# Learning More from Less: Weak Supervision and Beyond

**Rogerio Schmidt Feris**

Principal Research Scientist and Manager

IBM Research & MIT-IBM Watson AI Lab

# The battle against the long tail

- Training accurate deep neural network models usually requires lots of labeled data

- Data collection and annotation is expensive, tedious, time-consuming.
- Crowdsourcing may be infeasible for proprietary data.
- For some tasks, data may not be available at all (long tail distribution)

# This talk

- Weak supervised learning for fashion search
- Learning with less labels beyond weak supervision

# Street2Shop

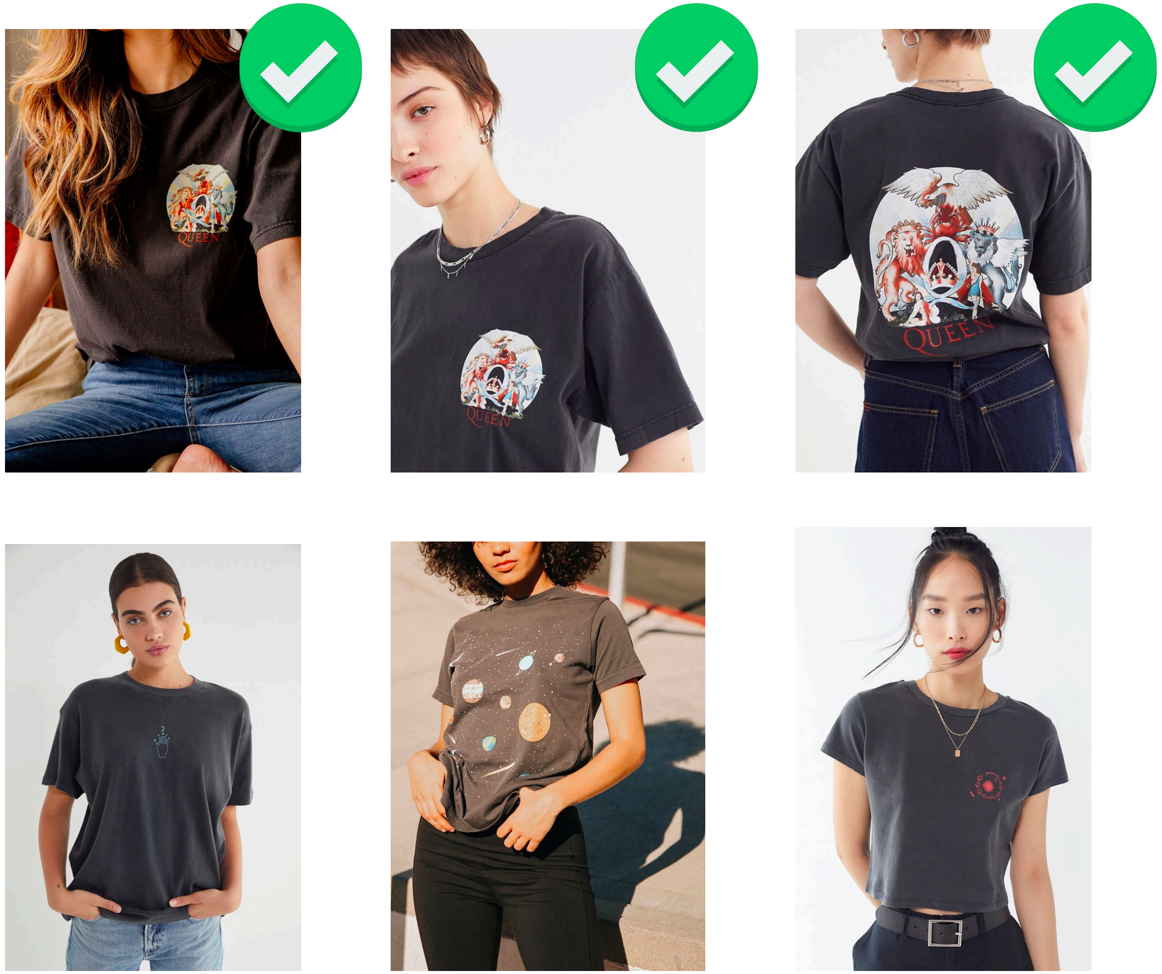

Hadi Kiapour, M., Han, X., Lazebnik, S., Berg, A. C., & Berg, T. L. Where to buy it: Matching street clothing photos in online shops. ICCV 2015

Slide credit: Tamara Berg

# Street2Shop Clothing Retrieval

Input: User Photo

Retrieved Images from Online Shopping Stores




[Liu et al, CVPR 2012] [Kiapour et al, ICCV 2015] [Huang et al, ICCV 2015]

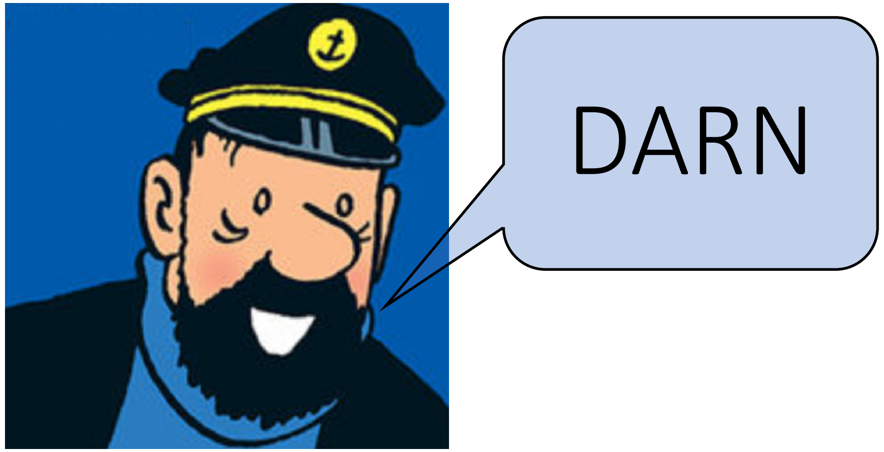
# Problem: Domain Discrepancy

Shopping Catalog

User Photo


## Proposed Approach:

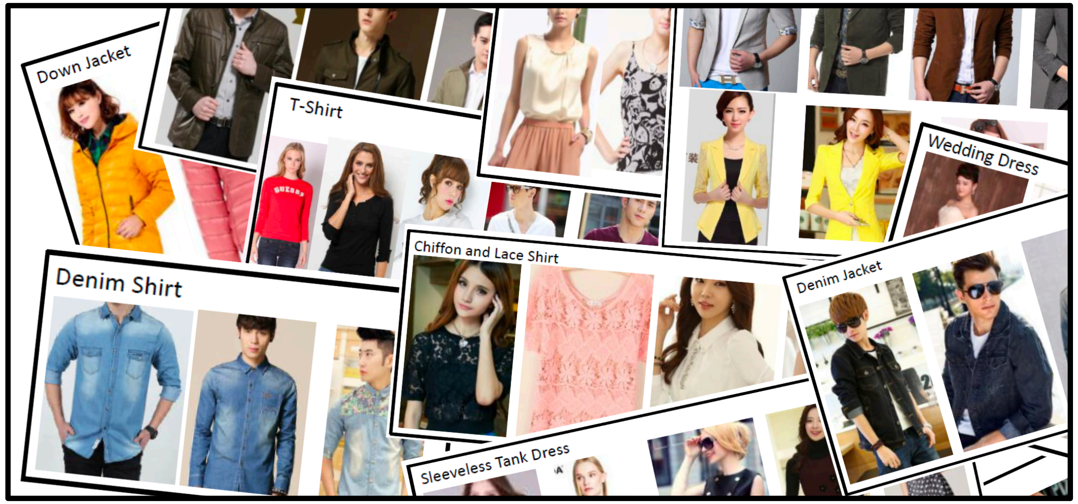
Dual Attribute-Aware Ranking Network (DARN)



# Weakly labeled data from shopping websites

- 9,000 image pairs mined from customer review websites (exact same clothing)

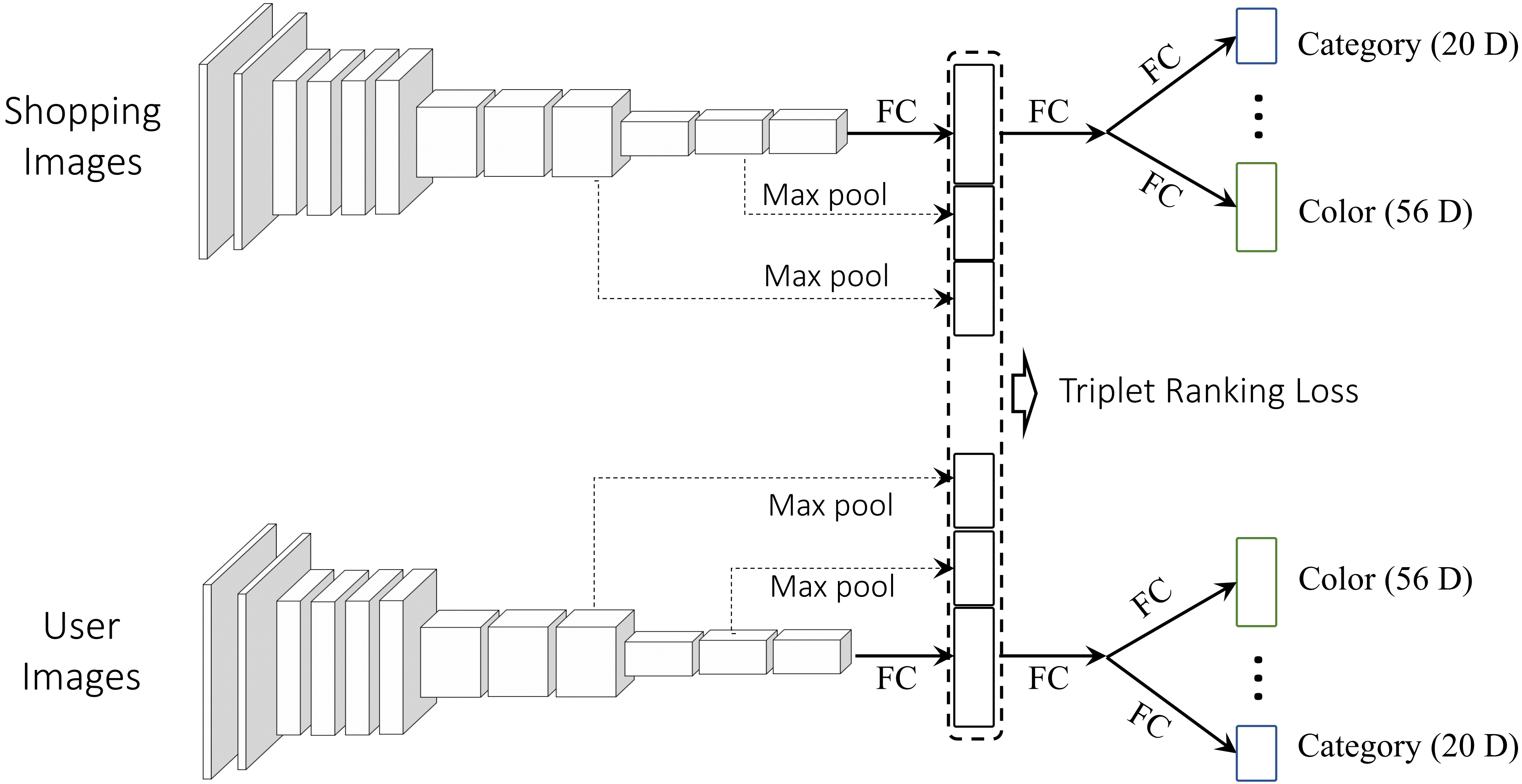
- Noisy attribute labels mined from online shopping stores (9 classes, 179 values)




| Attribute categories | Examples (total number) |
|---|---|
| Clothes Button | Double Breasted, Pullover, ... (12) |
| Clothes Category | T-shirt, Skirt, Leather Coat ... (20) |
| Clothes Color | Black, White, Red, Blue ... (56) |
| Clothes Length | Regular, Long, Short ... (6) |
| Clothes Pattern | Pure, Stripe, Lattice, Dot ... (27) |
| Clothes Shape | Slim, Straight, Cloak, Loose ... (10) |
| Collar Shape | Round, Lapel, V-Neck ... (25) |
| Sleeve Length | Long, Three-quarter, Sleeveless ... (7) |
| Sleeve Shape | Puff, Raglan, Petal, Pile ... (16) |

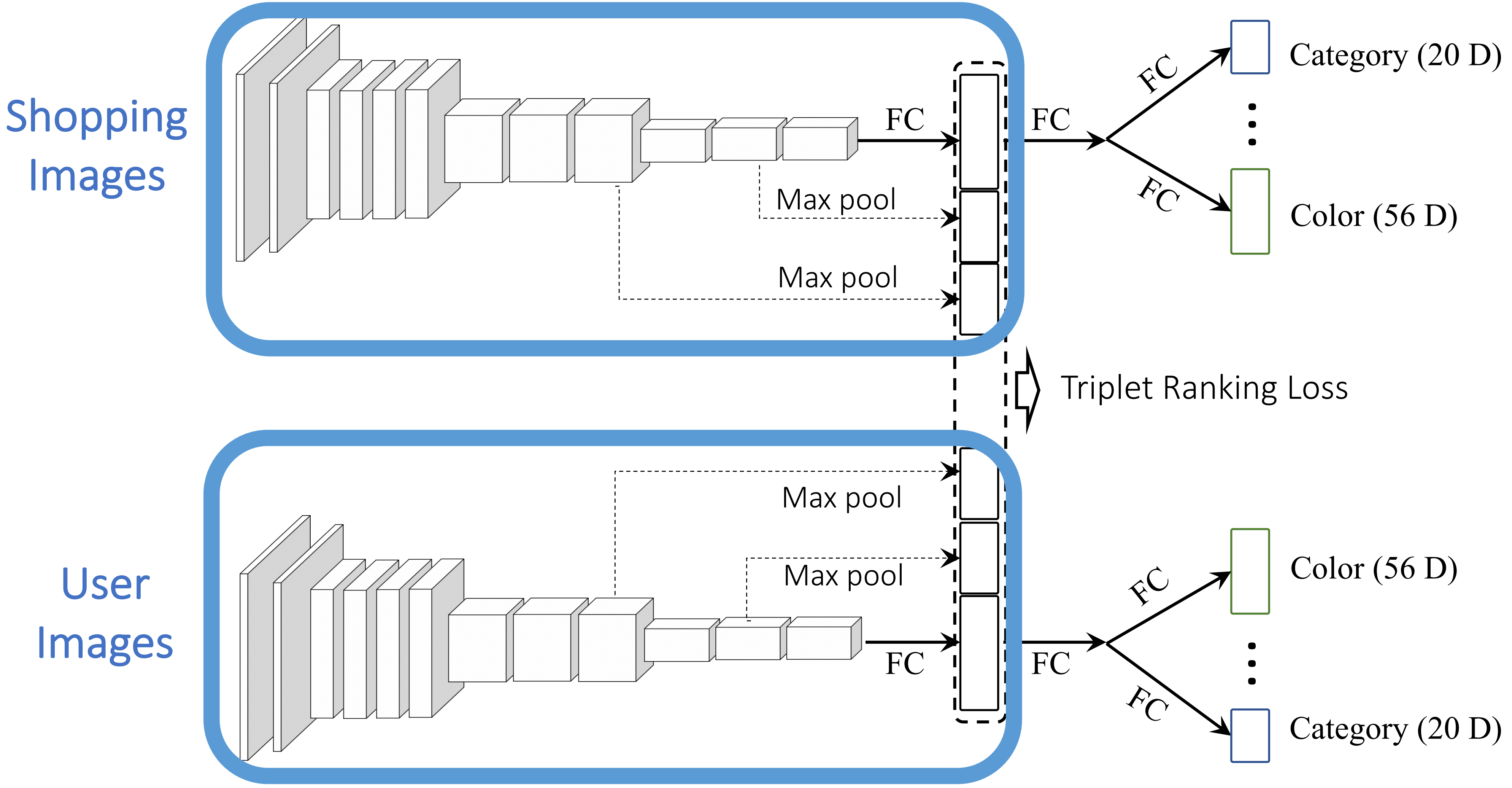
# Dual Attribute-Aware Ranking Network (DARN)

- Two sub-networks to model each domain (shopping and user images)



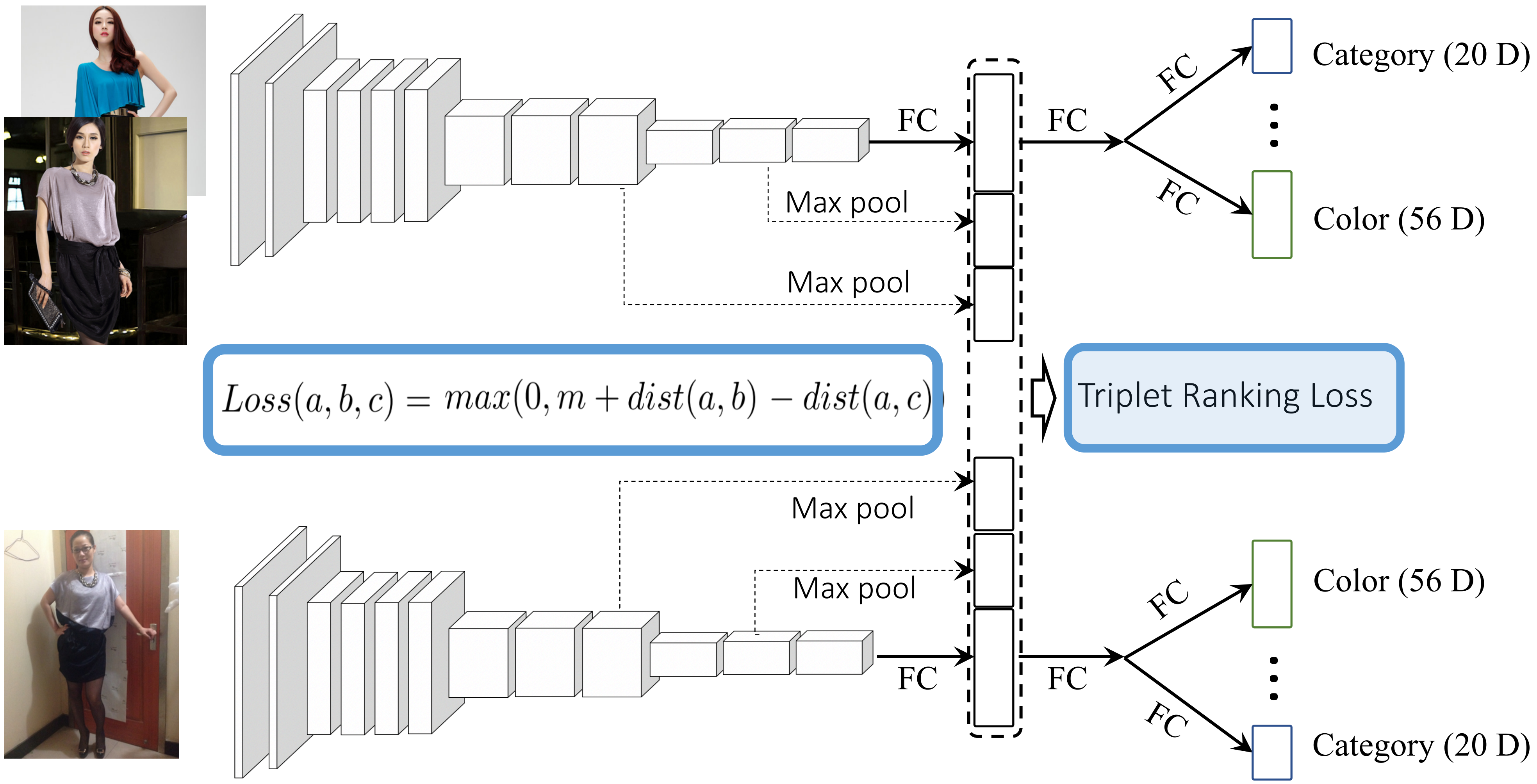
# Dual Attribute-Aware Ranking Network (DARN)

- Two sub-networks to model each domain (shopping and user images)



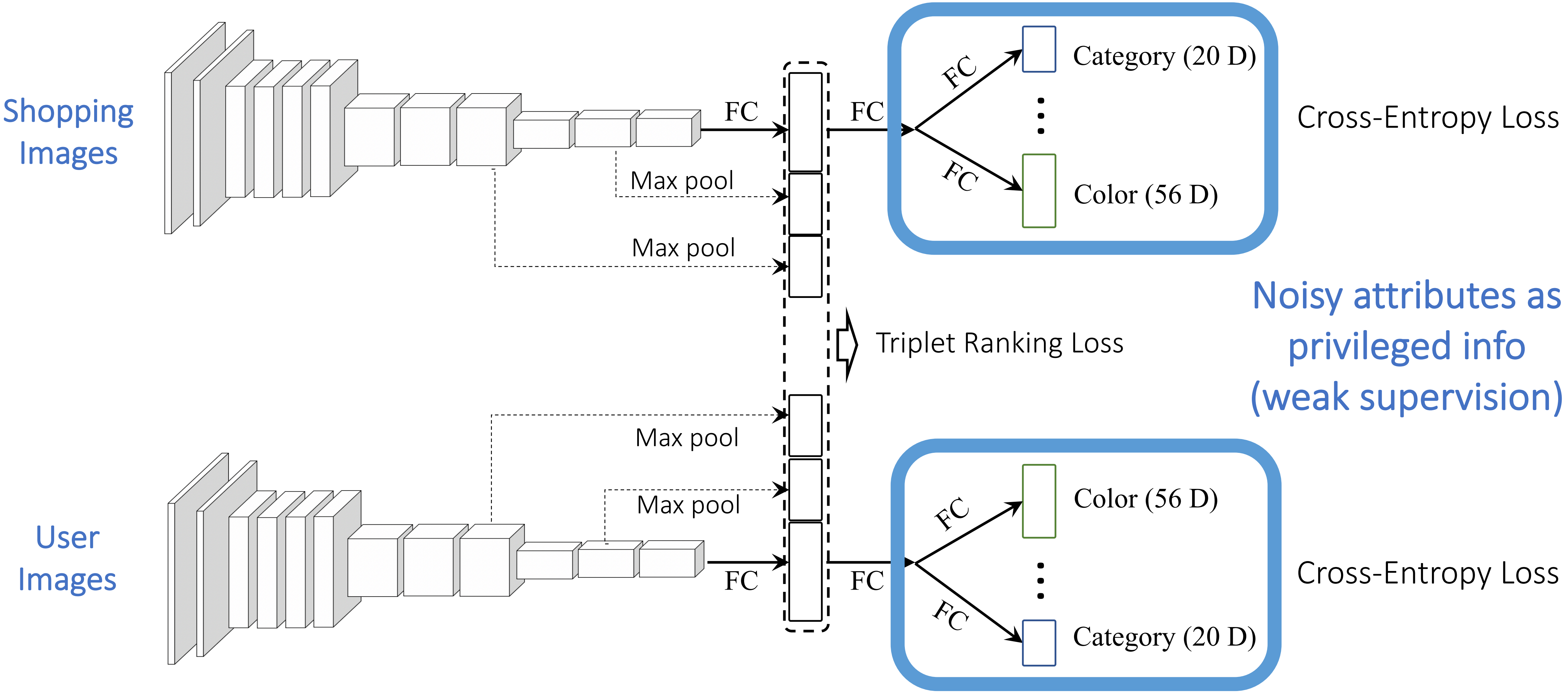
# Dual Attribute-Aware Ranking Network (DARN)

- Triplet Ranking loss function connecting the two sub-networks
- (visual similarity constraint)



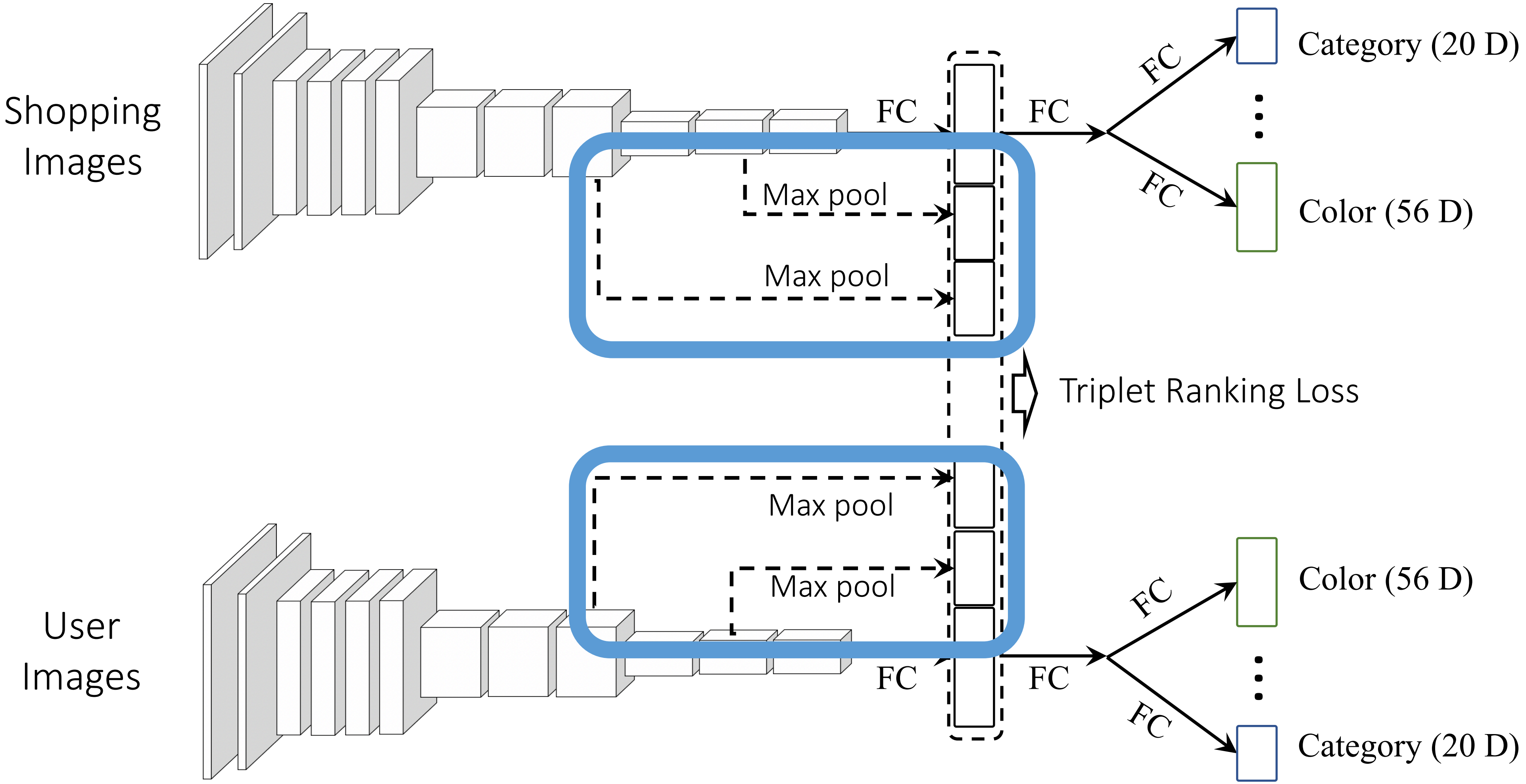
# Dual Attribute-Aware Ranking Network (DARN)

- Semantic embedding: simultaneous attribute learning and retrieval
- FC features are transmitted to multiple branches



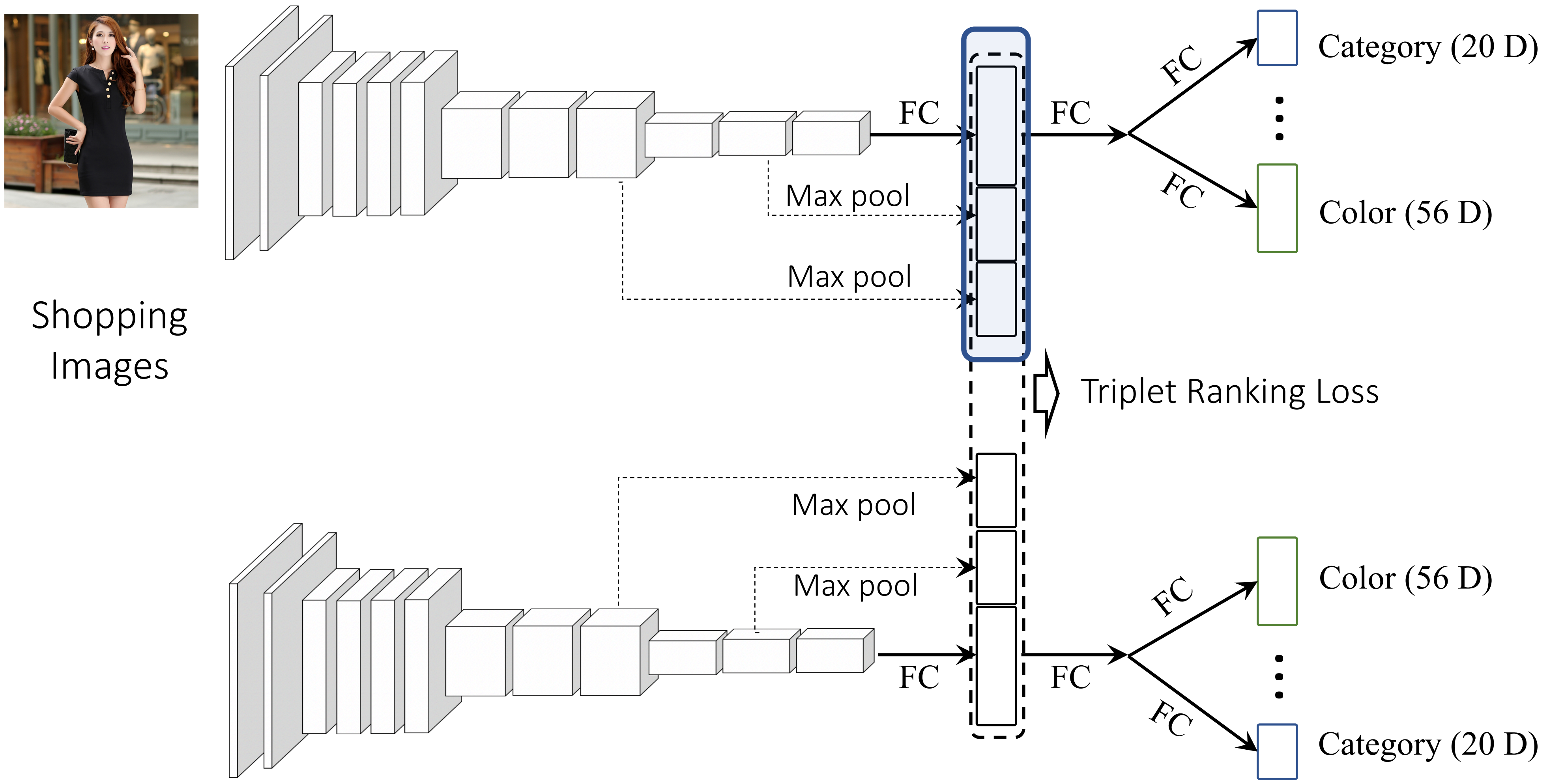
# Dual Attribute-Aware Ranking Network (DARN)

- Features from conv layers for encoding more localized information



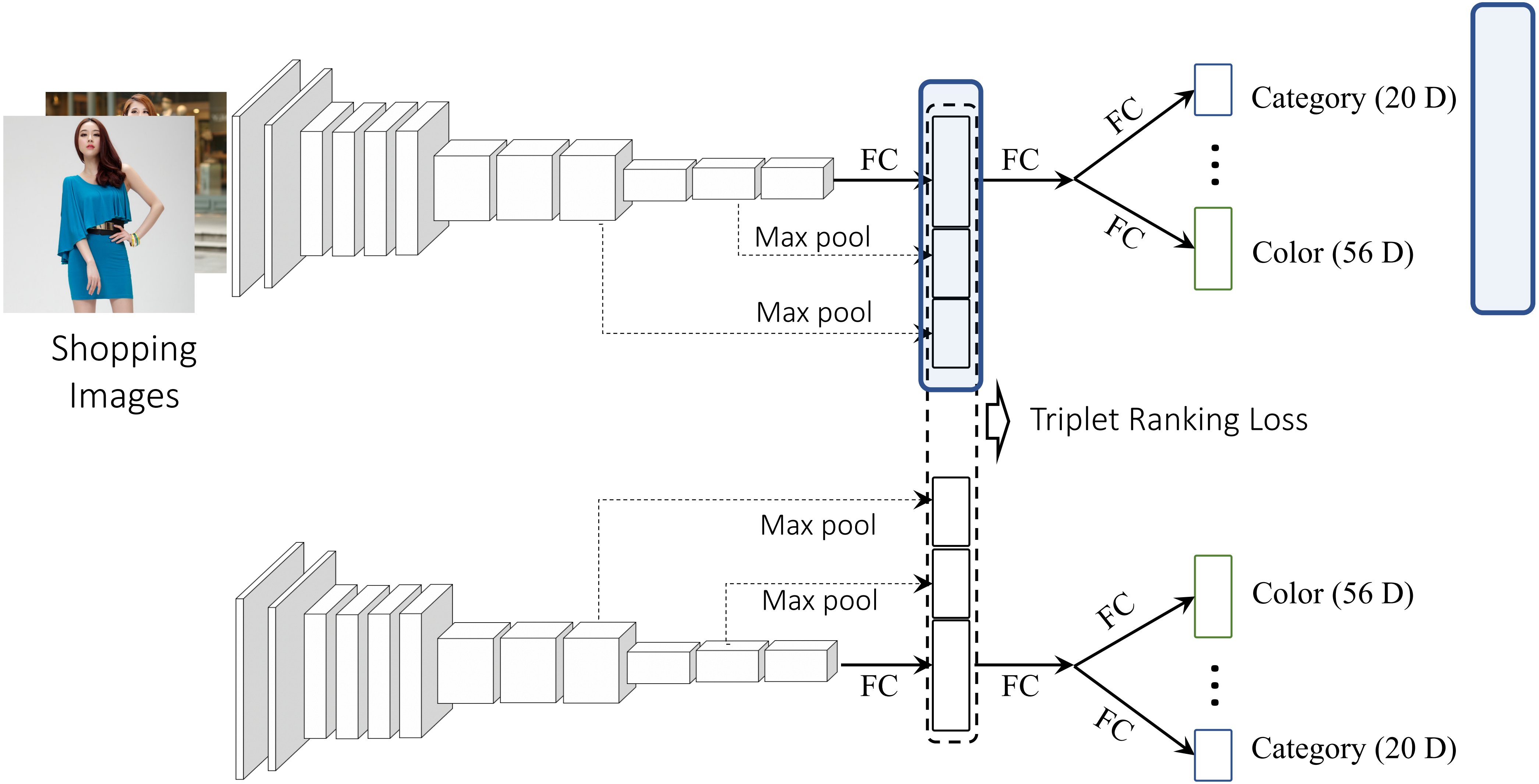
# Dual Attribute-Aware Ranking Network (DARN)

- <u>**Test time:**</u> Cross-domain Clothing Retrieval
- For each image in the gallery, compute features and store them in a database



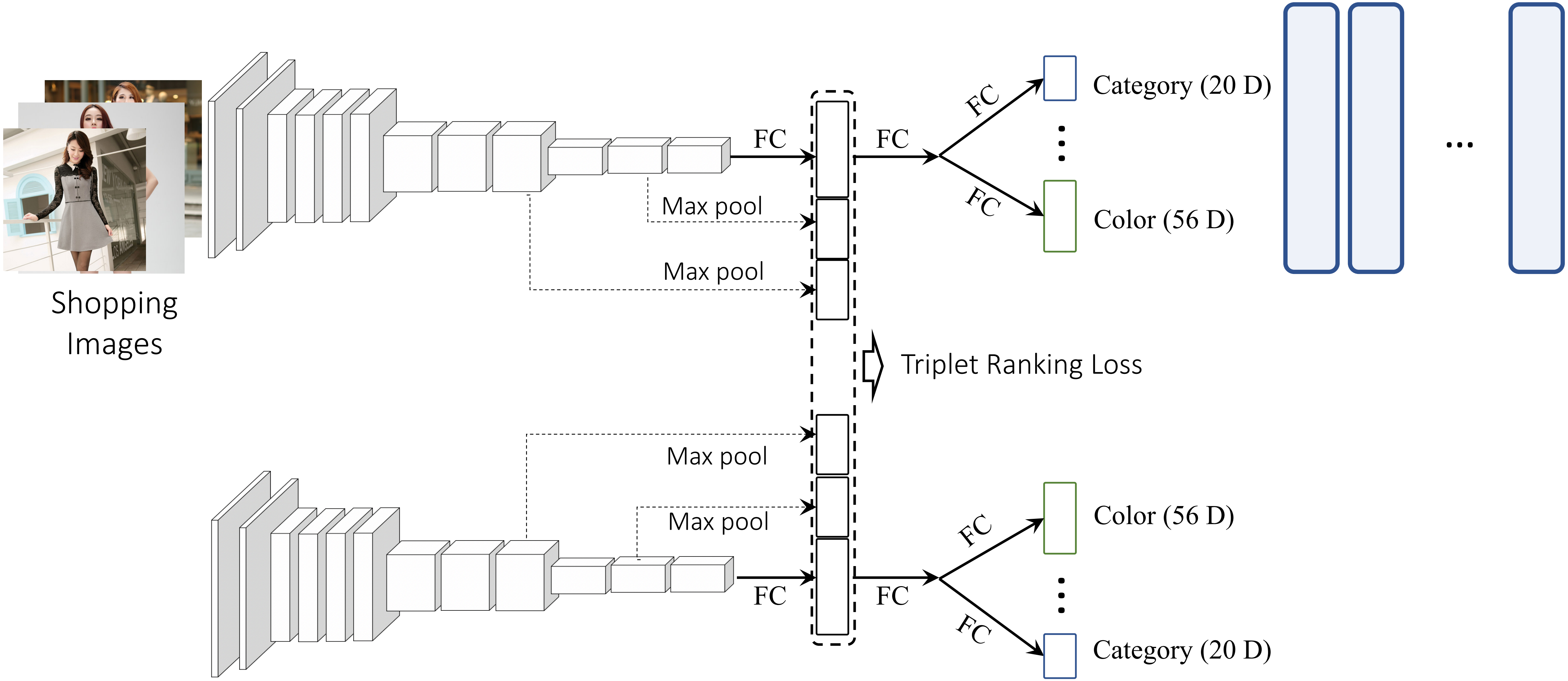
# Dual Attribute-Aware Ranking Network (DARN)

- **Test time:** Cross-domain Clothing Retrieval
- For each image in the gallery, compute features and store them in a database



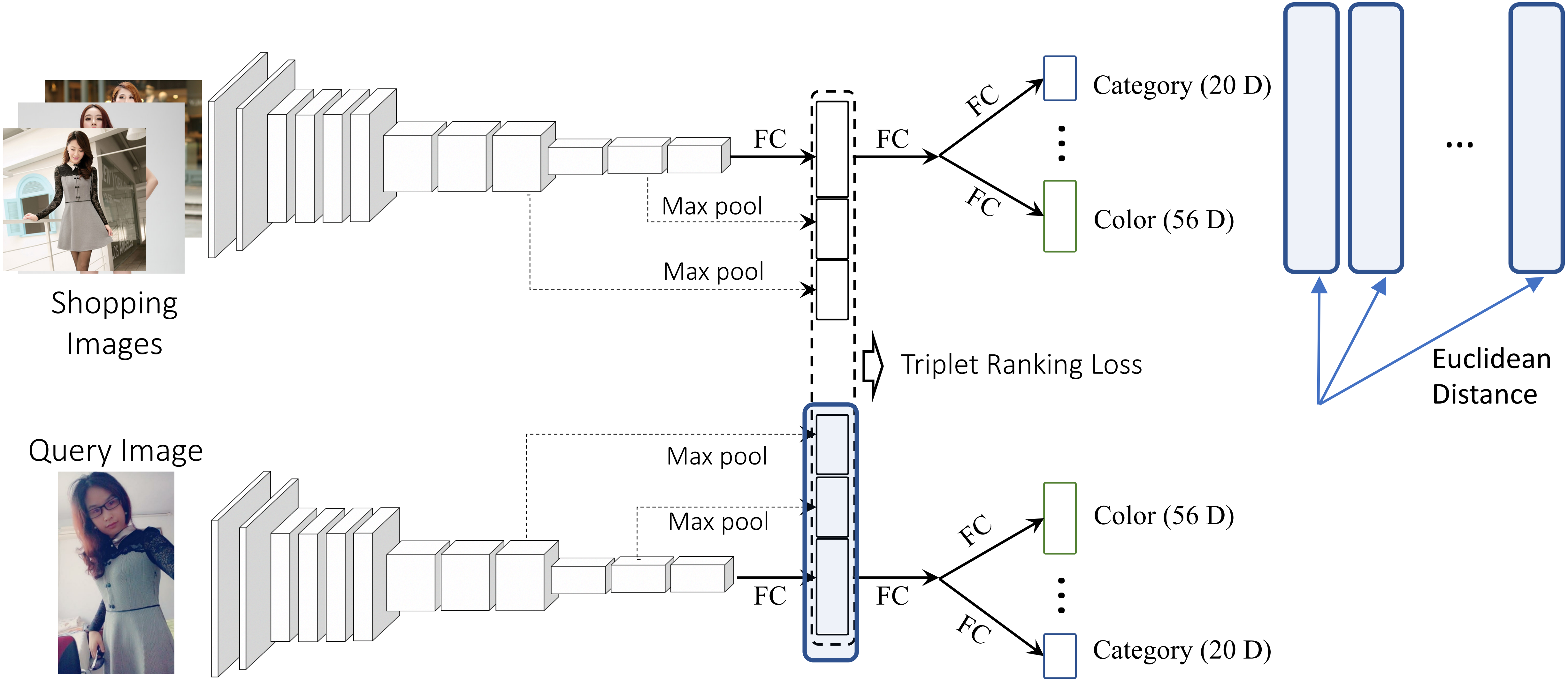
# Dual Attribute-Aware Ranking Network (DARN)

- **Test time:** Cross-domain Clothing Retrieval
- For each image in the gallery, compute features and store them in a database



# Dual Attribute-Aware Ranking Network (DARN)

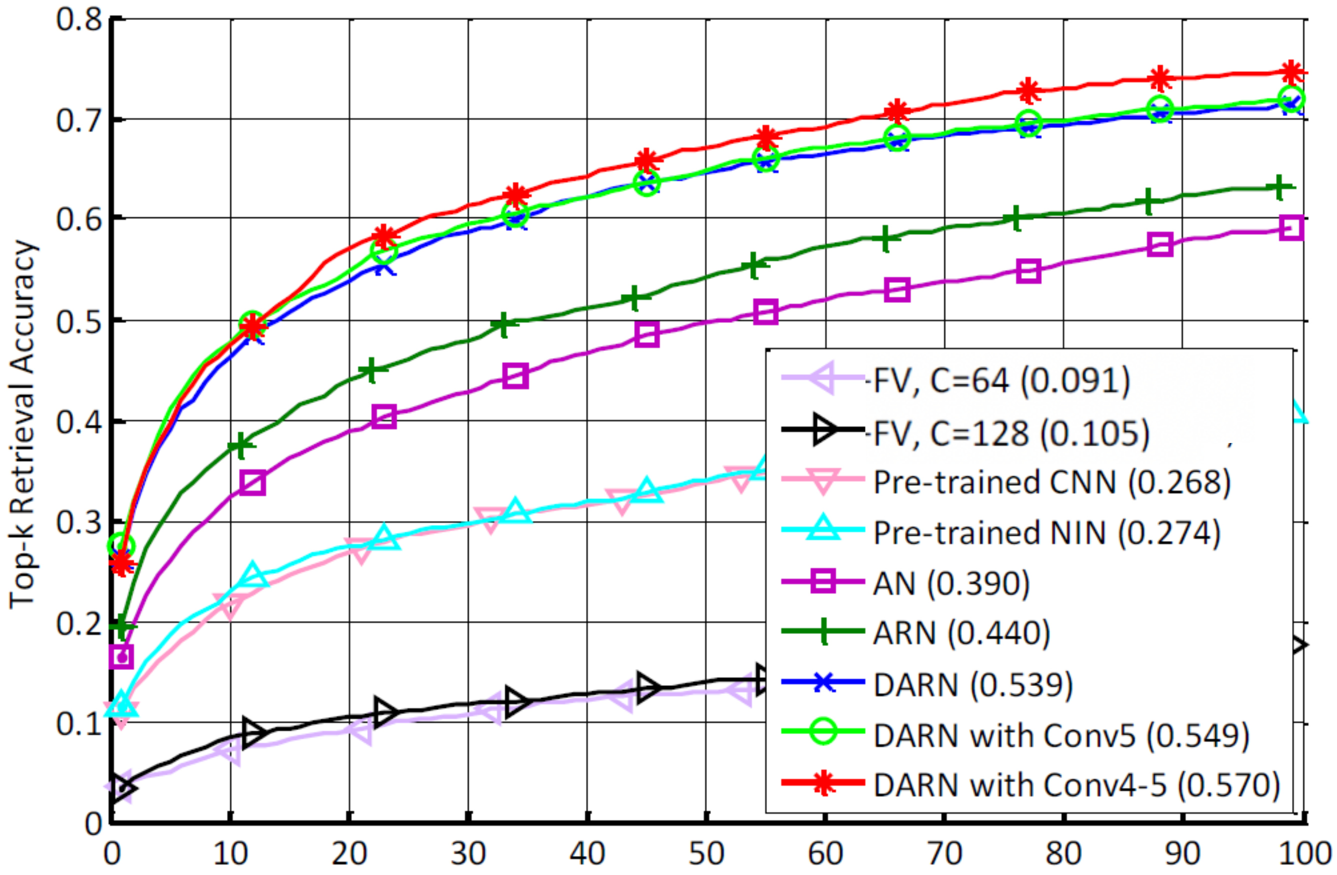
- <u>**Test time:**</u> Cross-domain Clothing Retrieval
- Given a query image, compute features and rank-order the gallery based on Euclidean distance



# Experimental Results

Our method (DARN) achieves the best results compared to other state-of-the-art approaches.

Top-k retrieval accuracy on 200,000 retrieval gallery.
The number in the parentheses is the top-20 retrieval accuracy.



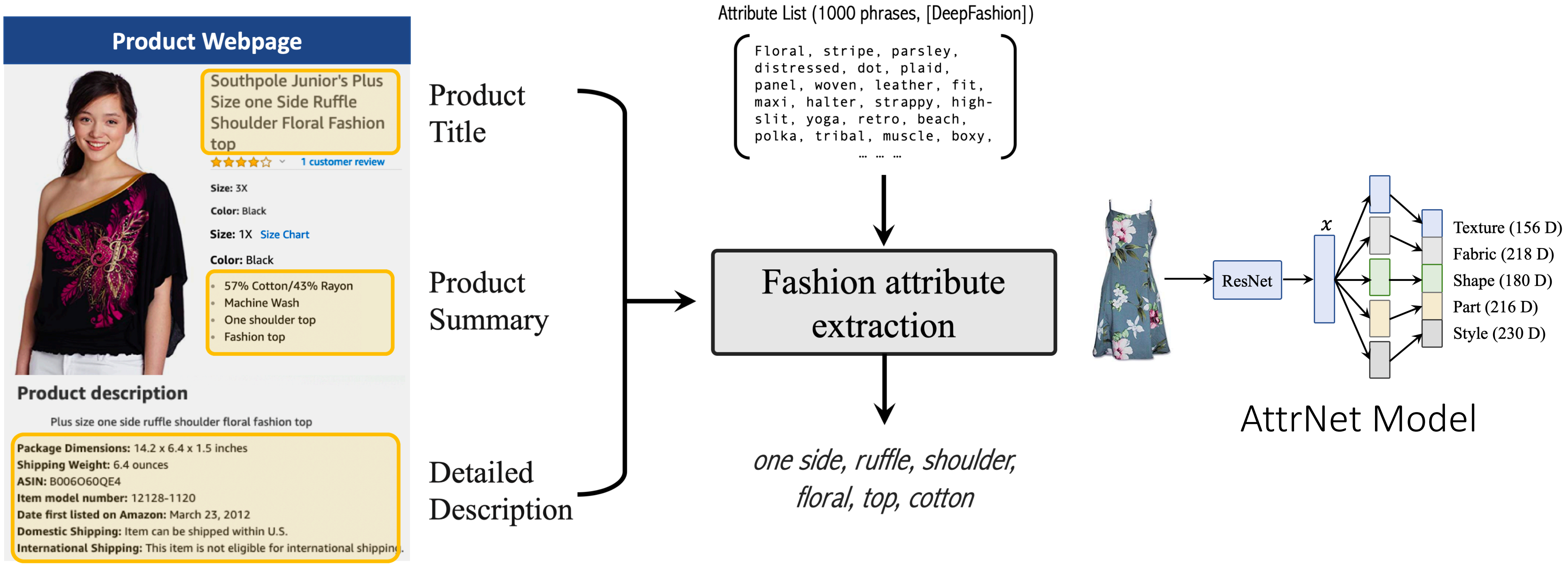

First Column: Query
Green Box: Exact same clothing

# FASHION IQ DEMO

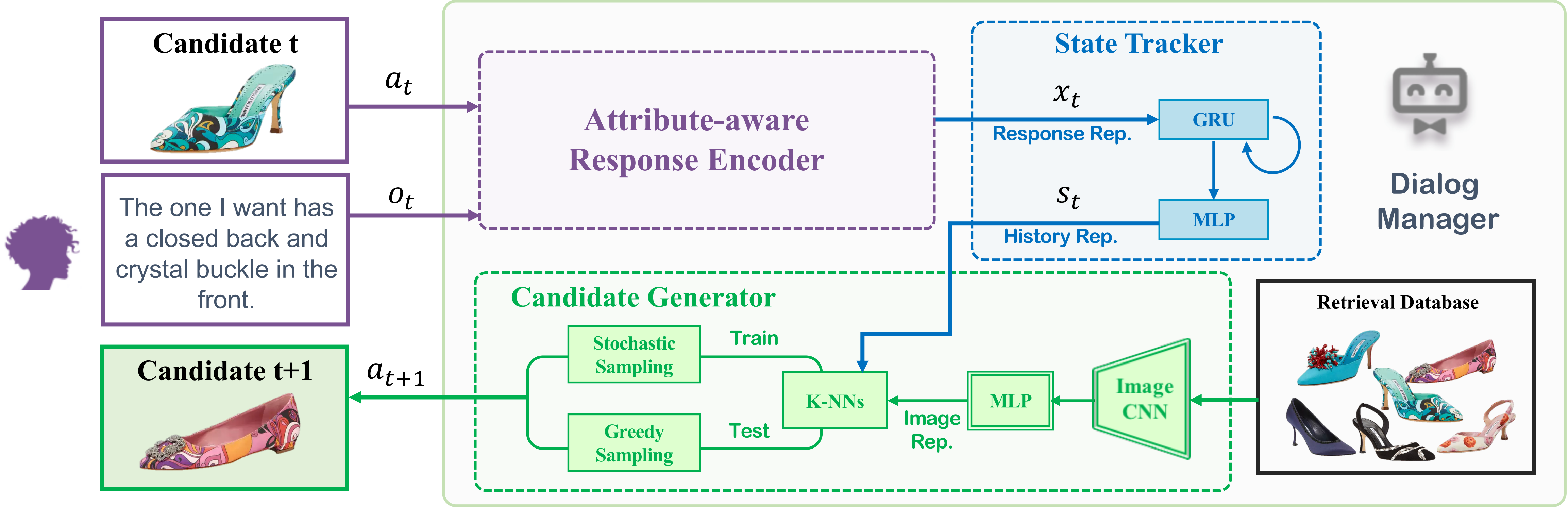
IBM RESEARCH AI

# Attributes as a weak supervisory signal

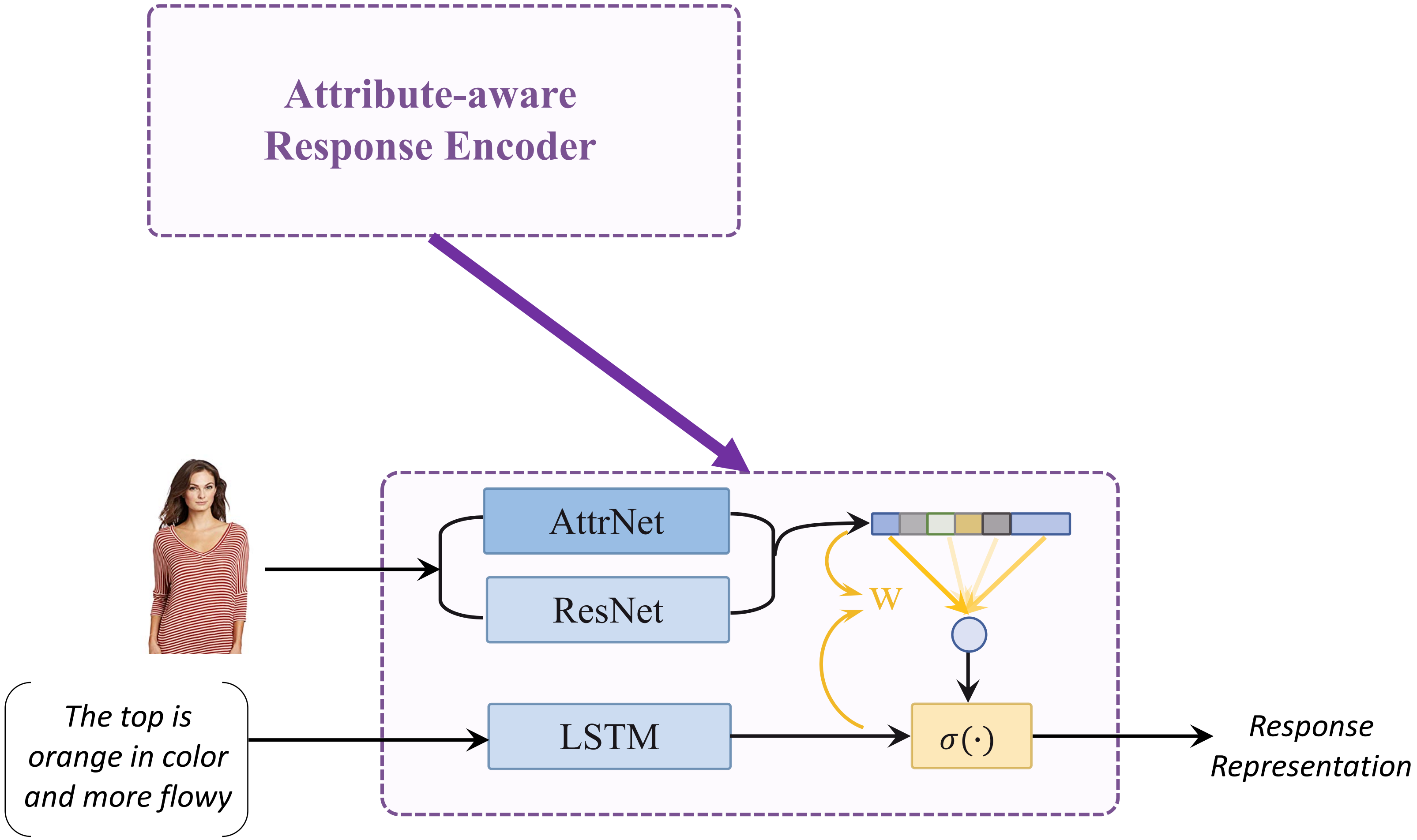
- Mining attributes from text surrounding the images



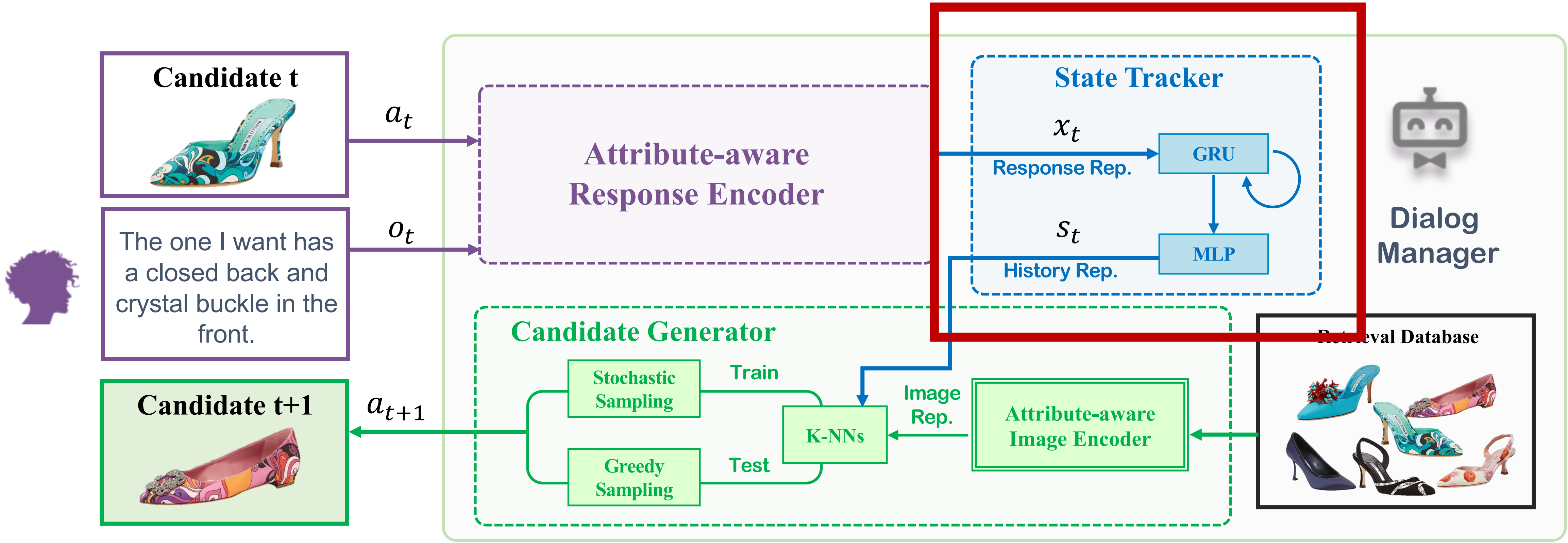
# Network Architecture



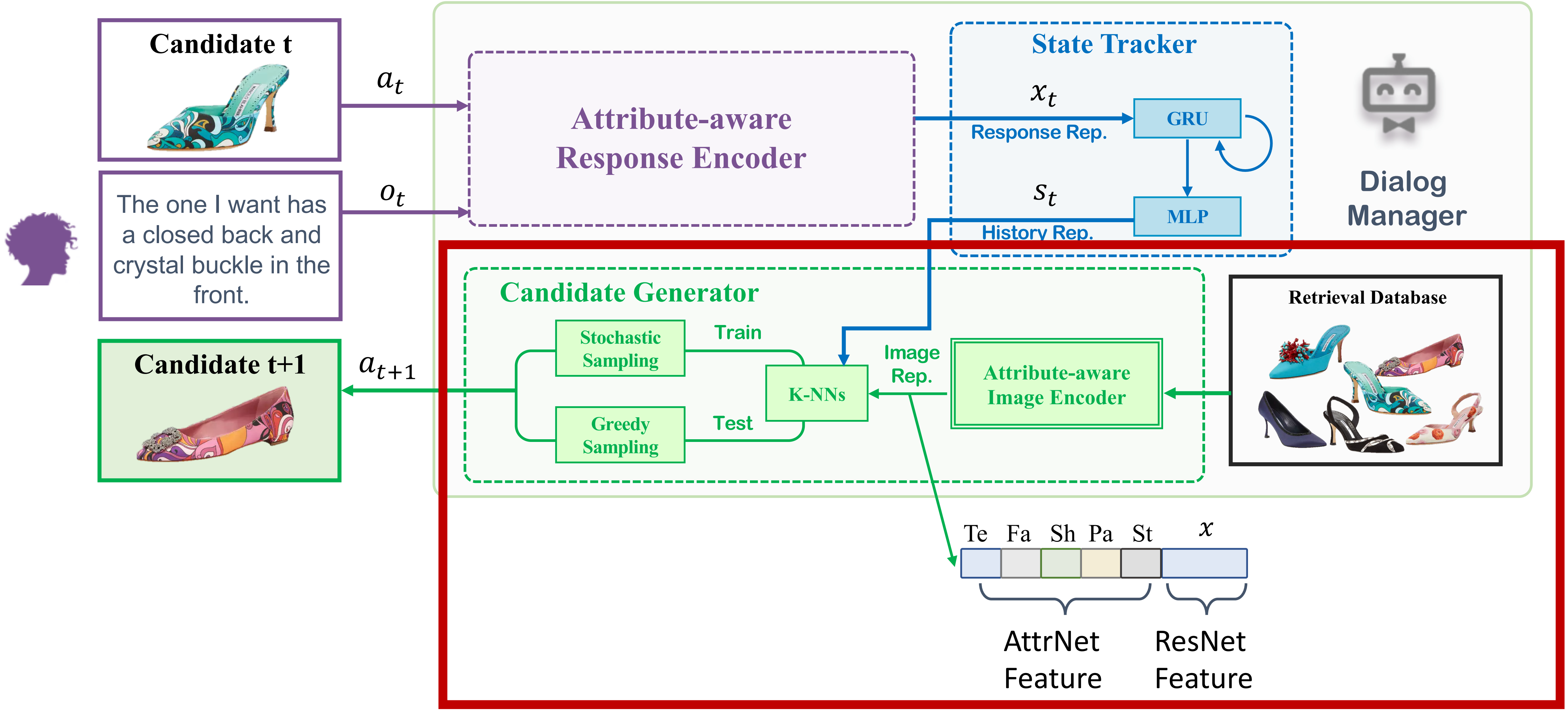
# Network Architecture



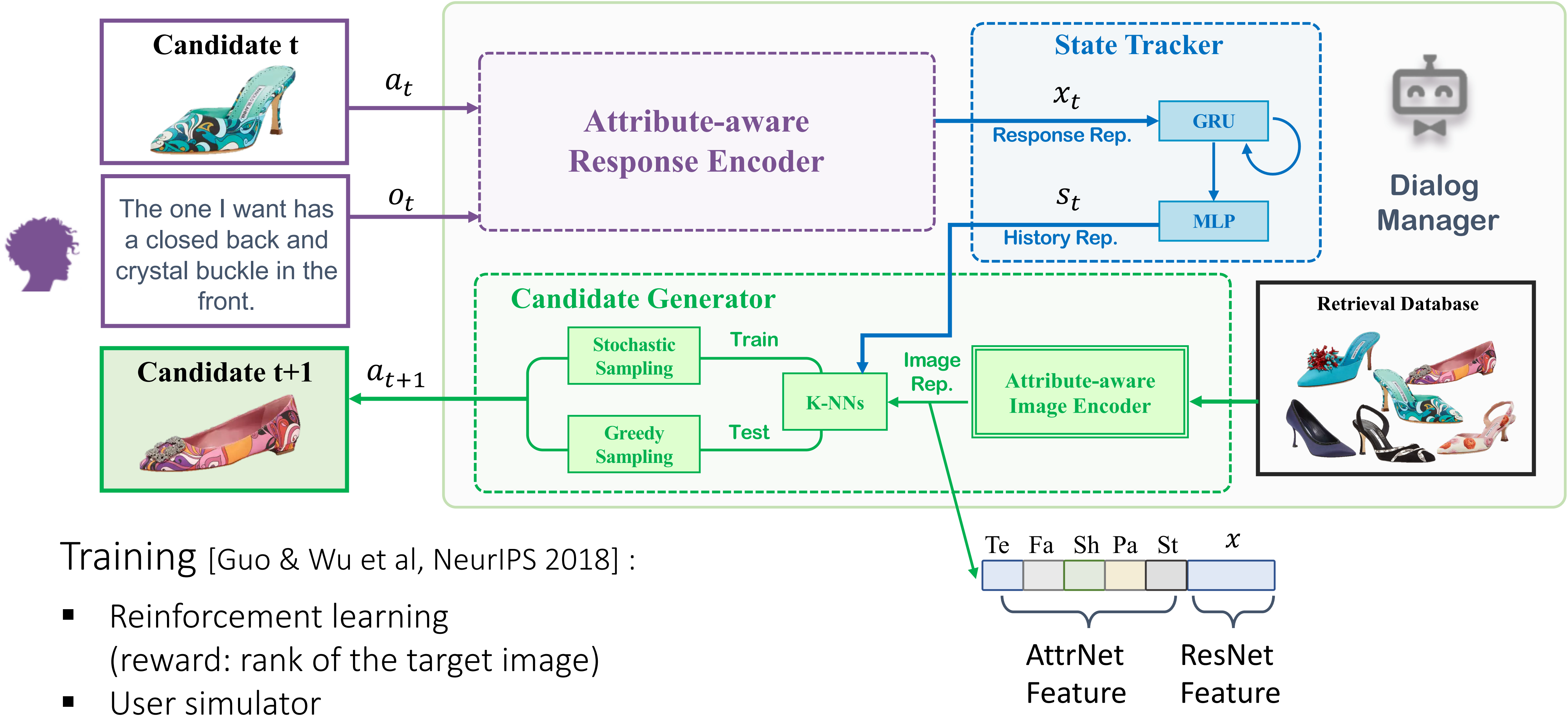
# Network Architecture



# Network Architecture



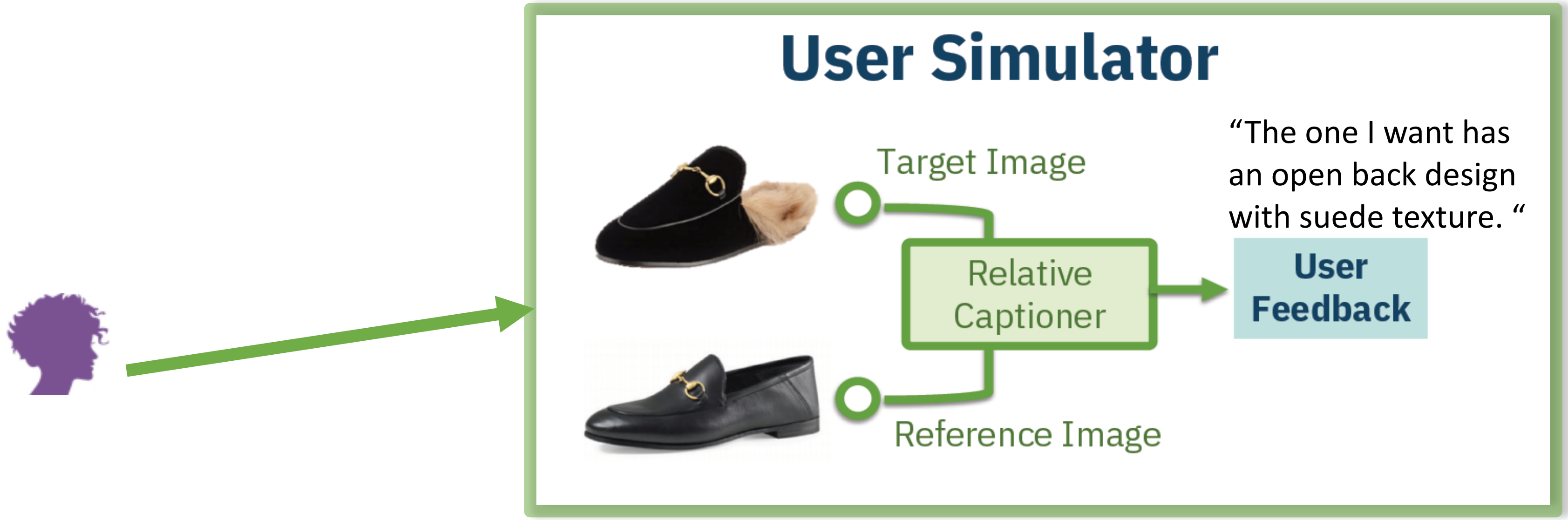
# Network Architecture




Training [Guo & Wu et al, NeurIPS 2018] :

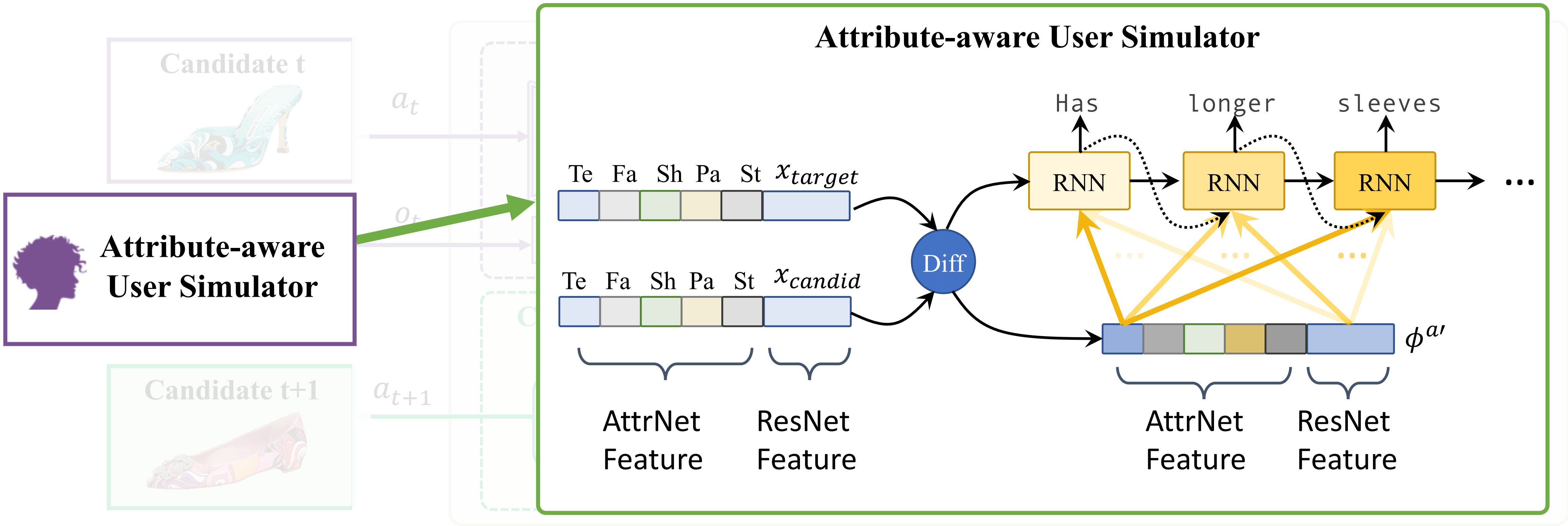
- Reinforcement learning
  (reward: rank of the target image)
- User simulator

# Training Dialog Manager with User Simulator




- Relative captioner: surrogate for real users
  - Automatically generates sentences describing the visual differences between target and reference images
  - New task and new dataset!

# Network Architecture



# Fashion IQ Dataset

https://www.spacewu.com/posts/fashion-iq/

- Images sourced from Amazon, including three classes, Dresses, Tops & Tees, and Shirts (~60K relative captions)

| | **Dresses** | | **Tops&Tees** | | **Shirts** | |
|---|---|---|---|---|---|---|
| | train / val / test | total | train / val / test | total | train / val / test | total |
| # Images | 11452 / 3817 / 3818 | 19087 | 16121 / 5374 / 5374 | 26869 | 19036 / 6346 / 6346 | 31728 |
| # Images with side info | 7741 / 2561 / 2653 | 12955 | 9925 / 3303 / 3210 | 16438 | 12062 / 4014 / 3995 | 20071 |
| # Relative Captions | 11970 / 4034 / 4048 | 20052 | 12054 / 3924 / 4112 | 20090 | 11976 / 4076 / 4078 | 20130 |

Relative Captions:
*"no sleeve flapping blouse"*
*"it has no sleeves and it is plain"*

Attribute Labels
*ruffle, wash, fit*

Relative Captions:
*" has a blue collar"*
*"has a blue color"*

Attribute Labels
*cotton, twill, wash, button-front, single-button*

Relative Captions:
*"is white with a black belt"*
*"is lighter in color"*

Attribute Labels
*stripe, cotton, gauze, tiered, wash, tube, braided*

# Results – Attribute-aware User Simulator

| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | Meteor | Rouge-L | CIDEr | SPICE |
|---|---|---|---|---|---|---|---|---|
| Attribute-aware (D) | **61.3** | **44.1** | **29.0** | **19.7** | **26.2** | **55.5** | **59.4** | **34.7** |
| with Attention (S) | **57.7** | **46.3** | **32.9** | **22.3** | **27.9** | **57.1** | **78.8** | **36.6** |
| (T) | **58.4** | **44.1** | **29.6** | **20.3** | **26.5** | **54.1** | **63.3** | **35.3** |
| Attribute-aware (D) | 58.5 | 42.0 | 26.7 | 17.5 | 24.0 | 53.2 | 42.7 | 30.8 |
| via Concatenation (S) | 54.5 | 42.6 | 29.1 | 19.4 | 25.8 | 53.5 | 47.1 | 31.8 |
| (T) | 55.9 | 41.0 | 26.0 | 17.0 | 25.4 | 51.5 | 40.7 | 31.1 |
| Image-Only (D) | 58.1 | 41.0 | 26.3 | 17.4 | 24.8 | 53.6 | 48.9 | 32.1 |
| (S) | 53.2 | 41.9 | 29.0 | 19.6 | 25.9 | 53.8 | 52.6 | 32.0 |
| (T) | 54.0 | 39.4 | 24.6 | 15.7 | 24.3 | 50.5 | 41.1 | 30.6 |

(D) Dresses, (S) Shirts, (t) Tops&Tees

- Attribute-aware methods outperform image-only baselines
- Attention mechanism can better utilize the additional attribute information

# Results – Interactive Image Retrieval

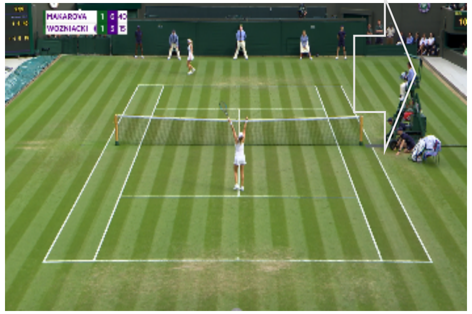
| | Dialog Turn 1 | | | | Dialog Turn 3 | | | | Dialog Turn 5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R@5 | R@10 | R@50 | P | R@5 | R@10 | R@50 | P | R@5 | R@10 | R@50 |
| Attribute-aware (D) | **90.52** | **4.74** | **7.73** | 23.94 | **98.09** | 26.45 | **36.19** | **67.72** | 98.92 | 40.71 | **52.43** | 79.91 |
| with Attention (S) | **90.87** | **2.88** | **4.96** | **17.32** | **98.02** | **18.95** | **27.33** | **55.49** | **98.87** | **29.49** | **40.07** | **69.71** |
| (T) | **90.37** | 3.07 | 5.16 | 17.27 | **98.04** | **21.93** | **30.18** | 59.06 | 99.03 | **36.97** | **47.87** | **77.30** |
| Attribute-aware (D) | 90.39 | 4.52 | 7.48 | **24.14** | 98.00 | **26.65** | 36.05 | 65.60 | **98.95** | **40.88** | 52.37 | **79.99** |
| via Concatenation (S) | 89.93 | 2.41 | 4.09 | 14.86 | 97.55 | 16.15 | 23.63 | 50.60 | 98.55 | 27.21 | 36.44 | 65.25 |
| (T) | 90.34 | **3.22** | **5.39** | **17.75** | 98.03 | 20.78 | 29.02 | **59.57** | **99.07** | 35.37 | 46.41 | 76.58 |
| Image-Only (D) | 89.45 | 3.79 | 6.25 | 20.26 | 97.49 | 19.36 | 26.95 | 57.78 | 98.56 | 28.32 | 39.12 | 72.21 |
| (S) | 89.39 | 2.29 | 3.86 | 13.95 | 97.40 | 14.70 | 21.78 | 47.92 | 98.48 | 23.99 | 32.94 | 62.03 |
| (T) | 87.89 | 1.78 | 3.03 | 12.34 | 96.82 | 10.76 | 17.30 | 42.87 | 98.30 | 20.57 | 29.59 | 60.82 |

- Attribute information and relative expressions jointly lead to better retrieval results
- More advanced techniques for composing side information, relative feedback and image features could lead to further performance gains.

# This talk

- Weak supervised learning for fashion search
- Learning with less labels beyond weak supervision

# IBM Research AI – Learning with Less Labels for Vision




## Data Augmentation

- Delta (Δ)-encoder
- Label Set Operations (LaSO)

## Visual Learning

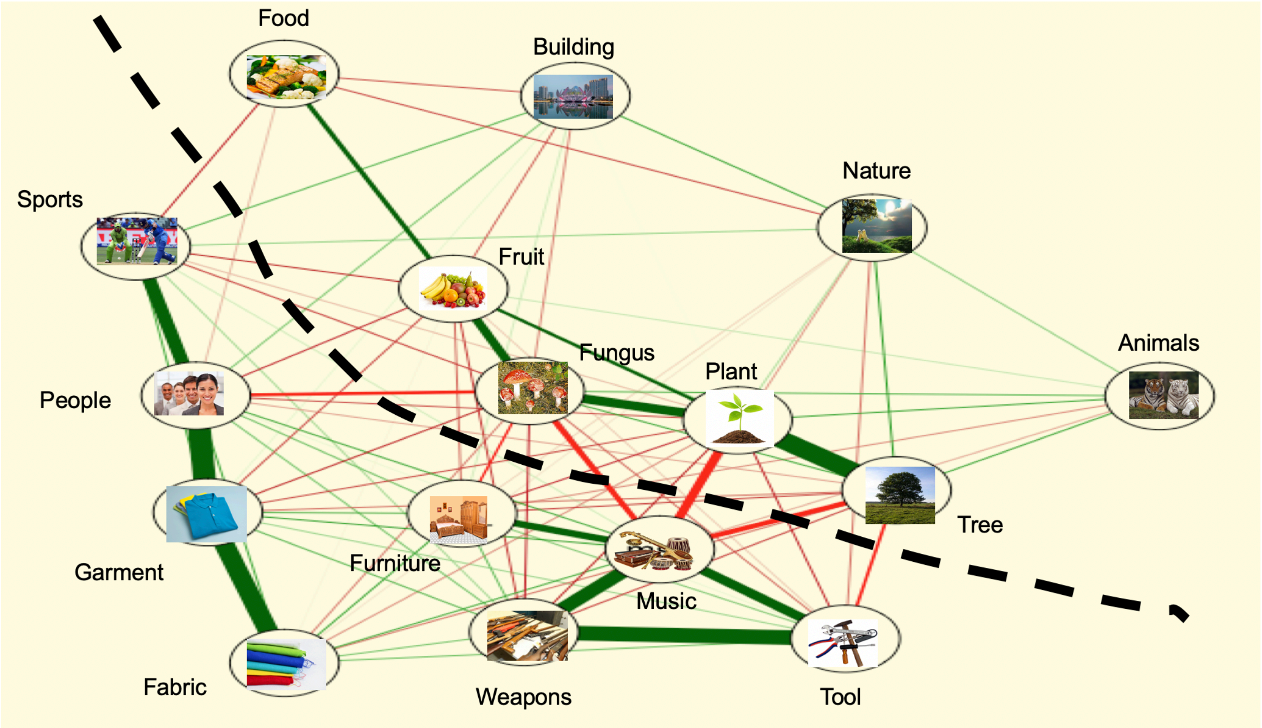
- Representative-based metric learning (RepMet)
- Learning with Semantics

## Transfer Learning

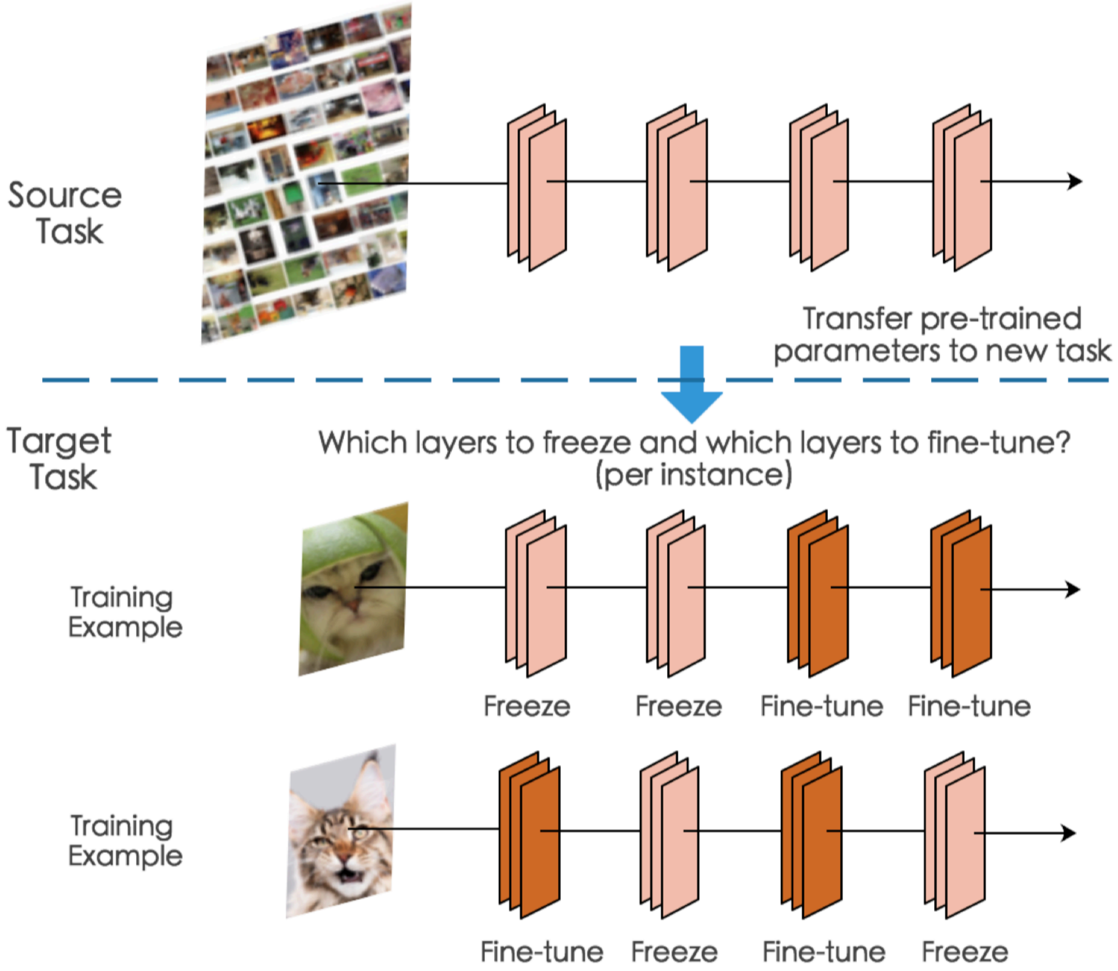
- Learning to Transfer
- SpotTune

# Transfer Learning

## Model Selection
[Dube et al, Deep Vision Workshop 2019]



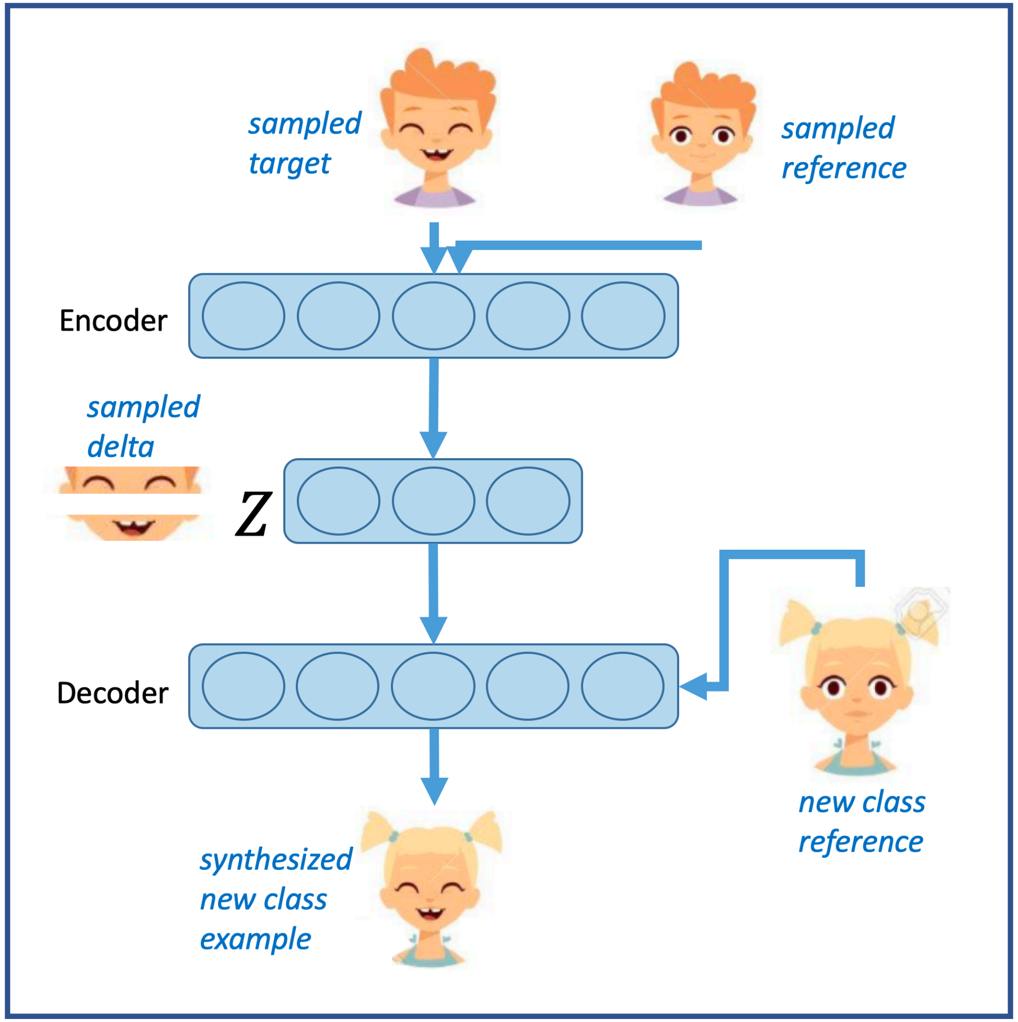

## SpotTune [Guo et al, CVPR 2019]



# Sample Synthesis for Few-Shot Learning

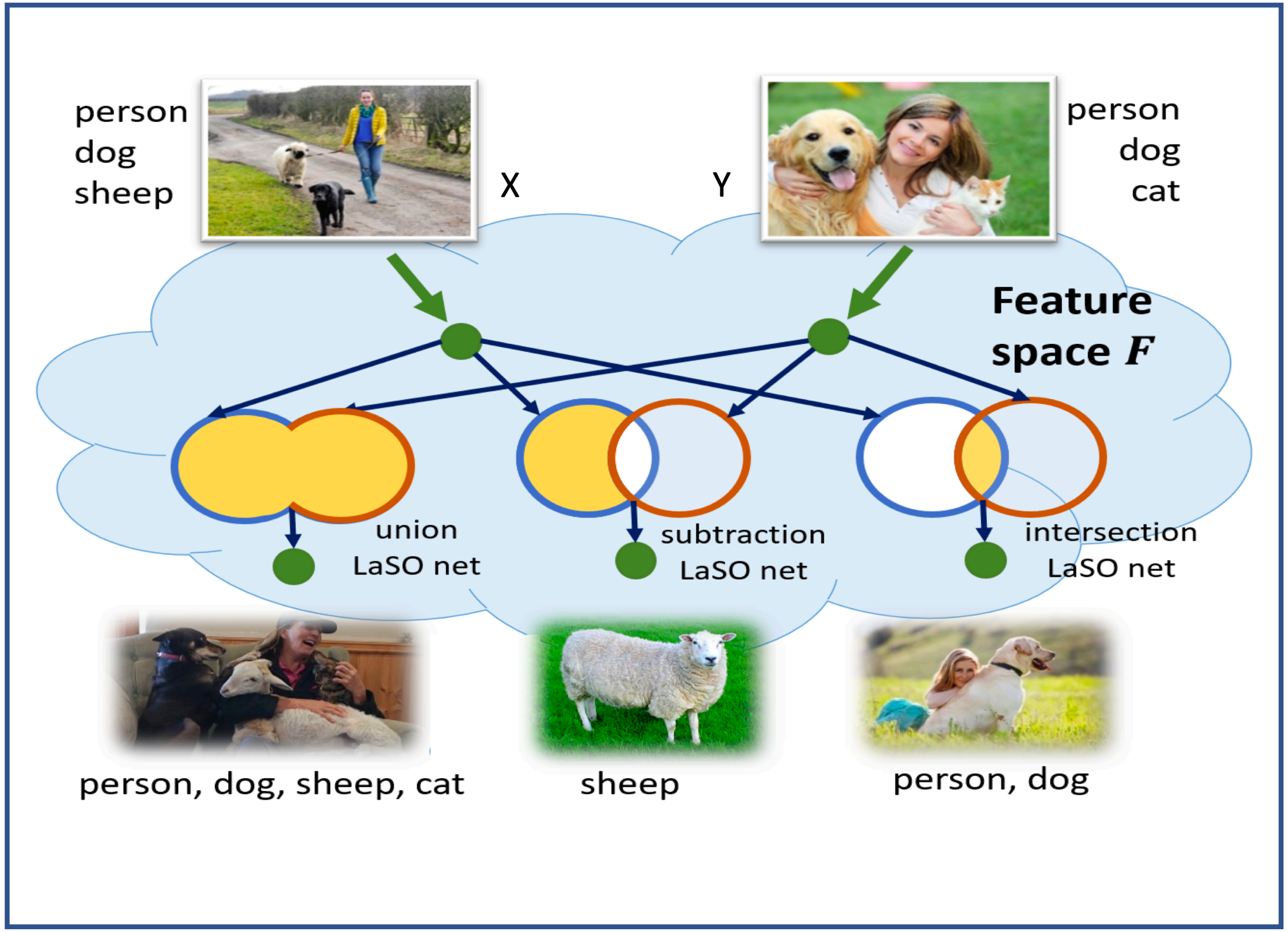
## Delta-Encoder

[Schwartz & Karlinsky et al, NeurIPS 2018]




## LaSO

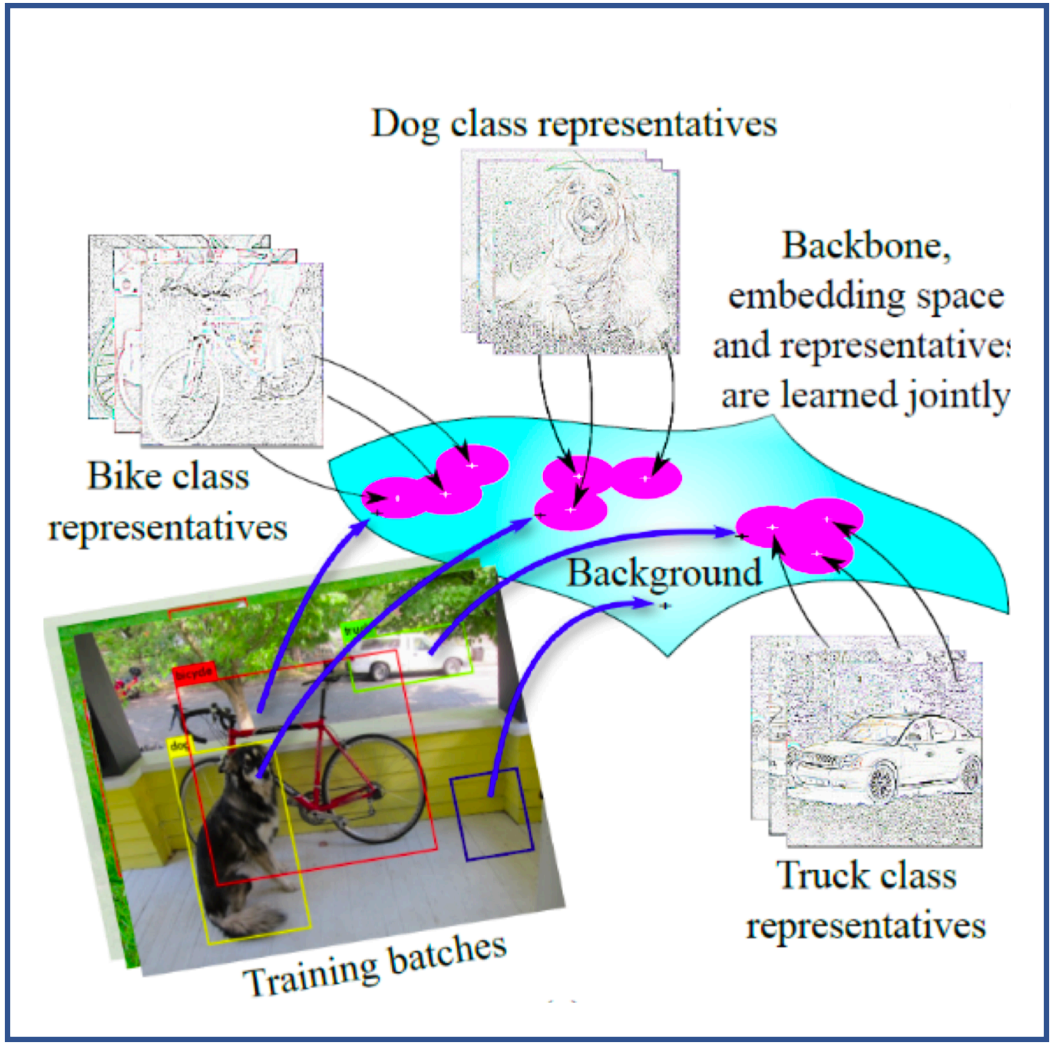
[Alfassy & Karlinsky et al, CVPR 2019]



# Few-shot Learning

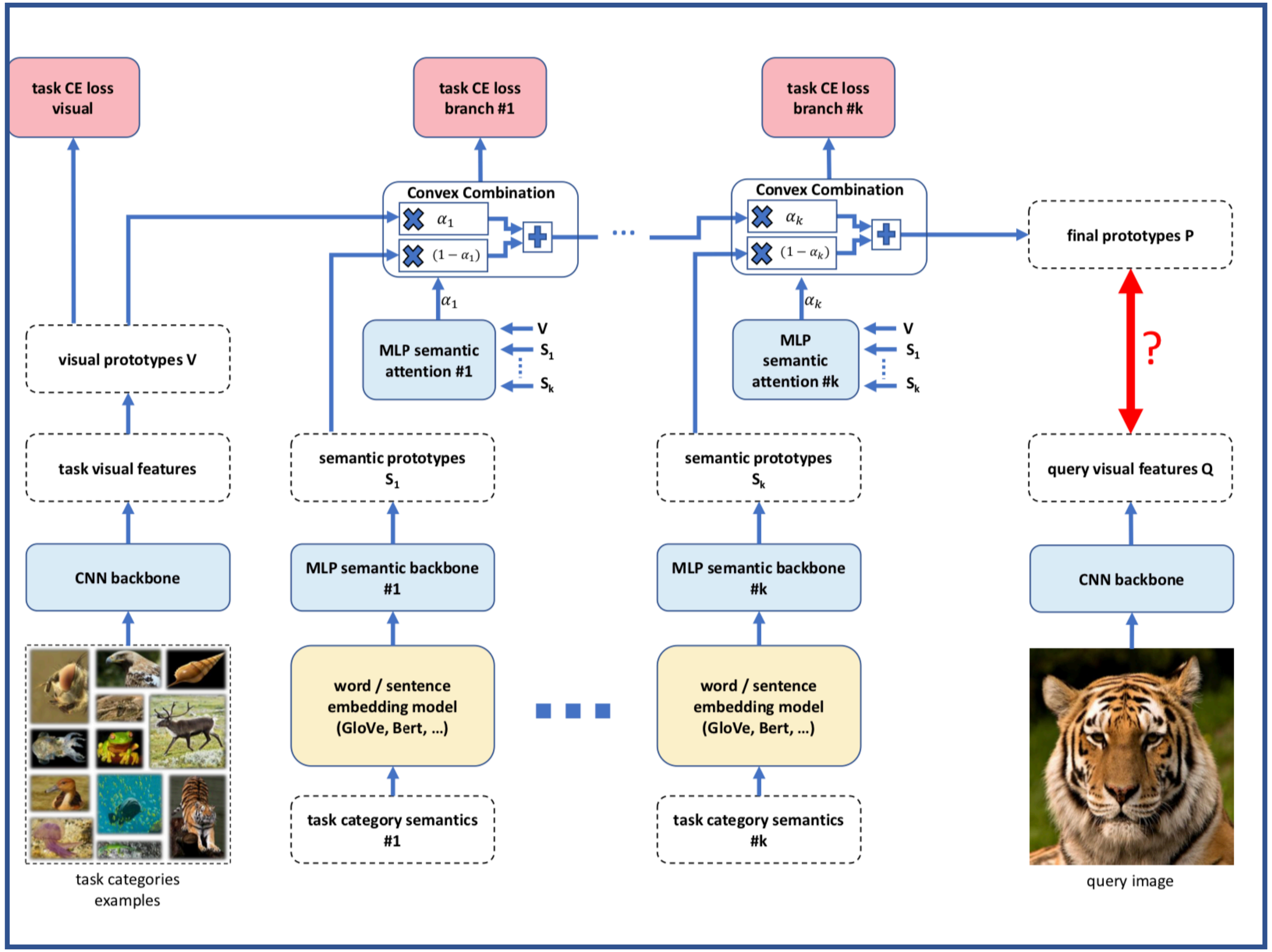
## RepMet

[Karlinsky et al, CVPR 2019]




## Learning with Semantics

[Schwartz & Karlinsky et al, Language & Vision Workshop, 2019]



# Summary

- Takeaway message: Noisy visual attribute labels mined from the web are useful as *privileged information* during training to improve image search:
  - Street2Shop fashion retrieval [Huang et al, ICCV 2015]
  - Dialog-based interactive fashion retrieval [Guo & Wu et al, NeurIPS 2018] [Guo & Wu et al, 2019]

- Check out our recent work on learning with less labels @CVPR

# IBM Research AI: Learning More from Less in Vision @ CVPR


1. A. Alfassy, L. Karlinsky, A. Aides, J. Shtok, S. Harary, R. Feris, R. Giryes, A. M. Bronstein, “LaSO: Label-Set Operations network for multi-label few-shot classification,” *CVPR-2019*, June 2019.

2. L. Karlinsky, J. Shtok, S. Harary, E. Schwartz, A. Aides, R. Feris, R. Giryes, A. M. Bronstein, “RepMet: Representative-based metric learning for classification and one-shot object detection”, *CVPR-2019*, June 2019.

3. Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, R. Feris, “SpotTune: Transfer Learning through Adaptive Fine-Tuning,” *CVPR-2019*, June 2019.

4. E. Schwartz, L. Karlinsky, R. Feris, R. Giryes, A. Bronstein, “Baby steps towards few-shot learning with multiple semantics,” *Language and Vision Workshop at CVPR-2019*, June 2019.

5. P. Dube, B. Bhattacharjee, S. Huo, P. Watson, B. Belgodere, J. R. Kender, “Automatic Labeling of Data for Transfer Learning”, Deep *Vision Workshop at CVPR-2019*, June 2019.