Deep Multi-task Attribute-driven Ranking for Fine-grained Sketch-based Image Retrieval

,BMVC 2016

Paper presentation
2018. 11 .22

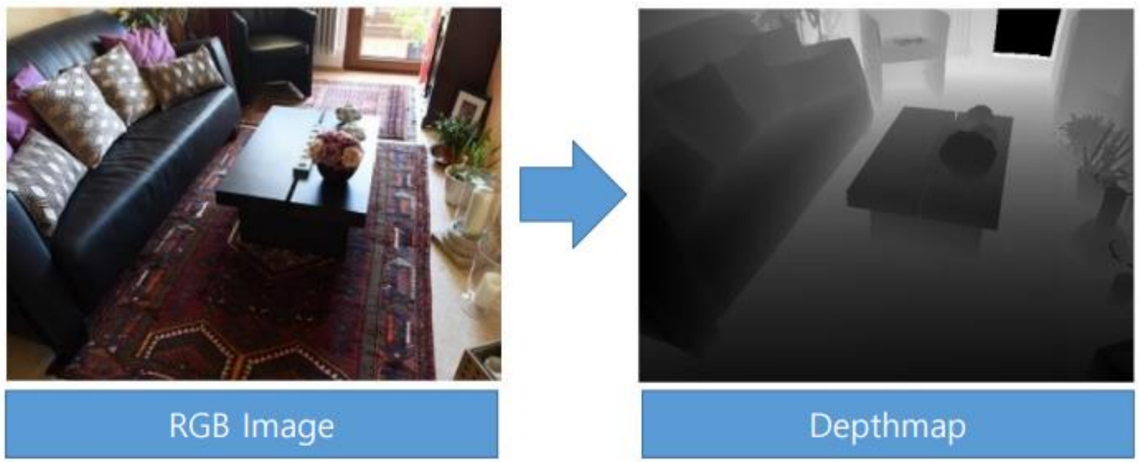
Taeun Hwang (황태운)

CS688: Web-scale image retrieval

KAIST

Review presentor: Taehee kim

- **Evaluation of CNN-based Single-Image Depth Estimation Methods, CVPR 18**



- In depthmap estimation about single image
  - Introduce a set of new error metrics
  - Present a new dataset from laser scan
  - Evaluate state-of-art methods

Contents

- Introduction
- Method
- Experiment & result

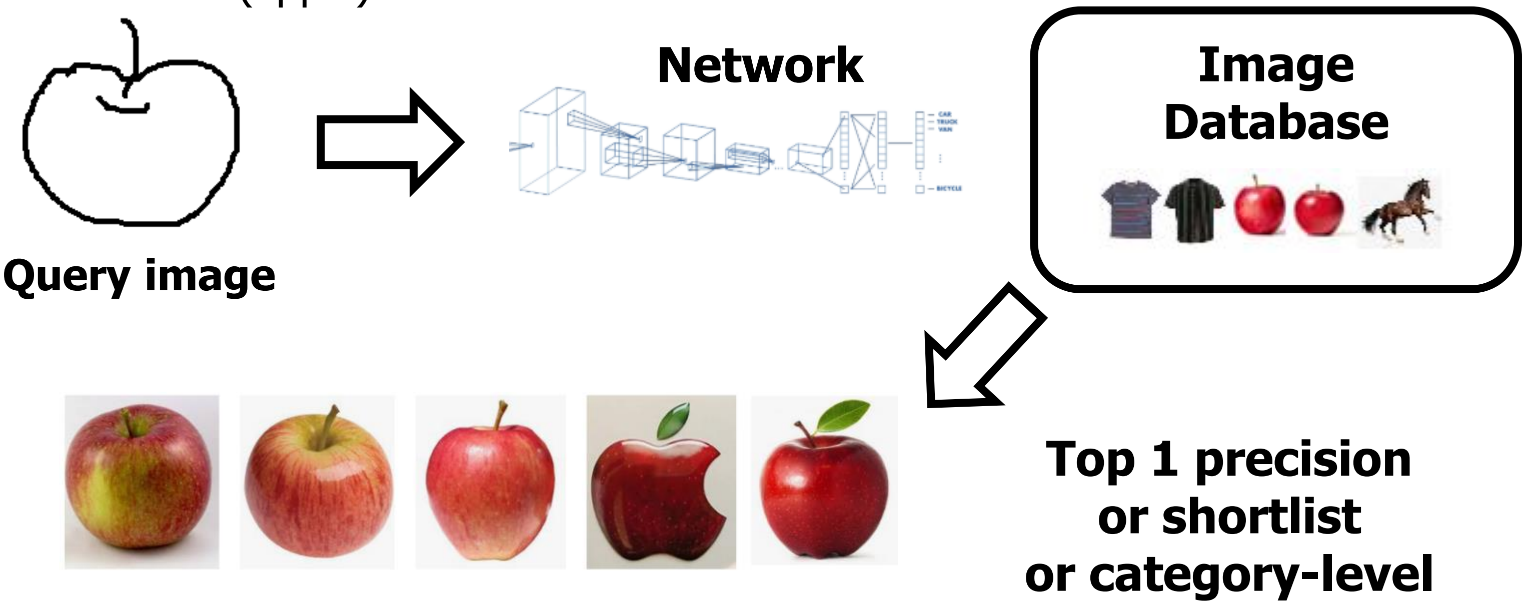
Introduction

KAIST

Introduction

- **Sketch-Based Image Retrieval(SBIR)**

Free-hand sketch (apple)

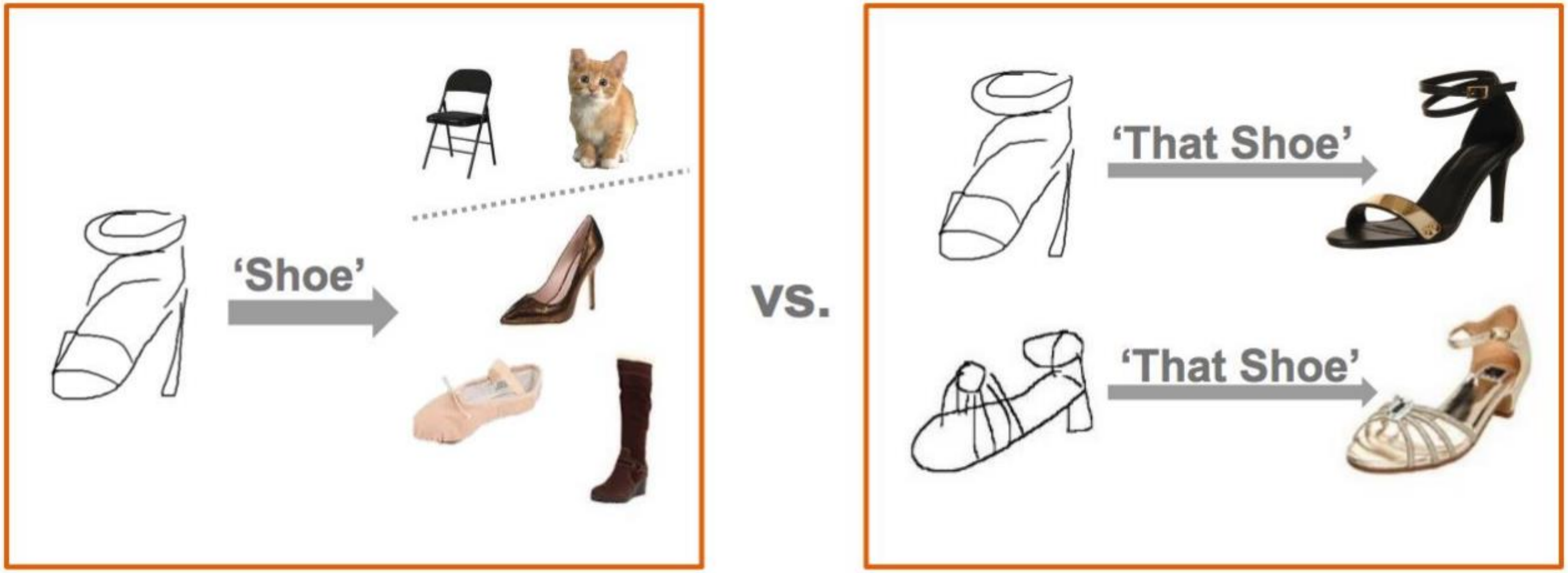


- People can quickly draw abstractly
- Sketch : have visual details

Introduction

- **Category-level SBIR** vs **Fine-grained SBIR**

From heechan's slide



Category-level SBIR

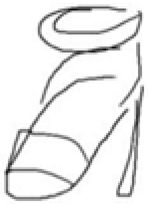
fine-grained SBIR

Just find category
→ More clearly and easily using "text" not "SBIR"

Find more detail
→ Top 1 or Top 10 precision

Purpose of this paper

- **Improve performance of**
  Fine-grained Sketch-based image retrieval
  - What **meaningful object properties** in sketch?
  - Exploits **Semantic attributes**
    - Ex) Shoe is high-heeled?
      Shoe has Shoelace?



    - Ex) Chair has arm-rest?

- **Learning Semantic attributes**

- **Method**

Method

- Perform 3-task deep learning
  - **Retrieval by fine-grained ranking**
  - **Attribute prediction**
    - Example of Attribute : Shoe is high-heeled?
  - **Attribute-level ranking**

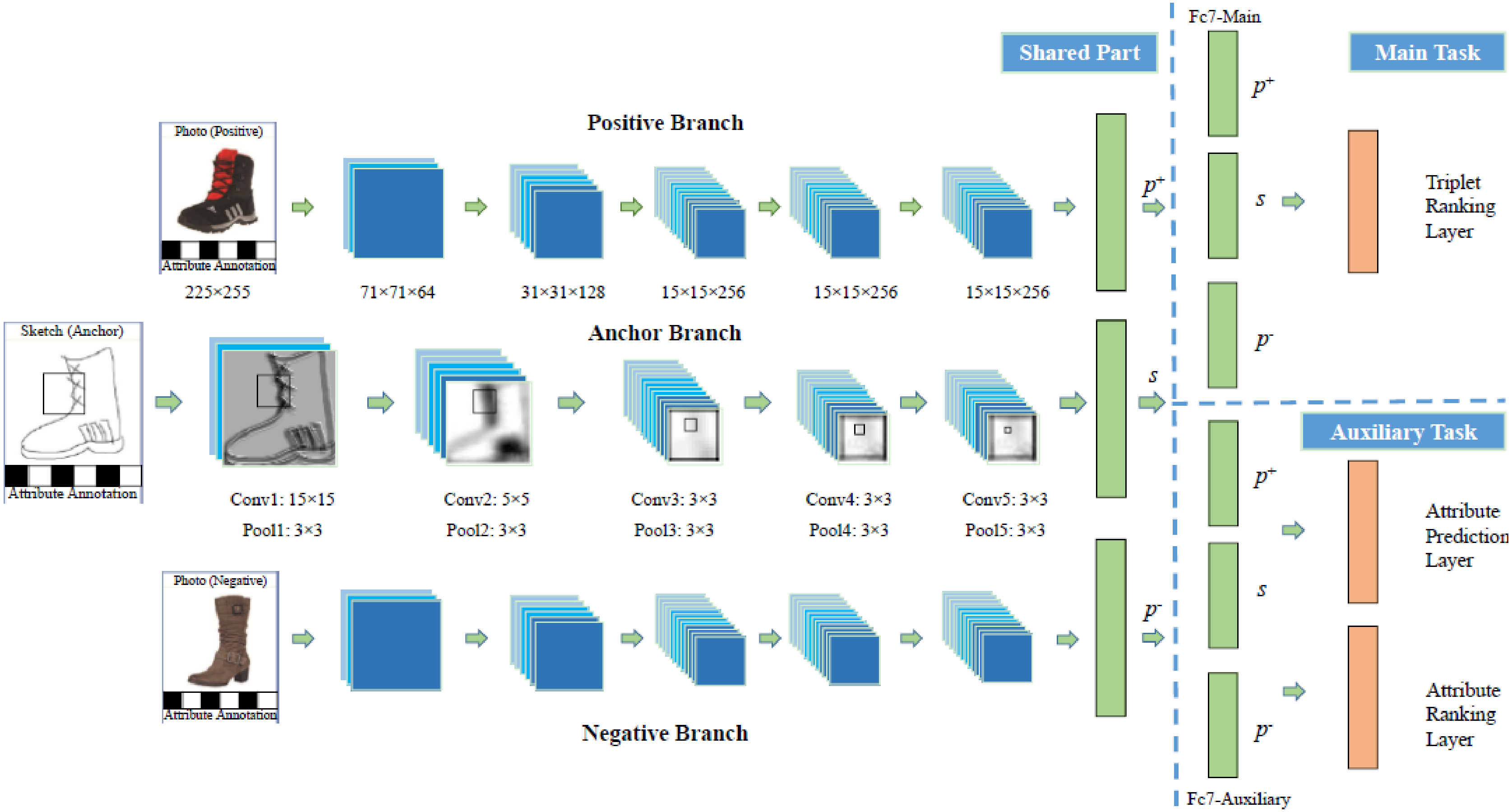
- Predicting **semantics attribute** and using this in the ranking procedure

  → Retrieval results to be more **semantically relevant**

KAIST

Network architecture

- **Multi-task : Do 3 tasks**



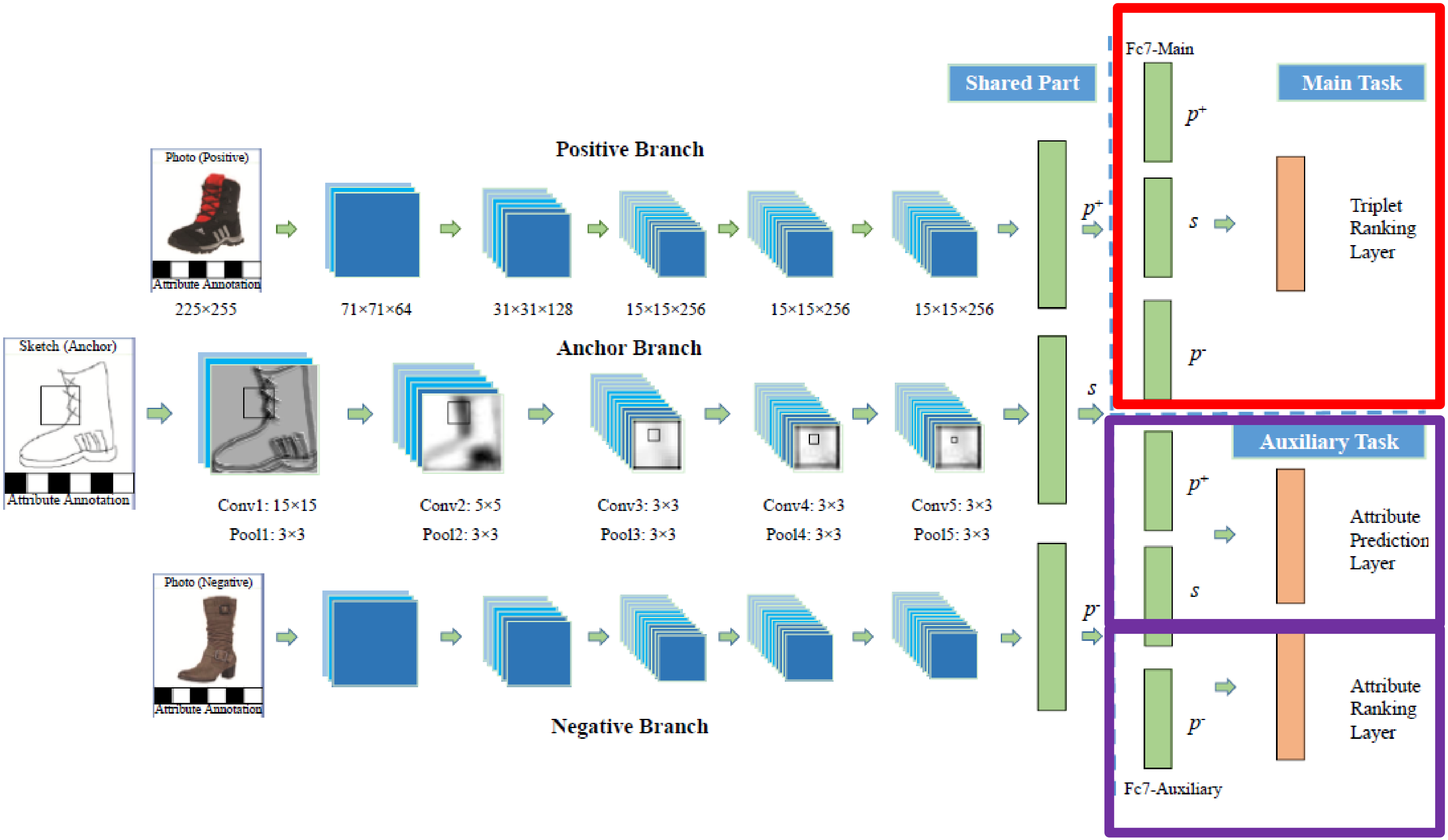
Multi-tasks

- **1. Main Triplet Ranking Task**
  - Main task : sketch-photo ranking

- **2. Attribute prediction Task (subtask)**
  - Predict semantic attributes
  - Example of attribute
    - Shoe is high-heeled
    - Chair has arm-rest
- **3. Attribute Ranking Task (subtask)**
  - Attribute-level sketch-photo matching

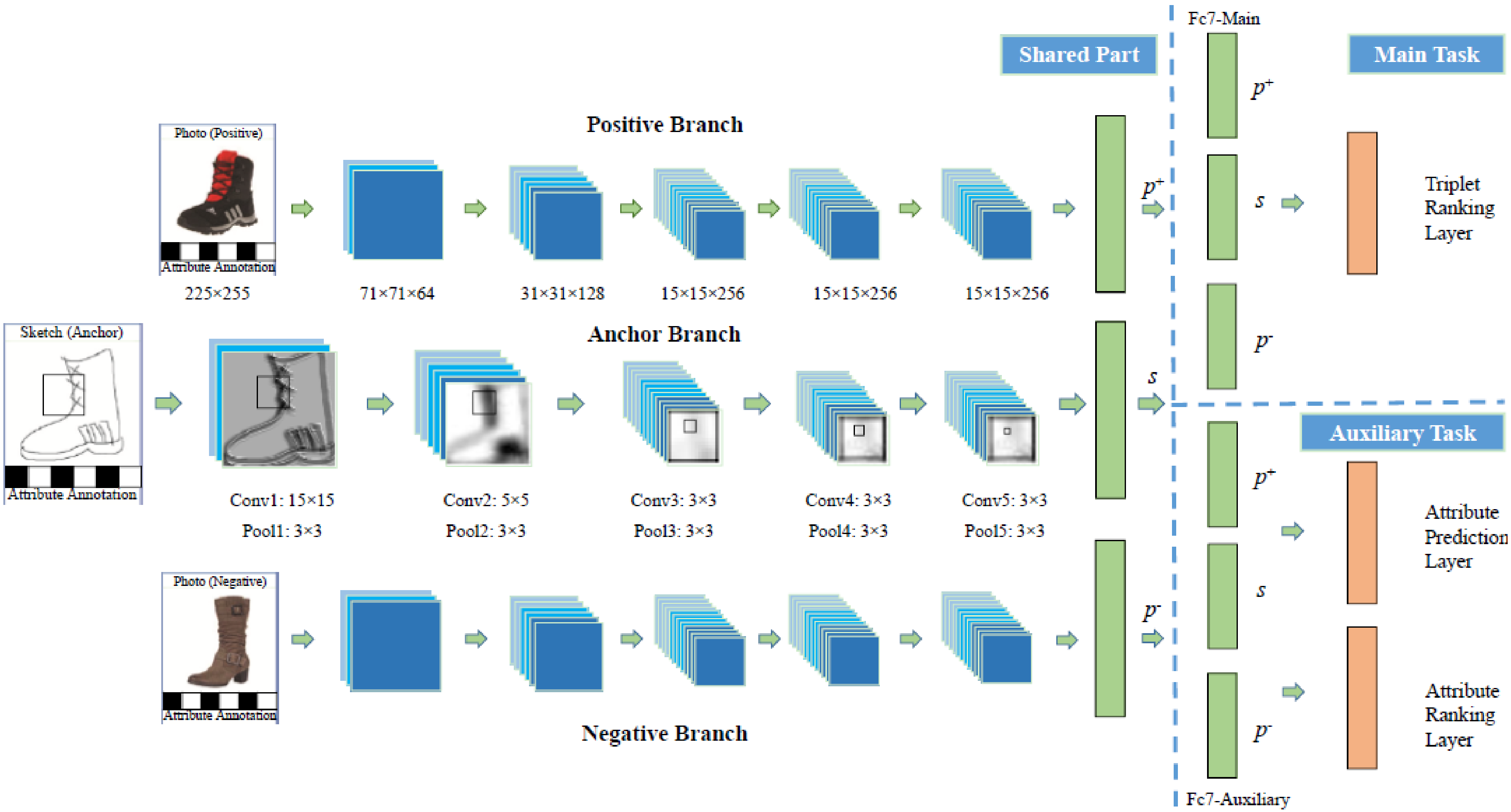
Overall network

1. Main Triplet Ranking Task

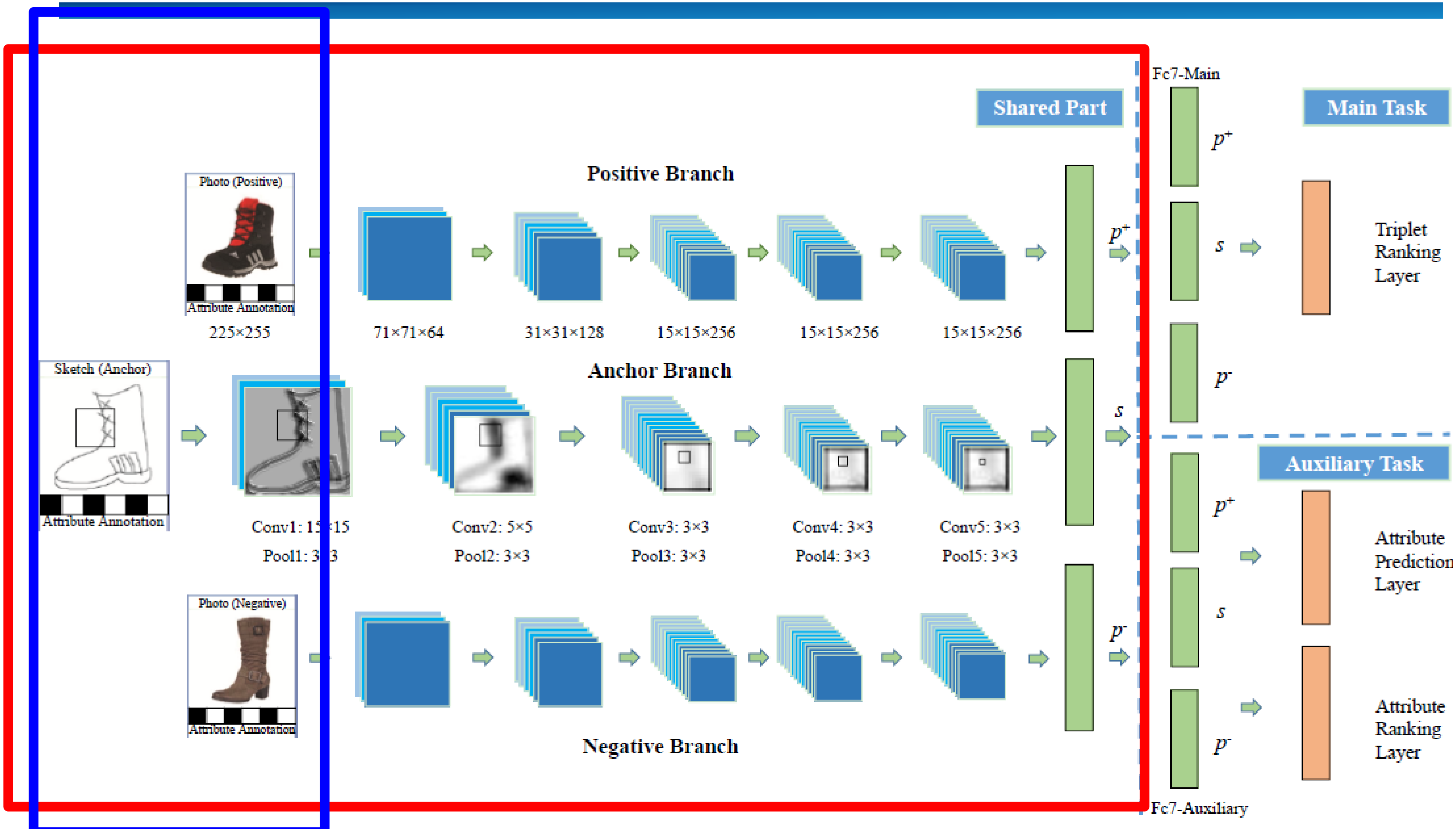


2, 3 sub task

Overall network



Overall network

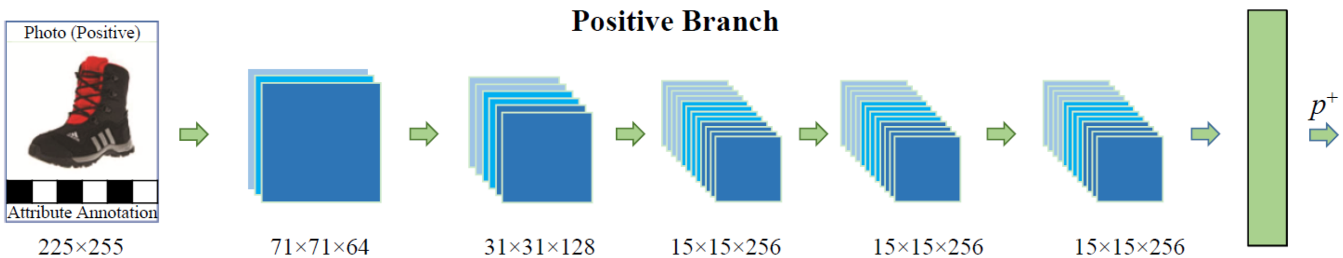


Input tuple : Sketch, P, N

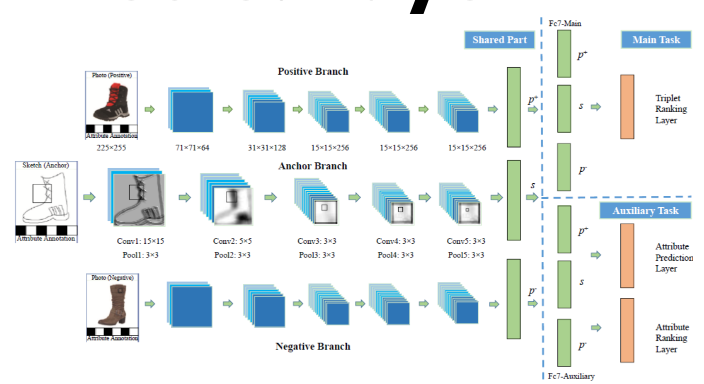
Task-Shared part

Task-Shared part

- There are **Three branch**
  - **For Sketch, positive image, negative image**

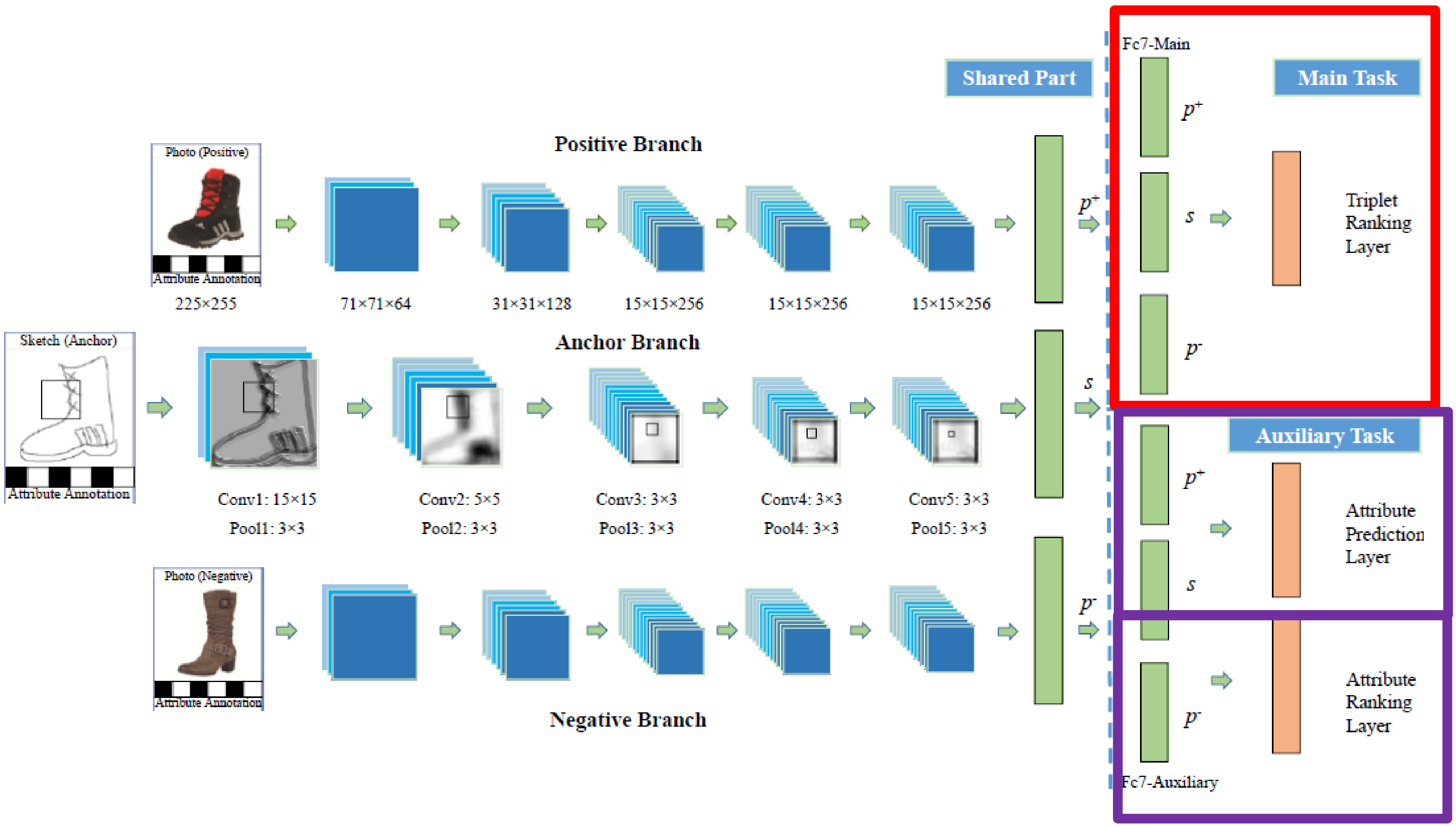


- Each branch consists of **five convolution layers** with max pooling + **a fully-connected layer**
  - **Make feature map**



Overall network

1. Main Triplet Ranking Task

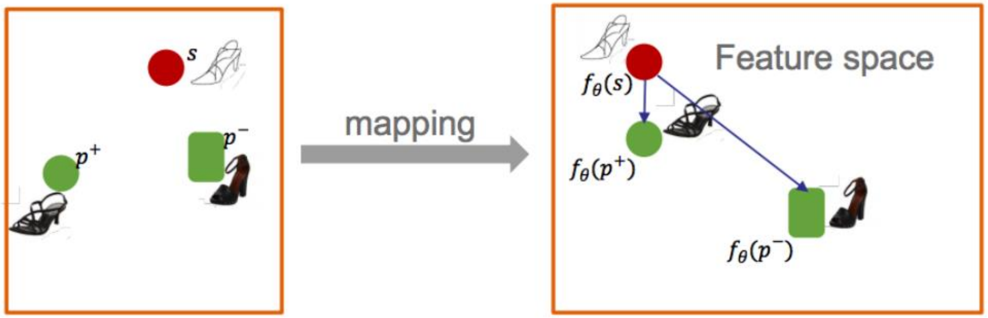


+ FC layer with dropout and RELU

2, 3 sub task

1. Main Triplet Ranking Task

- Trained by supervision in the form of triplet tuples
- Goal to learn : **p+** is **ranked above** the **p-**



- Loss function : **triplet ranking loss**

$$L_\theta\left(s,p^+,p^-\right) = \max\left(0, \Delta + D\left(f_\theta\left(s\right), f_\theta\left(p^+\right)\right) - D\left(f_\theta\left(s\right), f_\theta\left(p^-\right)\right)\right)$$

Triplet tuple instance :
Sketch **s**
Positive photo **p+**
Negative photo **p-**

f : feature
D : euclidean distance
Δ : margin

KAIST

2. Attribute prediction Task

- Predict semantics attributes (both sketch, image)
- Assume N different semantic attributes t
  - Training tuples for sketch : $\{s, t_1^s \ldots t_N^s\}$

- Attribute prediction loss : **cross-entropy** between **attribute label** and **prediction f**

$$L_p(s,t^s) = -\frac{1}{N}\sum_{n=1}^{N}\left[t_n^s \log f_{\theta,n}^{ap}(s) + (1-t_n^s)\log\left(1-f_{\theta,n}^{ap}(s)\right)\right]$$

similar for p+ and p- photos

- Trained simultaneously with the 1. main task

KAIST

3. Attribute Ranking Task

- 2. Attribute prediction task : would not be used in test-time
  - Also, not use Attributes

- But Attributes are good information for SBIR

- So, Attributes similarity between sketch and p+ used as a loss function

$$L_a\left(s, p^+, p^-\right) = H\left(f_\theta^{ap}(s), f_\theta^{ap}\left(p^+\right)\right)$$

H: cross-entropy

Multi-task Training and Testing

- Overall loss function for multi-task training

$$L\left(s,p^{+},p^{-}\right)=L_{\theta}\left(s,p^{+},p^{-}\right)+\lambda_{a}L_{a}\left(s,p^{+},p^{-}\right)$$
$$+\lambda_{s}L_{p}\left(s,t^{s}\right)+\lambda_{p^{+}}L_{p}\left(p^{+},t^{p^{+}}\right)+\lambda_{p^{-}}L_{p}\left(p^{-},t^{p^{-}}\right)$$
$$+\lambda_{\theta}\|\theta\|_{2}^{2}$$

weight hyper parameters $\lambda=\left(\lambda_{a},\lambda_{s},\lambda_{p^{+}},\lambda_{p^{-}}\right)$

- In test-time
  - Given query sketch s, the similarity of each image p in gallery is

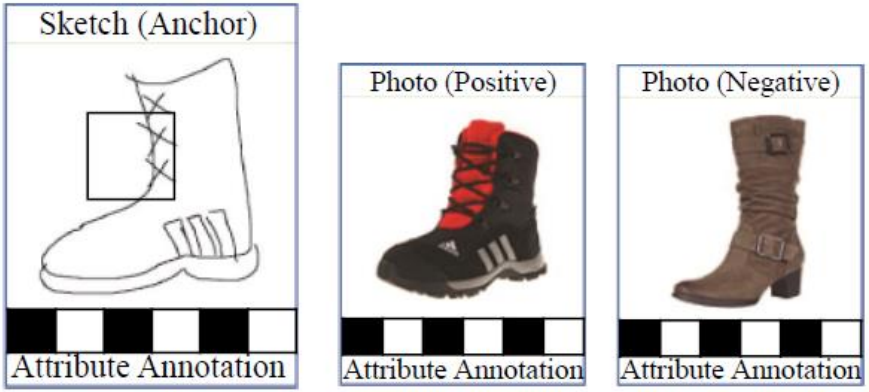
$$R_{s}\left(s,p\right)=D\left(f_{\theta}\left(s\right),f_{\theta}\left(p\right)\right)+\lambda_{a}H\left(f_{\theta}^{ap}\left(s\right),f_{\theta}^{ap}\left(p\right)\right)$$

D : euclidean distance
H : cross-entropy

Attribute-based sampling Strategy

- **Staged model pre-training strategy**

- **Attribute-based sampling Strategy**
  - **Triplet generation**
  - **Triplet sampling**



- **Experiment & result**

Experiments

- **Training and Evaluation Data**
  - **304 sketch-photo pairs of shoes**
  - **200 sketch-photo pairs of chairs**
    - Same dataset used in sketch-me-that-shoe

- **Evaluation metrics**
  - Top-K retrieval accuracy, K=1 K=10

Result

Figure 3: Retrieval results of our proposed method, compared with that of [19].

- Triplet Model : Sketch me that shoe, CVPR 16

Result

- Compare to other retrieval methods

Table 1: Comparative results against state of the art retrieval performance.

Shoe Dataset	top 1	top 10	trip-acc	Chair Dataset	top 1	top 10	trip-acc
BoW-HOG + rankSVM	17.39%	67.83%	62.82%	BoW-HOG + rankSVM	28.87%	67.01%	61.56%
Dense-HOG + rankSVM	24.35%	65.22%	67.21%	Dense-HOG + rankSVM	52.57%	93.81%	68.96%
ISN Deep + rankSVM	20.00%	62.61%	62.55%	ISN Deep + rankSVM	47.42%	82.47%	66.62%
3DS Deep + rankSVM	5.22%	21.74%	55.59%	3DS Deep + rankSVM	6.19%	26.80%	51.94%
Triplet model [19]	39.13%	87.83%	69.49%	Triplet model [19]	69.07%	97.94%	72.30%
Ours	**50.43%**	**91.30%**	**70.59%**	Ours	**78.35%**	**98.97%**	**73.13%**

- Comparison of w/o Attribute tasks usage

Table 2: Contribution of the proposed attribute side tasks.

Shoe Dataset	top 1	top 10	trip-acc	Chair Dataset	top 1	top 10	trip-acc
Ours - AP - AR	37.39%	82.61%	66.57%	Ours - AP - AR	50.52%	91.75%	69.62%
Ours - AR	45.22%	87.83%	**72.37%**	Ours - AR	72.16%	98.97%	72.00%
Ours - AP	44.35%	86.96%	71.34%	Ours - AP	72.16%	98.97%	72.10%
Ours	**50.43%**	**91.30%**	70.59%	Ours	**78.35%**	**98.97%**	**73.13%**

End

- **QnA**