Domain-Aware SE Network for Sketch-based Image Retrieval with Multiplicative Euclidean Margin Softmax

Peng Lu
Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
lup21@mails.tsinghua.edu.cn

Gao Huang
Department of Automation, Tsinghua University
Beijing, China
gaohuang@tsinghua.edu.cn

Hangyu Lin
School of Data Science, Fudan University
Shanghai, China
linhy960303@gmail.com

Wenming Yang∗
Shenzhen International Graduate School & Department of Electronic Engineering, Tsinghua University
Shenzhen, China
yang.wenming@sz.tsinghua.edu.cn

Guodong Guo
IDL, Baidu Research National Engineering Lab for Deep Learning Technology and Application
Beijing, China
guoguodong01@baidu.com

Yanwei Fu†
School of Data Science, Fudan University
Shanghai, China
yanweifu@fudan.edu.cn

ABSTRACT
This paper proposes a novel approach for Sketch-Based Image Retrieval (SBIR), for which the key is to bridge the gap between sketches and photos in terms of the data representation. Inspired by channel-wise attention explored in recent years, we present a Domain-Aware Squeeze-and-Excitation (DASE) network, which seamlessly incorporates the prior knowledge of sample sketch or photo into SE module and make the SE module capable of emphasizing appropriate channels according to domain signal. Accordingly, the proposed network can switch its mode to achieve a better domain feature with lower intra-class discrepancy. Moreover, while previous works simply focus on minimizing intra-class distance and maximizing inter-class distance, we introduce a loss function, named Multiplicative Euclidean Margin Softmax (MEMS), which introduce multiplicative Euclidean Margin into feature space and ensure that the maximum intra-class distance is smaller than the minimum inter-class distance. This facilitates learning a highly discriminative feature space and ensures a more accurate image retrieval result. Extensive experiments are conducted on two widely used SBIR benchmark datasets. Our approach achieves better results on both datasets, surpassing the state-of-the-art methods by a large margin.

CCS CONCEPTS
• Computing methodologies → Visual content-based indexing and retrieval.

†indicates corresponding author.
∗Yanwe Fu is with the School of Data Science, MOE Frontiers Center for Brain Science, and Shanghai Key Lab of Intelligent Information Processing, Fudan University.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:
Peng Lu, Gao Huang, Hangyu Lin, Wenming Yang, Guodong Guo, and Yanwei Fu. 2021. Domain-Aware SE Network for Sketch-based Image Retrieval with Multiplicative Euclidean Margin Softmax. In Proceedings of the 29th ACM International Conference on Multimedia (MM ’21), October 20–24, 2021, Virtual Event, China. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3474085.3475499

1 INTRODUCTION

Touch-screen devices, such as smartphones and iPad, enable users to draw free-hand sketches conveniently. These sketches are highly iconic, succinct, and abstract representations, and usually convey richer and more accurate information than texts in various scenarios. Consequently, many novel applications related with sketch have sprung up. One representative example is the Sketch-Based Image Retrieval (SBIR), which has attracted great attention from the computer vision community during the past decades [15, 24, 25, 35, 37, 48, 53]. For the SBIR task, learning good representations for both sketches and photos is vital and remains a challenging problem [31, 32, 37, 40, 48, 49].

Given a query sketch, the objective of SBIR is to find relevant photos that are semantically related, e.g., they come from the same category (as shown in Figure 1). This task appears to be easy for humans but is difficult for a machine. The main challenge comes from the fact that there is a huge gap between the data representation in the two domains: the sketches are represented by highly iconic, abstract, and sparse lines, while the photos are composed of dense color pixels with rich texture information. This domain gap obstructs the SBIR model in exploiting the shared semantics and discriminative representations for both photos and sketches.

Recently, plenty of works have been proposed to address this problem. A popular approach is constructing an excellent intermediate representation, i.e., converting photos to edge maps [24, 37, 48, 53] or translating sketches into the photo domain using generative models [53]. Another widely adopted approach is learning
With this modification, DASE module can emphasize different channels which surpass all existing algorithms by a large margin. The results obtained on several competitive SBIR tasks validate the effectiveness of our model.

**Contributions.** The main contributions of this paper include: (1) A novel network architecture DASE is introduced to incorporate the additional information about the domain attribution, which boosts the retrieval performance with both shallow and deep networks; (2) A novel loss function, the Multiplicative Euclidean Margin Softmax (MEMS), that optimizes a discriminative feature space where the maximum intra-class distance is smaller than the minimum inter-class distance; (3) Theoretical analysis is provided on the properties of the MEMS loss; (4) New state-of-the-art results have been obtained on several competitive SBIR tasks.

### 2 RELATED WORK

**SBIR.** Sketch-Based Image Retrieval (SBIR) aims to retrieve similar semantic meaning images as the query sketch. A typical solution is to learn a shared embedding space for both sketches and images. Such a common space facilitates the ranking of similarity of sketches and images. Previous methods transfer photos into sketch-tokens and then extract hand-craft sketch features [5, 12, 27, 31] or deep features [24, 37, 40, 48, 49] to represent the sketches. Recent deep learning based architectures [18, 21, 29, 53] enable cross-domain learning in an end-to-end manner. To accelerate the retrieval in a large-scale dataset, hashing based models [24, 35, 47, 53] have also been studied.

**Attention in CNNs.** Attention mechanism enables networks to assign different significance on different parts of the input. The important feature expressions are amplified while the less useful ones are suppressed. To better apprehend digital images, many convolutional networks with spatial attention [4, 13, 19, 44] and channel-wise attention [2, 11] are proposed to emphasize the informative region or channels. SENet [11] proposes a squeeze-and-excitation structure. Global context is aggregated by global average pooling for each channel. The aggregated information is then processed and used to reallocate attention over channels by feature fusion. To tackle the SBIR problem, spatial attention [37, 50, 53] has been utilized. However, channel-wise attention has never been explored in cross-modal embedding for sketches and photos so far.

**Loss functions.** Many metric learning based methods [10, 28, 42, 45] proposed learning deep features by the loss functions with Euclidean distances. To make the learned features more discriminative, other variants of loss functions have been investigated recently, such as the contrastive loss [3, 9], triplet loss [34], and softmax-based losses [6, 25, 26, 41]. Particularly, the contrastive and triplet losses aim at increasing the Euclidean margin in an additive manner, which might be neglected as there is no upper bound for Euclidean distance. A-softmax [25] optimize the inter-class angular distance to be several times the intra-class angular distance to guarantee that the least inter-class distance is more than the largest intra-class distance. Furthermore, prototypical loss [36] is also a variant of softmax which incorporates the Euclidean distance.
which the DASE module is inserted (see in Figure 2). In contrast to
we have the set \( D \) as the setting defined in [24], the same photo set
To facilitate learning feature representations for sketches and pho-
tive Euclidean Margin Softmax (MEMS) loss, which can efficiently
Further, to bridge the data representation gap between the sketch
to analyze the photos and sketches jointly. Such a SiameseNet-
like network is inspired by the fact that SiameseNet is efficient
in learning the embedding space across different domains (e.g.,
image-text embedding [43], or person re-identification [38]). The
semi-heterogeneous network is proposed in [24], including a shared
Siamese sub-network and separate sub-networks for each domain.
In contrast to [24], we further propose Domain-Aware Squeeze-
and-Excitation (DASE) module, which is shown in Figure 3, to learn
an effective feature representation. Particularly, our DASE module
is an extension of SE module [11] which provides an explicit mecha-
nism to re-weight the importance of channels after each block in
the network. As shown in Figure 3, our DASE module utilizes an
encoder-decoder structure followed by a sigmoid activation. Within
the intermediate space, a binary code is added to indicate whether
the input image is a sketch or a photo. The outputs of sigmoid
activation are the feature attention vector over channels. This con-
tional structure can thus help capture different characteristics of
input images conditioned on which domain they come from.
With the DASE module, our feature extractor, which is a Siamese
Network with the domain-conditioned structure, is learned over
two domains to explore shared semantics in a common feature
space. The binary code serves as a controlling gate and endows the
flexibility of network in learning from each individual domain.

3 METHOD
In this section, we present the proposed DASE network and the
MEMS loss for SBIR task in detail. The SBIR task is formulated as
\[ \mathcal{P} = \{(p_i, y_i) \mid y_i \in Y\} \]
and the sketch set \( S = \{(s_j, y_j) \mid y_j \in Y\} \) respectively. Given a
query sketch, the objective of SBIR is to find all of its relevant
photos that semantically belong to the same category as the sketch.
As the setting defined in [24], the same photo set \( \mathcal{P} \) is used for both
training and test, as the retrieval galleries. Sketch set \( S \) is split into
train and test sets, with the same label set \( Y \).

The feature representation networks of sketches and photos are
denoted as \( F_p \) and \( F_s \) respectively. The two networks project
sketches and photo to a shared space, i.e., \( F_p(p_i), F_s(s_j) \in X \); thus
we have the set \( \mathcal{D} = \{(x, y_x) \mid y_x \in Y\} \), where the feature \( x \in X \)
can come either from photo domain or sketch domain and the label
\( y_x \) is identical with the source sample of feature \( x \). We further
introduce the denotation, \( \mathcal{D}_y = \{ x \mid y_x = y \} \) for convenience.

3.1 Overview
To facilitate learning feature representations for sketches and pho-
tos, the proposed Domain-Aware SE (DASE) network is shared by
both domains. This network is composed of several ResBlocks in
which the DASE module is inserted (see in Figure 2). In contrast to
vanilla SE module [11], DASE module receives code about the do-
main of input image and emphasizes different channels individually.
Further, to bridge the data representation gap between the sketch
and photo domains, we present a novel loss function – Multiplica-
tive Euclidean Margin Softmax (MEMS) loss, which can efficiently
learn a unified embedding space. The MEMS loss optimizes the
larger inter-class and smaller intra-class distance over the photo
and sketch set. In other words, the instances of sketches and photos
in the same/different class should be close/far from each other.

3.2 Domain-Aware SE Network
Learning common feature representations for sketch and photo
domains is a non-trivial task. Obviously, sketches are iconic and
abstract with various deformation levels; while photos are realistic
images with rich color, texture, and shape information.
Rather than using independent sub-networks to process the
photos and sketches separately, our model learns a single network

Figure 2: Illustration of our framework. We mix sketches and photos as input of the DASE network, to which the proposed
DASE module is added. Domain knowledge is encoded in DASE modules in each ResBlock. After training, the output features or
binarized code can be used to conduct SBIR, where Euclidean distance and Hamming distance are used as metric, respectively.

Figure 3: The structure of DASE module. The activation function for the intermediate FC layer is Sigmoid.
3.3 Multiplicative Euclidean Margin Softmax Loss

The loss function is vital in successfully optimizing the networks over the recognition tasks, especially in our cross-domain scenario. To embed the photo and sketch into a shared space, we propose the Multiplicative Euclidean Margin Softmax loss, which exploits the strategy of maximizing the intra-class distance and minimizing the inter-class distance.

Given the category \( y \in \mathcal{Y} \), the maximum intra-class distance can be defined by \( \max_{y \in \mathcal{Y}} \sum \frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{-m\|x_i - c_{y_i}\|_2^2}}{e^{-m\|x_i - c_{y_i}\|_2^2} + \sum_{y \neq y_i} e^{-m\|x_i - c_{y}\|_2^2}} \) where \( m \) is a margin constant.

Moreover, the minimum inter-class distance is formulated by enforcing that the maximum intra-class distance to be smaller than minimum inter-class distance.

\[
\forall y \in \mathcal{Y}, \quad \max_{x \in \mathcal{D}_{\mathcal{Y}}} d(x, x') \leq \min_{x \in \mathcal{D}_{\mathcal{Y}}} d(x, x') \quad (1)
\]

Rather than directly optimizing Eq (1) over all instances, we compute the prototypes \( \{c_{y}\}_{y \in \mathcal{Y}} \subset \mathcal{X} \) of each class, and characterize the distribution of instances in feature space. If MEMS loss is well optimized, instances will be closer to their corresponding prototype than other prototypes in feature space. This is as,

\[
\forall (x, y_x) \in \mathcal{D}, \forall y \in \mathcal{Y} \land y \neq y_x, \ m \cdot d(x, c_y) \leq d(x, c_y) \quad (2)
\]

where \( m \geq 1 \) is referred to as the margin constant.

For convenience, we denote \( \mathcal{R}_{y,y'} \) as a region in the feature space,

\[
x \in \mathcal{R}_{y,y'} \iff m \cdot d(x, c_y) \leq d(x, c_{y'})
\]

Also, we denote \( \mathcal{R}_y \) as a region such that

\[
x \in \mathcal{R}_y \iff \forall y' \in \mathcal{Y} \land y' \neq y, m \cdot d(x, c_{y'}) \leq d(x, c_{y'})
\]

which takes account different data distributions of each class. It is easy to prove that \( \mathcal{R}_y = \bigcap_{y \neq y'} \mathcal{R}_{y,y'} \). Note that if Eq (2) holds,

\[
\forall (x, y_x) \in \mathcal{D}, \ x \in \mathcal{R}_{y_x}, \quad \text{and thus we can derive a sufficient condition for Eq (1)},
\]

\[
\forall y \in \mathcal{Y}, \max_{x \in \mathcal{R}_{y}} d(x, x') \leq \min_{x \in \mathcal{R}_{y}} d(x, x') \quad (3)
\]

MEMS loss. To incorporate this understanding and optimize Eq (2) by softmax loss, this gives us a novel loss function—Multiplicative Euclidean Margin Softmax (MEMS) loss:

\[
L_{ems} = \frac{1}{N} \sum_{i=1}^{N} - \log \frac{e^{-m\|x_i - c_{y_i}\|_2^2}}{e^{-m\|x_i - c_{y_i}\|_2^2} + \sum_{y \neq y_i} e^{-m\|x_i - c_{y}\|_2^2}} \quad (4)
\]

where \( x_i \) indicates the feature extracted by the last layer of feature representation network. \( c_j \) is the center of \( j \)-th category. To take the center \( c_j \) as the parameters, and update \( c_j \) dynamically, rather than directly using the average feature center. In Eq (4), we employ the negative squared Euclidean distance to measure the confidence of \( x_i \) being \( \{x_i - c_j\}_j \). Particularly, in binary classification, \( x_i \) is labelled as class 1 if \( m \|x - c_1\|_2^2 < \|x - c_2\|_2^2 \), and otherwise, as class 2.

3.4 Theoretical analysis and lower bound of \( m \)

The property of MEMS is largely determined by the value of \( m \). Intuitively, a larger \( m \) makes decision boundaries closer to corresponding prototypes and the distribution of features more compact. This produces a more discriminative metric space. However, overly large \( m \) tends to make the training process unstable due to the inherent variances among samples in each category.

Therefore, it is necessary to find the minimum \( m \) to ensure that, for every sample, and in metric space, the maximum intra-class distance is smaller than minimum inter-class distance. We give some theoretical analysis about the Eq (2) and Eq (3). Most importantly, if we adopt Euclidean distance as \( d(\cdot, \cdot) \), \( m \geq 2 + \sqrt{3} \) is the sufficient and necessary condition for Eq (3) given Eq (2). We show a brief proof below and the details can be found in Appendix.

\[\text{Sufficient.} \quad \text{If } d(x, x') = \|x - x'\|_2 \text{, } \mathcal{R}_{y,y'} \text{ is a } n\text{-ball in } n\text{-dimensional space with the center } (c_y + (c_y - c_{y'}))/(m^2 - 1) \text{ and the radius } \|m/(m^2 - 1)\|_2, \text{ Thus in binary categorization, the minimum value of } m \text{ is } 2 + \sqrt{3}. \text{ With the growth of the number of categories, the minimum value is reduced monotonously. So it is sufficient to have } m \geq 2 + \sqrt{3}, \text{ guaranteeing perfect discrimination.} \]

\[\text{Necessary.} \quad \text{We illustrate the necessity in multi-class cases, since, if two prototypes are far from the others, their relationship will resemble the one in binary case. So } m \geq 2 + \sqrt{3} \text{ is necessary regardless of the number of categories.} \]

3.5 Differences with other losses

We further discuss the differences between our MEMS loss and other losses.

**Prototypical loss** [36] It is defined as

\[
L_{\text{proto}} = \frac{1}{N} \sum_{i=1}^{N} - \log \frac{e^{-m\|x_i - c_{y_i}\|_2^2}}{e^{-m\|x_i - c_{y_i}\|_2^2} + \sum_{y \neq y_i} e^{-m\|x_i - c_{y}\|_2^2}} \quad (5)
\]

The prototypical loss is used in one-shot classification where only a few training instances are available for each class. Thus \( c_y \) is directly computed as the averaged mean of training instances; in contrast, our MEMS loss is a generalized softmax loss, and optimizes \( c_j \) via back-propagation. Furthermore, the prototypical loss is a special case of MEMS loss when the margin is 1. It is sufficient for classification task and insufficient for training a discriminative feature space for retrieval task. Therefore a margin constant \( m \) is introduced to make a balance between enlarging the inter-class distance and shrinking the intra-class distance.

**Angular Margin loss.** We further discuss the difference between MEMS loss and angular margin loss. Similar to Euclidean distance, angular distance based loss functions, such as A-Softmax [25] and LMCL [41], are also employed in learning a shared space in many tasks, e.g., face recognition. These angular margin losses aim at learning a discriminative distribution on a hypersphere. As shown in Table 1, they define different similarity functions for the instances of different classes. Note that \( \psi(\theta) \) is an artificial piece-wise function that serves as the extension of \( \cos (m \cdot \theta) \), to overcome its non-monotonicity. Nevertheless, it is non-trivial to define the \( \psi(\theta) \) in A-softmax function as stated in [41]. The scalar \( s \) in LMCL is used to expand the range of similarity function; otherwise, the output of softmax function would be closed to the uniform distribution over all categories.
Table 1: The similarity functions of A-Softmax, LMCL and MEMS loss. \( \theta_y \) represents angular distance between the feature vector and the prototype of category \( y \).

| Similarity          | intra-class | inter-class \( (i \neq y) \) |
|---------------------|-------------|-------------------------------|
| A-Softmax           | \( \|x_i\|_2 \cdot \psi(\theta_{i0}) \) | \( \|x_i\|_2 \cdot \cos(\theta_i) \) |
| LMCL                | \( s \cdot (\cos(\theta_{y_{ij}}) - m) \) | \( s \cdot \cos(\theta_i) \) |
| MEMS (ours)         | \(-m^2 \|x_i - c_y\|_2^2\) | \(-\|x_i - c\|_2^2\) |

4 EXPERIMENTS

4.1 Datasets and Settings

Datasets. Our model is evaluated on two large-scale sketch-photo datasets: TU-Berlin [8] Extension and Sketchy [33] Extension. The former includes 20,000 sketches uniformly distributed among 250 categories. Additionally, 204,489 natural images provided in [52] are utilized as the photo gallery. The Sketchy database consists of 75,471 hand-drawn sketches and 12,500 corresponding photos from 125 categories. It was extended by another 60,502 photos for SBIR task in [24]. Following the settings in [24, 53], 10/50 sketches from each category are picked as the query set for TU-Berlin/Sketchy datasets. Our model is evaluated on two large-scale sketch-photo datasets: TU-Berlin [8] Extension and Sketchy [33] Extension. Datasets: TU-Berlin [8] Extension and Sketchy [33] Extension. The former includes 20,000 sketches uniformly distributed among 250 categories. Additionally, 204,489 natural images provided in [52] are utilized as the photo gallery. The Sketchy database consists of 75,471 hand-drawn sketches and 12,500 corresponding photos from 125 categories. It was extended by another 60,502 photos for SBIR task in [24]. Following the settings in [24, 53], 10/50 sketches from each category are picked as the query set for TU-Berlin/Sketchy dataset, and the rest are used for training. All gallery photos are used in both training and testing phases.

Implementation. Our method is implemented using Pytorch with single 1080Ti GPU. We use Adam optimizer [16] with parameters \( \beta_1 = 0.9, \beta_2 = 0.999, \lambda = 0.0001 \). The learning rate is set to 0.0001 and linearly decays to 0 during the second half of training. The model converges after training for 200k iterations. We use the backbone network of ResNeXt-101, which is composed of 33 ResNeXt Blocks. The proposed DASE module is added in each block. We use \( m = 4 \) in the proposed MEMS loss. Our code is available at: https://github.com/Ben-Louis/SBIR-DASE-MEMS

SBIR hashing. SBIR can be conducted using either real-value or binary-value vectors as features. The latter is named SBIR hashing, which greatly accelerates the speed of SBIR tasks but requires auxiliary design or process. The proposed method can generate both styles of features. Particularly, for SBIR hashing, we utilize a simple hashing scheme to encode the features \( x_i \) generated by our network from one sketch \( s \) or a photo \( p_i \). Our hashing scheme uses a spectral normalized perception \( F_{SN} \) and a sign function to projects \( x_i \) into a low-dimensional binary space. This is a post-processing step after training CNNs. Specifically, the \( F_{SN} \) is optimized by

\[
L_{hash} = \frac{1}{K(K-1)} \sum_j \sum_{k \neq j} \left[ \frac{F_{SN}(e_j)}{|F_{SN}(e_j)|} - \frac{F_{SN}(e_k)}{|F_{SN}(e_k)|} \right]^2,
\]

where \( K \) is the number of categories and \( |\cdot| \) represents element-wise absolute value. Note that \( F_{SN}(x) = W_{SN} \cdot x + b \), where \( W_{SN} \) is a spectral normalized matrix and \( b \) represents bias. This loss forces the prototype \( e_j \) to have a large Hamming distance to each other in the low-dimensional space. As \( F_{SN} \) is spectral normalized, the Lipschitz constant of the mapping function from feature space to low-dimensional space (before binarization) is 1. Thus the property of Euclidean distance among instances, large inter-class distance and small intra-class distance, is kept in target low-dimensional space. After sign function, instances in the same class will have closed binary representation. The visualization of this process is shown in Figure 4.

Figure 4: Visualization of hashing process.
### 4.2 Results on Supervised SBIR

**Competitors.** We compare several competitors in Table 2: (1) hand-craft feature based models: LSK [32], SEHLO [31], HOG [5] and GF-HOG [12]; (2) cross-view feature embedding methods: CCA [39], PLSR [23], XQDA [20] and CVFL [46]; (3) deep learning based models: 3D Shape [40], Sketch-a-Net (SaN) [49], GN Triplet [33], Siamese CNN [30], Siamese-AlexNet, Triplet-AlexNet [24]. (4) Prototypical loss [36], Triplet loss [34], Softmax, A-Softmax [25] and LMCL [41] implemented with the same backbone network as our method. We use hyperparameters of A-Softmax [25] and LMCL [41] proposed in their papers. The Mean Average Precision (MAP) is reported.

**Results.** The results are summarized in Table 2. Obviously, our model outperforms all the competitors by a very large margin. It achieves a MAP improvement of 0.393/0.385 over the state-of-the-art real-valued based method – Triplet-AlexNet. This demonstrates the efficacy of our model. Note that the improved performance is due to the novel structure, and the MEMS loss function used here. We give further analysis in the ablation study in Section 4.4.

The effect of each component in our framework can be found in Table 2. We can conclude that the proposed MEMS loss and DASE module both improve the performance on SBIR task significantly.

### 4.3 Results on Supervised SBIR Hashing

**Competitors.** (1) Our hashing model is compared against 8 cross-modal hashing methods: Collective Matrix Factorization Hashing (CMFH) [7], Cross-Modal Semi-Supervised Hashing (CMSSH) [1], Cross-View Hashing(CVH) [17], Semantic Correlation Maximization (SCMSeq and SCM-Orth) [51], Semantics-Preserving Hashing(SePH) [22], Deep CrossModality Hashing (DCMH) [14], Deep Sketch Hash (DSH) [24] and Generative Domain-Migration Hashing (GDH) [53]. (2) Other softmax based loss functions implemented with the same backbone as our method: Prototypical loss [36], Triplet loss [34], Softmax, A-Softmax [25] and LMCL [41]. We use hyperparameters of A-Softmax [25] and LMCL [41] proposed in their papers. We still report the MAP.

**Training cost.** Our hashing scheme is taken as a post-processing step, in order to make our framework comparable to previous hashing based SBIR models. With our computed features, the hashing scheme is trained for 10000 steps; and the whole process can be finished in 1 minute on our computer.

**Results.** We summarize our results in Table 3. Our method achieves the best performance among all hashing-based methods and cross-modal learning methods. Critically, our model improves MAP with a scale over 0.13 in all conditions compared with GDH [53] which is the state-of-the-art method on this task. This further demonstrates the effectiveness of our framework in the SBIR hashing task. Some query examples with top-10 retrieval results are shown in Figure 5.

We also highlight that, although angular margin losses, particularly LMCL, can achieve comparable performance with MEMS in SBIR tasks, they suffer from a significant degradation in SBIR hashing. This degradation might be due to the property of angular distance, as it is more difficult to maintain in the dimension reduction mapping than Euclidean distance.
Comparison using same backbone. To emphasize the efficiency and efficacy of our model, we compare the MAP with state-of-the-art methods DSH [24] and GDH [53] using the same backbone networks, which is Alexnet and Resnet-18 respectively. The result is displayed in Table 4. The proposed method achieves higher MAP than other methods with much lower computation consumption.

Besides, our model can be trained efficiently. We only need to train a single network in an end-to-end manner without any domain translation process. On the contrary, GDH [53] utilized the cycle-consistent GANs to transfer sketches into photos. DSH [24] requires the pre-computed edge maps to bridge the gap between sketches and photos, and semantic representation (wordvec) is used as prior of inter-relationship among categories.

4.4 Ablation Study

To further investigate the propose approach, we conduct a series of ablation studies on both datasets.

Network modules. We compare two types of CNNs as well as their variants with SE modules [11] and the newly DASE modules. Intrinsically, the SE and DASE module can enhance the ability to learn different attention over feature channels, and thus enable a dynamic and implicit feature selection mechanism to our networks. The MAP results are shown in Figure 6. Both networks are trained and tested in the same setting; and the same MEMS loss is used for all the networks. We can find that SE module enhances the ability of CNNs to process inputs from multi-domains. Moreover, our DASE module is better than SE module on SBIR task, since the auxiliary binary code is introduced to make SE better learn to select important sketch/photo feature channels. DASE module can improve the performance of both deep and shallow networks. Particularly, on Resnet-18/ResNeXt-101, the auxiliary DASE modules have brought the MAP increase with 0.025/0.043 on TU-Berlin Extension and 0.011/0.009 on Sketchy Extension, respectively.

We also visualize the input feature maps of DASE modules in Figure 7. It is noticeable that some convolution kernels that can effectively extract content-related features from sketches might fail to extract such features from photos, and vice versa. Specifically, features in B2C239 generally respond to dark areas, and thus miss the key object in photos but capture crucial corners in sketches. On the contrary, in B6C17, sketch features show severe corner and edge effects which distort the semantic, while photo features respond to correct area. With prior knowledge, the DASE module can learn to amplify content-related features and suppress style-related features. The weight in the second fully-connected layer in DASE module associated with the binary code about input domain and B2C239 is -0.32, smaller than most channels, indicating that this channel is relatively amplified/suppressed when the input is a sketch/photo, respectively. As for B6C17, this weight is 0.30, and is larger than most channels, meaning that the photo features are amplified and the sketch features are suppressed.

The effects of different margin $m$. Performance of our model grows with the increase of value $m$, because it forces the network to learn discriminative representations. But when $m$ is too large, e.g., $m > 4$, the performance stops rising and even begins falling, which is shown in Figure 8.

As revealed in Figure 9: (1) when $m = 1$, the intra-class distances of part of instances are greater than the minimum inter-class distance, which leads to bad retrieval performance; (2) when $m = 4$ and $m = 16$, the minimum inter-class distances are greater than intra-class distances generally, which explains the better performance in these cases. Nevertheless, the standard deviation within some categories does not decrease with larger $m$ due to intrinsic variance, explaining why the performance stops rising or begins falling when $m$ is overly large.
Analysis of the mapping in hashing. We train a spectral normalized fully-connected layer $F_{SN}$ for hashing. But the mapping can be modeled by other structures. We try fully-connected networks with different depth and settings (w./w.o. spectral normalization) for the hashing task. Note that if we adopt multi-layer networks, the dimension of intermediate layers is set to 256, and the activation function is Rectified Linear Unit, as the Lipschitz constant of ReLU is smaller than 1. The results of MAP on both datasets are shown in Figure 10. We can found that multi-layer networks have worse performance than one-layer linear mapping. Moreover, the spectral normalized networks work better generally, which shows their advantage in maintaining discrimination property.

5 CONCLUSION

In this paper, we have introduced two innovations, including novel network architecture and a loss function. The Domain-Aware Squeeze-and-Excitation (DASE) network allows us to incorporate the domain information of each sample explicitly, while the proposed Multiplicative Euclidean Margin Softmax (MEMS) Loss enables us to learn highly discriminative features which facilitate highly accurate SBIR. Both the architecture and loss are intuitive and simple to implement. On two popular benchmark SBIR datasets, the proposed model has achieved new state-of-the-art results.

ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of Guangdong Province(No.2020A1515010711) and the Special Foundation for the Development of Strategic Emerging Industries of Shenzhen(Nos. JCYJ20200109143010272 and JCYJ20200109143035495). This work was supported in part by Science and Technology Commission of Shanghai Municipality Projects (1951120700, 2021SHZDZX0103).
REFERENCES

[1] Michael M Bronstein, Alexander M Bronstein, Fabrice Michel, and Nikos Paragios. 2010. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 3594–3591.

[2] Yue Cao, Jiari Xu, Stephen Lin, Fanyang Wei, and Han Hu. 2019. Gnet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the IEEE International Conference on Computer Vision Workshops. 0–9.

[3] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In CVPR.

[4] Jiefei Song, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. 2017. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In ECCV.

[5] Jingyi Zhang, Fumin Shen, Li Liu, Fan Zhu, Mengyang Yu, Ling Shao, Heng Tao Shen, and Luc Van Gool. 2018. Generative Domain-Migration Hashing for Sketch-to-Image Retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3598–3607.

[6] Jie Hu, Li Shen, and Gang Sun. 2017. Squeeze-and-excitation networks. In September.

[7] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. 2014. Discriminative deep metric learning for face verification in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1875–1882.

[8] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-Margin Softmax Loss for Convolutional Neural Networks. In ICLR.

[9] Jie Hu, Li Shen, and Gang Sun. 2017. Squeeze-and-excitation networks. In September.

[10] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. 2014. Discriminative deep metric learning for face verification in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1875–1882.

[11] Rui Hu, Mark Barnard, and John Collomosse. 2010. Gradient field descriptor for sketch based retrieval and localization. In Image Processing (ICIP), 2010 17th IEEE International Conference on. IEEE, 1025–1028.

[12] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In Advances in Neural Information Processing Systems. 4077–4087.

[13] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-Margin Softmax Loss for Convolutional Neural Networks. In ICLR.

[14] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. 2017. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In ECCV.

[15] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. 2016. Gated Siamese Convolutional Neural Network Architecture for Human Re-Identification. In ECCV.

[16] Qing-Yuan Jiang and Wu-Jun Li. 2016. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In ICCV.

[17] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In null. IEEE, 1375–1424.

[18] Jie Hu, Li Shen, and Gang Sun. 2017. Squeeze-and-excitation networks. In September.

[19] Junlin Hu, Jiwen Lu, and Yup-Feng Tan. 2014. Discriminative deep metric learning for face verification in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1875–1882.

[20] Jie Hu, Li Shen, and Gang Sun. 2017. Squeeze-and-excitation networks. In September.

[21] Nikolay Jitkoff, Shruti Chokshi, and Ankit Panda. 2017. A Randomized Feature Learning Solution for Zero-shot Sketch-based Image Retrieval. In RSS.

[22] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. 2016. Large-Margin Softmax Loss for Convolutional Neural Networks. In ICLR.

[23] Xiaohua Xia, Yixun Peng, and Jianshuo Xie. 2014. Cross-View Feature Learning for Scalable Image Similarity Analysis. In AAAI. 201–207.

[24] Yang Liu, Yi Jin, and Svetlana Lazebnik. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In CVPR.

[25] Jingyi Zhang, Fumin Shen, Li Liu, Fan Zhu, Mengyang Yu, Ling Shao, Heng Tao Shen, and Luc Van Gool. 2018. Generative Domain-Migration Hashing for Sketch-to-Image Retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 799–807.

[26] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 7794–7803.

[27] Xiaohua Xia, Yixun Peng, and Jianshuo Xie. 2014. Cross-View Feature Learning for Scalable Image Similarity Analysis. In AAAI. 201–207.

[28] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 7794–7803.

[29] Jingyi Zhang, Fumin Shen, Li Liu, Fan Zhu, Mengyang Yu, Ling Shao, Heng Tao Shen, and Luc Van Gool. 2018. Generative Domain-Migration Hashing for Sketch-to-Image Retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 297–314.
6 APPENDIX: THEORETICAL ANALYSIS OF MEMS LOSS

In this section, we will (1) give a formal definition of maximum intra-class distance and minimum inter-class distance; (2) show that for margin \( m \) in MEMS loss, \( m \geq 2 + \sqrt{3} \) is sufficient and necessary to ensure that the maximum intra-class distance is smaller than minimum inter-class distance, regardless of the number of categories.

6.1 Definition

Since we treat both sketch and photo as instances, we define the merged dataset as:

\[
D = \{(x_i, y_{x_i}) | y_{x_i} \in \mathcal{Y}_{1+n_2} \text{ where } \begin{cases} x_i = \mathcal{F}_p (p_i), y_{x_i} = y_{p_i} & \text{if } i \leq n_1 \\ x_i = \mathcal{F}_s (s_{i-n_1}), y_{x_i} = y_{s_{i-n_1}} & \text{if } i > n_1 \end{cases} \}
\]

where \( \mathcal{F}_p \) and \( \mathcal{F}_s \) are mappings that map photos/sketches into a feature space; \( n_1, n_2 \) represent the number of photos and sketches; \( p, s \) and \( y \) represent photo, sketch and category respectively. They are detailedly illustrated in Sec. 3. For convenience, we also define \( D_y = \{ x | (x, y_x) \in D \land y_x = y \} \).

Maximum Intra-class Distance and Minimum Inter-class Distance. For category \( y \in \mathcal{Y} \), the maximum intra-class distance can be defined by

\[
\max_{x, x' \in D_y} d(x, x')
\]

and the minimum inter-class distance:

\[
\min_{x \in D_y, x' \in \bigcup_{y' \neq y} D_{y'}} d(x, x')
\]

Here we give a formulation of our objective, which is the maximum intra-class distance being smaller than minimum inter-class distance, by proposition \( P_1 \):

\[
\forall y \in \mathcal{Y}, \max_{x, x' \in D_y} d(x, x') \leq \min_{x \in D_y, x' \in \bigcup_{y' \neq y} D_{y'}} d(x, x')
\]

Solve Problem with MEMS. Instead of optimizing the distance among instances directly as indicated by eq. 9, the proposed MEMS loss uses prototypes \( \{ c_{y} \}_{y \in \mathcal{Y}} \subset \mathcal{X} \) to characterize the distribution of instances in feature space. If this MEMS loss is well optimized, instances will be closer to their corresponding prototype than other prototypes in feature space. This relationship can be described as

\[
\forall (x, y_x) \in D, \forall y \in \mathcal{Y} \land y \neq y_x, m \cdot d(x, c_{y_x}) \leq d(x, c_y)
\]

For convenience, we denote \( R_{y, y'} \) as a region where

\[
x \in R_{y, y'} \iff m \cdot d(x, c_{y}) \leq d(x, c_{y'})
\]

Also, we denote \( R_y \) as a region where

\[
x \in R_y \iff \forall y' \in \mathcal{Y} \land y' \neq y, m \cdot d(x, c_{y}) \leq d(x, c_{y'})
\]

It is easy to prove that \( R_y = \bigcap_{y' \neq y} R_{y, y'} \). Note that if eq. 10 holds,

\[
\forall (x, y_x) \in D, \ x \in R_{y_x}
\]

and thus we can derive a sufficient condition for \( P_1 \) (eq. 9):

\[
\forall y \in \mathcal{Y}, \max_{x, x' \in R_y} d(x, x') \leq \min_{x \in R_y, x' \in \bigcup_{y' \neq y} R_{y'}} d(x, x')
\]

We denote eq. 11 as proposition \( P_2 \).

6.2 Finding Boundaries for \( m \)

The later induction is based on the assumption that our MEMS loss can be well optimized, i.e. eq. 10 holds, and the assumption that \( m \geq 1 \). Now the question is: what is the range of \( m \) that is sufficient and necessary for \( P_2 \)? In the rest of this section, we firstly calculate the closed form of \( R_{y, y'} \) and then prove that if \( m = m_0 \Rightarrow P_2 \), then \( m = m_0 + \varepsilon \Rightarrow P_2, \forall \varepsilon > 0 \). To this end, we only need to find the lower bound of \( m \), with regard to the number of categories: \( n_{\min}^{(\mathcal{Y})} \). Next, we prove \( n_{\min}^{(2)} = 2 + \sqrt{3} \). Finally, we prove \( n_{\min}^{(k)} = 2 + \sqrt{3}, \forall k \geq 3 \) is sufficient and necessary for \( P_2 \).

Lemma 1. If \( d(x, x') = \|x - x'\|_2 \), \( R_{y, y'} \) is a \( n \)-ball (ball in \( n \)-dimensional space) with center \( c_y + \frac{c_{y'} - c_y}{m-1} \) and radius \( \frac{m}{m-1} \|c_{y'} - c_y\|_2 \).
Proof If \( x \in \mathcal{R}_{y,y'} \),

\[
\begin{align*}
  m \| x - c_y \|_2 &\leq \| x - c_{y'} \|_2 \\
  m^2 \| x - c_y \|_2 &\leq \| x - c_{y'} \|_2 \\
  m^2 \left( x^T x - 2c_{y}^T x + c_{y}^T c_y \right) &\leq x^T x - 2c_{y'}^T x + c_{y'}^T c_{y'} \\
  (m^2 - 1) x^T x - 2 \left( m^2 c_{y}^T - c_{y'}^T \right) x &\leq c_{y'}^T c_{y'} - m^2 c_{y}^T c_y \\
  \left\| x - m^2 c_{y} - c_{y'} \right\|_2^2 &\leq \left( \frac{(m^2 - 1) (c_{y}^T c_{y} - m^2 c_{y}^T c_y) + m^2 c_{y}^T c_y}{(m^2 - 1)^2} \right) \\
  \left\| x - \left( \frac{m^2 c_{y} + c_{y'} - c_{y'}}{m^2 - 1} \right) \right\|_2^2 &\leq \left( \frac{m}{m^2 - 1} \right) \left\| c_{y} - c_{y'} \right\|_2^2 \\
  \left\| x - \left( c_{y}^T x + c_{y'}^T x \right) - \frac{m^2 c_{y} + c_{y'} - c_{y'}}{m^2 - 1} \right\|_2^2 &\leq \left( \frac{m}{m^2 - 1} \right) \left\| c_{y} - c_{y'} \right\|_2^2
\end{align*}
\]

Lemma 2. If \( m = m_0 \Rightarrow P_2 \), then \( m = m_0 + \epsilon \Rightarrow P_2 \), \( \forall \epsilon > 0 \)
Proof With \( m = m_0 \) slightly expanding to \( m = m_0 + \epsilon \), region \( \mathcal{R}_{y,y'} \) becomes \( \mathcal{R}_{y,y'}' \), where

\[
\forall x \in \mathcal{R}_{y,y'}, \quad d(x, c_{y'}) \geq (m_0 + \epsilon) d(x, c_y) \geq m_0 d(x, c_y)
\]

So we can conclude that \( \mathcal{R}_{y,y'}' \subseteq \mathcal{R}_{y,y'} \). Thus

\[
\forall y \in \mathcal{Y}, \quad \mathcal{R}_y' = \bigcap_{y' \neq y} \mathcal{R}_{y,y'}' \subseteq \bigcap_{y' \neq y} \mathcal{R}_{y,y'} = \mathcal{R}_y
\]

Now we rewrite \( P_2 \) as

\[
\forall y \in \mathcal{Y}, \forall y' \in \mathcal{Y} \land y' \neq y, \quad \max_{x,x' \in \mathcal{R}_y} d(x, x') \leq \min_{x \in \mathcal{R}_y, x' \in \mathcal{R}_{y'}} d(x, x')
\]

Since eq. 12,

\[
\max_{x,x' \in \mathcal{R}_y} d(x, x') \leq \max_{x,x' \in \mathcal{R}_y} d(x, x')
\]

\[
\min_{x \in \mathcal{R}_y, x' \in \mathcal{R}_{y'}} d(x, x') \geq \min_{x \in \mathcal{R}_y, x' \in \mathcal{R}_{y'}} d(x, x')
\]

which means if eq. 13 holds for \( m = m_0 \), it also holds for \( m = m_0 + \epsilon \).

Lemma 3. \( m_{(2)}^{(k)} = 2 + \sqrt{3} \) is sufficient and necessary for \( P_2 \).
Proof We can write \( \mathcal{Y} = \{1, 2\} \). Now \( \mathcal{R}_1 = \mathcal{R}_{1,1}, \mathcal{R}_2 = \mathcal{R}_{2,1} \), which are two n-balls with same radius and different centers. The maximum intra-class distance is the diameter of each n-ball:

\[
\forall y \in \mathcal{Y}, \quad \max_{x,x' \in \mathcal{R}_y} d(x, x') = \left( \frac{2m}{m^2 - 1} \right) \| c_1 - c_2 \|_2
\]

The minimum inter-class distance is the distance between two centers minus the diameter:

\[
\forall y \in \mathcal{Y}, \quad \min_{x \in \mathcal{R}_y, x' \in \mathcal{U}_{y' \neq y} \mathcal{R}_y} d(x, x') = \left( \frac{\left( c_1 + \frac{c_1 - c_2}{m^2 - 1} \right) - \left( c_2 + \frac{c_1 - c_2}{m^2 - 1} \right)}{m} \right) \| c_1 - c_2 \|_2
\]

Let \( \left( \frac{m^2 - 2m + 1}{m^2 - 1} \right) \| c_1 - c_2 \|_2 \geq \left( \frac{2m}{m^2 - 1} \right) \| c_1 - c_2 \|_2 \), we can get the result \( m \geq 2 + \sqrt{3} \) or \( m \leq 2 - \sqrt{3} \). We abandon the latter solution since only when \( m \geq 1 \) does it make sense. So in binary case class, \( m \geq m_{(2)}^{(k)} = 2 + \sqrt{3} \) if sufficient and necessary for \( P_2 \).

Lemma 4. \( m_{(k)}^{(k)} = 2 + \sqrt{3}, \forall k \geq 3 \) is necessary for \( P_2 \).
Proof Consider an extreme condition, where two prototypes \( c_{y_a}, c_{y_b} \) are far from the other prototypes. We notice that

\[
\forall y \in \mathcal{Y} \land y \neq y_a \land y \neq y_b \| c_{y_a} - c_{y_b} \|_2 \geq \frac{m+1}{m-1} \| c_{y_a} - c_{y_b} \|_2 \Rightarrow \min_{x \in \mathcal{R}_{y_a,y}} d(x, c_{y_a}) \geq \max_{x \in \mathcal{R}_{y_a,y}} d(x, c_{y_b}) \Rightarrow \mathcal{R}_{y_a,y} \subseteq \mathcal{R}_{y_a,y}
\]

Since we have no constraints on location of prototypes, this condition can always be likely to hold, regardless of the value of \( m \). When all the rest prototypes satisfy this condition for both \( c_{y_a}, c_{y_b} \), we have \( \mathcal{R}_{y_a} = \mathcal{R}_{y_a,y} \) and \( \mathcal{R}_{y_b} = \mathcal{R}_{y_b,y} \), which is same as in binary case. So \( m \geq m_{\min}^{(2)} \) becomes necessary to ensure the correctness of \( P_2 \), and thus Lemma 4 is true.

Lemma 5. \( m_{\min}^{(k)} = 2 + \sqrt{3}, \forall k \geq 3 \) is sufficient for \( P_2 \).

Proof If we want to prove that \( P_2 \) (eq. 13) holds, we have to show that every distinct pair \((y_a, y_b)\) satisfy \( \max_{x,x' \in \mathcal{R}_{y_a}} d(x, x') \leq \min_{x \in \mathcal{R}_{y_a}, x' \in \mathcal{R}_{y_b}} d(x, x') \). We remove a category \( y_c \), where \( y_c \neq y_a \) and \( y_c \neq y_b \), from \( \mathcal{Y} \) and forms \( \mathcal{Y}' \) such that \( |\mathcal{Y}'| = |\mathcal{Y}|-1 \).

Suppose \( m = m_{\min}^{(|\mathcal{Y}'|)} \) satisfies eq. 13, we have

\[
\max_{x,x' \in \mathcal{R}_{y_a}} d(x, x') \leq \min_{x \in \mathcal{R}_{y_a}, x' \in \mathcal{R}_{y_b}} d(x, x')
\]

where

\[
\mathcal{R}_{y_a} = \bigcap_{y \in \mathcal{Y}' \land y \neq y_a} \mathcal{R}_{y_a,y} = \bigcap_{y \in \mathcal{Y}' \land y \neq y_a} \mathcal{Y}_{y_a,y},
\]

\[
\mathcal{R}_{y_b} = \bigcap_{y \in \mathcal{Y}' \land y \neq y_b} \mathcal{R}_{y_b,y} = \bigcap_{y \in \mathcal{Y}' \land y \neq y_b} \mathcal{Y}_{y_b,y}
\]

When \( m \) is not changed and prototypes are not moved,

\[
\mathcal{R}_{y_a} \subseteq \mathcal{R}_{y_a}' \Rightarrow \max_{x,x' \in \mathcal{R}_{y_a}} d(x, x') \leq \max_{x \in \mathcal{R}_{y_a}, x' \in \mathcal{R}_{y_a}'} d(x, x')
\]

\[
\mathcal{R}_{y_b} \subseteq \mathcal{R}_{y_b}' \Rightarrow \min_{x,x' \in \mathcal{R}_{y_b}} d(x, x') \leq \min_{x \in \mathcal{R}_{y_b}, x' \in \mathcal{R}_{y_b}'} d(x, x')
\]

Thus, \( \max_{x,x' \in \mathcal{R}_{y_a}} d(x, x') \leq \min_{x \in \mathcal{R}_{y_a}, x' \in \mathcal{R}_{y_b}} d(x, x') \) is satisfied for any pair \((y_a, y_b)\) where \( y_a, y_b \in \mathcal{Y} \), even if we directly adopt \( m = m_{\min}^{(|\mathcal{Y}'|)} \) when \( |\mathcal{Y}'| = |\mathcal{Y}|-1 \). So we can conclude that \( m_{\min}^{(|\mathcal{Y}'|+1)} \) is sufficient for \( P_2 \). By Lemma 3, \( m_{\min}^{(2)} = 2 + \sqrt{3} \), we can conclude that \( m_{\min}^{(k)} = 2 + \sqrt{3}, \forall k \geq 3 \) is sufficient for \( P_2 \).