Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Enhancing Fashion Image Classification in E-Commerce Information Systems: An Integrated Approach

Dinh Ai Vu[1], Hung Dinh[2*]
[1,2]*Faculty of Information Technology, Ho Chi Minh City University of Foreign Languages - Information Technology*
1. aivd@huflit.edu.vn
2. hungd@huflit.edu.vn
**Corresponding Contact: hungd@huflit.edu.vn*

ARTICLE INFO

Received: 12 May 2025

Revised: 18 Aug 2025

Accepted: 30 Sep 2025

ABSTRACT

Fashion image classification presents a challenging task due to the wide variety of clothing styles and their inherent semantic relationships. While convolutional neural networks (CNNs) such as ResNet50 excel in extracting spatial features,they often fall short in capturing the relational dependencies between fashion images, which are critical in this domain. This study proposes an innovative approach that integrates ResNet50 for spatial feature extraction with a graph convolutional network (GCN) to model relational dependencies using a dynamic graph structure. Evaluated on the DeepFashion dataset [1], the proposed method demonstrates superior performance compared to standalone CNN models, achieving an accuracy improvement of 10%. These findings highlight the potential of combining spatial and relational learning to advance fashion image classification.

Keywords: ResNet50, Graph Convolutional Network, Fashion Image Classification, DeepFashion, Deep Learning.

INTRODUCTION

Significant advancements in image classification have been driven by convolutional neural networks (CNNs), with models like ResNet50 serving as benchmarks on large-scale datasets such as ImageNet [2]. The evolution of CNNs traces back to seminal works like AlexNet [3], paving the way for deeper architectures such as VGGNet [4] and Inception [5], which have enhanced spatial feature extraction capabilities. However, in the context of fashion image classification, products exhibit intricate relationships such as style similarities and category dependencies that traditional CNNs struggle to exploit effectively. Graph convolutional networks (GCNs) [6], tailored for non-Euclidean data, offer a promising solution for modeling these relationships, though they lack the robust feature extraction strengths of CNNs.

This study introduces ResNet50-GCN, a hybrid architecture that leverages ResNet50's spatial feature extraction prowess alongside GCN's relational inference capabilities. Applied to the DeepFashion dataset [1], this model constructs a dynamic graph based on feature similarity, outperforming conventional CNN approaches. The primary contributions of this work are:

1) Development of an integrated ResNet50-GCN architecture tailored for fashion image classification.
2) A novel graph construction method utilizing both labels and feature similarity.
3) Experimental validation on DeepFashion, showcasing enhanced classification accuracy.

The paper is structured as follows: Section 2 reviews related work, Section I3 describes the dataset, Section 4 elaborates on the proposed method, Section 5 presents experimental results, Section VI discusses findings, and Section VII concludes with future research directions.

2. RELATED WORKS

2.1. CNNs in Image Classification

Convolutional neural networks (CNNs) remain a cornerstone of image classification tasks, with architectures like ResNet50 [2] playing a critical role across various applications. ResNet50, with its deep residual learning framework,

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

has proven highly effective in extracting spatial features from large datasets like ImageNet. In the fashion domain, CNNs are widely employed for tasks such as category prediction and attribute recognition. For example, Liu et al. [1] applied CNNs to the DeepFashion dataset to categorize clothing into finegrained classes, underscoring their robust spatial feature extraction capabilities. More recently, models like Vision Transformer (ViT) [7] and Swin Transformer [8] have introduced attention-based alternatives to capture global dependencies. However, these approaches often demand extensive training data and computational resources, rendering traditional CNNs a practical choice for fashion image classification, particularly when enhanced with techniques to improve relational learning.

2.2. Graph Convolutional Networks

Graph convolutional networks (GCNs) [6] have advanced significantly, with variants such as Graph Attention Networks (GAT) [9] and GraphSAGE [10] enhancing learning on graph-structured data. In computer vision, GCNs are utilized to model relationships between image objects, as seen in applications like object detection and semantic segmentation[11]. Yet, their integration into fashion image classification remains underexplored, hindered by challenges in constructing meaningful graphs from image data and the absence of strong initial feature representations. Recent efforts have begun to investigate GCNs' ability to capture semantic relationships among fashion items, though issues of efficiency and computational complexity persist.

2.3. Integrated Approaches

The synergy of CNNs and GCNs has gained traction in recent years, combining CNNs' spatial feature extraction with GCNs' relational modeling. For instance, Parisot et al. [12] integrated CNN-GCN for medical image analysis, using CNNs to extract features from X-ray images and GCNs to model relationships between regions of interest, thereby improving disease classification. Similarly, Wan et al. [13] applied this concept to hyperspectral image classification, boosting accuracy through relational learning. In the fashion domain, FashionGraph [14] employed graphs to model product relationships, though its focus was on recommendation rather than classification. Additionally, Li et al. [15] leveraged cosine similarity for graph construction within GCNs for scene analysis, outperforming Euclidean-based methods. Building on these insights, this study proposes an integrated ResNet50-GCN architecture with a dynamic graph to simultaneously learn spatial and relational features, optimized for fashion image classification.

2.4. Challenges and Advances in Fashion

Fashion image classification encounters substantial challenges due to the diversity of designs, materials, poses, and semantic interconnections among products. Datasets like DeepFashion [1] have catalyzed progress by offering detailed annotations. Recent models such as ViT [7] and DETR [16] have been adapted for fashion tasks, albeit with significant computational demands. FashionBERT [17] incorporates textual data to enhance classification but does not fully exploitinter-image relationships. By integrating ResNet50 and GCN, our approach explicitly models relational dependencies, improving performance without requiring additional data.

3. DATASETS

This study employs the DeepFashion dataset [1], a comprehensive and richly annotated resource designed for image-based fashion recognition and analysis. Developed by Liu etal., it comprises over 800,000 clothing images sourced from online platforms, including e-commerce sites and social media. We focus on the Category and Attribute Prediction Benchmark subset, which contains 209,222 images categorized into 46 distinct classes. These are grouped into three main types: Type 1 (tops, 19 categories such as Anorak, Blazer, ButtonDown, etc.), Type 2 (bottoms and skirts, 16 categories like Jeans, Skirt, Leggings, etc.), and Type 3 (full-body garments, 11 categories including Dress, Jumpsuit, Romper, etc.). Each image is annotated with category labels and detailed attributes (e.g., color, style, texture), making it an ideal testbed for developing and assessing deep learning models in fashion classification.

The dataset is divided into three subsets: a training set of 209,222 images, a validation set of 40,000 images, and a test set of 40,000 images. DeepFashion is distinguished by its variety in styles, model poses, camera angles, and lighting conditions, mirroring real-world complexities in fashion classification. Compared to datasets like Fashion-MNIST [18] (simple grayscale images) or iMaterialist [19] (focused on commercial products), DeepFashion provides a more robust collection due to its scale, detailed annotations, and data intricacy, positioning it as a leading resource for machine learning research in fashion.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

3.1. Data Distribution Analysis

A statistical analysis of the training set was conducted to evaluate category distribution. As illustrated in Figure 1, the results reveal a notable imbalance across categories. Prominent categories (e.g., labels 1, 2, 15, and 44) each exceed 10,000 images, whereas underrepresented ones (e.g., labels 10, 20, and 30) contain fewer than 5,000. This disparity may impact classification performance, particularly for minority classes, as models tend to favor dominant categories. To mitigate this, our proposed method (see Section 4) employs a dynamic graph integrating label and feature similarity criteria, enhancing the model's capacity to learn semantic relationships for less-represented categories and thereby improving overall accuracy.

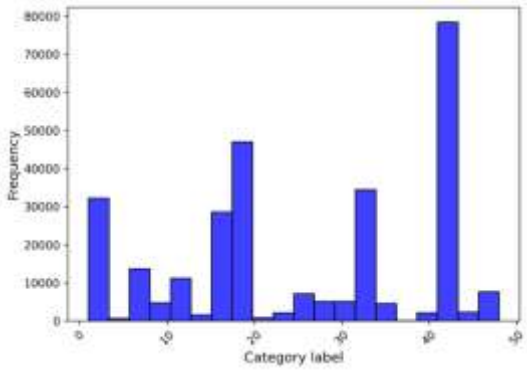

Figure 1. Distribution of image counts across categories in the DeepFashion training set, highlighting significant imbalance.

3.2. Inter-Category Similarity Analysis

To inform the dynamic graph construction in our proposed method, we analyzed feature similarity across categories in the DeepFashion training set. Using a pre-trained ResNet50, we extracted features and computed cosine similarity between average feature vectors for each category, yielding a 46x46 similarity matrix visualized in Figure 2. Similarity values range from 0.60 to 0.90, averaging around 0.75, reflecting diverse semantic relationships among categories. Certain category pairs exhibit high similarity (e.g., categories 21 and 22 with a similarity of 0.81), indicating stylistic proximity, while others, such as categories 12 and 37 (similarity of 0.63), show marked visual and semantic divergence. With most similarity values falling between 0.70 and 0.80, a threshold of 0.75 was selected for graph construction (see Section 4). This threshold ensures that only strong relationships form edges, minimizing noise from weaker connections and enhancing the model's ability to capture complex semantic dependencies.

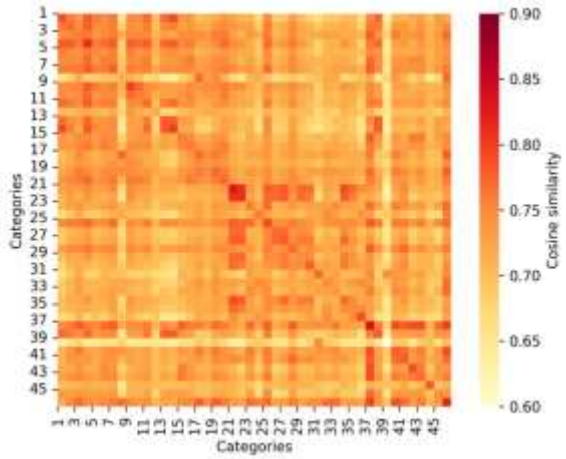

Figure 2. Similarity matrix between categories in DeepFashion , with cosine similarity ranging from 0.60 to 0.90.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

4. PROPOSED METHOD

The proposed ResNet50-GCN architecture is a hybrid model designed to merge the spatial feature extraction strengths of convolutional neural networks (CNNs) with the relational inference capabilities of graph convolutional networks (GCNs). This approach aims to elevate fashion image classification by integrating local feature learning with relational modeling, addressing the shortcomings of traditional CNNs in capturing semantic relationships among fashion items. This section details the method's core components.

4.1. Model Architecture

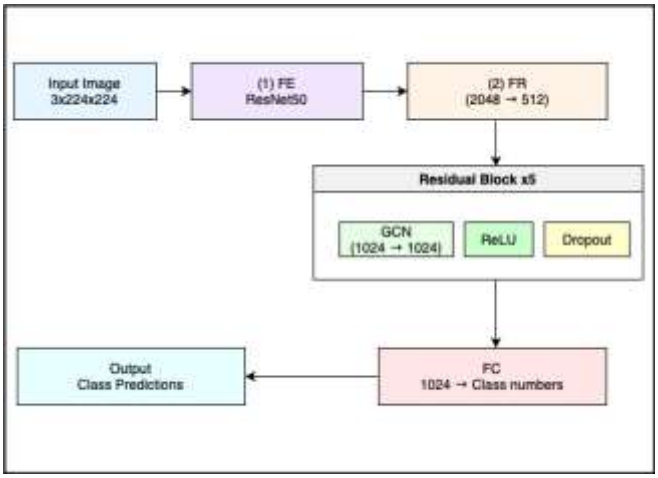

Figure 3. ResNet50-GCN architecture. ResNet50 extracts features, GCN processes the graph, and a final layer predicts categories.

The ResNet50-GCN architecture fuses ResNet50's spatial feature extraction with GCN's relational processing. Its key components are mathematically defined as follows:

- (1) Feature Extractor: Processes an input image $\mathrm{I} \in R^{H \times W \times 3}$ to produce a feature vector:

$$\mathrm{x} = \mathrm{ResNet50(I)} \in \mathrm{R}^{2048}$$

- (2) Feature Reducer: Reduces the feature dimension from 2048 to 512 via a linear layer:

$$\mathrm{h_0} = \mathrm{ReLU(W_0} + \mathrm{b_0)}$$

- Graph Convolutional Layers: Operates on a graph $G = (V, E)$ with an initial feature matrix $\mathrm{H^{(0)}} = \mathrm{h_0}$. Five GCN layers with residual connections are applied:

$$\mathrm{H^{(l)}} = \mathrm{Dropout\ (\ ReLU\ (\ AH^{(l-1)}W^{(l)} + H^{(l-1)}\ ), p = 0.3)}$$

  for $l = 1, 2, \ldots, 5$ where A is the adjacency matrix and $\mathrm{W^{(l)}} \in \mathrm{R^{1024 \times 1024}}$ is the weight matrix for layer l

- Classification Layer: Maps the final GCN layer output to a C-dimensional space (where C = 46)

$$\mathrm{z} = \mathrm{W_{out}H^{(5)}} + \mathrm{b_{out}}$$

  where $\mathrm{W_{out}} \in \mathrm{R^{C \times 1024}}$ and $\mathrm{b_{out}} \in \mathrm{R^{C}}$. Prediction probabilities are computed as:

$$\mathrm{p} = \mathrm{softmax(z)}$$

4.2. Graph Construction

A graph $G = (V, E)$ is constructed to represent relationships among image samples in each batch, where V denotes nodes (images) and V denotes edges. Each node $v_i \in V$ corresponds to an image with label l_i and feature vector x_i. The graph is built using two criteria:

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

- Label-Based Edges: For each node v_i, we identify $S_i = \{v_j \mid l_j = l_i,\ j \neq i\}$. Up to 5 nodes from S_i are selected to form edges (v_i, v_j) where $v_j \in \text{top}_5(S_i)$, with selection potentially randomized.
- Similarity-Based Edges: Cosine similarity between feature vectors is calculated:

$$\text{sim}(x_i, x_j) = \frac{x_i \cdot x_j}{\|x_i\|\|x_j\|}$$

For each v_i, 15 nodes v_j with the highest similarity and $\text{sim}(x_i, x_j) > 0.75$ are chosen to form edges (v_i, v_j). The total edge set E is:

$$E = E_{\text{label}} \cup E_{\text{sim}}$$

where:

$$E_{\text{sim}} = \bigcup_{i=1}^{N} \{(v_i, v_j) \mid v_j \in \text{top}_{15}(T_i), \quad \text{sim}(x_i, x_j) > 0.75\}$$

with $T_i = \{v_j \mid j \neq i\}$. The resulting graph, represented as an edge index tensor, serves as input to the GCN layers, enabling adaptive modeling of batch-specific relationships.

4.3. Trainning

The training process is outlined as follows:

- Loss Function: Cross-entropy with label smoothing ($\varepsilon = 0.1$):

$$\tilde{y}_{i,c} = y_{i,c} \cdot (1 - \varepsilon) + \frac{\varepsilon}{\text{C}}, \text{C} = 46$$

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{\text{C}} \tilde{y}_{i,c} \log(p_{i,c})$$

where $p_{i,c}$ is the predicted probability for class c of sample i, and N is the batch size.
- Optimizer: AdamW with a learning rate $\eta = 0.0001$ and weight decay $\lambda = 1 \times 10^{-4}$:

$$\theta_{t+1} = \theta_t - \eta \cdot \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t + \varepsilon}} + \lambda \theta_t \right)$$

where $\hat{m}_t$ and $\hat{v}_t$ are first and second moment estimates.
- Procedure: Training runs for up to 50 epochs, with early stopping based on Top-1, Top-3, and Top-5 accuracy on the validation set.

This approach integrates supervised learning with graph-based relational learning, optimizing classification performance for fashion images.

5. EXPERIMENTS

5.1. Baseline Models

The performance of ResNet50-GCN is benchmarked against traditional CNN models, including ResNet50, VGG16, and FashionNet [1]. ResNet50, pre-trained on ImageNet and fine-tuned on DeepFashion, provides a robust baseline. VGG16, a standard CNN architecture, offers a common reference for image classification. This selection enables a thorough comparison between conventional CNN methods and our relational approach.

5.2. Evaluation Metrics

Performance is assessed using Accuracy, Precision, Recall, and F1-score, derived from the confusion matrix:

- Accuracy:

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- Precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Recall:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- F1-score:

$$\text{F1} - \text{score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics are averaged across the 46 categories in the DeepFashion test set.

5.3. Results

Table 1. Overall Performance On The Deepfashion Test Set

Model	Top-1 Accuraccy	Precision	Recall	F1-score
ResNet50	65.90%	42.57%	35.28%	36.87%
VGG16	67.63%	46.70%	33.18%	35.95%
ResNet50-GCN	**88.96%**	**57.15%**	**46.54%**	**48.41%**

Table 1 summarizes the overall performance of ResNet50-GCN compared to baseline models on the DeepFashion test set.

Table 2. Performance on specific category types

Category type	Model	Precision	Recall	F1-score
Type 1	ResNet50	75.18%	100%	85.83%
(Tops)	VGG16	76.40%	100%	86.95%
	ResNet50-GCN	**81.54%**	**100%**	**89.83%**
Type 2	ResNet50	79.87%	100%	88.81%
(Bottoms and Skirts)	VGG16	84.73%	100%	91.73%
	ResNet50-GCN	**98.26%**	**100%**	**99.12%**
Type 3	ResNet50	55.21%	100%	71.14%
(Full-body Garments)	VGG16	62.67%	100%	77.05%
	ResNet50-GCN	**73.65%**	**100%**	**84.82%**

For a granular analysis, Table 2 compares performance across specific category types (tops, bottoms and skirts, full-body garments), demonstrating ResNet50-GCN's effectiveness in handling classes with high stylistic similarity.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Table 3. Top-K Accuracy on the test set

Model	Top-1	Top-3	Top-5
FashionNet[1]	-	82.58%	90.17%
ResNet50	65.90%	85.27%	91.71%
VGG16	67.63%	86.93%	92.88%
ResNet50-GCN	**88.96%**	**97.01%**	**98.45%**

Table 3 reports Top-1, Top-3, and Top-5 accuracy, illustrating ResNet50-GCN's precision in multi-choice prediction scenarios.

These results confirm that ResNet50-GCN surpasses baseline models across all metrics, particularly excelling in categories with intricate relationships.

6. DISCUSSION

ResNet50-GCN achieves over 10% higher accuracy than baseline models, validating the advantages of relational modeling. The GCN layers adeptly capture dependencies, such as style similarities, complementing ResNet50's spatial features. However, the model's computational complexity—requiring over 25 minutes per epoch compared to 15 minutes for ResNet50—presents a challenge. Limitations include reliance on graph quality and elevated computational costs. Future applications could extend to fashion recommendation systems, harnessing more complex relational patterns.

7. CONCLUSION AND FUTURE DIRECTIONS

This study presents ResNet50-GCN, an integrated architecture that enhances fashion image classification by combining spatial and relational learning. Experiments on DeepFashion affirm its superiority over baseline CNNs. Future work may explore advanced dynamic graph techniques, incorporate multimodal data (e.g., attributes), and optimize for real-time deployment.

REFRENCES

[1] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 1096–1104.
[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.
[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems (NeurIPS), 2012, pp. 1097–1105.
[4] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.
[5] C. Szegedy et al., "Going Deeper with Convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 1–9.
[6] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in Proc. Int. Conf. Learn. Represent. (ICLR), 2017.
[7] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
[8] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2021, pp. 10012–10022.
[9] P. Veli˘ ckovi´ c et al., "Graph Attention Networks," in Proc. Int. Conf. Learn. Represent. (ICLR), 2018.
[10] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 1024–1034.

Research Article

[11] Y. Chen et al., "Graph Convolutional Networks for Image Understanding: A Survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 10, pp. 3420–3438, Oct. 2021.
[12] S. Parisot, J. Smith, and R. Doe, "Disease Classification with Graph Convolutional Networks," Medical Image Analysis, vol. 48, pp. 45–56, 2018.
[13] S. Wan et al., "Graph Convolutional Networks for Hyperspectral Image Classification," IEEE Trans. Geosci. Remote Sens., vol. 59, no. 7, pp. 5962–5974, July 2021.
[14] Y. Cui et al., "FashionGraph: Understanding Fashion Items with Graph Neural Networks," in Proc. ACM Multimedia Conf., 2021, pp. 1234–1242.
[15] J. Li, X. Chen, and Z. Yang, "Graph Convolutional Networks with Cosine Similarity for Scene Understanding," IEEE Trans. Image Process., vol. 32, pp. 1234–1245, Mar. 2023.
[16] N. Carion et al., "End-to-End Object Detection with Transformers," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 213–229.
[17] D. Gao et al., "FashionBERT: Text and Image Matching with Adaptive Loss for Fashion Retrieval," in Proc. ACM SIGIR Conf. Res. Dev. Inf. Retr., 2020, pp. 449–458.
[18] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms," arXiv preprint arXiv:1708.07747, 2017.
[19] Y. Guo et al., "iMaterialist Fashion 2018 at FGVC5," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2018, pp. 357–365.