SCH-GAN: Semi-Supervised Cross-Modal Hashing by Generative Adversarial Network

Jian Zhang; Yuxin Peng; Mingkuan Yuan
IEEE Transactions on Cybernetics, Vol. 50, No. 2, February 2020

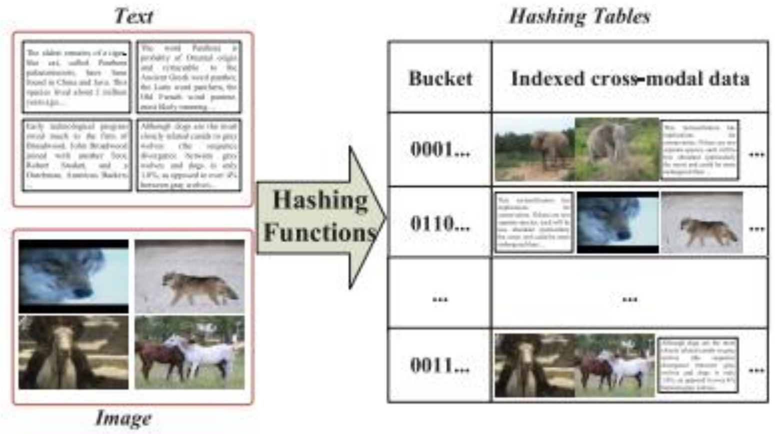
2022.05.23 Fikriansyah Adzaka

Vision Language Intelligence Lab

Sejong University

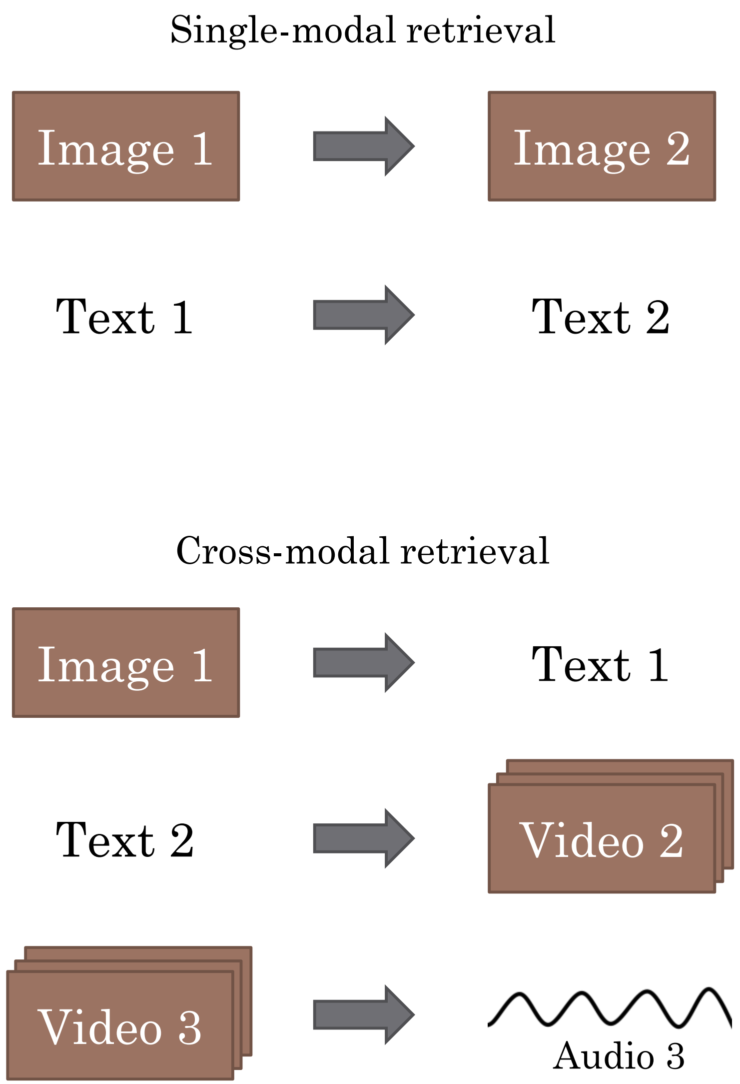
Background

- Efficient retrieval of multimedia data from **large scale databases** has become an urgent need and a big challenge
  - Solution: use **hashing methods**
- Hashing methods aim to transfer high dimensional features into **short binary codes**
  - Similar data can have similar binary codes
- Advantages: **fast retrieval with less storage**
- Disadvantages: **collisions**



Background

- Most hashing methods are designed for single modality retrieval
  - However, multimedia data are usually presented with different modalities
- Retrieving data across different modalities is called **cross-modal retrieval**
  - This way, users can retrieve whatever they want by submitting whatever they have
- Challenge: **heterogeneity gap**
  - Text: $\mathbb{R}^{Token}$
  - Audio: $\mathbb{R}^{Sample \times Channel}$
  - Images: $\mathbb{R}^{Height \times Width \times Channel}$
  - Videos: $\mathbb{R}^{Frame \times Height \times Width \times Channel}$



Background

- Cross-modal hashing classification:
  - Supervised: learn the hash function that preserve the semantic correlations provided by the labels
  - Unsupervised: project the data from different modalities into a common space, then find the hash function that maximize the correlation
- Supervised methods typically achieve better retrieval accuracy, but **very labor intensive** to obtain the labels
  - It is even harder to label cross-modal data since there are multiple modalities
- Solution: exploit the unlabeled data too via **semi-supervised learning**

- Why semi-supervised learning?
  - Reduce the needs to label new data
  - Can prevent model overfitting
- However, few efforts have been done for semi-supervised cross-modal hashing
  - How to exploit informative unlabeled data to promote hashing learning?
- One popular method for semi-supervised learning is by using Generative Adversarial Network (GAN)
- This paper propose a novel Semi-supervised Cross-modal Hashing approach by GAN (SCH-GAN)

Notation and Problem Formulation

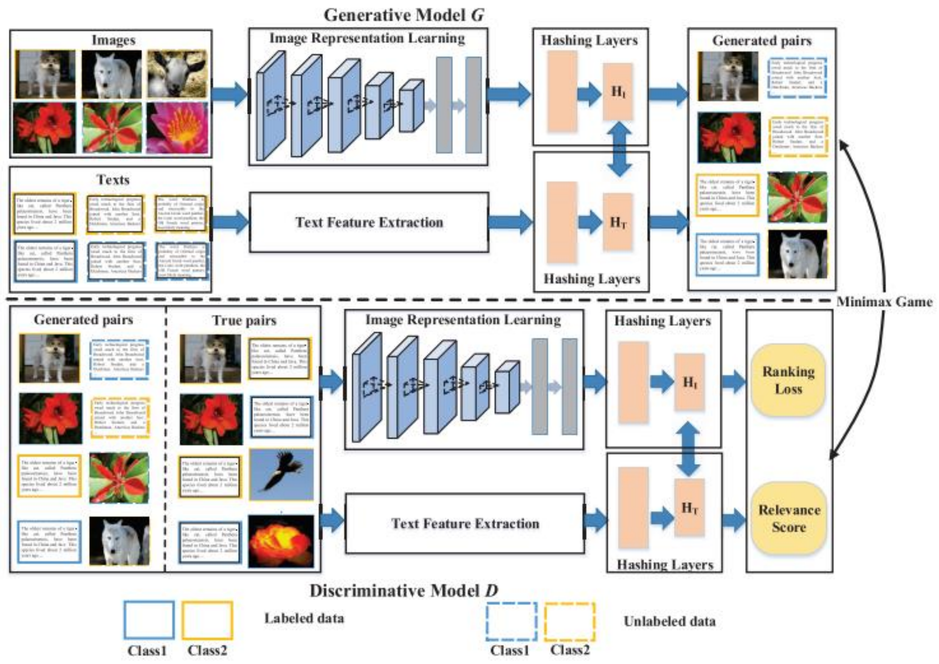
- I and T, list of **I**mage and **T**ext
- $D = \{I, T\}$, the multimodal **D**ataset
- D is further split into D_{db} and D_q
  - D_{db}, the **d**ata**b**ase set, consist of
    - $D_{db}^U = \{I_{db}^U, T_{db}^U\}$, the **U**nlabeled data
      - $I_{db}^U = \{i_p^U\}_{p=1}^m$, **m** unlabeled individual image
      - $T_{db}^U = \{t_p^U\}_{p=1}^m$, **m** unlabeled individual text
    - $D_{db}^L = \{I_{db}^L, T_{db}^L\}$, the **L**abeled data
      - $I_{db}^L = \{i_p^L\}_{p=1}^n$, **n** labeled individual image
      - $T_{db}^L = \{t_p^L\}_{p=1}^n$, **n** labeled individual text
      - $\{c_p^I\}_i^n$, corresponding **c**lass for each image
      - $\{c_p^T\}_i^n$, corresponding **c**lass for each text
    - $m \gg n$
  - $D_q = \{I_q, T_q\}$, the **q**uery set
    - $I_q = \{i_p\}_{p=1}^t$, **t** individual query image
    - $T_q = \{t_p\}_{p=1}^t$, **t** individual query text
- The goal of cross-modal hashing is to learn two mapping functions that maps into the common **H**amming space:
  - $H_I : \mathbb{R}^I \to \mathbb{R}^H$
  - $H_T : \mathbb{R}^T \to \mathbb{R}^H$
- Idea: semantically similar data of different modalities should be close in H
  - Once the mapping function is learned, given a query of any modality, we can retrieve the closest data in H
- Goal: can we leverage D_{db}^U to train H?

Network Overview

- In general, GAN consist of two models that compete with one another: generative and discriminative models
    - These two models play a minimax game to iteratively optimize each other, hence increasing the accuracy

- In this paper, the generative model learns to fit the **relevance distribution** of the unlabeled data and select **margin examples** from the unlabeled data based on the query

- The discriminative model learns to **distinguish** the selected data from generative model and the true positive data

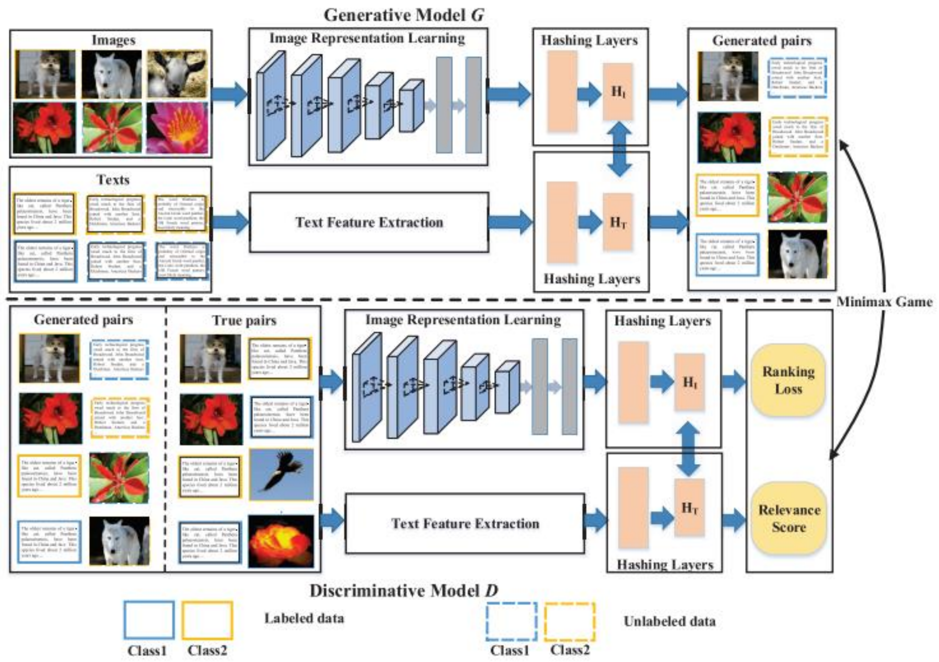


Generative Model

- Input: $I_{db}^{U}, T_{db}^{U}, I_q, T_q$
- Feature representation layers
  - CNN for images
  - Bag-of-words for texts
- Hashing layers
  - Fully-connected for both modalities
    - $f(x)$, the first layer, maps the input into a common space
    - $h(x) \in [0,1]^q$, the second layer, is the hashing functions with dimension q

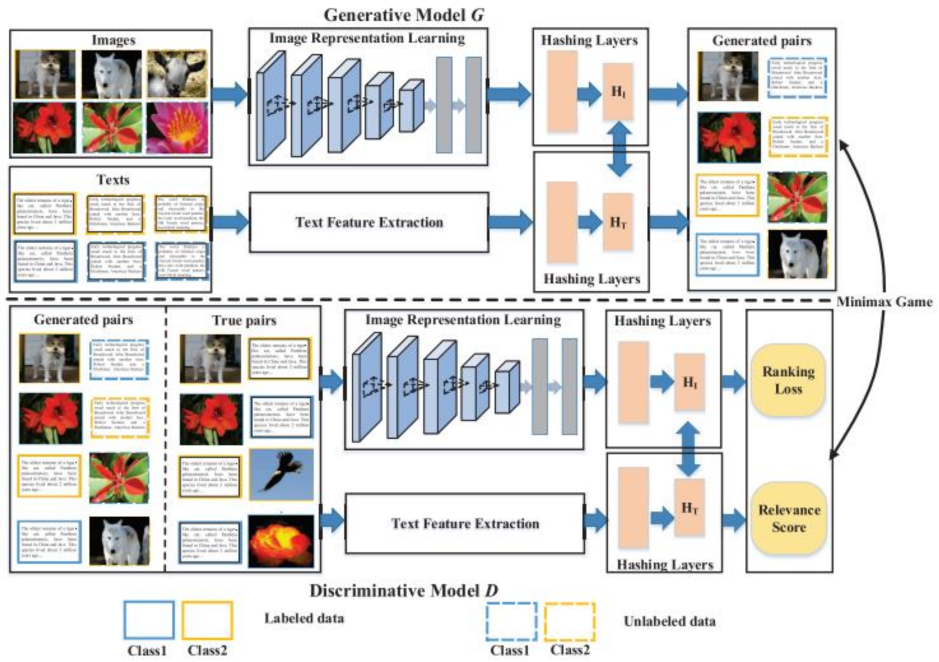
$$h(x) = \text{sigmoid}\left(W^T f(x) + v\right)$$

      - W and v, is the weights and bias
- Similarity calculation between different modalities using Hamming distance
- Output: $i^U | q_t$ and $t^U | q_i$



Discriminative Model

- Input: $(i, q_t), (t, q_i)$
- Same feature representation layers, hashing layers, and similarity calculation with generative model
- Output: [0,1]
  - If it is the generated pairs, the value should be as close to 0 as possible
  - If it is the true pairs, the value should be as close to 1 as possible



Objective Function

- Let:
  - p_θ, the generative model
  - p_{true}, the true distribution
  - $f_\emptyset$, the discriminative model
  - $\theta, \emptyset, r$, the model parameters
- The adversarial process is a minimax game for both image and text query
- Because both equations are symmetric, from now we will see the text query in detail, and that should also apply to image query

Text query:

$$\mathcal{V}(G, D) = \min_{\theta} \max_{\phi} \sum_{j=1}^{n} \Big(E_{i \sim p_{\text{true}}\left(i^L | q_t^j, r\right)} \left[\log\left(D\left(i^L | q_t^j \right) \right) \right] + E_{i \sim p_\theta\left(i^U | q_t^j, r\right)} \left[\log\left(1 - D\left(i^U | q_t^j \right) \right) \right] \Big).$$

Image query:

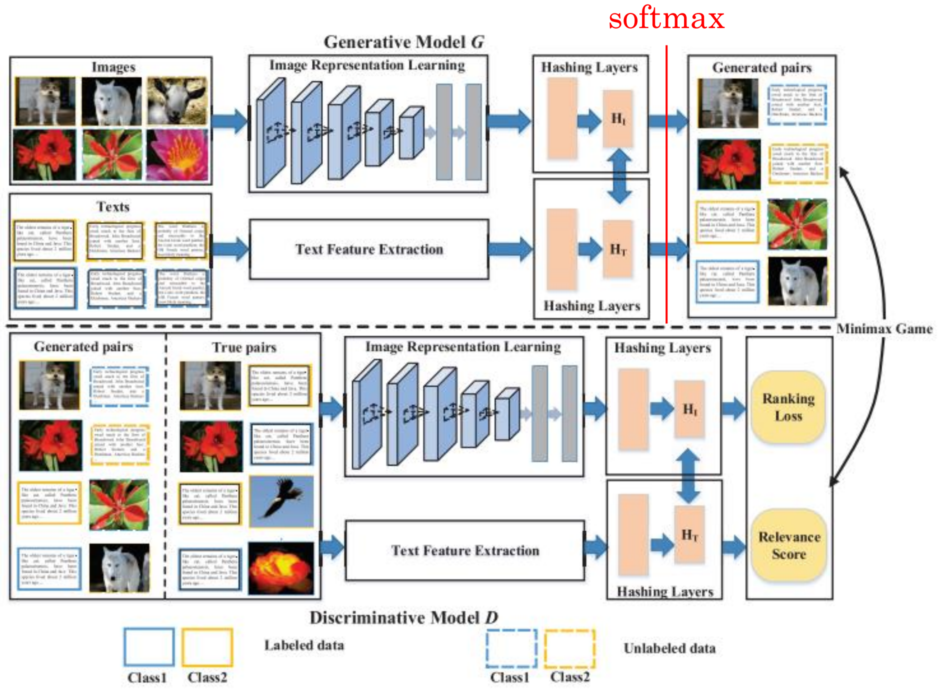
$$\mathcal{V}(G, D) = \min_{\theta} \max_{\phi} \sum_{j=1}^{n} \Big(E_{t \sim p_{\text{true}}\left(t^L | q_i^j, r\right)} \left[\log\left(D\left(t^L | q_i^j \right) \right) \right] + E_{t \sim p_\theta\left(t^U | q_i^j, r\right)} \left[\log\left(1 - D\left(t^U | q_i^j \right) \right) \right] \Big)$$

Objective Function

- Generative model output:

$$p_\theta(i^U|q_t, r) = \frac{\exp\left(-\|h_T(q_t) - h_I(i^U)\|^2\right)}{\sum_{i^U} \exp\left(-\|h_T(q_t) - h_I(i^U)\|^2\right)}$$

  - $h_I(*)$ the hashing functions of image
  - $h_T(*)$ the hashing functions of text



Objective Function

- Discriminative model output:

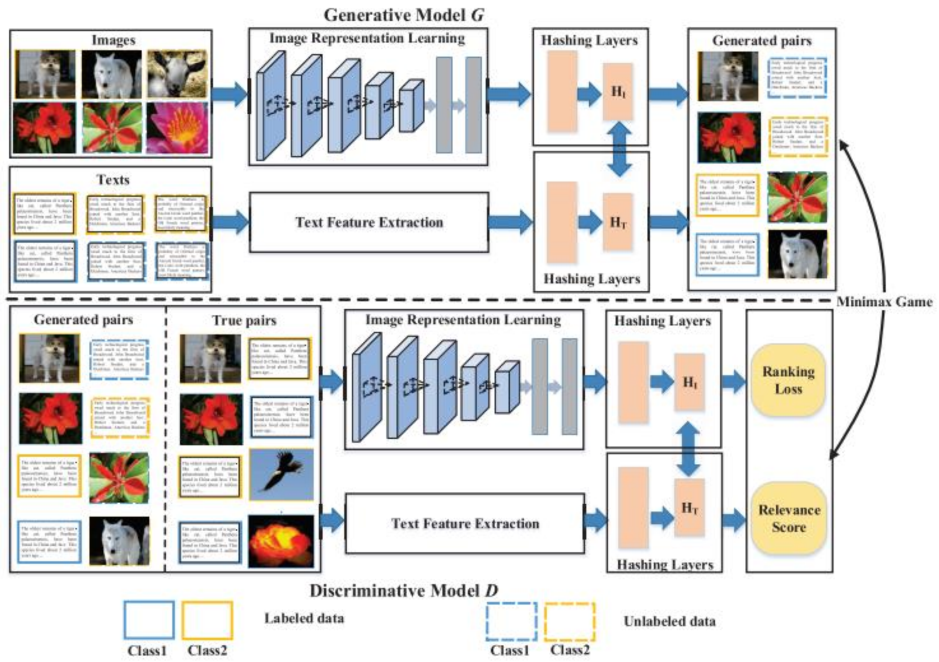
$$D(i^U|q_t) = \text{sigmoid}(f_\phi(i^U, q_t)) = \frac{\exp(f_\phi(i^U, q_t))}{1 + \exp(f_\phi(i^U, q_t))}$$

$$D(i^L|q_t) = \text{sigmoid}(f_\phi(i^L, q_t)) = \frac{\exp(f_\phi(i^L, q_t))}{1 + \exp(f_\phi(i^L, q_t))}.$$

$$f_\phi(i^U, q_t) = \max\Big(0, m_i + \|h_T(q_t) - h_I(i^+)\|^2 - \|h_T(q_t) - h_I(i^U)\|^2\Big)$$

$$f_\phi(i^L, q_t) = \max\Big(0, m_i + \|h_T(q_t) - h_I(i^L)\|^2 - \|h_T(q_t) - h_I(i^-)\|^2\Big)$$

  - i^+, the semantically similar image with q^t, sampled from labeled data
  - i^-, the semantically dissimilar image from q^t, sampled from labeled data
  - i^U selected by generative model
  - m^i, the margin parameter, set to be 1



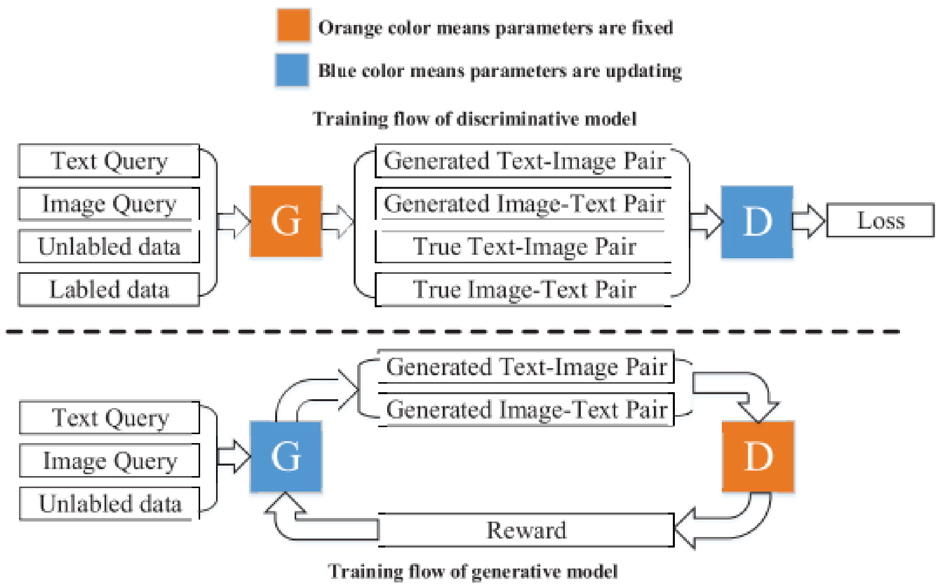
Optimization

- Discriminative model:

$$\phi^* = \arg\max_{\phi} \sum_{j=1}^{n} \Big(E_{i\sim p_{\text{true}}(i^U|q_t^j,r)} \Big[\log\Big(\text{sigmoid}\Big(f_\phi\big(i^L, q_t^j\big)\Big)\Big)\Big] + E_{i\sim p_{\theta^*}(i^U|q_t^j,r)} \Big[\log\Big(1 - \text{sigmoid}\Big(f_\phi\big(i^U, q_t^j\big)\Big)\Big)\Big]\Big)$$

- Generative model:

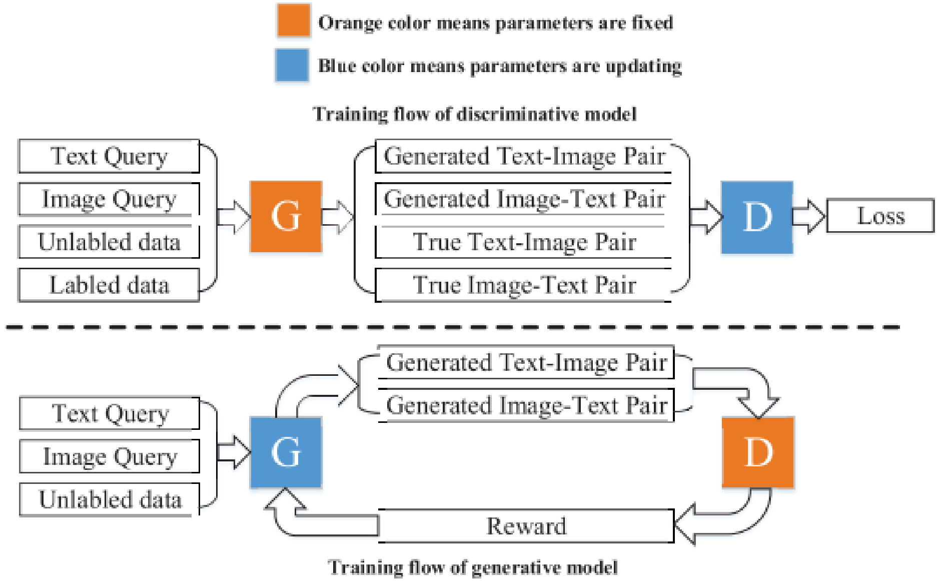
$$\begin{aligned}\theta^* &= \arg\min_{\theta} \sum_{j=1}^{n} \Big(E_{i\sim p_{\text{true}}(i^U|q_t^j,r)} \Big[\log\Big(\text{sigmoid}\Big(f_{\phi^*}\big(i^L, q_t^j\big)\Big)\Big)\Big] + E_{i\sim p_{\theta}(i^U|q_t^j,r)} \Big[\log\Big(1 - \text{sigmoid}\Big(f_{\phi^*}\big(i^U, q_t^j\big)\Big)\Big)\Big]\Big) \\ &= \arg\min_{\theta} \sum_{j=1}^{n} E_{i\sim p_{\theta}(i^U|q_t^j,r)} \left[\log\left(1 - \frac{\exp(f_\phi(i^U, q_t))}{1+\exp(f_\phi(i^U, q_t))}\right)\right] \\ &= \arg\max_{\theta} \sum_{j=1}^{n} E_{i\sim p_{\theta}(i^U|q_t^j,r)} \left[\log(1+\exp(f_\phi(i^U, q_t)))\right] \qquad (10)\end{aligned}$$



Optimization

- Generative model selects data from unlabeled data
  - Because the selective strategy is **discrete**, it is not differentiable
- The author propose a policy gradient-based reinforcement learning method:

$$
\begin{aligned}
&\nabla_\theta E_{i\sim p_\theta\left(i^U|q_t^j,r\right)}\left[\log\left(1+\exp\left(f_\phi\left(i^U,q_t^j\right)\right)\right)\right]\\
&=\sum_{k=1}^{m}\nabla_\theta p_\theta\left(i_k^U|q_t^j,r\right)\log\left(1+\exp\left(f_\phi\left(i_k^U,q_t^j\right)\right)\right)\\
&=\sum_{k=1}^{m}p_\theta\left(i_k^U|q_t^j,r\right)\nabla_\theta\log p_\theta\left(i_k^U|q_t^j,r\right)\log\left(1+\exp\left(f_\phi\left(i_k^U,q_t^j\right)\right)\right)\\
&=E_{i\sim p_\theta\left(i^U|q_t^j,r\right)}\left[\nabla_\theta\log p_\theta\left(i^U|q_t^j,r\right)\log\left(1+\exp\left(f_\phi\left(i^U,q_t^j\right)\right)\right)\right]\\
&\simeq\frac{1}{m}\sum_{k=1}^{m}\nabla_\theta\underbrace{\log p_\theta\left(i_k^U|q_t^j,r\right)}_{\text{Policy}}\underbrace{\log\left(1+\exp\left(f_\phi\left(i_k^U,q_t^j\right)\right)\right)}_{\text{Reward}}\qquad(11)
\end{aligned}
$$



Optimization

Algorithm 1 Training algorithm of proposed SCH-GAN

Input: The generative model $p_\theta(i|q_t, r)$, the discriminative model $f_\phi(i, q_t)$, training data D^L_{db} and D^U_{db}

Randomly initialize the parameters of $p_\theta(i|q_t, r)$ and $f_\phi(i, q_t)$

repeat

  for d-step **do**

    Generate m text-image pairs by $p_{\theta^*}(i^U|q^j_t, r)$ given text query q^j_t

    Sampled m true text-image pairs from D^L_{db} based on labels

    Train discriminative model $f_\phi(i, q_t)$ by equation 9

  end for

  for g-step **do**

    Generate m text-image pairs by $p_\theta(i^U|q^j_t, r)$ given text query q_t

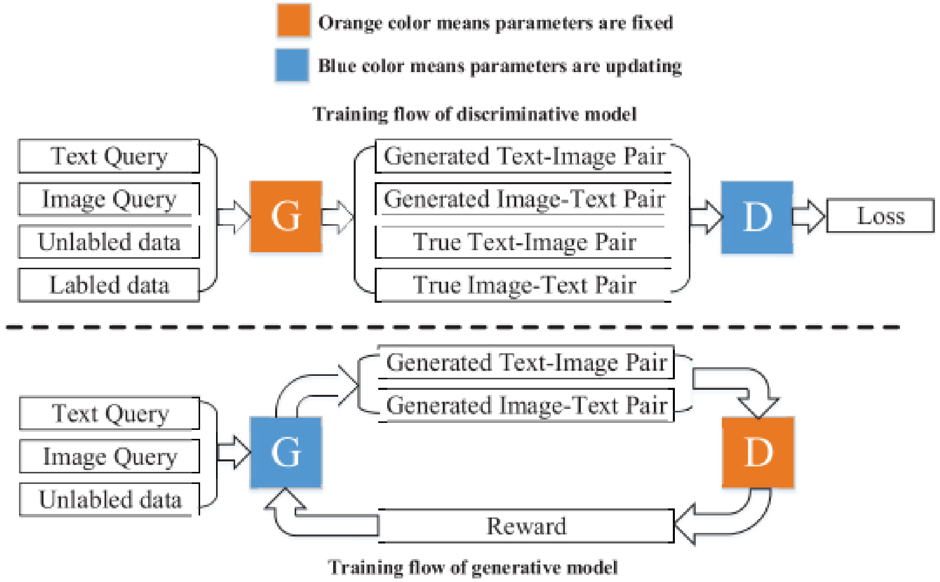
    Calculate reward by $log(1 + \exp(f_{\phi^*}(i^U_k, q^j_t)))$

    Update parameters of generative model $p_\theta(i^U|q^j_t, r)$ by equation 11

  end for

until SCH-GAN converges

Output: Optimized generative model $p_{\theta^*}(i|q_t, r)$ and discriminative model $f_{\phi^*}(i, q_t)$



Evaluation

- Dataset
  - Wikipedia: 2866 image/text pairs, 10 categories
  - NUSWIDE: 269498 image/tag pairs, 81 concepts
  - MIRFlickr: 25000 image/tag pairs, 24 semantic labels
- The data is further split into query set – labeled set – and unlabeled set
  - The label info is removed on the unlabeled set
  - The unlabeled set is much bigger than the labeled and query set

- Evaluation metrics
  - Mean Average Precision (MAP)
    - The mean of average precisions (AP) of all queries
  - Precision Recall curve (PR-curve)
    - The precision at certain level of recall of the retrieved ranking list
  - Precision at top k (topK-precision)
    - The precision with respect to different numbers of retrieved samples
- Tasks
  - Image-to-text: Using image as query to retrieve semantically similar texts from retrieval database
  - Text-to-image: Using text as query to retrieve semantically similar images from retrieval database

Quantitative Result – MAP

TABLE I
THE MAP SCORES OF TWO RETRIEVAL TASKS ON WIKIPEDIA DATASET WITH DIFFERENT LENGTH OF HASH CODES.

Methods	image→text				text→image			
	16	32	64	128	16	32	64	128
CVH [11]	0.193	0.161	0.144	0.134	0.297	0.225	0.187	0.167
PDH [13]	0.483	0.483	0.494	0.497	0.842	0.842	0.838	0.851
CMFH [14]	0.439	0.496	0.473	0.461	0.484	0.548	0.573	0.568
CCQ [46]	0.463	0.471	0.470	0.456	0.744	0.788	0.785	0.741
CMSSH [35]	0.160	0.159	0.157	0.156	0.206	0.208	0.206	0.205
SCM_orth [17]	0.229	0.192	0.171	0.161	0.238	0.171	0.145	0.131
SCM_seq [17]	0.396	0.459	0.462	0.442	0.442	0.557	0.538	0.510
SePH [19]	0.515	0.518	0.533	0.538	0.748	0.781	0.792	0.805
DCMH [50]	0.475	0.508	0.507	0.503	0.819	0.828	0.788	0.720
SCH-GAN (Ours)	0.525	0.530	0.551	0.546	0.860	0.876	0.889	0.888

Quantitative Result – MAP

TABLE II
THE MAP SCORES OF TWO RETRIEVAL TASKS ON NUSWIDE DATASET WITH DIFFERENT LENGTH OF HASH CODES.

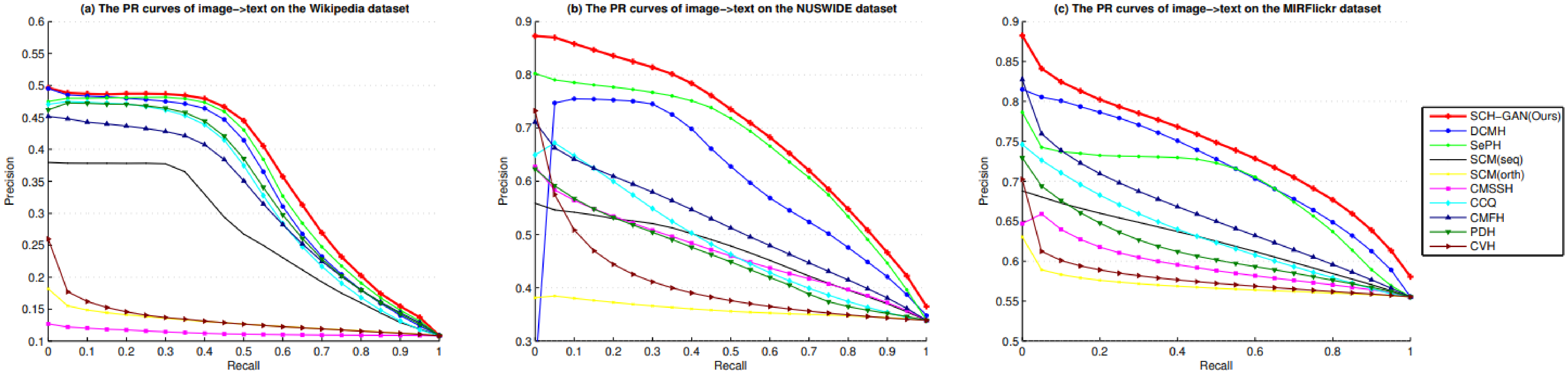
Methods	image→text				text→image			
	16	32	64	128	16	32	64	128
CVH [11]	0.458	0.432	0.410	0.392	0.474	0.445	0.419	0.398
PDH [13]	0.475	0.484	0.480	0.490	0.489	0.512	0.507	0.517
CMFH [14]	0.517	0.550	0.547	0.520	0.439	0.416	0.377	0.349
CCQ [46]	0.504	0.505	0.506	0.505	0.499	0.496	0.492	0.488
CMSSH [35]	0.512	0.470	0.479	0.466	0.519	0.498	0.456	0.488
SCM_orth [17]	0.389	0.376	0.368	0.360	0.388	0.372	0.360	0.353
SCM_seq [17]	0.517	0.514	0.518	0.518	0.518	0.510	0.517	0.518
SePH [19]	0.701	0.712	0.719	0.726	0.642	0.653	0.657	0.662
DCMH [50]	0.631	0.653	0.653	0.671	0.702	0.695	0.694	0.693
SCH-GAN (Ours)	0.713	0.726	0.734	0.748	0.741	0.743	0.771	0.779

Quantitative Result – MAP

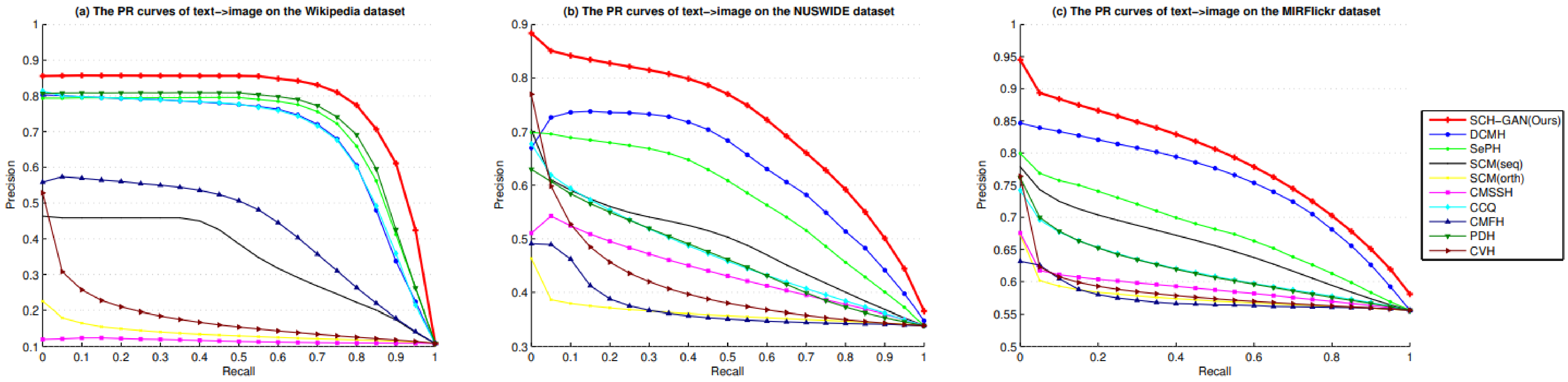
TABLE III
THE MAP SCORES OF TWO RETRIEVAL TASKS ON MIRFLICKR DATASET WITH DIFFERENT LENGTH OF HASH CODES.

Methods	image→text				text→image			
	16	32	64	128	16	32	64	128
CVH [11]	0.602	0.587	0.578	0.572	0.607	0.591	0.581	0.574
PDH [13]	0.623	0.624	0.621	0.626	0.627	0.628	0.628	0.629
CMFH [14]	0.659	0.660	0.663	0.653	0.611	0.606	0.575	0.563
CCQ [46]	0.637	0.639	0.639	0.638	0.628	0.628	0.622	0.618
CMSSH [35]	0.611	0.602	0.599	0.591	0.612	0.604	0.592	0.585
SCM_orth [17]	0.585	0.576	0.570	0.566	0.585	0.584	0.574	0.568
SCM_seq [17]	0.636	0.640	0.641	0.643	0.661	0.664	0.668	0.670
SePH [19]	0.704	0.711	0.716	0.711	0.699	0.705	0.711	0.710
DCMH [50]	0.721	0.729	0.735	0.731	0.764	0.771	0.774	0.760
SCH-GAN (Ours)	0.739	0.747	0.755	0.769	0.775	0.790	0.798	0.799

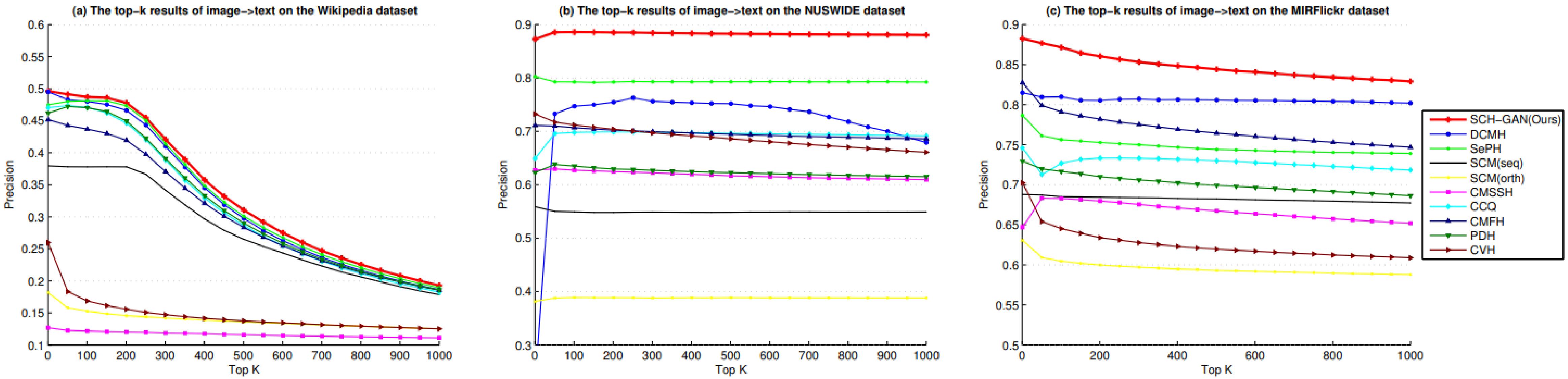
Quantitative Result – PR-Curve



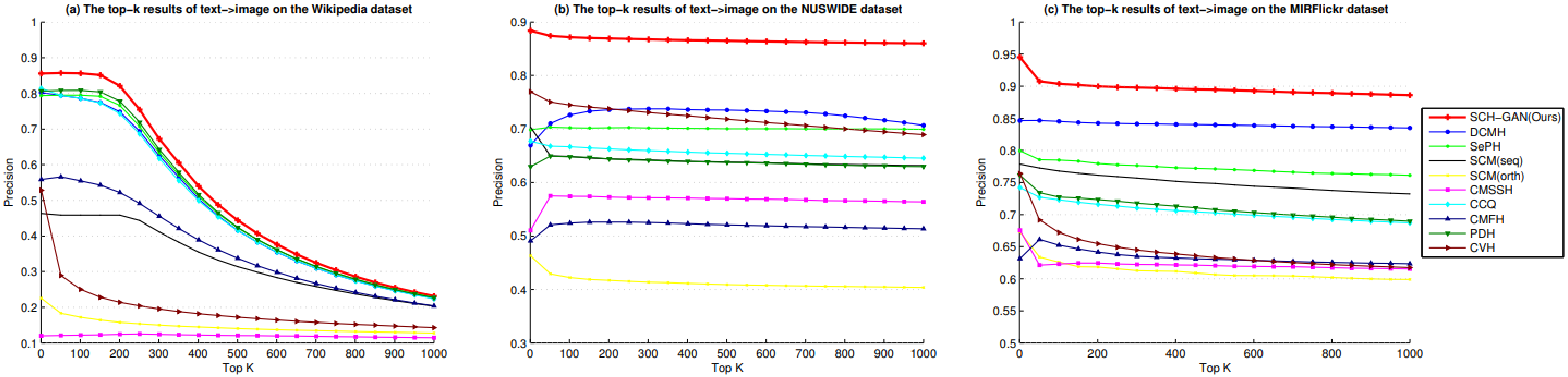
Quantitative Result – PR-Curve



Quantitative Result – TopK-Precision



Quantitative Result – TopK-Precision



Qualitative Result

Task	Query	Method	Top 5 Results				
Image→Text		SCH-GAN	road bear park camping wild vacation usa mountain holiday mountains nature animal alaska clouds america landscape dangerous	flowers brazil naturaleza flores flower color bird nature colors birds animal animals brasilia brasil cores out flying colorful	blue black color green nature leaves yellow butterfly insect wings colorful searchthebest butterflies magicwings naturesfinest	park flowers flower macro bird birds closeup canon bug insect landscape rebel fly flying petals newjersey dragon wildlife seagull	england nature animal closeup landscape countryside cow cattle natural britain farm derbyshire highland calf mooo bovine farmanimal
		DCMH [54]	reflection bird nature water swimming duck flood waterbird reflexions coot oxbow treereflection springrains superaplus aplusphoto oxbowinthespring	holland green bird water netherlands amsterdam television weird canal tv garbage telly pollution rubbish environment fowl discarded waterfowl	fish water azul aquarium goldfish curiosity melpin thelittledoglaughed	water pool swimming canon butterfly jon photoshoot goggles explore swimmer splash 28135 40d	california door nature water bay shadows salt sanjose doorway walkway saltpond marsh railing alviso 1740 evaporation
		CDQ [53]	fish island shark marine paradise underwater scuba diving whaleshark maldives	uk wild portrait england blackandwhite baby lake playing elephant love wet water animal animals swimming swim canon river children	plants green 20d nature wet water canon ball succulent drops globe waterdrop grow spray photophilosophy mireasrealm godaspicks	england white brick bird window water bike ilovenature canal swan ride lancashire reflect galaxy commute lancaster cob mute birdwatcher	pink flowers flores flower reflection water gardens garden botanical pond colorado lily denver botanic flickrelite

Qualitative Result

Task	Query	Method	Top 5 Results
Text→Image	sunset germany evening dusk balloon meschede crescentmoon communicationstower	SCH-GAN	
	sunset germany evening dusk balloon meschede crescentmoon communicationstower	DCMH [54]	
	sunset germany evening dusk balloon meschede crescentmoon communicationstower	CDQ [53]	

Conclusion

- The author proposed a novel GAN for cross-model hashing
  - Generative model tries to select margin examples of another modality from unlabeled data given a query of one modality
  - Discriminative model tries to predict the correlation between query and selected examples of generative model
  - Both models play a minimax game to optimize each other in an adversarial way
- The author also proposed a RL-based algorithm to handle non-differentiable generative model
- Experiments compared with nine SOTA methods on three widely used datasets verify the effectiveness of the proposed approach