

# Complete the Look: Scene-based Complementary Product Recommendation (Supplementary Material)

Wang-Cheng Kang<sup>†\*</sup>, Eric Kim<sup>‡</sup>, Jure Leskovec<sup>‡§</sup>, Charles Rosenberg<sup>‡</sup>, Julian McAuley<sup>‡</sup>

<sup>‡</sup>Pinterest, <sup>§</sup>Stanford University, <sup>†</sup>UC San Diego

{wckang, jmcauley}@ucsd.edu, {erickim, jure, crosenberg}@pinterest.com

## 1. Performance Analysis

### 1.1. Ablation Study

We perform an ablation study to analyze the effect of the global and local compatibility measuring components. Table 1 shows the accuracy of our method and three variants on all datasets. ‘L’ and ‘G’ represents the variants with only the local component and the global component (respectively). We also exam the performance of the variant ‘G+L<sup>0</sup>’ which assigns equal weights on each region (instead of using attention weights). ‘G+L’ is the default model which uses both components and attention weights. We can see the hybrid model ‘G+L’ is better than using the two components individually. Also, the attention is helpful as it can boost performance compared with ‘G+L<sup>0</sup>’.

Method	Fashion-1	Fashion-2	Home
L	67.6	73.8	77.1
G	66.9	74.2	77.8
G+L <sup>0</sup>	66.9	74.1	78.6
G+L (Default)	<b>68.5</b>	<b>75.3</b>	<b>79.6</b>

Table 1. Ablation Study

### 1.2. The Effect of Local Features

As our method utilizes intermediate feature maps as local features, we exam the effect of using features from different layers and networks. In addition to different blocks of ResNet-50 [1], we also consider the VGG-16 [3] network where the `fc7` layer is used as the visual feature vector. In Table 2, we can observe that the ResNet-50 network with the `block3` feature achieves the best performance on all three datasets. Hence we choose to use the `block3` layer by default. Moreover, we find using VGG features is generally worse than using ResNet features.

### 1.3. Performance on Each Category

We exam the performance on each category, compared with the strongest baseline Siamese Nets [4]. Figure 1

Local Feature	Fashion-1	Fashion-2	Home
<b>ResNet-50</b>			
<code>block1</code> (28×28×256)	68.0	74.5	77.7
<code>block2</code> (14×14×512)	67.6	73.3	78.3
<code>block3</code> (7×7×1024)	<b>68.5</b>	<b>75.3</b>	<b>79.6</b>
<code>block4</code> (7×7×2048)	64.9	74.2	78.5
<b>VGG-16</b>			
<code>pool3</code> (28×28×512)	64.0	71.1	76.2
<code>pool4</code> (14×14×512)	63.3	71.7	76.7
<code>pool5</code> (7×7×512)	63.1	72.0	75.5

Table 2. The effect of local features

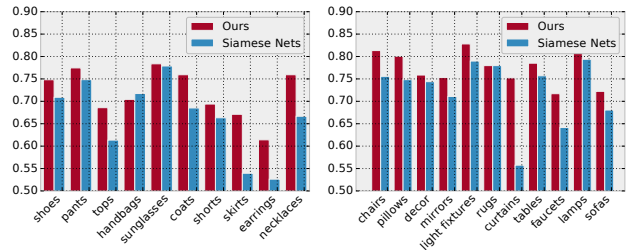


Figure 1. Accuracy per category. Left: STL-Fashion, right: STL-Home.

shows the performance comparison per category, based on the STL-Fashion and STL-Home datasets. We can see that in most categories, our method outperforms Siamese Nets. Moreover, the baseline has severe performance drops on categories like skirts and curtains. This verifies the effectiveness of our method when recommending products from various categories.

## 2. Implementation Details

The architecture of the two-layer network  $g(\Theta; \cdot)$  is Linear-BN-Relu-Dropout-Linear-L2Norm, where the dimensionality is set to  $4 \times d$  for the first linear layer, and  $d$  for the last linear layer. The dropout rate is set to 0.5, and the learning rate is set to 0.001. For all methods, we update the statistics of batch normalization [2] layers in ResNet during training, and find it generally improves the performance.

\*Work done while intern at Pinterest.



Figure 2. Recommended items with scenes *outside* of our datasets. Left image is query, right composes top recommended products.

### 3. Qualitative Examples

In addition to the examples where the query scenes are from the test set (as shown in the paper), we also test the model with scenes outside of the dataset (Figure 2). We obtained them as top search results for queries ‘street fashion’, ‘women fashion’, and ‘beach’ on Google. For each query, we show top-1 products from three categories.

### 4. Human Study Interface


Figure 3 shows screenshots of the interface when conducting the user study. We first provide a description of the task and a small sample of tests to the fashion experts, and then ask them to answer the questions (i.e., choose one of the two products that is more compatible with the given scene).

### References

- [1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [2] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.
- [3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- [4] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, and S. J. Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *ICCV*, 2015.


Task 3/20 ^

The Scene:




Given the scene image above, which product (category: Apparel & Accessories|Clothing|Pants) is more compatible? (Please use your best judgement and fashion sense to choose one.)

Product A




Product B




Task 5/20 ^

The Scene:




Given the scene image above, which product (category: Apparel & Accessories|Clothing|Outerwear|Coats & Jackets) is more compatible? (Please use your best judgement and fashion sense to choose one.)

Product A




Product B




Task 7/20 ^

The Scene:



Given the scene image above, which product (category: Apparel & Accessories|Shoes) is more compatible? (Please use your best judgement and fashion sense to choose one.)

Product A



Product B




Figure 3. Screenshots of the user study interface.