Is It All Relative?

Interactive Fashion Search based on Relative Natural Language Feedback

Rogerio Schmidt Feris

Principal Research Scientist and Manager

IBM Research & MIT-IBM Watson AI Lab

Computer Vision for Fashion

Style discovery and analysis

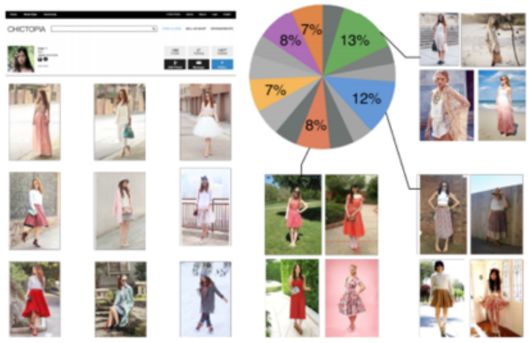
(MH Kiapour, et al, ECCV 2014)

Trend modeling and forecast

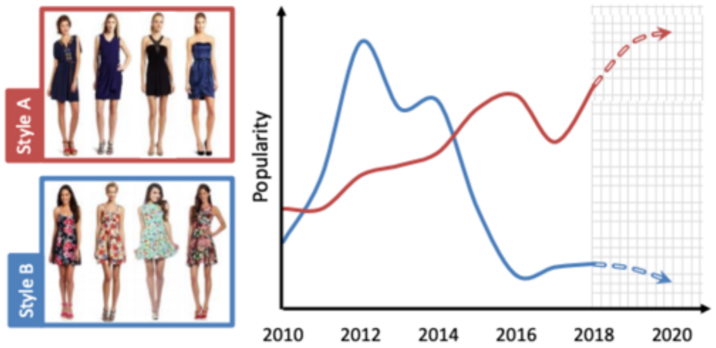
(R He, et al, WWW 2016)

Outfit recommendation

(WL Hsiao, et al, CVPR 2018)



(WL Hsiao, et al, ICCV 2017)



(Z Al-Halah, et al, ICCV 2017)

Virtual Try-on

(X Han, et al, CVPR 2018)

This Talk: Fashion Image Search

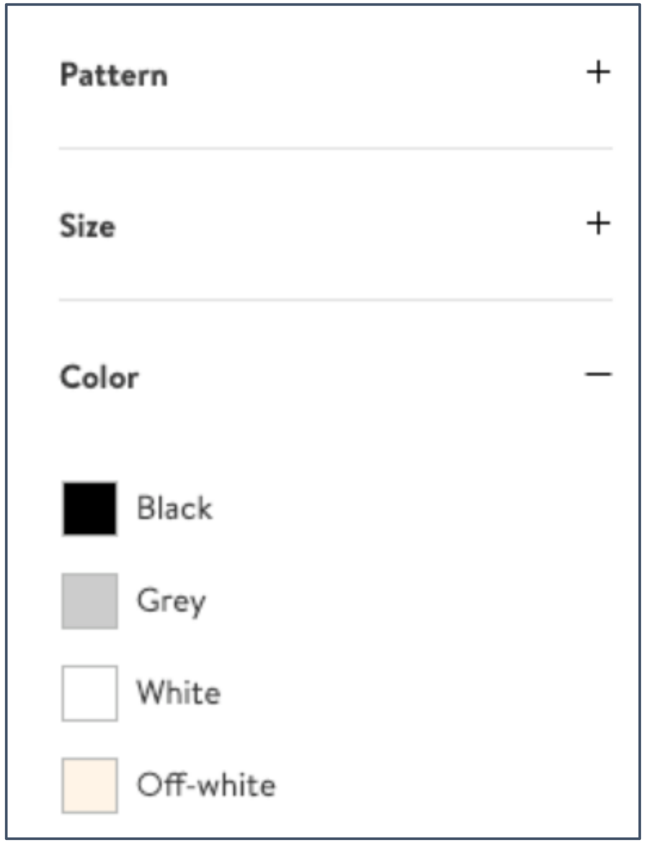
with (subjective) visual attributes

Fashion Search: Challenges

- Subjective Attributes

- Hard to describe the desired fashion item in words and resolve user intent

- Filter choices are limited. Hard to narrow down search results to the desired style.



Interactive Personal Shopper

HI! What Are you interested in shopping today?

OR show me a photo

User drag-n-drop a look that is similar to what she/he is looking for.

Interactive Personal Shopper

(Street2shop) Retrieved results based on user input image.

Pick the one you are most interested in

Or tell me you preferences

I prefer black color.

Refine search results by taking user feedback.

Interactive Personal Shopper

Pick the one you are most interested in

Or tell me you preferences

Like the right one but with different neckline

User can iteratively interact with the search interface

Interactive Personal Shopper

Pick the one you are most interested in

Or tell me you preferences

Outline

- Street2Shop

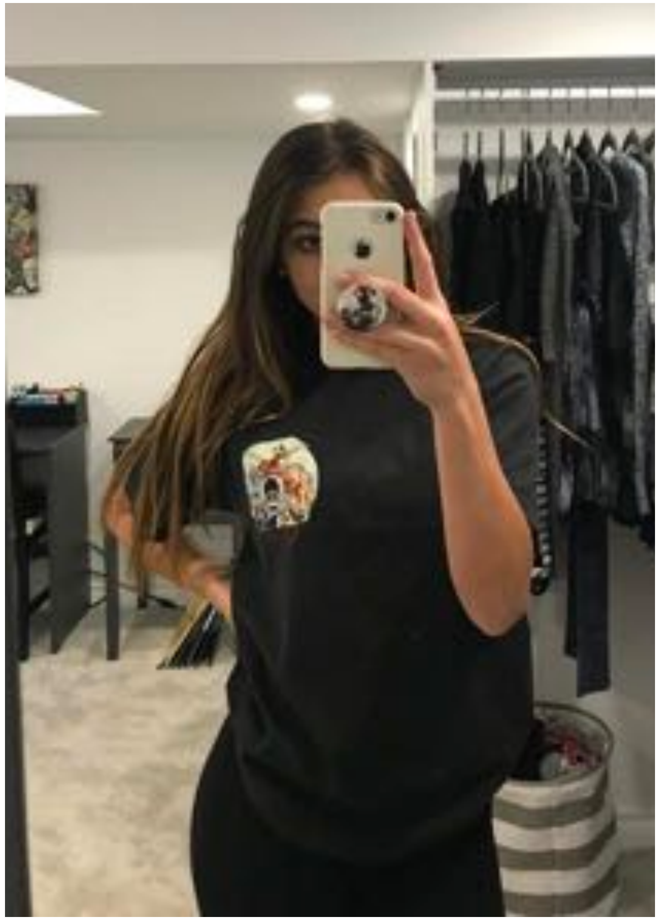
[Huang et al, ICCV 2015]

- Interactive image search using natural language feedback

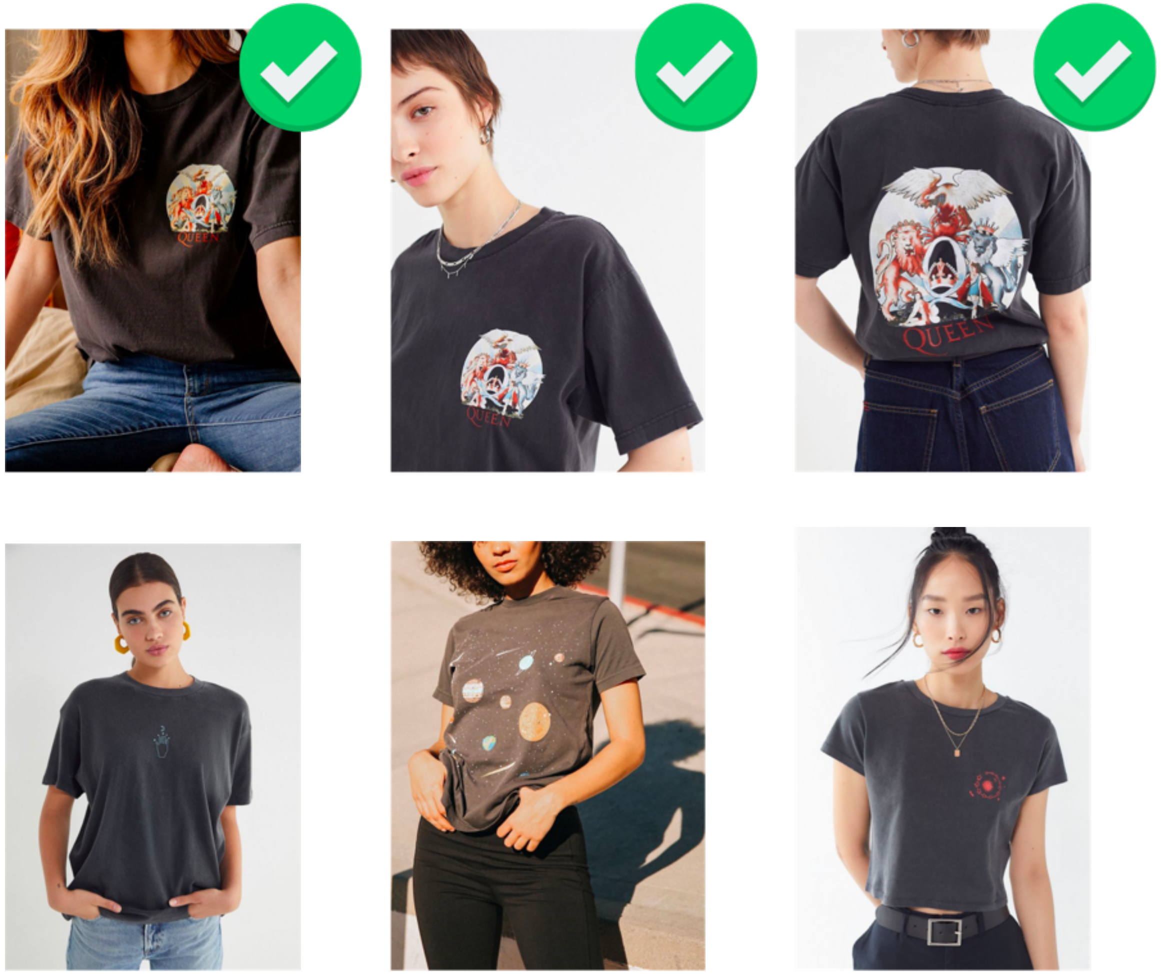
[Guo & Wu et al, NeurIPS 2018]
[Guo & Wu et al, 2019]

Clothing Retrieval (Street2Shop)

Input: User Photo



Retrieved Images from Online Shopping Stores



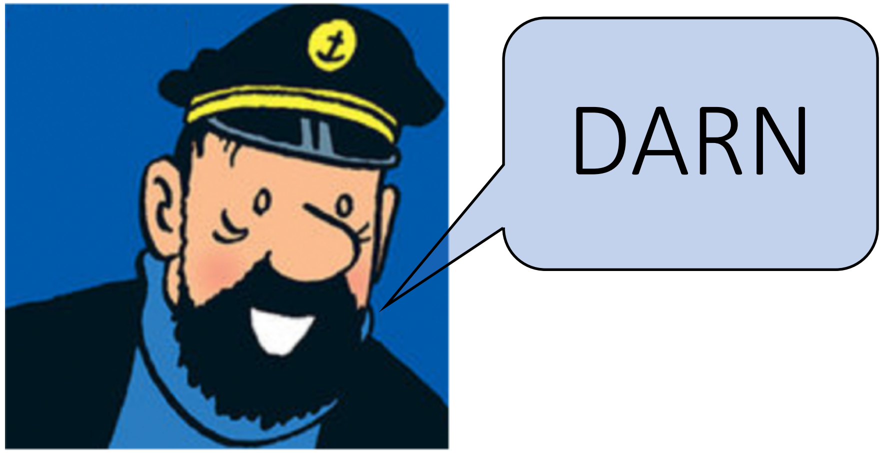
[Liu et al, CVPR 2012] [Kiapour et al, ICCV 2015] [Huang et al, ICCV 2015]

Problem: Domain Discrepancy

Shopping Catalog User Photo

Proposed Approach:

Dual Attribute-Aware Ranking Network (DARN) [Huang et al, ICCV 2015]



Weakly labeled data from shopping websites

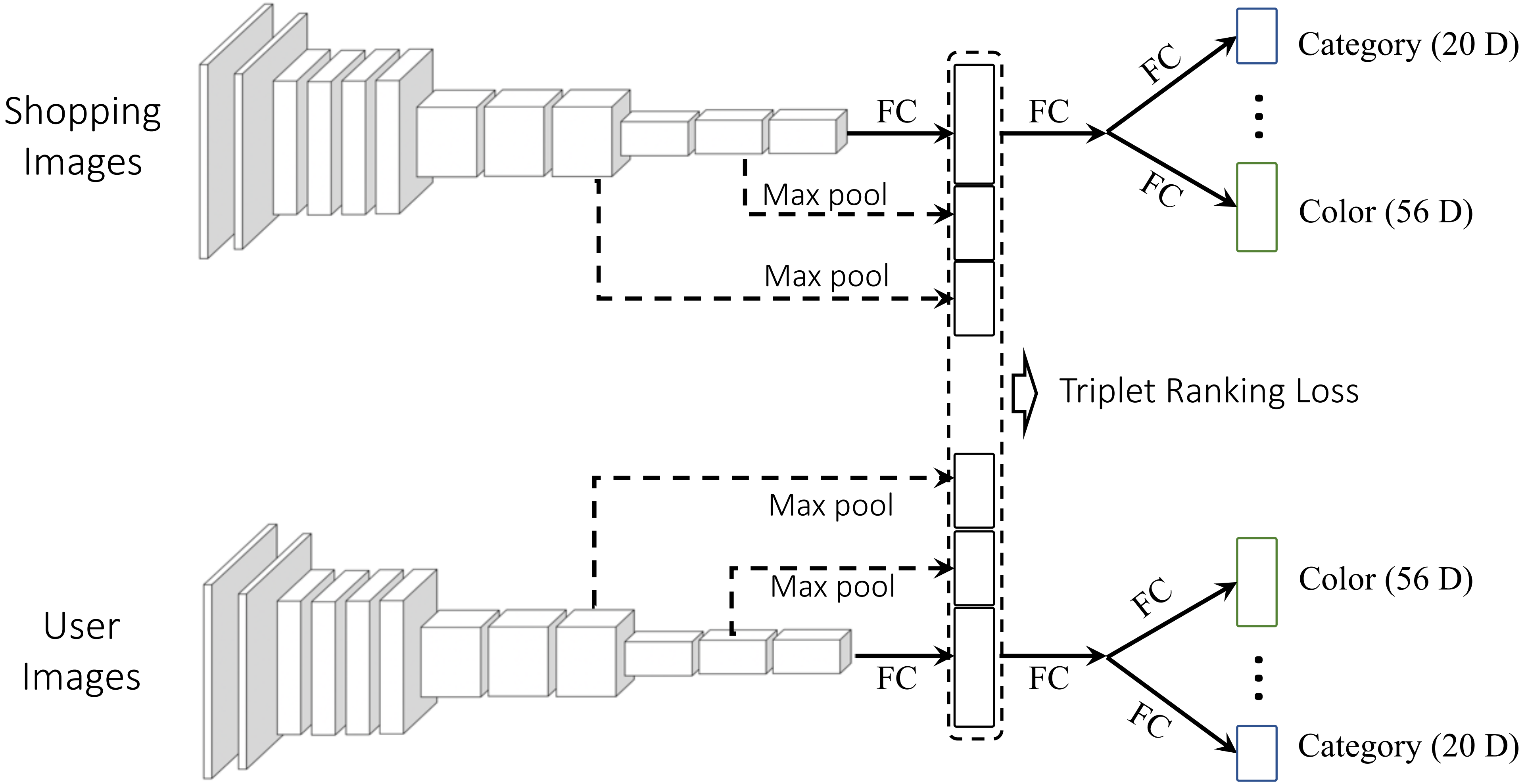
- 9,000 image pairs (exact same clothing)

- Noisy attribute labels (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)

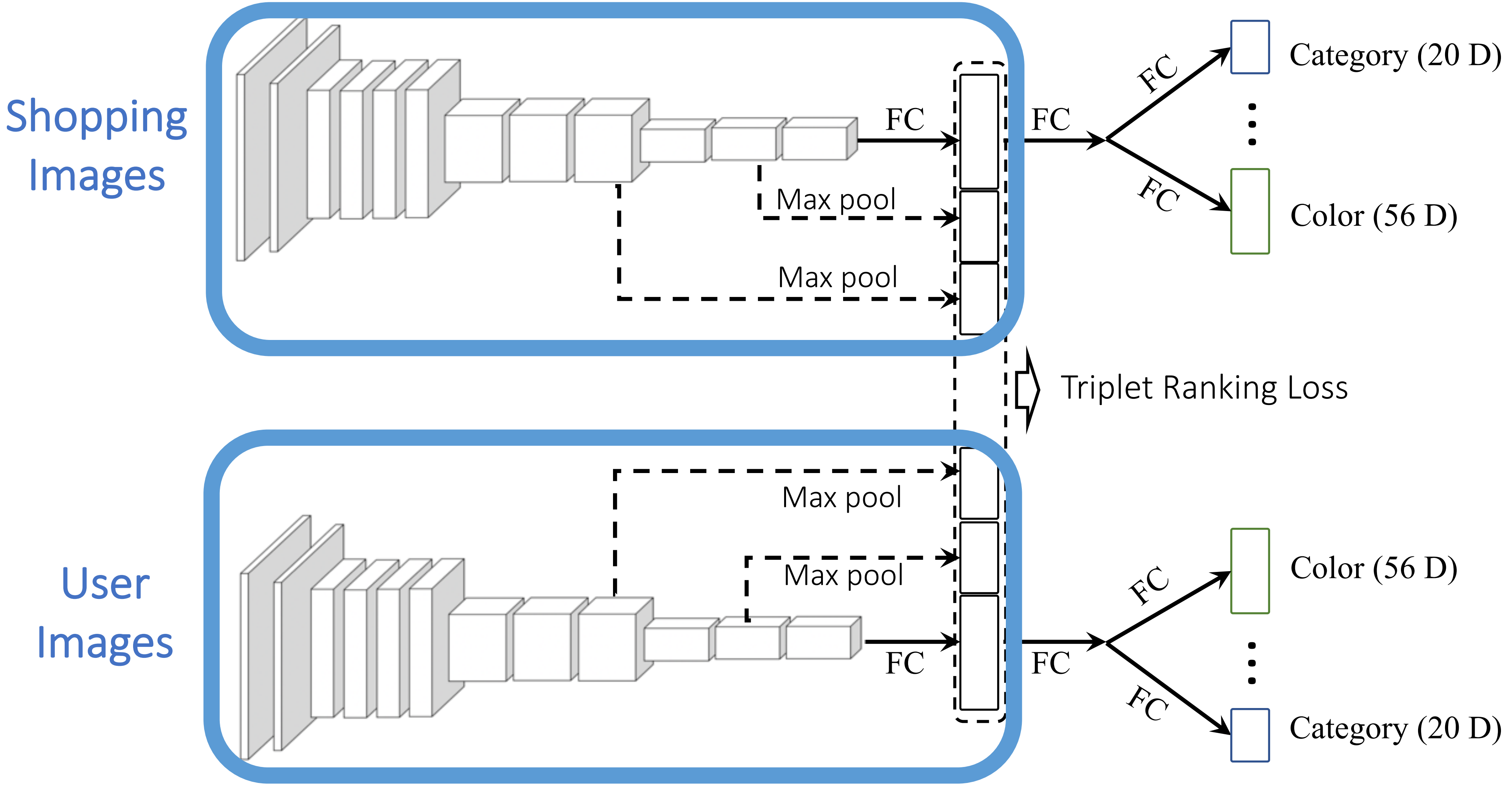
Dual Attribute-Aware Ranking Network (DARN)

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



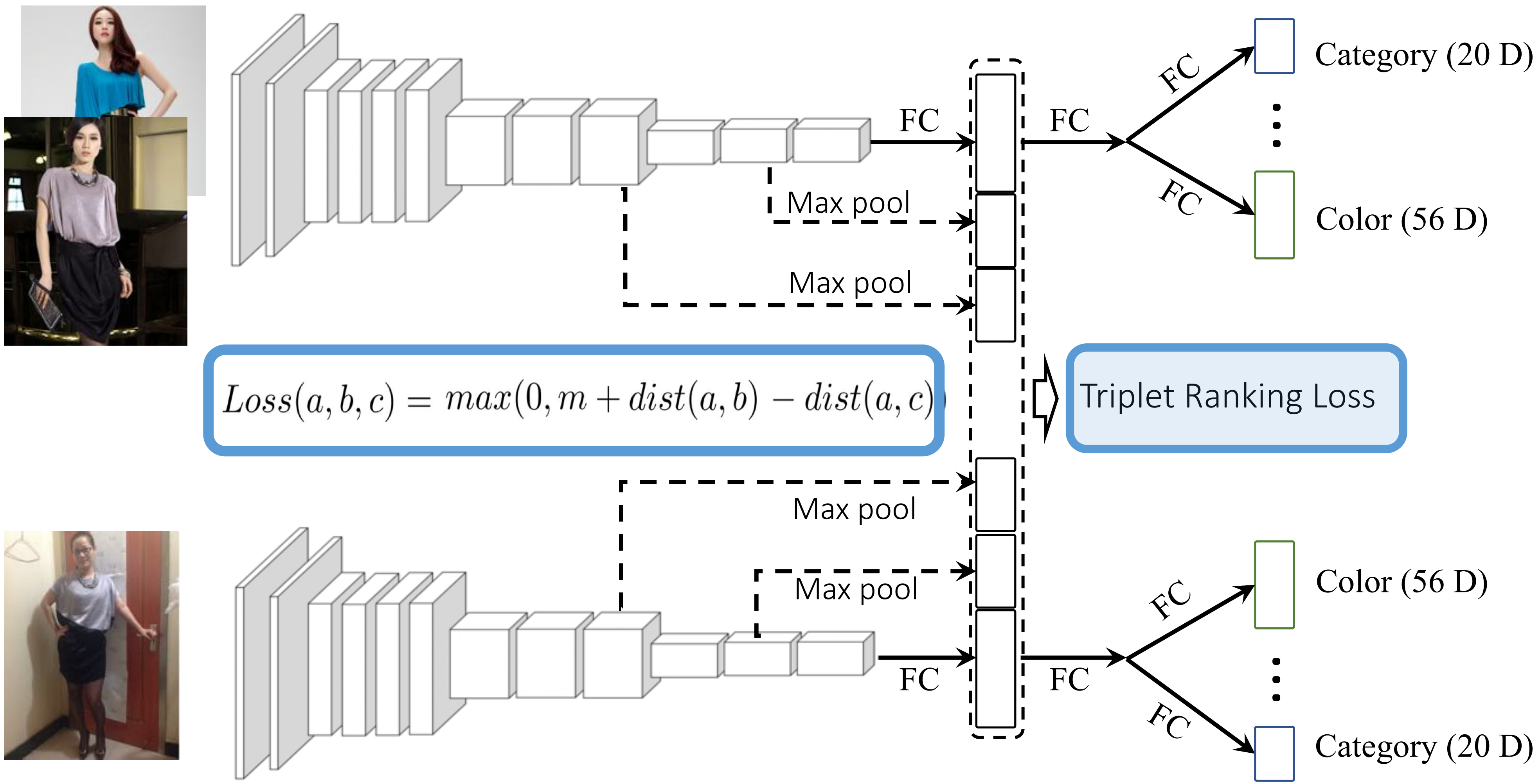
Dual Attribute-Aware Ranking Network (DARN)

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



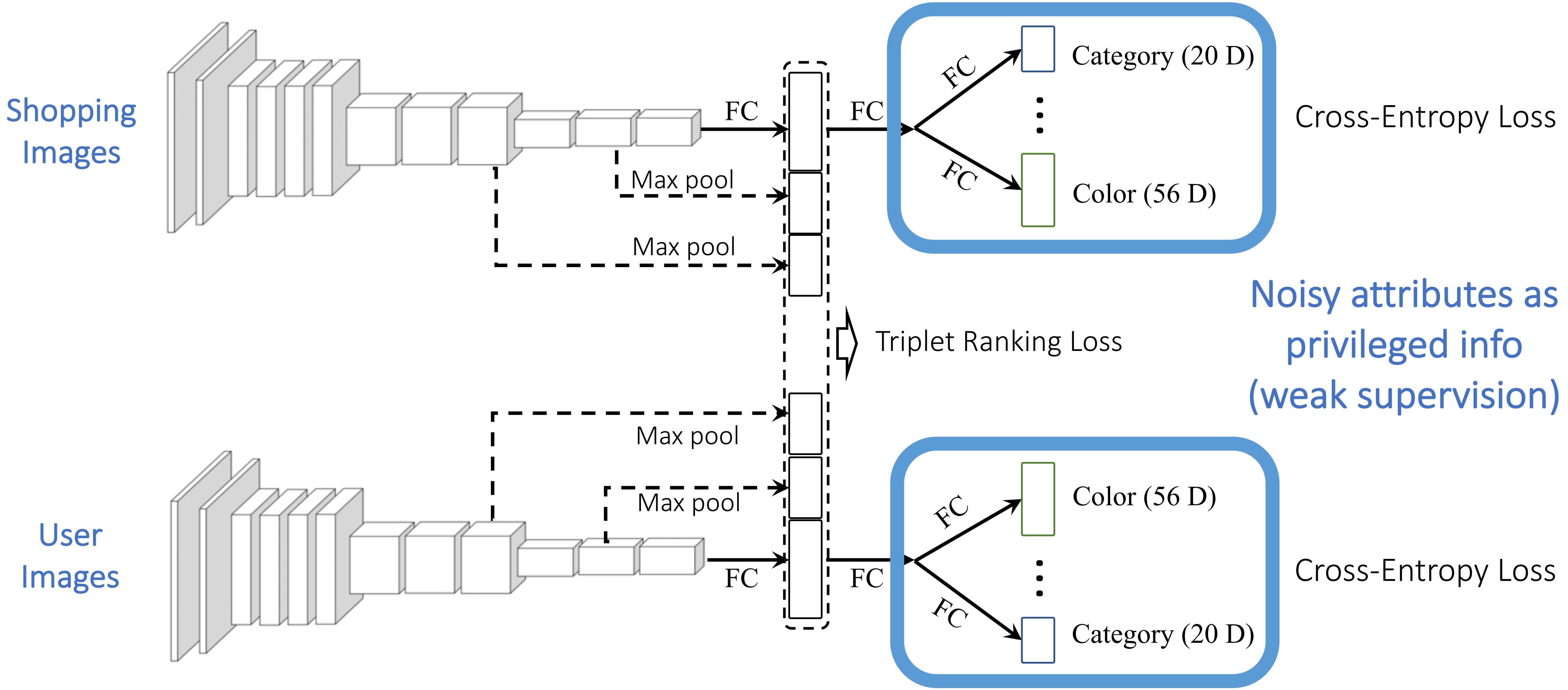
Dual Attribute-Aware Ranking Network (DARN)

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



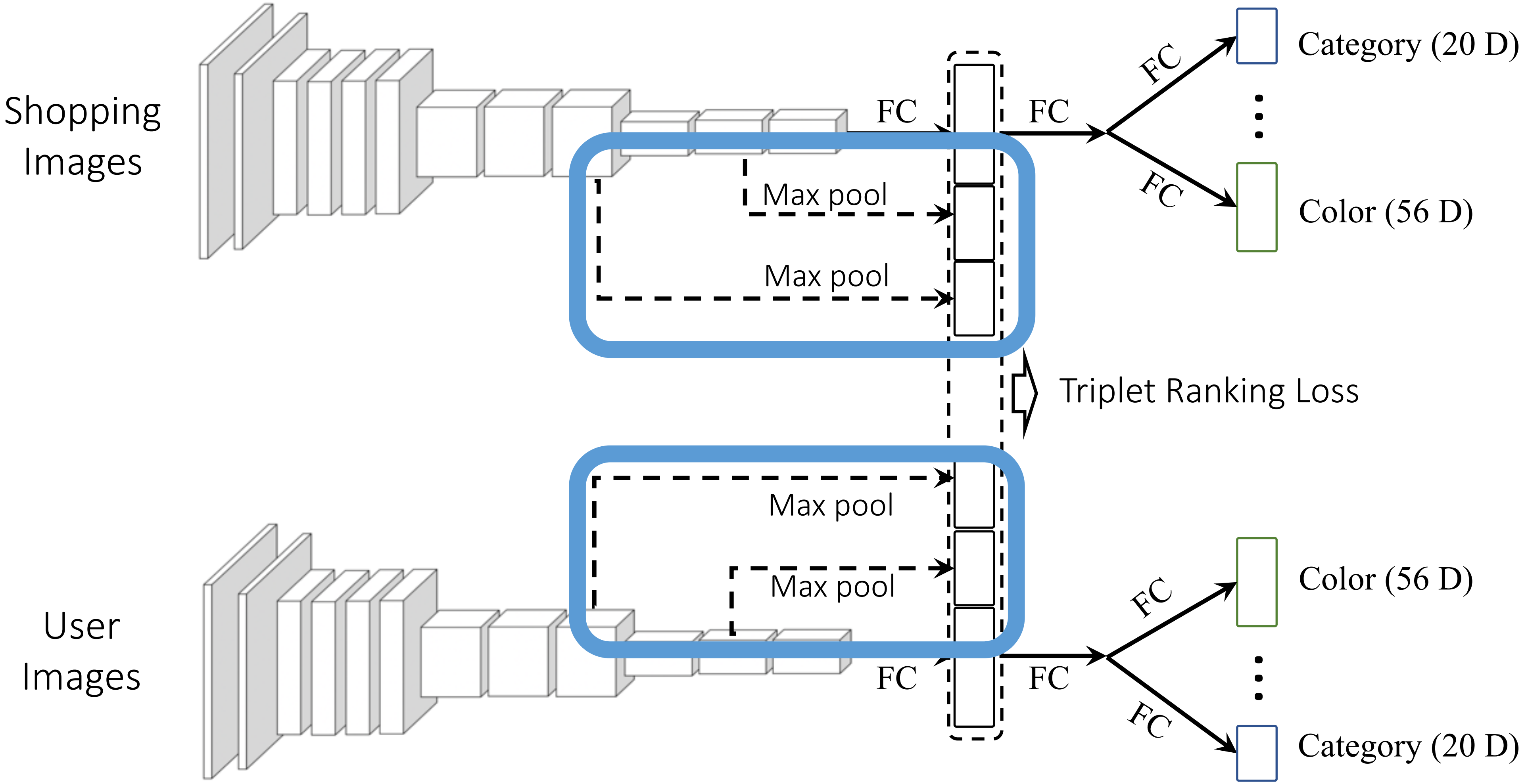
Dual Attribute-Aware Ranking Network (DARN)

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



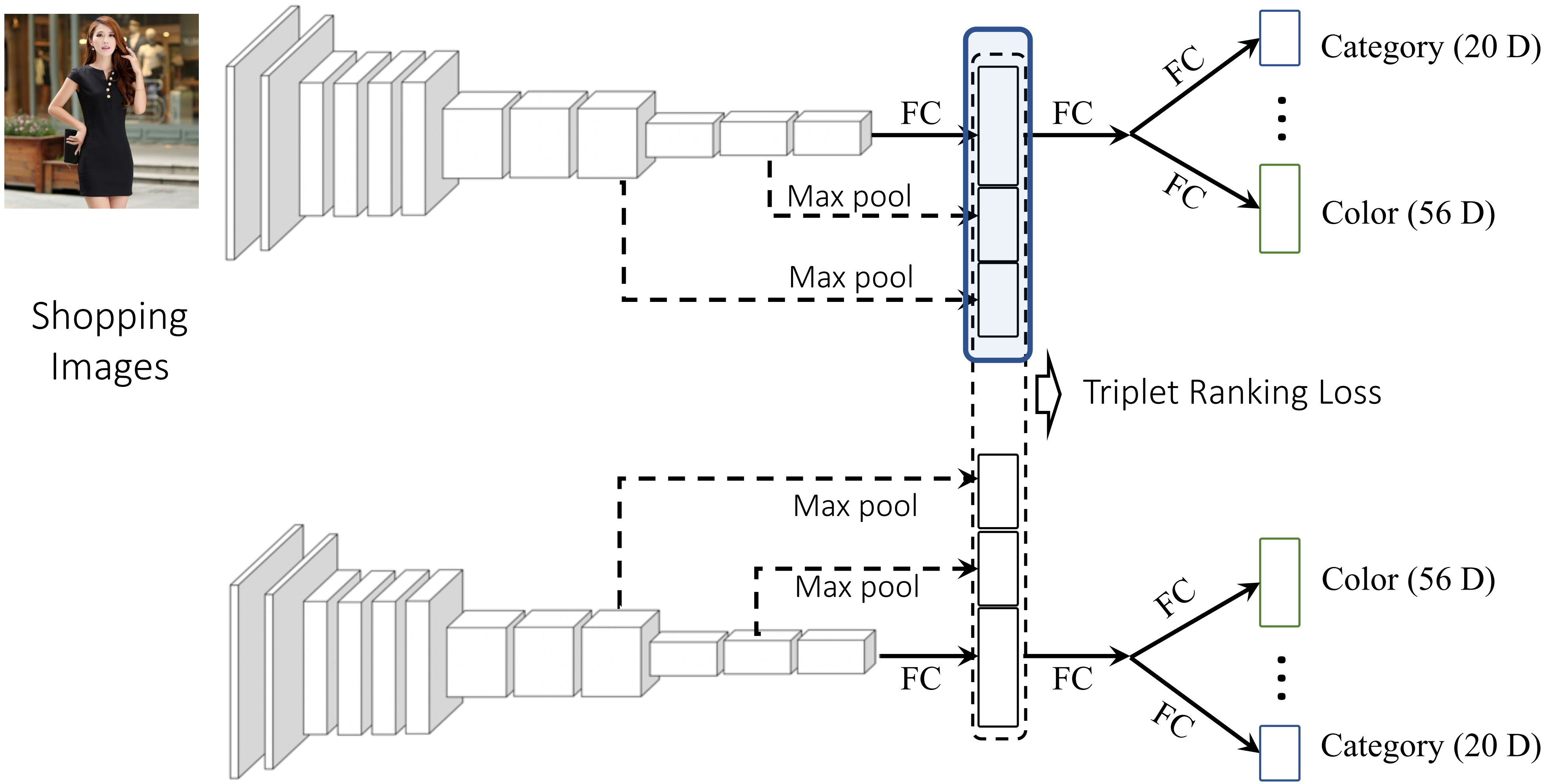
Dual Attribute-Aware Ranking Network (DARN)

- Features from conv layers for encoding more localized information



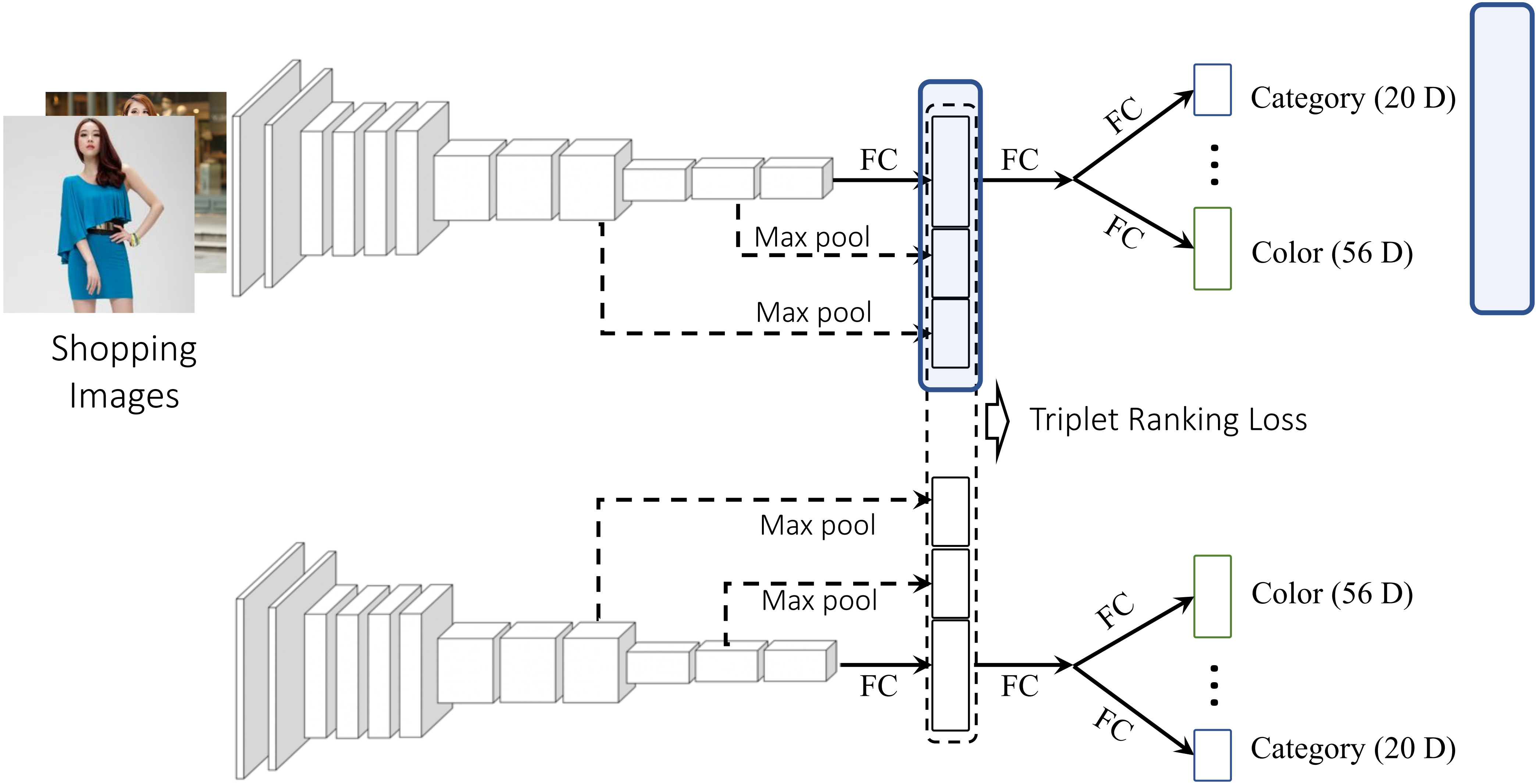
Dual Attribute-Aware Ranking Network (DARN)

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



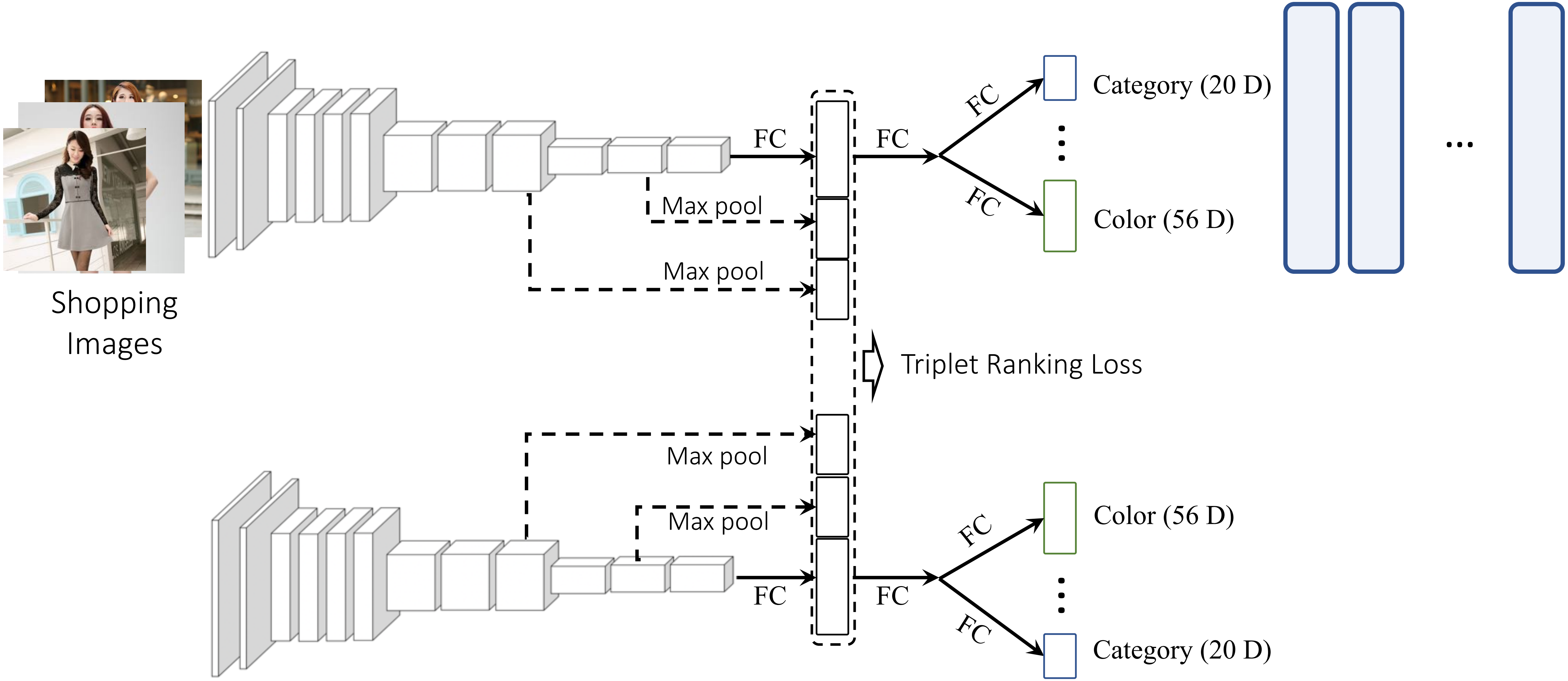
Dual Attribute-Aware Ranking Network (DARN)

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



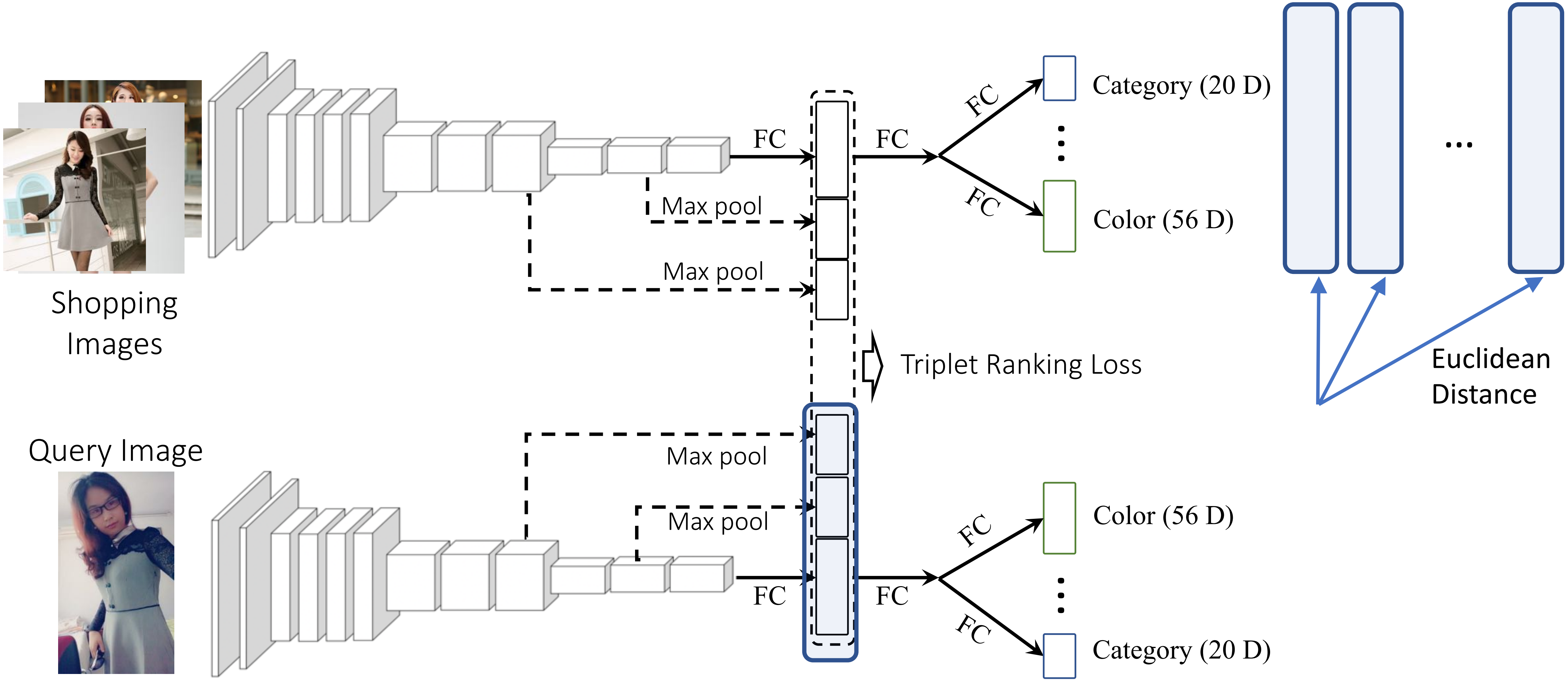
Dual Attribute-Aware Ranking Network (DARN)

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



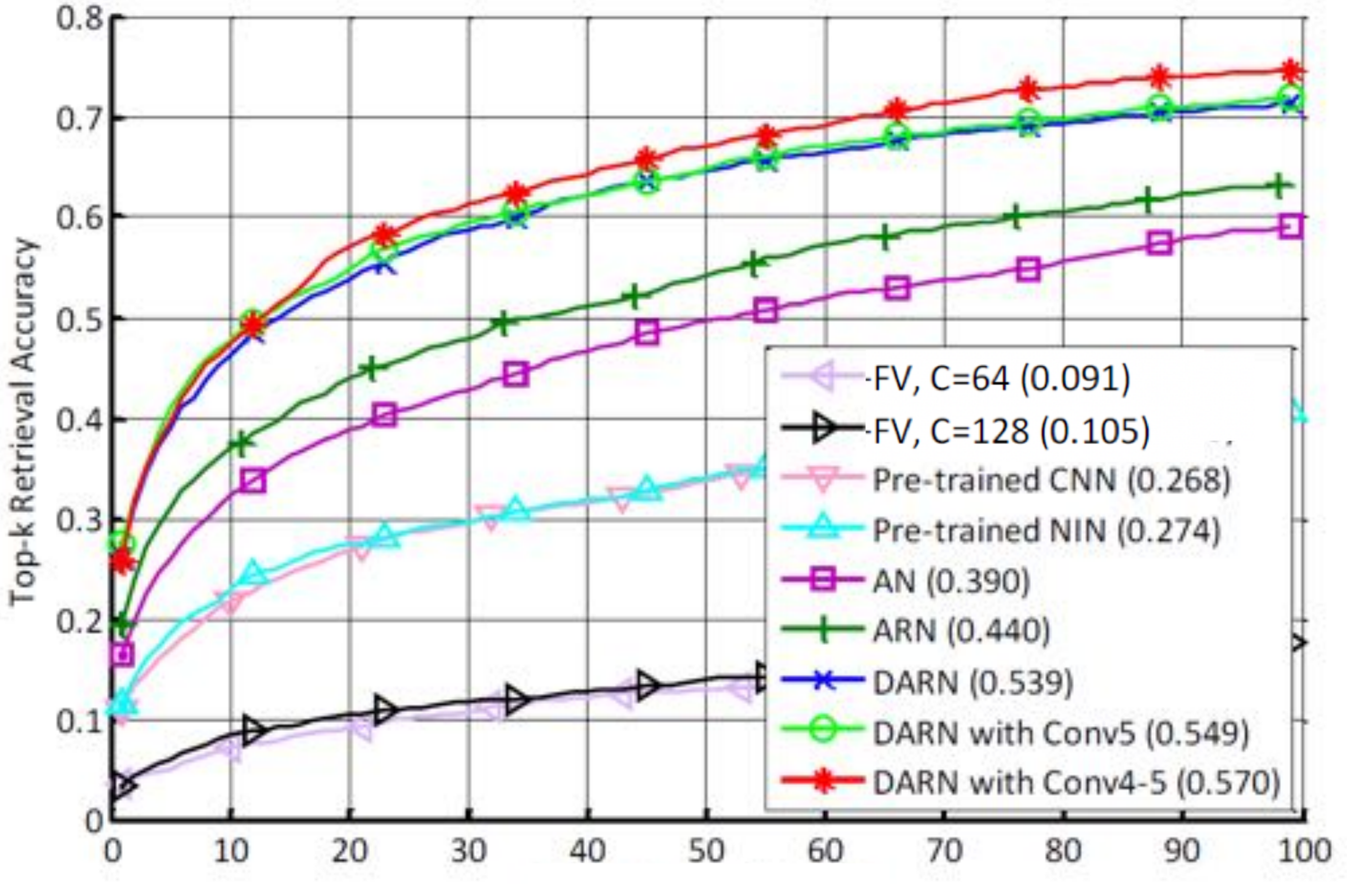
Dual Attribute-Aware Ranking Network (DARN)

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



Experimental Results

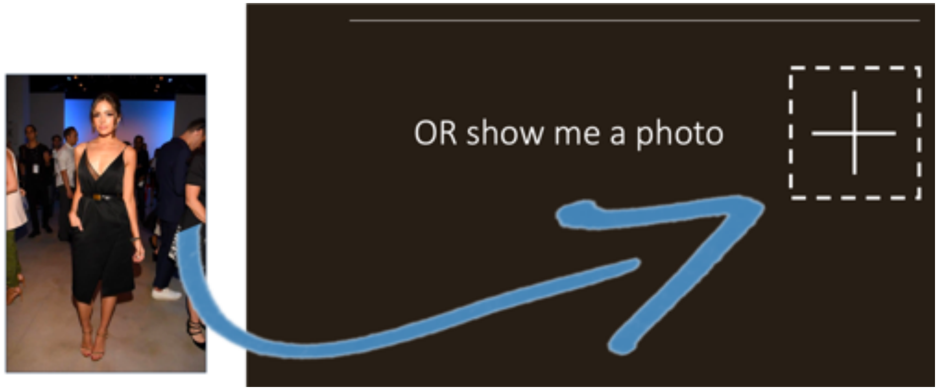
Top-k retrieval accuracy on 200,000 retrieval gallery.
The number in the parentheses is the top-20 retrieval accuracy.



First Column: Query
Green Box: Exact same clothing

Outline

- Street2Shop



[Huang et al, ICCV 2015]

- Interactive image search using natural language feedback

[Guo & Wu et al, NeurIPS 2018]
[Guo & Wu et al, 2019]

Fashion Search using Interactive Feedback

Relevance Feedback [Rui et al, 1998]

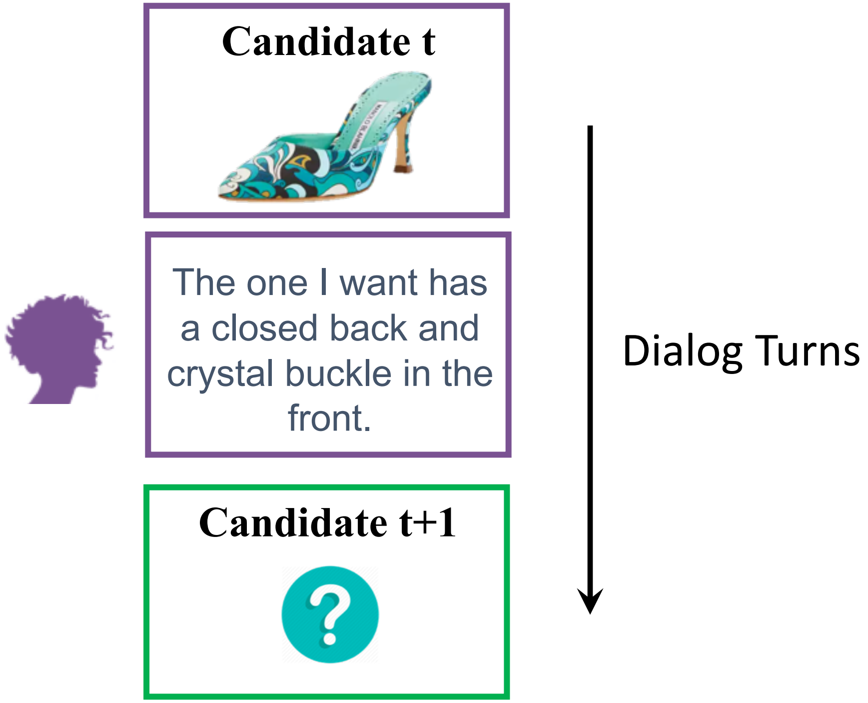
Relative Attribute Feedback [Kovashka et al, 2012]

- Limit the information the user can convey about an image
- Pre-defined set of attributes (limited vocabulary, cumbersome interface)

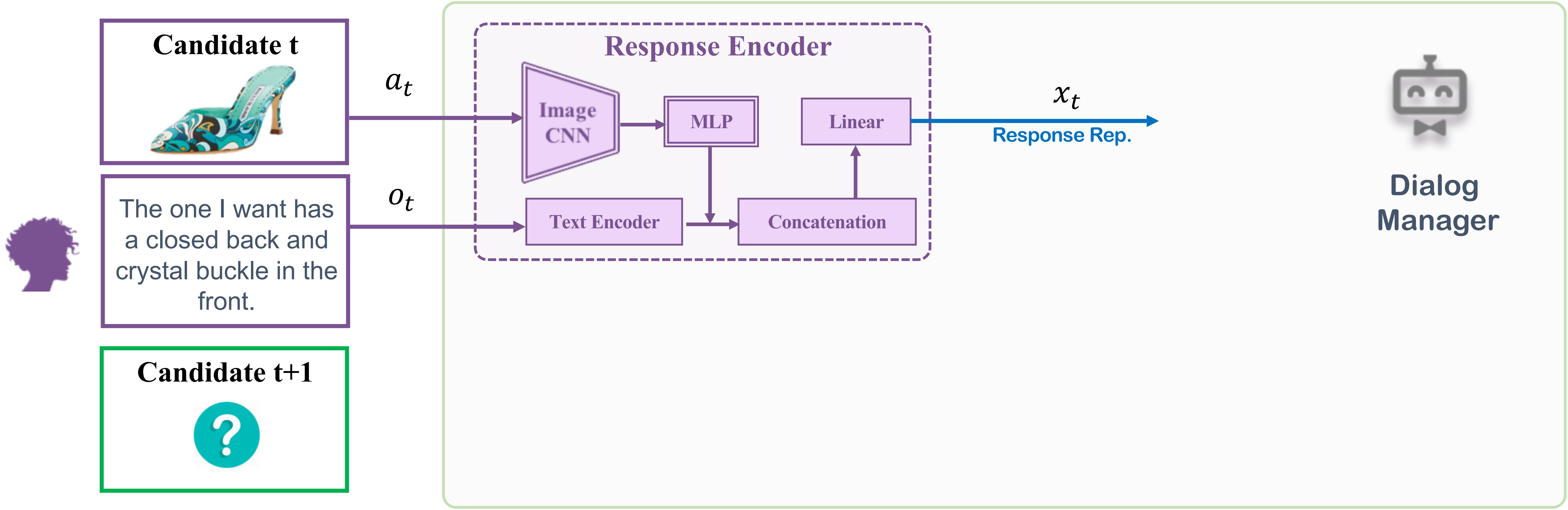
FASHION IQ DEMO

IBM RESEARCH AI

Network Architecture

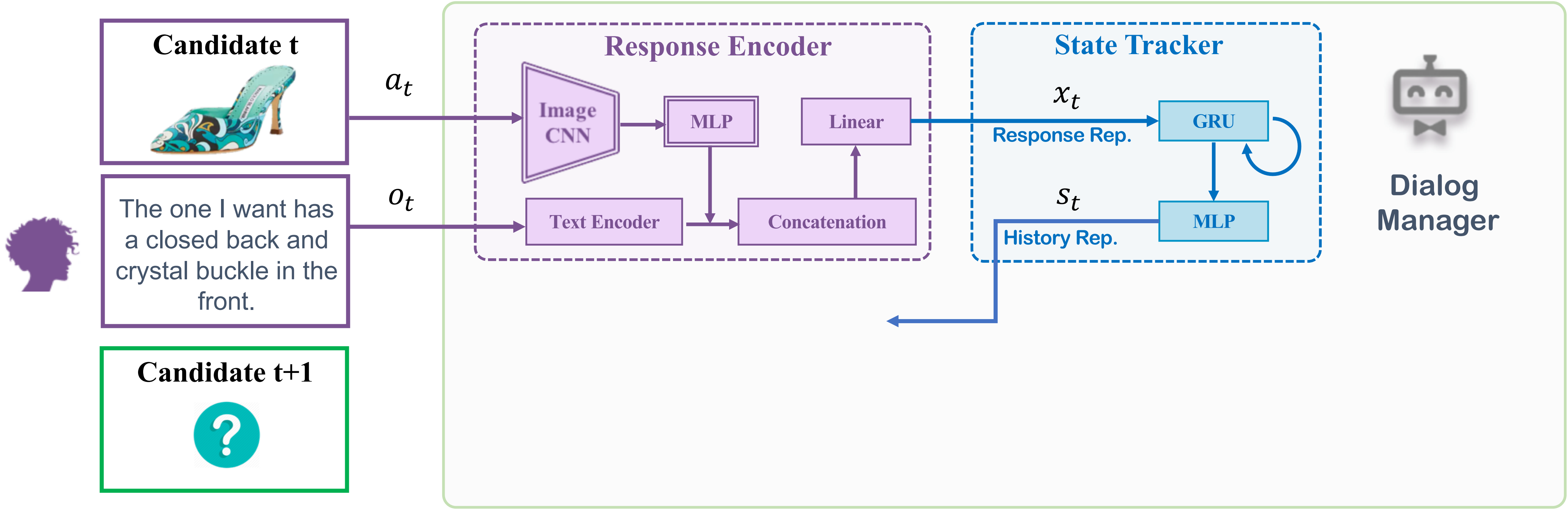


Network Architecture



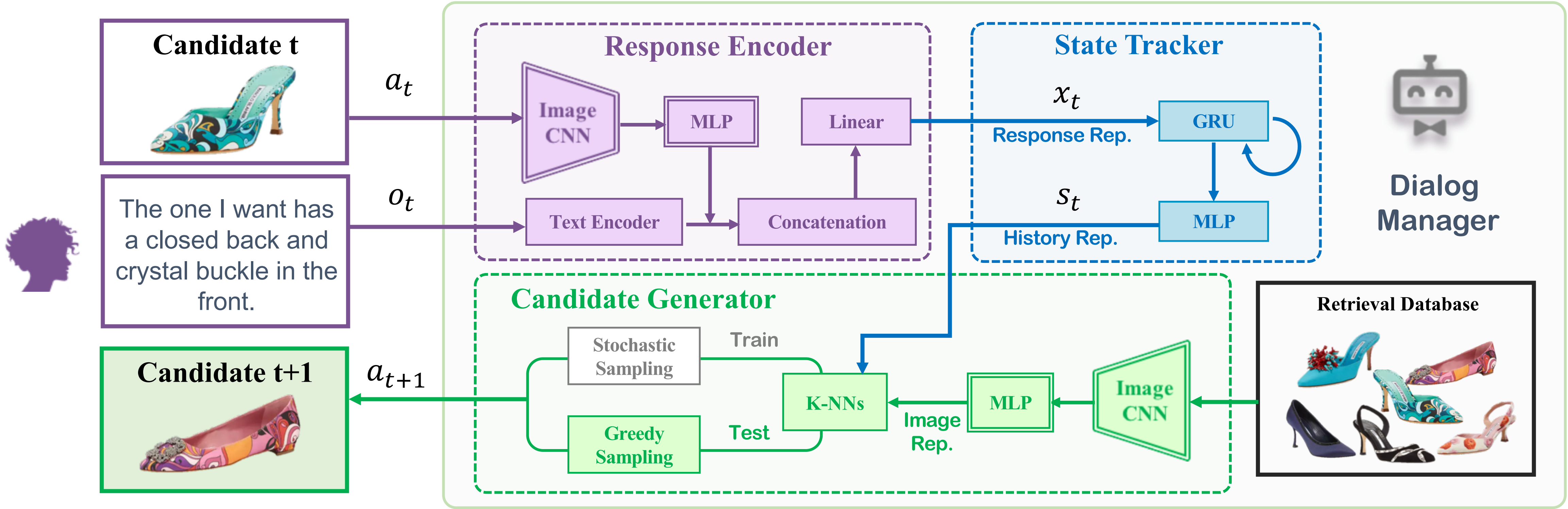
The goal of the Response Encoder is to embed the information from the t-th dialog turn to a joint visual semantic representation.

Network Architecture



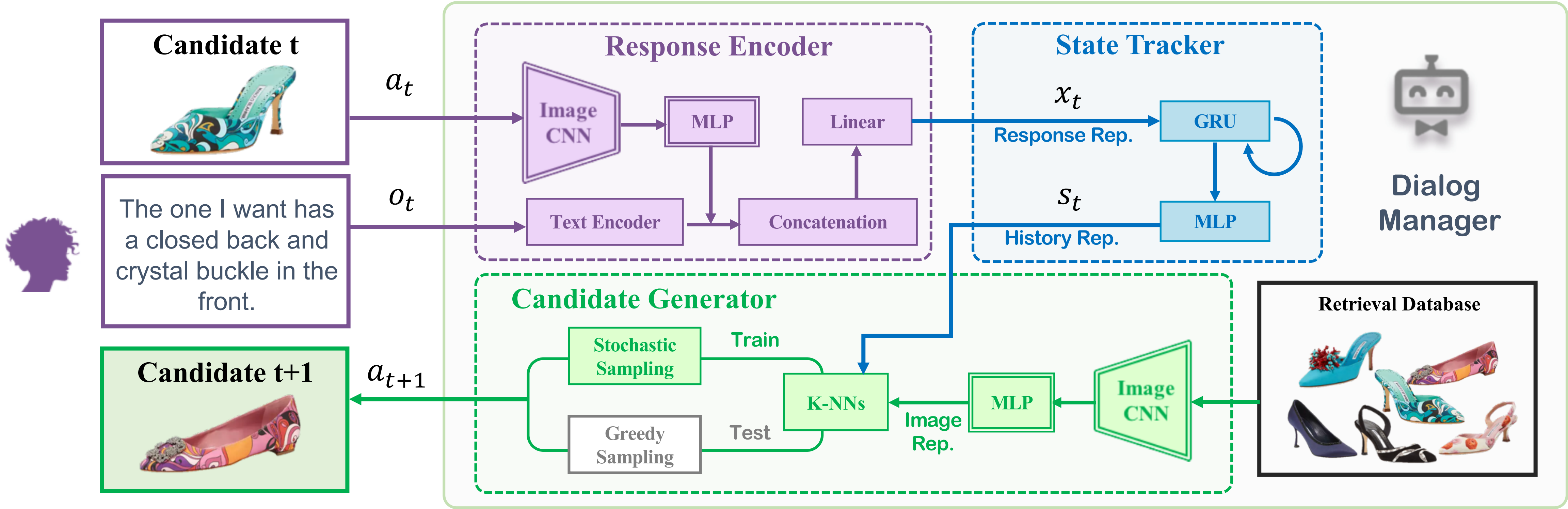
The State Tracker receives as input the response representation, combines it with the history representation of previous dialog turns, and outputs the history representation.

Network Architecture



During testing, the candidate for t+1 round is selected by finding the closest database feature to the history representation.

Network Architecture



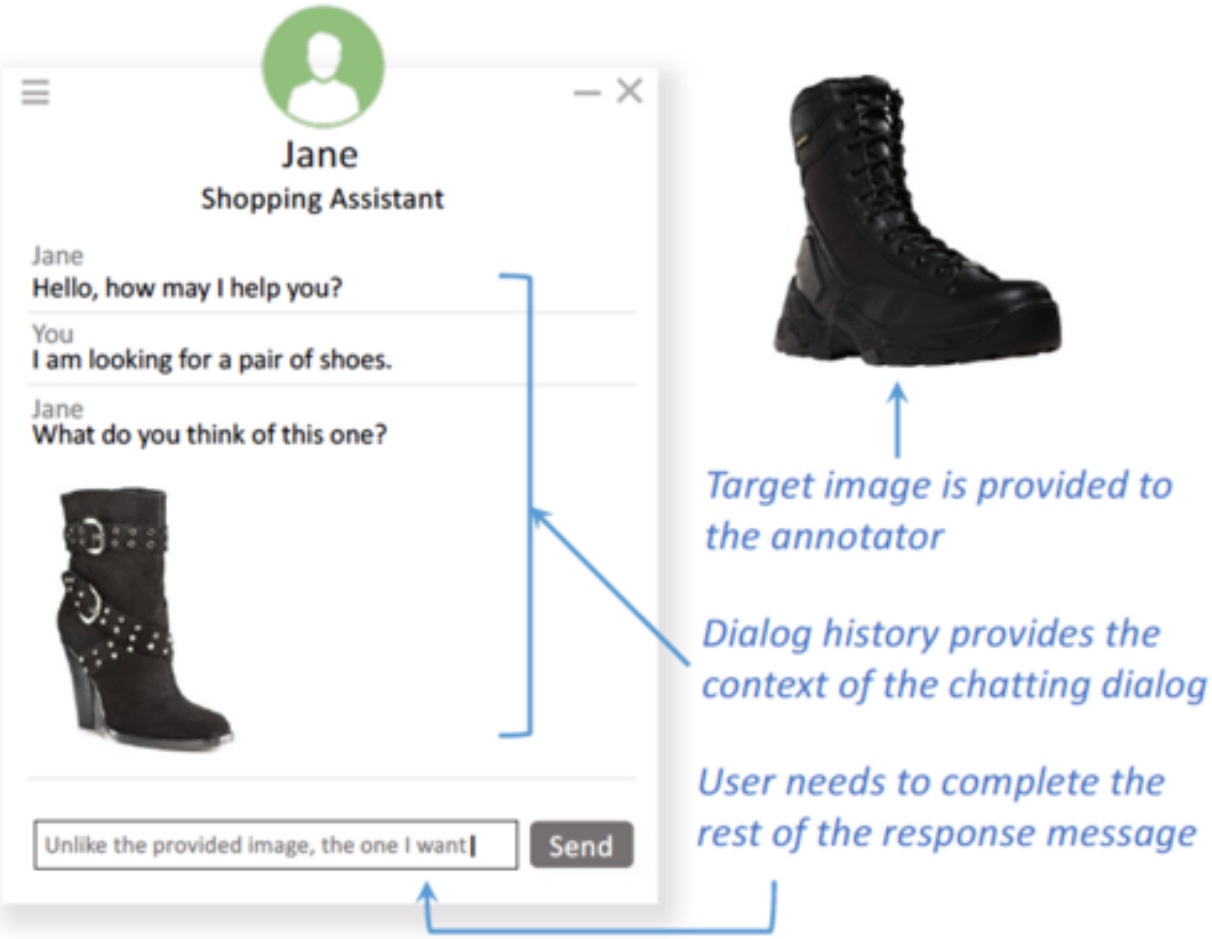
Training the network

- How to obtain training data? Expensive and slow to collect dialog data from real users.

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!

AMT task to collect human-written relative expressions



Shoes Relative Captions Dataset:

- ~10K training images, ~5K testing images
- 1 relative expression per image

Relative Captioner (User Simulator) Model

- Feature concatenation of target and reference images
- Show, Attend, and Tell model [Xu et al, 2015] to generate relative captions

Example predictions:

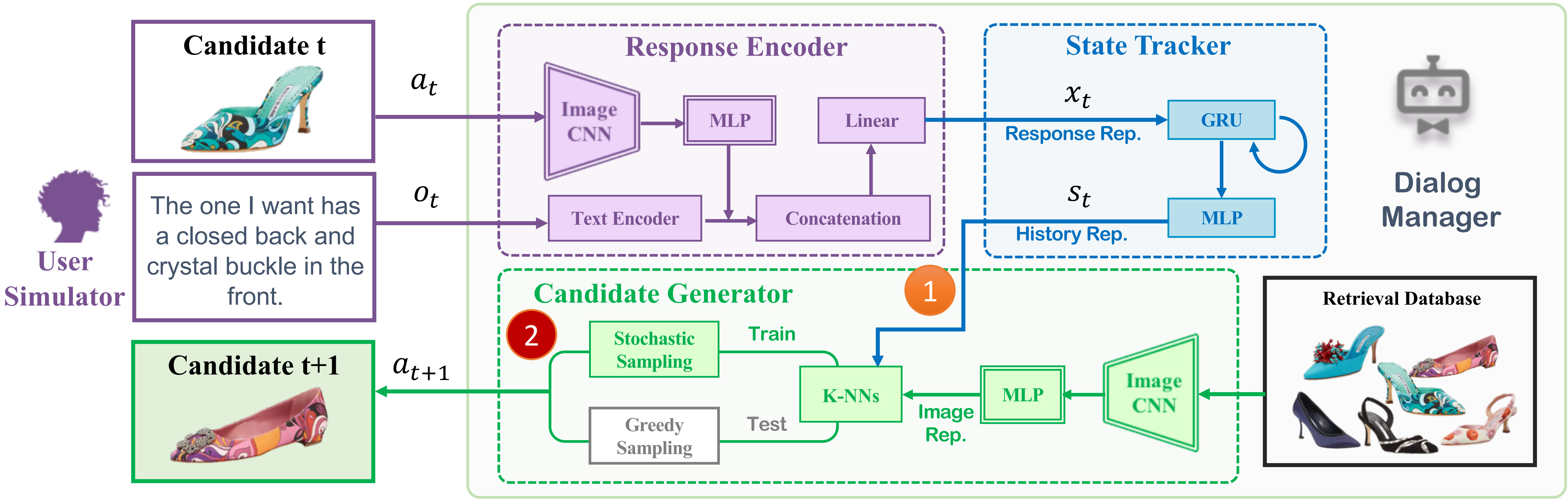
Unlike the provided image, the one(s) I want ***are blue and green sneakers***

are floral print with an all-over floral pattern

are brown with a higher heel

are black with a thicker heel

Training the network



1 Supervised pre-training (triplet loss)

$$\mathcal{L}^{\text{sup}} = \mathbb{E}\left[\sum_{t=1}^{T} \max(0, \|s_t - x^+\|_2 - \|s_t - x^-\|_2 + \mathrm{m})\right]$$

History representation

Target feature

Random image feature

2 Reinforcement Learning to maximize the rank of the target image, with model-based policy improvement

Results

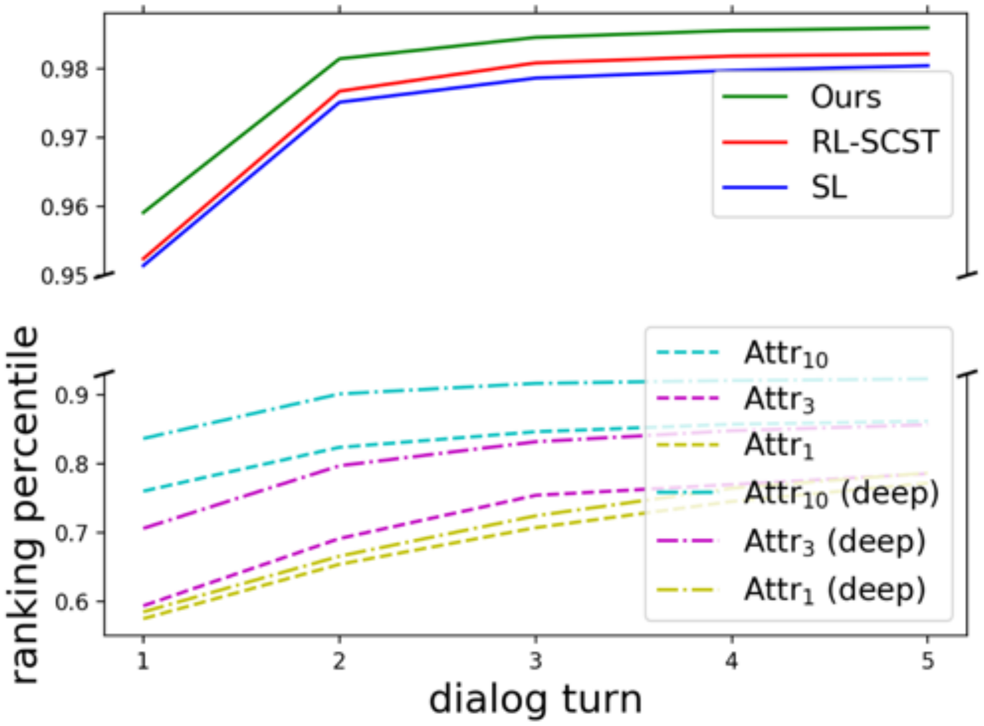
Policy Learning Results

SL: supervised learning where the agent is trained only using triplet loss;

RL-SCST: policy learning using Self-Critical Sequence Training after pre-training using SL.

Effectiveness of Natural Language Feedback

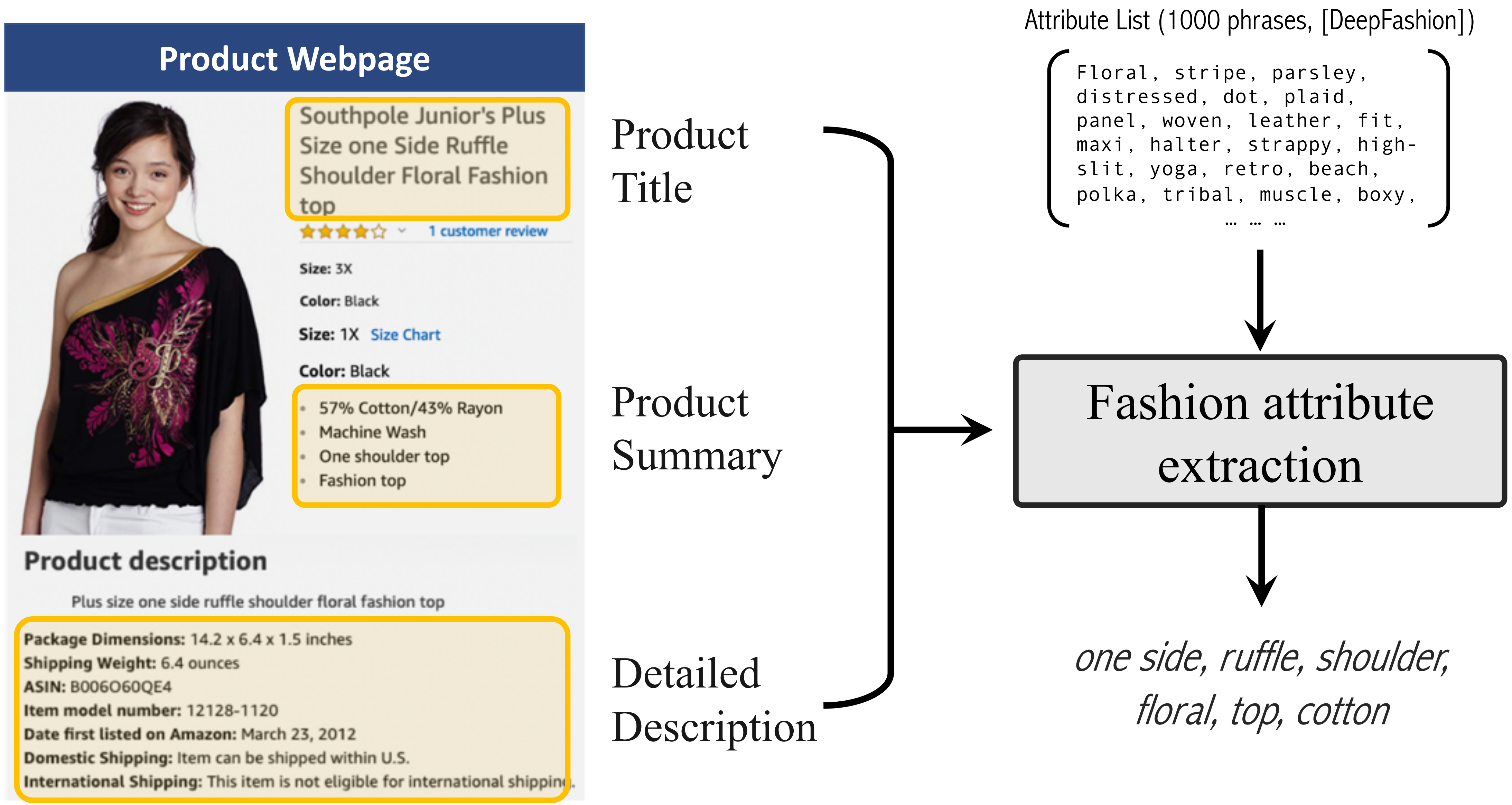
$\mathbf{Attr_n}$ and **$\mathbf{Attr_n}$(deep)**: dialog managers trained with relative attribute feedback . A rule based feedback generator concatenates respective attribute words with “more” or “less”.



Leveraging Side Information

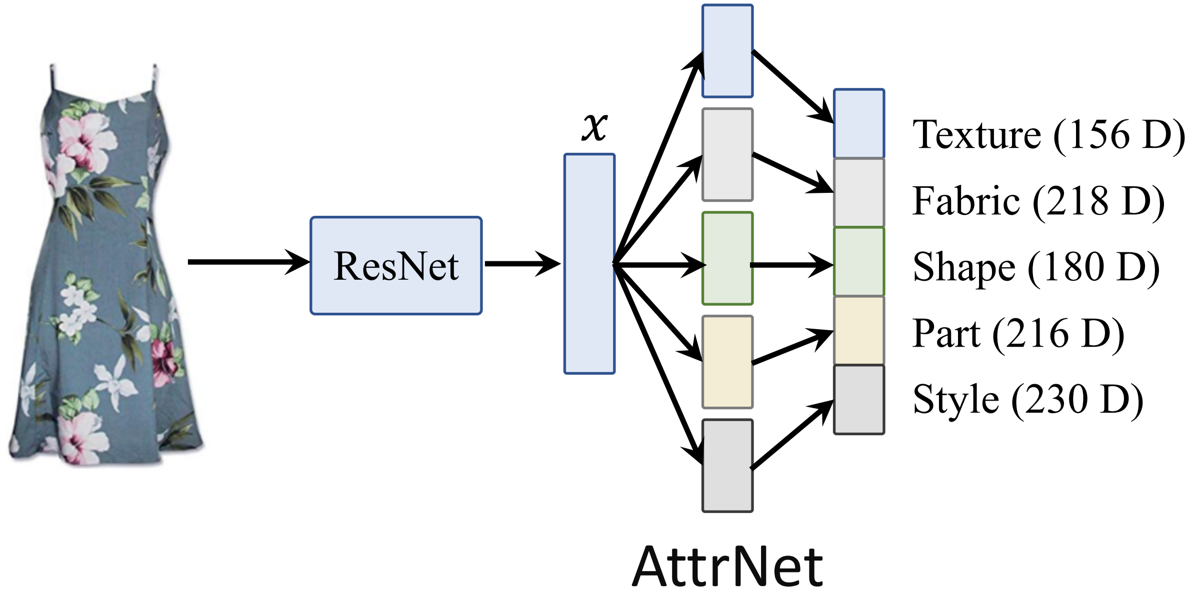
Text surrounding fashion images as weak supervision

Extracting Visual Attributes from Text



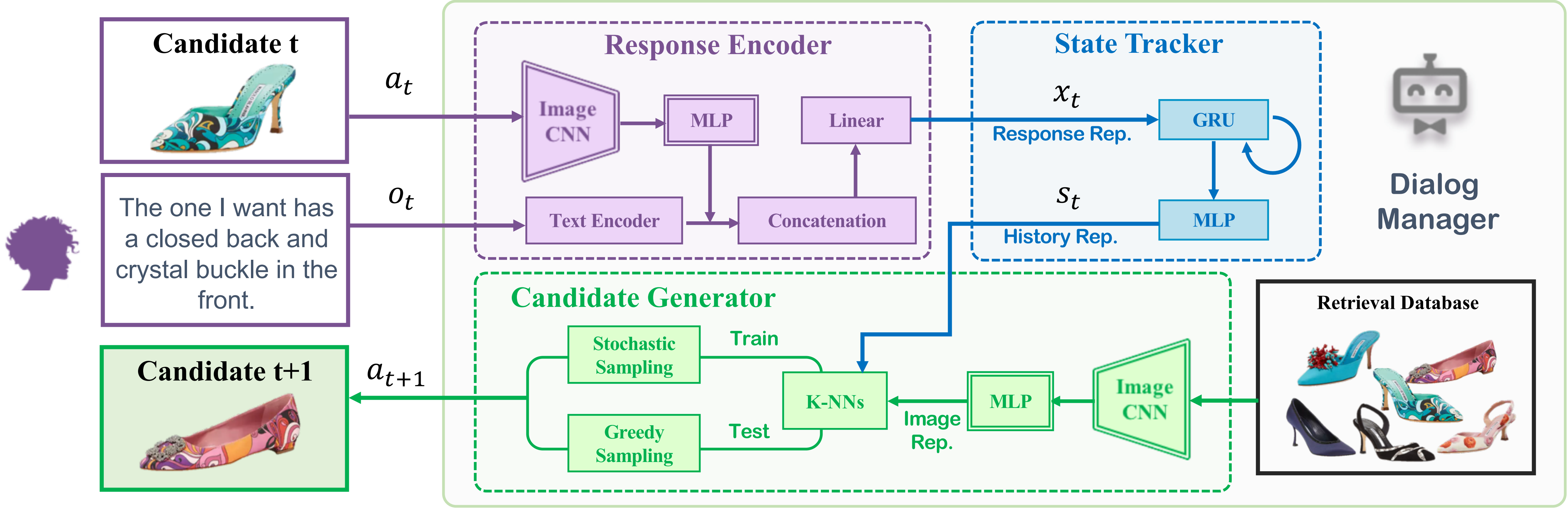
Attribute Prediction Network

- Similar to our DARN work we used an attribute prediction network to obtain attribute-aware visual features
- Use this information as a weak supervisory signal

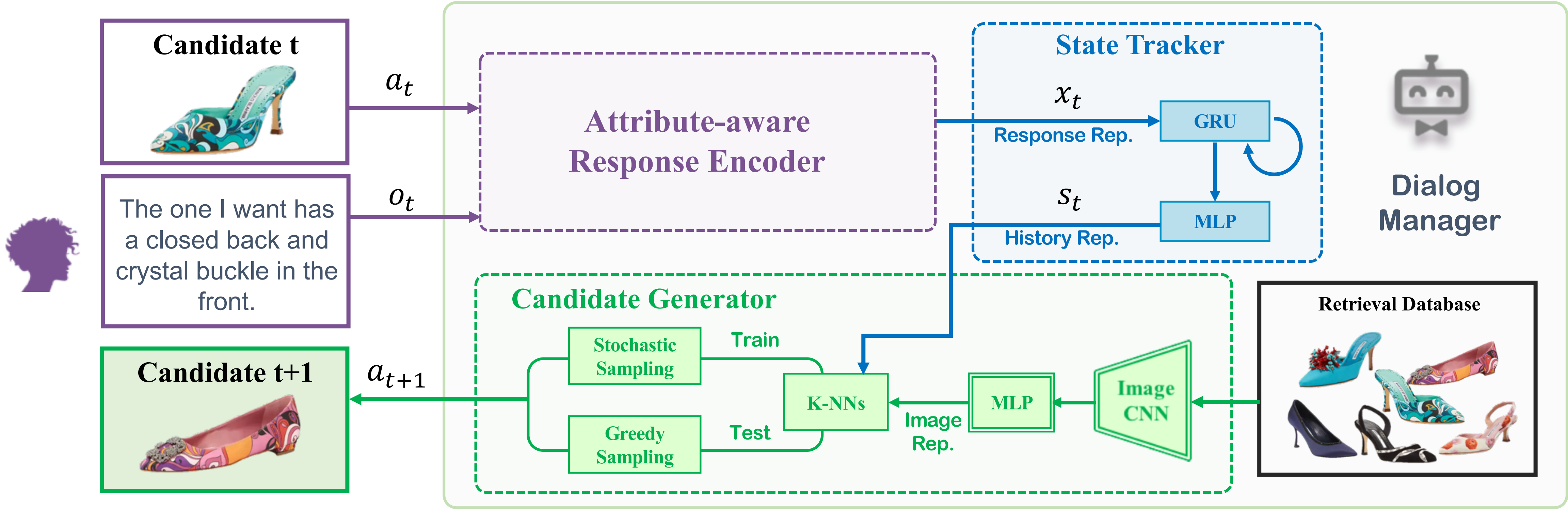


	Dresses		Shirts		Tops&Tees	
	top-3	top-5	top-3	top-5	top-3	top-5
Texture	0.50	0.60	0.69	0.78	0.54	0.65
Fabric	0.45	0.53	0.70	0.76	0.52	0.58
Shape	0.36	0.47	0.69	0.78	0.51	0.61
Part	0.31	0.44	0.51	0.66	0.37	0.49
Style	0.19	0.28	0.26	0.36	0.21	0.28
All	0.36	0.46	0.57	0.66	0.43	0.51

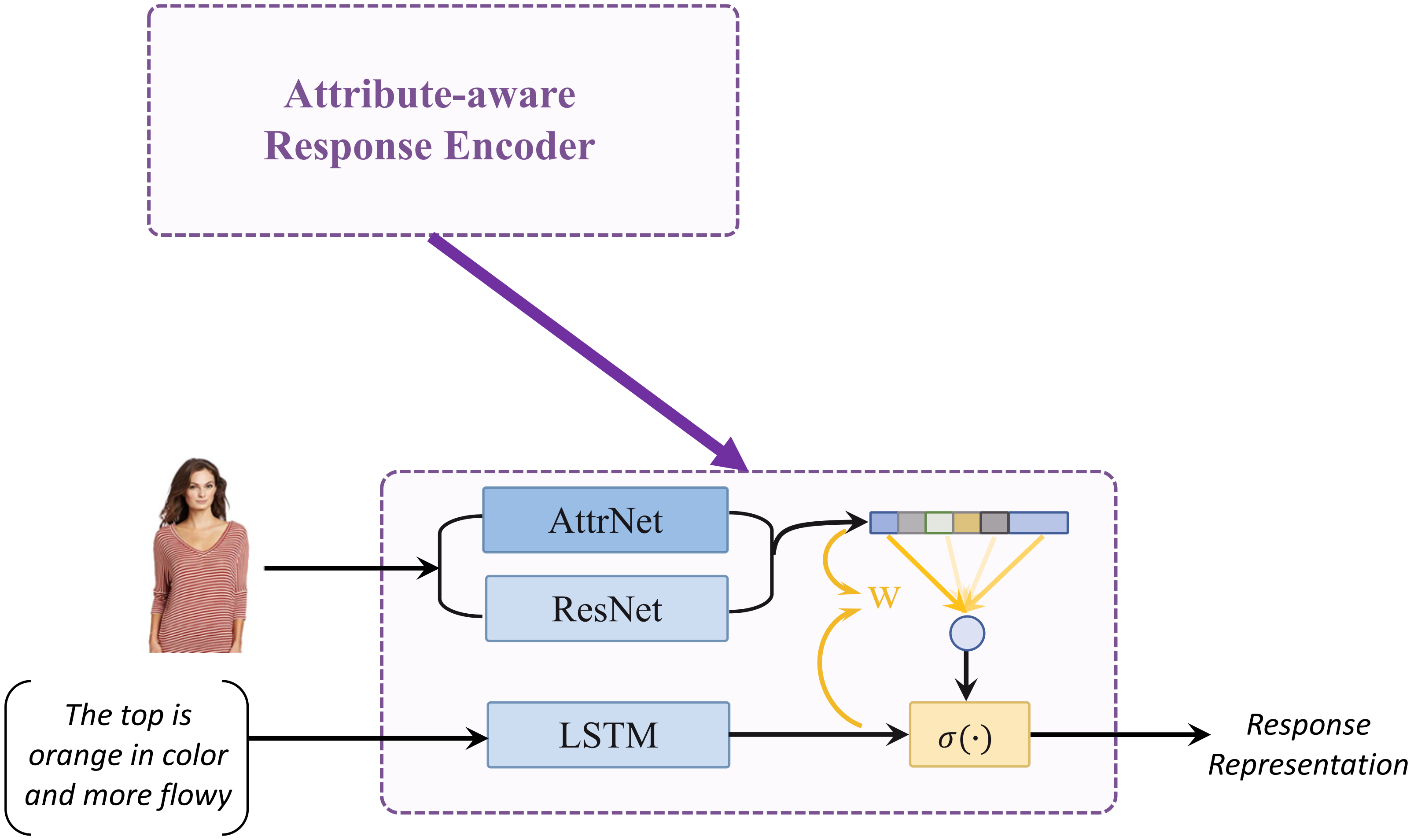
Network Architecture



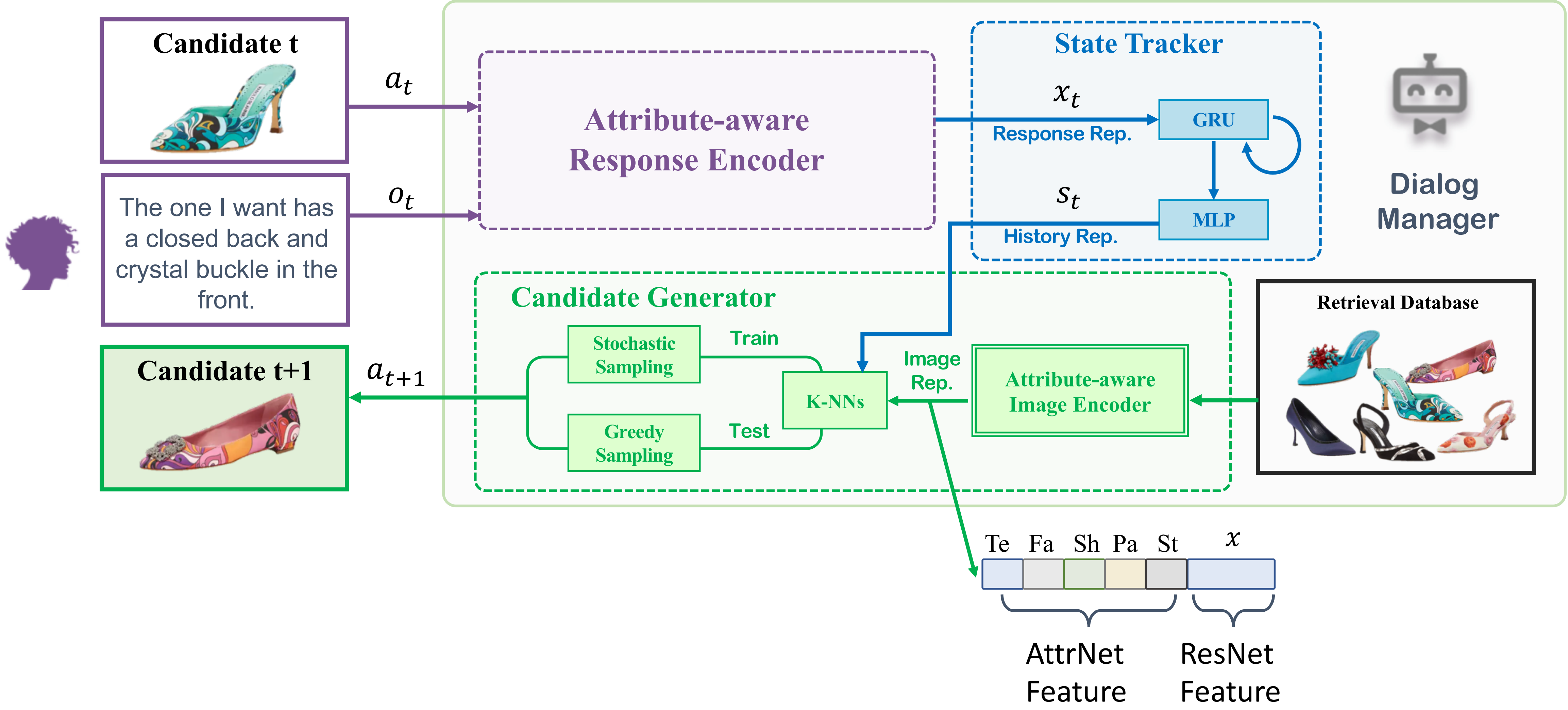
Network Architecture



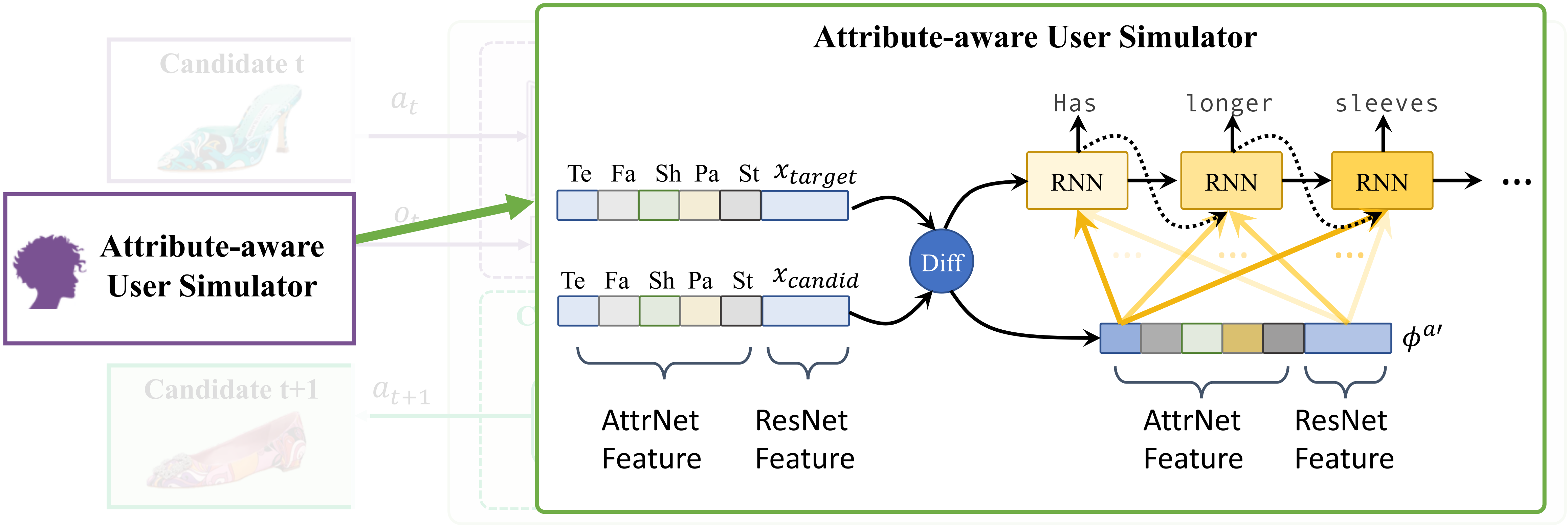
Network Architecture



Network Architecture



Network Architecture



Fashion IQ Dataset

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

- 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

Dresses

Top & Tees

Shirts

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.

Summary

- Natural language user feedback provides a more natural, expressive, and effective way to interactive image search
- Incorporating side information is a low-cost, effective technique to further improve retrieval results

- Challenges ahead
  - The data issue: user simulator does not accurately model real-user behavior (personal preference, fashion expertise, history, …)
  - Users can communicate better if the agent can ask informative questions in addition to showing images

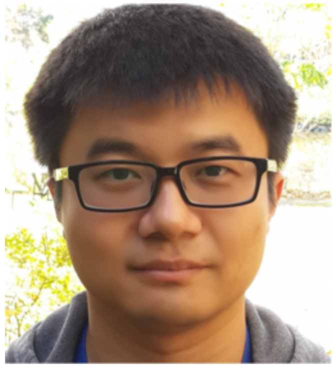
Thank you!

- [ICCV2015] Huang, Junshi, Rogerio S. Feris, Qiang Chen, and Shuicheng Yan. "Cross-domain image retrieval with a dual attribute-aware ranking network."
- [NeurIPS 2018] Guo, Xiaoxiao*, Hui Wu*, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio S. Feris. "Dialog-based Interactive Image Retrieval."
- [Arxiv 2019] Guo, Xiaoxiao*, Hui Wu*, Steven Rennie, and Rogerio S. Feris. "The Fashion IQ Dataset: Retrieving Images by Combining Side Information and Relative Natural Language Feedback"

* (equal contribution)

Hui Wu

Xiaoxiao Guo



Check out the fashion IQ challenge at ICCV 2019!