A Regression Deep Learning Approach for Fashion Compatibility

Luís Silva[1], Ivan Gomes[1], C.Mendes Araújo[2], Tiago Cepeda[1], Francisco Oliveira [1] and João Oliveira [1]
[1] *Department of Digital Transition, CITEVE - Centro Tecnológico das Indústrias Têxtil e do Vestuário de Portugal, V.N.Famalicão, Portugal*
[2]*CMAT – Centro de Matemática and Departamento de Matemática, Universidade do Minho, Braga, Portugal*
{*lfsilva, fdoliveira, igomes*}*@citeve.pt, clmendes@math.uminho.pt*

Keywords: Visual Search, Deep learning, outfit, BiLSTM, CNN, compatibility learning, similarity learning, transformer.

Abstract: In the ever-evolving world of fashion, building the perfect outfit can be a challenge. We propose a fashion recommendation system, which we call Visual Search, that uses computer vision and deep learning to ensure that it has a co-ordinated set of fashion recommendations. It looks at photos of incomplete outfits, recognizes existing items, and suggests the most compatible missing piece. At the heart of our system lies a compatibility model made of a Convolutional Neural Network and bidirectional Long Short Term Memory to generate a complementary missing piece. To complete the recommendation process, we incorporated a similarity model, based on Vision Transformer. This model meticulously compares the generated image to the catalog items, selecting the one that most closely matches the generated image in terms of visual features.

1 INTRODUCTION

The phenomenal rise of e-commerce has revolutionized traditional distribution channels, unleashing a wave of innovative business models that have fundamentally reshaped the retail landscape (Xiao, Wu, Xie, & Hu, 2019). However, this transformative digital revolution has not yet permeated the entire spectrum of businesses, particularly within the traditional sector, where many enterprises continue to lag behind in terms of online visibility and engagement with the digital marketplace. The COVID-19 pandemic, a far more recent phenomenon, has had a notably profound impact. The outbreak led to a precipitous decline of sales in traditional retail stores, while worldwide online sales of clothing and textiles surged to unprecedented heights (Çiçek & Muzaffar, 2021), compelling several companies to re-evaluate their direct online sales strategies.

The digital realm is poised to become the primary driver of growth for the fashion industry, presenting a wealth of opportunities for textile companies to thrive.

This strategic move would enable the sector to complement its traditional distribution channels with a direct-to-consumer approach, fostering stronger customer connections and brand loyalty. Simultaneously, the Business-to-Business segment should not be overlooked, as it represents a significant source of revenue and growth potential. Regardless of whether the focus is on Business-to-Consumer or Business-to-Business, many of the underlying challenges remain the same, particularly in creating a seamless and personalized digital experience for both consumer and business customers. In fashion analysis, visual compatibility refers to the extent to which clothing items complement visually across different categories. For instance, the compatibility between a “suit” and “oxfords” is typically higher than with “trainers”.

Visual Search emerges as a groundbreaking pipeline of deep learning models for fashion recommendation, ushering in a new era of innovation that transforms the way individuals curate their sartorial identities.

It comprises of four key modules. The first handles image prepossessing, ensuring necessary manipulations, the second, utilizing a compatibility model, assesses features against a fashion dataset, pinpointing the ideal clothing item. A third module finds real-world analogs, and the fourth validates predictions. Users can upload multiple photos of their current outfit and the service will swiftly offer suggestions of the store’s catalog for the missing clothing item to ensure the overall outfit coordination, or they can simply upload an image of the desired piece and find the most similar item in the store’s catalog. Thus enhancing the overall user experience by facilitating the discovery of items that closely align with their styling preferences.

Our main purpose with Visual Search is to optimize the shopping experience by functioning as a recommendation system through sophisticated image processing, compatibility and similarity analysis.

Recognizing each user's distinct style fingerprint, Visual Search tailors recommendations to the current wardrobe. This personalized approach ensures that users receive relevant and useful suggestions, enhancing their overall shopping experience. In the realm of online shopping, Visual Search eliminates guesswork and frustration. By providing tailored recommendations, the system simplifies the search for the perfect outfit, reducing the time spent sifting through countless options. This heightened convenience fosters user satisfaction and encourages repeat purchases.

The paper follows a structured approach consisting of six sections. The second section is focused in the exploration of related works, elucidating existing research that has served as the inspiration for this study. Section three is all about the methodology, describes the datasets used, and details the architecture and complexity of the two main models, the compatibility model and the similarity model. This section expounds upon their respective architectures, the mathematical formulas behind, while also explaining the method employed for performance evaluation. The fourth section showcases the results obtained from the implementation of each model, offering insights and the implications derived from the findings. This section also highlights the enhancements achieved through the proposed methodologies. In the fifth section, Conclusion, the accomplishments of the study are reviewed alongside expectations for future research. Lastly, the sixth section serves as an acknowledgment of the contributions and support received.

2 RELATED WORK

We explore various strands of research closely associated with our methodology.

Recommendation Systems in Fashion. Various methodologies have been proposed for suggesting fashion items (Hwangbo, Kim, & Cha, 2018; Yethindra & Deepak, 2021; Bellini, Palesi, Nesi, & Pantaleo, 2023). (Hwangbo et al., 2018) introduced a recommendation system that compiles data from online shopping mall databases, gather purchase history (offline) and click history (online) to feed a so called "K-RecSys" model that takes in consideration this parameters. (Yethindra & Deepak, 2021) provide a personalized clothing recommendations for men using logistic regression classification and semantic similarity computation through fashion ontology. (Bellini et al., 2023) introduced a recommendation system tailored for fashion retail shops. It employs a multi-clustering approach, considering items and users' profiles across both online and physical stores. By leveraging mining techniques, the system predicts the purchase behavior of newly acquired customers.

Visual Compatibility Extraction. In this field (Yin, Li, Lu, & Zhang, 2019) proposed a fashion compatibility knowledge learning method that incorporates visual compatibility relationships as well as style information using a Convolutional Neural Network (CNN), the 'convolutional' part references the use of convolutional layers, which apply convolution operations to detect and extract features from input data. (Han, Wu, Jiang, & Davis, 2017) employed a Bidirectional Long Short Term Memory (BiLSTM) to capture the compatibility relationships of fashion items by considering an outfit as a sequence from top to bottom and then accessories and images in the collection as individual time steps, "named Long Short-Term Memory (LSTM)" due to its ability to capture and retain long-term dependencies in data while handling short-term information through memory cells. (Revanur, Kumar, & Sharma, 2021) used a semi-supervised learning approach where was leveraged large unlabeled fashion corpus to create pseudo positive and negative outfits on the fly during training. For each labeled outfit in a training batch, a pseudo-outfit is obtained by matching each item in the labeled outfit with unlabeled items. More recently, (Jing, Cui, Guan, Nie, & Su, 2023) delved into a fashion compatibility modeling approach with a category-aware multimodal attention network, termed as FCM-CMAN. In this paper, the focus is on the visual compatibility of entire outfits, where items in a fashion collection are expected to exhibit similar styles, forming a cohesive and stylish composition. To achieve this, a BiLSTM model is employed to discern compatibility relationships within outfits, capturing the dependencies among various fashion items. This approach goes beyond by using a visual-semantic embedding. This capability enhances individual item recommendations by understanding and revealing their relationships within the given context.

Similarity in Fashion. The focus is on identifying items that have similarity to those shown. This entails uncovering apparel pieces that share common visual attributes or style elements, providing cus-

tomers with recommendations that align closely with their preferences (Dong et al., 2021; Gao et al., 2020; Manandhar, Bastan, & Yap, 2018). (Dong et al., 2021). introduce an Attribute-Specific Embedding Network to predict fine-grained fashion similarity by jointly learning multiple attribute-specific embeddings. (Gao et al., 2020) propose a novel graph reasoning network (GRNet) on a similarity pyramid, which learns similarities between a query and a gallery cloth by using both initial pairwise multi-scale feature representations and matching propagation for unaligned representations. (Manandhar et al., 2018) introduced a new attribute-guided metric learning (AGML) with multitask CNN that jointly learns fashion attributes and image embeddings while taking category and brand information into account.

.

3 APROACH/METHODS

In our work, we use this concept of similarity to enhance our compatibility learning. After generating a fashion item that is compatible within a given context, we use similarity metrics to identify real-world items that closely resemble our generated prediction.

A CNN approach was employed to capture and extract features from images, specifically focusing on generating a unseen image from a given contextual representation, in this case, an outfit represented as a sequence. Subsequently, we used a BiLSTM for fashion compatibility modeling, which processes the sequential nature of the outfit to generate the final image. Following the generation process, a pre-trained model was employed to identify the most similar real image to our generated one, completing the comprehensive workflow of our approach.

3.1 Dataset

The Cleaned Maryland dataset was developed by the Fashion Team at the Laboratory for Artificial and it is a clean version of the Maryland Polymores dataset. It was also used in (Han et al., 2017; Zou, Pang, Zhang, & Wong, 2022). Polyvore, a widely used fashion website, serves as a platform where users share outfits, providing information including images, descriptions, likes, hashtags, and more. The Cleaned Maryland dataset comprises 21,889 outfits carefully extracted from Polyvore, having the fashion items been re-organized into 20 categories. For the purpose of our specific study, however, we adopted a more focused approach. Instead of utilizing all 20 categories, we narrowed down our selection to a more streamlined set of four key categories: ‘top‘, ‘bottom‘, ‘other‘, and ‘feet‘. This reduction was a deliberate choice aimed at simplifying the dataset and transforming it into a sequence of items, aligning with the learning capabilities of our BiLSTM model. As a result of this, the outfits in our dataset have been condensed to a more manageable number of 1356. One notable feature of the dataset is the exclusion of background information. By removing unnecessary background details, the dataset minimizes extraneous noise. An example of an outfit in this categories is shown in **Figure1**.

Figure 1: **Compatible Outfit from Maryland dataset**

In addition to a evaluation on the datasets, testing is extended to include another dataset named ‘Community Pictures‘. This dataset comprises 5000 images of 20 categories and it was collaboratively constructed by the community, with members contributing images of their clothing items.

3.2 Fashion outfit compatibility model

Our compatibility model initiates with a CNN as a feature extractor. The primary goal of the CNN is to capture visual features from input images, with a specific focus on identifying and extracting key features crucial for understanding the outfit sequence, determining their importance for future feeding into the BiLSTM. The CNN model is extended to operate on each image in the sequence independently. The feature maps acquired from individual images are subsequently either combined or further processed to capture temporal dependencies.

Let I_t represent the t-th image in the input sequence, and $F(I_t)$be the feature map obtained after passing $I_t$through the CNN. The convolutional operation is as follows:

$$F(I_t)_{i,j,k} = \sigma\left(\sum_m \sum_n \sum_p W_{m,n,p,k} \cdot I_{t,i+m,j+n,p} + b_k\right)$$

- $F(I_t)_{i,j,k}$ is the activation at position (i, j) in the k-th feature map for the t-th image.
  - $W_{m,n,p,k}$ is the weight of the k-th filter at position (m,n) in channel p.
  - $I_{t,i+m,j+n,p}$ denotes the pixel intensity at position $(i+m, j+n)$ in channel p of the t-th image.
  - b_k is the bias term for the k-th filter.
  - σ is the ReLU activation function.

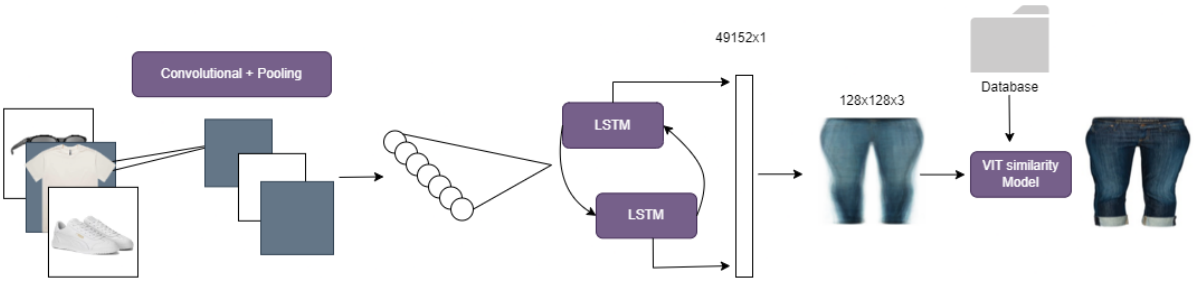

Figure 2: **Model architecture**

The output of the CNN, $F(I_t)$, serves as input to the BiLSTM model. The BiLSTM then processes the sequence of feature maps to generate predictions for the next item in the outfit sequence. The loss function $E_f(F;\Theta_f)$ is computed based on the negative log probability of observing the next item x_{t+1} given the previous items $x_1,...,x_t$. The LSTM equations involve the use of the ReLU activation function:

$$E_f(F;\Theta_f) = -\frac{1}{N}\sum_{t=1}^{N} \log Pr(x_{t+1}|x_1,...,x_t;\Theta_f)$$

where Θ_f denotes the model parameters of the forward prediction model, and $Pr(\cdot)$, computed by the LSTM model, is the probability of observing x_{t+1} conditioned on previous inputs.

More specifically, the LSTM model maps an input sequence $\{x_1,x_2,...,x_N\}$ to outputs via a sequence of hidden states by computing the following equations recursively from $t=1$ to $t=N$:

$$\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i),\\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f),\\
c_t &= f_t c_{t-1} + i_t \sigma(W_{xc}x_t + W_{hc}h_{t-1} + b_c),\\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o),\\
h_t &= o_t \sigma(c_t),
\end{aligned}$$

where: - x_t is the input at time t, - h_t is the hidden state at time t, - c_t is the cell state at time t, - i_t, f_t, o_t are the input, forget, and output gates' activations, - σ is the ReLU activation function, - W and b are the weight matrices and bias vectors for different gates in the LSTM.

- Forward LSTM:

  $$\overrightarrow{H_t} = \text{LSTM}_{\text{forward}}(X_t, \overrightarrow{H_{t-1}})$$

  where

  X_t is the input at time t,

  $\overrightarrow{H_t}$ is the hidden state at time t.

  The forward LSTM processes the input sequence from the beginning to the end, capturing dependencies in the forward direction.

- Backward LSTM:

  $$\overleftarrow{H_t} = \text{LSTM}_{\text{backward}}(X_t, \overleftarrow{H_{t+1}})$$

  where

  X_t is the input at time t,

  $\overleftarrow{H_t}$ is the hidden state at time t.

  The backward LSTM processes the input sequence from the end to the beginning, capturing dependencies in the backward direction.

- Final Hidden State:

  $$H_t = [\overrightarrow{H_t}; \overleftarrow{H_t}]$$

  The final hidden state at time t is the concatenation of the forward and backward hidden states. This combined representation captures both forward and backward context, enabling the model to understand the sequential dependencies within the input sequence.

3.3 Fashion similarity model

Our similarity approach, takes in the compatibility model output, our generated image, to identify the most visually similar image from a pre-loaded set. This process is orchestrated with a pre-trained Vision Transformer (ViT) as the backbone for similarity assessment. The vit_base_patch16_224.mae follows the ViT architecture, which represents a departure from CNN. Vision Transformers use a transformer-based architecture, originally designed for natural language processing tasks, to process image data. Notably, the model employs a patch-based approach, breaking down the input image into smaller patches and treating them as sequence for processing.

- **Parameters (M):** 85.8
- **GMACs:** 17.6
- **Activations (M):** 23.9
- **trained Images:** 224x224

The number of parameters (85.8M) indicates the model's complexity, while the GMACs (17.6) reflect its computational workload in terms of Giga Multiply-Accumulates. Additionally, the activations (23.9M) represent the total number of activations during inference, offering insights into the model's computational efficiency. Lastly, the size of trained images (224x224) underscores the scale at which the model operates.

One of features of vit_base_patch16_224.mae is its pretraining methodology. The Self-Supervised Masked Autoencoder (MAE) technique involves training the model to predict masked-out portions of the input image.

3.4 Performance evaluation

In evaluating the accuracy of our approach, we employ the Universal Image Quality Index (UIQI) as the metric used for this task. The UIQI measures the similarity and quality of images, enabling a evaluation beyond a mere binary comparison (Wang & Bovik, 2002).

The UIQI is calculated using the following formula:

$$UIQI = \frac{4 \cdot \text{covariance}(I_1, I_2) \cdot \text{mean}(I_1) \cdot \text{mean}(I_2)}{(\text{variance}(I_1) + \text{variance}(I_2)) \cdot (\text{mean}^2(I_1) + \text{mean}^2(I_2))}$$

Here: I_1 and I_2 are the intensity values of the two images being compared.

Approach and Threshold

To evaluate the accuracy of our approach, we consider the most similar image identified by our similarity model. We calculate the UIQI of this identified image with the test image belonging to the outfit in the first place. Instead of relying on a simplistic comparison of identical images, we set a threshold for the UIQI. If the UIQI value exceeds this threshold, we classify the prediction as accurate.

Why UIQI over Image Similarity

The choice of UIQI over a straightforward image similarity check is motivated by the nature of fashion. Similar images may exhibit subtle differences, such as distinct patterns or textures, yet still be suitable for the same outfit. For instance, two white t-shirts might differ in design but share compatibility within the context of an outfit. The UIQI accounts for it, providing a more refined assessment of image quality and similarity (Wang & Bovik, 2002). This approach allows us to capture the essence of fashion compatibility beyond strict visual identity.

4 RESULTS

In this section, we discuss the benchmark and conclusions aimed at enhancing the accuracy of our approach." We systematically compare results across various variables, offering insights into the effectiveness of our method.

4.1 Compatibility

Table 1: Comparison between different datasets

CNN + BiLSTM	Maryland	CP
top	**91.18%**	42.64%
bottom	**84.24%**	49.49%
other	79.56%	**82.64%**
feet	77.06%	**83.74%**

Based on the evaluation results, it is evident that the model's performance varies significantly across the two datasets, Maryland and Community Pictures, and their respective categories.

In the Maryland dataset, where outfits are made with purpose and exhibit a logical and consistent composition, the model consistently demonstrates accuracy ranging from 75% to 90% across all four categories (‘top‘, ‘bottom‘, ‘other‘, and ‘feet‘). This consistent performance suggests that the model effectively generalizes to the structured composition of the Maryland dataset. The variation in accuracy from ‘top‘ and ‘bottom‘ to ‘feet‘ and ‘other‘ could be from the fact that feet and other have more variability in their shapes and colors.

On the other hand, in the Community Pictures dataset, which involves randomly assembled outfits without a predetermined logic, due to the fact that the dataset was just made out of random images from the community, the model exhibits notable differences in accuracy among the categories. Remarkably high accuracy is observed for ‘other‘ and ‘feet‘, where interpolation was applied due to a reduced number of images in these categories. The interpolation led to the model encountering the same images multiple times, enabling it to recognize and classify these categories effectively, resulting in high accuracy. However, the categories ‘top‘ and ‘bottom‘ show considerably lower accuracy. This can be attributed to the inherent randomness and lack of consistency in the outfit compositions within the Community Pictures dataset. As outfits were generated by mixing clothes without a structured approach, the model struggled to find meaningful patterns in these categories, leading to low accuracy.

We present the results obtained from our exper-

imental evaluation of different compositional models applied to the task at hand. Table 2 showcases the performance metrics in terms of accuracy for each model configuration.

VGG is a CNN architecture introduced by the Visual Geometry Group at the University of Oxford. The key characteristic of VGG is its simplicity and uniform architecture. The network consists of multiple layers with small receptive fields Proposed by (He, Zhang, Ren, & Sun, 2016), ResNet introduces a concept of residual learning, where shortcut connections allow the network to learn the residual functions, making it easier to train extremely deep networks.

Table 2: Models Benchmark Accuracy Results

Comp Model	CNN	ResNet	Resnet+CNN	VGG	VGG+CNN
LSTM	81,33%	80,29%	78,56%	80,29%	78,45%
BiLSTM	**91,18%**	85,65%	80,92%	84,56%	79,37%
No LSTM	65,29%	64,56%	62,11%	62,35%	61,98%

The LSTM model performs significantly better than the No LSTM counterpart across all the compared models. This suggests that the inclusion of LSTM layers in the model architecture contributes positively to the overall accuracy. Models without LSTM exhibit lower accuracy compared to their LSTM counterparts. This indicates that the temporal dependencies captured by LSTM layers are beneficial for this task. BiLSTM consistently outperforms LSTM, achieving the highest accuracy among all the models. This indicates that bidirectional temporal context is crucial for the task, as BiLSTM considers information from both past and future time steps (Han et al., 2017). The use of pre-trained models (ResNet and VGG) did not lead to an improvement in accuracy compared to the standalone CNN model. This unexpected result suggests that, in this particular task, the transfer learning process may not have effectively leveraged the pre-learned features from these architectures. It could be related to the domain of the pre-trained models, the specifics of the transfer learning process, or the characteristics of the dataset.

4.2 Similarity

To determine the best-performing model among a batch of 600 backbones, an evaluation process was undertaken. The evaluation aimed to assess each model's ability to identify the five most similar images from a large batch of diverse data. For each evaluation instance, a single image was provided as input to the model, which was then tasked with retrieving the five most similar images from the given dataset. This process was repeated for multiple images, to quantify the performance of each model, the Universal Quality Index (UIQI) was employed to compute the mean similarity index for the top five retrieved images across all input images. The decision to evaluate the top five images comes from the nature of recommending items for outfits. In fashion, there often exist multiple suitable clothing options that can complement a particular look. By considering the top five recommendations, the evaluation process acknowledges the variability and subjective nature of those recommendations. Based on the evaluation results of the similarity models, the performance varies across different resolutions and models. The table below presents the UIQI mean values for the top four similarity models at various image resolutions.

Figure 3: **Generated Image as input of Similarity model**

In Figure 3 its a representation of the generated image being compared with real world images, the similarity model compares it against a database of real images, evaluating features, textures, colors, and overall composition to determine the closest match, in this example, the model determined that the most similar image had a UIQI of 0.71 in comparison to ours. Since this value exceeds our threshold, the image is labeled to be well predicted.

Table 3: Similarity Accuracy

Sim Models	vit	tresnet	convnext	efficientnet
128x128	0.8729	0.8646	0.8634	0.8603
256x256	0.8744	0.8678	0.8659	0.8625
364x364	0.8749	0.8691	0.8662	0.8634
512x512	**0.8749**	0.8693	0.8674	0.8640

While higher resolutions generally lead to better UIQI mean values (Wang & Bovik, 2002), it's essential to note that the improvements in model accuracy may not always justify the increased complexity and time consumption associated with handling those highest resolutions. Notably, the observations indicate that the 512x512 resolution tends to exhibit the highest UIQI mean, underscoring the influence of resolution on model performance, however the improvement is not enough to justify the use of higher resolutions since it will slow down the pipeline too much.

5 CONCLUSION

This paper uses a approach to fashion compatibility learning by simultaneously training a CNN and a Bi-

LSTM model. the method treats an outfit as a sequence, with each item serving as a time step. What's new about this approach is the fact that this model task is to generate an image via pixel predicting values, a regression. Additionally, a similarity model is used on top of this generated image to recommend a real-world image. To validate the accuracy of this approach, the Universal Image Quality Index is employed on the recommended image and the actual image. This metric serves as a measure of how closely our recommendation aligns with real outfits.

The outcomes showcase the effectiveness of this approach in learning the compatibility of fashion outfits. Recognizing that fashion compatibility is subjective, varying from one individual to another, our future research will explore modeling user-specific compatibility and style preferences, our goal is to construct a more personalized system that caters to individual tastes and preferences, thereby enhancing the overall user experience.

Expanding our approach to include keywords like style, mood, artist, material, texture, and brand adds depth to our model's decision-making process improving our latent space of features extracted. This approach holds immense promise in the realm of generative AI, particularly in the domain of image generation. Our aspirations go beyond outfits because compatibility isn't exclusive to fashion. Just as there are compatible pieces in clothing, there are combinations in other industries too. By exploring diverse fashion domains, we're talking of compatibility that exists beyond clothing. Moreover, our curiosity extends beyond fashion, by the potential for compatibility in other sectors, recognizing that the principles guiding our recommendation system can be applied across various industries. Whether it's in home goods, technology, or beyond.

ACKNOWLEDGEMENTS

TexP@CT Mobilizing Pact - Innovation Pact for the Digitalization of Textiles and Clothing, project no. 61, to Reinforce the Competitiveness and Resilience of the National Economy, financed through Component 5 - Capitalization and Business Innovation, of the European funds allocated to Portugal by the Recovery and Resilience Plan (PRR), under the European Union's (EU) Recovery and Resilience Mechanism, as part of Next Generation EU (https://recuperarportugal.gov.pt/), for the period 2021 - 2026. CMAT: partially supported by FCT-'Fundação para a Ciência e a Tecnologia', within projects UIDP/00013/2020 and UIDB/00013/2020 (DOI 10.54499/UIDP/00013/2020 and DOI 10.54499/UIDB/00013/2020)

REFERENCES

Bellini, P., Palesi, L. A. I., Nesi, P., & Pantaleo, G. (2023). Multi clustering recommendation system for fashion retail. *Multimedia Tools and Applications*, *82*(7), 9989–10016.

Dong, J., Ma, Z., Mao, X., Yang, X., He, Y., Hong, R., & Ji, S. (2021). Fine-grained fashion similarity prediction by attribute-specific embedding learning. *IEEE Transactions on Image Processing*, *30*, 8410–8425.

Gao, Y., Kuang, Z., Li, G., Luo, P., Chen, Y., Lin, L., & Zhang, W. (2020). Fashion retrieval via graph reasoning networks on a similarity pyramid. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Han, X., Wu, Z., Jiang, Y.-G., & Davis, L. S. (2017). Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th acm international conference on multimedia* (pp. 1078–1086). doi: https://doi.org/10.48550/arXiv.1707.05691

He, K., Zhang, X., Ren, S., & Sun, J. (2016, June). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr).*

Hwangbo, H., Kim, Y. S., & Cha, K. J. (2018). Recommendation system development for fashion retail e-commerce. *Electronic Commerce Research and Applications*, *28*, 94–101.

Jing, P., Cui, K., Guan, W., Nie, L., & Su, Y. (2023). Category-aware multimodal attention network for fashion compatibility modeling. *IEEE Transactions on Multimedia*.

Manandhar, D., Bastan, M., & Yap, K.-H. (2018, September). Tiered deep similarity search for fashion. In *Proceedings of the european conference on computer vision (eccv) workshops.*

Revanur, A., Kumar, V., & Sharma, D. (2021). Semi-supervised visual representation learning for fashion compatibility. In *Proceedings of the 15th acm conference on recommender systems* (pp. 463–472).

Wang, Z., & Bovik, A. C. (2002). A universal image quality index. *IEEE signal processing letters*, *9*(3), 81–84.

Xiao, J., Wu, Y., Xie, K., & Hu, Q. (2019). Managing the e-commerce disruption with it-based in-

novations: Insights from strategic renewal perspectives. *Information & Management*, *56*(1), 122–139.

Yethindra, D. N., & Deepak, G. (2021). A semantic approach for fashion recommendation using logistic regression and ontologies. In *2021 international conference on innovative computing, intelligent communication and smart electrical systems (icses)* (pp. 1–6).

Yin, R., Li, K., Lu, J., & Zhang, G. (2019). Enhancing fashion recommendation with visual compatibility relationship. In *The world wide web conference* (pp. 3434–3440).

Zou, X., Pang, K., Zhang, W., & Wong, W. (2022). How good is aesthetic ability of a fashion model? In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 21200–21209).

Çiçek, Y., & Muzaffar, H. (2021). The impact of covid-19 pandemic crisis on online shopping. *AYBU Business Journal*, *1*(1), 16-25.