A^3-FKG: Attentive Attribute-Aware Fashion Knowledge Graph for Outfit Preference Prediction

Huijing Zhan, Jie Lin, Kenan Emir Ak, Boxin Shi, Ling-Yu Duan, and Alex C. Kot

***Abstract*—With the booming development of the online fashion industry, effective personalized recommender systems have become indispensable for the convenience they brought to the customers and the profits to the e-commercial platforms. Estimating the user's preference towards the outfit is at the core of a personalized recommendation system. Existing works on fashion recommendation are largely centering on modelling the clothing compatibility without considering the user factor or characterizing the user's preference over the single item. However, how to effectively model the outfits with either few or even none interactions, is yet under-explored. In this paper, we address the task of personalized outfit preference prediction via a novel Attentive Attribute-Aware Fashion Knowledge Graph (A^3-FKG), which is incorporated to build the association between different outfits with both outfit- and item- level attributes. Additionally, a two-level attention mechanism is developed to capture the user's preference: 1) User-specific relation-aware attention layer, which captures the user's fine-grained preferences with different focus on relations for learning outfit representation; 2) Target-aware attention layer, which characterizes the user's latent diverse interests from his/her behavior sequences for learning user representation. Extensive experiments conducted on a large-scale fashion outfit dataset demonstrate significant improvements over other methods, which verify the excellence of our proposed framework.**

***Index Terms*—Personalized Preference Prediction, Knowledge Graph, Attribute-Aware, Attention, Multi-Modal.**

I. Introduction

Recent years have witnessed the explosive development and the tremendous profits online fashion shopping has brought. It has been reported that the global fashion e-commerce revenue is expected to embrace double growth compared to that in 2018, reaching $872 billion by 2023 [1]. Despite the promising economic benefits, the enormous amount of fashion products on the Internet make it intractable for online shoppers to seek their preferred outfits. This inspires us to develop an accurate and high-quality personalized outfit preference prediction system that assists to capture customer's fashion tastes and further facilitates the recommendation services.

Research studies on fashion recommendation [2]–[10] mainly focus on learning the general compatibility between different items within a fashion outfit. Vasileva *et al.* [3] proposed to embed items of different categories into separated subspaces for compatibility modeling. Li *et al.* [7] employed Recurrent Neural Network (RNN) to represent the outfits with variable length items for quality scoring. Cui *et al.* [8] established a fashion graph to model the co-occurrence dependency between items of different categories. However, the problem of personalized fashion recommendation (*i.e.*, preference prediction), especially outfit recommendation, which considers the user's factor is still under-investigated.

Fig. 1: Examples of outfits created by two users. User A likes casual-style clothes while User B prefers black suits and high-heeled shoes. Prior works mainly focus on the interactions (denoted by blue solid arrows) between users and items, and ignore the connections between different outfits (denoted by orange dashed edges). Best viewed in color.

Even though there have been numerous research attempts [11]–[23] devoted to the personalized fashion, most of them are item-based approaches which model the user's interest patterns towards the individual items. And research studies [15], [17], [21], [23] on personalized outfit recommendation is still at the preliminary stage. Chen and McAuley *et al.* [17], [23] attempted to aggregate the item features into a unified outfit representation via a pre-defined strategy and performed the recommendation in the same manner as the item-based recommendation. Song *et al.* [15] defined the user-outfit preference score as the weighted sum of user's preference scores towards individual items and the pairwise item relationships. Hidayati *et al.* [21] developed a personalized style recommendation framework given the user's body measurements. The most similar work to ours is [23], which also adopted Graph Neural Networks (GNNs) for outfit representation learning. Whereas, we propose to effectively inject the fashion semantics into the knowledge graph and the attentive mechanism is integrated in different-level hierarchy, which enables us to learn enhanced user/outfit feature representation.

H. Zhan, J. Lin and K. E. Ak are with the Institute for Infocomm Research (I2R), Singapore 138634. (e-mail: zhan-huijing@i2r.a-star.edu.sg, lin-j@i2r.a-star.edu.sg, kenan_emir_ak@i2r.a-star.edu.sg).

B. Shi and L.-Y. Duan are with the National Engineering Laboratory for Video Technology, Department of Computer Science and Technology, Peking University, Beijing 100871, China (e-mail: shiboxin@pku.edu.cn, lingyu@pku.edu.cn).

A. C. Kot is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: eackot@ntu.edu.sg).

Despite the effectiveness of [23], the inter-relations between outfits (*e.g.,* fashion items with same colors or outfits sharing similar styles) are ignored in learning the user-outfit preference predictor. As shown in Fig. 1, given the user-created outfits in the leftmost column with white tops, it is highly possible that user A would show interest in the outfits on the right, due to the similar clothing style and tops in the same color. The same rule also applies to user B who has a preference on black suits and high heels. It is worth mentioning that the exploration of connectivity among outfits is of great importance in dealing with few interactions between users and outfits. Without building relationships with existing outfits, it poses tremendous challenges to infer the user's preference toward "fresh" outfits with no history ratings.

This motivates us to take a closer look into the following challenges when modeling the user preference: Q1) How to capture the connectivity between different outfits with few mutual interactions and further quantify their mutual effects? Q2) How to explore different user's interests toward distinctive perspectives of the outfit from their historical interaction records? Q3) How to provide convincing reasons with semantics for the preference prediction results?

In this paper, we construct a novel **A**ttentive **A**ttribute-**A**ware **F**ashion **K**nowledge **G**raph (A^3-FKG) to capture the structural connectivity between entity nodes, which can be outfits, items and attributes (answer to Q1). Moreover, the outfit representation is enriched by the entity embedding, propagation from the entities' connected neighbors and multi-modal fashion items' content features, *e.g.,* visual image and textual description. More specifically, a relation-aware attention layer is developed to characterize the individual user's fine-grained interests, *i.e.,* attributes of the outfits, which facilitates learning the outfit representation (answer to Q2). Also, the user's implicit and diverse tastes are modeled via the weighted aggregation of his/her previously interacted outfits via a novel target-aware attention layer (answer to Q2), which is also part of the user representation. Finally, the enriched user and outfit representations are utilized to estimate the personalized preference score. The connections within the knowledge graph bridged by attributes and the attentive network structure jointly contribute to make the prediction explainable (answer to Q3). The effectiveness of the proposed system and the advantages of the components are demonstrated on the real-world large-scale fashion outfit dataset with high sparsity interactions.

Our main contributions of this work can be summarized as follows:

- To the best of our knowledge, we are the first to construct a comprehensive attribute-aware fashion knowledge graph for personalized outfit preference prediction problem with thorough experimental analysis.
- We develop a two-level attention network to discriminate individual user's fine-grained interests and latent multiple tastes for outfit composition.
- We take advantage of the multi-modality content features and structural entity/relation representation for a more enhanced outfit representation.
- Extensive experimental results demonstrate the significant improvements over other baselines and the effectiveness of different components of our framework.

II. Related Works

A. Knowledge Graph Guided Representation Learning

Existing KG-based representation learning approaches can be generally categorized into three types according to how the information of knowledge graph is utilized: 1) Embedding-based approaches [24], [25] which employ knowledge graph embedding (KGE) algorithm to encode the KG into low-dimensional entity and relation embedding, which are further incorporated into recommendation. Representative KGE algorithms include TransE [26], TransH [27], TransR [28], TransD [29], *etc*. However, these methods infer the user's preference in an implicit manner by representation learning thus the recommendation objective isn't optimized directly; 2) Path-based approaches [30], [31] which take advantage of the semantic connectivity patterns of the graph nodes for the recommendation. However, it is not practical to include all the candidate paths in the large-scale scenario and 3) Hybrid approaches which leverage both the semantic connectivity information and entity representation [32]–[34]. Our proposed method can be seen as a hybrid approach to learn the structured entity representations with semantics enriched by fine-grained clothing attributes. Moreover, to the best of our knowledge, our work is among the pioneering attempt to extract the domain knowledge from meta-data and take advantage of the external dataset to build a comprehensive attribute-aware fashion knowledge graph for modeling users' personalized preference toward outfits.

B. Attention Mechanism

Recent years have witnessed the success of attention-based neural networks in a variety of tasks, ranging from question answering [35], image caption [36] to image generation [37], sentiment analysis [38], *etc*. As to its impact on the recommender system, it not only greatly boosts the performance but also offers the reasons for the recommendation in terms of learnt attentive weights. He [39] adopted an item- and outfit-level attention network to handle the implicit feedbacks in the context of micro-video and image-based recommendation. Feng [40] developed a personalized food recipe recommendation system via a hierarchical attention network at the ingredient and component level, which jointly considered the interaction between food ingredients, recipe images and historical interactions. Inspired by the increasing popularity to incorporate attention-based into Graph Convolutional Networks (GCNs) [41], such as Graph Attention Networks (GATs) [42], Relational Graph Convolutional Network (R-GCN) [43], we also develop a user-specific relation-aware attention to aggregate neighborhood information. Distinct from [33], our method considers the neighborhood entity and its connected relation as a holistic set to calculate relation-level weights. Moreover, a target-aware attention scheme is developed to discriminate the impact of historically interacted outfits over characterizing the user's diverse interests.

C. Interaction Sparsity

One of the most common problems in the personalized preference prediction is the high level of sparsity within the user-item interaction matrix. Latent factor-based models [44], [45] might fail to capture the collaborative signals between users or items. To tackle the issue, many solutions have been raised, which can be further categorized into three forms: 1) Item's side information including textual modality data [46] (reviews, description, *etc*), visual modality data (image, *etc*); 2) User's profile [47] (age, friend circle, social relationships, *etc*) or purchase behavior [48]; and 3) Transforming the user-interaction matrix to user-item bipartite graph in order to connect potential similar users/items in multi-hop neighbor interlinks [49]. The first and second types of side information injected into the preference embedding suffer the problem of additional regularization term which makes the optimization more complicated. As to the graph-based learning strategies of the third type usually cannot capture enough connections especially in the extremely sparsity case. Our approach takes advantage of their merits by incorporating the fine-grained outfit- and item-level attributes into knowledge graph as well as item's multi-modal content features for outfit representation learning. Also, calculating the similarity between the target outfit and the users' historically engaged ones helps to infer their preference patterns.

III. Problem Formulation

Suppose we have a set of users $\mathcal{U} = \{u_1, u_2, \cdots, u_{|\mathcal{U}|}\}$ and a set of outfits $\mathcal{O} = \{o_1, o_2, \cdots, o_{|\mathcal{O}|}\}$. Each outfit o_j, $1 \leq j \leq |\mathcal{O}|$ is composed of a sequence of items $S_j = \{s_{j1}, s_{j2}, \cdots, s_{j|o_j|}\}$, where $|o_j|$ denotes the number of items in the j-th outfit. Each item $s_{jk}, 1 \leq k \leq |o_j|$ is associated with an image x_{jk} and the textual description t_{jk}, *e.g.,* item's meta-data or title. The user-outfit interaction matrix $\mathcal{Y} \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{O}|}$ represents the implicit feedback. Here $y_{uo_j} = 1$ denotes an engagement between user u and outfit o_j. In addition, we also have the outfit's side information (*e.g.,* attributes of fashion items, number of likes for outfits) which are utilized for the construction of attribute-aware fashion knowledge graph $\mathcal{G}$, denoted as $\mathcal{G} = \{(h, r, t)|h, t \in \mathcal{E}, r \in \mathcal{R}\}$. Here (h, r, t) indicates the KG triple, which means that there is an edge characterized by r from head entity h to tail entity t.

With the user-outfit interaction matrix $\mathcal{Y}$, fashion knowledge graph $\mathcal{G}$ and item set $\mathcal{S} = \{(x_{jk}, t_{jk})|o_j \in \mathcal{O}\}$, our goal is to learn a personalized preference prediction function $\hat{y}_{uo_j} = \mathcal{F}(u, o|\Theta, \mathbf{Y}, \mathcal{G}, \mathcal{S})$ between user u and outfit o with which he has never interacted before. Here Θ represents the model parameters. For ease of reading, Table II lists the notations throughout the paper.

IV. The Proposed Framework

A. An Overview of A^3-FKG

The overall pipeline of the proposed framework A^3-FKG is demonstrated in Fig. 2. Given a user u_1 and a candidate outfit o_3, an attribute-aware fashion knowledge graph $\mathcal{G}$ is constructed (see Section IV-B) based on the outfit/item's description from the dataset and predicted attribute labels with the classifier pre-trained on the external clothing data. Our proposed framework encompasses three components: 1) Outfit encoder (see Section IV-C); 2) User encoder (see Section IV-D) and 3) Outfit preference predictor (see Section IV-E). In learning the outfit representation, a user-specific relation-aware attention layer (see Section IV-C-1) is utilized to accumulate the features from the neighborhood entities of o_3 considering user's interest on different types of relation. An aggregation layer is leveraged to combine the self-feature $\mathbf{e}_{o_3}$ and propagated feature $\tilde{\mathbf{e}}_{\mathbf{o_3}}$ into a unified vector $\mathbf{g}_{o_3}$ before feeding into the weighting matrix $\mathbf{W}_a$. Meanwhile, its multi-modal content features $\mathbf{z}_{o_3}$ are extracted (see Section IV-C-2) on the sequence of fashion items. Finally, a knowledge-aware image-word feature aggregation sub-module (see Section IV-C-3) is introduced to aggregate $\mathbf{g}_{o_3}$ and $\mathbf{z}_{o_3}$ into a joint outfit representation $\mathbf{F}_{o_3}$. In learning the user representation, it is composed of two terms: general user ID embedding and fine-grained preference term, calculated by the summation of historically interacted outfits via personalized weights. Finally, the overall preference score $\hat{y}(u_1, o_3)$ is computed as the inner product between the learnt user and outfit embedding. In the following sections, more details are provided about each component of the framework.

B. Attribute-aware Fashion Knowledge Graph Construction

Fig. 3 illustrates the construction of the proposed attribute-aware heterogeneous fashion knowledge graph $\mathcal{G} = \{(h, r, t)|h, t \in \mathcal{E}, r \in \mathcal{R}\}$ with different types of entities and relations. Here the entity set $\mathcal{E}$ is comprised of two types of nodes, corresponding to outfit ID e_o and attribute value e_a of distinct types. For example, "white" and "floral " indicate the attribute values for color and pattern, respectively. The attributes set $\mathcal{A} = \{\mathcal{A}_o, \mathcal{A}_i\}$ is further divided into outfit-level and item-level attribute, corresponding to $\mathcal{A}_o = \{a_o^1, a_o^2, \cdots, a_o^{|\mathcal{A}_o|}\}$ and $\mathcal{A}_i = \{a_i^1, a_i^2, \cdots, a_i^{|\mathcal{A}_i|}\}$, respectively. And the relation set $\mathcal{R}$ is defined in two different forms: 1) outfit_outfit-level attribute. For example, r_1 : outfit_style and r_2 : outfit_popularity; 2) outfit_item-level attribute. For example, r_9 : outfit_color_of_top and r_{11} : outfit_collar_of_top. Accordingly, there are altogether nine types of entity nodes (see Table I) and fourteen types of relations (see Fig. 3) in the fashion knowledge graph $\mathcal{G}$. It is worth mentioning that the entity nodes are composed of outfit ID and eight types of attributes.

For the attribute type, we borrow the fashion domain knowledge from the internal and external dataset. More specifically, the outfit-level attributes are defined based on the meta-data provided by the experimental dataset. While the item-level attributes are extracted with a well-performing clothes attribute classifier with 14 different attribute types and 194 different values, which are pre-trained on a large-scale commercial clothing dataset. And by filtering out noisy attribute types with low classification accuracy and minority labels, five dominant attribute types, including local attributes (*e.g.,* collar sleeve) and global attributes (*e.g.,* color, pattern, category) are estimated. It is worth mentioning that due to the lack

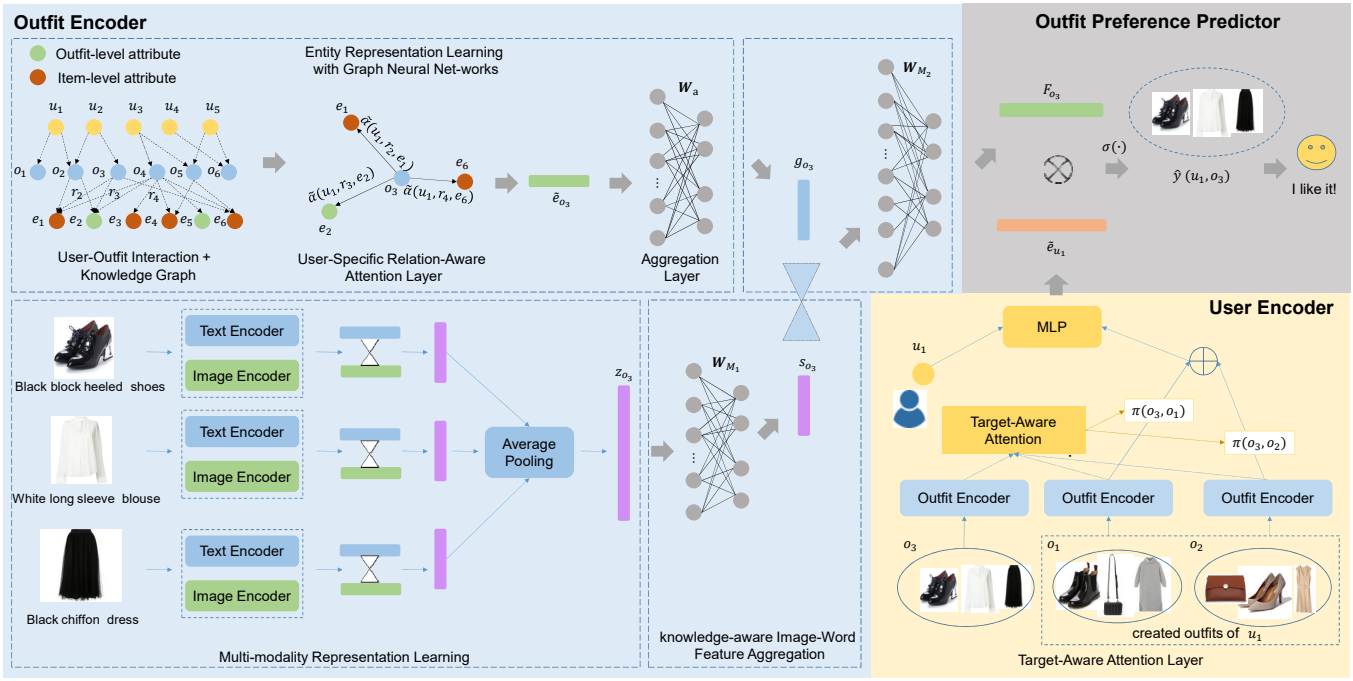


Fig. 2: Overview of the proposed A^3-FKG framework. The blue and yellow branches denote the details of outfit and user encoder, respectively. And the gray branch is the overall preference predictor module. Here hourglass represents the concatenation operation.

TABLE I: Summarization of outfit- and item-level attribute classes as well as the number of values for each class.

Outfit-Level Attribute	popularity (13), price (51), style (9)
Item-Level Attribute	color (20), pattern (27), collar (26), sleeve (8), category (6)

TABLE II: Summarization of notations.

General Symbols	Description	Symbols in Equations	Description
$\mathcal{U}$, $\mathcal{O}$, $\mathcal{S}$	user, outfit, item set	$\mathrm{W}_r, \mathbf{W}_a, \mathbf{b}_a$	relation and aggregation weight matrix, bias
Θ	model parameters	$\mathbf{W}_T$,$\mathbf{W}_V$	textual and visual projection matrices
$\mathcal{Y}$	user-outfit interaction matrix	$\mathbf{W}_{M1}$, $\mathbf{W}_{M2}$, $\mathbf{b}_{M1}$	multi-modal projection matrices
$\mathcal{G}$	attribute-aware fashion knowledge graph	$\mathbf{W}_{2u}, \mathbf{W}_{1u}, \mathbf{W}_{1t}, \mathbf{W}_{1k}$	user-level attention projection matrix
h, t, r, $\mathcal{E}$, $\mathcal{R}$	head, tail entity, relation, entity set, relation set	$\mathbf{s}_{o_j}$, $\mathbf{g}_{o_j}$	multi-modal and structural representation of o_j
$\hat{y}_{uo_j}$	preference score between user u and outfit o_j	$\mathbf{e}_u$, $\mathbf{e}_f$, $\tilde{\mathbf{e}}_u$	global, fine-grained and overall user representation
e_o, e_a	outfit ID, attribute entity	(u, o_j^+, o_j^-)	training triples
$\mathcal{A}$, $\mathcal{A}_o$, $\mathcal{A}_i$	attribute, outfit- and item-level attribute set	$\tilde{\mathbf{e}}_{o_j}$	aggregated local neighborhood representation
$\mathcal{N}_{o_j}^L$, $\mathcal{N}_{o_j}$	sampled and full set of neighborhood of o_j	$\mathbf{F}_{o_j}$	overall representation of o_j
$\mathbf{e}_o, \mathbf{e}_r, \mathbf{e}_t$	entity embedding for head, relation and tail nodes	λ, λ_1, λ_2, γ	trade-off parameters
$\tilde{\alpha}(u, r, t)$	user-specific relation-aware coefficients	$\ell_{kg}, \ell_{bpr}, \ell_{total}$	kg, bpr, and total loss

of the groundtruth attribute annotation, we randomly sample 100 images and manually scrutinize the predicted results. For a given attribute type, if almost half of the sampled set demonstrate the wrong prediction, then we won't incorporate this attribute into the attribute set $\mathcal{A}_i$. To summarize, the detailed attribute types and the number of respective values are shown in Table I.

For the style attribute, it is non-trivial to leverage the off-the-shelf classifier [50] to identify the outfit style. What is more, most of the existing style classifiers are trained on the fashion dataset with full-body human images, which doesn't fit our scenario with clean shop images. To address this issue, we categorize the style of the outfit based on the color statistics of the items within an outfit, such as pairwise element-wise difference vectors, *etc*. Finally, we cluster all the statistical color vectors of the dataset into 13 styles.

C. Outfit Encoder

In this subsection, we aim to learn the outfit encoder, which is further made up of three components: 1) Entity representation learning with graph neural networks (GNNs);

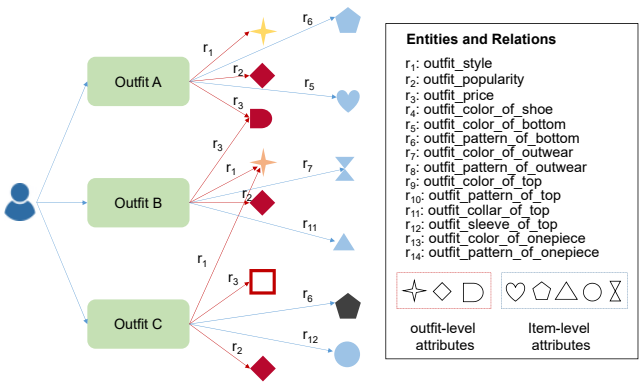


Fig. 3: A toy example of constructed attribute-aware fashion knowledge graph. Different attribute categories are denoted in different shapes and their values are differentiated by colors (*e.g.,* the star symbol connected to outfit A and B via relation r_1 represent different outfit styles.

2) Multi-modality representation learning; and 3) Knowledge-aware image-word feature aggregation.

1) Entity Representation Learning with Graph Neural Networks: For the task of outfit recommendation, the architecture of outfit entity representation component includes two layers: user-specific relation-aware attention layer to accumulate the information propagated from the neighboring entities and aggregation layer to selectively aggregate the representation of the entity itself and that of its neighborhood.

User-specific Relation-Aware Attention Layer. Different users exhibit distinctive personal tastes and are likely to purchase the same outfit with focus on different aspects. For example, user A focuses on the price of the outfit while user B cares more about the popularity, that is, the number of likes in our paper. To model the personalized interests toward multiple perspectives, we develop an attentive user-specific relation-aware attention scheme. Given a target user u, a candidate outfit entity node e_{o_j} and its neighboring set $\mathcal{N}_{o_j}^L$, the local aggregated neighborhood representation $\tilde{\mathbf{e}}_{o_j} \in \mathbb{R}^d$ is formulated as follows:

$$\tilde{\mathbf{e}}_{o_j} = \sum_{(e_{o_j}, r, e_t) \in \mathcal{N}_{o_j}^L} \tilde{\alpha}(u, r, t)\mathbf{e}_t, \tag{1}$$

Note that to keep computation efficient, we randomly select L neighbors among the original neighborhood set $\mathcal{N}_{o_j}$ of entity e_{o_j}, denoted as $\mathcal{N}_{o_j}^L$. The user-specific attention coefficients $\tilde{\alpha}(u, r, t)$ of the neighborhood entity is defined as follows:

$$\tilde{\alpha}(u, r, t) = \frac{\exp\big(\beta\pi(u, r, t)\big)}{\sum_{(e_{o_j}, r', e_t') \in \mathcal{N}_{o_j}^L} \exp(\beta\pi(u, r', t'))}, \tag{2}$$

where $\tilde{\alpha}(u, r, t)$ is the normalized coefficients of $\pi(u, r, t)$ by softmax function across all triples in $\mathcal{N}_{o_j}$ and β is the scaling ratio. And $\pi(u, r, t)$ is defined as follows:

$$\pi(u, r, t) = \mathbf{e}_u^T \tanh\big(\mathbf{W}_r(\mathbf{e}_r || \mathbf{e}_t)\big), \tag{3}$$

where $\mathbf{e}_u \in \mathbb{R}^d$, $\mathbf{e}_r \in \mathbb{R}^d$ and $\mathbf{e}_t \in \mathbb{R}^d$ represent the d-dimension user vector, relation and entity vectors, respectively. And $||$ denotes the concatenation operation between vectors. We select $\tanh$ as the activation function [42]. And $\mathbf{W}_r \in \mathbb{R}^{2d \times d}$ is the trainable transformation matrix to distill informative signal for the subsequent aggregation procedure.

Aggregation Layer. An aggregation layer is designed to aggregate the entity's own feature $\mathbf{e}_{o_j}$ and the propagated feature $\mathbf{e}_{\mathcal{N}_{o_j}^L}$ from its neighborhood. We empirically choose the neighbor aggregator function, which takes the neighborhood aggregated representation $\tilde{\mathbf{e}}_{o_j}$ before applying a nonlinear transformation:

$$\mathbf{g}_{o_j} = \phi(\mathbf{W}_a \tilde{\mathbf{e}}_{o_j} + \mathbf{b}_a), \tag{4}$$

where ϕ is the activation function, empirically set as LeakyRELU [51]. Here we take the neighbor aggregator which is experimentally proven to be powerful in mitigating the overfitting issues and stablizing the training process. $\mathbf{W}_a \in \mathbb{R}^{d \times d}$ and $\mathbf{b}_a \in \mathbb{R}^d$ denote the projection matrix and bias for the non-linear transformation, respectively. After the aggregator function, each entity node can be represented as $\mathbf{g}_{o_j}$.

2) Multi-modality Representation Learning: Given an outfit o_j with a sequence of $|o_j|$ items with the image and textual description, denoted as $S_j = \{(x_{j1}, t_{j1}), \cdots, (x_{jk}, t_{jk}), \cdots, (x_{j|o_j|}, t_{j|o_j|})\}, 1 \le k \le |o_j|$, we first tokenized each item description t_{jk} into separated Japanese words. For example, each word within t_{jk} is denoted as one-hot vector and then fed into the word embedding matrix to obtain its dense vector representation. Then average pooling operation is performed on the sequence of words to generate the textual representation $\mathbf{w}_{jk} \in d^w$. In the similar way, each item image is forwarded into the image embedding matrix to obtain the visual representation $\mathbf{v}_{jk} \in d^v$. Here d^w and d^v indicate the embedding sizes of individual item's textual and visual features, respectively.

To enable the consistency between item's textual and visual representation, we further transform $\mathbf{w}_{jk}$ and $\mathbf{v}_{jk}$ into the joint visual-semantic space through projection matrix $\mathbf{W}_T \in \mathbb{R}^{d_t \times d_w}$ and $\mathbf{W}_V \in \mathbb{R}^{d_t \times d_v}$, expressed as below:

$$\begin{aligned} \tilde{\mathbf{v}}_{jk} &= \mathbf{W}_V \mathbf{v}_{jk}, \\ \tilde{\mathbf{w}}_{jk} &= \mathbf{W}_T \mathbf{w}_{jk}. \end{aligned} \tag{5}$$

Thus, o_j is denoted as $\mathbf{W}_{o_j} = \big[[\tilde{\mathbf{v}}_{j1}, \tilde{\mathbf{w}}_{j1}], \cdots, [\tilde{\mathbf{v}}_{j|o_j|}, \tilde{\mathbf{w}}_{j|o_j|}]\big] \in \mathbb{R}^{2d_t \times |o_j|}$, where d_t is the dimension of projected latent feature. Then the output multi-modal feature for the outfit o_j is to take the average pooling over the concatenated visual and textual representation of all the items, denoted as $\mathbf{z}_{o_j} = \frac{1}{|o_j|} \sum_{k=1}^{|o_j|} (\tilde{\mathbf{v}}_{jk} || \tilde{\mathbf{w}}_{jk})$,

3) Knowledge-aware Image-Word Feature Aggregation: For the given candidate outfit o_j, after deriving the knowledge-aware image-word content features $\mathbf{z}_{o_j}$ through multi-modal representation learning, together with its entity embedding $\mathbf{g}_{o_j}$ introduced in Section IV-C1, a transformation matrix $\mathbf{W}_{M1} \in \mathbb{R}^{2d_t \times d}$ is employed to map $\mathbf{z}_{o_j}$ into the same latent entity space as $\mathbf{g}_{o_j}$,

$$\mathbf{s}_{o_j} = \phi(\mathbf{W}_{M1} \mathbf{z}_{o_j} + \mathbf{b}_{M1}), \tag{6}$$

then we concatenate the transformed feature vector $\mathbf{s}_{o_j} \in \mathbb{R}^d$ with the outfit entity representation $\mathbf{g}_{o_j}$. Again, the trainable transformation matrix $\mathbf{W}_{M2} \in \mathbb{R}^{2d \times d}$, is utilized to project

$\mathbf{s}_{o_j}$ and $\mathbf{g}_{o_j}$ into the user latent space. The final outfit representation $\mathbf{F}_{o_j} \in \mathbb{R}^d$ for computing the preference score is denoted as below:

$$\mathbf{F}_{o_j} = \phi(\mathbf{W}_{M_2}(\mathbf{s}_{o_j}||\mathbf{g}_{o_j}) + \mathbf{b}_{M_2}). \tag{7}$$

D. User Encoder

The user's previous interactions reveal his/her behavior patterns as well as latent interests, based on which effective user embedding can be learnt to facilitate preference modeling. However, the user's diverse interests toward outfits are not one-dimensional. That is, different outfit compositions have varying impacts in characterizing users.

Taking a closer look into the historically clicked outfit set $\{o_1^t, o_2^t, \cdots, o_{|u_t|}^t\}$ of a particular user u_t, they may be composed of multiple styles. Here $|u_t|$ is the number of outfits that the specific user has interacted. If each outfit depicts a specific style, the user preferences can be computed as a linear aggregation of $\sum_{k=1}^{|u_t|} w_{o_k}\mathbf{F}_{o_k}$. It is not practical to assume the attention weights w_{o_k} for clicked outfits are equal. For example, if we are considering about whether to purchase a set of top and bottom in black and white. It is highly likely we will take the outfit into the shopping cart if several items with similar styles are discovered in the previous activities. Therefore, it is essential to design an attentive layer which aims to model user's diverse aesthetic interests.

Target-Aware Attention Layer. Given a target outfit o_j, the interacted set of user u_t has varying impacts on determining the personalized preference towards o_j. The similarity between the target and clicked outfits, modeled as the inner product between two vectors, as shown below:

$$\pi(k,j) = \mathbf{W}_{2u}^T \tanh(\mathbf{W}_{1\mathrm{t}}\mathbf{F}_{\mathrm{o_j}} + \mathbf{W}_{1\mathrm{k}}\mathbf{F}_{\mathrm{o_k}} + \mathbf{W}_{1\mathrm{u}}\mathbf{e}_{\mathrm{u}} + \mathbf{b}_{\mathrm{u}}), \tag{8}$$

where $\mathbf{F}_{o_k} \in \mathbb{R}^d$ and $\mathbf{F}_{o_j} \in \mathbb{R}^d$ are the query and target outfit vectors for outfit o_k and o_j. And $\mathbf{W}_{2u} \in \mathbb{R}^{d/2}$, $\mathbf{W}_{1t}$, $\mathbf{W}_{1k}$ and $\mathbf{W}_{1u} \in \mathbb{R}^{d\times d/2}$ are trainable projection matrices. For better understanding, the term $\pi(k,j)$ can be regarded as the activation value and o_k as the user's interest units. Note that we don't take the traditional softmax operation by scaling the value of $\pi(k,j)$ to the range of [0, 1], which severely suppresses the intensity characterizing the users' interests. Thus the user's fine-grained embedding can be further expressed as below:

$$\mathbf{u}_f = \sum_{k=1}^{|u_t|} \pi(k,j)\mathbf{F}_{o_k}, \tag{9}$$

here the weighted sum of the user's clicked outfits $\mathbf{u}_f$ can be also considered as the fine-grained user embedding which relates with the target outfit. And the user representation is composed of two terms: 1) global user representation, that is, user ID embedding $\mathbf{e}_u$ and 2) local fine-grained user representation denoted by the weighted aggregation $\mathbf{u}_f$. Then we transform them into a joint latent space and obtain a unified user embedding $\tilde{\mathbf{e}}_\mathbf{u} \in \mathbb{R}^{d_m}$ as shown below:

$$\tilde{\mathbf{e}}_\mathbf{u} = (1-\lambda)\mathbf{e}_u + \lambda\phi(\mathbf{W}_f\mathbf{u}_f + \mathbf{b}_f), \tag{10}$$

where $\mathbf{W}_f \in \mathbb{R}^{d_m\times d}$ is the trainable projection matrix, and $\mathbf{b}_f$ is the bias. λ is the trade-off parameter balancing the importance of the global and fine-grained user representation, which is set empirically by maximizing the Pt-AUC score over the validation set.

E. Outfit Preference Prediction and Loss Function

Given a target user u and outfit o_j, the predicted personalized preference score of (u, o_j) is calculated by:

$$\hat{y}_{uo_j} = \sigma(\tilde{\mathbf{e}}_u^\top \mathbf{F}_{o_j}), \tag{11}$$

where $\tilde{\mathbf{e}}_u$ is the user representation and σ denotes the sigmoid function for normalizing the preference score. Note that the negative outfits utilized for training are the ones created by other users rather than random mixture of items [14]. Thus, there is no need to introduce the general compatibility term into the final preference score modeling as in [14], [15].

To optimize the knowledge-aware recommendation model, several losses are incorporated into the overall objective function. Since the task of preference prediction can be regarded as the binary classification problem, in which the cross-entropy loss ℓ_{kg} is adopted, shown as below:

$$\ell_{kg} = -\frac{1}{N}\sum_{u,o_j}(y_{uo_j}\log\hat{y}_{uo_j} + (1-y_{uo_j})\log(1-\hat{y}_{uo_j})), \tag{12}$$

where $\hat{y}_{uo_j}$ and y_{uo_j} denote the predicted personalized preference score and the ground-truth label. BPR loss [52] is also introduced with triplets of one user u and two outfits (o_j^+, o_j^-), in which the user has engaged with o_j^+ and has no interaction with o_j^-. Furthermore, it is assumed that the user demonstrates a larger preference over the observed ones than non-observed ones. Therefore, the BPR loss ℓ_{bpr} is formulated as below:

$$\ell_{bpr} = \sum_{(u,o_j^+,o_j^-)\in\mathcal{T}} -\ln\sigma(\mathbf{u}^\top o_j^+ - \mathbf{u}^\top o_j^-), \tag{13}$$

where $\sigma(\cdot)$ is the sigmoid function. The triplet set is sampled from $\mathcal{T}$, defined as below:

$$\mathcal{T} = \{(u, o_j^+, o_j^-)|y_{uo_j^+} = 1, y_{uo_j^-} = 0\}, \tag{14}$$

The overall objective loss L_{total} to optimize is described as below:

$$L_{total} = \lambda_1\ell_{kg} + \lambda_2\ell_{bpr} + \gamma||\Theta||_2^2, \tag{15}$$

where λ_1 and λ_2 denote trade-off parameters to balance pointwise based loss ℓ_{kg} and pairwise ranking loss ℓ_{bpr}. The last term is the L_2 regularizer loss, which imposes the constraint on the model parameters to prevent overfitting.

V. EXPERIMENTS

In this section, we perform extensive experiments and report the results to shed light on the following questions:

- **RQ1:** How does the proposed A^3-FKG perform compared with other baseline methods w/o or w side information?
- **RQ2:** What is the contribution of each component? How do the integration of the knowledge graph, the contribution of different modality and two-level attention mechanism perform complementary to each other?

- **RQ3:** How do different parameters affect the preference prediction accuracy of A^3-FKG?
- **RQ4:** Does the attribute-aware fashion knowledge graph qualitatively assist in enhancing the recommendation performance?

A. Dataset and Experiment Settings

Dataset. Despite a variety of fashion-oriented datasets [14], [15], [50], [53]–[55] are available for different research tasks, most of them are not well-suited for our task. We conducted experiments on IQON3000 [15], a real-world benchmark dataset with outfit collections created by different users. Most of the outfits in the dataset are forming one-to-one mapping with the users. That is, each outfit is associated with a particular user only. This makes IQON3000 a suitable testbed to evaluate the real-life scenario where there are few interactions between outfits. The raw dataset is composed of $308,747$ outfit compositions created by 3,568 users. Several post-processing procedures are performed on the raw dataset to remove the outfits with merely one item or invalid outfit descriptions. Table III presents the basic statistics of the post-processed dataset, namely IQON3000$_c$.

Implementation Details. For the visual embedding, each item image is fed forward into pre-trained ResNet50 network [56], we take the output of the last average pooling layer and obtain the 2048-dimensional visual feature vector. Then it is further mapped to 512-dimension with one fully-connected (fc) layer. For the textual embedding, we adopt the same strategy as [15], which utilizes the Japanese word2vec *Nwjc2vec* to obtain the textual embedding matrix. The number of tokens in the word embedding and the dimension of each token are set to $54,275$ and 300, respectively. For entity embedding, graph convolutional neural networks is utilized and the dimension d of the entity is set to 64. The size of the user embedding d_m is also set to 64. For the attributes, the outfit-level attribute is extracted from the product descriptions and the item-level attribute is estimated utilizing the fine-tuned VGG19 network trained on a large-scale commercial clothing dataset. The learning rate is set to 0.01. The number of neighboring size L of its neighboring set $\mathcal{N}_{o_j}^L$ is set as 16. The trade-off parameter between different losses are set to $\gamma = 1e-5$, $\lambda_1 = 1.0$, and $\lambda_2 = 1.0$. LeakyReLU is set as the activation function, denoted as ϕ in Section IV and Adam [57] is utilized for parameter updating within the mini-batch samples, the size of which is set as 512. The trade-off parameters are determined through the grid search strategy on the validation set.

Following the typical data splitting protocol [58], the interaction history is partitioned into three parts: training, validation and testing sets. And the ratio is set to 7:1:2. The positive set of each user is composed of the clicked outfits and the negative set is randomly sampled over other user's clicked outfits. The reported performance is the averaged version with the experiment repeated 3 times on the test set. It is worth noting that the algorithm will be terminated automatically when number of training epoch reaches the specified maximum value or evaluation accuracy demonstrates repeated decrease. We consider outfit recommendation as a preference prediction problem and four evaluation metrics are utilized: 1) Area under the ROC curve (Pt-AUC); 2) Average accuracy value (Pt-ACC); 3) F1 value (Pt-F1); and 4) Area under the ROC curve with outfit pairs (positive and negative user-interacted data), denoted as Pr-AUC. It is worth mentioning here that the prefix "Pt" and "Pr" represent "point-wise" and "pairwise", respectively.

B. Compared Baselines

We compared the proposed A^3-FKG with different approaches as baselines. Among them, TransE [26], KGCN [34] and KGNN-LS [33] are KG-aware methods, while MMGCN [59] converts the user-item interaction matrix into the bipartite graph and HFCN [23] proposes a hierarchical graph structure to enable message passing in the user-outfit-item flow. Then the rest are KG-free methods, which can be further divided into two classes according to whether the side information is incorporated or not.

KG-free methods w/o side info

- **Random (RAND)**: The preference score toward the outfit is randomly assigned.
- **PopRank (POP)**: The "like" score is an indicator of popularity, which is utilized to assess the degree of user preference.
- **BPR** [52]: A pairwise ranking approach, which aims to maximize the difference of the posterior probability between the positive and negative outfit pairs.

KG-free methods w/ side info

- **VBPR** [13]: Visual Bayesian Personalized Ranking which incorporates the product's visual features into the BPR framework.
- **TBPR**: This baseline integrates the textual features into the BPR framework instead of visual features.
- **GP-BPR** [15]: State-of-the-art method on the dataset. The prominent difference is that we conduct experiments on the outfit-level, in which each outfit is comprised of several fashion items from a variety of categories, not strictly limited to tops and bottoms. Note that we don't include the general compatibility score into the final preference prediction.

KG-aware methods

- **TransE** [26]: State-of-the-art multi-relational knowledge graph embedding approach which considers the relations as translation on the entity embeddings.
- **KGCN** [34] and **KGNN-LS** [33]: State-of-the-art KG-guided hybrid methods which exploit the high-order connectivity to capture the entity embedding. Sum and concatenation aggregation strategy are utilized respectively to enrich the node's representation with its local neighborhood features.
- **MMGCN** [59]: Recent work on multi-modal graph-based recommendation approach, which proposes to learn the modal-specific user and item representation instead of combining different modality features as whole.
- **HFCN** [23]: Recent work which develops a novel hierarchical fashion graph structure. It enables the information

TABLE III: Basic statistics of $IQON3000_c$ and the constructed variants of knowledge graph structure. ("#" indicates the "the number of")

Basic statistics of $IQON3000_c$				Variants of KG Structure			
				KG-O		KG-I	
# users	3,568	avg. # items per outfit	5.8	# entities	307,834	# entities	1,023,786
# outfits	307,680	avg. # words per item	7.9	# relations	14	# relations	9
# items	715,946	avg. # outfits per user	86	# triples	3,075,337	# triples	6,655,387

propagation from item to the user in three different levels: 1) item-item 2) item-outfit and 3) outfit-user.

C. On Performance Comparison (RQ1)

The comparisons with the baselines and state-of-the-art methods are presented in Table IV. Several conclusions can be drawn as follows:

- The proposed framework A^3-FKG consistently and significantly outperforms other baselines in all cases. It also reveals the significant advantage of incorporating multi-modal side information and two-level attention mechanism component into our system.
- Our superiority over MMGCN and HFCN indicates the importance of attribute-aware knowledge graph as the side information. MMGCN merely leverages the user-outfit interaction to build the bipartite graph, the effectiveness of which is deteriorated when there are few interlinks between user-user and outfit-outfit.
- The state-of-the-art hybrid KG-aware approaches, *i.e.,* KGCN and KGNN-LS, outperformed the PRs, *i.e.,* pairwise ranking methods, by a large margin in personalized preference prediction task. It again serves as a strong evidence that the incorporation of the knowledge-graph is of vital importance in characterizing the user preference over outfits.
- Both the KG-aware methods and latent factor models within the Bayesian Personalized Ranking (BPR) framework have demonstrated competitive capability in the raking-oriented evaluation (*i.e.,* Pr-AUC). It again verifies that both item's side information and structural interlinks in the KG play an important role in pairwise ranking.
- Textual feature (*i.e.,* TBPR) are found to be more superior than utilizing visual features (*i.e.,* VBPR). And both of them significantly improved over merely utilizing the latent factor features (*i.e.,* BPR).

D. Model Ablation (RQ2)

To explore the importance of different components in the proposed system, we compare among the variants of A^3-FKG by replacing the associated modules with alternatives. In the following subsections, we will investigate the effect of two-level attention mechanism, multi-modal side information, the impact of different choices of the aggregation layer and different types of losses.

1) Effect of Two-Level Attention Network: Firstly, we want to validate the impact of personalized attention with respect to the outfit- and user-level, corresponding to user-specific relation-aware (RAA) and target-aware attention (TAA) module, respectively. To independently investigate the importance of attention schemes, free of the impact of multi-modal data, we remove the visual/textual content in the outfit encoding, represented by A^3-FKG (-MM). Table V displays the performance with different combinations of attention modules on both the simplified and complete version of A^3-FKG, from which we can obtain the following findings:

- The two-level attention mechanism consistently improves the performance, which strongly proves the necessity of modeling the users' fine-grained preference on different relations and their diverse interests from their behavior sequences.
- A^3-FKG with the user-specific relation-aware attention consistently outperforms other variants w/o attention networks. One possible reason is that users demonstrate distinctive interests on the different aspects of the outfit. And the user-specific attention mechanism is capable of attending to important outfit traits which satisfy user preferences. Also, it is important to choose the architecture of the attention network. In this case, the inner product between the user embeddings and relations can well capture the user preference as opposed to complex structures like the multi-layer perceptron (MLP).

2) Effect of Multi-Modal Content Features: To mitigate the negative impact of outfits which have few interactions with the users, we introduce the content features of items - 1) Image (I) and 2) Text description (W). Fig. 4 illustrates the results without any side information of items (-I-W), visual-modality only (-W), textual-modality only (-I) and with both modality data. We can find that the textual signal is comparatively more important in personalized preference modeling, which is consistent with the comparison between the VBPR and TBPR as shown in Table. IV. One possible reason is that the textual description is an explicit way to learn the preference matching rules than the visual signal.

3) Effect of the Aggregation Layer: Here after the neighborhood feature aggregation with a different focus on various types of relations, we adopt variants of aggregators to explore their impacts on the framework. More specifically, four different kinds of aggregation schemes are considered and compared, corresponding to 1) Sum 2) Concatenation 3) Neighbor 4) Ego aggregator, denoted as A^3-FKG$_{sum}$, A^3-FKG$_{concat}$, A^3-FKG$_{neighbor}$, and A^3-FKG$_{ego}$, respectively. From the results shown in Fig. 5, we have the observations below:

- A^3-FKG$_{neighbor}$ achieves the best results among the four metrics, which demonstrates that the neighbor feature

TABLE IV: Overall performance comparison between the proposed A^3-FKG and other approaches.

Methods	Model	Pt-AUC	Pt-ACC	Pt-F1	Pr-AUC
KG-free w/o side info	RAND	0.5000	0.4996	0.4996	0.4959
	POP	0.5013	0.5000	0.6203	0.4930
	BPR [52]	0.5660	0.5527	0.5553	0.5625
KG-free w/ side info	VBPR [13]	0.6510	0.5859	0.5638	0.7383
	TBPR	0.7286	0.6494	0.7274	0.7402
	GP-BPR [15]	0.7486	0.6602	0.7085	0.8086
KG-aware	TransE [26]	0.7412	0.6725	0.6879	0.7658
	KGNN-LS [33]	0.7916	0.7205	0.7334	0.7918
	KGCN [34]	0.8132	0.7340	0.7429	0.8144
Graph-based	MMGCN [59]	0.6536	0.6404	0.6688	0.6572
	HFCN [23]	0.8839	0.8113	0.8232	0.8857
Proposed	A^3-FKG	**0.9289**	**0.8575**	**0.8606**	**0.9276**

TABLE V: Performance with different combinations of attention modules. Here ✓ and ✗ denote w/ and w/o attention module, respectively. And A^3-FKG (-MM) and A^3-FKG represent the simplified version without incorporating the multi-modal content and the complete version. RAA and TAA represents the relation-aware attention and target-aware attention, respectively.

Model	Attention Level		Evaluation Metrics			
	RAA	TAA	Pt-AUC	Pt-ACC	Pt-F1	Pr-AUC
A^3-FKG (-MM)	✗	✗	0.8725	0.7991	0.8117	0.8718
	✓	✗	0.8822	0.8055	0.8158	0.8792
	✗	✓	0.8809	0.8047	0.8134	0.8775
	✓	✓	**0.8840**	**0.8086**	**0.8180**	**0.8809**
A^3-FKG	✗	✗	0.9169	0.8443	0.8503	0.9205
	✓	✗	0.9263	0.8542	0.8572	0.9270
	✗	✓	0.9275	0.8559	0.8575	0.9277
	✓	✓	**0.9289**	**0.8575**	**0.8606**	**0.9276**

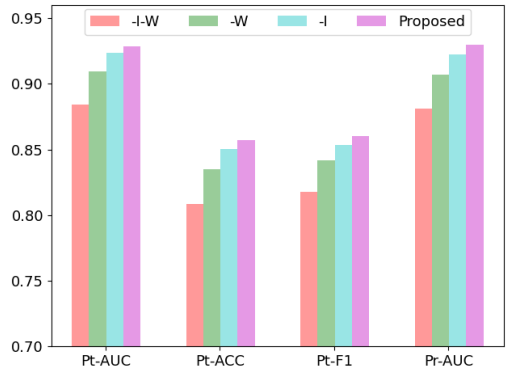


Fig. 4: Effects of multi-modal Content Features. Here (-I) and (-W) mean without the visual and textual information.

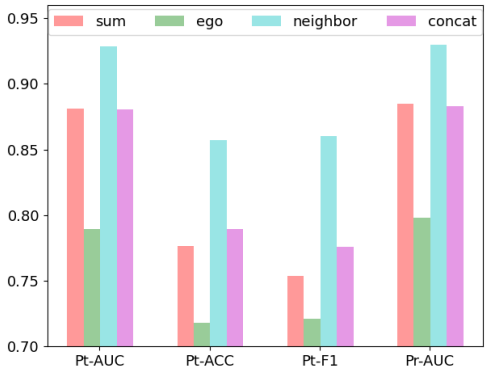


Fig. 5: Effect of variants of aggregators.

aggregation plays a significant role in modeling the user preference. The neighboring attribute-level entity nodes are capable of conveying more semantics than the ego node itself in our case. It is reasonable since our dataset is highly sparse and the representation capability of the entity id (without much collaborative information) is limited.

- The sum and concatenation aggregators are experimentally proven to be more effective than the ego aggregator, which reveals that both the outfit entity and the attribute nodes jointly can enhance the representation power of outfit representation. It also validates the effectiveness of graph neural networks for representation learning.

4) Effect of different losses: Three different types of losses are included in Eq 15), the weighted sum of the cross-entropy loss (ℓ_{kg}), BPR pairwise loss (ℓ_{bpr}) and L_2 regularization loss. In most of the recommender system, either ℓ_{bpr} or ℓ_{lc} is chosen for parameter learning, we propose to combine these two losses, which experimentally found to bring further

TABLE VI: Effect of outfit- and item-level attributes.

Hierarchy	Variants	Pt-AUC	Pt-ACC	Pt-F1	Pr-AUC
Outfit-level	Popularity	0.9231	0.8496	0.8514	0.9230
	Price	0.8664	0.7860	0.7857	0.8651
	Style	0.8684	0.7882	0.7879	0.8660
	Popularity+Price	0.9235	0.8519	0.8543	0.9220
	Popularity+Style	0.9258	0.8497	0.8491	0.9237
	Price+Style	0.8678	0.7871	0.7856	0.8654
	All	0.9267	0.8548	0.8576	0.9251
Item-level	All outfit-level + color	0.9277	0.8548	0.8557	0.9268
	All outfit-level + pattern	0.9286	0.8570	0.8601	0.9299
	All outfit-level + collar	0.9275	0.8547	0.8575	0.9261
	All outfit-level + sleeve	0.9272	0.8561	0.8595	0.9267
	All outfit-level + sleeve + collar + pattern + color	**0.9289**	**0.8575**	**0.8606**	**0.9276**

TABLE VII: Effect of different losses.

Losses	ℓ_{kg} + L_2	ℓ_{bpr} + L_2	L_{total}
Pt-AUC	0.9268	0.8623	**0.9289**
Pt-ACC	0.8555	0.7366	**0.8575**
Pt-F1	0.8587	0.6921	**0.8606**
Pr-AUC	0.9273	0.9165	**0.9276**

improvements and accelerates the convergence of the model, as shown in Table. VII.

E. Attribute-Level Analysis

In this section, we quantitatively explore the impact of both outfit- and attribute-level attributes on the performance of the proposed framework. Since our outfit entity learning is based on the aggregation of its connected neighbors, it is necessary for each outfit node to have at least one attribute. We experiment with the following order: 1) we conduct the experiments on the outfit-level attributes; 2) the item-level attributes are incorporated one by one on top of outfit-level attributes. From the results shown in Table VI, we have the following observations:

- With respect to the outfit-level attribute, popularity consistently outperforms the rest attributes. And most of the pairwise attribute groups have achieved better results than utilizing merely one attribute. By integrating all the attributes, a subtle improvement can be obtained.
- Global attributes such as color and pattern demonstrate more representative capability for describing the user's preference, compared with local fine-grained attributes, such as collar of the top.
- Based on the reported results, the style attribute achieves inferior performance compared with other outfit-level attributes extracted directly from the dataset. However, we still remain the style attribute since it provides an overall impression of the outfit which is experimentally helpful in the qualitative analysis.

Because our target is to construct a comprehensive attribute list to provide the plausible reasons about the preference estimation results. Even though some combinations with minor performance decrease is still acceptable. In this paper, the presented results are based on the attribute all-in strategy.

F. Parameter Sensitivity (RQ3)

In this subsection, the effects of hyper-parameters are investigated on the performance of user-outfit preference prediction.

Impact of attribute-aware fashion KG structure. We modify the structure of established fashion knowledge graph by linking the item's attributes (*i.e.*, outfit- and item-level attributes) to the item entity node rather than the outfit entity node. For example, given an outfit P including a red sweather (item A), for the originally constructed KG-O (utilized throughout the paper), it connects the outfit node "outfit P" to color node "red" via the color-category relation "color of the sweather". While the adapted KG-I, it links the attribute node "red" and category node "sweather" to item node "item A". From Table IX, KG-O achieves better performance partially because of its more comprehensive local structure with an enhanced representation capability. Taking the computation and memory into account, we only consider the one-hop neighbor instead of multi-hop high-order connectivity. Compared to KG-I, the attribute nodes of KG-O are directly connected with the outfit entity node, the aggregation of which thus incorporates more semantics to the outfit embedding. It is worthwhile mentioning that the experiments throughout the paper are conducted on KG-O.

Impact of user embedding dimension We investigate how the dimension of user embedding d_m affects the performance by searching the candidate value in the range of $\{4, 8, 16, 32, 64, 128, 256\}$. The results are shown in Table VIII. From the results, with d_m increasing from 4, the performance gradually increases and it achieves the best performance when d_m reaches 64. It is reasonable as larger embedding size means better representation capability.

Impact of number of neighbor size We vary the number of sampled entity neighbors from 2 to 32 to investigate the importance of sampled neighbors. The results are shown in Table X. When the number of neighbors is small, the outfit entity representation is not comprehensive to aggregate the information from its connected nodes. We find that even with a small neighborhood size as 2, the network still demonstrates

TABLE VIII: Performance in terms of different user embedding dimension d_m.

Metrics \ d_m	4	8	16	32	64	128	256
Pt-AUC	0.8691	0.9004	0.9142	0.9260	**0.9289**	0.9281	0.9214
Pt-ACC	0.7985	0.8308	0.8441	0.8546	**0.8575**	0.8556	0.8473
Pt-F1	0.8049	0.8383	0.8504	0.8584	**0.8606**	0.8574	0.8489
Pr-AUC	0.8684	0.8975	0.9128	0.9249	**0.9276**	0.9275	0.9241

TABLE IX: Performance in terms of different attribute-aware fashion KG structure.

KG structure	KG-I	KG-O
Pt-AUC	0.9165	**0.9289**
Pt-ACC	0.8414	**0.8575**
Pt-F1	0.8424	**0.8606**
Pr-AUC	0.9156	**0.9276**

TABLE X: Performance in terms of number of neighbor size L.

Metrics \ L	2	4	8	16	32
Pt-AUC	0.8801	0.8918	0.9118	**0.9289**	0.9255
Pt-ACC	0.8004	0.8121	0.8337	**0.8575**	0.8509
Pt-F1	0.8038	0.8129	0.8342	**0.8606**	0.8510
Pr-F1	0.8799	0.8928	0.9114	**0.9276**	0.9258

excellence in the prediction performance. Taking a closer look into the dataset statistics, we find that about outfit nodes with no more than 8, 16 neighbors occupy by 29% and 98% of the overall size, respectively. Therefore, setting the neighbor size L as 16 is the best choice to incorporate all types of connected nodes.

G. Case Study (RQ4)

One of the potential application of the proposed system is the explainable personalized recommendation, which offers plausible reasons for the returned results. We aim to explore the fine-grained aesthetic tastes of users as well as their multi-dimensional interest. A user with ID u2407505 is randomly selected and several exemplar outfit compositions are displayed. From the behavior sequence, we can find that from the global aspect, he/she demonstrates preference on the light color (*e.g.*, gray, white) clothes, casual/sporty style. From the local aspect, stripped pattern, hoodies and blue-colored long trousers are to the user's appetite. To validate whether the returned results are in consistency with our observation, the top-3 attentive weights are extracted for demonstration, from Fig. 6, we can find that the top-1 returned outfit is among the user's interacted outfits, which proves the success of our system in the recommendation task. In addition, the top-3 recommended outfits share similar styles with the history record. And the reason for recommendation is also in line with our observation. One interesting finding is that the system is capable of discovering that the user is fond of denim material, as shown in the last two rows.

Fig. 6: Visualization of top-3 recommendation results (last three rows). The first and second rows show examples of the user's historically clicked outfits. The matched one is highlighted with dashed green bounding box. Attribute preference ordering are illustrated in bars of different colors. Best viewed in color.

VI. Conclusion and Future Work

In this paper, we propose an end-to-end personalized outfit recommender system (A^3-FKG), which investigates the usage of knowledge graph in capturing the connectivity between entities (which can be outfit, items, attributes) and exploit the complementary benefits of the multi-modal information. To differentiate the varying contribution of outfit-attribute relations in the knowledge graph as well as the activation of user's diverse interests with respect to the target item, we develop two-level attention modules, corresponding to user-specific relation-aware and target-aware networks. The effectiveness of the proposed system is validated on the real-world dataset. This work represents an initial attempt to integrate the knowledge graph into the recommender system in the fashion domain. In the future, we will attempt to borrow the domain knowledge of clothing matching rules into the construction of fashion-aware knowledge graph.

VII. Acknowledgement

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funds (Project No.A1892b0026).

REFERENCES

[1] G. Charlton. Global fashion ecommerce market. https://www.salecycle.com/blog/featured/online-fashion-retail-11-essential-statistics/.

[2] X. Liu, Y. Sun, Z. Liu, and D. Lin. Learning diverse fashion collocation by neural graph filtering. *arXiv preprint arXiv:2003.04888*, 2020.

[3] M. I. Vasileva, B. A. Plummer, K. Dusad, S. Rajpal, R. Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. *arXiv preprint arXiv:1803.09196*, 2018.

[4] Andreas V., Balazs K., Sean B., Julian M., Kavita B., and Serge B. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *ICCV*, 2015.

[5] X. Yang, Y. Ma, L. Liao, M. Wang, and T.-S Chua. Transnfcm: Translation-based neural fashion compatibility modeling. In *AAAI*, 2019.

[6] X. Han, Z. Wu, Y.-G Jiang, and L.-S Davis. Learning fashion compatibility with bidirectional lstms. In *MM*, 2017.

[7] Y. Li, L. Cao, J. Zhu, and J. Luo. Mining fashion outfit composition using an end-to-end deep learning approach on set data. *TMM*, pages 1–1.

[8] Z. Cui, Z. Li, S. Wu, X.-Y Zhang, and L. Wang. Dressing as a whole: Outfit compatibility learning based on node-wise graph neural networks. In *WWW*, 2019.

[9] R. Tan, M. I Vasileva, K. Saenko, and B. A Plummer. Learning similarity conditions without explicit supervision. In *ICCV*, 2019.

[10] H. Zhan, B. Shi, J. Chen, Q. Zheng, L.-Yu Duan, and A. C Kot. Fashion recommendation on street images. In *ICIP*, 2019.

[11] W.-L Hsiao and K. Grauman. Creating capsule wardrobes from fashion images. In *CVPR*, 2018.

[12] X. Dong, X. Song, F. Feng, P. Jing, X.-S. Xu, and L. Nie. Personalized capsule wardrobe creation with garment and user modeling. In *MM*, 2019.

[13] R. He and J. McAuley. Vbpr: visual bayesian personalized ranking from implicit feedback. In *AAAI*, 2016.

[14] Z. Lu, Y. Hu, Y. Jiang, Y. Chen, and B. Zeng. Learning binary code for personalized fashion recommendation. In *CVPR*, 2019.

[15] X. Song, X. Han, Y. Li, J. Chen, X.-S Xu, and L. Nie. Gp-bpr: Personalized compatibility modeling for clothing matching. In *MM*, 2019.

[16] Y. Lin, M. Moosaei, and H. Yang. Outfitnet: Fashion outfit recommendation with attention-based multiple instance learning. In *WWW*, 2020.

[17] W. Chen, P. Huang, J. Xu, X. Guo, C. Guo, F. Sun, C. Li, A. Pfadler, H. Zhao, and B. Zhao. Pog: Personalized outfit generation for fashion recommendation at alibaba ifashion. In *KDD*, 2019.

[18] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, and H. Zha. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *SIGIR*, 2019.

[19] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel. Image-based recommendations on styles and substitutes. In *SIGIR*, 2015.

[20] Y. Lin, P. Ren, Z. Chen, Z. Ren, J. Ma, and M. de Rijke. Improving outfit recommendation with co-supervision of fashion generation. In *WWW*, 2019.

[21] S. Hidayati, T. Goh, J.-S Chan, C.-C Hsu, J. See, W. Kuan, K.-L Hua, Y. Tsao, and W.-H Cheng. Dress with style: Learning style from joint deep embedding of clothing styles and body shapes. *TMM*, 2020.

[22] S. Hidayati, C.-C Hsu, Y.-T Chang, K.-L Hua, J. Fu, and W.-H Cheng. What dress fits me best? fashion recommendation on the clothing style for personal body shape. In *MM*, 2018.

[23] X. Li, X. Wang, X. He, L. Chen, J. Xiao, and T.-S Chua. Hierarchical fashion graph network for personalized outfit recommendation. *arXiv preprint arXiv:2005.12566*, 2020.

[24] F. Zhang, N. J. Yuan, D. Lian, X. Xie, and W.-Y Ma. Collaborative knowledge base embedding for recommender systems. In *KDD*, 2016.

[25] H. Wang, F. Zhang, M. Hou, X. Xie, M. Guo, and Q. Liu. SHINE: signed heterogeneous information network embedding for sentiment link prediction. In *ICDM*, 2018.

[26] A. Bordes, N. Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, 2013.

[27] Z. Wang, Jianwen Zhang, J. Feng, and Z. Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, 2014.

[28] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, 2015.

[29] S. Xiong, W. Huang, and P. Duan. Knowledge graph embedding via relation paths and dynamic mapping matrix. In *Advances in Conceptual Modeling*, volume 11158, pages 106–118, 2018.

[30] C. Shi, J. Liu, F. Zhuang, P. S. Yu, and B. Wu. Integrating heterogeneous information via flexible regularization framework for recommendation. *Knowl. Inf. Syst.*, 49(3):835–859, 2016.

[31] Y. Wang, Y. Xia, S. Tang, F. Wu, and Y. Zhuang. Flickr group recommendation with auxiliary information in heterogeneous information networks. *Multimedia Syst.*, 23(6):703–712, 2017.

[32] C. Qin, H. Zhu, F. Zhuang, Q. Guo, Q. Zhang, L. Zhang, C. Wang, E. Chen, and H. Xiong. A survey on knowledge graph based recommender systems. *SCIENTIA SINICA Informationis*.

[33] H. Wang, F. Zhang, M. Zhang, J. Leskovec, M. Zhao, W. Li, and Z. Wang. Knowledge-aware graph neural networks with label smoothness regularization for recommender systems. In *KDD*, 2019.

[34] H. Wang, M. Zhao, X. Xie, W. Li, and M. Guo. Knowledge graph convolutional networks for recommender systems. In *WWW*, 2019.

[35] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

[36] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[37] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. In *ICML*, 2019.

[38] Gaël Letarte, Frédérik Paradis, Philippe Giguère, and François Laviolette. Importance of self-attention for sentiment analysis. In *EMNLP*, 2018.

[39] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, and T.-S. Chua. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *SIGIR*, 2017.

[40] X Gao, F. Feng, X. He, H. Huang, X. Guan, C. Feng, Z. Ming, and T.-S Chua. Hierarchical attention network for visually-aware food recommendation. *TMM*, 22(6):1647–1659, 2019.

[41] Thomas N K. and Max W. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[42] X. Wang, X. He, Y. Cao, M. Liu, and T.-S. Chua. Kgat: Knowledge graph attention network for recommendation. In *KDD*, 2019.

[43] Michael S., Thomas N K., Peter B., Rianne V. D. B., Ivan T., and Max W. Modeling relational data with graph convolutional networks. In *ESWC*, 2018.

[44] David Goldberg, David Nichols, Brian M Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12).

[45] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

[46] D. Liang, J. Altosaar, L. Charlin, and D. M. Blei. Factorization meets the item embedding: Regularizing matrix factorization with item co-occurrence. 2016.

[47] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. In *ICDM*, 2011.

[48] J. Li, M. Jing, K. Lu, L. Zhu, Y. Yang, and Z. Huang. From zero-shot learning to cold-start recommendation. In *AAAI*, 2019.

[49] S. Liu, I. Ounis, C. Macdonald, and Z. Meng. A heterogeneous graph neural model for cold-start recommendation. In *SIGIR*, 2020.

[50] Edgar S.-S. and Hiroshi I. Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *CVPR*, 2016.

[51] A. Maas, A. Hannun, and A. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013.

[52] Steffen R., Christoph F., Zeno G., and Lars S.-T. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

[53] H. Zhan, B. Shi, and A. C. Kot. Cross-domain shoe retrieval with a semantic hierarchy of attribute classification network. *TIP*, 26(12):5867–5881, 2017.

[54] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.

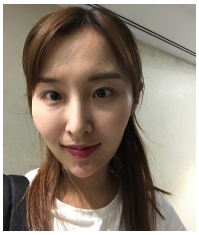
[55] H. Zhan, C. Yi, B. Shi, L.-Y. Duan, and A. C. Kot. Pose-normalized and appearance-preserved street-to-shop clothing image generation and feature learning. *TMM*, 2020.

[56] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[57] Diederik P. K. and Jimmy B. Adam: A method for stochastic optimization. 2015.

[58] H. Wang, F. Zhang, J. Wang, M. Zhao, W. Li, X. Xie, and M. Guo. Ripplenet: Propagating user preferences on the knowledge graph for recommender systems. In *CIKM*, 2018.

[59] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *MM*, 2019.



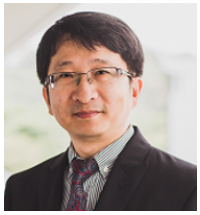
Huijing Zhan is a research scientist with the Institute of Infocomm Research, A*STAR, Singapore. She received the BE degree from the Special Class for the Gifted Young, Huazhong University of Science and Technology and the PhD degree from Nanyang Technological University, in 2012 and 2018. Her research interests include personalized fashion recommendation and retrieval.

Jie Lin is a research scientist with the Institute of Infocomm Research, A*STAR, Singapore. He received his B.S and Ph.D. from the School of Computer Science and Technology, Beijing Jiaotong University. His research interests include deep learning, AI hardware, computer vision, data compression and privacy preserving machine learning. His previous work on image feature coding has been recognized as core contributions to the MPEG-7 Compact Descriptors for Visual Search (CDVS) and Compact Descriptors for Video Analysis (CDVA) standard. His current work is hardware-software co-optimization for deep learning, the key to enable next-generation AI hardware for a wide range of applications at the edge.

Kenan Emir Ak received the Ph.D. degree in electrical and computer engineering from the National University of Singapore, in 2020. He is a Research Scientist at The Visual Intelligence lab of the Institute for Infocomm Research, A*STAR, Singapore. Prior to joining A*STAR, he was at the Adobe Deep Learning research team in California. His research interests include machine learning, computer vision, and image processing.

Boxin Shi (M'14) received the BE degree from the Beijing University of Posts and Telecommunications, the ME degree from Peking University, and the PhD degree from the University of Tokyo, in 2007, 2010, and 2013. He is currently a Boya Young Fellow Assistant Professor and Research Professor at Peking University, where he leads the Camera Intelligence Group. Before joining PKU, he did postdoctoral research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University from 2013 to 2016, and worked as a researcher in the National Institute of Advanced Industrial Science and Technology from 2016 to 2017. He won the Best Paper Runner-up award at International Conference on Computational Photography 2015. He has served as an editorial board member of IJCV and an area chair of CVPR.

Ling-Yu Duan (M'06) is a Full Professor with the National Engineering Laboratory of Video Technology (NELVT), School of Electronics Engineering and Computer Science, Peking University (PKU), China, and has served as the Associate Director of the Rapid-Rich Object Search Laboratory (ROSE), a joint lab between Nanyang Technological University (NTU), Singapore, and Peking University (PKU), China since 2012. He is also with Peng Cheng Laboratory, Shenzhen, China, since 2019. He received the Ph.D. degree in information technology from The University of Newcastle, Callaghan, Australia, in 2008. His research interests include multimedia indexing, search, and retrieval, mobile visual search, visual feature coding, and video analytics, etc. He was a Co-Editor of MPEG Compact Descriptor for Visual Search (CDVS) Standard (ISO/IEC 15938-13), and is serving as a Co-Chair of MPEG Compact Descriptor for Video Analytics (CDVA). Currently he is an Associate Editor of ACM Transactions on Intelligent Systems and Technology (ACM TIST) and ACM Transactions on Multimedia Computing, Communications, and Applications (ACM TOMM), and serves as the area chairs of ACM MM and IEEE ICME. He is a member of the MSA Technical Committee in IEEE-CAS Society.



Alex C. Kot (S'85-M'89-SM'98-F'06) has been with the Nanyang Technological University, Singapore since 1991. He was Head of the Division of Information Engineering and Vice Dean Research at the School of Electrical and Electronic Engineering. Subsequently, he served as Associate Dean for College of Engineering for eight years. He is currently Professor and Director of Rapid-Rich Object SEarch (ROSE) Lab and NTU-PKU Joint Research Institute. He has published extensively in the areas of signal processing, biometrics, image forensics and security, and computer vision and machine learning.

Dr. Kot served as Associate Editor for more than ten journals, mostly for IEEE transactions. He has served the IEEE SP Society in various capacities such as the General Co-Chair for the 2004 IEEE International Conference on Image Processing and the Vice-President for the IEEE Signal Processing Society. He received the Best Teacher of the Year Award and is a co-author for several Best Paper Awards including ICPR, IEEE WIFS and IWDW, CVPR Precognition Workshop and VCIP. He was elected as the IEEE Distinguished Lecturer for the Signal Processing Society and the Circuits and Systems Society. He is a Fellow of IES, a Fellow of IEEE, and a Fellow of Academy of Engineering, Singapore.