

Stylistic Compatibility Learning with Deep Neural Networks for Indoor Scene

Siqi Wang
siqi_wang2@brown.edu



Figure 1: A Typical Indoor Scene with Product Style Tags.

ABSTRACT

Learning the stylistic compatibility between the object and the scene is a new area in indoor scene analysis. In this work, we constructed a dataset for the problem and proposed a deep neural network with conditioning method to learn the style. We show that the network is not only effective in assessing the stylistic compatibility between the single object and a set of objects, but could also provide reasonable recommendations of the objects to add to the existing scene.

KEYWORDS

stylistic compatibility, indoor scene, deep neural network, conditioning method

1 INTRODUCTION

With the fast development of e-commerce, there is an increased demand for well-designed scenes, especially for companies that sell furniture and home decorations online like Wayfair. Displaying synthetic, photo-realistic 3D scenes is the best advertising method to attract customers, however, the interior design process takes great effort from highly skilled designers and is very time-consuming and expensive. Achieving aesthetically pleasing, consistent style is one

of the biggest challenges in the design process: aesthetic criteria are subjective and the elements of style are difficult to describe explicitly with precision. The big gap between the designers' limited productivity and the significant demand for scenes makes it necessary to develop models that could assist in the scene generation process. Also, for people without any interior style knowledge, the style compatibility learning model allows them to design their own rooms. However, most prior work in 3D indoor scene synthesis[17], focuses purely on the position and category (for example, sofas, desks, etc.) of the object. As for style learning, there is also related research in stylistic compatibility in outfit fashion[5], but limited to stylistic relation between two individual items.

In this research project, we develop the first model that can learn the stylistic compatibility between individual objects and scenes. Given the object's category, our model can give suggestions of products that are most stylistically compatible with the scene in real-time. In order for the model to be an efficient tool for assisting scene generation, the input for the scene is not a real-time rendering, which is time-consuming and computational-expensive. Instead, the scene refers to a set of objects (henceforth, referred to as context). The raw data include the 3D professionally-designed scenes from Wayfair, product information for every individual product (henceforth, referred to as SKU) that appears in the scene, such as the SKU iconic images and category labels, and the scene information such as scene style labels. The SKU category labels and

scene style labels are not directly used in the model training but are utilized for constructing the dataset.

From the raw data, context SKUs - query SKU pairs are designed for training the model. A set of SKUs represents the context, and the SKU in the context are referred to as context-SKU; one SKU from the same scene as the context is the “positive example”, which makes a context-positive SKU pair; SKU from a different scene is the “negative example”, which makes a context-negative SKU pair. This positive/negative example SKU is referred to as query SKU. The model learns stylistic compatibility by training a logistic regression model from these pairs. For the positive pairs, the regression model outputs 1 and 0 for the negative ones. The output of the regression model shows the probability of putting this query SKU in the context to make a stylistic consistent scene.

There are two sub-networks in the model: context-SKU CNN and query-SKU CNN. the context-SKU CNN is for processing the image of every single SKU in the context. The one-hot encoding of the query SKU category is used for generating conditional layer parameters for this CNN. query-SKU CNN is for processing the images of the query SKU in the pair. The aggregation of the output from the previous context-SKU CNN, which is the context embedding, is used for generating conditional layer parameters for query-SKU CNN. The regression model is optimized via binary cross-entropy loss.

As for the evaluation, there are two tasks for the model: 1. Predicting the stylistic compatibility between a single query SKU and the context. 2. Recommending the top K stylistically compatible SKUs: Given context SKUs and the category of the query SKU, our model must provide K stylistically compatible SKU from the given category. We assume that the SKU that appears in the same scene as the context is a “ground truth”, and we use this SKU to evaluate the K suggestions given by our model.

The contributions of this project are the following:

- To our best of knowledge, this is the first model learning the stylistic compatibility between objects and scenes in the scene understanding domain.
- A dataset of stylistic compatible objects is constructed and ready for use.
- Demonstrated application value of our model for autonomous 3D scene generation and human-in-the-loop design tools.

2 RELATED WORK

Our work is studying the stylistic compatibility within a scene completion task. We reviewed related work on scene completion, component suggestion & fashionability, and conditioning methods in neural network.

Scene completion Given a partial scene and a query position, scene completion models could generate the possibility for different categories of objects to be put in that location. Previous work could be divided into three groups: rule-based methods, data-driven statistical models, and graph neural networks. The rule-based method follows the guidelines for furniture layout as terms in density functions[7]. These kinds of methods take care of both functional and visual criteria, however, they have the assumption of pre-specified models and have difficulty integrating interaction

inputs; generally speaking, rule-based design approaches are brittle, require a large amount of domain knowledge, and generate output that lacks diversity. Statistical models apply a Bayesian network with object co-occurrence data. Many models based upon co-occurrence have been shown to have limited creativity and generalization ability ; these models tend to only work on small scale scene generation given large amounts of clean relational data. Neural networks, especially graph neural networks[11][3], have been the most popular method in recent years. The scene is parsed as a tree-structure graph[18], where each node represents an object and each edge represents spatial and semantic relationships. However, all of the methods above focus only on the category-level of the object in the scene completion and include no explicit information concerning style.

Component suggestion and fashionability Existing work for learning compatibility could be divided into two categories. The first category concerns functional compatibility such as component suggestions in 3D modeling assembly. The second category concerns fashionability, such as studying the stylistic compatibility in outfits[12][16]. Sung et al[13] predicted the probability distribution of components to recommend by using a mixture density network (MDN). Given a triplet of the partial assembly X, positive sample Y and negative sample Z, their work first maps samples to an embedding space and then generates a Gaussian mixture distribution on the space given X. The novel idea in this work is that the compatibility should be represented as a continuous probability distribution rather than discrete values. Vasileva et al[15] did some work related to fashion compatibility. Their type-aware embeddings could learn the compatibility between specific types of clothes, such as the fashionability between a pair of pants and a pair of shoes. However, their method could not be directly applied to our problem for two key reasons. First, fashion compatibility in their model was defined to be pair-wise. Their model could answer whether two pieces of clothing match, but not whether an additional piece of clothing would compliment an existing outfit. Second, there are hundreds of object categories for furniture in indoor scenes (an order of magnitude greater than the number of object categories in clothing). This makes it infeasible to have category aware pair-wise compatibility for each of the existing categories. In order to accomplish this, we would need to construct an embedding space for each individual category and then map the joint relationships between each of these embedding spaces, which is computationally expensive.

Conditioning methods in neural network Attentions are added in the network with conditioning methods of some conditioning information. There are two popular conditioning methods: the first one is concatenating the conditioning information in the input or simply adding the conditional bias[4][2][6]; Conditional DCGANs[10], WaveNet[8], and Conditional PixelCNN[14] are the classical networks of this method. the second one is influencing the network with a learned affine transformation from the conditioning information. In FiLM[9], a Feature-wise Linear Modulation (FiLM) layer applies an affine transformation on the network’s intermediate features. These conditioning methods are effective for the visual question answering (VQA)[1] tasks. In our project, conditional methods are applied in both CNN to strengthen the relationship between the context SKUs and the query SKUs.

3 METHOD

3.1 Overview

Our model is preferment across two primary tasks: Given the context and the query SKU, our model could predict the probability of adding this SKU to make the scene stylistic consistent. Also, given the context and a query object category, our model could propose objects that look stylistically compatible in the scene. To achieve this, there are two essential networks in the model: a context-SKU CNN and a query-SKU CNN. The context-SKU CNN learns the feature of each SKU in the context with the conditioning information of the query SKU's category. Then, the context embedding is generated by aggregating all the single context-SKU embedding from the context-SKU CNN output. The context-SKU CNN and the query-SKU CNN are linked with a conditioning method that applies the context embedding as the conditioning information to the query-SKU CNN. Finally, the query-SKU CNN outputs the probability of adding the query SKU to the context.

One of the biggest challenges is designing a dataset. Since our research topic is novel, there is no reference and existing open-source dataset for our problem. Due to our partnership with Wayfair, we received two thousand professionally-designed 3D scenes, which are annotated with style labels. For all the objects in the scene, we also received valuable SKU information including front-perspective images and furniture categories. From the raw 3D data, we first get the list of objects in the scene and match the related SKU information with the object and the scene. Secondly, we generate context SKUs - positive query SKU- negative query SKU triplets, which are used to generate context SKUs - query SKU pairs for training. The subset of objects in the scene is defined as context, which serves as the anchor in the triplet. A positive sample is a single object selected from the same scene as the context (henceforth, referred to as P) and a negative sample (henceforth, referred to as N) is a single object in the same category as a positive example and selected from a different scene. The subjective property of stylistic compatibility makes it difficult to construct a negative query SKU. The negative query SKU could probably be a good fit for the scene but was missed by the designer. Since the number of distinct objects is very limited in all the scenes, we perform various data augmentation to expand the training dataset. More details about the dataset are described in section 3.2 data processing.

3.2 Data processing

Data source Stylistic compatibility is a less invested area and there is no available dataset for this project. To train the network efficiently, the training data should be professionally stylistic compatible. To satisfy this, we select raw data of the 3D scenes designed by Wayfair designers. We assume that every scene is consistent in style and every object in the scene is stylistic compatible with any other objects or objects set. There are 2461 scenes of 27 styles. For each scene, there is a style label and the main SKU, which is the oriented object that the scene is made for. The scenes are in 3Ds Max file format and every object is a node in the scene structure. We parse the scene to get the list of SKUs with max scripts and match SKUs with category information and front images from the Wayfair database. To train the model, each SKU is represented as

an RGB image. 7751 distinct objects of 305 categories are collected from all the scenes.

Challenges and solutions Data is the key to the networks and constructing good datasets is one of the most difficult parts of this project. Firstly, the raw data is noisy. There are object nodes in the scene structure that are not shown in the scene rendering; the names of the node may not match the SKU in the database; A single node may be a grouping of several objects. Secondly, the style distribution of the scene and the category distribution of the SKUs are unbalanced. The most popular SKU categories are pillows and rugs, which are general objects to different styles and could be hard examples to learn styles. There are also very minor categories like 'Flatware Single Pieces' and 'Hooks', which contribute little to the style of the scene and do not represent any details. To eliminate the influences from those minor categories, we manually selected 17 categories' objects as P/N and 194 categories objects to be the main SKU of the scene. (we only restrict the category of the main SKU in the scene so that at least a SKU in the scene could have some style information. It is possible that objects of other categories are included in the context.) There are also a lot of concerns when generating the triplets. As the quantity, more triplets are expected from the limited number of scenes. To address this, subsets of the scene objects are selected as the context, which means nearly $\frac{n^2}{2}$ of triplets could be generated from the scene having n SKUs; As for the quality, contexts from the same scene should not appear at the training and testing dataset at the same time. So we split training and testing datasets at the scene level; The P for the context sometimes could be very subjective as well, so we introduce simple N and hard N. Simple N is the object that appears in a different style of scenes with context and hard N is the object appears in the same style of scenes with context.

Data processing pipeline For the balance of the positive and negative pairs, triplets of context-positive SKU- negative SKU are generated first. The data processing pipeline starts with splitting scenes for training and testing, which guarantees that no same context appears in both training and testing datasets. To construct the triplet, there are two key parameters: The category of the query SKU and the number of SKUs in the context. A scene is selected with the filters of those two parameters, and the context and positive-query SKU are generated from this scene. Then the negative query SKU of the same category is selected from other scenes. Data augmentation is added to the data generator during the training time. The augmentations include rotation, width/height shift, changing brightness, zoom-in/out, vertical/horizontal flip. All those augmentations are robust to the style information in the image and are helpful to the generalization of the model.

3.3 Model

The stylistic compatibility regression model consists of two essential networks: context-SKU CNN and query-SKU CNN. The input for both networks is the RGB image of the single SKU. The architectures of these two networks are similar, but they have different conditioning information and are designed for different purposes.

The conditioning information for context-SKU CNN is the category of the query-SKU. It is expected that the features extracted from the context SKU images are relative to the query SKU category.

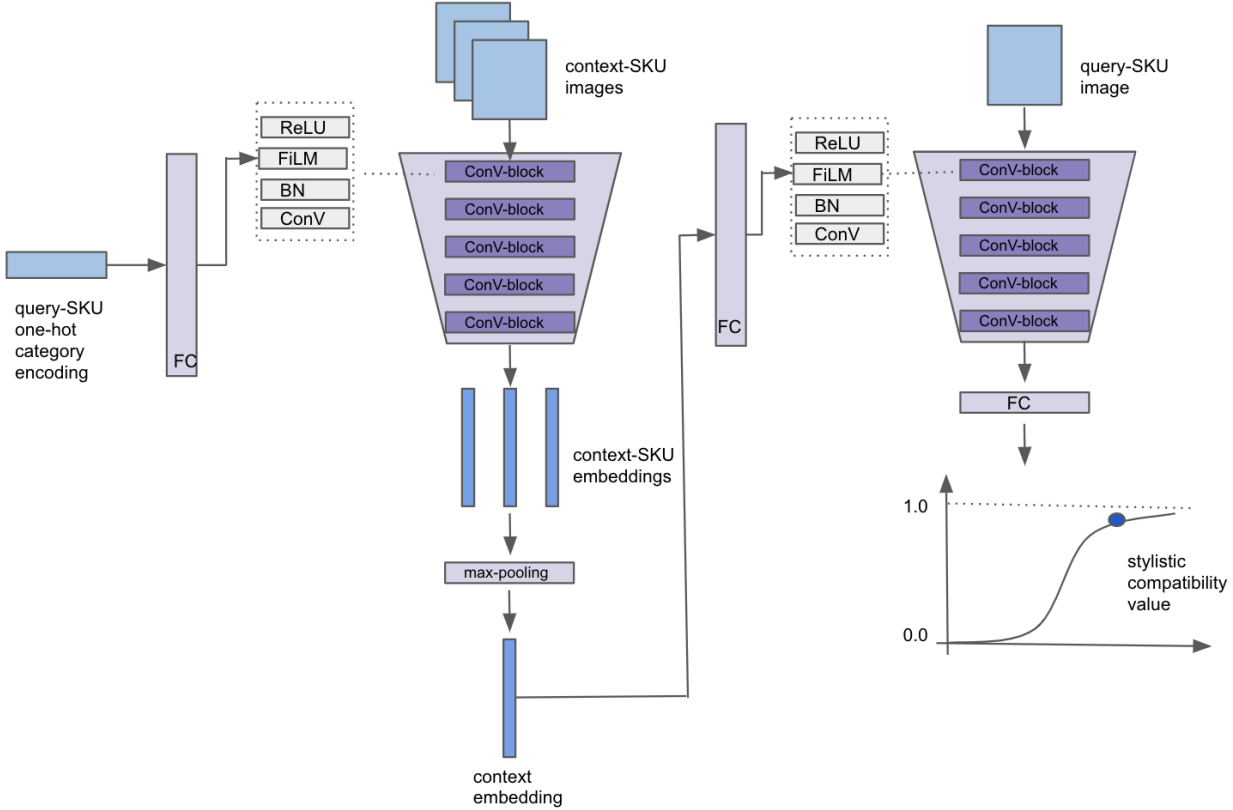


Figure 2: The architecture of the stylistic compatibility regression model

For example, if the query SKU is an armchair, the context-SKU CNN may raise more attention to the color of the SKUs in the context; if the query SKU is an end table, shape and material features may be more important to the network. The category of the query-SKU is represented as an one-hot encoding and it generates parameters of affine transformations for context-SKU CNN’s intermediate features. This Feature-wise linear modulation is added after convolutional layers in the network as the Figure 2 shown.

The embedding for each context SKU is generated from the context-SKU CNN. To generate a single embedding to represent the context, all the context SKU embeddings need aggregating. Since the context embedding should be irrelevant to the appearing order of SKUs in the context, global max-pooling is finally selected as the aggregation method.

The context embedding from the global max-pooling layer is the conditioning information for the query-SKU CNN. query-SKU CNN is a regression model, which output the stylistic compatibility between the query SKU and the context. The compatibility is learned from the conditioning context embedding information and the query SKU features. Similar to the conditioning method of the context-SKU CNN, Feature-wise linear modulation is also added after every convolutional layer in query-SKU CNN.

4 EXPERIMENT

We tested our model on two tasks. The first one is given the context and the query SKU and then output of the probability of adding this SKU to make the scene stylistic consistent. The second task is recommending top-K SKUs of the given category to the context.

In all the experiments below, the number of SKUs in the context is fixed as three. Also, the negative query SKUs are selected from the different styles from the positive query SKU, which are “simple negatives”.

Figure 3, 4, 5 show the result of the first task. In all the successful examples, the positive query SKUs show the stylistic consistent features with the context SKUs. Such as the color consistency in row 2 in Figure 3. While the negative query SKUs are not compatible to the scene because of color (row4) or materials (row 5). For the false positive examples shown in Figure 4, some of predictions look reasonable (row 2). That is one of biggest challenges in labelling the “true negative”. The negative sample could also be stylistic consistent with the context SKUs even though no such match has appeared before. For the false negative examples shown in Figure 5, one possible reason is the lack of style information in the context SKUs. In row 2, row 3, all the context SKUs are decorations and the model is not confident in all the predictions.


	Context(test dataset)	SKU	Label	Prediction
0			1.0	0.997
1			1.0	0.998
2			1.0	0.999
3			0.0	0.003
4			0.0	0.001
5			0.0	0.001

Figure 3: The quantitatively results of evaluating the compatibility of a single SKU to the context (succesful examples)

	Context(test dataset)	SKU	Label	Prediction
0			0.0	1.000
1			0.0	1.000
2			0.0	1.000
3			0.0	1.000
4			0.0	1.000
5			0.0	1.000

Figure 4: The quantitatively results of evaluating the compatibility of a single SKU to the context (False positives)

Figure 6 show the result for the second task. Images on the left most is the scene rendering for the context. next three rows show the context SKUs, the top 5 recommendations and their respective prediction confidence values. GT on the right represents the ground truth SKU, which is the SKU that originally appears in the scene, and it's prediction value. From the results, we could find that GT

SKU always has high probability and some of the them even appear in the recommendations (row 3, row 4, row 5, row 6). The existing problem is that the prediction values for the top K recommendations might be very close, and top K are probably not representative for some scenes.

	Context(test dataset)	SKU	Label	Prediction
0			1.0	0.331
1			1.0	0.337
2			1.0	0.428
3			1.0	0.431
4			1.0	0.465
5			1.0	0.487

Figure 5: The quantitatively results of evaluating the compatibility of a single SKU to the context (False negatives)

5 CONCLUSION AND FUTURE WORK

This project is a pioneer work in studying the interior scene design. A regression model with feature-wise linear modulation layers is proposed to learn the stylistic compatibility between the scene context and the SKU. The dataset constructed in this project is also very important resource to other studies in this stylistic learning field.

Without a doubt, we are still at the very initial stage of this problem and more could be explored in future works.

5.1 Future work

5.1.1 Dataset. Have higher standard for the raw data The quality of the 3D scene data is essential to the following parsing and dataset construction. Considering the current challenges with the dataset, here are some suggestions for the improvements of the data quality: The SKU should have meaningful node names in 3D scene files; No discarded SKUs in the scene, which are the SKUs that are not stylistic compatible with other SKUs in the scene; each SKU should have a single node and no wrong groupings for several SKUs together; The bounding boxes of the SKUs should be able to represent the size of the object.

Add other information to the context Only the node name is used from the 3D scenes right now. Actually, there is more information could be utilized. Such as the positions in the scene: the camera position, the SKU positions in the camera space. When selecting the SKUs from the scene to construct the context, the supporting relationships should also be considered. For example, a pillow should not be included in the context if the supporting sofa is not included.

Balance the distribution of the SKU category and style

One of the biggest challenges for the dataset now is the unbalance of the SKU category and styles. The most common categories are rugs and pillows now, which do not carry too much style information and could be compatible with many different styles.

Overcome the subjective in positive/negative sample evaluation Due to the subjective property in evaluating the style, the negative query SKU examples may not be “true” negatives. In future work, other methods could be proposed for selecting negative examples. Such as selecting the SKU that is most visually dissimilar to the positive as the negative; the SKU with the largest embedding distance to the positive in some visual embedding space.

5.1.2 Modifications of network architecture and improvements in training. Support more SKUs in the context The only successful experiment now is restricted to only three SKUs in the context. More SKUs in the context are tested, however, the performance shows serious overfitting. One possible training improvement is to start training the model from fewer context SKUs and adding the SKUs iteratively. Another attempt could be optimizing the context SKU embedding aggregation methods. Right now, there are no trainable parameters in the max-pooling and all the learning ability for the context is in the first context-SKU CNN. When the number of SKU increasing, this CNN may tend to “memorize” more rather than learning.

Add additional loss for constructing better SKU embedding space In the ideal SKU embedding space, similar SKUs should be clustered together, especially for the same SKU with different augmentations. The current model has no loss to constrain the distance of similar SKUs.

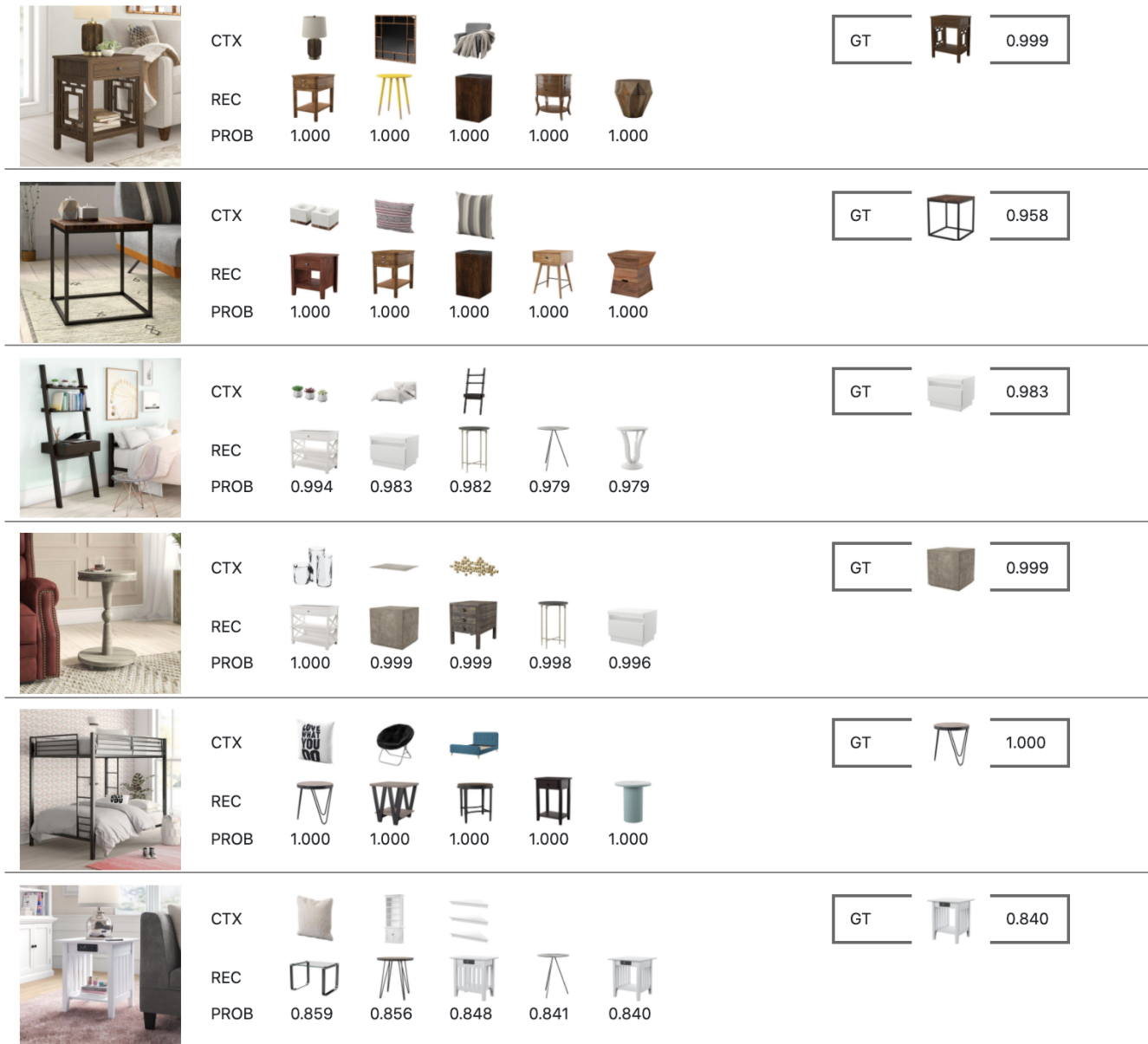


Figure 6: The quantitatively results of recommending 5 SKUs to the context

5.1.3 Context embedding and SKU embedding exploration. It is interesting to see the visualization of the embedding space. How the SKU embedding space will change conditioning on different contexts.

5.1.4 Human preference evaluation. One of the biggest challenges of the evaluation is no “ground truth”. The quantitative and qualitative analysis may not be accurate based on the current “ground truth”. Let professional designers and common users provide more reliable “ground truth”. Their feedback could also be useful for future improvements.

ACKNOWLEDGMENTS

I would like to express my great appreciation to Wayfair, for the valuable scene data access and warm assistance. To professor Daniel Ritchie, manager Tim Zhang, for their thoughtful and inspiring instructions through the whole project. To my research partner Naveen Srinivasan, for all the support.

REFERENCES

- [1] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. 2017. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*. 6594–6604.

- [2] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1243–1252.
- [3] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584* (2017).
- [4] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [5] Wei-Lin Hsiao, Isay Katsman, Chao-Yuan Wu, Devi Parikh, and Kristen Grauman. 2019. Fashion++: Minimal edits for outfit improvement. In *Proceedings of the IEEE International Conference on Computer Vision*. 5047–5056.
- [6] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.
- [7] Paul Merrell, Eric Schkufza, Zeyang Li, Maneesh Agrawala, and Vladlen Koltun. 2011. Interactive furniture layout using interior design guidelines. *ACM transactions on graphics (TOG)* 30, 4 (2011), 1–10.
- [8] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).
- [9] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- [10] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).
- [11] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE Transactions on Neural Networks* 20, 1 (2008), 61–80.
- [12] Xuemeng Song, Fuli Feng, Jinhuan Liu, Zekun Li, Liqiang Nie, and Jun Ma. 2017. Neurostylist: Neural compatibility modeling for clothing matching. In *Proceedings of the 25th ACM international conference on Multimedia*. 753–761.
- [13] Minhyuk Sung, Hao Su, Vladimir G Kim, Siddhartha Chaudhuri, and Leonidas Guibas. 2017. Complementme: weakly-supervised component suggestions for 3D modeling. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 1–12.
- [14] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. 2016. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*. 4790–4798.
- [15] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. 2018. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 390–405.
- [16] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. 2015. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Proceedings of the IEEE International Conference on Computer Vision*. 4642–4650.
- [17] Kai Wang, Manolis Savva, Angel X Chang, and Daniel Ritchie. 2018. Deep convolutional priors for indoor scene synthesis. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–14.
- [18] Yang Zhou, Zachary White, and Evangelos Kalogerakis. 2019. SceneGraphNet: Neural Message Passing for 3D Indoor Scene Augmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 7384–7392.