A Metric Learning-Based Approach to Consumer-to-Shop Fashion Image Retrieval

2271013 Minji Kim
2271033 Youngchae Song
2391028 Seungyeon Jo
2491025 Jiwon Lee

Abstract

This project addresses the Consumer-to-Shop (C2S) fashion image retrieval problem by formulating it as a metric learning–based retrieval task. Using a sampled subset of the DeepFashion C2S dataset, we construct an embedding-based retrieval pipeline and analyze the impact of backbone selection, loss functions, and preprocessing strategies. EfficientNet-B3 with Batch-Hard Triplet Loss achieves the best overall performance among the tested configurations. We further demonstrate that careful preprocessing design and embedding dimension selection improve retrieval performance without modifying the model architecture. The final model achieves Recall@1, Recall@5 and Recall@10 of 0.6402*, 0.7280 and 0.7751 on the test set, highlighting the importance of embedding design in consumer-to-shop fashion retrieval.*

1. Introduction

1.1 Background & Motivation

With the rapid growth of online fashion commerce, there is increasing demand for image-based product search that allows users to find similar items using their own photos. In particular, Consumer-to-Shop (C2S) fashion image retrieval aims to match consumer images of worn clothing with corresponding shop images, playing a crucial role in improving user experience and purchase conversion rates.

However, C2S retrieval is inherently challenging due to significant domain gaps between consumer and shop images, including differences in background, lighting, pose, and garment appearance. To address these challenges, this project focuses on a metric learning–based retrieval approach using feature embeddings, rather than conventional classification, and experimentally analyzes a C2S retrieval pipeline on the DeepFashion dataset.

1.2 Problem Statement

The goal of this project is to develop a fashion image retrieval model that takes a consumer image as input and retrieves the same or visually most similar shop item with high accuracy. To achieve this, we construct a preprocessing pipeline based on the DeepFashion C2S dataset, extract image embeddings using various backbone networks, and learn an embedding space through metric learning–based loss functions.

Rather than targeting a large-scale commercial system, this undergraduate-level study focuses on systematically analyzing key factors that influence retrieval performance, including backbone selection, loss functions, preprocessing strategies, and embedding design, under limited data and time constraints. Through this step-by-step evaluation, we aim to gain practical insights into the characteristics of the consumer-to-shop retrieval problem and the design considerations for effective retrieval models.

2. Intuition & Proposed Approach

The core of the Consumer-to-Shop (C2S) fashion retrieval problem is not image classification, but measuring visual similarity between consumer and shop images. Accordingly, this task is more naturally formulated as a retrieval problem rather than a classification problem. Instead of predicting predefined classes, the model must learn an embedding space in which visually similar items are placed close together.

To address this, we adopt a metric learning approach that directly optimizes distances between image embeddings. By pulling consumer–shop pairs of the same item closer while pushing apart dissimilar items, metric learning aligns well with the retrieval objective and Recall@K evaluation.

The choice of backbone network is critical due to the substantial domain gap between consumer and shop images in terms of background, lighting, and pose. We therefore employ EfficientNet, which provides strong generalization capability and a favorable trade-off between performance and computational efficiency, making it suitable for an undergraduate project with limited resources.

Finally, providing the best balance between representational capacity and generalization, selecting an appropriate embedding dimension is applied for retrieval performance, with a 256-dimensional embedding.

Dataset & Preprocessing

3.1 Dataset

We use the Consumer-to-Shop Clothes Retrieval benchmark from the DeepFashion dataset. The benchmark contains consumer and shop images with the following annotations:

- Bounding boxes for clothing regions
- Item IDs linking consumer images to their corresponding shop images
- Split labels (train/val/test) for consumer images

The original C2S dataset comprised four categories: Tops, Dresses, Clothing, and Trousers. However, given the constraint of the dataset's large volume for our specific project, we opted to use only two categories: Tops, which had the highest proportion in the original data, and Dresses, which is comparatively easier for shape recognition.

Even when restricted to only the Tops and Dresses categories, the dataset still contained a large volume of data, with item IDs distributed across the training, validation, and test sets as follows: 12,214, 6,235, and 6,134, respectively. This scale was further amplified by each item possessing between 2 and 10 images. Considering our limited computational resources, we constituted our dataset by sampling the original item IDs: 10% for the training set, and 2% each for the validation and test sets. Through this sampling process, we retained 1,221, 124, and 122 unique item IDs for the training, validation, and test sets, respectively. Using the images associated with these selected items, we constructed 12,421, 1,097, and 956 positive pairs for each corresponding split.

3.2 Preprocessing

All images were resized to 224×224 to enforce a consistent spatial resolution across the dataset, which supports stable optimization and ensures architectural compatibility with standard convolutional models. After resizing, the images were converted into tensor format and normalized using fixed channel-wise mean and standard deviation values, allowing the input distribution to remain numerically stable during training. For the training split, a random horizontal flip was applied with a probability of 0.5, and controlled variations in brightness, contrast, and saturation were introduced. These stochastic transformations provide moderate appearance diversity, reducing the risk of overfitting and improving robustness to variations in viewpoint and illumination. In contrast, the validation and test splits employed only resizing, tensor conversion, and normalization, ensuring consistent evaluation without altering the intrinsic visual characteristics of the samples.

3.3 Challenges in Preprocessing

In the early stages of the project, an attempt was made to improve the model's generalization capability by employing a sampling strategy that aimed to maintain item diversity while reducing the number of images within each item ID. However, when we tried to sufficiently reduce the overall dataset volume using this methodology, we encountered a significant problem: only one or two images were left per item ID, which severely hindered the model's ability to learn effectively. Consequently, we pivoted our approach and successfully reconfigured the dataset using a method that samples based on the Item ID itself.

3. Method

Experimental Setup

All models were trained for up to 40 epochs using the Adam optimizer (learning rate $1\text{x}10^{-4}$) and batch size 32. The embedding dimension for the feature vectors was fixed at 128, and all embeddings were L2-normalized before distance computation. Early stopping was applied based on validation Recall@5 (patience = 5). Data augmentation including RandomHorizontalFlip (0.5), and ColorJitter (0.2, 0.2, 0.2) was applied. Performance was evaluated using Recall@K (K = 1, 5, 10). For reproducibility, all experiments were conducted with a fixed random seed of 42.

4.1 Backbone Selection

We conducted an experiment to determine the optimal backbone model, fixing the loss function as Semi-Hard Triplet Loss. We compared the performance of two candidate models: ResNet-34, which is a relatively light yet stable model within the classical CNN structure of the ResNet family, and EfficientNet-B3, a structure that maximizes efficiency and is considered one of the latest State-of-the-Art (SOTA) models.

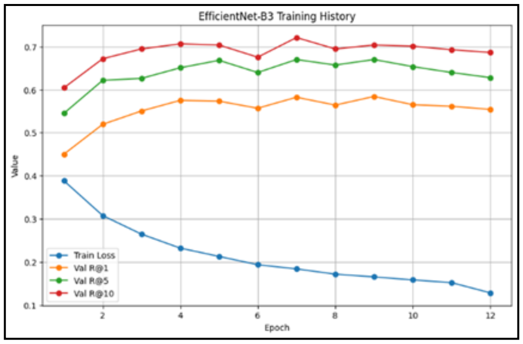

Figure 1: EfficientNet-B3 Training History

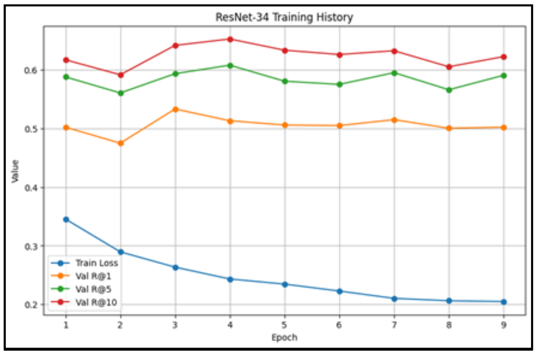

Figure 2: ResNet-34 Training History

The experimental results demonstrated clear differences. EfficientNet-B3 showed a rapid and sustained decrease in loss, while its R@K values increased steadily, maintaining a comparatively high level until early stopping was triggered. In contrast, ResNet-34 exhibited a slower rate of loss reduction, and its R@K values briefly increased early on before becoming unstable or stagnant, leading to premature termination.

Backbone	R@1	R@5	R@10
EfficientNet-B3	0.5825	**0.6700**	0.7211
ResNet-34	0.5132	0.6080	0.6527

Consequently, EfficientNet-B3 showed significantly higher performance than ResNet-34 across all Recall@K metrics. This suggests that EfficientNet-B3 learned more appropriate feature embeddings for the C2S task. Furthermore, its consistent decrease in loss and steady increase in R@K imply higher training efficiency and greater model potential compared to ResNet-34.

We hypothesize that in image retrieval tasks like C2S, which require robustness to domain shifts, EfficientNet's ability to efficiently process high resolution and its subsequent fine-grained feature extraction capabilities—achieved through Compound Scaling provide a substantial advantage over ResNet-34's traditional architecture. [1]

4.2 Loss Function Selection

This experiment compares the effectiveness of three representative metric learning loss functions—Semi-Hard Triplet Loss, Batch-Hard Triplet Loss, and InfoNCE Loss—for fashion image retrieval. The goal is to learn embeddings that bring *consumer* and *shop* images of the same item closer in feature space while separating different items.

Semi-Hard Triplet Loss Negative samples are chosen such that they are farther than positives but still within the margin boundary:

$$d(a,p) < d(a,n) < d(a,p) + m \quad (1)$$

Batch-Hard Triplet Loss For each anchor sample, the hardest positive and hardest negative samples are selected within a batch.

$$\mathcal{L}_{BH} = \max(0, d(a,p) - d(a,n) + m) \quad (2)$$

Where $d(\cdot)$ denotes Euclidean distance and m is a margin (set to 0.3).

InfoNCE Loss Contrastive learning objective where all positive and negative pairs are used via a softmax:

$$\mathcal{L}_{NCE} = -\log \frac{\exp(s_{pos}/\tau)}{\sum_j \exp(s_j/\tau)} \quad (3)$$

Where s_j denotes cosine similarity and $\tau = 0.07$ is the temperature parameter.

All experiments were conducted using the EfficientNet-B3 backbone under identical training settings. Each loss function was tested over three independent runs, and the mean and standard deviation of validation performance were reported to ensure statistical reliability.

Loss Function	R@1	R@5	R@10
Semi-Hard Triplet	0.5910 ± 0.0104	0.6712 ± 0.0014	0.7092 ± 0.0105
Batch-Hard Triplet	0.6095 ± 0.0071	**0.6834 ± 0.0055**	0.7171 ± 0.0005
InfoNCE	0.5820 ± 0.0122	0.6645 ± 0.0026	0.6905 ± 0.0019

The above table summarizes the validation results. Batch-Hard Triplet Loss achieved the highest retrieval performance, producing compact and well-separated embeddings through effective hard sample mining. Semi-Hard Triplet showed stable convergence but underutilized difficult negatives, leading to slightly lower recall. InfoNCE Loss exhibited consistent training behavior but performed worse due to its reliance on large batch sizes; with a batch size of 32, its contrastive supervision was less effective.

Overall, Batch-Hard Triplet Loss demonstrated the strongest discriminative capability in our retrieval framework.

4. Advanced Preprocessing and Embedding Experiments (Stage 2)

With the best-performing EfficientNet-B3 backbone and Batch-Hard Triplet Loss fixed, this stage aimed to analyze how much retrieval performance could be improved by refining the input preprocessing and embedding configuration. In the consumer-to-shop fashion retrieval task, both preprocessing and embedding design significantly influence embedding quality and generalization. Therefore, we analyzed performance variations by modifying only preprocessing- and embedding-related factors while keeping the model architecture constant. The experiments covered four aspects:

1. the strength and method of data augmentation
2. margin adjustments during bounding box cropping
3. domain normalization between consumer and shop images
4. embedding dimension settings (128, 256, 512)

Each experiment was conducted independently, and the following sections describe their setups and results.

5.1 Augmentation

To evaluate the effect of image augmentation strength on retrieval performance, we compared four augmentation modes — *Weak*, *Medium*, *Strong*, and *Super-Strong* — applied during training.

Mode	Crop	Flip	Color Jitter	Erasing
Basline	Resize	0.5	0.1	X
Weak	Resize	0.5	0.2	X
Medium	RandomResized Crop(0.8~1.0)	0.5	0.2	X
Strong	RandomResized Crop(0.6~1.0)	0.5	0.3	RandomErasing (p=0.25, scale=0.02~0.4)
Super-strong	RandomResized Crop(0.5~1.0)	0.5	0.4	RandomErasing (p=0.4, scale=0.02~0.4)

Mode	R@1	R@5	R@10
Baseline	0.6095	0.6834	0.7171
Weak	0.6226	0.6627	0.7165
Medium	0.5816	0.6609	0.7037
Strong	0.6135	**0.6837**	0.7174
Super-strong	0.6062	0.6727	0.7211

As shown in Table, *Strong* augmentation achieved the highest Recall@5 (0.6837), but its improvement over the *Baseline* (0.6834) was marginal. Other augmentation modes (*Weak*, *Medium*, and *Super-Strong*) even resulted in slightly lower performance. These findings indicate that the baseline augmentation—simple resizing with mild color jitter and flipping—is already sufficient for robust representation learning in our fashion retrieval setup.

5.2 Margin adjustment during bounding box-based cropping

In this experiment, we investigated how adjusting the margin size during bounding box–based cropping affects retrieval performance in the consumer-to-shop fashion image retrieval task. While bounding box annotations provide a reliable way to localize clothing items, using overly tight crops may remove contextual visual cues, whereas overly loose crops may introduce irrelevant background noise. Therefore, an appropriate margin around the bounding box is expected to play a crucial role in learning robust and discriminative embeddings.

To analyze this effect, we fixed the backbone network (EfficientNet-B3), loss function (Batch-Hard Triplet Loss), and all training hyperparameters, and varied only the margin ratio applied to the bounding box during cropping. Specifically, three margin settings were evaluated: 0.0 (tight crop), 0.1, and 0.2, where the margin ratio indicates the proportional expansion of the bounding box in all directions.

All experiments were trained with identical data splits and learning configurations, and performance was evaluated using Recall@K on the validation set. The results are summarized in the table below.

Margin Ratio	R@1	R@5	R@10
0.0 (Baseline)	0.6095	**0.6834**	0.7171
0.1	0.5834	0.6582	0.7074
0.2	0.5871	0.6554	0.6955

As shown in the results, the baseline tight crop setting (margin = 0.0) achieved the best performance across all Recall@K metrics. In particular, introducing additional margin did not lead to performance gains; instead, Recall@5 slightly decreased when a margin of 0.1 was applied. This indicates that, for the DeepFashion consumer-to-shop retrieval task, tightly cropped bounding boxes already capture the most discriminative garment information necessary for effective embedding learning.

One possible explanation is that the dataset annotations provide sufficiently accurate bounding boxes that precisely localize the target garments. As a result, including surrounding context—even in small amounts—may introduce irrelevant visual elements such as background regions, body parts, or occlusions, which can interfere with the model's ability to focus on fine-grained garment-specific features. This effect becomes more pronounced as the margin increases to 0.2, where further performance degradation is observed.

These results suggest that, contrary to the initial hypothesis, additional contextual information is not always beneficial for fine-grained fashion retrieval. Instead, preserving clean and tightly focused garment regions appears to be more important than capturing broader contextual cues. The experiment highlights a trade-off between contextual enrichment and feature purity, with the latter playing a more critical role in this setting. Based on these observations, the baseline configuration with a margin ratio of 0.0 was retained for all subsequent experiments in Stage 2.

5.3 Lightweight Domain Normalization

To evaluate how domain-aware preprocessing influences retrieval robustness, four Domain Normalization modes were implemented: Baseline, CLAHE, Gray-World, and a Combined method.

The Baseline mode applies only standard resizing and augmentation without domain-specific correction. CLAHE-based normalization enhances local contrast by converting the image to YUV, applying contrast-limited histogram equalization to the luminance channel, and converting back to RGB. The gray-world mode adjusts each color channel so that their average intensities match a shared global mean, reducing color cast and illumination imbalance. The combined mode applies CLAHE followed by gray-world adjustment to correct both contrast and color inconsistencies across domains.

Mode	R@1	R@5	R@10
Baseline	0.6095	**0.6834**	0.7171
CLAHE	0.6026	0.6727	0.7311
Gray-world	0.6108	0.6828	0.7366
Both	0.6117	0.6782	0.7284

When evaluated in terms of Recall@5, Gray-World normalization achieved performance comparable to the baseline, while CLAHE and the combined approach resulted in degraded performance. This indicates that color and contrast normalization are not decisive factors for improving retrieval performance on the given dataset. The results further suggest that lightweight domain normalization alone is insufficient to yield consistent improvements in retrieval robustness.

5.4 Embedding Dimension

Embedding Dimension	R@1	R@5	R@10
128(Baseline)	0.6095	0.6834	0.7171
256	0.6080	**0.6892**	0.7238
512	0.6126	0.6755	0.7119

Our experimental results clearly showed that increasing the embedding dimension from 128 to 256 led to a notable rise in R@5 performance, from 0.6834 to 0.6892. This suggests that for a fine-grained retrieval task like C2S, a 128-dimensional space lacked the representational power needed to compress the rich features extracted by EfficientNet-B3. Conversely, performance declined to 0.6755 when the dimension was excessively increased to 512. This drop is likely due to the Curse of Dimensionality, where the sparsity of the feature space increases relative to the sampled dataset size, thereby hindering the model's generalization capability. Therefore, the 256-dimensional embedding space appears to provide the optimal trade-off between representational power and generalization.

5. Final Model Evaluation on Test Set

Based on the experimental results presented in Section 5, a **single best-performing configuration, featuring an Embedding Dimension of 256,** was selected to construct the final model. Rather than exhaustively evaluating all possible combinations of preprocessing and embedding strategies, we chose the most effective setting observed in the Stage 2 experiments, considering the limited experimental time and computational resources.

The final model configuration is summarized as follows:

- Backbone: EfficientNet-B3
- Loss Function: Batch-Hard Triplet Loss
- Embedding: L2-normalized embedding space
- Embedding Dimension: 256

The final model's performance on the completely unseen test set was measured as Recall@5 = 0.7280, significantly surpassing the validation score. This uplift suggests that the test set, while fully independent, possessed an inherently more uniform or less challenging distribution of item pairs compared to the validation set. Crucially, the high test performance validates that the optimized embedding structure (256-dim) generalizes robustly to real-world domain shifts, confirming the effectiveness of our design choices.

Metric	R@1	R@5	R@10
	0.6402	**0.7280**	0.7751

It should be noted that, due to time constraints, not all combinations of preprocessing and embedding settings were exhaustively evaluated. Instead, this study focused on identifying a strong and practically effective configuration, demonstrating that meaningful performance improvements can be achieved through careful preprocessing and embedding space optimization while keeping the core model architecture fixed.

6. Conclusion

This project presented a metric learning–based consumer-to-shop image retrieval pipeline built on a sampled subset of the DeepFashion C2S dataset. Through systematic experiments, EfficientNet-B3 and Batch-Hard Triplet Loss were selected as the most effective backbone and training objectives.

Through further preprocessing and embedding experiments, a 256-dimensional embedding was selected as the final configuration, achieving a favorable balance between representational capacity and generalization under limited computational resources. The final model achieved Recall@1, Recall@5, and Recall@10 scores of 0.6402, 0.7280, and 0.7751 on a held-out test set, highlighting the effectiveness of embedding design and preprocessing without changes to the core architecture.

7. Discussion and Future Work

Despite the promising results, this study has several limitations. The experiments were conducted on a reduced subset of the DeepFashion C2S dataset, and the exploration of preprocessing options, embedding configurations, and training hyperparameters was limited. Moreover, advanced metric learning losses and mining strategies were not investigated due to time and computational constraints.

Future work may focus on scaling up the training data, performing systematic searches over preprocessing and hyperparameters, and incorporating alternative metric learning losses such as ArcFace, Circle Loss, or ProxyNCA++. These extensions could further improve retrieval performance in challenging cross-domain fashion retrieval scenarios.

References

[1] M. Tan and Q. V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Networks. In *CVPR*, pages 234-778, 2019.