TVT: Three-Way Vision Transformer through Multi-Modal Hypersphere Learning for Zero-Shot Sketch-Based Image Retrieval

Jialin Tian[1], Xing Xu[1]*, Fumin Shen[1], Yang Yang[1], and Heng Tao Shen[1,2]

[1]Center for Future Media and School of Computer Science and Engineering
University of Electronic Science and Technology of China, China
[2]Peng Cheng Lab, China

Abstract

In this paper, we study the zero-shot sketch-based image retrieval (ZS-SBIR) task, which retrieves natural images related to sketch queries from unseen categories. In the literature, convolutional neural networks (CNNs) have become the de-facto standard and they are either trained end-to-end or used to extract pre-trained features for images and sketches. However, CNNs are limited in modeling the global structural information of objects due to the intrinsic locality of convolution operations. To this end, we propose a Transformer-based approach called ***T**hree-Way **V**ision **T**ransformer (TVT)* to leverage the ability of Vision Transformer (ViT) to model global contexts due to the global self-attention mechanism. Going beyond simply applying ViT to this task, we propose a token-based strategy of adding fusion and distillation tokens and making them complementary to each other. Specifically, we integrate three ViTs, which are pre-trained on data of each modality, into a three-way pipeline through the processes of distillation and multi-modal hypersphere learning. The distillation process is proposed to supervise fusion ViT (ViT with an extra fusion token) with soft targets from modality-specific ViTs, which prevent fusion ViT from catastrophic forgetting. Furthermore, our method learns a multi-modal hypersphere by performing inter- and intra-modal alignment without loss of uniformity, which aims to bridge the modal gap between modalities of sketch and image and avoid the collapse in dimensions. Extensive experiments on three benchmark datasets, *i.e.*, Sketchy, TU-Berlin, and QuickDraw, demonstrate the superiority of our TVT method over the state-of-the-art ZS-SBIR methods.

Introduction

Sketch-based image retrieval (SBIR) (Eitz et al. 2010; Saavedra, Barrios, and Orand 2015) is a practical problem that the sketch is used as a query to retrieve relevant images from the gallery. The conventional SBIR scenario assumes that training and testing data come from the distributions of the same categories. Many methods (Sangkloy et al. 2016; Liu et al. 2017a) have achieved satisfying performance in this scenario with the help of a large number of annotated samples. However, annotating samples is labor-intensive and time-consuming, as well as these methods perform poorly on data of unseen classes. Consequently, there has been some work (Shen et al. 2018; Dey et al. 2019) focused on studying the SBIR problem in the zero-shot setting, which assumes that the training class set and the test class set are disjoint. So *zero-shot SBIR (ZS-SBIR)* is a more challenging problem for the inherent modal gap as well as the semantic gap brought by the zero-shot setting.

So far, most existing ZS-SBIR methods largely rely on convolutional neural networks (CNNs), *i.e.*, they either fine-tune the pre-trained CNNs to extract features and then build projection models to learn a shared embedding space (Dutta and Akata 2019; Hwang et al. 2020), or train the whole model in an end-to-end manner (Liu et al. 2019; Wang et al. 2021). In spite of the excellent representational power of CNNs, these methods are also limited in modeling the global structural information due to the inherent local nature of

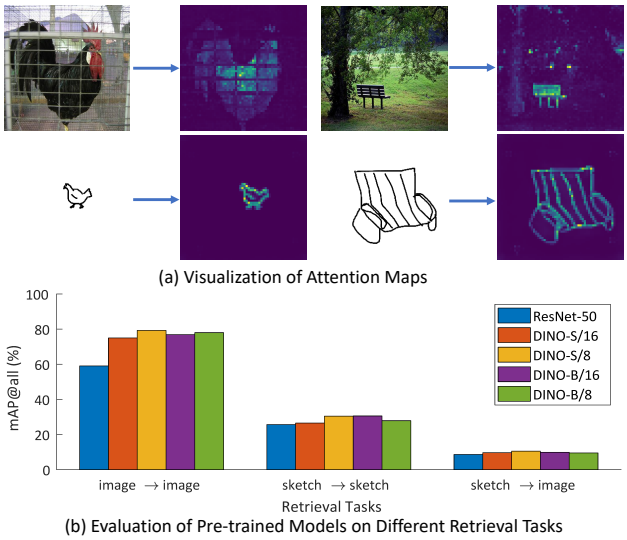


Figure 1: Illustration of (a) the visualization of the attention maps of images and sketches in the self-attention modules from the last layer of the latest pre-trained model DINO-S/8 (Caron et al. 2021), and (b) the comparison between pre-trained ResNet-50 and DINO variants under different tasks with unseen data of Sketchy (non-overlapping with ImageNet).

*Corresponding author

convolution operations. However, global structural information is essential for the ZS-SBIR task since the only information that the image and sketch together contain is the global structural information of the object. Vision Transformer (ViT) has demonstrated that it is a advanced alternative to the CNN framework, with the global structural information modeling capability and exceptional transferability. In particular, ViT pre-trained in a self-supervised manner (*e.g.*, DINO (Caron et al. 2021)) surprisingly shows a segmentation property. The suffixes of DINO variants indicate the model size and input patch size, where DINO-S/8 (DINO-B/16) means the "Small" ("Base") variant with 8×8 (16×16) patch size (as shown in Fig. 1). We can see that the DINO model explicitly learns the object boundaries of images and sketches (Fig. 1(a)), ignoring occlusions and backgrounds. The effect of the segmentation property is also reflected in the retrieval tasks (Fig. 1(b)): DINO-S/8 outperforms ResNet-50 (He et al. 2016) by a margin in both intra- and inter-modal retrieval tasks on unseen data of Sketchy (Yelamarthi et al. 2018).

Motivated by the above observation, we take the first step in this paper towards utilizing the global structure modeling capability of ViT for the ZS-SBIR task. Specifically, we propose a novel approach named ***T**hree-Way **V**ision **T**ransformer (TVT)*, which integrates two modality-specific ViTs and a fusion ViT into a three-way pipeline by a token-based strategy. As the general framework of our proposed TVT model shown in Fig. 2, the modality-specific ViTs, pre-trained in a self-supervised manner, is to provide global structural information specific to each image and sketch for fusion ViT. In addition, fusion ViT is the model that an extra fusion token is added to interact with other tokens through the self-attention mechanism, whose output is mapped into a common hypersphere to alleviate the modal gap. Besides, we design a novel method for multi-modal hypersphere learning. The representations of each class are required to be well clustered regardless of modalities (both inter- and intra-modal alignment), but representations of each modality are only individually encouraged to approach the uniform distribution (intra-modal uniformity). As a result, the distributions of each class for both modalities are overlapped on the hypersphere, while avoiding collapse in dimensions. Finally, the fusion ViT is optimized to perform multi-modal hypersphere learning through the fusion token and preserve global structural information through the distillation tokens. Extensive experiments on three benchmark datasets of ZS-SBIR verify the superiority of our TVT method.

We summarize the main contributions of this work as:

- To the best of our knowledge, we are the first to model the global structural information in the field of ZS-SBIR using Vision Transformer, which is critical for the alignment between sketches and images.
- We propose a novel Three-Way Vision Transformer method termed TVT based on the distillation tokens and fusion token. These two types of tokens play the same role as the normal class token, except that the former is used for knowledge distillation and the latter for eliminating the modal gap.
- We devise a novel multi-modal hypersphere learning process that effectively leverages the representational power of the hypersphere by inter- and intra-modal alignment and intra-modal uniformity.

Related Work

Zero-Shot Sketch-Based Image Retrieval. ZS-SBIR is a challenging task that simultaneously addresses the inherent modal gap and the semantic gap brought by the zero-shot setting. Pioneer work (Shen et al. 2018) first studied the SBIR problem under the zero-shot setting by cross-modal learning (Shen et al. 2021; Xu et al. 2020b). The subsequent work mainly used fine-tuned pre-trained CNNs to extract features and then built projection models to learn a joint embedding space (Xu et al. 2021, 2020a) with the help of semantic information, including the generative adversarial network (Dutta and Akata 2019), the adversarial network with Gradient Reversal Layer (Dey et al. 2019), the content-style disentanglement model (Dutta and Biswas 2019), and so on. (Yelamarthi et al. 2018; Hwang et al. 2020) adopted variational auto-encoder (VAE) to learn latent embedding space but without semantic information. Unlike the above methods, (Liu et al. 2019) presented a framework that trains CSE-ResNet-50 (Lu et al. 2018) with knowledge distillation in an end-to-end manner, through which features are extracted and cross-modal retrieval is conducted. (Wang et al. 2021) improved this model by tackling the large intra-class diversity of sketches with a category-specific memory bank. However, all these methods largely rely on CNNs and are consequently limited in modeling global structural information, which is greatly important for ZS-SBIR. In this paper, we take the first step to use ViT's global structure modeling capability for the ZS-SBIR task.

Vision Transformer. The architecture of Transformer was firstly introduced by (Vaswani et al. 2017) for machine translation and has currently become the de-facto standard for its tremendous success. Subsequently, several attempts (Hu, Shen, and Sun 2018; Wang et al. 2018b; Li et al. 2019; Ramachandran et al. 2019; Zhang et al. 2020) have been devoted to adapting the mechanism of Transformer to CNNs. More recently, (Dosovitskiy et al. 2021) proposed a convolution-free method that directly applies Transformer to the sequence of image patches, which achieved state-of-the-art results on the image recognition task. (Touvron et al. 2021) subsequently addressed the problem of ViT requiring huge amounts of data and computation from the perspective of knowledge distillation, producing competitive results by training on ImageNet (Deng et al. 2009) solely. (Caron et al. 2021) investigated the impact of self-supervised pre-training for ViT and the resulting model showed a superior segmentation property and performed particularly well with a k-NN classifier alone. In this paper, we leverage the segmentation property to align sketches and images by captured global structural information.

Representation Learning on the Hypersphere. (Liu et al. 2017b; Davidson et al. 2018; Xu and Durrett 2018; Wang et al. 2018a) have shown that learning presentations on hypersphere performs better than Euclidean space since

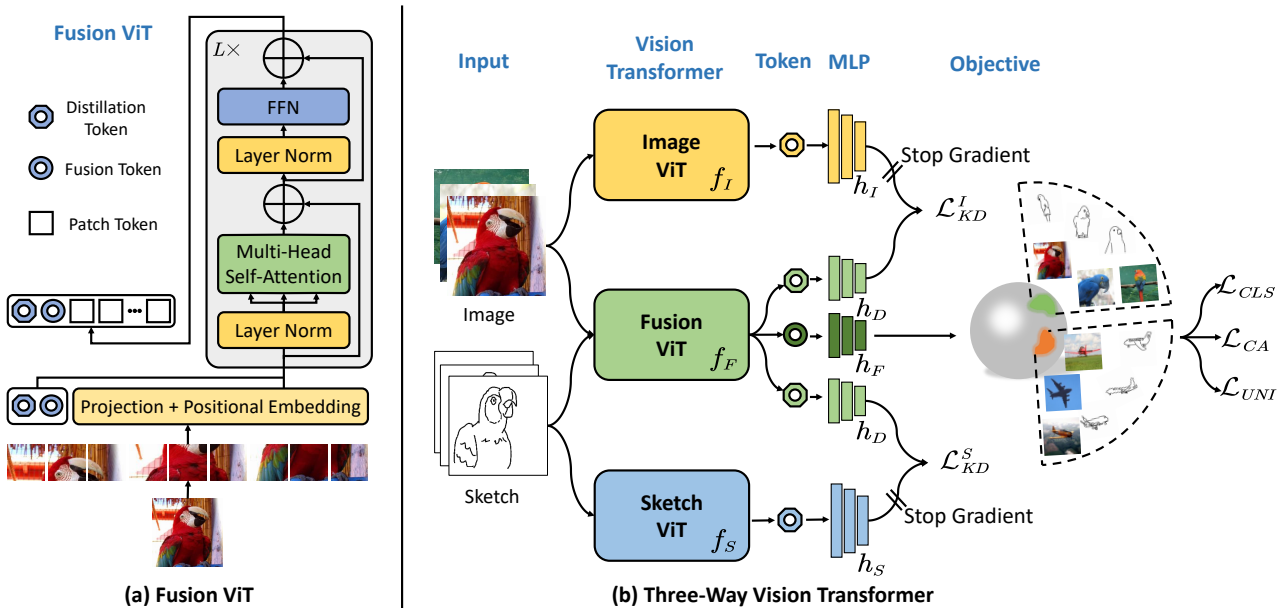


Figure 2: The illustration of basic architectures of (a) our proposed fusion ViT with an additional fusion token and (b) Three-Way Vision Transformer, respectively. Images and sketches are fed into the fusion ViT and modality-specific ViTs, which are pre-trained in a self-supervised manner. The output of the distillation token of fusion ViT is to predict those of modality-specific ViTs, while the output of the fusion token is to learn a common hypersphere. For clarity, we have drawn the distillation token of fusion ViT and the associated MLP (multi-layer perceptron) symmetrically twice.

angular information preserves key semantics rather than the magnitude. (Sablayrolles et al. 2019) presented a differential entropy regularizer derived from the estimator by (Kozachenko and Leonenko 1987), which was subsequently applied to image retrieval with contrastive loss by (El-Nouby et al. 2021). (Wang and Isola 2020) analyzed the behavior of contrastive learning theoretically and experimentally and argued that optimizing contrastive loss is equivalent to optimizing the two properties of alignment and uniformity. Inspired by this research, we propose multi-modal hypersphere learning to perform intra- and inter-modal alignment with intra-modal uniformity.

Proposed Method

Problem Definition

We first describe the definition of zero-shot sketch-based image retrieval. The goal of this task is to train a model on the training images and sketches from seen classes and then apply it to extract common representations of unseen data for retrieval. The training dataset of seen classes is denoted as $\mathcal{D}_s = \{\mathcal{I}_s, \mathcal{S}_s\}$, where $\mathcal{I}_s$ and $\mathcal{S}_s$ represent sets of natural images and sketches from seen classes $\mathcal{Y}_s$, respectively. Mathematically, they are formulated as $\mathcal{I}_s = \{(x_i^I, y_i)|y_i \in \mathcal{Y}_s\}_{i=1}^{N_1}$ and $\mathcal{S}_s = \{(x_j^S, y_j)|y_j \in \mathcal{Y}_s\}_{j=1}^{N_2}$, where N_1 and N_2 mean the cardinality of the $\mathcal{I}_s$ and $\mathcal{S}_s$, respectively. Similarly, the test dataset can be consistently defined as $\mathcal{D}_u = \{\mathcal{I}_u, \mathcal{S}_u\}$ for unseen categories $\mathcal{Y}_u$. Note that under the zero-shot scenario, the scope of seen and unseen classes are disjoint, *i.e.*, $\mathcal{Y}_s \cap \mathcal{Y}_u = \phi$. This setting implies that we need to improve the generalization of the model trained on limited data.

Network Architecture

The overall framework of the proposed TVT method is illustrated in Fig.2. The DINO model $g = h \circ f$ is composed of a ViT backbone f and an additional projection head h (an MLP), whose output is a K-D vector treated as probabilities to achieve self-distillation training. In this way, our TVT model consists of two modality-specific ViTs (f_I and f_S for modalities of image and sketch, respectively) with their corresponding projection heads (h_I and h_S) and a fusion ViT (f_F) with two projection heads (h_D and h_F for the distillation and fusion token, respectively). For brevity, we hereafter use g_I, g_S, g_D, and g_F to denote the joint operations of corresponding f and h. Then, we integrate them into a three-way pipeline through the processes of distillation and hypersphere learning. The distillation process allows the fusion ViT to reconcile the outputs of h_D with those of h_I and h_S to prevent catastrophic forgetting. Furthermore, the hypersphere learning process aims to learn good representations by performing inter- and intra-modal alignment without loss of uniformity on the unit hypersphere. It is implemented by a token-based strategy that adds a new fusion token to the initial embedding, as shown in Fig. 2(a). The detailed procedure of our TVT method will be described in the remainder of this section.

Image ViT

Firstly, let us briefly review the mechanism of ViT. It consists of alternating L layers of multi-head self-attention (MSA) and Feed-Forward Network (FFN) blocks. Both MSA and FFN layers contain "pre-norm" layer normalization and are paralleled with skip connections. ViT takes as input a sequence of image patches of fixed resolution $n \times n$. These patches are then linearly projected and added a learnable positional embedding to form a sequence of vector-shaped tokens. An extra learnable class token is incorporated into the sequence to aggregate information from other tokens such that it serves as a global image description. We refer to (Vaswani et al. 2017) for the basic theory of Transformer and (Dosovitskiy et al. 2021) for its adaptation to vision tasks.

In this paper, we choose the DINO-S/8 variant (Caron et al. 2021) as the basic architecture for the sake of its excellent performance (shown in Fig. 1) and compact model size, which is even less than ResNet-50 in terms of parameters count. Since the class token of DINO is not attached to any label nor supervision and is instead used for distillation, we renamed it to distillation token to avoid ambiguity.

Sketch ViT

Since the original DINO-S/8 model is not trained on sketches data, we firstly fine-tune it on sketches of seen categories in the same self-supervised manner as DINO. Notably, the labels of sketches are excluded from this process to avoid the loss of the pre-trained model's generalization. This is because the sketch only abstractly depicts the structural information of the object, without the complex textures and background variations like an image. If we fine-tune the pre-trained model based on the supervised signal, this inevitably results in modality-specific overfitting, which is detrimental to the subsequent training process.

More specifically, we utilize the multi-crop strategy to generate a set V of various views for each sketch, which consists of two global views ($x^S_{g,1}$ and $x^S_{g,2}$) with a resolution of 224^2 and ten local views with a resolution of 96^2. Then, we build the teacher-student architecture that the teacher and student are both initialized from the same pre-trained weights. Specifically, the optimization follows the "local-to-global" strategy by feeding all views of V into the student while only feeding $x^S_{g,1}$ and $x^S_{g,2}$ into the teacher. We denote by Z_t, τ_t, and θ_t (Z_s, τ_s, and θ_s) the logits, temperature, and parameters for the teacher (student) and ψ the softmax operation. Finally, the objective can be formulated as:

$$\min_{\theta_s} \sum_{x \in \left\{x^S_{g,1}, x^S_{g,2}\right\}} \sum_{\substack{x' \in V \\ x' \neq x}} \mathrm{KL}\left(\psi\left(Z_t(x)/\tau_t\right), \psi\left(Z_s\left(x'\right)/\tau_s\right)\right), \tag{1}$$

where $\theta_t = \zeta\theta_t + (1-\zeta)\theta_s$ is updated by the exponential moving average of θ_s and taken as the sketch ViT for the subsequent training.

Fusion ViT

Distillation through Tokens. After obtaining two modality-specific ViTs with associated heads, we start to train fusion ViT with supervision from them. Since the two modality-specific ViTs are pre-trained in a self-supervised manner, they are encouraged to discover global structural information specific to each image and sketch. However, the fusion ViT aims to reduce the modal gap between images and sketches of the same category, which will inevitably require the model to pay more attention to the more discriminative local structures shared by the whole category, gradually forgetting the structural information specific to each instance. Therefore, we avoid this catastrophic forgetting phenomenon by knowledge distillation. Given a batch of N images, we reconcile the probability vectors given by h_D with those of h_I through the distillation tokens, which is formulated as follows:

$$\mathcal{L}^I_{KD} = \sum_{i=1}^{N} \mathrm{KL}\left(\psi\left(g_I(x^I_i)/\tau_t\right), \psi\left(g_D(x^I_i)/\tau_s\right)\right), \tag{2}$$

where τ_t, τ_s, and ψ are the same as defined previously. Similarly, $\mathcal{L}^S_{KD}$ is the knowledge distillation loss for the modality of sketch. Then, we define $\mathcal{L}_{KD}$ as follows:

$$\mathcal{L}_{KD} = \mathcal{L}^I_{KD} + \mathcal{L}^S_{KD}. \tag{3}$$

In this way, we prevent our model from reducing the rich visual information to a limited number of concepts selected from the thousands of object classes acquired by pre-training.

Inter- and Intra-Modal Alignment. As shown in Fig. 2(b), the fusion tokens of images and sketches are jointly projected into a unit hypersphere in which the images and sketches of the same class are expected to be well clustered. When all classes are well clustered, they are linearly separable in the hypersphere space. Therefore, we classify the samples using a linear classifier:

$$\mathcal{L}_{CLS} = -\mathbb{E}\left[\log P\left(y_i \mid g_F(x_i); \theta_c\right)\right], \tag{4}$$

where x_i can be an image or a sketch, θ_c is the parameters of the shared classifier. Consequently, $\mathcal{L}_{CLS}$ can perform intra-modal alignment as well as inter-modal alignment by the shared classifier. We also propose a center alignment loss that explicitly requires the distributions of sketches and images to overlap on the hypersphere:

$$\begin{aligned} c^*_{y_i} &= \lambda c^*_{y_i} + (1-\lambda) \sum_{j=1}^{N_{y_i}} \left[g_F\left(x^*_j\right)\right], \\ c^*_{y_i} &= \frac{c^*_{y_i}}{\left\|c^*_{y_i}\right\|_2}, \quad * \in \{I, S\}, \\ \mathcal{L}_{CA} &= \sum_{i \in \mathcal{Y}_s} (c^I_i - c^S_i)^2. \end{aligned} \tag{5}$$

Here λ is the weight of the exponential moving average. N_{y_i} is the number of samples x^*_j with the label y_i in the batch. I and S indicate the modalities of image and sketch. The centers are l_2-normalized to map back to the hypersphere. Equipped with L_{CLS} and L_{CA}, we align the distributions of the bi-modal data from both inter-modal and intra-modal aspects.

Algorithm 1: Overall training procedure of TVT.

Phase 1: Fine-tuning sketch ViT.

Input: $\mathcal{S}_s = \{(x_j^S, y_j)|y_j \in \mathcal{Y}_s\}_{j=1}^{N_2}$, batch size N, exponential moving average ζ.

Output: Model Parameters θ_t

1: Build a teacher-student architecture (θ_t and θ_s).
2: **repeat**
3: Sample a batch of sketches.
4: Update θ_s using Adam optimizer with Eq. 1.
5: $\theta_t = \zeta\theta_t + (1-\zeta)\theta_s$.
6: **until** Reach maximum iterations.
7: Take θ_t as parameters of Sketch ViT.

Phase 2: Training TVT.

Input: $\mathcal{I}_s = \{(x_i^S, y_i)|y_i \in \mathcal{Y}_s\}_{i=1}^{N_1}$, $\mathcal{S}_s = \{(x_j^S, y_j)|y_j \in \mathcal{Y}_s\}_{j=1}^{N_2}$, batch size N, image ViT g_I, sketch ViT g_S, learning rate μ, hyper-parameters λ, λ_1, λ_2.

Output: Model parameters θ_{f_F}, θ_{h_F}, θ_{h_D}.

1: Build the three-way pipeline.
2: **repeat**
3: Sample a batch of images and sketches.
4: Compute the objective $\mathcal{L} \leftarrow$ Eq. 7.
5: $\theta_{f_F} \leftarrow \theta_{f_F} - 0.1 * \mu\nabla_{\theta_{f_F}}\mathcal{L}$.
6: $\theta_{h_*} \leftarrow \theta_{h_*} - \mu\nabla_{\theta_{h_*}}\mathcal{L}, \quad * \in \{F, D\}$.
7: **until** Reach maximum iterations.
8: Take trained fusion ViT to conduct ZS-SBIR.

Intra-Modal Uniformity. Both alignment and uniformity are key properties of representations in the hypersphere, where uniformity implies an efficient use of the representational power of the hypersphere. Specifically, we adopt the average Gaussian potential to encourage the uniformity of sketches or images:

$$\begin{aligned} G_t(x_i^*, x_j^*; g_F) &= e^{-t\|g_F(x_i^*) - g_F(x_j^*)\|_2^2}, \\ \mathcal{L}_{UNI}^* &= \log \mathbb{E}\left[G_t(x_i^*, x_j^*; g_F)\right], \quad * \in \{I, S\}, \\ \mathcal{L}_{UNI} &= \mathcal{L}_{UNI}^I + \mathcal{L}_{UNI}^S, \end{aligned} \tag{6}$$

where t is a fixed parameter. It is worth noting that L_{UNI} is separately applied on the distribution of each modality, rather than constraining all representations regardless of modalities. Such a design is reasonable because the distributions of both modalities on the hypersphere are expected to approach the uniform distribution, but with overlapping positions learned by inter- and intra-modal alignment.

Overall Objective. Finally, the overall objective of the fusion ViT is the linear combination of the four losses as:

$$\mathcal{L} = \mathcal{L}_{KD} + \mathcal{L}_{CLS} + \lambda_1\mathcal{L}_{CA} + \lambda_2\mathcal{L}_{UNI}, \tag{7}$$

where λ_1 and λ_2 are the hyper-parameters of center alignment loss and uniformity loss. The training procedure of our TVT method is shown in Algorithm 1.

Experiments

Experimental Setup

Datasets. We verify the effectiveness of our TVT method on three benchmark datasets of SBIR, *i.e.*, Sketchy (Sangkloy et al. 2016), TU-Berlin (Eitz et al. 2010), and QuickDraw (Dey et al. 2019). **Sketchy** is originally composed of 75,471 sketches and 12,500 natural images from 125 classes. Then (Liu et al. 2017a) extended this dataset with additional 60,502 images, so yielding in total 73,002 images. **TU-Berlin** consists of sketches of 250 categories, with 80 sketches each. It is extended by the collection of 204,489 images provided by (Liu et al. 2017a). **QuickDraw** contains 330,000 sketches and 204,000 images from 110 classes, which makes it the largest dataset among three datasets with the most abstract sketches drawn by the amateur.

Evaluation Setting. There are two kinds of seen and unseen class divisions for Sketchy: the one proposed by (Liu et al. 2017a) randomly selects 25 classes as unseen classes, while the one proposed by (Yelamarthi et al. 2018) selects classes that do not overlap with ImageNet categories as unseen classes. For simplicity, we refer to the former one as **Sketchy** and the latter one as **Sketchy-NO**. **TU-Berlin** is similar to Sketchy in that 30 randomly selected classes are used as unseen classes. However, **QuickDraw** is similar to Sketchy-NO in that it selects 30 classes that do not overlap with ImageNet categories as unseen classes. The output of the fusion token is taken as the retrieval feature. In addition, we binarize the real features by iterative quantization (ITQ) (Gong et al. 2012) for comparison. The cosine and hamming distance metrics are used to compute the similarities for real and binary embedding, respectively.

Implementation Details. We implement TVT with the popular PyTorch toolkit. For our network architecture, the ViTs consist of 12 Transformer blocks (an MSA and an FFN) with 6 heads in multi-head self-attention. The projection heads contain three fully connected layers with dimensions [2048, 2048, 256] followed by l_2 normalization and an additional output layer for distillation (fixed 65536-D) or classification. We train the model in 50 epochs with Adam optimizer with weight decay that is initially 0.04 and is ramped up to 0.4 by a cosine schedule. The batch size of 512 samples is distributed over two GPUs with 16 steps of gradient accumulation. The base value of the learning rate μ is set to 0.0005 * (batch size/256). μ is raised linearly to the base value during the first 5 epochs and is decayed to 1e-6 by a cosine schedule as well. Especially, the learning rates are set to 0.1*μ for the ViTs but μ for the projection heads. The temperature τ_s is always 0.1 while τ_t increases linearly from 0.04 to 0.07 during the initial 5 epochs. We follow the data augmentations of DINO (Caron et al. 2021), which consist of color jittering, Gaussian blur, and solarization. The fixed t in Eq. 6 is set to 2 according to (Wang and Isola 2020). What's more, sketch ViT and class centers are updated with ζ = 0.996 and λ = 0.9, respectively. Finally, λ_1 and λ_2 are set to 2.0 and 0.5 in all experiments, unless specified otherwise. Further implementation codes and additional experimental analyses can be found in the **supplementary material**.

Comparing with the State-of-the-Arts

We compare our TVT method with 10 state-of-the-art methods relevant to the ZS-SBIR task, including CAAE (Yelamarthi et al. 2018), CVAE (Yelamarthi et al. 2018), SEM-

Table 1: Comparison of our method and 10 compared approaches on Sketchy and TU-Berlin. The subscript "b" denotes results obtained by binary hashing codes, and "-" means that the results are not reported in the original papers. "Sketchy-NO" is short for Sketchy with non-overlapping classes. The best and second-best results are marked in bold and underlined, respectively.

Methods	Dim	Sketchy-NO		Sketchy		TU-Berlin	
		mAP@200	Prec@200	mAP@all	Prec@100	mAP@all	Prec@100
CAAE (ECCV'2018)	4096	0.156	0.260	0.196	0.284	-	-
CVAE (ECCV'2018)	4096	0.225	0.333	-	-	0.005	0.001
ZSIH (CVPR'2018)	64	-	-	0.254	0.340	0.220	0.291
SEM-PCYC$_b$ (CVPR'2019)	64	-	-	0.344	0.399	0.293	0.392
SEM-PCYC (CVPR'2019)	64	-	-	0.349	0.463	0.297	0.426
Dey *et al.* (CVPR'2019)	256	0.369	0.370	-	-	0.110	0.121
SAKE$_b$ (ICCV'2019)	64	0.356	0.477	0.364	0.487	0.359	0.481
SAKE (ICCV'2019)	512	0.497	<u>0.598</u>	0.547	0.692	0.475	<u>0.599</u>
LCALE (AAAI'2020)	64	-	-	0.476	0.583	-	-
OCEAN (ICME'2020)	64	-	-	0.462	0.590	0.333	0.467
IIAE (NeurIPS'2020)	64	0.373	0.485	0.573	0.659	0.412	0.503
DSN$_b$ (IJCAI'2021)	64	0.367	0.481	0.436	0.553	0.385	0.497
DSN (IJCAI'2021)	512	<u>0.501</u>	0.597	<u>0.583</u>	<u>0.704</u>	<u>0.481</u>	0.586
TVT$_b$ (Ours)	64	0.447	0.554	0.553	0.727	0.396	0.606
TVT (Ours)	384	**0.531**	**0.618**	**0.648**	**0.796**	**0.484**	**0.662**

Table 2: Overall comparison of TVT and 2 compared approaches on large-scale QuickDraw. The best results are shown in bold.

Methods	QuickDraw		
	mAP@all	mAP@200	Prec@200
CVAE	0.003	0.006	0.003
Dey *et al.*	0.075	0.090	0.068
TVT (Ours)	**0.149**	**0.191**	**0.293**

PCYC (Dutta and Akata 2019), Dey *et al.* (Dey et al. 2019), SAKE (Liu et al. 2019), IIAE (Hwang et al. 2020), LCALE (Lin et al. 2020), OCEAN (Zhu et al. 2020), and DSN (Wang et al. 2021). We report the results on Sketchy-NO, Sketchy, and TU-Berlin in Table 1 and the results on QuickDraw in Table 2. Since IIAE and DSN are the two latest competitive approaches, we implement them according to their public codes and instructions, and we report their results on Sketchy-NO, in addition to IIAE on TU-Berlin.

As we can see, our TVT method shows a consistent and significant improvement over all of the state-of-the-art (SOTA) methods. Most of ZS-SBIR methods only experimented on Sketchy and TU-Berlin, which share the same way of randomly selecting unseen classes. Specifically, on these two datasets, TVT consistently beats the SOTA (DSN) with 11.1% and 0.5% improvements of mAP@all scores, respectively. However, few of them experimented on more realistic and challenging datasets: Sketchy-NO and QuickDraw guarantee that the unseen classes do not overlap with ImageNet, in addition to QuickDraw being a very large dataset. On Sketchy-NO, our approach improves the mAP@200 score from 0.501 to 0.531 compared with DSN. Moreover, on the large-scale QuickDraw, it achieves a huge improvement of almost 100% mAP@all score. Given the large-scale nature of these datasets and the limitation of the fixed class splits, these results effectively prove that the dramatic improvement of our method is not by chance or by split bias. Compared with hashing methods, our method also gets the best results. When we compare the results using metrics that consider only top k candidates, the improvement achieved by our method is more pronounced. On Sketchy and TU-Berlin, our approach surpasses DSN with 13.1% and 13.0% improvements of Prec@100 scores, respectively. On QuickDraw, it gains increases of 112.2% and 330.9% of mAP@200 and Prec@200 scores, respectively. These results mean that the true positive examples have a higher probability of appearing in the top 100 (or 200) retrieved results, which is well suited to the retrieval task.

All these comparisons can demonstrate that our method can effectively align intra- and inter-modal distributions without loss of uniformity and then achieves satisfactory generalization on unseen classes.

Further Analysis on TVT

Table 3: Ablation results (mAP@all) for each loss term on Sketchy and TU-Berlin. The best results are shown in bold.

Models	Sketchy	TU-Berlin
DINO-S/8	0.101	0.084
TVT w/o $\mathcal{L}_{CLS}$	0.286	0.244
TVT w/o $\mathcal{L}_{KD}$	0.599	0.452
TVT w/o $\mathcal{L}_{CA}$	0.630	0.476
TVT w/o $\mathcal{L}_{UNI}$	0.634	0.479
Full TVT	**0.648**	**0.484**

Ablation Study. We first investigate the effect of each loss term in Eq. 7 by ablating it in Eq. 7 in the training phase. The

results of these variants, the full TVT and pre-trained DINO-S/8 on Sketchy and TU-Berlin are shown in Table 3, where "w/o" means the ablating behavior.

From the comparison of these models, we can draw the following conclusions: 1) TVT w/o $\mathcal{L}_{CLS}$ performs worse than the other variants as it fails to consider inter-modal alignment. However, it is better than DINO-S/8, demonstrating that the center alignment and three-way training pipeline can align inter-modal distributions to some extent. 2) The performance of TVT w/o $\mathcal{L}_{KD}$ shows that learning by focusing only on the fusion ViT will inevitably lead to catastrophic forgetting, which clarifies the need for three-way training through distillation tokens. 3) The results of TVT w/o $\mathcal{L}_{CA}$ indicate that explicitly required overlap of class centers on the hypersphere facilitates the elimination of the modal gap. 4) The results of TVT w/o $\mathcal{L}_{UNI}$ suggest that uniformity effectively prevents the reduction of generalization caused by dimensional collapse on the hypersphere. 5) The full model achieves the best results with both advantages of knowledge distillation and hypersphere learning.

Qualitative Analysis. Fig. 3 shows the top 10 retrieved candidates of sketches queries, where correct and incorrect candidates are marked with checkmarks and crosses, respectively. Our model successfully retrieves the correct candidates in most cases, except for some structurally similar incorrect candidates. For example, the hot air balloons (penultimate row) are so similar to the parachutes in structure and background that they are retrieved incorrectly.

Fig. 4 visualizes the distributions of seen and unseen data of Sketchy by t-SNE (Van der Maaten and Hinton 2008). We can see that the seen data are well clustered together regardless of modalities, but with a certain degree of uniformity. Besides, all classes are separated by proper distances. The unseen data are not involved in the training, but they are also able to cluster together at relatively small distances based on the classes.

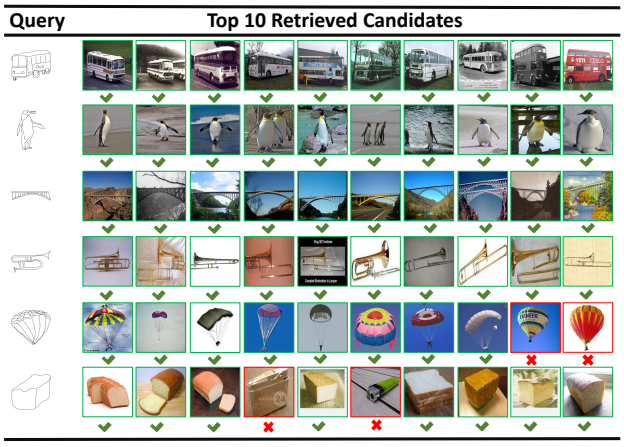
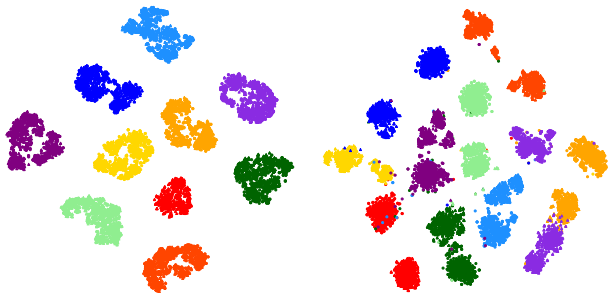


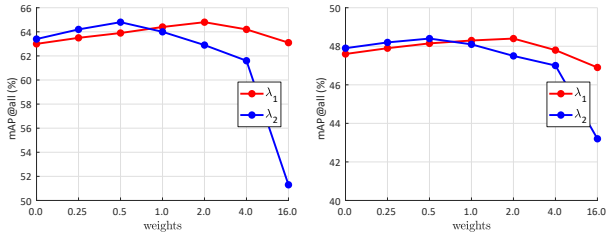
Figure 3: Retrieval examples of ZS-SBIR results on unseen data of TU-Berlin.

Analysis on Parameter Sensitivity. As shown in Fig. 5, we analyze the effect of center alignment and uniformity with varying hyper-parameters λ_1 and λ_2 in Eq. 7 on Sketchy and TU-Berlin. We can observe that the effect of center alignment is less influenced by λ_1 and reach the peak at $\lambda_1 = 2$. However, the effect of uniformity shows a different trend: it accelerates the deterioration of the retrieval results when λ_2 grows too large. It indicates the different importance of alignment and uniformity.



(a) Seen data (b) Unseen data

Figure 4: The t-SNE visualization for seen and unseen data of Sketchy, where the colored circles (•) and upper triangles (▲) represent images and sketches (zoom in for better viewing), respectively.



(a) Sketchy (b) TU-Berlin

Figure 5: ZS-SBIR mAP@all scores on Sketchy and TU-Berlin with different values of λ_1 and λ_2 for center alignment and uniformity, respectively

Conclusions

In this paper, we took the first step to leverage ViT to model the global structure of objects, which is essential for ZS-SBIR. We firstly proposed a novel yet effective Three-Way Vision Transformer that integrates modality-specific ViTs and our proposed fusion ViT into a three-way pipeline. Then, we trained the fusion ViT by a devised token-based strategy with distillation, which aims to prevent catastrophic forgetting, and multi-modal hypersphere learning, which encourages the representations to be well clustered without loss of uniformity according to their class. We conducted extensive experiments on three benchmark datasets to demonstrate the superiority of our approach and establish new state-of-the-art performance. In the future, we will investigate the performance of our approach on other multi-modal multi-view datasets.

Acknowledgments

This work was supported in part by the Sichuan Science and Technology Program, China (No. 2019ZDZX0008, 2019YFG0003, 2019YFG0533 and 2020YFS0057); National Natural Science Foundation of China under Grants (No. 61976049, 62072080 and U20B2063).

References

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. arXiv:2104.14294.

Davidson, T. R.; Falorsi, L.; De Cao, N.; Kipf, T.; and Tomczak, J. M. 2018. Hyperspherical Variational Auto-Encoders. *34th Conference on Uncertainty in Artificial Intelligence (UAI-18)*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dey, S.; Riba, P.; Dutta, A.; Llados, J.; and Song, Y.-Z. 2019. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2179–2188.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Dutta, A.; and Akata, Z. 2019. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5089–5098.

Dutta, T.; and Biswas, S. 2019. Style-Guided Zero-Shot Sketch-based Image Retrieval. In *British Machine Vision Conference 2019*, 209–213.

Eitz, M.; Hildebrand, K.; Boubekeur, T.; and Alexa, M. 2010. An evaluation of descriptors for large-scale image retrieval from sketched feature lines. *Computers & Graphics*, 34(5): 482–498.

El-Nouby, A.; Neverova, N.; Laptev, I.; and Jégou, H. 2021. Training Vision Transformers for Image Retrieval. arXiv:2102.05644.

Gong, Y.; Lazebnik, S.; Gordo, A.; and Perronnin, F. 2012. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 35(12): 2916–2929.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Hwang, H.; Kim, G.-H.; Hong, S.; and Kim, K.-E. 2020. Variational Interaction Information Maximization for Cross-domain Disentanglement. *Advances in Neural Information Processing Systems*, 33.

Kozachenko, L.; and Leonenko, N. N. 1987. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2): 9–16.

Li, X.; Wang, W.; Hu, X.; and Yang, J. 2019. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 510–519.

Lin, K.; Xu, X.; Gao, L.; Wang, Z.; and Shen, H. T. 2020. Learning Cross-Aligned Latent Embeddings for Zero-Shot Cross-Modal Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11515–11522.

Liu, L.; Shen, F.; Shen, Y.; Liu, X.; and Shao, L. 2017a. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2862–2871.

Liu, Q.; Xie, L.; Wang, H.; and Yuille, A. L. 2019. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, 3662–3671.

Liu, W.; Zhang, Y.-M.; Li, X.; Liu, Z.; Dai, B.; Zhao, T.; and Song, L. 2017b. Deep Hyperspherical Learning. In *NIPS*, 3953–3963.

Lu, P.; Huang, G.; Fu, Y.; Guo, G.; and Lin, H. 2018. Learning large euclidean margin for sketch-based image retrieval. *arXiv preprint arXiv:1812.04275*.

Ramachandran, P.; Parmar, N.; Vaswani, A.; Bello, I.; Levskaya, A.; and Shlens, J. 2019. Stand-Alone Self-Attention in Vision Models. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Saavedra, J. M.; Barrios, J. M.; and Orand, S. 2015. Sketch based Image Retrieval using Learned KeyShapes (LKS). In *Proceedings of the British Machine Vision Conference 2015*, volume 1, 1–11.

Sablayrolles, A.; Douze, M.; Schmid, C.; and Jégou, H. 2019. Spreading vectors for similarity search. In *International Conference on Learning Representations*.

Sangkloy, P.; Burnell, N.; Ham, C.; and Hays, J. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4): 1–12.

Shen, H. T.; Liu, L.; Yang, Y.; Xu, X.; Huang, Z.; Shen, F.; and Hong, R. 2021. Exploiting Subspace Relation in Semantic Labels for Cross-Modal Hashing. *IEEE Transactions on Knowledge and Data Engineering*, 33(10): 3351–3365.

Shen, Y.; Liu, L.; Shen, F.; and Shao, L. 2018. Zero-shot sketch-image hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3598–3607.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347–10357. PMLR.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018a. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5265–5274.

Wang, T.; and Isola, P. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 9929–9939. PMLR.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018b. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Wang, Z.; Wang, H.; Yan, J.; Wu, A.; and Deng, C. 2021. Domain-Smoothing Network for Zero-Shot Sketch-Based Image Retrieval. arXiv:2106.11841.

Xu, J.; and Durrett, G. 2018. Spherical Latent Spaces for Stable Variational Autoencoders. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Xu, X.; Lin, K.; Yang, Y.; Hanjalic, A.; and Shen, H. 2021. Joint Feature Synthesis and Embedding: Adversarial Cross-modal Retrieval Revisited. *IEEE Transactions on Pattern Analysis Machine Intelligence*.

Xu, X.; Lu, H.; Song, J.; Yang, Y.; Shen, H. T.; and Li, X. 2020a. Ternary Adversarial Networks With Self-Supervision for Zero-Shot Cross-Modal Retrieval. *IEEE Transactions on Cybernetics*, 50(6): 2400–2413.

Xu, X.; Wang, T.; Yang, Y.; Zuo, L.; Shen, F.; and Shen, H. T. 2020b. Cross-Modal Attention With Semantic Consistence for Image-Text Matching. *IEEE Trans. Neural Networks Learn. Syst.*, 31(12): 5412–5425.

Yelamarthi, S. K.; Reddy, S. K.; Mishra, A.; and Mittal, A. 2018. A zero-shot framework for sketch based image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 300–317.

Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; Li, M.; and Smola, A. 2020. ResNeSt: Split-Attention Networks. arXiv:2004.08955.

Zhu, J.; Xu, X.; Shen, F.; Lee, R. K.-W.; Wang, Z.; and Shen, H. T. 2020. Ocean: A Dual Learning Approach For Generalized Zero-Shot Sketch-Based Image Retrieval. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.