. “Symmetric Cross Entropy for Robust Learning With Noisy Labels”. In: *International Conference on Computer Vision (ICCV)*. 2019.

[58] Ling Carlos García, Elizabeth, FridaRim, and Ferrando Jaime. *HM Personalized Fashion Recommendations*. https://kaggle.com/competitions/h-and-m-personalized-fashion-recommendations. Accessed: 2024-09-16.

[59] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. “Score-Based Generative Modeling through Stochastic Differential Equations”. In: *International Conference on Learning Representations (ICLR)*. 2020.

[60] Albert Gu, Karan Goel, and Christopher Ré. “Efficiently Modeling Long Sequences with Structured State Spaces”. In: *arXiv preprint arXiv:2111.00396* (2021).

[61] A Vaswani. “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2017).

[62] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016.

[63] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2009.

[64] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems (NeurIPS)* 32 (2019).

[65] Ilya Loshchilov and Frank Hutter. “Decoupled Weight Decay Regularization”. In: *International Conference on Learning Representations*. 2018.

[66] Jerome H Friedman. “Greedy function approximation: A gradient boosting machine”. In: *The Annals of Statistics* 29.5 (Oct. 2001).

[67] Ryotaro Shimizu, Yuki Saito, Megumi Matsutani, and Masayuki Goto. "Fashion intelligence system: An outfit interpretation utilizing images and rich abstract tags". In: *Expert Systems with Applications* 213 (Mar. 2023), p. 119167.

[68] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. "Deep Learning-based Human Pose Estimation: A Survey". In: *ACM Computing Surveys* 56.1 (2023), pp. 1–37.

[69] Andrea Avogaro, Federico Cunico, Bodo Rosenhahn, and Francesco Setti. "Markerless human pose estimation for biomedical applications: a survey". In: *Frontiers in Computer Science* 5 (2023), pp. 1153–1167.

[70] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. "Segment Anything". In: *International Conference on Computer Vision (ICCV)*. 2023.

[71] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations (ICLR)*. 2019.

[72] Noam Shazeer and Mitchell Stern. "Adafactor: Adaptive Learning Rates with Sublinear Memory Cost". In: *35th International Conference on Machine Learning (ICML)*. 2018.

[73] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[74] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft COCO: Common Objects in Context". In: *European Conference on Computer Vision (ECCV)*. 2014.

[75] Igor Ilic, Berk Görgülü, Mucahit Cevik, and Mustafa Gökçe Baydoğan. "Explainable boosted linear regression for time series forecasting". In: *Pattern Recognition* 120 (2021), p. 108144.

[76] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. "Human motion trajectory prediction: A survey". In: *The International Journal of Robotics Research* (2020).

[77] Parth Kothari, Sven Kreiss, and Alexandre Alahi. "Human trajectory forecasting in crowds: A deep learning perspective". In: *IEEE Transactions on Intelligent Transportation Systems* (2021).

[78] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer. 2020.

[79] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. "Social lstm: Human trajectory prediction in crowded spaces". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[80] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. "Social gan: Socially acceptable trajectories with generative adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[81] Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. "Transformer Networks for Trajectory Forecasting". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. IEEE, Jan. 2021, pp. 10335–10342.

[82] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. "Stochastic trajectory prediction via motion indeterminacy diffusion". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[83] Chenxin Xu, Weibo Mao, Wenjun Zhang, and Siheng Chen. "Remember Intentions: Retrospective-Memory-based Trajectory Prediction". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 6478–6487.

[84] Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. "Learning Pedestrian Group Representations for Multi-modal Trajectory Prediction". In: *Computer Vision – ECCV 2022*. Springer Nature Switzerland, 2022, pp. 270–289.

[85] Pranav Mantini and Shishir K Shah. "Human trajectory forecasting in indoor environments using geometric context". In: *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*. 2014.

[86] Andrey Rudenko, Tomasz P Kucner, Chittaranjan S Swaminathan, Ravi T Chadalavada, Kai O Arras, and Achim J Lilienthal. "Thör: Human-robot navigation data collection and accurate motion trajectories dataset". In: *IEEE Robotics and Automation Letters* (2020).

[87] Luca Rossi, Marina Paolanti, Roberto Pierdicca, and Emanuele Frontoni. "Human trajectory prediction and generation using LSTM models and GANs". In: *Pattern Recognition* (2021).

[88] Peng Wang, Jing Yang, and Jianpei Zhang. "Location prediction for indoor spaces based on trajectory similarity". In: *2021 4th International Conference on Data Science and Information Technology*. 2021.

[89] Geri Skenderi, Alessia Bozzini, Luigi Capogrosso, Enrico Carlo Agrillo, Giovanni Perbellini, Franco Fummi, and Marco Cristani. "Dohmo: Embedded computer vision in co-housing scenarios". In: *2021 Forum on specification & Design Languages (FDL)*. IEEE. 2021.

[90] Peng Wang, Jing Yang, and Jianpei Zhang. "Indoor trajectory prediction for shopping mall via sequential similarity". In: *Information* (2022).

[91] Andrea Toaiari, Federico Cunico, Francesco Taioli, Ariel Caputo, Gloria Menegaz, Andrea Giachetti, Giovanni Maria Farinella, and Marco Cristani. "SCENE-pathy: Capturing the Visual Selective Attention of People Towards Scene Elements". In: *International Conference on Image Analysis and Processing*. Springer. 2023.

[92] Patrizia Gabellini, Mauro D'Aloisio, Matteo Fabiani, and Valerio Placidi. "A large scale trajectory dataset for shopper behaviour understanding". In: *New Trends in Image Analysis and Processing–ICIAP 2019: ICIAP International Workshops, BioFor, PatReCH, e-BADLE, DeepRetail, and Industrial Session, Trento, Italy, September 9–10, 2019, Revised Selected Papers 20*. Springer. 2019.

[93] Luigi Capogrosso, Geri Skenderi, Federico Girella, Franco Fummi, and Marco Cristani. "Toward smart doors: A position paper". In: *International Conference on Pattern Recognition*. Springer. 2022.

[94] Chiho Choi and Behzad Dariush. "Looking to relations for future trajectory forecast". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

[95] Sheng Guo, Hanjiang Xiong, and Xianwei Zheng. "A novel semantic matching method for indoor trajectory tracking". In: *ISPRS international journal of geo-information* (2017).

[96] A ASHRAE. *Guideline 10P, Interactions Affecting the Achievement of Acceptable Indoor Environments*. 2010.

[97] Sirin Haddad, Meiqing Wu, He Wei, and Siew Kei Lam. "Situation-aware pedestrian trajectory prediction with spatio-temporal attention model". In: *arXiv preprint arXiv:1902.05437* (2019).

[98] Görkay Aydemir, Adil Kaan Akan, and Fatma Güney. "Adapt: Efficient multi-agent trajectory prediction with adaptation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

[99] Dirk Helbing and Peter Molnar. "Social force model for pedestrian dynamics". In: *Physical review E* (1995).

[100] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. "Activity forecasting". In: *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*. Springer. 2012.

[101] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. "Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

[102] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. "Spatio-temporal graph transformer networks for pedestrian trajectory prediction". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer. 2020.

[103] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. "It Is Not the Journey But the Destination: Endpoint Conditioned Trajectory Prediction". In: *Computer Vision – ECCV 2020*. Springer International Publishing, 2020, pp. 759–776.

[104] Taco Cohen and Max Welling. "Group equivariant convolutional networks". In: *International conference on machine learning*. PMLR. 2016.

[105] Yiding Yang, Zunlei Feng, Mingli Song, and Xinchao Wang. "Factorizable graph convolutional networks". In: *Advances in Neural Information Processing Systems* (2020).

[106] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. "Learning to simulate complex physics with graph networks". In: *International conference on machine learning*. PMLR. 2020.

[107] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. "Se (3)-transformers: 3d roto-translation equivariant attention networks". In: *Advances in neural information processing systems* (2020).

[108] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. "E (n) equivariant graph neural networks". In: *International conference on machine learning*. PMLR. 2021.

[109] Wenbing Huang, Jiaqi Han, Yu Rong, Tingyang Xu, Fuchun Sun, and Junzhou Huang. "Equivariant Graph Mechanics Networks with Constraints". In: *International Conference on Learning Representations (ICLR)*. 2021.

[110] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2020.

[111] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. "Training data-efficient image transformers & distillation through attention". In: *International conference on machine learning*. PMLR. 2021.

[112] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. "BEiT: BERT Pre-Training of Image Transformers". In: *International Conference on Learning Representations*. 2021.

[113] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. "Improving data association by joint modeling of pedestrian trajectories and groupings". In: *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*. Springer. 2010.

[114] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. "Crowds by Example". In: *Computer Graphics Forum* 26.3 (Sept. 2007), pp. 655–664.

[115] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. "Stgat: Modeling spatial-temporal interactions for human trajectory prediction". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

[116] Yue Hu, Siheng Chen, Ya Zhang, and Xiao Gu. "Collaborative motion prediction via neural motion message passing". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

[117] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris Kitani. "AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021, pp. 9793–9803.

[118] Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. "Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[119] Mikkel Knudsen and Jari Kaivo-oja. "Collaborative Robots: Frontiers of Current Literature". In: *Journal of Intelligent Systems: Theory and Applications* 3 (June 2020), pp. 13–20.

[120] ISO. *ISO/TS 15066:2016. Robots and robotic devices — Collaborative robots*. https://www.iso.org/obp/ui/#iso:std:iso:ts:15066:ed-1:v1:en. 2021.

[121] Diego Rodriguez-Guerra, Gorka Sorrosal, Itziar Cabanes, and Carlos Calleja. "Human-Robot Interaction Review: Challenges and Solutions for Modern Industrial Environments". In: *IEEE Access* 9 (2021), pp. 108557–108578.

[122] Julie Shah, James Wiken, Cynthia Breazeal, and Brian Williams. "Improved human-robot team performance using Chaski, a human-inspired plan execution system". In: *HRI 2011 - Proc. 6th ACM/IEEE Int. Conf. Human-Robot Interact.* (2011), pp. 29–36.

[123] Luca Gualtieri, Ilaria Palomba, Erich J. Wehrle, and Renato Vidoni. *The Opportunities and Challenges of SME Manufacturing Automation: Safety and Ergonomics in Human–Robot Collaboration*. Springer International Publishing, 2020, pp. 105–144.

[124] Björn Matthias and Thomas Reisinger. "Example application of ISO/TS 15066 to a collaborative assembly scenario". In: *47th Int. Symp. Robot. ISR 2016* 2016 (2016), pp. 88–92.

[125] Qiongjie Cui, Huaijiang Sun, and Fei Yang. "Learning Dynamic Relationships for 3D Human Motion Prediction". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 6518–6526.

[126] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. "MSR-GCN: Multi-Scale Residual Graph Convolution Networks for Human Motion Prediction". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021.

[127] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. "Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 211–220.

[128] Xin Li and Dawei Li. "GPFS: A Graph-Based Human Pose Forecasting System for Smart Home with Online Learning". In: *ACM Trans. Sen. Netw.* 17.3 (2021).

[129] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. "Learning Trajectory Dependencies for Human Motion Prediction". In: *The IEEE International Conference on Computer Vision (ICCV)*. 2019.

[130] Chenxi Wang, Yunfeng Wang, Zixuan Huang, and Zhiwen Chen. "Simple Baseline for Single Human Motion Forecasting". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2021, pp. 2260–2265.

[131] Yang Zhao and Yong Dou. "Pose-Forecasting Aided Human Video Prediction With Graph Convolutional Networks". In: *IEEE Access* 8 (2020), pp. 147256–147264.

[132] Guokun Lai, Hanxiao Liu, and Yiming Yang. "Learning Depthwise Separable Graph Convolution from Data Manifold". In: *abs/1710.11577* (2018).

[133] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Mo Zhou, Zhenxing Niu, and Gang Hua. "Sparse Graph Convolution Network for Pedestrian Trajectory Prediction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

[134] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. "History Repeats Itself: Human Motion Prediction via Motion Attention". In: *The European Conference on Computer Vision (ECCV)*. 2020.

[135] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (2014).

[136] Praveen Kumar Reddy Maddikunta, Quoc-Viet Pham, Prabadevi B, N Deepa, Kapal Dev, Thippa Reddy Gadekallu, Rukhsana Ruby, and Madhusanka Liyanage. "Industry 5.0: A survey on enabling technologies and potential applications". In: *Journal of Industrial Information Integration* 26 (Mar. 2022), p. 100257.

[137] Shirine El Zaatari, Mohamed Marei, Weidong Li, and Zahid Usman. "Cobot programming for collaborative industrial tasks: An overview". In: *Robotics and Autonomous Systems* 116 (2019), pp. 162–180.

[138] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. "Deep 3D human pose estimation: A review". In: *Computer Vision and Image Understanding* 210 (2021), p. 103225.

[139] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. "Habitat 3.0: A co-habitat for humans, avatars and robots". In: *arXiv preprint arXiv:2310.13724* (2023).

[140] Priyaranjan Biswal and Prases K Mohanty. "Development of quadruped walking robots: A review". In: *Ain Shams Engineering Journal* 12.2 (2021), pp. 2017–2031.

[141] Wolfgang Merkt, Vladimir Ivan, Yiming Yang, and Sethu Vijayakumar. "Towards shared autonomy applications using whole-body control formulations of locomanipulation". In: *Proceedings of the IEEE International Conference on Automation Science and Engineering*. 2019, pp. 1206–1211.

[142] Matthias Guertler, Laura Tomidei, Nathalie Sick, Marc Carmichael, Gavin Paul, Annika Wambsganss, Victor Hernandez Moreno, and Sazzad Hussain. "WHEN IS A ROBOT A COBOT? MOVING BEYOND MANUFACTURING AND ARM-BASED COBOT MANIPULATORS". In: *Proceedings of the Design Society* 3 (2023), pp. 3889–3898.

[143] Srijeet Halder, Kereshmeh Afsari, Erin Chiou, Rafael Patrick, and Kaveh Akbari Hamed. "Construction inspection & monitoring with quadruped robots in future human-robot teaming: A preliminary study". In: *Journal of Building Engineering* 65 (2023), p. 105814.

[144] Seyed S Mohammadi, Nuno F Duarte, Dimitrios Dimou, Yiming Wang, Matteo Taiana, Pietro Morerio, Atabak Dehban, Plinio Moreno, Alexandre Bernardino, Alessio Del Bue, et al. "3dsgrasp: 3d shape-completion for robotic grasp". In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 3815–3822.

[145] Nuno Ferreira Duarte, Mirko Raković, Jovica Tasevski, Moreno Ignazio Coco, Aude Billard, and José Santos-Victor. "Action anticipation: Reading the intentions of humans and robots". In: *IEEE Robotics and Automation Letters* 3.4 (2018), pp. 4132–4139.

[146] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling". In: *ArXiv* abs/1803.01271 (2018).

[147] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. "Convolutional Sequence to Sequence Learning". In: *The International Conference on Machine Learning (ICML)*. 2017.

[148] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. "Convolutional Sequence to Sequence Model for Human Dynamics". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[149] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. "3D human pose estimation in video with temporal convolutions and semi-supervised training". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[150] Federico Cunico, Federico Girella, Andrea Avogaro, Marco Emporio, Andrea Giachetti, and Marco Cristani. "OO-dMVMT: A Deep Multi-View Multi-Task Classification Framework for Real-Time 3D Hand Gesture Classification and Segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2023, pp. 2745–2754.

[151] Marco Emporio, Ariel Caputo, Deborah Pintani, Federico Cunico, Federico Girella, Andrea Avogaro, Marco Cristani, and Andrea Giachetti. "gesture based interaction with the Hololens 2". In: *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter*. 2023, pp. 1–2.

[152] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. "Recurrent Network Models for Human Dynamics". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 4346–4354.

[153] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G. Ororbia. "A Neural Temporal Model for Human Motion Prediction". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 12108–12117.

[154] Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. "Structural-RNN: Deep Learning on Spatio-Temporal Graphs". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 5308–5317.

[155] Judith Bütepage, Hedvig Kjellström, and Danica Kragic. "Anticipating many futures: Online human motion prediction and synthesis for human-robot collaboration". In: *ArXiv* abs/1702.08212 (2017).

[156] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. "Structured Prediction Helps 3D Human Motion Modelling". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019.

[157] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, Ding Liu, Jing Liu, and Nadia Magnenat Thalmann. "Learning Progressive Joint Propagation for Human Motion Prediction". In: *The European Conference on Computer Vision (ECCV)*. 2020.

[158] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*. 2017.

[159] Muhammet Balcilar, Guillaume Renton, Pierre Héroux, Benoit Gaüzère, Sébastien Adam, and Paul Honeine. "Spectral-designed depthwise separable graph neural networks". In: *Proceedings of Thirty-seventh International Conference on Machine Learning (ICML 2020)-Workshop on Graph Representation Learning and Beyond (GRL+ 2020)*. 2020.

[160] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. "AMASS: Archive of Motion Capture as Surface Shapes". In: *International Conference on Computer Vision*. 2019.

[161] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. "Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera". In: *European Conference on Computer Vision (ECCV)*. 2018.

[162] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. "Multi-Person Extreme Motion Prediction with Cross-Interaction Attention". In: *arXiv preprint arXiv:2105.08825* (2021).

[163] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. "Three-dimensional reconstruction of human interactions". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 7214–7223.

[164] Mejdi Dallel, Vincent Havard, David Baudry, and Xavier Savatier. "InHARD - Industrial Human Action Recognition Dataset in the Context of Industrial Collaborative Robotics". In: *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. 2020.

[165] Juan Antonio Corrales Ramon, Francisco A. Candelas Herias, and Fernando Torres. "Safe human-robot interaction based on dynamic sphere-swept line bounding volumes". In: *Robot. Comput. Integr. Manuf.* 27.1 (2011), pp. 177–185.

[166] Andrea Bauer, Dirk Wollherr, and Martin Buss. "Human–robot collaboration: a survey". In: *International Journal of Humanoid Robotics* 5.01 (2008), pp. 47–66.

[167] Afonso Castro, Filipe Silva, and Vitor Santos. "Trends of Human-Robot Collaboration in Industry Contexts: Handover, Learning, and Metrics". In: *Sensors* 21.12 (2021), p. 4113.

[168] Juan Alberto Garcia-Esteban, Luis Piardi, Paulo Leitao, Belen Curto, and Vidal Moreno. "An interaction strategy for safe human Co-working with industrial collaborative robots". In: *Proc. - 2021 4th IEEE Int. Conf. Ind. Cyber-Physical Syst. ICPS 2021* (2021), pp. 585–590.

[169] Kai Lemmerz, Paul Glogowski, Phil Kleineberg, Alfred Hypki, and Bernd Kuhlenkötter. "A Hybrid Collaborative Operation for Human-Robot Interaction Supported by Machine Learning". In: *Int. Conf. Hum. Syst. Interact. HSI* 2019-June (2019), pp. 69–75.

[170] Jen Hao Chen and Kai Tai Song. "Collision-Free Motion Planning for Human-Robot Collaborative Safety under Cartesian Constraint". In: *IEEE Int. Conf. Robot. Autom.* (2018), pp. 4348–4354.

[171] Akira Kanazawa, Jun Kinugawa, and Kazuhiro Kosuge. "Adaptive Motion Planning for a Collaborative Robot Based on Prediction Uncertainty to Enhance Human Safety and Work Efficiency". In: *IEEE Trans. Robot.* 35.4 (2019), pp. 817–832.

[172] Hugo Nascimento, Martin Mujica, and Mourad Benoussaad. "Collision avoidance in human-robot interaction using kinect vision system combined with robot's model and data". In: *IEEE Int. Conf. Intell. Robot. Syst.* (2020), pp. 10293–10298.

[173] Emmanuel P. Beltran, Arthur Akira S. Diwa, Benedict Troy B. Gales, Christian E. Perez, Carlo Antonio A. Saguisag, and Kanny Krizzy D. Serrano. "Fuzzy Logic-based Risk Estimation for Safe Collaborative Robots". In: *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM)*. 2018, pp. 1–5.

[174] Marco Costanzo, G. De Maria, Gaetano Lettera, and Ciro Natale. "A Multimodal Approach to Human Safety in Collaborative Robotic Workcells". In: *IEEE Transactions on Automation Science and Engineering* PP (Jan. 2021), pp. 1–15.

[175] Sangseung Kang, Minjong Kim, and Kyekyung Kim. "Safety Monitoring for Human Robot Collaborative Workspaces". In: *Int. Conf. Control. Autom. Syst.* 2019-October.Iccas (2019), pp. 1192–1194.

[176] Jiwoong Lim, Jihyun Lee, Changjoo Lee, Gunwoo Kim, Younghoon Cha, Joonhyung Sim, and Sungsoo Rhim. "Designing path of collision avoidance for mobile manipulator in worker safety monitoring system using reinforcement learning". In: *ISR 2021 - 2021 IEEE Int. Conf. Intell. Saf. Robot.* (2021), pp. 94–97.

[177] Chris Torkar, Saeed Yahyanejad, Horst Pichler, Michael Hofbaur, and Bernhard Rinner. "RNN-based human pose prediction for human-robot interaction". In: *Proceedings of the ARW & OAGM Workshop 2019*. 2019, pp. 76–80.

[178] Jianjing Zhang, Hongyi Liu, Qing Chang, Lihui Wang, and Robert X Gao. "Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly". In: *CIRP annals* 69.1 (2020), pp. 9–12.

[179] Lorenzo Vianello, Jean-Baptiste Mouret, Eloise Dalin, Alexis Aubry, and Serena Ivaldi. "Human Posture Prediction during Physical Human-Robot Interaction". In: *IEEE Robotics and Automation Letters* (2021).

[180] Javier Laplaza, Albert Pumarola, Francesc Moreno-Noguer, and Alberto Sanfeliu. "Attention deep learning based model for predicting the 3D Human Body Pose using the Robot Human Handover Phases". In: *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE. 2021, pp. 161–166.

[181] Andrey Rudenko, Tomasz P. Kucner, Chittaranjan S. Swaminathan, Ravi T. Chadalavada, Kai O. Arras, and Achim J. Lilienthal. "THÖR: Human-Robot Navigation Data Collection and Accurate Motion Trajectories Dataset". In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 676–682.

[182] Tim Schreiter, Tiago Rodrigues de Almeida, Yufei Zhu, Eduardo Gutierrez Maestro, Lucas Morillo-Mendez, Andrey Rudenko, Tomasz P Kucner, Oscar Martinez Mozos, Martin Magnusson, Luigi Palmieri, et al. "The magni human motion dataset: Accurate, complex, multi-modal, natural, semantically-rich and contextualized". In: *arXiv preprint arXiv:2208.14925* (2022).

[183] Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. "Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[184] Zhi Yan, Li Sun, Tom Duckctr, and Nicola Bellotto. "Multisensor online transfer learning for 3d lidar-based human detection with a mobile robot". In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2018, pp. 7635–7640.

[185] Fernando Amodeo, Noé Pérez-Higueras, Luis Merino, and Fernando Caballero. "FROG: A new people detection dataset for knee-high 2D range finders". In: *arXiv preprint arXiv:2306.08531* (2023).

[186] Arthur Zhang, Chaitanya Eranki, Christina Zhang, Ji-Hwan Park, Raymond Hong, Pranav Kalyani, Lochana Kalyanaraman, Arsh Gamare, Arnav Bagad, Maria Esteva, et al. "Towards Robust Robot 3D Perception in Urban Environments: The UT Campus Object Dataset". In: *arXiv preprint arXiv:2309.13549* (2023).

[187] Xiaoxiong Zhang, Adarsh Ghimire, Sajid Javed, Jorge Dias, and Naoufel Werghi. "Robot-Person Tracking in Uniform Appearance Scenarios: A New Dataset and Challenges". In: *IEEE Transactions on Human-Machine Systems* (2023).

[188] Mejdi Dallel, Vincent Havard, David Baudry, and Xavier Savatier. "Inhard-industrial human action recognition dataset in the context of industrial collaborative robotics". In: *Proceedings of the IEEE International Conference on Human-Machine Systems*. 2020, pp. 1–6.

[189] Edward Vendrow, Duy Tho Le, Jianfei Cai, and Hamid Rezatofighi. "JRDB-pose: A large-scale dataset for multi-person pose estimation and tracking". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 4811–4820.

[190] Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. "SACSoN: Scalable Autonomous Control for Social Navigation". In: *IEEE Robotics and Automation Letters* (2023).

[191] Snehesh Shrestha, Yantian Zha, Saketh Banagiri, Ge Gao, Yiannis Aloimonos, and Cornelia Fermuller. "NatSGD: A Dataset with Speech, Gestures, and Demonstrations for Robot Learning in Natural Human-Robot Interaction". In: *arXiv preprint arXiv:2403.02274* (2024).

[192] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. "Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation". In: *IEEE Robotics and Automation Letters* 7.4 (2022), pp. 11807–11814.

[193] Yuhao Chen, Yue Luo, Chizhao Yang, Mustafa Ozkan Yerebakan, Shuai Hao, Nicolas Grimaldi, Song Li, Read Hayes, and Boyi Hu. "Human mobile robot interaction in the retail environment". In: *Scientific Data* 9.1 (2022), p. 673.

[194] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. "Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction". In: *The European Conference on Computer Vision (ECCV)*. 2020.

[195] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. "Is Space-Time Attention All You Need for Video Understanding?" In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2021.

[196] François Chollet. "Xception: Deep Learning with Depthwise Separable Convolutions". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1800–1807.

[197] Kenta Oono and Taiji Suzuki. "Graph Neural Networks Exponentially Lose Expressive Power for Node Classification". In: *International Conference on Learning Representations*. 2020.

[198] VYann LeCun, John Denker, and Sara Solla. "Optimal Brain Damage". In: *Advances in Neural Information Processing Systems*. 1989.

[199] George Michalos, Sotiris Makris, Panagiota Tsarouchi, Toni Guasch, Dimitris Kontovrakis, and George Chryssolouris. "Design considerations for safe human-robot collaborative workplaces". In: *Procedia CIrP* 37 (2015), pp. 248–253.

[200] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. "Voxelpose: Towards multi-camera 3d human pose estimation in wild environment". In: *European Conference on Computer Vision*. Springer. 2020, pp. 197–212.

[201] Julieta Martinez, Michael J. Black, and Javier Romero. "On human motion prediction using recurrent neural networks". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[202] Sebastian Hjorth and Dimitrios Chrysostomou. "Human–robot collaboration in industrial environments: A literature review on non-destructive disassembly". In: *Robotics and Computer-Integrated Manufacturing* 73 (2022), pp. 102–208.

[203] Sami Haddadin, Alin Albu-Schaffer, Mirko Frommberger, Jurgen Rossmann, and Gerd Hirzinger. "The "DLR Crash Report": Towards a standard crash-testing protocol for robot safety-Part I: Results". In: *2009 IEEE International Conference on Robotics and Automation*. IEEE. 2009, pp. 272–279.

[204] Emanuele Magrini, Federica Ferraguti, Andrea Jacopo Ronga, Fabio Pini, Alessandro De Luca, and Francesco Leali. "Human-robot coexistence and interaction in open industrial cells". In: *Robotics and Computer-Integrated Manufacturing* 61 (2020).

[205] Shuying Liu and Weihong Deng. "Very deep convolutional neural network based image classification using small training sample size". In: *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. 2015, pp. 730–734.

[206] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated Residual Transformations for Deep Neural Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5987–5995.

[207] Matthew D. Zeiler and Rob Fergus. “Visualizing and Understanding Convolutional Networks”. In: *European Conference on Computer Vision (ECCV)*. 2014.

[208] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. *XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks*. 2016.

[209] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. *Pruning Convolutional Neural Networks for Resource Efficient Inference*. 2017.

[210] Geoffrey Hinton, Jeff Dean, and Oriol Vinyals. “Distilling the Knowledge in a Neural Network”. In: *NIPS*. 2014, pp. 1–9.

[211] Muyang Li, Ji Lin, Yaoyao Ding, Zhijian Liu, Jun-Yan Zhu, and Song Han. “GAN Compression: Efficient Architectures for Interactive Conditional GANs”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 5283–5293.

[212] Katarina Benesova, Andrej Svec, and Marek Suppa. *Cost-effective Deployment of BERT Models in Serverless Environment*. 2021.

[213] Marco Minelli, Alessio Sozzi, Giacomo De Rossi, Federica Ferraguti, Francesco Setti, Riccardo Muradore, Marcello Bonfè, and Cristian Secchi. “Integrating Model Predictive Control and Dynamic Waypoints Generation for Motion Planning in Surgical Scenario”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2020, pp. 3157–3163.

[214] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. “Deep high-resolution representation learning for human pose estimation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 5693–5703.

[215] Pin-Ling Liu and Chien-Chi Chang. “Simple method integrating OpenPose and RGB-D camera for identifying 3D body landmark locations in various postures”. In: *International Journal of Industrial Ergonomics* 91 (2022), p. 103354.

[216] Yi Yang and Deva Ramanan. “Articulated human detection with flexible mixtures of parts”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.12 (2012), pp. 2878–2890.

[217] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. “Panoptic studio: A massively multiview system for social motion capture”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 3334–3342.

[218] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. “Space-time-separable graph convolutional network for pose forecasting”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11209–11218.

[219] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. "Back to MLP: A simple baseline for human motion prediction". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 4809–4819.

[220] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. "CSDI: Conditional score-based diffusion models for probabilistic time series imputation". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 24804–24816.

[221] Young Jin Heo, Dayeon Kim, Woongyong Lee, Hyoungkyun Kim, Jonghoon Park, and Wan Kyun Chung. "Collision detection for industrial collaborative robots: A deep learning approach". In: *IEEE Robotics and Automation Letters* 4.2 (2019), pp. 740–746.