Eberhard Karls Universität Tübingen

Faculty of Economics and Social Sciences

Master Thesis

Evaluating Perceptual Alignment in Deep Vision Models for Fashion Sales Prediction

Supervisor: Dr. Aseem Behl
Chair of Marketing

Author:
Nils Großepieper
Study program: Data Science in Business and Economics (M.Sc.)
Student ID: 6299477
Email: nils.grossepieper@student.uni-tuebingen.de

Date of submission: 29 December 2025

Contents

List of Figures

List of Tables

1 Introduction

Fast fashion has become increasingly prevalent in recent years, driven by consumers' constant demand for the latest trends. This accelerates product cycles and increases market volatility, as fashion items often have shorter lifespans and consumers' preferences can shift rapidly. These dynamics make demand prediction difficult, particularly because new products usually lack sufficient historical sales data (I. F. Chen et al. 2021, p. 3). Poor sales prediction accuracy can lead to ecological and economic challenges: the fashion industry is responsible for roughly 10% of global greenhouse gas emissions and generates approximately 92 million tons of waste annually (Z. Li et al. 2024, pp. 2–3). Nearly half of all fashion items produced are sold at a discount or never sold at all due to poor sales prediction accuracy (Banerjee et al. 2021, p. 167). Recent research suggests that the application of artificial intelligence in fashion sales prediction could reduce prediction errors by up to 50%, leading to more sustainable production and greater profitability (Banerjee et al. 2021, p. 167).

State-of-the-art approaches to fashion sales prediction incorporate image embeddings from deep vision models (Skenderi et al. 2022, p. 12). This is intuitive, as purchasing decisions in fashion are heavily influenced by visual impressions. Consumers often decide spontaneously based on the visual appearance of the product whether to buy it (Cakici et al. 2020, pp. 74–75). However, a core challenge arises from the fact that deep vision models do not perceive images in the same way as humans do. If an embedding space does not capture the same features as human customers when evaluating a product, the prediction performance may decline.

To address this limitation, Fu et al. (2023, p. 2) introduced the Novel Image Generations with Human-Tested Similarity (NIGHTS) dataset, which contains a human similarity judgment. The goal of this dataset is to align the embeddings of deep vision models more closely with human perception. Previous work showed that fine-tuning deep vision models on the NIGHTS dataset can improve computer vision tasks such as segmentation and image retrieval (Sundaram et al. 2024, p. 2). Despite these promising results, no prior research has examined whether perceptual alignment based on human similarity judgments improves fashion sales prediction. This represents an important gap, as visual perception plays a central role in consumer decision-making.

The main objective of this thesis is to evaluate whether perceptual alignment in deep vision models improves the accuracy of fashion sales prediction. The analysis examines whether perceptually aligned deep vision models outperform standard non-aligned models. The analysis further examines whether domain-specific human similarity judgments provide additional benefits beyond general-purpose data, whether different deep vision architectures and fine-tuning methods respond differently to perceptual alignment, how model capacity influences the effect

of alignment, and whether various prediction model types benefit differently from changes in the embedding space.

To research these questions, the accuracy of sales prediction for new products without prior sales history is investigated. The prediction models are trained mainly with image embeddings and contextual information from previous fashion items, including historical sales numbers. Predictive accuracy based on embeddings from non-aligned deep vision models is compared with accuracy obtained from perceptually aligned models. A detailed description of the datasets, deep vision models, fine-tuning methods, and prediction models is provided in Chapters 3 and 4.

This thesis contributes to the literature in several ways. First, it provides the first empirical evaluation of perceptual alignment in the context of fashion sales prediction, thus expanding research on perceptual alignment beyond traditional computer vision benchmarks. Second, it compares multiple deep vision model architectures, and fine-tuning methods within a unified experimental setting. This allows conclusions to be drawn about how different models respond to alignment. Third, it evaluates whether domain-specific human similarity judgment datasets yield an advantage over a general-purpose dataset that encapsulates human judgment. Fourth, it evaluates the effect of perceptual alignment on deep vision models with different capacities. Finally, it investigates how various prediction model types leverage aligned embedding spaces, providing evidence on whether distance-based, linear, or non-linear prediction models benefit differently from changes in the embedding representation.

The remainder of this thesis is organized into seven chapters. The theoretical background of fashion sales prediction, deep vision models, and perceptual alignment is explained in Chapter 2. In Chapter 3, an overview of the different datasets used in this thesis is given, followed by a chapter that explains the methodology and experiment design of this thesis. Chapter 5 discusses the results, and Chapter 6 places the results in the context of the current research landscape. Finally, Chapter 7 gives a conclusion and summarizes the contributions, limitations, and possible future research. In the appendix, more details about the variable selection, model training, and supplementary results can be found.

2 Background and Related Work

2.1 Fashion Sales Prediction

Fast fashion has become an increasingly significant business model in the fashion industry. The concept of fast fashion is based on the continuous introduction of new items that reflect current trends and popular designs in a fast-changing market (I. F. Chen et al. 2021, p. 2). The fast-fashion market is characterized by several factors that make sales prediction particularly

challenging. First, product life cycles are extremely short, as items are designed to capture the momentary style. Second, replenishment times are long, since most products are manufactured outside the primary consumer markets. Third, purchases are highly impulsive; many customers decide only at the point of sale whether they will buy an item. Fourth, demand is highly volatile, as the fashion market can be disrupted by unpredictable events—for example, when an influencer or celebrity wears a particular item, demand may increase suddenly. Lastly, product assortments change continuously, resulting in a broad and frequently updated range of designs (Nenni et al. 2013, p. 3). Due to the rapidly changing nature of the fast fashion market, there is often insufficient historical sales information for new products. As a result, sales predictions frequently rely on the assumption that visually or stylistically similar products exhibit similar sales patterns (Giri et al. 2022, p. 565).

Machine learning models have been shown to outperform traditional forecasting methods, such as ARIMA or SIMPLE Exponential Smoothing (SES) in fashion sales prediction (Fildes et al. 2022, p. 1322). Traditional methods struggle to capture nonlinear relationships, outliers, or missing values, whereas machine learning models can utilize qualitative information like text descriptions or product images (Giri et al. 2022, p. 567). Skenderi et al. (2024, p. 12) showed that incorporating image embeddings from deep vision models into prediction models such as kNN, gradient boosting, or transformer-based architectures significantly improve fashion sales prediction.

Using image embeddings as input variables for prediction models is important for fashion sales prediction since buying decisions in fashion heavily depend on the visual perception of the customers. An item's perception depends on attributes such as texture, shape, layout, and color, all of which strongly affect impulsive purchasing behavior (Cakici et al. 2020, pp. 74–75). Other important factors such as packaging, product quality, store atmosphere, and brand perception also influence buying decisions. However, these aspects cannot be fully captured in an image of a product and therefore cannot be included in the design of the experiment in this thesis (Zhang et al. 2018, p. 4). Rawat (2024, pp. 12–13) discovered that image embeddings using pre-trained deep vision models can explain half the variation in sales and two-thirds of variation in price.

2.2 Deep Vision Models and Embeddings

According to Dubey (2021, pp. 1–2), deep vision models transform raw images into numerical feature vectors, so-called embeddings. These embeddings summarize the semantic and visual content of the image. Over the past decade, a shift away from hand-engineered feature descriptors toward learned representations has been observed. In traditional computer vision, features such as shape, color, texture, or gradients were manually defined. In contrast, deep vision models automatically learn hierarchical representations that capture increasingly complex

visual patterns. These learned features are more robust to variations in illumination, scale, and viewpoint, and can be transferred across visual tasks. As a result, embeddings have become a standard interface between visual perception and downstream prediction problems (Dubey 2021, pp. 1–2).

To gain a deeper understanding of the effect of human perceptual alignment, this thesis investigates the two most widely used deep vision model architectures: Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). These architectures differ fundamentally in how they process visual information. In a standard CNN, convolutional layers apply learnable kernels that slide across the image to extract local features. Early layers extract low-level visual properties such as edges, textures, and color gradients, while deeper layers extract high-level visual properties such as shapes, objects parts, and complete objects. Pooling layers reduce the spatial resolution, while keeping the most salient information from previous layers. Finally, the flattened representations produced by the convolutional and pooling layers serve as input to the fully connected layers (Maurício et al. 2023, pp. 2–4).

ViTs on the other hand, treat an image as a sequence of patches. Each image is divided into smaller patches, flattened and turned into a vector. The patches receive positional encodings before being fed into a transformer encoder. In the transformer encoder, all patches are processed together, allowing each patch to attend to all other patches to capture global relationships (Dosovitskiy 2020, pp. 2–4).

Raghu et al. (2021, pp. 1–4) highlight the architectural differences and the distinct ways in which the two model types process visual information. CNNs incorporate strong inductive biases due to local connectivity and translation equivariant. CNNs assume that neighboring pixels are related and therefore learn hierarchical features that are built progressively from local to global structures. Because of this architectural design, CNNs tend to focus primarily on local textures. ViTs, by contrast, operate with weaker inductive biases. ViTs model relationships between all patches through global self-attention. They can capture long-range dependencies within an image. As a result, ViTs integrate information globally, while CNNs build local features that gradually expand in scope. A drawback of ViTs is their greater data requirements during training (Raghu et al. 2021, pp. 1–4) and their higher computational demand (Caron et al. 2021, p. 1).

2.3 Perceptual Alignment and Human Judgment

Perceptual distance is a measurement of how similar two images are in a way that they conform with human judgment. Creating a metric that measures perceptual distance is difficult since human judgment is multi-faceted and depends on multiple factors. The human sense of similarity can depend on the context of the image or high-order image structure. For this reason, it might not be possible to capture human judgment by a simple distance metric (Zhang et al. 2018,

pp. 586–587). There is no single form of human judgment with a clearly correct answer: Is a blue circle more similar to a blue square or to a red circle?

It has been shown that deep convolutional networks trained on high-level image classification tasks develop internal activations that reflect perceptual similarity in the embedding space. Perceptual alignment is not represented by a dedicated function, rather it emerges from visual representations that capture meaningful structure in the world. Deep vision models trained on semantic representation tasks often produce embeddings whose Euclidean distance aligns with human judgment of similarity (Zhang et al. 2018, p. 587).

Deep vision models can approximate human perception; however, the way in which deep vision models and humans process images differs fundamentally. Humans can use the context of a scene to guide attention. When observing a scene, humans pay attention to semantic relationships (whether an object's identity fits in a scene) as well as spatial relationships (whether an object's location is appropriate) (Wu et al. 2014, p. 4). On the other hand, the perception of deep vision models depends on local spatial features in the case of CNNs, or on self-attention mechanisms in the case of ViTs (Poonam et al. 2025, p. 1).

Poonam et al. (2025, pp. 9–11) showed that ViTs perform worse than CNNs and significantly worse than humans on low-level perceptual tasks, particularly those involving comparative judgments (position–length, bars and point clouds). Additional studies have identified further systematic differences between human perception and deep vision model representations. (Geirhos et al. 2018, p. 8) found that humans classify objects primarily based on their shape, whereas CNNs trained on ImageNet data rely mainly on an object's texture. Research also indicates that the similarity between human perception and CNN representations varies across layers: deeper layers more closely resemble human perception than earlier layers, although this similarity declines again at the final layer (Peterson et al. 2018, pp. 8–9). Muttenthaler et al. (2022, p. 9) further reported that increasing model size does not enhance alignment in human similarity judgments; however, models trained on more diverse datasets exhibit improved alignment with human visual perception.

Sundaram et al. (2024, p. 2) demonstrated that aligning pretrained deep vision models with human judgment data can improve various downstream tasks, such as depth estimation, segmentation, and retrieval-based tasks. Furthermore, fine-tuning on the NIGHTS dataset led to enhanced performance on fashion retrieval tasks. These fine-tuned models achieved superior results on the Consumer-to-Shop benchmark instance retrieval task of the DeepFashion2 dataset (Sundaram et al. 2024, p. 8).

2.4 Datasets Used in Related Work

In this thesis, the NIGHTS dataset and a small Fashion Triplet (FT) dataset are used to train deep vision models based on human judgment. However, there exist other datasets that captures human judgment. The Berkeley-Adobe Perceptual Patch Similarity (BAPPS) dataset captures low-level conceptual similarity, as it consists of image patches with distortions such as color shifts, blurring, and compression artifacts (Zhang et al. 2018, p. 588). In contrast, the THINGS dataset captures high-level conceptual similarity, as its triplets are constructed across different object categories and require humans to make concept-level similarity judgments (Hebart et al. 2023, p. 4).

One of the most widely used datasets in academia for fashion sales prediction is the Visuelle 2.0 dataset, which is used in this thesis. Alternative datasets used for fashion sales prediction include the H&M Personalized Fashion Recommendations dataset, which contains more than 100,000 fashion product images and millions of customer purchase records (H&M et al. 2022). Another relevant dataset is Styles and Substitutes, which includes millions of product images, metadata, and sales-rank metrics (McAuley et al. 2015, p. 2).

There are many fashion-specific datasets for image retrieval tasks, such as DeepFashion (Liu et al. 2016, p. 1097-1099), DeepFashion2 (Ge et al. 2019, pp. 2–5), and ModaNet (Zheng et al. 2018, pp. 3–4), which can all be used for clothing-specific visual recognition tasks.

2.5 Gap Analysis

Previous studies such as Skenderi et al. (2024, p. 12) and Rawat (2024, pp. 12–13) have shown that image embeddings are essential components of modern sales prediction models. Sundaram et al. (2024, p. 8) further demonstrated that deep vision models fine-tuned on datasets reflecting human similarity judgment achieve improved performance on downstream tasks such as fashion retrieval. However, no prior work has examined whether perceptual alignment learned either from a large, domain-agnostic similarity dataset or from a small, fashion-specific human-judgment dataset can enhance fashion sales prediction.

This thesis addresses this research gap by answering five research questions. In this thesis the effect of perceptual alignment in deep vision models on fashion sales prediction is evaluated. The thesis also examines the effect of different datasets that contain human judgments. In addition, the study systematically compares different deep vision architectures (CNNs and ViTs) and fine-tuning approaches (a Multi-Layer Perceptron head and Low-Rank Adaptation). Further it examines the effect of perceptual alignment on variations in model capacity. Finally, several prediction models (kNN regression, ridge regression, random forest regression, and gradient boosting regression) are compared to examine how models with different learning behaviors

make use of the aligned visual embeddings. This includes investigating whether these prediction models benefit differently when trained on embeddings from perceptually aligned deep vision models. The goal of this thesis is not to identify the optimal prediction model, but rather to quantify the effect of perceptual alignment on fashion sales prediction.

3 Data

3.1 NIGHTS Dataset

The first dataset used for training the deep vision models is the Novel Image Generations with Human-Tested Similarity (NIGHTS) dataset. The NIGHTS dataset captures human similarity judgments in image triplets. Each triplet contains a reference image and two images from which participants were asked to choose which image looks more like the reference image. All three images were created by a modern diffusion model using the same prompt, so each image contains the same object. The images vary in camera poses, viewing angles, or general image layout (Fu et al. 2023, p. 2).

The novelty of the NIGHTS dataset is that it captures mid-level variation in images such as style, color, pose, and other visual factors that humans consider in their visual assessment. Previous datasets primarily focused on low-level image distortions, such as blurring, or on high-level variations related to categorical changes (Fu et al. 2023, pp. 2–3).

All images were generated by the Stable Diffusion v1.4 model and the categories for the prompts were sampled from prominent datasets: ImageNet, CIFAR-10, CIFAR-100, Oxford 102 Flowers, Food-101, and SUN397. This sampling strategy results in a diverse set of categories. Classes containing humans such as baby or man were filtered out due to malformed generated images (Fu et al. 2023, pp. 3–4).

To obtain robust perceptual alignment, three conditions must be met: human judgments should be automatic (requiring little to no conscious reasoning), stable (invariant to changes in mental representation), and shared across individuals (Fu et al. 2023, p. 3). To measure perceptual similarity, participants answered two-alternative forced-choice (2AFC) questions, in which they were asked to select which of two images appeared more similar to a given reference image, without a time constraint. To validate the reliability of these judgments, an additional just noticeable difference (JND) test was conducted, in which participants assessed similarity under time pressure (Fu et al. 2023, pp. 4–5).

Each image triplet was evaluated by at most ten participants, with each participant contributing one vote indicating the image they perceived as more similar to the reference image. Triplets for which fewer than six participants agreed on the same image were discarded, as low agreement

indicates ambiguous perceptual similarity. This filtering step ensured that only triplets with a clear and consistent human judgment were retained for fine-tuning the deep vision models. For the remaining triplets, the winning image received, on average, approximately 70% of the votes. After removing uncertain triplets, the NIGHTS dataset consists of 20,019 image triplets, split into training, validation, and test sets using an 80/10/10 split (the test split is not used since testing is done based on the accuracy of fashion sales prediction) (Fu et al. 2023, pp. 4–5).

3.2 Fashion Triplet Dataset

The Fashion Triplets (FT) dataset was collected by the Marketing Department of the University of Tübingen and consists of image triplets depicting dresses on a white background without a person wearing them. Each triplet contains one reference image and two candidate images, whose similarity to the reference image was judged by participants using a 2AFC task. In total, the dataset comprises 266 image triplets.

Like the NIGHTS dataset, the FT dataset captures human similarity judgments. In contrast to NIGHTS, however, it is domain-specific and exclusively contains images of dresses. The central motivation behind this dataset is that, despite its small size, fashion-specific human similarity judgments may provide more relevant perceptual supervision for fashion-related tasks than larger, domain-agnostic datasets.

On average, each image triplet was evaluated by 56 participants. To ensure reliable perceptual supervision, only triplets for which at least 70% of participants selected the same candidate image as more similar to the reference image were retained, following a procedure analogous to the NIGHTS dataset. This filtering step removed ambiguous triplets with low agreement and ensured that only triplets with a clear and consistent human judgment were used for fine-tuning the deep vision models. After discarding uncertain triplets, the FT dataset consists of 190 image triplets. The data set is divided 90/10 into a training and validation set.

3.3 Visuelle 2.0 Dataset

The datasets described in the previous sections are used to train the deep vision models. For fashion sales prediction, the Visuelle 2.0 dataset (hereafter referred to as Visuelle) is used. Visuelle contains real fashion sales data from the Italian fast fashion company Nuna Lie for six fashion seasons (two fashion seasons per year: Spring/Summer and Autumn/Winter) from 2017 to 2019. The dataset includes sales data for 5,355 clothing items across 110 stores (Skenderi et al. 2022, p. 2241).

For each item, there exists an image of the product on a white background without a person

wearing it and time series data disaggregated at the store level such as sales, inventory stock, normalized price, discounts, and release date. The dataset also includes additional external time series information like weather information and textual tags, which contain information about product category, fabric, and main color of the product. Sales data are available for each item for twelve weeks (Skenderi et al. 2022, pp. 2241–2243).

The predictions in this thesis are performed at the product level rather than at the product–store level. Accordingly, sales numbers for each product are aggregated across all stores in which the product was sold. Using the same image embedding for hundreds of store-level observations with varying sales values would reduce the explanatory power of the visual features, as the prediction model would learn that identical image inputs correspond to inconsistent outcomes (Barz et al. 2020, p. 6). Because of this aggregation, certain store-specific variables, such as release dates or store-level weather information, cannot be used in the analysis.

Since some products are sold in only a few stores, while others are sold in nearly every store, dummy variables for each store were introduced in the Visuelle dataset. This allows the prediction models to account for products that are sold across a varying number of stores. In addition, dummy variables for each sales week were introduced, which enables the prediction models to estimate sales for a specific product in each week.

To ensure that the impact of visual embeddings on forecasting performance remains interpretable, only a limited set of non-visual features was retained for the prediction models. Store-level and week-level dummy variables were included to account for differences in distribution breadth and seasonal sales patterns. Other available variables, such as year, price, fashion season, item category, fabric, and main color, exhibited little to no significant effect when predicting aggregated product-level sales and were therefore excluded from the main analysis. Details on the feature selection procedure are provided in the appendix.

The primary goal of this thesis is to predict the sales performance of previously unseen fashion products, thereby evaluating how well prediction models trained on visual embeddings generalize to new images. For this purpose, the data are split into training, validation, and test sets based on product release dates. Since release dates may vary across stores for the same product, the earliest release date is used for the split.

The 10% of products released most recently form the test set, the preceding 10% constitute the validation set, and the remaining 80% are used for training (Skenderi et al. 2022, p. 2244). This results in 51,444 observations in the training set and 6,408 observations each in the validation and test sets. All three splits contain approximately equal proportions of products from the Spring/Summer and Autumn/Winter seasons.

4 Methodology

4.1 Selected Deep Vision Models

The CNN model used in this thesis is the Residual Network with 50 layers (ResNet-50), introduced by He et al. (2016, p. 1) with the goal of extending the depth of CNNs through residual learning. Instead of learning a direct mapping, each residual block learns a residual function that is added to the input via an identity shortcut, resulting in an output of the form $y = F(x) + x$ (He et al. 2016, pp. 2–3). These shortcut connections mitigate the degradation problem observed in very deep networks, where increasing depth leads to higher training error, by facilitating gradient flow through the layers (He et al. 2016, pp. 1–2). ResNet-50 was trained on the ImageNet-2012 dataset, which contains 1.28 million training images across 1,000 labeled categories. The model was optimized to maximize the predicted probability of the correct class label, encouraging the network to learn category-discriminative features that capture shapes and visual patterns relevant for object recognition (He et al. 2016, pp. 3–4).

Two different ViT–based model types are evaluated in this thesis, which differ substantially in their training objectives and representation learning. The first of these models is Contrastive Language–Image Pretraining (CLIP) ViT. In this thesis the model variants CLIP-B/16 is used. The model consists of two transformer-based encoders: one for images and one for text. A new dataset was constructed to train CLIP, which consists of 400 million image-text pairs collected from publicly available sources on the internet (Radford et al., 2021, pp. 3-4). Unlike ResNet-50, CLIP was not trained to predict fixed class labels. Instead, it was trained using a contrastive learning objective that encourages embeddings of matching image–text pairs to have high cosine similarity, while embeddings of mismatched pairs are pushed apart. This objective is implemented via a cross-entropy–based contrastive loss over batches of image–text pairs (Radford et al. 2021, p. 4). Through this large-scale natural language supervision, CLIP learns a semantically rich and flexible representation that supports zero-shot generalization for unseen tasks. In contrast to category-discriminative features learned by ResNet-50, CLIP embeddings capture higher-level semantic relationships grounded in language, making them particularly relevant for studying perceptual alignment. In this thesis, only the ViT component of CLIP model for image embedding extraction was used. The CLIP base model configuration with twelve self-attention layers, twelve attention heads, and a patch size of 16 was used for experiments (Radford et al. 2021, p. 6).

The other ViT model tested in this thesis is DINO which adopts a fundamentally different training concept compared to CLIP. Caron et al. (2021, pp. 2–4) describe DINO as a self-supervised ViT model trained without human-annotated data. DINO does not predict classes like ResNet-

50; instead, it adopts a self-distillation framework. In this method, a student network learns to match the output distribution of a teacher network that shares the same architecture. Both networks process multiple augmented views of the same image. The student receives both a local and a global crop, while the teacher model only sees the global crop. Their output similarity is optimized using a cross-entropy loss between the normalized probability distributions across K output dimensions, which act as pseudo-class prototypes. The student is updated through standard gradient descent, while the teacher's weights are updated as an exponential moving average (EMA) of the student's parameters (Caron et al. 2021, p. 3-4). In this thesis DINOv1 ViT-B/16 is used which is based on a standard ViT architecture with twelve transformer self-attention blocks and twelve attention heads per block. On top of the backbone, DINO employs a three-layer MLP projection head, followed by L_2 normalization. The model was trained on the ImageNet dataset without labels. The training process followed a standard data augmentation and multi-crop strategy to enforce local-global feature consistency (Caron et al. 2021, p. 5). Through this label-free distillation process, DINO learns semantic and spatial structure in images, such as object boundaries and relationships between objects. DINO's training focuses entirely on intra-visual consistency, which produces embeddings that are perceptually structured and semantically meaningful (Caron et al. 2021, p. 1).

In addition to the DINOv1-ViT-B/16 model, DINOv2-ViT-B/14 and DINOv3-ViT-B/16 are also evaluated in this thesis to test whether newer self-supervised training strategies influence the effect of perceptual alignment. DINOv2 replaces the unlabeled ImageNet dataset with a substantially larger curated dataset containing 142 million images with better data quality and a more diverse set of images (Oquab et al. 2023, pp. 8–9). DINOv3 further scales the training dataset to 1.689 billion images. Additionally, DINOv3 has an increased model size, introduces Gram anchoring to stabilize long-run training and uses a custom ViT architecture with modern positional embeddings (Siméoni et al. 2025, pp. 4–8). Using successive generations of the same model family enables a systematic comparison of how perceptual alignment behaves in increasingly large and modern ViT models.

Taken together, ResNet-50, CLIP, and DINO represent three fundamentally different representation-learning paradigms. ResNet-50 learns visual features through supervised category-level classification, encouraging discriminative representations optimized for object recognition. CLIP, in contrast, learns visual representations through alignment with natural language, embedding images in a semantically grounded space shaped by linguistic supervision. DINO follows a third paradigm, relying entirely on self-supervised learning through intra-visual consistency, without labels or language. Comparing these models enables a systematic analysis of how different training signals influence perceptual alignment. Specifically, it allows assessing the effects of category supervision, language-based semantic supervision, and self-supervised visual learning on the alignment between image embeddings and human similarity judgments.

For ResNet-50, embeddings were extracted from the global average pooling layer by removing the final classification head. For CLIP and DINO, embeddings were obtained from the token of the final transformer layer after the respective projection stages [CLS]. These embeddings serve as the basis for perceptual alignment fine-tuning and as input features for the downstream prediction models (Fu et al. 2023, p. 5).

4.2 Perceptual Alignment Fine-Tuning

To improve perceptual alignment in deep vision models, a fine-tuning stage is introduced in which feature representations are aligned with human perceptual judgments before applying the models to downstream tasks such as fashion sales prediction. The goal is to investigate whether fine-tuning on human similarity datasets leads to improved embeddings for general-purpose visual representations.

During this fine-tuning stage, the parameters θ of a pretrained vision backbone f_θ are optimized using human perceptual similarity judgments. Each training example consists of an image triplet $(x, \tilde{x}_0, \tilde{x}_1)$, where x denotes a reference image and $\tilde{x}_0$ and $\tilde{x}_1$ denote two variant images. All images are embedded using the vision model f_θ (Sundaram et al. 2024, pp. 3–4).

Human judgments are collected via a 2AFC test and encoded as a binary label $y \in \{0, 1\}$, where $y = 0$ indicates that $\tilde{x}_0$ is perceived as more similar to the reference image x, and $y = 1$ indicates that $\tilde{x}_1$ is perceived as more similar. A single training sample is thus defined as $D = \{(x, \tilde{x}_0, \tilde{x}_1), y\}$ (Sundaram et al. 2024, pp. 3–4).

To quantify perceptual similarity, the cosine distance between the embedding of the reference image and each variant image is computed. The cosine distance is defined as:

$$d_i(x, \tilde{x}_i) = 1 - \frac{f_\theta(x) \cdot f_\theta(\tilde{x}_i)}{\|f_\theta(x)\| \, \|f_\theta(\tilde{x}_i)\|} \tag{4.1}$$

The model's preference between the two variant images is expressed by the difference in cosine distances, denoted as Δd. This value indicates which variant is predicted by the deep vision model to be more similar to the reference image:

$$\Delta d = d_0(x, \tilde{x}_0) - d_1(x, \tilde{x}_1) \tag{4.2}$$

For the alignment loss, the binary label $y \in \{0, 1\}$ is mapped to $\bar{y} \in \{-1, 1\}$ to indicate which variant image is perceived as more similar to the reference image. In addition, a margin parameter m is introduced to enforce a minimum separation between similarity scores. In all

experiments, the margin is fixed to $m = 0.05$. The alignment loss is then defined as:

$$L_{\text{alignment}}(\theta) = \max(0,\, m - \Delta d \cdot \bar{y}) \tag{4.3}$$

The alignment loss encourages the embedding of the reference image to be closer to the embedding of the variant image judged as more similar by humans, while increasing the distance to the less similar variant. By optimizing this objective, the model learns an embedding space in which relative similarity relationships derived from human judgments are preserved. This formulation is closely related to the triplet loss commonly used in metric learning, but here the supervision signal is provided directly by human perceptual similarity judgments rather than predefined class labels (Sundaram et al. 2024, pp. 3–4).

Two different fine-tuning methods are tested in this thesis. A multi-layer perceptron (MLP)-based fine-tuning approach is applied to both CNN- and ViT-based models, while Low-Rank Adaptation (LoRA) is applied only to ViT-based models. Both MLP and LoRA use the same alignment loss during fine-tuning (Fu et al. 2023, p. 6).

An MLP is a standard feedforward neural network that is added to the final layer of the vision model. Such networks can approximate any measurable function to arbitrary accuracy when provided with enough hidden units (Hornik et al. 1989, p. 359). In this thesis, all parameters of the pretrained vision backbone are frozen during the alignment stage, and only the parameters of a single hidden-layer MLP head are fine-tuned (Fu et al. 2023, p. 6).

LoRA is a parameter-efficient fine-tuning method that adapts large pretrained models without updating all their weights. Instead of training full weight matrices, LoRA introduces two low-rank trainable matrices A and B for the attention and projection matrices W_q and W_v of the transformer. The adapted weights are given by $W = W_0 + BA$, where W_0 denotes the frozen pretrained weights and BA represents a low-rank update. During training, only the parameters of A and B are optimized, which greatly reduces the number of trainable parameters while maintaining the expressive capacity of the model. This approach introduces only minimal additional computational overhead. By applying low-rank updates within the model's attention and projection layers, LoRA enables efficient adaptation with reduced resource requirements while preserving stable model behavior. In this thesis, LoRA serves as an alternative alignment method to the MLP head, enabling a comparison between internal and external adaptation of visual embeddings (Hu et al. 2022, pp. 1–5)

4.3 Selected Prediction Models

This section covers four prediction models that use image embeddings and dummy variables to predict the fashion sales numbers. The effect of perceptual alignment is tested based on k-Nearest Neighbors (kNN) regression, ridge regression (RR), random forest (RF) regression, and gradient boosting (GB) regression.

The first prediction model used in this thesis is kNN regression. It predicts the sales of a new product by exploiting similarity relationships in the embedding space. This approach assumes that products with similar characteristics, such as visual appearance and sales-related context (e.g., the number of stores in which a product is sold) exhibit similar sales patterns. kNN regression is a nonparametric method in which the sales predictions for a new product are computed as a weighted average of the sales outcomes of the k most similar products from a historical dataset. The k-nearest neighbors are determined using cosine similarity between the feature representations of the new product and products from the historical dataset (Ekambaram et al. 2020, pp. 3112–3113). Since kNN relies on local distances in the embedding space, its performance directly depends on how well perceptual alignment preserves meaningful similarity relationships between the products.

The next prediction model is ridge regression, which extends ordinary least squares regression by introducing an L_2 penalty on the regression coefficients. This regularization stabilizes predictions in settings with highly correlated or high-dimensional input variables. Ridge regression minimizes the sum of the squared prediction error and a penalty on the magnitude of the coefficients, which shrinks coefficients toward zero and reduces variance at the cost of increased bias. This bias–variance trade-off helps to prevent overfitting by limiting the influence of noisy or weakly informative features (Hastie 2020, pp. 426–428). Ridge regression captures linear relationships between embedding features and the target variable.

The third model used in this thesis is the random forest regression, which is an ensemble learning method. Random forests combine the predictions of many randomized decision trees, where each tree is trained on a random subset of training samples and input features. This process introduces diversity through bootstrapping and feature subsampling. The final prediction is obtained by averaging the outputs of all individual trees. Compared to a single tree-based model, this ensemble approach reduces variance and mitigates overfitting (Biau 2012, pp. 1063–1064). Random forest models can capture non-linear relationships and interaction effects. In the context of perceptual alignment, this allows the model to learn more complex dependencies between the embedding space and the target variable than linear or local proximity–based models.

The final prediction model used in this thesis is gradient boosting regression. Like random forests, gradient boosting is an ensemble method, but it differs fundamentally in how the ensem-

ble is constructed. Instead of training trees independently, gradient boosting builds the model sequentially by adding weak learners that correct the residual errors of the current ensemble. Each new tree is trained to be maximally correlated with the negative gradient of the loss function, effectively performing gradient descent in function space (Natekin et al. 2013, pp. 1–3). Through this sequential error-correction mechanism, gradient boosting can capture complex non-linear interactions and refine decision boundaries in the embedding space. This makes it particularly well suited for assessing whether perceptually aligned embeddings provide more informative and structured signals for sales prediction.

All experiments are conducted with four different prediction models for two main reasons. The first reason is robustness and generalization. Evaluating multiple model types ensures that any observed effects of perceptual alignment are not specific to a single prediction approach, but instead reflect a consistent pattern across different model families.

The second reason for using multiple prediction model types is to examine how perceptual alignment affects different predictive structures within the embedding space. Specifically, this setup allows an assessment of whether perceptual alignment benefits are primarily captured by local proximity–based models (kNN), linear mappings (ridge regression), or more expressive non-linear ensemble models (random forest and gradient boosting). By covering models with fundamentally different inductive biases, this evaluation provides insight into how perceptually aligned embeddings interact with both simple and complex prediction mechanisms.

4.4 Experimental Protocol

This section describes the experimental protocol used in this thesis. It first provides a detailed overview of the training pipeline applied to the different deep vision models, embedding extraction, and downstream prediction models. A schematic overview of this pipeline is shown in Figure 4.1. The second part of this section describes how experimental results are evaluated and how statistical significance is assessed.

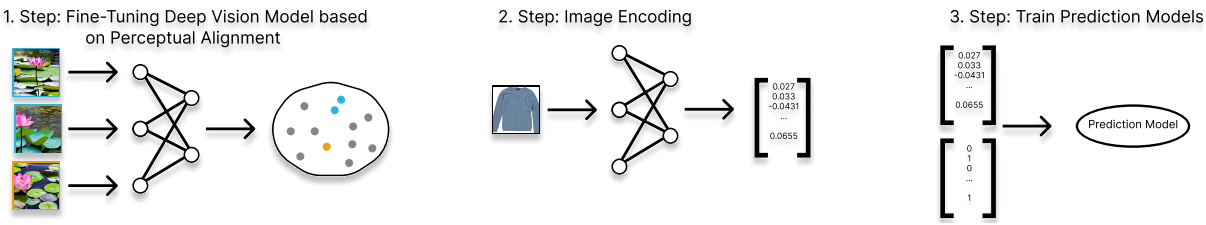

Figure 4.1: Overview of the training pipeline used in this thesis, illustrating perceptual alignment fine-tuning of deep vision models, subsequent embedding extraction and prediction models training for fashion sales prediction.

To test the effect of perceptual alignment, a diverse set of deep vision models was chosen. The

models used in this thesis are ResNet-50, CLIP-B/16, and different versions of the DINO model; DINOv1 ViT-B/16, DINOv2 ViT-B/14, and DINOv3 ViT-B/16.

For each model, multiple alignment conditions were considered: a non-aligned baseline, fine-tuning on the NIGHTS dataset, fine-tuning on the FT dataset, and fine-tuning on a mixed dataset in which 190 NIGHTS triplets were replaced with all 190 FT triplets. ViT models were fine-tuned using both MLP-based heads and LoRA, whereas the CNN models were fine-tuned using an MLP head only.

To improve robustness, each model configuration was trained three times using different random seeds on the same train–validation split. For experiments involving the FT dataset, different train–validation splits were used across runs due to the limited dataset size.

After fine-tuning, image embeddings are extracted for all products in the Visuelle dataset using each deep vision model. To ensure comparability across models and to reduce computational complexity, the resulting embeddings are reduced to 256 dimensions using principal component analysis (PCA). This dimensionality reduction serves multiple purposes: it mitigates noise in the embedding space, alleviates the curse of dimensionality, and accelerates the training of downstream prediction models (Jolliffe et al. 2016, p. 1).

The prediction models take as input the 256-dimensional image embeddings, 110 store-level dummy variables, and twelve week-level dummy variables. Image embeddings are normalized to unit length, while dummy variables are centered at zero and scaled by the standard deviation of the embedding features. This normalization is particularly important for scale-sensitive models such as kNN regression and ridge regression.

Subsequently, the four prediction models are trained. For each experimental setting, hyperparameter tuning is conducted separately for each prediction model to ensure a fair comparison across different vision-model embeddings. This procedure prevents performance differences from arising due to favorable hyperparameter choices for specific embeddings rather than from the embeddings themselves. All prediction models are optimized using mean squared error (MSE) as the training objective.

After training, each prediction model is evaluated based on its predictive performance on the test dataset. Model performance is assessed using the mean absolute error (MAE) and the weighted absolute percentage error (WAPE), which are computed over the entire test set.

The mean absolute error (MAE) measures the average magnitude of prediction errors without regard to their direction. It is computed as the mean of the absolute differences between predicted and observed values, providing an intuitive measure of prediction accuracy in the same units as the target variable (here, the number of fashion products sold).

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{4.4}$$

The weighted absolute percentage error (WAPE) expresses the total absolute prediction error as a percentage of the total observed values. By normalizing the absolute error by the sum of true values, WAPE enables meaningful comparisons across datasets or products with different sales volumes.

$$\text{WAPE} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{\sum_{i=1}^{N} y_i} \tag{4.5}$$

In the final step, the results are evaluated to determine whether prediction models using embeddings from perceptually aligned deep vision models differ in performance from models using embeddings from non-aligned deep vision models. Statistical significance is assessed using paired bootstrap resampling, a non-parametric method for comparing two models evaluated on the same test set (Koehn 2004, pp. 4–7).

Let $e^{(i)}_{\text{aligned}}$ and $e^{(i)}_{\text{non-aligned}}$ denote the prediction errors (e.g., MAE or WAPE) of the aligned and non-aligned models, respectively, for product i. For each product, the paired difference in prediction error is computed as:

$$\Delta_i = e^{(i)}_{\text{aligned}} - e^{(i)}_{\text{non-aligned}}. \tag{4.6}$$

Negative values of Δ_i indicate lower prediction error for the aligned model, whereas positive values indicate worse performance.

The null hypothesis tested in this thesis is

$$H_0 : \mathbb{E}[\Delta] = 0, \tag{4.7}$$

and the two-sided alternative hypothesis

$$H_1 : \mathbb{E}[\Delta] \neq 0. \tag{4.8}$$

This hypothesis test allows for the possibility that perceptual alignment may either improve or degrade predictive performance.

Paired bootstrap resampling is performed by repeatedly drawing, with replacement, n products from the test set to form B bootstrap samples. For each bootstrap sample b, the mean prediction error difference is computed as:

$$\bar{\Delta}^{(b)} = \frac{1}{n} \sum_{i=1}^{n} \Delta_i^{(b)}. \tag{4.9}$$

The empirical distribution of $\{\bar{\Delta}^{(1)}, \ldots, \bar{\Delta}^{(B)}\}$ is used to assess statistical significance. A two-sided bootstrap p-value is estimated as:

$$p = \frac{2}{B} \min \left(\sum_{b=1}^{B} \mathbb{I}\big(\bar{\Delta}^{(b)} \leq 0\big), \sum_{b=1}^{B} \mathbb{I}\big(\bar{\Delta}^{(b)} \geq 0\big) \right), \tag{4.10}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

A difference in performance is considered statistically significant at level $\alpha = 0.05$ if the corresponding two-sided bootstrap confidence interval for $\bar{\Delta}$ does not include zero (equivalently, if $p < 0.05$). This testing framework is used throughout the experimental analysis to assess whether perceptual alignment yields statistically significant differences in fashion sales prediction performance.

5 Results

This chapter presents the results of an experimental evaluation of perceptual alignment in deep vision models for fashion sales prediction. Prediction models trained on perceptually aligned embeddings are compared to models using non-aligned embeddings to assess whether fine-tuning on human perceptual similarity judgments improves downstream sales prediction performance.

The analysis addresses five research questions. First, it is examined whether perceptual alignment improves fashion sales prediction accuracy compared to non-aligned embeddings. Second, it is examined whether the effect of perceptual alignment differs across perceptual judgment datasets (NIGHTS, FT, and a mixed setting in which 190 NIGHTS samples are replaced with FT samples). Third, it examines whether the effect of perceptual alignment varies across deep vision models and fine-tuning methods. Fourth, it tests whether the complexity of the deep vision model influences the effectiveness of perceptual alignment for fashion sales prediction. Finally, it analyzes whether the impact of perceptual alignment depends on the complexity of the downstream prediction model.

Prediction performance is evaluated using mean absolute error (MAE) and weighted absolute

percentage error (WAPE), where lower values indicate better performance. Statistical significance is assessed using paired bootstrap resampling. Performance differences are marked according to their p-values: $p < 0.05(*)$, $p < 0.01(**)$, and $p < 0.001(***)$, indicating whether models using aligned embeddings outperform their non-aligned counterparts.

The results are organized into four sections. The first section analyzes the performance of ResNet-50, the second examines CLIP-B/16, and the third focuses on DINOv1-ViT-B/16. In the fourth section, DINOv1-ViT-B/16, DINOv2-ViT-B/14, and DINOv3-ViT-B/16 are compared to assess how model scale and training data influence the impact of perceptual alignment.

Figure 5.1 provides an illustrative comparison of sales predictions for three example fashion items based on embeddings extracted from the DINOv1 vision model. Predictions using non-aligned embeddings are contrasted with those based on perceptually aligned embeddings fine-tuned with LoRA on the NIGHTS and FT datasets across all four prediction models. A systematic quantitative evaluation of all model configurations follows in the subsequent sections.

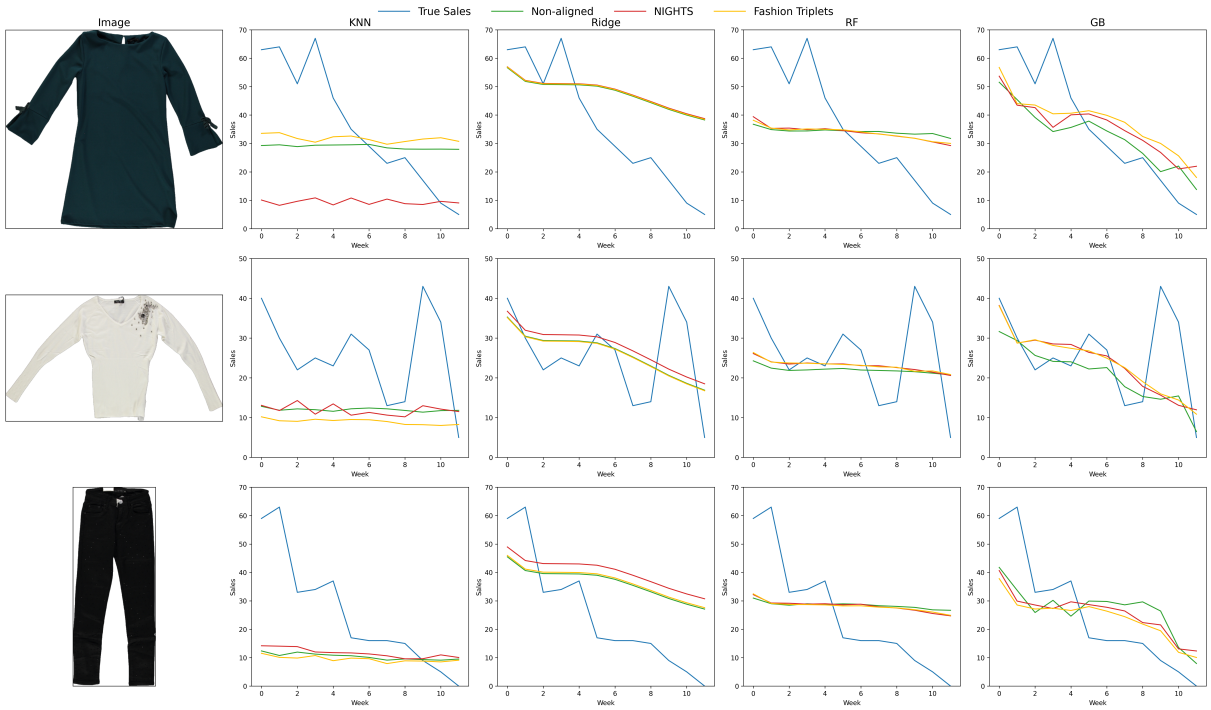

Figure 5.1: Sales predictions for three example fashion items based on embeddings from the DINOv1-ViT-B/16 vision model. Predictions are shown for non-aligned embeddings and for embeddings fine-tuned using LoRA on the NIGHTS and FT datasets, across four different prediction models.

5.1 Results for ResNet-50

Table 5.1 presents the MAE of the four prediction models trained on embeddings extracted from different ResNet-50 variants. No consistent trend can be observed regarding whether fine-tuning on NIGHTS, FT, or a mixed dataset yields the strongest improvements. For instance, the random

forest model achieves its best performance when using embeddings fine-tuned on NIGHTS only, whereas for gradient boosting the lowest MAE is obtained when using embeddings fine-tuned on the combined NIGHTS and FT dataset. This pattern, in which no single human judgment dataset consistently outperforms the others, is also observed for the remaining deep vision models.

Furthermore, the results show that differences in prediction performance are driven more strongly by the choice of downstream prediction model than by the specific embedding variant. The variance in MAE between prediction models is substantially larger than the variance between different ResNet-50 embeddings within the same prediction model. In particular, kNN regression shows no statistically significant benefit from perceptual alignment, while ridge regression exhibits only minor improvements. In contrast, random forest regression and gradient boosting regression achieve statistically significant performance gains when trained on perceptually aligned embeddings, indicating that perceptual alignment primarily benefits more expressive prediction models. Across all experiments, kNN regression yields the highest MAE values, followed by ridge regression, random forest regression, and gradient boosting regression, which consistently achieves the lowest MAE. These performance patterns are observed consistently across all evaluated deep vision models.

One aspect that stands out in the ResNet-50 results is that perceptual alignment yields a stronger improvement in prediction performance compared to the ViT models evaluated in this thesis. The best-performing gradient boosting model (trained on the NIGHTS and FT combination) achieves, on average, an MAE reduction of 0.816 per product per week compared to the same model trained on non-aligned ResNet-50 embeddings. This represents the largest performance gain attributable to perceptual alignment observed among the evaluated deep vision models.

Table 5.1: Global mean absolute error (MAE) across prediction models using embeddings from **ResNet-50** with an **MLP** head.

Model	**kNN**	**Ridge**	**RF**	**GB**
ResNet-50-Non-Aligned	**18.977**	17.008	15.165	14.344
ResNet-50MLP-NIGHTS	19.192	16.968	**14.631*****	13.624**
ResNet-50MLP-FT	19.134	**16.924***	14.875***	13.617***
ResNet-50MLP-N->FT190	19.230	16.974	14.810***	**13.528*****

Note: Statistical significance was assessed using a paired bootstrap test (5 000 resamples). Symbols *, **, *** indicate significant differences (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$) relative to the vanilla model; positive and negative changes denote improvements or deteriorations, respectively.

Table 5.2 reports the results of the same experiments using WAPE as the evaluation metric.

Across all experiments and deep vision models, MAE and WAPE exhibit consistent performance patterns, with both metrics ranking the prediction models in the same order and indicating similar improvements due to perceptual alignment. This consistency suggests that relative prediction errors scale proportionally with actual sales volumes and that performance differences are not driven by a small number of high-selling products. Based on this consistency, WAPE results for the remaining experiments are reported in the appendix, while MAE is used as the primary evaluation metric in the main text.

Table 5.2: Global weighted absolute percentage error (WAPE) across prediction models using embeddings from **ResNet-50** with an **MLP** head.

Model	**kNN**	**Ridge**	**RF**	**GB**
ResNet-50-Non-Aligned	**51.641**	46.283	41.266	39.032
ResNet-50MLP-NIGHTS	52.224	46.172	**39.814*****	37.072**
ResNet-50MLP-FT	52.068	**46.055***	40.477***	37.056***
ResNet-50MLP-N->FT190	52.329	46.190	40.302***	**36.812*****

Note: Statistical significance was assessed using a paired bootstrap test (5 000 resamples). Symbols *, **, *** indicate significant differences (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$) relative to the vanilla model; positive and negative changes denote improvements or deteriorations, respectively.

5.2 Results for CLIP-B/16

Tables 5.3 and 5.4 summarize the performance of the CLIP deep vision model, with Table 5.3 reporting results for MLP-based fine-tuning and Table 5.4 for LoRA-based fine-tuning. Overall, the results exhibit similar patterns to those observed for ResNet-50; however, the improvements resulting from perceptual alignment are less pronounced. The prediction models trained on non-aligned CLIP embeddings generally outperform those using non-aligned ResNet-50 embeddings. The performance gap between CLIP and ResNet-50 narrows, and in some experiments even reverses after perceptual alignment, as the MAE differences between the fine-tuned models decrease. Perceptual alignment leads to improvements for prediction models trained on CLIP embeddings; however, these improvements are less pronounced than those observed for ResNet-50.

Across prediction models, embeddings obtained via LoRA-based fine-tuning consistently yield slightly better performance than those obtained through MLP-based fine-tuning. One notable exception in the CLIP results is that no statistically significant improvement is observed for gradient boosting regression when using MLP-based fine-tuned embeddings. This behavior

is not observed for LoRA-fine-tuned CLIP embeddings nor for any other deep vision model evaluated in this thesis.

Table 5.3: Global mean absolute error (MAE) across prediction models using embeddings from **CLIP-B/16** with an **MLP** head.

Model	kNN	Ridge	RF	GB
CLIP-B16-Non-Aligned	19.749	16.908	15.060	13.700
CLIP-B16MLP-NIGHTS	20.204	**16.861***	14.877***	13.505
CLIP-B16MLP-FT	**19.687**	16.898	**14.588*****	**13.481**
CLIP-B16MLP-N->FT190	20.406*	16.870	14.764***	13.608

Note: Statistical significance was assessed using a paired bootstrap test (5 000 resamples). Symbols *, **, *** indicate significant differences (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$) relative to the vanilla model; positive and negative changes denote improvements or deteriorations, respectively.

Table 5.4: Global mean absolute error (MAE) across prediction models using embeddings from **CLIP-B/16** fine-tuned with **LoRA**.

Model	kNN	Ridge	RF	GB
CLIP-B16-Non-Aligned	19.749	16.908	15.060	13.700
CLIP-B16LoRA-NIGHTS	20.157	16.831	**14.622*****	**13.429****
CLIP-B16LoRA-FT	20.955***	16.968	14.789***	13.646
CLIP-B16LoRA-N->FT190	**19.360**	**16.805**	14.659***	13.597

Note: Statistical significance was assessed using a paired bootstrap test (5 000 resamples). Symbols *, **, *** indicate significant differences (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$) relative to the vanilla model; positive and negative changes denote improvements or deteriorations, respectively.

5.3 Results for DINOv1-ViT-B/16

Tables 5.5 and 5.6 summarize the performance of the DINOv1 vision model. The DINOv1 model exhibits results that are broadly similar to those observed for CLIP; however, perceptual alignment leads to a larger number of statistically significant improvements. Embeddings extracted from perceptually aligned DINOv1 models result in significant performance gains not only for random forest and gradient boosting regression, but also for ridge regression.

Consistent with the CLIP results, DINOv1 embeddings fine-tuned using LoRA achieve slightly better performance than those fine-tuned using an MLP head. In contrast to the other deep vision models, DINOv1 shows a distinct pattern in which fine-tuning with LoRA on the NIGHTS dataset outperforms fine-tuning on alternative datasets. The best overall prediction performance is achieved by the gradient boosting model trained on embeddings from the DINOv1 model fine-tuned with LoRA on the NIGHTS dataset, resulting in an MAE of 13.360 units per product per week. Compared to the corresponding gradient boosting model using non-aligned DINOv1 embeddings, this represents an improvement in prediction accuracy of 3.52%.

Table 5.5: Global mean absolute error (MAE) across prediction models using embeddings from **DINOv1-ViT-B/16** with an **MLP** head.

Model	kNN	Ridge	RF	GB
DINOv1-vitb16-Non-Aligned	19.875	17.070	14.996	13.848
DINOv1-vitb16MLP-NIGHTS	**19.284****	17.019*	14.820***	13.431***
DINOv1-vitb16MLP-FT	19.491**	17.027*	14.676***	**13.428****
DINOv1-vitb16MLP-N->FT190	19.597	**17.007***	**14.619*****	13.446**

Note: Statistical significance was assessed using a paired bootstrap test (5 000 resamples). Symbols *, **, *** indicate significant differences (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$) relative to the vanilla model; positive and negative changes denote improvements or deteriorations, respectively.

Table 5.6: Global mean absolute error (MAE) across prediction models using embeddings from **DINOv1-ViT-B/16** fine-tuned with **LoRA**.

Model	kNN	Ridge	RF	GB
DINOv1-vitb16-Non-Aligned	19.875	17.070	14.996	13.848
DINOv1-vitb16LoRA-NIGHTS	19.876	**16.900***	**14.627*****	**13.360****
DINOv1-vitb16LoRA-FT	**19.723**	17.031*	14.702***	13.537*
DINOv1-vitb16LoRA-N->FT190	19.890	16.932*	14.846***	13.494**

Note: Statistical significance was assessed using a paired bootstrap test (5 000 resamples). Symbols *, **, *** indicate significant differences (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$) relative to the vanilla model; positive and negative changes denote improvements or deteriorations, respectively.

5.4 Scaling Effects Across DINO Model Variants

Finally, the performance of different DINO model variants is investigated. Table 5.7 reports results for prediction models trained on embeddings from non-aligned DINO models, while Table 5.8 presents results for prediction models using embeddings from DINO models fine-tuned with LoRA on the NIGHTS dataset.

For prediction models trained on non-aligned embeddings, a clear pattern emerges across most model types: with the exception of ridge regression, the prediction models using embeddings from DINOv3 achieve the lowest MAE. In contrast, when comparing prediction models trained on embeddings from DINO models fine-tuned on the NIGHTS dataset, the performance ranking becomes less consistent. For kNN regression, embeddings from DINOv3 continue to perform best. However, for ridge regression and random forest regression, the lowest MAE is achieved by models using embeddings from DINOv2, while for gradient boosting regression the best performance is obtained using embeddings from DINOv1. Overall, fine-tuning on the NIGHTS dataset with LoRA reduces the dominance of the DINOv3 embeddings.

Table 5.7: Global mean absolute error (MAE) across prediction models using embeddings from non-aligned DINO vision models.

Model	kNN	Ridge	RF	GB
DINOv1-vitb16-Non-Aligned	19.875	17.070	14.996	13.848
DINOv2-vitb14-Non-Aligned	19.754	**16.980**	14.745	14.378
DINOv3-vitb16-Non-Aligned	**19.702**	17.121	**14.738**	**13.655**

Note: The reported values correspond to prediction errors obtained using non-aligned deep vision models. No statistical significance testing is performed for these comparisons.

Table 5.8: Global mean absolute error (MAE) across prediction models using embeddings from the **DINO** vision model fine-tuned with **LoRA** on the **NIGHTS** dataset.

Model	kNN	Ridge	RF	GB
DINOv1-vitb16LoRA-NIGHTS	19.876	16.900*	14.627***	**13.360****
DINOv2-vitb14LoRA-NIGHTS	19.585	**16.807***	**14.609****	13.542***
DINOv3-vitb16LoRA-NIGHTS	**19.535**	16.963*	14.804	13.484

Note: Statistical significance was assessed using a paired bootstrap test (5 000 resamples). Symbols *, **, *** indicate significant differences (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$) relative to the vanilla model; positive and negative changes denote improvements or deteriorations, respectively.

Figure 5.2 illustrates the relationship between model size and performance improvements resulting from perceptual alignment. The figure reports MAE for non-aligned and perceptually aligned models across all five deep vision models evaluated in this thesis, based on the gradient boosting prediction model. Results are shown separately for MLP-based and LoRA-based fine-tuning, with all models fine-tuned on the NIGHTS dataset.

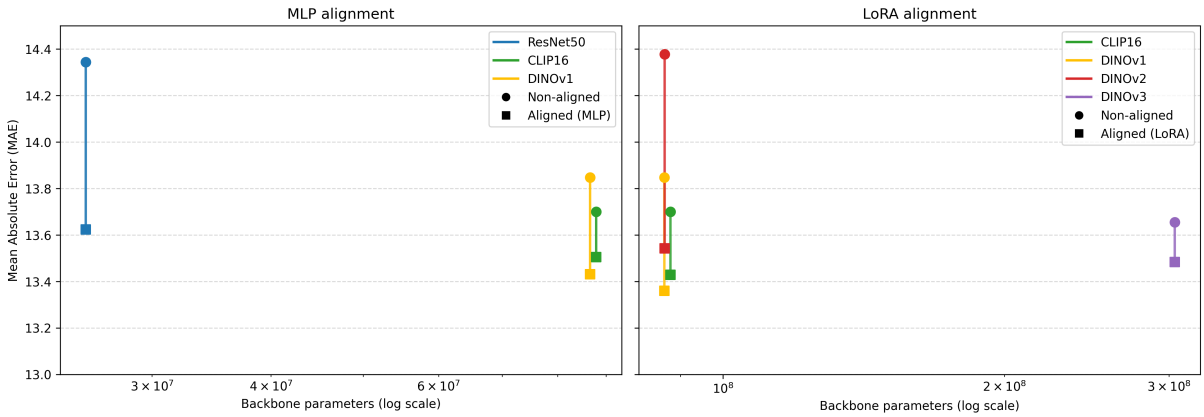

Figure 5.2: Relationship between model size and the effect of perceptual alignment on prediction performance. Mean absolute error (MAE) is shown for non-aligned and perceptually aligned deep vision models across all five deep vision models evaluated in this thesis, based on the gradient boosting prediction model. Results are reported separately for MLP-based and LoRA-based fine-tuning on the NIGHTS dataset.

In summary, the results show that perceptual alignment can improve fashion sales prediction performance. No consistent advantage is observed across the different perceptual judgment datasets (NIGHTS, FT, or their combination), although for DINOv1 a slightly stronger improvement is observed when fine-tuning on NIGHTS. The effect of perceptual alignment tends to be more pronounced for smaller deep vision models with fewer parameters. Both CNN-based and ViT–based models benefit from perceptual alignment.

The choice of fine-tuning method has only a minor influence on prediction performance, although LoRA-based fine-tuning consistently performs slightly better than MLP-based fine-tuning. Most importantly, the results demonstrate that the downstream prediction model plays a more critical role for fashion sales prediction than the choice of the deep vision model itself. Prediction models that can capture non-linear relationships, such as random forest regression and gradient boosting regression, consistently outperform simpler models such as ridge regression and kNN regression. These more expressive models also benefit most strongly from perceptual alignment.

6 Discussion

6.1 Interpretation of Findings

The results of this thesis provide several insights into the effect of perceptual alignment on deep vision models in fashion sales prediction. The four most important findings are discussed in the following section.

The first key finding is that fine-tuning deep vision models on datasets that encapsulate human similarity judgments improves the embedding space for fashion sales prediction. This effect holds for both large, domain-agnostic datasets such as NIGHTS and small, domain-specific datasets such as the fashion triplet (FT) dataset. Replacing 190 randomly selected image triplets from NIGHTS with fashion triplets does not lead to additional improvements compared to fine-tuning exclusively on the NIGHTS dataset. These results demonstrate that human perceptual information is highly sample efficient. Aligning visual representations with human similarity judgments produces embeddings that better capture aspects of fashion items relevant to consumer perception.

The second key finding concerns the interaction between perceptual alignment and the choice of downstream prediction model. Across all experiments, variation in prediction accuracy between different prediction models is substantially larger than variation attributable to the choice of deep vision model embeddings. More expressive prediction models, such as random forest and gradient boosting regression, consistently outperform simpler approaches like kNN and ridge regression, regardless of the embeddings used for training. Importantly, simpler models that rely on local similarity or linear relationships do not benefit significantly from perceptual alignment, whereas non-linear models are able to exploit the refined structure of aligned embedding spaces for more accurate predictions. This pattern suggests that perceptual alignment primarily enhances higher-order relationships within the embedding space, which non-linear prediction models are better equipped to capture.

Another key finding relates to the role of model capacity and architecture in determining the effectiveness of perceptual alignment. Larger and more complex deep vision models generally achieve stronger performance in their non-aligned form, reflecting their ability to capture a broad range of visual semantics. However, perceptual alignment tends to yield stronger relative improvements for smaller and less complex models. As a result, the performance gap between prediction models using embeddings from large and small deep vision models narrows after fine-tuning. In some cases prediction models using embeddings from perceptually aligned smaller deep vision models even outperform their larger counterparts. A similar pattern is observed across architectural families: while both CNNs and ViTs benefit from perceptual alignment,

the effect is more pronounced for CNN-based models. This suggests that models with lower parameter capacity or less expressive baseline representations have greater flexibility to adjust their internal feature spaces in response to human perceptual supervision, whereas larger models experience more limited marginal gains due to already well-structured representations.

The final key finding indicates that LoRA-based fine-tuning yields slightly stronger performance than MLP-based fine-tuning in these experiments. While the magnitude of this difference is modest compared to the overall gains achieved through perceptual alignment, the effect is consistently observable across the evaluated configurations. This suggests that, although the perceptual alignment objective is the primary driver of performance improvements, the choice of adaptation mechanism can provide an additional, small benefits.

Overall, perceptual alignment leads to significant improvements in image embeddings for fashion sales prediction. While the magnitude of these improvements varies across deep vision models, the choice of downstream prediction model plays an even more important role.

6.2 Relation to Previous Work

The findings of this thesis both confirm and extend prior research on perceptual alignment. Sundaram et al. (2024, pp. 5–8) demonstrated that fine-tuning ViT models on human judgment data improves performance across a wide range of downstream tasks, including semantic segmentation, depth estimation, counting, and image retrieval. The results of this thesis are consistent with these findings and extend them by showing that perceptual alignment also improves representation quality in applied forecasting settings, such as fashion sales prediction. This goes beyond classical computer vision benchmarks.

Another finding of this thesis is that random forest and gradient boosting prediction models outperform kNN regression and ridge regression models. This result is consistent with broader findings in machine learning research showing that tree-based ensemble methods, such as random forest and gradient boosting, often outperform other model types on tasks involving tabular or mixed-type data (Grinsztajn et al. 2022, pp. 6–8).

This thesis also shows that larger deep vision models tend to outperform smaller deep vision models in their non-aligned form, but that this difference shrinks after fine-tuning the models on human judgment data. Similar results were found in Sundaram et al. (2024, pp. 5–8), although to a lesser extent. In their study, the fine-tuned DINOv1 model outperformed the non-aligned DINOv1 model in every downstream task tested, while the fine-tuned DINOv2 model only outperformed the non-aligned DINOv2 model in counting and instance retrieval tasks. These findings suggest that more modern (and larger) models do not always benefit proportionally from perceptual alignment and that the relative gains from fine-tuning depend strongly on model capacity

and task characteristics.

6.3 Practical Implications

The findings of this thesis offer several practical implications for both fashion retailers and machine learning researchers. First, the results demonstrate that perceptual alignment has a meaningful and statistically significant effect on fashion sales prediction. This holds true for both small, domain-specific datasets and large, domain-agnostic datasets. In practical terms, the best-performing prediction model achieved an improvement of approximately 3–4% compared to its non-aligned counterpart. Although this improvement may appear modest, it can translate into substantial economic effects in a highly competitive fashion market, where better forecasting helps reduce overproduction, minimize discounting, and optimize inventory planning. The findings further suggest that the choice of prediction model is even more important than the choice of deep vision model. For fashion retailers aiming to reduce forecasting error, improving the downstream prediction model therefore appears more critical than further fine-tuning the vision model that produces the image embeddings.

For machine learning researchers, the results highlight that perceptual alignment has value beyond classical computer vision tasks such as retrieval, segmentation, and depth estimation. The improvements observed in sales prediction indicate that aligning the embedding space with human judgment can improve the capture of semantic information that is relevant for downstream tasks such as fashion sales prediction. The fact that even very small datasets containing human judgment data can improve the embedding space for specific tasks underscores the sample efficiency of human perceptual information. Additionally, the finding that nonlinear models such as random forest regression and gradient boosting regression benefit more from aligned embeddings has implications for model selection in multimodal prediction tasks.

6.4 Limitations

The research conducted in this thesis is primarily limited by the availability and scope of suitable data. All results are based on sales predictions derived from the Visuelle dataset, which contains sales data for a single Italian fashion retailer over the period from 2017 to 2019. As a result, the findings may differ for retailers operating in other segments of the fashion market or under different market conditions. In addition, it remains unclear whether the observed effects of fine-tuning deep vision models would persist over longer time horizons in which more substantial changes in fashion styles occur.

Another limitation concerns the image setup of the Visuelle dataset. All fashion images were captured on a white background without models wearing the garments. Images containing more

complex visual cues, such as different models, poses, or non-neutral backgrounds, could influence the effectiveness of perceptual alignment and may lead to different results.

A different limitation of this thesis concerns the number of images containing human judgment data. In particular, the FT dataset is limited in size due to monetary and time constraints. While the results indicate that even a small domain-specific dataset can positively influence fashion sales prediction, the effect of perceptual alignment may differ for larger domain-specific datasets. Collecting a dataset comparable in size to NIGHTS could therefore provide a more robust assessment of the impact of perceptual alignment in the fashion domain.

Furthermore, the number of deep vision models and prediction models tested in this thesis is limited. Although they capture a wide range of model types, they by no means represent the full spectrum of available architectures. The findings may therefore differ for more recent deep vision models, alternative fine-tuning strategies, or other prediction models not included in this thesis.

7 Conclusion

This thesis investigates the effect of perceptual alignment on deep vision models for fashion sales prediction. The central research question is whether aligning image embeddings with human similarity judgments leads to improved forecasting performance compared to non-aligned deep vision models.

The main findings in this thesis demonstrate that fine-tuning deep vision models using perceptual alignment leads to significant improvements in fashion sales prediction. This effect is observed for both large, domain-agnostic datasets and small, domain-specific datasets. The results further show that the choice of downstream prediction model has a greater influence on forecasting accuracy than the choice of the deep vision backbone. Non-linear prediction models consistently outperform simpler alternatives and are better able to exploit improvements in the embedding space. While larger and more complex deep vision models achieve stronger performance in their non-aligned form, the performance gap between large and small models narrows after fine-tuning, indicating that smaller, less complex models benefit disproportionately from perceptual alignment. These results collectively indicate that human-aligned visual representations capture semantic information relevant for consumer perception and sales dynamics.

This thesis contributes to research on deep vision models and their alignment with human judgment by demonstrating that human perceptual similarity data can serve as a highly sample-efficient supervision signal for representation learning. The findings establish that perceptual alignment meaningfully reshapes embedding spaces in ways that interact with downstream pre-

diction model architectures, highlighting the importance of jointly considering representation learning and predictive modeling. By extending perceptual alignment from classical computer vision benchmarks to the applied domain of fashion sales prediction, this work broadens the empirical scope of alignment-based representation learning. Beyond methodological contributions, the results also have practical relevance for the fashion industry. Improved alignment between visual representations and human perception can support more accurate demand forecasting, improved production planning, optimized discount strategies, and reductions in overproduction, waste, and greenhouse gas emissions.

Future research could extend this work by collecting larger domain-specific human judgment datasets and by evaluating the impact of perceptual alignment across different fashion sales datasets featuring more diverse clothing styles, market segments, and longer observation periods. Additional extensions include investigating alternative deep vision architectures, downstream prediction models, and fine-tuning strategies. Overall, the findings of this thesis indicate that perceptual alignment is a promising and data-efficient approach for improving the embedding spaces of deep vision models in fashion sales prediction. These findings motivate further research into other downstream applications that may benefit from perceptual alignment.

References

Banerjee, Satya Shankar, Sanjay Mohapatra, and Goutam Saha (2021). "Developing a framework of artificial intelligence for fashion forecasting and validating with a case study". In: *International Journal of Enterprise Network Management* 12.2, pp. 165–180.

Barz, Björn and Joachim Denzler (2020). "Do we train on test data? purging cifar of near-duplicates". In: *Journal of Imaging* 6.6, p. 41.

Biau, Gérard (2012). "Analysis of a random forests model". In: *The Journal of Machine Learning Research* 13.1, pp. 1063–1095.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Cakici, A Celil and Sena Tekeli (2020). "Consumers' perceptions of visual product aesthetics based on fashion innovativeness and fashion leadership levels: A research study in Mersin". In: *Journal of Global Business Insights* 5.1, pp. 73–86.

Caron, Mathilde, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin (2021). "Emerging properties in self-supervised vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660.

Chen, I Fei and Chi Jie Lu (2021). "Demand forecasting for multichannel fashion retailers by integrating clustering and machine learning algorithms". In: *Processes* 9.9, p. 1578.

Dosovitskiy, Alexey (2020). "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929*.

Dubey, Shiv Ram (2021). "A decade survey of content based image retrieval using deep learning". In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.5, pp. 2687–2704.

Ekambaram, Vijay, Kushagra Manglik, Sumanta Mukherjee, Surya Shravan Kumar Sajja, Satyam Dwivedi, and Vikas Raykar (2020). "Attention based multi-modal new product sales time-series forecasting". In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3110–3118.

Fildes, Robert, Stephan Kolassa, and Shaohui Ma (2022). "Post-script—Retail forecasting: Research and practice". In: *International Journal of Forecasting* 38.4, pp. 1319–1324.

Fu, Stephanie, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola (2023). "Dreamsim: Learning new dimensions of human visual similarity using synthetic data". In: *arXiv preprint arXiv:2306.09344*.

Ge, Yuying, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo (2019). "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5337–5345.

Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel (2018). "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness". In: *International conference on learning representations*.

Giri, Chandadevi and Yan Chen (2022). "Deep learning for demand forecasting in the fashion and apparel retail industry". In: *Forecasting* 4.2, pp. 565–581.

Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux (2022). "Why do tree-based models still outperform deep learning on typical tabular data?" In: *Advances in neural information processing systems* 35, pp. 507–520.

H&M and Kaggle (2022). *H&M Personalized Fashion Recommendations Dataset*. `https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/data`. Accessed: 2025-02-07.

Hastie, Trevor (2020). "Ridge regularization: An essential concept in data science". In: *Technometrics* 62.4, pp. 426–433.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hebart, Martin N, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker (2023). "THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior". In: *Elife* 12, e82580.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5, pp. 359–366.

Hu, Edward J, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. (2022). "Lora: Low-rank adaptation of large language models." In: *ICLR* 1.2, p. 3.

Jolliffe, Ian T and Jorge Cadima (2016). "Principal component analysis: a review and recent developments". In: *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* 374.2065, p. 20150202.

Koehn, Philipp (2004). "Statistical significance tests for machine translation evaluation". In: *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 388–395.

Li, Zhikun, Ya Zhou, Minyi Zhao, Dabo Guan, and Zhifeng Yang (2024). "The carbon footprint of fast fashion consumption and mitigation strategies-a case study of jeans". In: *Science of The Total Environment* 924, p. 171508.

Liu, Ziwei, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang (2016). "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1096–1104.

Maurício, José, Inês Domingues, and Jorge Bernardino (2023). “Comparing vision transformers and convolutional neural networks for image classification: A literature review”. In: *Applied Sciences* 13.9, p. 5521.

McAuley, Julian, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel (2015). “Image-based recommendations on styles and substitutes”. In: *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pp. 43–52.

Muttenthaler, Lukas, Jonas Dippel, Lorenz Linhardt, Robert A Vandermeulen, and Simon Kornblith (2022). “Human alignment of neural network representations”. In: *arXiv preprint arXiv:2211.01201*.

Natekin, Alexey and Alois Knoll (2013). “Gradient boosting machines, a tutorial”. In: *Frontiers in neurorobotics* 7, p. 21.

Nenni, Maria Elena, Luca Giustiniano, and Luca Pirolo (2013). “Demand forecasting in the fashion industry: a review”. In: *International Journal of Engineering Business Management* 5, p. 37.

Oquab, Maxime, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. (2023). “Dinov2: Learning robust visual features without supervision”. In: *arXiv preprint arXiv:2304.07193*.

Peterson, Joshua C, Joshua T Abbott, and Thomas L Griffiths (2018). “Evaluating (and improving) the correspondence between deep neural networks and human representations”. In: *Cognitive science* 42.8, pp. 2648–2669.

Poonam, Poonam, Pere-Pau Vázquez, and Timo Ropinski (2025). “Evaluating graphical perception capabilities of Vision Transformers”. In: *Computers & Graphics*, p. 104458.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. (2021). “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PmLR, pp. 8748–8763.

Raghu, Maithra, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy (2021). “Do vision transformers see like convolutional neural networks?” In: *Advances in neural information processing systems* 34, pp. 12116–12128.

Rawat, Pranjal (2024). “A Deep Learning Approach to Heterogeneous Consumer Aesthetics in Retail Fashion”. In: *arXiv preprint arXiv:2405.10498*.

Siméoni, Oriane, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. (2025). “Dinov3”. In: *arXiv preprint arXiv:2508.10104*.

Skenderi, Geri, Christian Joppi, Matteo Denitto, and Marco Cristani (2024). “Well googled is half done: Multimodal forecasting of new fashion product sales with image-based google trends”. In: *Journal of Forecasting* 43.6, pp. 1982–1997.

Skenderi, Geri, Christian Joppi, Matteo Denitto, Berniero Scarpa, and Marco Cristani (2022). "The multi-modal universe of fast-fashion: the Visuelle 2.0 benchmark". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2241–2246.

Sundaram, Shobhita, Stephanie Fu, Lukas Muttenthaler, Netanel Tamir, Lucy Chai, Simon Kornblith, Trevor Darrell, and Phillip Isola (2024). "When does perceptual alignment benefit vision representations?" In: *Advances in Neural Information Processing Systems* 37, pp. 55314–55341.

Wu, Chia-Chien, Farahnaz Ahmed Wick, and Marc Pomplun (2014). "Guidance of visual attention by semantic information in real-world scenes". In: *Frontiers in psychology* 5, p. 54.

Zhang, Richard, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang (2018). "The unreasonable effectiveness of deep features as a perceptual metric". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595.

Zheng, Shuai, Fan Yang, M Hadi Kiapour, and Robinson Piramuthu (2018). "Modanet: A large-scale street fashion dataset with polygon annotations". In: *Proceedings of the 26th ACM international conference on Multimedia*, pp. 1670–1678.

A Appendix

The Appendix is divided into four subchapters, each providing additional information about the experimental procedures used in this thesis. The first subchapter gives an overview of the parameters used for fine-tuning the deep vision models. The second subchapter describes the variable selection method applied to the fashion sales prediction data. The third subchapter provides further details on the hyperparameter tuning of the prediction models. The final subchapter contains supplementary tables reporting additional experimental results that are not essential for the main findings of this thesis.

A.1 Deep Vision Model Fine-Tuning

The fine-tuning parameters used in this thesis vary depending on the deep vision model architecture (CNN or ViT), the fine-tuning method applied (MLP head or LoRA), and the dataset used for fine-tuning. Different parameter settings are employed when models are fine-tuned on large-scale datasets such as NIGHTS or on a combination of NIGHTS and FT, compared to fine-tuning on the small FT dataset.

This thesis evaluates five different ViT model variants. All models are fine-tuned using consistent parameter settings within each fine-tuning configuration, with variations only arising from the fine-tuning method and dataset size.

The following tables summarize the fine-tuning parameters used for each model configuration.

Table A.1: Fine-tuning parameters for CNN-based deep vision models using an MLP head on large-scale perceptual alignment datasets.

Parameter	Value
Hidden size (MLP)	512
Learning rate	0.0003
Weight decay	0.0001
Batch size	64
Number of epochs	20
Triplet loss margin	0.05
Early stopping patience	3
Early stopping min. delta	0.0

Table A.2: Fine-tuning parameters for CNN-based deep vision models using an MLP head on small-scale perceptual alignment datasets.

Parameter	Value
Hidden size (MLP)	512
Learning rate	0.0001
Weight decay	0.0001
Batch size	32
Number of epochs	40
Triplet loss margin	0.05
Early stopping patience	5
Early stopping min. delta	0.0

Table A.3: Fine-tuning parameters for ViT-based deep vision models using an MLP head on large-scale perceptual alignment datasets.

Parameter	Value
Hidden size (MLP)	512
Learning rate	0.0003
Weight decay	0.0
Batch size	32
Number of epochs	20
Triplet loss margin	0.05
Early stopping patience	3
Early stopping min. delta	0.0

Table A.4: Fine-tuning parameters for ViT-based deep vision models using an MLP head on small-scale perceptual alignment datasets.

Parameter	Value
Hidden size (MLP)	512
Learning rate	0.0001
Weight decay	0.0001
Batch size	128
Number of epochs	40
Triplet loss margin	0.05
Early stopping patience	5
Early stopping min. delta	0.0

Table A.5: Fine-tuning parameters for ViT-based deep vision models using LoRA on large-scale perceptual alignment datasets.

Parameter	Value
Learning rate	0.0003
Weight decay	0.0
Batch size	32
Number of epochs	8
Triplet loss margin	0.05
LoRA rank	16
LoRA scaling factor (α)	32
LoRA dropout rate	0.2
Early stopping patience	3
Early stopping min. delta	0.0

Table A.6: Fine-tuning parameters for ViT-based deep vision models using LoRA on small-scale perceptual alignment datasets.

Parameter	Value
Learning rate	0.0001
Weight decay	0.0001
Batch size	16
Number of epochs	25
Triplet loss margin	0.05
LoRA rank	8
LoRA scaling factor (α)	8
LoRA dropout rate	0.1
Early stopping patience	5
Early stopping min. delta	0.0

These parameter settings ensure comparability across models while accounting for differences in model architecture, dataset size, and fine-tuning strategy.

A.2 Variable Selection

Variable selection for the Visuelle dataset was conducted using a random forest regression model with permutation feature importance, computed over 5,000 random permutations. Permutation importance measures the increase in prediction error when a feature is randomly permuted, providing an intuitive and model-agnostic estimate of feature relevance (Breiman 2001, pp. 13–14).

Figure A.1 shows the relative importance of the different variable groups. The number of stores in which an item is sold exhibits the highest importance for predicting sales volumes. Despite this strong effect, this variable is not used in the prediction models. Instead, store-identifier

dummy variables are included, as they implicitly contain the same information while also capturing store-specific effects, such as differences in customer demand across stores.

Most remaining variables do not contribute substantially to predictive performance. Consequently, only store dummy variables and sales-week dummy variables are retained for fashion sales prediction.

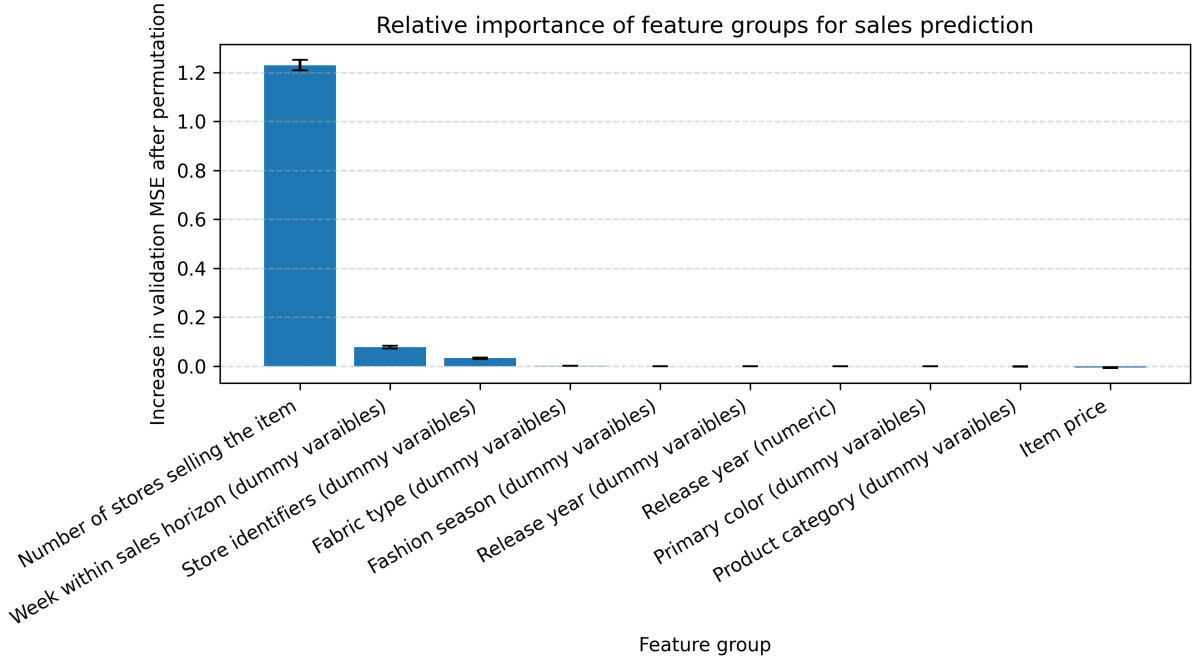

Figure A.1: Group permutation feature importance for fashion sales prediction based on the Visuelle 2.0 dataset.

A.3 Prediction Model Training

The following tables provide an overview of how hyperparameter tuning for the prediction models was conducted. They summarize the hyperparameters that were evaluated and the corresponding search ranges. For the kNN regression model and ridge regression model, a grid search was employed due to the limited number of hyperparameters.

For the random forest regression model and the gradient boosting regression model, a tree-structured Parzen estimator (TPE) search was used, as these models have a larger hyperparameter space.

Hyperparameter configurations were selected based on performance on the validation set, which contains 10% of the total dataset, using mean squared error as the optimization criterion.

Table A.7: Hyperparameter search space for the k-Nearest Neighbors (kNN) regression model.

Hyperparameter	Search Space
Number of trials	400
Maximum number of neighbors (k)	$\{1, \ldots, 200\}$
Neighbor weighting scheme	{uniform, distance}

Table A.8: Hyperparameter search space for the ridge regression prediction model.

Hyperparameter	Search Space
Number of trials	9
Regularization strength (α)	$\{10^{-4}, \ldots, 10^{4}\}$

Table A.9: Hyperparameter search space for the random forest regression model.

Hyperparameter	Search Space
Number of trials	40
Number of trees ($n_{\text{estimators}}$)	$\{300, 500, 700, 900\}$
Maximum tree depth	$\{7, 9, 11, 13\}$
Minimum samples to split a node	$\{2, 5, 10\}$
Minimum samples per leaf node	$\{1, 3, 5\}$

Table A.10: Hyperparameter search space for the gradient boosting regression model.

Hyperparameter	Search Space
Number of trials	40
Number of trees ($n_{\text{estimators}}$)	$\{300, 500, 700, 900\}$
Learning rate	$\{0.05, 0.07, 0.09, 0.11\}$
Maximum tree depth	$\{5, 7, 9, 11\}$
Minimum samples to split a node	$\{2, 5, 10\}$
Minimum samples per leaf node	$\{1, 3, 5\}$

A.4 Supplementary Results

This section contains supplementary tables reporting weighted absolute percentage error (WAPE) results for selected experiments that are not discussed in detail in the main text. Specifically, WAPE results are reported for the CLIP-B/16 and DINOv1-ViT-B/16 vision models, each evaluated with both MLP-based and LoRA-based fine-tuning.

In addition, supplementary WAPE tables are provided for comparisons between non-aligned DINO models and perceptually aligned DINO models, as well as for comparisons between

DINO models fine-tuned with LoRA on the NIGHTS dataset. These tables complement the main analysis by assessing whether the observed performance patterns are robust with respect to the choice of evaluation metric.

Across all supplementary experiments, WAPE results closely mirror the trends observed for mean absolute error (MAE) in the main Results section. For consistency and ease of comparison, all supplementary tables follow the same ordering and structure as the corresponding MAE tables presented in the main text.

Table A.11: Global weighted absolute percentage error (WAPE) across prediction models using embeddings from **CLIP-B/16** with an **MLP** head.

Model	kNN	Ridge	RF	GB
CLIP-B16-Non-Aligned	53.741	46.010	40.982	37.280
CLIP-B16MLP-NIGHTS	54.978	**45.883***	40.483***	36.748
CLIP-B16MLP-FT	**53.571**	45.983	**39.698*****	**36.684**
CLIP-B16MLP-N->FT190	55.529*	45.905	40.176***	37.029

Note: Statistical significance was assessed using a paired bootstrap test (5 000 resamples). Symbols *, **, *** indicate significant differences (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$) relative to the vanilla model; positive and negative changes denote improvements or deteriorations, respectively.

Table A.12: Global weighted absolute percentage error (WAPE) across prediction models using embeddings from **CLIP-B/16** fine-tuned with **LoRA**.

Model	kNN	Ridge	RF	GB
CLIP-B16-Non-Aligned	53.741	46.010	40.982	37.280
CLIP-B16LoRA-NIGHTS	54.851	45.801	**39.789*****	**36.543****
CLIP-B16LoRA-FT	57.022***	46.174	40.244***	37.134
CLIP-B16LoRA-N->FT190	**52.683**	**45.730**	39.889***	36.999

Note: Statistical significance was assessed using a paired bootstrap test (5 000 resamples). Symbols *, **, *** indicate significant differences (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$) relative to the vanilla model; positive and negative changes denote improvements or deteriorations, respectively.

Table A.13: Global weighted absolute percentage error (WAPE) across prediction models using embeddings from **DINOv1-ViT-B/16** with an **MLP** head.

Model	kNN	Ridge	RF	GB
DINOv1-vitb16-Non-Aligned	54.084	46.452	40.807	37.682
DINOv1-vitb16MLP-NIGHTS	**52.476****	46.312*	40.329***	36.549***
DINOv1-vitb16MLP-FT	53.039**	46.334*	39.937***	**36.540*****
DINOv1-vitb16MLP-N->FT190	53.326	**46.280***	**39.781*****	36.589**

Note: Statistical significance was assessed using a paired bootstrap test (5 000 resamples). Symbols *, **, *** indicate significant differences (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$) relative to the vanilla model; positive and negative changes denote improvements or deteriorations, respectively.

Table A.14: Global weighted absolute percentage error (WAPE) across prediction models using embeddings from **DINOv1-ViT-B/16** fine-tuned with **LoRA**.

Model	kNN	Ridge	RF	GB
DINOv1-vitb16-Non-Aligned	54.084	46.452	40.807	37.682
DINOv1-vitb16LoRA-NIGHTS	54.086	**45.989***	**39.802*****	**36.354*****
DINOv1-vitb16LoRA-FT	**53.670**	46.344	40.008***	36.837*
DINOv1-vitb16LoRA-N->FT190	54.124	46.075*	40.398***	36.720**

Note: Statistical significance was assessed using a paired bootstrap test (5 000 resamples). Symbols *, **, *** indicate significant differences (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$) relative to the vanilla model; positive and negative changes denote improvements or deteriorations, respectively.

Table A.15: Global weighted absolute percentage error (WAPE) across prediction models using embeddings from non-aligned DINO vision models.

Model	kNN	Ridge	RF	GB
DINOv1-vitb16-Non-Aligned	54.084	46.452	40.807	37.682
DINOv2-vitb14-Non-Aligned	53.755	**46.205**	40.123	39.126
DINOv3-vitb16-Non-Aligned	**53.612**	46.590	**40.105**	**37.157**

Note: The reported values correspond to prediction errors obtained using non-aligned deep vision models. No statistical significance testing is performed for these comparisons.

Table A.16: Global weighted absolute percentage error (WAPE) across prediction models using embeddings from the **DINO** vision model fine-tuned with **LoRA** on the **NIGHTS** dataset.

Model	**kNN**	**Ridge**	**RF**	**GB**
DINOv1-vitb16LoRA-NIGHTS	54.086	45.989*	39.802***	**36.354****
DINOv2-vitb14LoRA-NIGHTS	53.293	**45.735***	**39.754****	36.850***
DINOv3-vitb16LoRA-NIGHTS	**53.158**	46.160*	40.285	36.693

Note: Statistical significance was assessed using a paired bootstrap test (5 000 resamples). Symbols *, **, *** indicate significant differences (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$) relative to the vanilla model; positive and negative changes denote improvements or deteriorations, respectively.

Eidesstattliche Erklärung

Ich versichere an Eides Statt, dass ich die Arbeit selbständig verfasst, keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe, alle wörtlich oder sinnesgemäß aus anderen Werken übernommenen Aussagen als solche gekennzeichnet habe und dass die Arbeit weder vollständig noch wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen ist und dass ich die Arbeit weder vollständig noch in wesentlichen Teilen bereits veröffentlicht habe sowie dass das in Dateiform eingereichte Exemplar mit den eingereichten gebundenen Exemplaren übereinstimmt.

Ort, Datum: Thuine, 29.12.25

Unterschrift: N. Graßhoff