2025 -2 Computer Vision

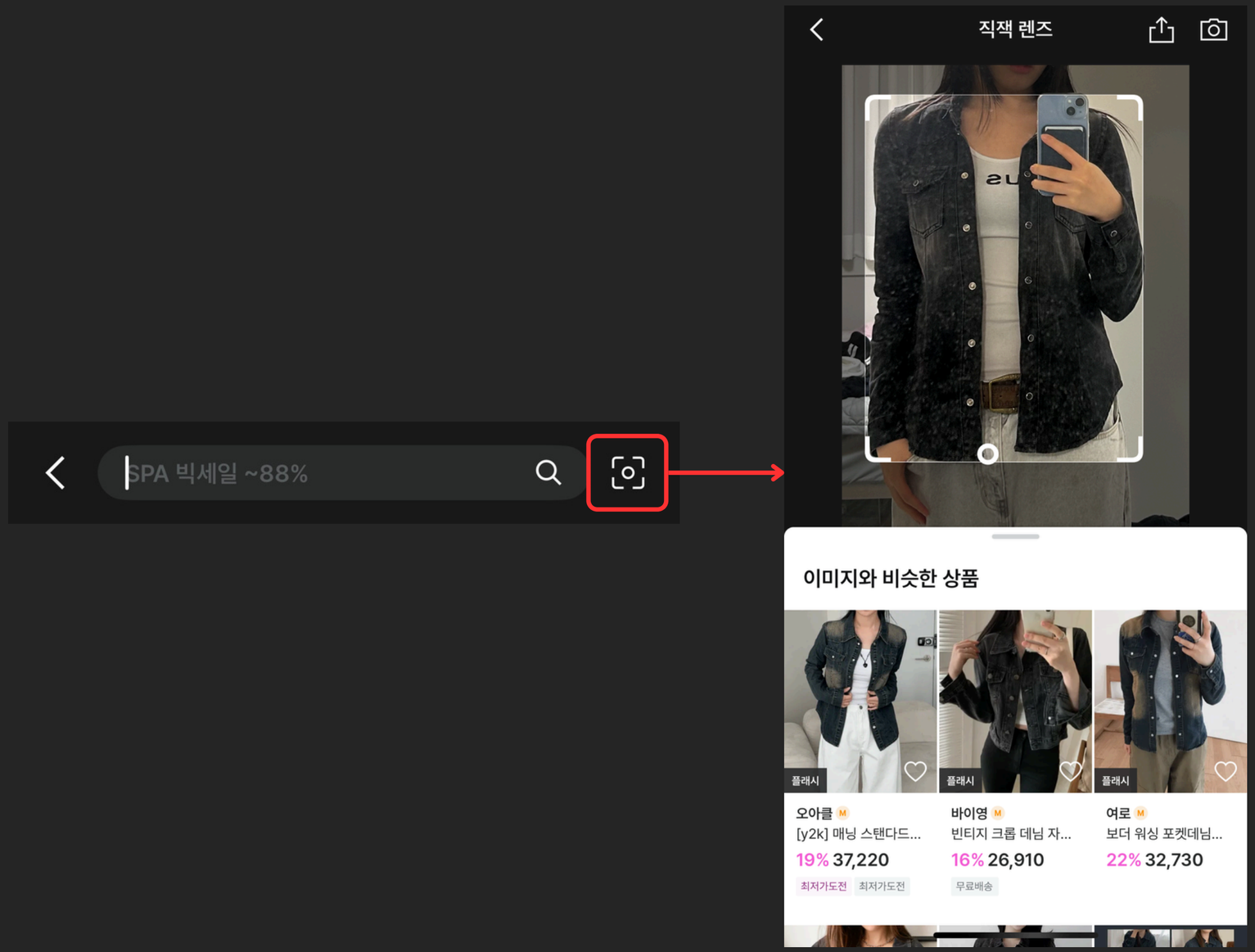
A Metric Learning-Based Approach to Consumer-to-Shop Fashion Image Retrieval

Team 2 김민지, 송영채, 이지원, 조승연

Overview

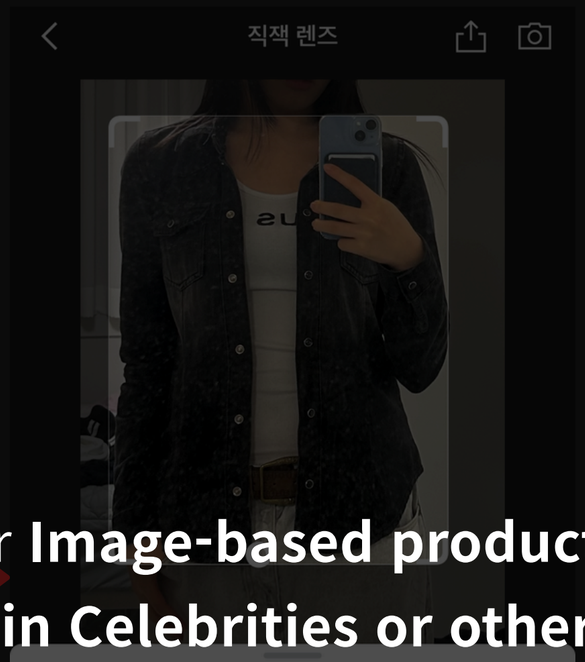
- **Introduction and Approach**
- **Dataset & Sampling & Preprocessing**
- **Stage 1：Backbone and Loss function selection**
- **Stage 2：Advanced Preprocessing and Embedding Experiments**
- **Final Model Evaluation on Test Set**
- **Conclusion**

Introduction and Approach



→ Consumer Image

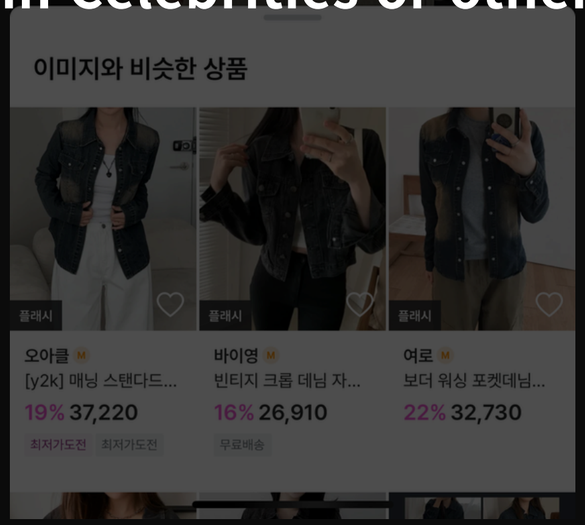
→ Shop Image



→ Consumer Image

→ Increasing demand for **Image-based product search**

“Users want to find items in Celebrities or others photos!”



→ Shop Image

01 Introduction and Approach

Goal

Consumer-to-Shop (C2S) retrieval

Given a consumer image,

Retrieve the same or most similar shop item.

Difficulty

Large domain gap between Consumer image ⇔ Shop image

(by Background, Lighting, Pose, Appearance)

→

Approach

Metric learning–based image **retrieval** instead of classification

Focus of analysis

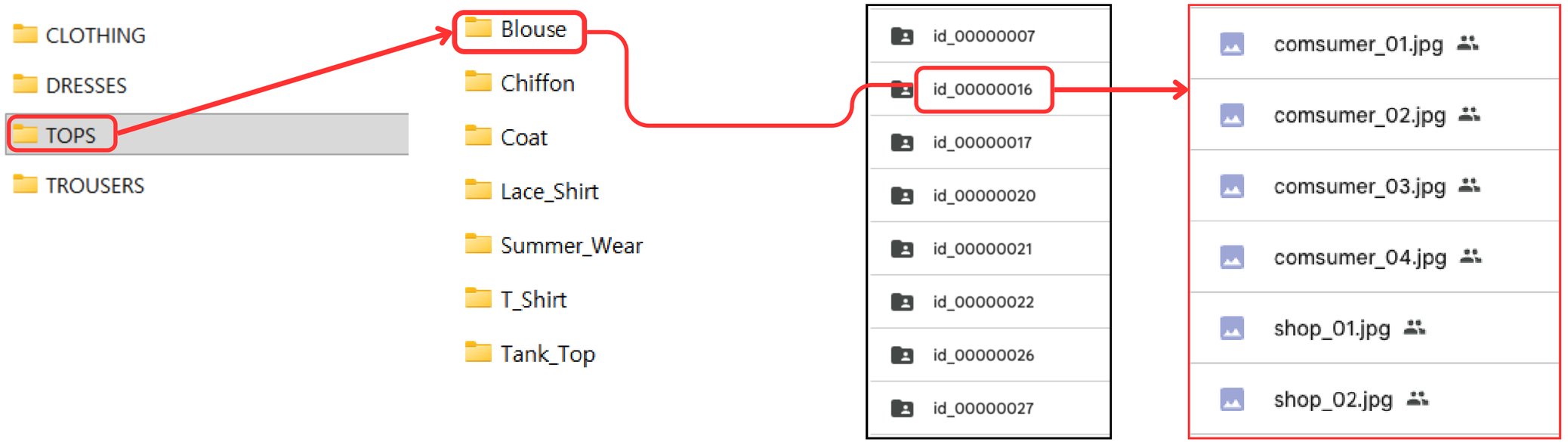
- **Backbone** networks
- **Loss functions**
- **Preprocessing** strategies
- Advanced **embedding design**

Dataset & Sampling & Preprocessing

02-1 Dataset

Dataset: DeepFashion Consumer-to-Shop (C2S)

- Provides **paired consumer and shop images** of the same clothing item
- **Categories**: Clothing, Dresses, Tops, Trousers



02-1 Dataset

Dataset: DeepFashion Consumer-to-Shop (C2S)

Labels

- **Item ID** linking consumer ↔ shop images
- Clothing **bounding boxes**
- Predefined **train / val / test splits**

item_id	consumer_path	cons_x1	cons_y1	cons_x2	cons_y2	shop_path	shop_x1	shop_y1	shop_x2	shop_y2	split
id_00006282	img/CLOTHING/Polo_Shirt/id_00006282/consumer_0...	1	1	223	300	img/CLOTHING/Polo_Shirt/id_00006282/shop_01.jpg	1	39	200	292	train
id_00006282	img/CLOTHING/Polo_Shirt/id_00006282/consumer_0...	1	1	169	230	img/CLOTHING/Polo_Shirt/id_00006282/shop_01.jpg	1	39	200	292	train
id_00006282	img/CLOTHING/Polo_Shirt/id_00006282/consumer_0...	6	8	215	300	img/CLOTHING/Polo_Shirt/id_00006282/shop_01.jpg	1	39	200	292	train
id_00006282	img/CLOTHING/Polo_Shirt/id_00006282/consumer_0...	1	1	225	300	img/CLOTHING/Polo_Shirt/id_00006282/shop_01.jpg	1	39	200	292	train
id_00006695	img/CLOTHING/Polo_Shirt/id_00006695/consumer_0...	37	66	164	211	img/CLOTHING/Polo_Shirt/id_00006695/shop_01.jpg	10	60	178	259	test
...	...	...	...	...	...	...	...	...	...	...	...
id_00026713	img/DRESSES/Dress/id_00026713/consumer_12.jpg	8	61	223	272	img/DRESSES/Dress/id_00026713/shop_01.jpg	63	1	154	295	val
id_00026713	img/DRESSES/Dress/id_00026713/consumer_13.jpg	19	1	211	285	img/DRESSES/Dress/id_00026713/shop_01.jpg	63	1	154	295	val
id_00026717	img/DRESSES/Dress/id_00026717/consumer_01.jpg	61	72	146	247	img/DRESSES/Dress/id_00026717/shop_01.jpg	1	1	200	300	val
id_00026719	img/DRESSES/Dress/id_00026719/consumer_01.jpg	84	53	158	171	img/DRESSES/Dress/id_00026719/shop_01.jpg	50	44	178	187	train
id_00026719	img/DRESSES/Dress/id_00026719/consumer_02.jpg	83	1	197	162	img/DRESSES/Dress/id_00026719/shop_01.jpg	50	44	178	187	train

02-2 Dataset Sampling

Limited to two Categories

- Tops (**largest portion** - 64%)
- Dresses (relatively **easier shape recognition**)

Image Sampling

- Within **each item ID**

item_id	consumer_path	cons_x1	cons_y1	cons_x2	cons_y2	shop_path	shop_x1	shop_y1	shop_x2	shop_y2	split
id_00006282	img/CLOTHING/Polo_Shirt/id_00006282/consumer_0...	1	1	223	300	img/CLOTHING/Polo_Shirt/id_00006282/shop_01.jpg	1	39	200	292	train
id_00006282	img/CLOTHING/Polo_Shirt/id_00006282/consumer_0...	1	1	169	230	img/CLOTHING/Polo_Shirt/id_00006282/shop_01.jpg	1	39	200	292	train
id_00006282	img/CLOTHING/Polo_Shirt/id_00006282/consumer_0...	6	8	215	300	img/CLOTHING/Polo_Shirt/id_00006282/shop_01.jpg	1	39	200	292	train
id_00006282	img/CLOTHING/Polo_Shirt/id_00006282/consumer_0...	1	1	225	300	img/CLOTHING/Polo_Shirt/id_00006282/shop_01.jpg	1	39	200	292	train
id_00006695	img/CLOTHING/Polo_Shirt/id_00006695/consumer_0...	37	66	164	211	img/CLOTHING/Polo_Shirt/id_00006695/shop_01.jpg	10	60	178	259	test
...	...	...	...	...	...	...	...	...	...	...	...
id_00026713	img/DRESSES/Dress/id_00026713/consumer_12.jpg	8	61	223	272	img/DRESSES/Dress/id_00026713/shop_01.jpg	63	1	154	295	val
id_00026713	img/DRESSES/Dress/id_00026713/consumer_13.jpg	19	1	211	285	img/DRESSES/Dress/id_00026713/shop_01.jpg	63	1	154	295	val
id_00026717	img/DRESSES/Dress/id_00026717/consumer_01.jpg	61	72	146	247	img/DRESSES/Dress/id_00026717/shop_01.jpg	1	1	200	300	val
id_00026719	img/DRESSES/Dress/id_00026719/consumer_01.jpg	84	53	158	171	img/DRESSES/Dress/id_00026719/shop_01.jpg	50	44	178	187	train
id_00026719	img/DRESSES/Dress/id_00026719/consumer_02.jpg	83	1	197	162	img/DRESSES/Dress/id_00026719/shop_01.jpg	50	44	178	187	train

Split		Total Length
Train	50.5%	32240
Val	24.99%	
Test	24.49%	

02-2 Dataset Sampling

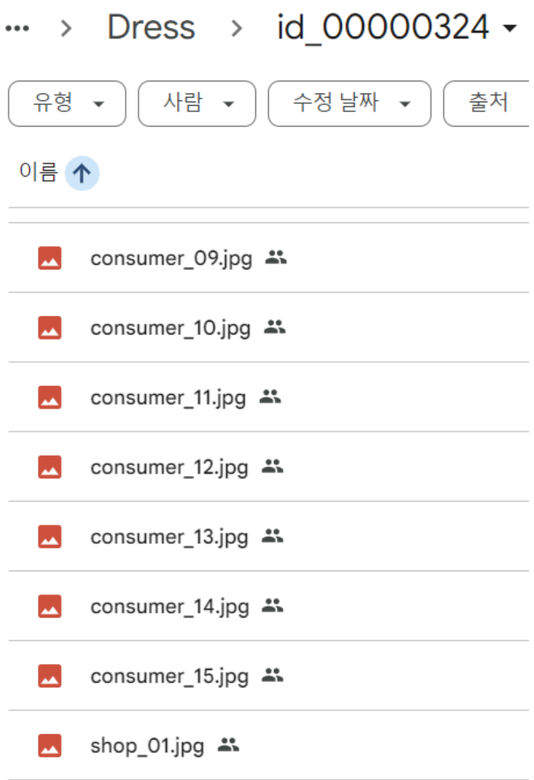
Problem : Some items have 1–2 images

- **Insufficient positive pairs**
- Model **failed to learn** meaningful embeddings

Solution : **Item ID–level sampling**

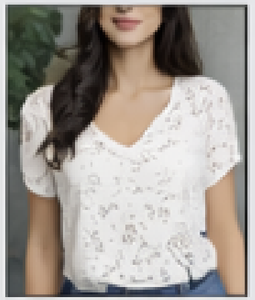
- Preserve intra-item variability

vs.

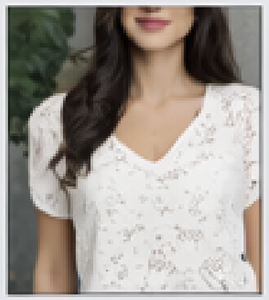


02-3 Preprocessing

Raw Image
(Consumer / Shop)



Bounding Box Crop
(Clothing Region)



Resize
(224 × 224)

To Tensor
+ Normalization

Train

- Horizontal Flip (p=0.5)
- Brightness, Contrast, Saturation Jitter

Val/Test

- Resize
- Normalize

Stage 1
Backbone and Loss function selection

03 Backbone and Loss function selection

Experimental Setup

1.Training Configuration

- Optimizer: Adam (learning rate = 1×10^{-4})
- Batch Size: 32
- Training Epochs: 40
- Early Stopping: Based on validation **Recall@5** (patience = 5)

2. Data Augmentation

- RandomHorizontalFlip (p = 0.5)
- ColorJitter (brightness = 0.2, contrast = 0.2, saturation = 0.2)

3. Embedding Settings

- Embedding Dimension: 128
- L2 normalization applied before distance computation

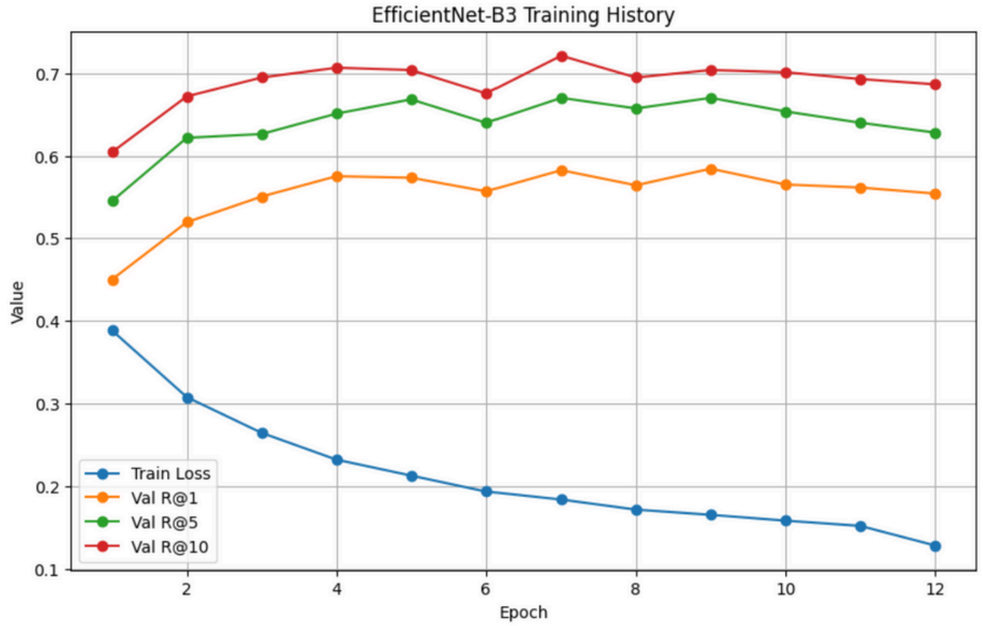
4. Evaluation

- Metric: Recall@K (K = 1, 5, 10)
- Fixed random seed = 42 for reproducibility

03-1
Backbone selection
(**fixing the loss function** as Semi-Hard Triplet Loss)

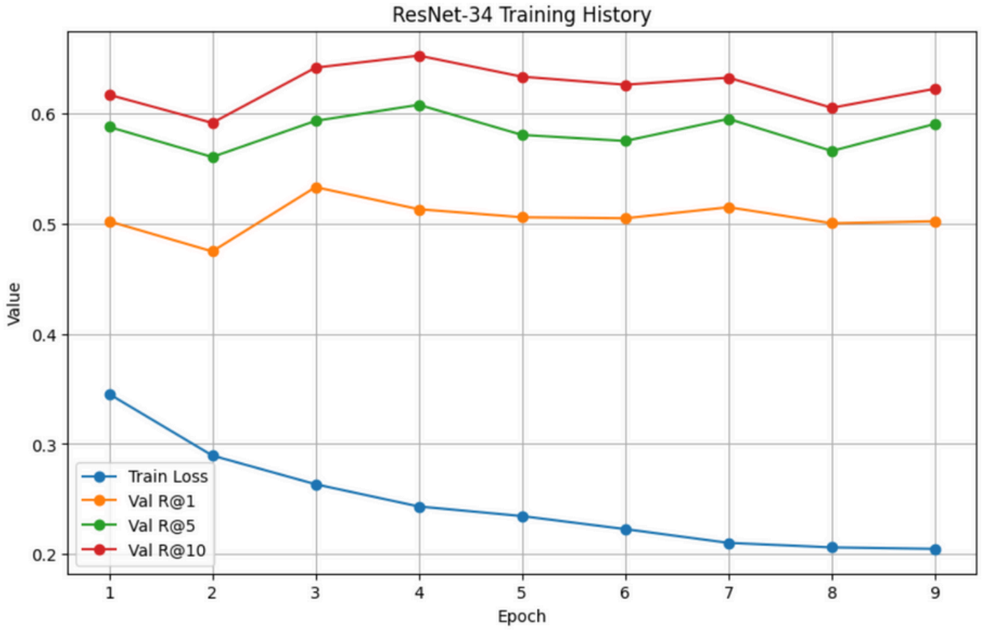
EfficientNet-B3

A structure that **maximizes efficiency** and is considered one of the latest State-of-the-Art (SOTA) models



ResNet-34

A relatively **light yet stable** model within the classical CNN structure of the ResNet family



03-1
Backbone selection
(**fixing the loss function** as Semi-Hard Triplet Loss)

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

Result

EfficientNet-B3 showed significantly higher performance.

Implication

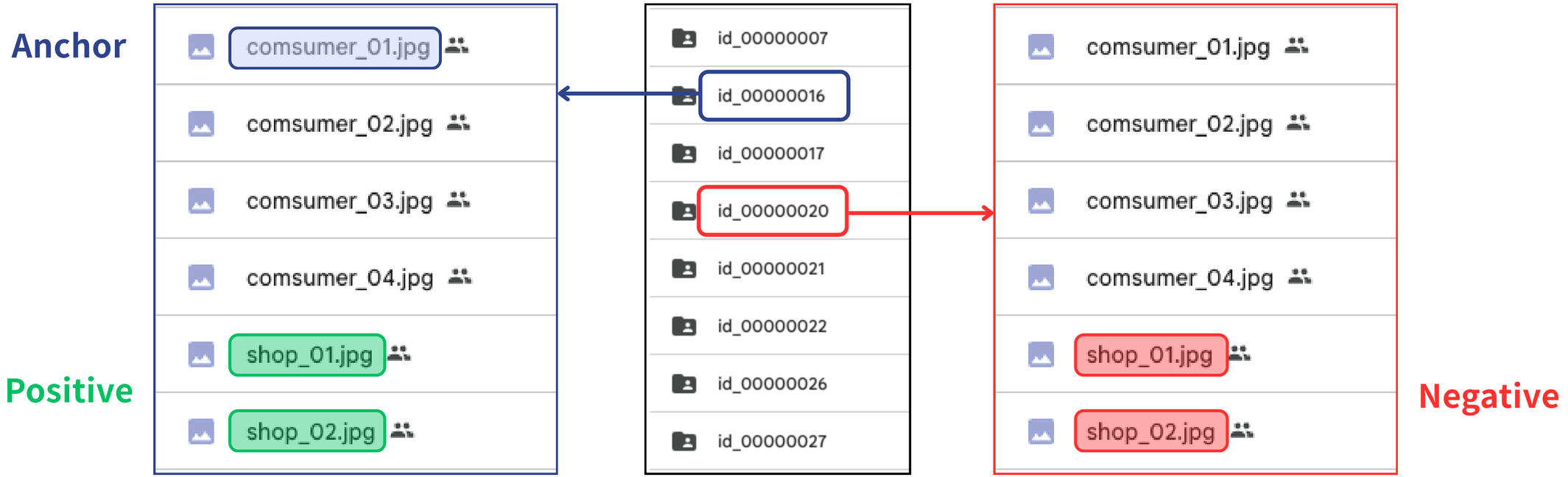
This suggests that EfficientNet-B3 learned **more appropriate feature embeddings** for the C2S task.

EfficientNet’s subsequent fine-grained feature extraction capabilities— achieved through **Compound Scaling** provide a substantial advantage over ResNet-34's traditional architecture.

03-2
Loss function selection

(**fixing the backbone** as EfficientNet-B3)

- **Metric Learning Loss Function** (e.g., Triplet, InfoNCE)
  - Anchor–Positive pairs → **closer**
  - Anchor–Negative pairs → **farther**



03-2
Loss function selection

(fixing the backbone as EfficientNet-B3)

Semi-Hard Triplet

Negative samples are chosen such that they are **farther than** positives but still within the margin boundary :

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

$$d(a,p) < d(a,n) < d(a,p) + m$$

- $d(\cdot)$: Euclidean distance
- m : margin (0.3)

Batch-Hard Triplet

For each anchor sample, the **hardest positive** and **hardest negative** samples are selected within a batch :

$$\mathcal{L}_{\text{BH}} = \max(0,\, d(a,p_{\text{hard}}) - d(a,n_{\text{hard}}) + m)$$

InfoNCE

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

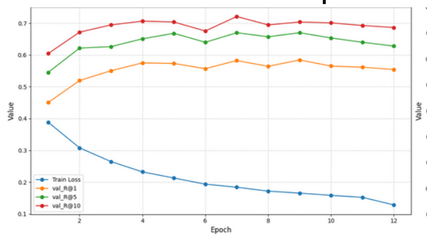
$$\mathcal{L}_{\text{NCE}} = -\log \frac{\exp(s_{\text{pos}}/\tau)}{\sum_j \exp(s_j/\tau)}$$

- s_j : cosine similarity
- τ : temperature (0.07)

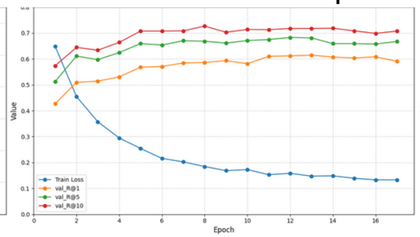
03-2
Loss function selection

Backbone	R@1	R@5	R@10
Semi-Hard Triplet	0.5910	0.6712	0.7092
Batch-Hard Triplet	**0.6095**	**0.6834**	**0.7171**
InfoNCE	0.5820	0.6645	0.6905

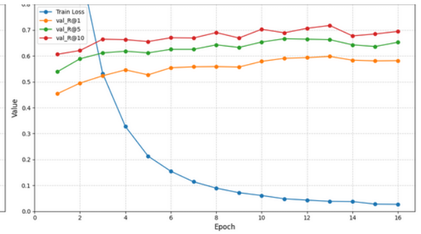
Semi-Hard Triplet



Batch-Hard Triplet



InfoNCE



Result

Batch-Hard Triplet achieved the highest retrieval performance.

Implication

This suggests that Batch-Hard Triplet demonstrate the **strongest discriminative capability**.

Semi-Hard Triplet Loss showed stable training, but produced slightly lower performance due to the limited use of hard negatives.

InfoNCE Loss showed underperformed compared to triplet-based methods,
likely because contrastive learning benefits from larger batch sizes, which were constrained in this setup (batch size = 32).

Stage 2
Advanced Preprocessing and Embedding Experiments

04
Advanced Preprocessing and Embedding Experiments

Purpose:
Optimize preprocessing and embedding design select the best overall configuration
(fixing the backbone as EfficientNet-B3 & loss function as Batch-Hard Triplet)

Data Augmentation

Bounding Box Cropping Margin

Domain Normalization between Consumer and Shop images

Embedding Dimension

04-1
Augmentation

Goal:

Analyze the impact of **data augmentation strength.**

Setting

Mode	Crop	Flip	Color Jitter	Erasing
Basline	Resize	0.5	0.2	X
Weak	Resize	0.5	0.1	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)

04-1
Augmentation

Goal:

Analyze the impact of **data augmentation strength.**

Setting

Mode	Crop	Flip	Color Jitter	Erasing
Basline	Resize	0.5	0.2	X
Weak	Resize	0.5	0.1	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)

Result

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

04-1
Augmentation

Goal:

Analyze the impact of **data augmentation strength.**

- No meaningful performance difference between baseline and strong
- Other augmentation modes resulted in slightly lower performance

→ **Baseline augmentation alone proved sufficient**

Preserving natural visual characteristics > **Stronger Augmentation**

Result

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

04-2
Bounding box cropping Margin

Goal:

Evaluate whether **additional contextual information around garments** improves retrieval performance

Setting

- Bounding box margin ratios tested:
  - 0.0 (Baseline)
  - 0.1
  - 0.2

Result

Mode	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

04-2
Bounding box cropping Margin

Goal:

Evaluate whether **additional contextual information around garments** improves retrieval performance

Tight Object Localization > **Additional contextual information**

may includes irrelevant visual elements
(background clutter, body parts, or occlusions)

Result

Mode	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

04-3
Lightweight Domain Normalization

Goal

- Evaluate the effectiveness of **reducing the domain gap** between consumer-shop images **through simple color and contrast normalization**

Experimental Setup

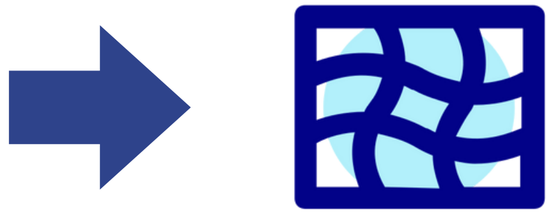
- Compared normalization strategies:
  - Baseline (no domain normalization)
  - CLAHE (contrast normalization)
  - Gray-World (color normalization)
  - Both (CLAHE + Gray-World)

04-3
Lightweight Domain Normalization

Aggressive contrast

&

Aggressive color correction



distortion

Negatively impact the model's ability to learn discriminative, garment-specific features

Result

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

04-4
Embedding Dimension

Goal

- Analyze the effect of embedding dimensionality on **representational capacity and generalization**

Experimental Setup

- Embedding dimensions compared:
  - 128 (Baseline)
  - 256
  - 512

Result

Mode	R@1	R@5	R@10
128 (Baseline)	0.6095	0.6834	0.7171
256	0.608	**0.6892**	**0.7238**
512	**0.6126**	0.6755	0.7119

04-4
Embedding Dimension

128

- Lacked the representational power **needed to compress the rich features** extracted by EfficientNet-B3.

512

- This drop is likely due to the **Curse of Dimensionality**

Result

Mode	R@1	R@5	R@10
128 (Baseline)	0.6095	0.6834	0.7171
256	0.608	**0.6892**	**0.7238**
512	**0.6126**	0.6755	0.7119

256-dimensional embedding yields the best retrieval performace

Final Model Evaluation on Test Set

05 Final Model Evaluation on Test Set

Final Model Configuration

- Backbone: **EfficientNet-B3**
- Loss function: **Batch-Hard Triplet Loss**
- Embedding: L2-normalized
- Embedding Dimension: **256**

Test Set Evaluation

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

Conclusion

06 Conclusion

Experimental Focus

- Built a metric learning–based C2S image retrieval framework
- Fixed EfficientNet-B3 and Batch-Hard Triplet Loss, and analyzed the impact of preprocessing and embedding design

Key Findings

- 256-d embedding achieves the best balance between representation and generalization
- Retrieval performance improves without modifying the backbone

Future Work

- Perform systematic and automated exploration of preprocessing and embedding configurations
- Scale experiments to the full C2S dataset and broader fashion categories
- Investigate advanced metric learning losses (e.g., ArcFace, Circle Loss, ProxyNCA++)

Live Demo

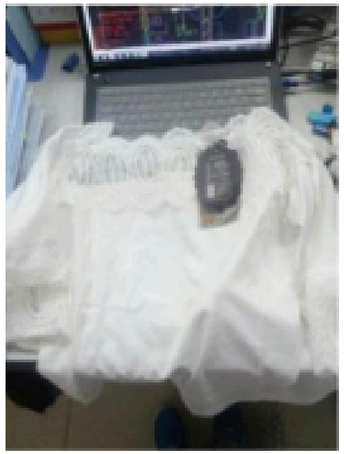
Run Demo

Demo

```
Query: img/TOPS/Lace_Shirt/id_00025743/consumer_01.jpg

Top-K unique shop paths:
Top-1: img/TOPS/Lace_Shirt/id_00025743/shop_01.jpg (sim=0.5745)
Top-2: img/TOPS/Summer_Wear/id_00029049/shop_02.jpg (sim=0.5236)
Top-3: img/TOPS/Summer_Wear/id_00029049/shop_01.jpg (sim=0.4480)
Top-4: img/DRESSES/Dress/id_00023791/shop_01.jpg (sim=0.3418)
Top-5: img/TOPS/Blouse/id_00010960/shop_01.jpg (sim=0.2383)
```

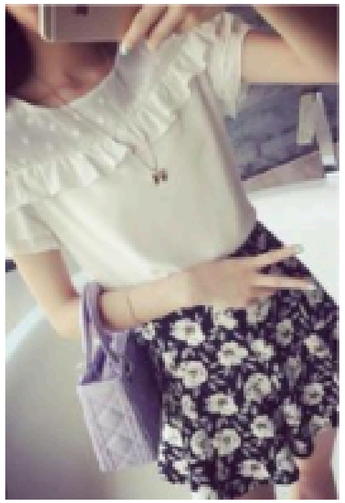
Query (Consumer)



Top-1

Top-2

Top-3



Top-4

Top-5

Same clothing

Similar clothings with Lace detail

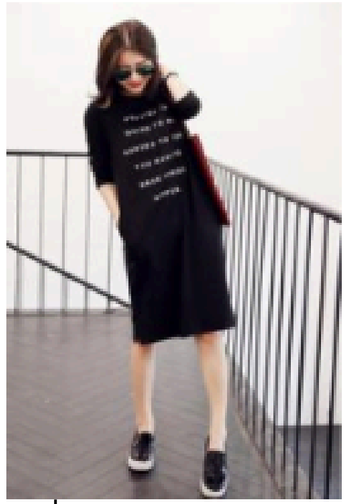
Demo

```
Query: img/DRESSES/Dress/id_00008175/consumer_01.jpg

Top-K unique shop paths:
Top-1: img/DRESSES/Dress/id_00008175/shop_01.jpg (sim=0.6954)
Top-2: img/DRESSES/Dress/id_00008175/shop_02.jpg (sim=0.6363)
Top-3: img/TOPS/T_Shirt/id_00008306/shop_01.jpg (sim=0.4904)
Top-4: img/TOPS/Summer_Wear/id_00007478/shop_01.jpg (sim=0.3989)
Top-5: img/TOPS/Summer_Wear/id_00005352/shop_01.jpg (sim=0.3777)
```

Query (Consumer)

Top-1



Top-2

Top-3

Top-4

Top-5

Same clothing

Successfully retrieved same clothing in **different colors**.

THANK YOU