MEDAL: multi-modal MEta-space Distillation and ALignment for Visual Compatibility Learning

[1]Dween Rabius Sanny, [2]Vinay Kumar Verma, [1]Prateek Sircar, [1]Deepak Gupta
International Machine Learning, Amazon India[1], Private Brands & Discovery[2]
{drsanny, sircarp, dgupt, vkvermaa} @amazon.com

Abstract

Visual compatibility recommendation systems aim to surface compatible items (e.g. pants, shoes) that harmonise with a user-selected product (e.g., shirt). Existing methods struggle in three key aspects: they rely on global CNN representations that overlook fine-grained local cues critical for visual pairing; they force all categories into a single latent space, ignoring the fact that compatibility rules differ across product-type pairs; and they demand costly, expert-annotated outfit labels. We introduce MEDAL(Meta-space Distillation and Alignment), a self-supervised framework that addresses all three challenges simultaneously. MEDAL (i) employs a local–global augmentation curriculum inside a teacher–student ViT to emphasise patch-level texture and pattern similarities while suppressing confounding global shape cues; (ii) partitions the joint feature manifold into learnable, pair-specific meta-spaces so that, for example, {shirt,pants} and {pants,shoes} relationships are modelled with distinct projection masks; and (iii) replaces manual labels with distantly supervised KD, harvesting pseudo-compatible sets via object detection on web images, thus scaling to millions of real-world examples. We further fuse perceptually uniform LUV colour histograms to capture global colour harmony often missed by pure vision transformers. Extensive experiments on Polyvore disjoint/non-disjoint and a 2M-image in-house dataset show state-of-the-art gains of up to +3.72/+2.7FITB and +9.58R@10 over the strongest baseline, whilst cutting annotation cost to zero. Qualitative studies confirm that MEDAL retrieves stylistically coherent outfits and correctly penalises mismatched colour palettes.

1. Introduction

In fashion recommendation systems, two fundamental tasks, outfit compatibility prediction (CP) and complementary item retrieval (CIR) underpin the ability to generate personalized, coherent ensembles. CP evaluates whether a set of items forms a visually cohesive outfit, quantifying their collective aesthetic appeal, while CIR addresses the challenge of completing partial outfits by retrieving compatible items from large-scale databases. For instance, given a combination such as a top, pants, and shoes, CIR systems recommend harmonizing accessories like handbags, thereby supporting both retailers in delivering engaging, tailored suggestions and consumers in assembling stylish, well-coordinated outfits. When trained on extensive fashion datasets, such systems effectively learn style preferences and the underlying principles of outfit coordination, leading to recommendations that are increasingly aligned with individual tastes.

Early research [8, 33, 38] focused on leveraging CNN-based feature extractors [9, 32] for fashion recommendation, but these models were ultimately limited in their capacity and generalizability to complex, diverse e-commerce settings. More recent efforts [3, 20, 22, 34] have turned to transformer encoders to model entire outfits holistically, often fusing visual and textual modalities to capture intricate relationships among garments and accessories. Nevertheless, most current visual compatibility models are built atop classification-based, pre-trained architectures that prioritize global features such as overall shape and style. This approach fails to account for the fact that, in practice, local features including texture, pattern, and colour gradients are often more decisive when determining the compatibility of items like shirts and pants. Patch-level embeddings and end-to-end learning pipelines are therefore better suited for capturing the nuanced cues that govern visual compatibility.

A further limitation of existing models [8, 18, 20] is their reliance on a single joint embedding space for all compatible items, despite the clear observation that compatibility rules can vary significantly across different product categories. For example, the visual signals that define shirt–pants compatibility differ markedly from those that guide pants–shoes matching. This "one-space-fits-all" strategy can result less generalizability, especially as the diversity of compatible items increases. Compounding these modelling challenges, datasets such as Hypatia-OutfitBuilder [10] and Polyvore [8] contain too few samples for robust generalization to real-

world scenarios, while manual annotation is expensive and requires domain expertise.

To address these limitations, we propose a framework that combines local and global data augmentation specifically tailored for visual compatibility learning. Our teacher-student architecture aligns locally and globally augmented views of compatible pairs, optimizing the embedding space to prioritize local feature matching while preserving global context. Rather than using a single joint space, our approach defines pairwise meta-spaces, allowing the model to learn specialized, compact representations for each type of compatible pair. To overcome annotation bottlenecks, we adopt distantly supervised learning by mining complete outfit images from the web. Individual clothing items are extracted using the Grounding-DINO [21] detector, forming sets considered as compatible pairs. We further refine these sets and augment existing datasets such as Hypatia-OutfitBuilder using the SkiLL [42] similarity model. For textual attributes, a pretrained attribute extraction module identifies key item descriptors that are then encoded as text features. Given the critical role of colour in fashion compatibility, we explicitly incorporate LUV-space colour embeddings to further enhance performance.

Our contributions are as follows: **1.** A teacher-student framework that leverages local and global feature matching for robust visual compatibility representations; **2.** The use of multiple style meta-spaces to model compatibility across diverse fashion categories; **3.** Explicit computation and integration of colour embeddings in LUV colour space; **4.** A distantly supervised approach to produces large-scale, real-world fashion data without requiring costly human annotation; **5.** An adaptive triplet loss informed by negative sample similarity.

2. Related Work

Outfit Compatibility Prediction has been addressed from two perspectives: pairwise item-to-item compatibility[23, 39] and overall outfit compatibility[8, 33, 38]. Pairwise methods learn a shared style space across item categories using co-purchase and co-view data [23, 35, 39], where pairwise distances are used to measure compatibility. However, these approaches only model individual item pairs and not full outfits. Later research explored outfit-level compatibility using the Polyvore dataset [8, 18, 33, 38]. Han et al.[8] treated outfits as sequences and used LSTMs with a fill-in-the-blank (FITB) task. Vasileva et al.[38] introduced multiple style spaces to capture nuanced similarity notions across categories. Cucurull et al. [4] applied graph convolutional networks, though their applicability is limited for dynamic catalogs lacking item connectivity.

Recent works [33, 38, 40] have explored learning subspace embeddings for visual compatibility. Conditional Similarity Networks (CSN)[40], used in[38], learn type-aware embeddings for item category pairs. Tan et al.[33] introduced shared subspace learning with adaptive importance weights. CSA-Net[20] further builds on this idea by learning weighted subspaces conditioned on item categories for Cross-domain Item Retrieval (CIR).

The above methods [23, 38, 39] can be applied to CIR, though they were originally developed for outfit compatibility. Generative methods like GANs [13, 31, 45] generate item representations based on a given input and retrieve complementary items through similarity with an indexed database. Recently, [44] leveraged diffusion models with user history for generative visual compatibility.

Attention mechanisms [3, 11, 20, 22, 24, 34] have gained popularity in fashion recommendation. Specifically,[20, 34] use attention to model pairwise compatibility. Transformer self-attention[15] models higher-order interactions across items, and transformer encoders [30] are used to represent entire outfits for scalable retrieval. These methods largely emphasize global features, which often fail to capture the fine-grained attributes essential for visual compatibility.

Recent work further advances the field through multimodal and transformer-based designs. VICTOR[27] applies contrastive vision–language pretraining to detect both overall and item-level mismatch. Fashion-GPT[2] integrates large language models with a modular retrieval pipeline that exploits multi-view textual embeddings. CANN[19] models compositional coherence with multi-head attention and localized focal regions for fine-grained attribute alignment. Attention-Based Fusion[14] combines patch-level visual and textual features using soft attention for outfit recommendation. MLCN[43] predicts and diagnoses compatibility by comparing features across multiple CNN layers. Finally, HAT[12] incorporates user history using a dual-transformer setup for personalized retrieval.

In our proposed model (MEDAL), we learn both local and global features using the knowledge distillation framework. Since color plays a significant role in visual compatibility, we explicitly incorporate color embeddings and pairwise meta-space alignment into the model to help learn more robust representations.

3. Proposed Method

This paper proposes a novel method for learning visual compatibility. The proposed model comprises four key components: (1) local and global feature learning, (2) distantly supervised knowledge distillation, (3) meta-space learning for each style pair, and (4) explicit color fusion. We employ a knowledge distillation and triplet loss formulation, utilizing three distinct images: an anchor, a positive, and a negative sample. The following section provides brief descriptions of the proposed model.

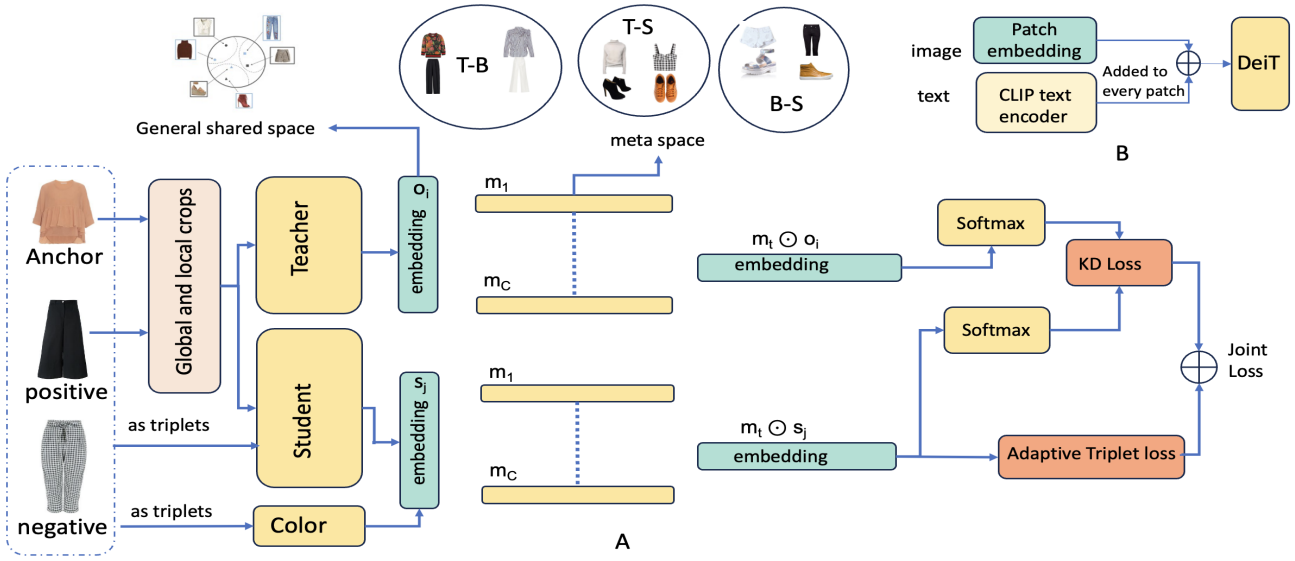

Figure 1. **A:** The block diagram of the proposed model. The Teacher model accepts the local and global images, while the Student model also accepts the color embeddings. These are further aligned with the meta-space alignment m_t. Compatible pairs lie nearby in their respective meta-spaces: T-S (tops-bottoms), T-S (tops-shoes), B-S (bottoms-shoes). **B:** Our modified DeiT, which fuses both visual and textual information.

3.1. Local and Global Augmentation

Recent visual compatibility models primarily focus on capturing global information high-level descriptors such as style, shape, and overall color. While such features are adequate for assessing visual similarity (for example, comparing two shirts for general resemblance), they are often insufficient for true visual compatibility between items from different categories. For instance, when determining whether a shirt and a pair of pants are compatible, subtle local features such as fabric texture, weave, fine-grained patterns, or trim become critical. These local cues govern how well different garments coordinate in real-world fashion, capturing nuances that global representations typically miss. Motivated by this, we propose a feature matching strategy based on both local and global augmentations: local crops ensure the model attends to patch-level details essential for compatibility, while global views preserve overall context. Empirical ablations (Table 4) confirm that integrating local features significantly enhances compatibility prediction accuracy over global features alone, validating the importance of our approach.

Let $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_K$ be the compatible item sets, where $\forall \mathcal{P}_i = \{x_i, y_i\}$, and x_i and y_i are the item pairs belonging to a compatible set. Assume that $\mathcal{G}$ represent the global augmentations, and $\mathcal{A}$ is the local augmentation. The global augmentation contains high-resolution images (224*224) with various transformations such as *RandomCrop*, *RandomGray*, *ColorJitter*, and *GaussianBlur*, while the local augmentation contains low-resolution images (a small random crop of the whole image with resolution 96*96) with the same image transformations applied mentioned for global augmentation. For the $\mathcal{P}_i^{\text{th}}$ compatible set, the local and global augmentations are defined as:

$$\boldsymbol{a}_1, \boldsymbol{a}_2 = \mathcal{G}(x_i), \mathcal{G}(y_i) \tag{1}$$

$$\boldsymbol{l}_1, \boldsymbol{l}_2 \ldots \boldsymbol{l}_k = \mathcal{A}(x_i), \mathcal{A}(x_i), \ldots \mathcal{A}(x_i) \tag{2}$$

Where k is the number of local augmentations, in the Eq-2 we can replace x_i with the y_i since it is symmetric. In practice, We keep two global and 8 local augmentations. The obtained local and global augmentation set for the compatible pair are used to train the model using the knowledge distillation. The following section discuss the details about the method and training.

3.2. Distantly Supervised Knowledge Distillation

The obtained local and global augmented images are used to distill knowledge from the teacher to the student network. Let $\mathcal{T}_\phi$ and $\mathcal{S}_\theta$ be the teacher and student models with parameters ϕ and θ, respectively. $\mathcal{T}_\phi$ and $\mathcal{S}_\theta$ are identical copies of the same vision transformer (ViT) [5]-based architecture. Here, we leverage the DeiT [37] pretrained model, which is highly efficient compared to the vanilla ViT model.

The global and local augmented images obtained using Eqs. 1 and 2 are passed to $\mathcal{T}_\phi$ and $\mathcal{S}_\theta$, respectively. The outputs from both networks are used to compute the loss, and the outputs from $\mathcal{T}_\phi$ and $\mathcal{S}_\theta$ are given as:

$$\boldsymbol{o}_i = \mathcal{T}_\phi(\boldsymbol{a}_i) \quad \& \quad \boldsymbol{s}_j = \mathcal{S}_\theta(f_c([\boldsymbol{l}_j, c_j]) \tag{3}$$

The loss over the output obtained is defined as:

$$\mathcal{L}(\theta) = \sum_{\forall \boldsymbol{o}_i} \sum_{\forall \boldsymbol{s}_j} CE\left(SM(\boldsymbol{o}_i), SM(\boldsymbol{s}_j)\right) \tag{4}$$

Here, CE is the cross-entropy loss, $SM(\boldsymbol{o}_i)$ and $SM(\boldsymbol{s}_j)$ are the softmax outputs of the teacher and student logits, respectively, and c_j is the color embedding[1]. Also, $f_c([\boldsymbol{l}_j, c_j])$ represents the concatenation of $\boldsymbol{l}_j$ and c_j, followed by a transformation with f_c, a fully connected layer. In Eq. 4, $\boldsymbol{o}_i$ acts as the ground truth since the parameter ϕ is frozen. Eq. 4 aligns the local and global features of the compatible pair, which significantly helps in learning the local features. The loss is optimized with respect to the student model parameters, i.e., θ, while the teacher model parameter ϕ is updated using the exponential moving average (EMA) with a rate of α, as given by:

$$\phi_t = \alpha\theta_t + (1-\alpha)\phi_{t-1} \tag{5}$$

The model update using EMA helps to overcome the mode collapse problem [1, 7], as it is highly likely that the model may project all the embeddings to the same point and minimize the loss to zero. This knowledge distillation leverages distantly supervised pairs of compatible item sets. The compatible item sets are obtained from images collected in the wild, where each clothing item detected in the image is considered a compatible set. Section 4 discusses the details of data collection; please refer to the same for further details.

3.3. Style Meta-Space Alignment

Let's assume that we have a compatible set $\{\text{shirt}, \text{pant}, \text{shoes}, \text{belt}\}$. To learn compatibility, we need to project these items into a compatible space where the items within the set are very close to each other, while non-compatible items are far apart. In this space, the model learns a common feature from the compatible set, which shows high similarity between all the embeddings learned in the compatible space. Learning a common feature from such a highly diverse set of compatible items is a difficult task, and the model may be unable to learn robust embeddings or may collapse to a single data point. To solve this issue, we can assume that the single joint space is composed of a collection of meta-spaces, where each meta-space represents a smaller portion of the whole space. We divide the compatible set into pairs of compatible items, such as $\{\text{shirt}, \text{pant}\}$, $\{\text{pant}, \text{shoes}\}$, $\{\text{shirt}, \text{shoes}\}$, etc. Each compatible pair is projected into a single meta-space. Previous work by Tan et al. [33] learned shared subspaces and achieved better performance than independent subspaces. However, their method requires input image pairs for subspace selection during inference, which is impractical for retrieval tasks. In contrast, our subspace indexing mechanism depends solely on item categories. Since our model requires only a single image and two category labels, we can construct a practical indexing approach suitable for retrieval applications.

The meta-space alignment based learning can be constructed by simply choosing the various compatible set from the whole possible compatible item pair and following the max margin loss. Assume that we have C type of compatible item pair, assume that one type pair is {shirt, pant} which is represented as (u, v) is the t^{th} pair. Lets assume that m_t is the learnable meta-space alignment vector for the t^{th} pair type. It projects the shared space embedding to the t^{th} meta-space. For compatible set $\{\boldsymbol{x}_i^u, \boldsymbol{y}_i^v\}$ meta-space projection is defined as:

$$\boldsymbol{e}_t = m_t \odot \mathcal{T}_\phi(\boldsymbol{x}_i^u) \tag{6}$$

$$\boldsymbol{e}_s = m_t \odot \mathcal{S}_\theta(f_c([\boldsymbol{y}_i^v, c_i])) \tag{7}$$

where $\odot$ is the element wise product which learns a weight to each dimensions.

3.4. Color and Text Information fusion

Color information plays a crucial role in visual compatibility, however, during knowledge distillation, the model focuses on learning style, texture, pattern, etc., and pays minimal attention to the color [16, 26]. To preserve the color information, we explicitly incorporate the color embedding into the model. To calculate the color embedding for each image, we transform the RGB space to the LUV [25] space and calculate the histogram in the LUV space. The LUV color space is a perceptually uniform color space, i.e., the perceived color difference between two points in the LUV space corresponds to the Euclidean distance between them. This property makes the LUV space particularly suitable for color-based computations, as it better aligns with human perception of color differences compared to other color spaces like RGB or CIELAB. The color vector is obtained by calculating the histogram of pixel values in the LUV space for each image. This color vector captures the distribution of colors present in the image, providing a compact yet informative representation of the color information. The color vector is concatenated with the output of the student backbone and passed through a linear layer to match the final embedding size, which is incorporated in the Eq.3 and Eq. 6.

Textual attributes such as material, print type, or style descriptors can provide complementary information that visual features may not fully capture. To effectively integrate text and image signals, we employ an early fusion mechanism inspired by recent multimodal transformers. Let t denote the attribute text associated with an item (e.g., "blue denim jeans" or "striped cotton shirt"). We use a pretrained CLIP [28] text encoder to obtain a fixed-length embedding vector $\boldsymbol{t} = \text{CLIP}_{\text{text}}(t) \in \mathbb{R}^d$. In our early fusion approach, for each image patch embedding $\boldsymbol{p}_i \in \mathbb{R}^d$ from the DeiT [37] visual encoder, we add a scaled CLIP text embedding $\lambda\boldsymbol{t}$ to obtain the fused representation $\boldsymbol{h}_i = \boldsymbol{p}_i + \lambda\boldsymbol{t}$. This addition is performed for every patch, allowing the textual information to modulate the representation of all image regions from the very first transformer layer.

[1]Color embeddings are defined in Section 3.4

3.5. Negative Sampling with Adaptive Margin

The max margin between the meta-space leads to improved model performance and robust to the noisy pairs. We leverages the Triplet Loss to achieve the maximum margin between the various meta-spaces. Triplet loss requires negative samples, and mining hard negatives typically yields the best performance. The annotation of hard negative samples are difficult. We incorporate a distantly supervised strategy to get the proxy hard negative samples. To obtained the approximate negative samples we first calculate the image embedding with the help of SkiLL [42] pretrained model. This image embedding are used for the tree construction using the agglomerative clustering approach. At the leaf level we choose the nearest cluster (other cluster of the same parent node) to draw the negative samples.

3.5.1. Approximate Hard Negative sampling:

Assume that we have a compatible set $\{\boldsymbol{x}_i^u, \boldsymbol{y}_i^v\}$ as positive pair of type (u, v). From the constructed tree we identify the cluster l_0 from the category v where the sample $\boldsymbol{y}_i^v$ resides. Then we sample l number of nearest clusters from the positive sample l_0. These cluster are close to positive cluster however they contains the negative samples compared to the compatible set $\{\boldsymbol{x}_i^u, \boldsymbol{y}_i^v\}$. Therefore these cluster samples provides the approximately hard negative samples. We sample n number of samples from each negative clusters and collected total of $l \times n$ negative samples. Now, negative sample $\boldsymbol{z}_i^u$ can become any of the $l \times n$ images or we can can take every sample and compute average loss for triplet.

3.5.2. Adaptive Margin for Triplet Loss:

For a compatible positive pair $\{\boldsymbol{x}_i^u, \boldsymbol{y}_i^v\}$ lets assume $T_i = \{\boldsymbol{z}_j^v)\}_{j=1}^{l \times n}$ is the set of negative pair (we keep small value $l = 2$ and $n = 2$) for the model's training efficiency. Now the Triplet Loss can be computed as:

$$\mathcal{M}_{\theta, m_t}(\boldsymbol{x}_i^u, \boldsymbol{y}_i^v, \boldsymbol{z}_i^v) = \frac{1}{Z_T} \sum_{z \in T_M} \max\{0, d(\boldsymbol{e}_t, \boldsymbol{e}_s) - d(\boldsymbol{e}_t, \boldsymbol{e}_z) + \alpha_z\}$$

where $\boldsymbol{e}_z = m_t \odot \mathcal{S}_\theta(f_c([\boldsymbol{y}_i^v, c_i]))$ and α_z is a dynamic violate margin, which is different from the constant margin of traditional triplet loss. It is computed according to the class relationship between the anchor class y_a and the negative class y_n over the constructed hierarchical class tree. Specifically, for triplet the violate margin α_z is computed as:

$$\alpha_z = \beta + d_H(c_y, c_z) - s_{c_y} \tag{8}$$

where β (= 0.4) is a constant parameter that encourages the image clusters to reside further apart from each. c_x, c_y, c_z are the clusters where $\boldsymbol{x}_i^u, \boldsymbol{y}_i^v, \boldsymbol{z}_i^v$ resides. $d_H(c_x, c_z)$ is the threshold for merging the clusters c_y and c_z at a certain level on the hierarchical tree. s_{c_y} is the average distance between samples in the cluster c_y. In our adaptive triplet loss, a sample from c_y is encouraged to push the nearby points with different semantic meanings apart from itself (a formal trouser with a matching formal shirt should push a Jeans pant apart). Furthermore, it also contributes to the gradients of positive and negative data points which are very far from each other. To note that, in contrast to [6], we computed the whole tree structure and the thresholds for merging nodes only once.

Figure 2. The data collection strategy involves the following steps: each item from the image is detected and cropped, and this set is considered as a compatible set. Further, we retrieve similar items from the database and augment the compatible set.

4. Data Collection

The deep learning model requires a huge amount of labeled data. However, for the visual compatibility model, a large number of compatible item sets are required, which necessitates domain experts and the collection of a large amount of samples. This is costly and time-consuming. A strong pre-trained model may reduce the amount of data required, but most pre-trained models are available for classification tasks, which has significantly different objective compared to visual compatibility. Therefore, these pre-trained models do not work and show limited impact on model performance. There are a few available datasets, such as Polyvore [8] and Hypatia-OutfitBuilder [10]. These datasets have limited samples, and Polyvore dataset is highly biased towards color.

To overcome the above limitations and enrich the dataset, we follow a distantly supervised strategy. We collected complete outfit images from various open source sites, which captures real-world scenarios of human-worn outfits. We employed the Grounding-DINO [21] model to detect individual clothing objects, then cropped and mapped them to product catalog images using a visual similarity model SkiLL [42]. We ensured color harmony in outfit recommendations by matching product retrievals to original outfit colors. We expanded the Hypatia-OutfitBuilder dataset by retrieving visually similar products from softlines image catalog for each item and augmented the dataset with additional relevant options. Our final dataset comprised 2MN outfit pairs, providing a rich and diverse foundation for training our models. This process enabled us to establish connections between outfit components and available catalogs, enriching the dataset and improving model performance. Figure 2, shows the overview of the distantly supervised strategy.

Methods	Polyvore disjoint				Polyvore nondisjoint			
	FITB	R@10	R@30	R@50	FITB	R@10	R@30	R@50
Type-Aware [38]	55.65	3.66	8.26	11.98	57.83	3.50	8.56	12.66
SCE-Net Averag [33]	53.67	4.41	9.85	13.87	59.07	5.10	11.20	15.93
CSA-Net [20]	59.26	5.93	12.31	17.85	63.73	8.27	15.67	20.91
Outfit transformer (Vision only) [30]	-	6.03	12.20	16.51	58.92	9.29	16.94	21.82
MEDAL (Vision only)	61.70	7.97	14.81	20.03	66.50	9.22	17.44	23.34
Outfit transformer [30]	59.48	6.53	12.12	16.64	67.10	9.58	17.96	21.98
HAT [12]	57.32	5.13	10.04	15.29	64.87	7.46	15.74	20.38
MEDAL (Vision+Text)	**63.20**	**8.32**	**16.38**	**21.83**	**69.80**	**11.05**	**20.87**	**27.66**

Table 1. Comparison of our model with state-of-the-art methods on the FITB (using accuracy) and CIR tasks (using %recall@top-k).

5. Implementation details

We adopted the DeiT architecture [36] as the backbone for our student and teacher models, owing to its promising performance in low-data regimes. Our models employed patch sizes of eight, which, although more computationally expensive compared to patch sizes of 16, yielded superior results. To incorporate color information, we calculated a color histogram over 560 different colors, resulting in a 560-dimensional color vector. This vector was then passed through a linear layer to reduce its dimensionality to 64. The training process involved a batch size of 12, distributed across eight GPUs. We utilized the AdamW optimizer with an initial learning rate of 0.00001. Our learning rate schedule followed a similar approach to that in [42], where linear scaling was applied for the first 10 epochs, followed by cosine scheduling for learning rate decay. The temperature parameter, τ, played a crucial role in our model, and we set it to a lower value within the range $\tau \in [0.04, 0.07]$. As the training progressed, we increased the l_2 penalty from an initial value of 0.04 to $10\times$ its original value. The merging threshold for nodes at different tree levels and the whole hierarchical tree structure in Section 3.5 were computed using the $children\_$ and $distances\_$ functionality provided by the sklearn agglomerative clustering package. this is a one time offline calculation during training. We exclusively deploy the Student model for inference. This inference pipeline incorporates the LUV color embeddings, textual information, and Meta-Space alignment, while the computationally intensive local-global augmentations and approximate negative sampling are restricted to the training phase. Please refer to supplement for more details.

6. Results and Evaluations

We conducted the rigorous experiment over the various task and metric to evaluate the proposed model. The following section discuss the details about the various task, evaluation metric and results.

6.1. Baselines and Evaluation Metric

For evaluation, we utilized disjoint and nondisjoint test set from the Polyvore to calculate the Compatible Fill-in-the-Blank (FITB), Complementary Prediction (CP) and Complementary Item Retrieval (CIR) scores, which are widely used metrics for assessing the performance of outfit recommendation systems. Additionally, we calculated only CIR scores in our in-house test dataset which consists of 200,000 samples.

In our study, we conducted a comprehensive evaluation of our proposed method by comparing its performance against several state-of-the-art approaches, including the Type-aware Net [38], SCE-Net [33], CSA Net [20], Outfit Transformer [30] and HAT [12]. HAT was mainly evaluated for Polyvore-U [12] dataset. We reproduce it and report the metric on Polyvore-disjoint and Polyvore-nondisjoint datasets [38]. To assess the effectiveness of our method, we employed widely-used evaluation metrics, including Area Under the Curve (AUC) for complementary prediction task, accuracy for Fill-in-the-Blank (FITB) task and Recall@K for complementary item retrieval task. The AUC metric measures the ability of a model to distinguish between positive and negative instances, while the FITB metric evaluates the model's capability to predict the correct item given a partially observed outfit. We calculated these metrics on both the disjoint and non-disjoint sets of the Polyvore dataset, which is a widely-used benchmark for fashion recommendation tasks.

6.2. Complementary prediction and FITB

We evaluated our proposed model on outfit compatibility prediction (OCP) and fill-in-the-blank (FITB) tasks using the Polyvore Outfit [8] test datasets. We trained our proposed model on Polyvore disjoint (16995 training outfits) and nondisjoint (53306 train outfits) training set and utilized respective disjoint (15,145 test outfits) and non-disjoint (10,000 test outfits) test sets to evaluate the metrics. The disjoint set ensures a strict separation between training and test/validation sets, while the non-disjoint set allows some individual items to overlap. In FITB, incorrect choices are sampled from the same category as the correct choice, with fine-grained item type annotations. The task is to select the most compatible candidate item, evaluated by overall accuracy. For OCP, we predict the compatibility of fashion items in an outfit, reporting AUC. Our model is benchmarked against state-of-the-art methods, and we report performance on both FITB and OCP tasks using the Polyvore Outfit dataset. We have not calculated CP and FITB tasks in our internal test

	R@1	R@5	R@10	R@20	R@50	R@100
Type-aware [38]	0	0	0	0	0	0
CSA-Net [20]	0.26	1.22	2.14	3.67	2.34	4.03
Oufit Transformer [30]	0.58	1.99	3.98	5.84	3.31	4.26
HAT [12]	0.52	1.36	3.53	5.61	3.18	3.86
MEDAL (Ours)	**3.86**	**11**	**13.56**	**15.6**	**12.55**	**12.13**

Table 2. Evaluation for the retrieval task, where we have shown relative %Recall@K (R@K) wrt Type-aware, and $K = \{1, 5, 10, 20, 50, 100\}$ (Outfit Transformer and HAT are multi-modal model which takes both text and image, for fair comparison we have removed the text from them.)

dataset as we don't have any annotation for that.

The results for the FITB over the Polyvore dataset are shown in Table 1. We can observe that for both the Polyvore disjoint and non-disjoint test sets, the proposed approach (MEDAL) shows consistently better results. For the Polyvore disjoint and non-disjoint sets, our approach shows 3.72% and 2.7% absolute improvement in the FITB task respectively when we incorporated both textual and visual information. When we incorporated vision only information we see that our model has outperformed the previous baseline CSA-Net in the FITB task. In the CP task given in Table 3 we have seen our model outperforming the recent baseline by 1.0% and 2.0% for the Polyvore disjoint and nondisjoint test set respectively and seen similar results in case of vision only mode. In the Polyvore disjoint dataset, our model outperformed recent baseline in the CP task in both vision only and vision plus textual mode.

6.3. Complementary item retrieval

We trained our model on the Polyvore disjoint and nondisjoint train sets and performed the CIR task on the respective test sets. For this task, we used recall@top-k (abbreviated as R@k) as the metric. For the calculation of R@k, we adopted the same methodology described in CSA-Net to evaluate the Polyvore disjoint and nondisjoint sets. We indexed our dataset by category, treating each category as the target category. For indexing, we used the same methodology described in CSA-Net. The results are shown in Table 1. We observe that our model outperforms all recent baselines in almost all R@k metrics for both vision-only and joint vision-text models. In the Polyvore disjoint set, we saw a significant improvement of 1.79%, 4.26%, and 5.19% in R@10, R@30, and R@50, respectively, compared to recent baselines. Another significant improvement of 1.44%, 2.91%, and 5.68% was observed in R@10, R@30, and R@50 for Polyvore nondisjoint set. We also noticed that our vision-only model outperformed the Outfit Transformer in both the textual and visual modes for the disjoint set by 1.44%, 2.68%, and 3.39% for R@10, R@30, and R@50, respectively. However, our vision-only model performed slightly worse than its Outfit Transformer counterpart in the Polyvore nondisjoint set for R@10.

Method	Features	PO-D	PO
BiLSTM + VSE [8]	Image + Text	0.62	0.65
GCN (k=0) [17]	Image	0.67	0.68
SiameseNet [38]	Image	0.81	0.81
Type-Aware [38]	Image + Text	0.84	0.86
SCE-Net [33]	Image + Text	-	0.91
CSA-Net [20]	Image	0.87	0.91
OutfitTransformer [30]	Image	0.87	0.92
OutfitTransformer [30]	Image + Text	0.88	0.93
HAT [12]	Image + Text	0.88	0.92
Ours	Image	0.89	0.92
Ours	Image + Text	**0.89**	**0.95**

Table 3. Comparison of our model with state-of-the-art methods on the compatibility prediction task using the AUC metric [8]. The methods are evaluated on Polyvore-Outfits (where -D denotes the disjoint dataset).

We trained the proposed model on our in-house data described in Section 4. We did not use any textual information in this dataset, as obtaining textual descriptions for real-world fashion images is challenging. For testing, we indexed 200,000 outfits as the test set, as explained earlier in Section 4. For each outfit in the test set, we masked or removed one item, such as a shirt, pants, or accessory. The goal is to retrieve the masked item from the respective category within the indexed dataset, mimicking a real-world scenario where a user might have a partially constructed outfit, and the system needs to recommend the missing item(s) to complete the ensemble. We employed the R@k metric to assess the retrieval performance, as in the Polyvore dataset. To determine the relevance of the retrieved products with the ground truth, we utilized a visual similarity model, SkiLL [42], which compares the visual features (e.g., color, texture, style) of the retrieved products with the ground truth masked item to quantify their similarity. For each query, we computed the similarity or relevance score of the retrieved items with respect to the ground truth data. We considered a retrieval successful if at least one of the top-k retrieved items had a relevance score greater than 0.7. Subsequently, we calculated the overall recall by aggregating the successful retrievals across the entire dataset. We calculated R@K, where $K = 1, 5, 10, 20, 50, 100$, numbers are in absolute percent difference. In contrast to the Polyvore dataset, we used lower

Figure 3. Qualitative retrieval result of the proposed model over the Polyvore dataset. **Left:** is the query outfit with item in black box in the right being ground truth. **Right:** top-10 retrieval results for the ground truth category are shown.

Methods	Polyvore disjoint				Polyvore nondisjoint			
	FITB	R@10	R@30	R@50	FITB	R@10	R@30	R@50
MEDAL (Vision Only)	61.70	7.97	14.81	20.03	66.50	9.22	17.44	23.34
Without Negative clustering	61.30	7.86	15.35	20.11	68.46	10.92	20.13	26.39
Without Color	61.10	7.75	15.14	20.17	67.70	10.64	19.33	25.10
With RGB	60.72	7.46	14.67	19.53	67.44	10.58	19.16	24.72
With LAB color Space	62.88	8.19	15.74	20.91	68.79	10.71	20.03	26.38
Without Meta-Space	62.40	8.13	15.56	20.72	68.20	10.89	19.94	25.85
Without Local Augmentations	58.92	5.23	11.41	16.75	65.87	8.33	17.27	23.05
Without Triplet Loss	57.70	4.87	10.65	15.03	64.40	7.90	16.07	21.40
MEDAL (Vision+Text+LUV)	**63.20**	**8.32**	**16.38**	**21.83**	**69.80**	**11.05**	**20.87**	**27.66**

Table 4. Ablations for the Polyvore dataset for the retrieval and FITB task over the various proposed proposed components. R@k: Recall@k

top-k values, as in an e-commerce setting we generally do not recommend more than 10 items. In Table 2, we show the results for the retrieval task. We observe that MEDAL outperformed all the retrieval baselines by a significant margin with +9.58 absolute percentage difference compared to state of the art Outfit Transformer.

7. Ablations

Table 4 isolates the specific contribution of each component. Our analysis identifies Triplet Loss and Local Augmentation as the dominant factors; removing them causes the sharpest performance degradation (dropping FITB by $\sim$4.0% and $\sim$2.8% respectively), confirming their critical role in capturing fine-grained compatibility. We also validated our Approximate Hard Negative sampling against a standard random sampling baseline (Without Negative Clustering in Table 4); the clustering approach yields superior generalization by forcing the model to distinguish between semantically similar but stylistically incompatible items. While LUV color and Meta-Space provide smaller individual gains, they remain essential for resolving specific aesthetic ambiguities. Crucially, the full MEDAL framework outperforms any partial configuration, demonstrating that these components are synergistic rather than merely additive. Finally, integrating text further robustifies the model against categorical confusion, validating the necessity of our design.

8. Conclusion

Developing robust models for visual compatibility remains challenging for real-world deployment. This paper proposes a robust approach to predicting visual compatibility, applicable to both the hardlines and softlines categories. In this work, we introduce a novel approach that leverages distantly supervised knowledge distillation, meta-space alignment, and approximate negative sampling to build the visual compatibility model. The proposed local-global augmentation captures fine-grained local information, which plays a key role in extracting compatible features while ignoring irrelevant global features. Since color significantly influences visual compatibility, we incorporate explicit color embedding and triplet loss to train a model that effectively distinguishes between compatible and non-compatible items. Also, to conduct experiments in the wild, we constructed a large-scale in-house visual compatibility dataset using an object detection approach. We conducted extensive experiment over the in-house and publicly available dataset. The proposed approach across various tasks and metrics demonstrate its superiority compared to the baseline approaches. Ablation studies of the various components disentangle the contributions of each element in the proposed model.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[2] Qiangqiang Chen, Tianyi Zhang, Mengwei Nie, Zhiwei Wang, Shicheng Xu, Wen Shi, and Zhen Cao. Fashion-gpt: Integrating llms with fashion retrieval system. In *Proceedings of the First Workshop on Large Generative Models Meet Multimodal Applications (LGM3A)*, pages 69–78, 2023.

[3] WenFeng Chen, Pipei Huang, Jiaming Xu, Xin Guo, Cheng Guo, Fei Sun, Chao Li, Andreas Pfadler, Huan Zhao, and Binqiang Zhao. Pog: Personalized outfit generation for fashion recommendation at alibaba ifashion. In *SIGKDD*, 2019.

[4] Guillem Cucurull, Perouz Taslakian, and David Vazquez. Context-aware visual compatibility prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[6] Weifeng Ge, Weilin Huang, Dengke Dong, and Matthew R. Scott. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European conference on computer vision (ECCV)*, pages 269–285, 2018.

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc.

[8] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1078–1086, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Hypatia. Outfit builder outfits. https://mldatasets.aka.corp.amazon.com/OutfitBuilderOutfits/1.

[11] Junkyu Jang, Eugene Hwang, and Sung-Hyuk Park. Lost your style? navigating with semantic-level approach for text-to-outfit retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8066–8075, 2024.

[12] Myong Chol Jung, Julien Monteil, Philip Schulz, and Volodymyr Vaskovych. Personalised outfit recommendation via history-aware transformers. In *Proceedings of the 18th ACM International Conference on Web Search and Data Mining (WSDM '25)*, pages 633–641. Association for Computing Machinery, 2025.

[13] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian McAuley. Visually-aware fashion recommendation and design with generative image models. In *ICDM*, 2017.

[14] Katrien Laenen and Marie-Francine Moens. Attention-based fusion for outfit recommendation. In *Proceedings of the Workshop on Recommender Systems for Fashion (FashionXRecSys)*, 2019.

[15] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer. *arXiv preprint arXiv:1810.00825*, 2018.

[16] Attila Lengyel, Ombretta Strafforello, Robert-Jan Bruintjes, Alexander Gielisse, and Jan van Gemert. Color equivariant convolutional networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Kedan Li, Chen Liu, and David Forsyth. Coherent and controllable outfit generation, 2019.

[18] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo. Mining fashion outfit composition using an end-to-end deep learning approach on set data. *IEEE Transactions on Multimedia*, 2017.

[19] Zhi Li, Bo Wu, Qi Liu, Likang Wu, Hongke Zhao, and Tao Mei. Learning the compositional visual coherence for complementary recommendations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1333–1342, 2020.

[20] Yen-Liang Lin, Son Tran, and Larry S Davis. Fashion outfit complementary item retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3311–3319, 2020.

[21] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[22] Alexander Lorbert, David Neiman, Arik Poznanski, Eduard Oks, and Larry Davis. Scalable and explainable outfit generation. In *CVPR Workshop*, 2021.

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

[24] Soham Mitra, Atri Sukul, Swalpa Kumar Roy, Pravendra Singh, and Vinay Kumar Verma. Scorecam++: Gated score-weighted visual explanations for cnns. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2700–2709, 2025.

[25] Yoshi Ohno. Cie fundamentals for color measurements. In *NIP & Digital Fabrication Conference*, pages 540–545. Society of Imaging Science and Technology, 2000.

[26] Felix O'Mahony, Yulong Yang, and Christine Allen-Blanchette. Color equivariant network. *arXiv preprint arXiv:2406.09588*, 2024.

[27] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Ioannis Kompatsiaris. Victor: Visual incompatibility detection with transformers and fashion-specific contrastive pre-training. *Journal of Visual Communication and Image Representation*, 90:103741, 2023.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[29] Soumya Roy, Vinay Verma, and Deepak Gupta. Efficient expansion and gradient based task inference for replay free incremental learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1165–1175, 2024.

[30] Rohan Sarkar, Navaneeth Bodla, Mariya Vasileva, Yen-Liang Lin, Anurag Beniwal, Alan Lu, and Gerard Medioni. Outfittransformer: Outfit representations for fashion recommendation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2263–2267, 2022.

[31] Yong-Siang Shih, Kai-Yueh Chang, Hsuan-Tien Lin, and Min Sun. Compatibility family learning for item recommendation and generation. In *AAAI*, 2018.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[33] Reuben Tan, Mariya I. Vasileva, Kate Saenko, and Bryan A. Plummer. Learning similarity conditions without explicit supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10373–10382, 2019.

[34] Meet Taraviya, Anurag Beniwal, Yen-Liang Lin, and Larry Davis. Personalized compatibility metric learning. In *KDD Workshop*, 2021.

[35] Sambeet Tiady, Arihant Jain, Dween Rabius Sanny, Khushi Gupta, Srinivas Virinchi, Swapnil Gupta, Anoop Saladi, and Deepak Gupta. Merlin: Multimodal & multilingual embedding for recommendations at large-scale via item associations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, 2024.

[36] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[37] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[38] Mariya I. Vasileva, Bryan A. Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European conference on computer vision*, pages 390–405, 2018.

[39] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Proceedings of the IEEE international conference on computer vision*, pages 4642–4650, 2015.

[40] Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2017.

[41] Vinay Verma, Dween Sanny, Abhishek Singh, and Deepak Gupta. Cod: Coherent detection of entities from images with multiple modalities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8015–8024, 2024.

[42] Vinay K Verma, Dween Rabius Sanny, Shreyas Sunil Kulkarni, Prateek Sircar, Abhishek Singh, and Deepak Gupta. Skill: Skipping color and label landscape: self supervised design representations for products in e-commerce. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3502–3506, 2023.

[43] Xin Wang, Bo Wu, and Yueqi Zhong. Outfit compatibility prediction and diagnosis with multi-layered comparison network. In *Proceedings of the 27th ACM International Conference on Multimedia (MM)*, pages 2022–2031, 2019.

[44] Yiyan Xu, Wenjie Wang, Fuli Feng, Yunshan Ma, Jizhi Zhang, and Xiangnan He. Diffusion models for generative outfit recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1350–1359, 2024.

[45] Cong Yu, Yang Hu, Yan Chen, and Bing Zeng. Personalized fashion design. In *ICCV*, 2019.