Contrastive Learning of Semantic Concepts for Open-set Cross-domain Retrieval

Aishwarya Agarwal[1,2], Srikrishna Karanam[2], Balaji Vasan Srinivasan[2], and Biplab Banerjee[1]
[1]Indian Institute of Technology Bombay, Mumbai India
[2]Adobe Research, Bangalore India
{aishagar,skaranam,balsrini}@adobe.com, bbanerjee@iitb.ac.in

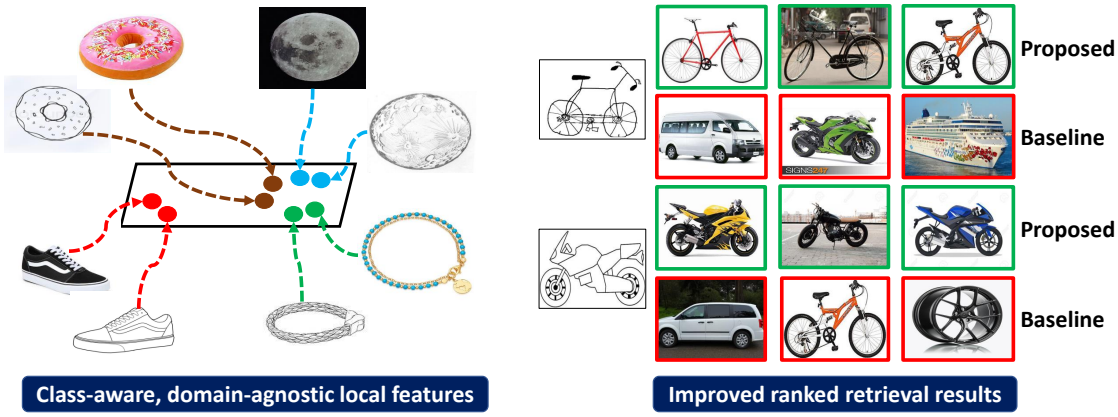

Figure 1: We propose a new method for learning semantic local features for unseen-class/domain image retrieval that shows substantial performance improvements. Left: Our method learns discriminative and domain-agnostic local features (all classes well separated and cross-domain images close) while also being representative ("circle"- themed classes like *donut*, *bracelet*, and *moon* grouped close vs. unrelated *shoe* further apart). Right: Sample retrieval results. Best viewed in color.

Abstract

We consider the problem of image retrieval where query images during testing belong to classes and domains both unseen during training. This requires learning a feature space that has the ability to generalize across both classes and domains together. To this end, we propose semantic contrastive concept network (SCNNet), a new learning framework that helps take a step towards class and domain generalization in a principled fashion. Unlike existing methods that rely on global object representations, SCNNet proposes to learn local feature vectors to facilitate unseen-class generalization. To this end, SCNNet's key innovations include (a) a novel trainable local concept extraction module that learns an orthonormal set of basis vectors, and (b) computes local features for any unseen-class data as a linear combination of the learned basis set. Next, to enable unseen-domain generalization, SCNNet proposes to generate supervisory signals from an adjacent data modality, i.e., natural language, by mining freely available textual label information associated with images. SCNNet derives these signals from our novel trainable semantic ordinal distance constraints that ensure semantic consistency between pairs of images sampled from different domains. Both the proposed modules above enable end-to-end training of the SCNNet, resulting in a model that helps establish state-of-the-art performance on the standard DomainNet, PACS, and Sketchy benchmark datasets with average Prec@200 improvements of 42.6%, 6.5%, and 13.6% respectively over the most recently reported results.

1. Introduction

We consider the problem of image retrieval where query images come from both classes as well as domains *unseen* during training. This is a natural extension of both domain generalization (where test data comes from unseen domains) and open set learning (test data from unseen classes).

Such a problem is motivated from a practical perspective since real-world image retrieval systems will need to work well on *both* domains and classes unseen during training [35, 11, 38].

While there is individually much work in both open-set learning and domain generalization for image retrieval, extending them to our setting where query samples belong to both unseen classes and unseen domains is a non-trivial problem. For instance, while the sketch retrieval line of work in [36, 3, 5, 19, 7, 31] shows strong performance on unseen categories, they require model retraining on new data if the query comes from a new domain (e.g., clipart instead of sketch). Similarly, recent methods in domain generalization [2, 13, 18] assume domain shift to be the only kind of data variation and have no mechanisms to handle test data from unseen classes. This suggests addressing both unseen-class generalization and unseen-domain generalization *jointly* is key in our context.

Paul et al. [22] was the first work to consider this open-set (unseen classes) and cross-domain (unseen domains) retrieval problem and proposed a method based on mixup [39] to simulate train-time data samples for training the model. However, this method has some limitations that motivate our proposed algorithm. First, instead of real images, they use samples from the mixup [39] operation for training, which has been shown in prior work [10, 37] to be very sensitive to the mixing hyperparameter (which is hard to determine, e.g., multiple trial-and-error rounds). Further, this also leads to augmented samples colliding with real samples on the data manifold [10], leading to confusion during training and underperformance during testing. Next, since this method uses a global object representation that relies on per-class image content, they tend to misrepresent out-of-distribution data. On the other hand, there is recent evidence in favor of local features [29, 32, 4, 33, 28] since they tend to capture semantic attributes intuitively common across seen and unseen data (e.g., wheels in a car vs. ambulance). However, learning local features for unseen-class and unseen-domain image retrieval is an open problem. Consequently, we ask: *can we learn semantic local features that can generalize across both domains and classes for retrieval?*

To address the aforementioned issues, we propose `Semantic Contrastive Concept Network` (SCNNet) with two key innovations. First, inspired by the success of detecting local groups of image pixels as "concepts" for visual explainability [12, 9], we ask if (a) such concepts can be trained in an end-to-end fashion without needing additional steps such as segmentation as done in previous work [9], and (b) such concepts can be mapped into vector representations for images of unseen classes unlike existing work [9] that seeks to compute importance scores for explainability of seen-class images. We answer both questions in the affirmative with our novel concept extraction and representation module (CEM). Our key novelty lies in training CEM to produce an orthonormal set of basis vectors, corresponding to a set of local visual concepts, that enables representing an image from a new class as a linear combination of the basis set, thereby computing local features and This way, CEM uses the learned concepts in a principled fashion to compute local features for unseen-class images, addressing a key limitation of our baselines [22, 20].

Next, to handle the domain mismatch between training and testing data for unseen-domain generalization, we exploit information from adjacent data modalities to mine supervision signals, e.g., natural language in the form of textual image labels. Since standard text embedding models are aplenty [21, 24], our key insight is to use this feature space to map the local features above to a domain-agnostic space. Unlike related work [25] that seeks to make image and text features similar, we adopt an ordinal contrastive approach. Given input image triplets during training, our proposed approach, called the semantic ordinal distance module (SOM), seeks to maximize the relative ordinality of distances in the image and text feature spaces. Specifically, from the triplet, we sample both inter- and intra-domain pairs of images belonging to different classes and explicitly constrain relative image feature distances to be consistent with those from the text modality, ensuring local features of same-class-different-domain images are mapped close.

We evaluate SCNNet on the standard DomainNet [23], PACS [14], and Sketchy [26] benchmarks and demonstrate significant gains, including establishing a new state of the art with average Prec@200 improvements of 42.6%, 6.5%, and 13.6% respectively. Our key contributions are:

- We propose a new approach, SCNNet, for generalized unseen-class and unseen-domain image retrieval with two key modules: a concept extraction and represention module (CEM) for handling unseen-class images and a semantic ordinal distance module (SOM) for handling unseen-domain images.
- CEM's key innovation includes a data-driven strategy for learning, end-to-end, an orthogonal set of local visual concepts that can be used to generate local feature representations for any unseen-class image.
- SOM's key innovation includes the use of natural language in a novel contrastive distance learning framework to ensure features of same-class-different-domain images are mapped close in the feature space.

2. Related Work

As discussed in Section 1, our problem of interest, i.e., image retrieval where query images come from both unseen classes and unseen domains, is relatively new, with Paul et

al. [22] being the first method that tackled this problem. Since this new problem has flavors of both domain generalization and zero-shot learning, we briefly review work in these areas below.

Domain generalization. Research efforts under this theme seek to learn models that can generalize to domains unseen during training. Much recent work has used classification as a proxy task (i.e., classifying images from unseen domains) and proposed techniques based on self-supervised learning [2], metric learning [30], adversarial learning [17], meta-learning [15], and episodic training [16] to learn models for domain-invariant representation learning. However, these methods do not have handling for unseen-class images during testing while also being restricted to the classification problem. On the other hand, our method is specifically designed for both a different problem (retrieval instead of classification) while also handling queries from unseen classes and unseen domains during testing.

Zero-shot retrieval. Much recently published research under this theme [27, 36, 5, 7, 3, 6, 34] has used data from only two domains (sketch and image), and generally followed a two-branch architecture design for learning a shared feature space for the two domains. Such architecture designs, however, would not scale in the context of our problem where data can come from any domain (not just that two seen ones). While Liu et al. [19] proposed a single-branch design for processing data from both domains, the principle of using a classification-guided supervision mechanisms where data from all domains are needed for training limits its ability to handle unseen-domain data at test time.

Our closest baseline is the work of Paul et al. [22] that proposed a mixup-based [39] training strategy where new samples were generated using mixup [39]. However, as noted above, this is shown to be prone to issues like manifold collision [10, 37], leading to model underperformance. Our method removes the need for this operation by mining supervision signals from large amounts of textual labels available in the natural language domain. Further, Paul et al. [22] proposed to use global features that tend to rely on overall class-specific image content, leading to brittle representations for out-of-distribution test cases [4, 33, 28]. Instead, our intuition is that local features can help capture semantic visual concepts that are more generalizable, and propose a novel concept extraction and representation module that helps learn these local features.

3. Approach

3.1. Preliminaries

We seek to address the relatively newer problem of unseen-class and unseen-domain image retrieval first proposed in the recent work of Paul et al. [22]. A certain set of *seen* classes and domains are assumed during training. We denote these as $\mathcal{C}_{\text{seen}}$ and $\mathcal{D}_{\text{seen}}$ respectively with the cardinality $N_{\mathcal{D}_{\text{seen}}} \geq 2$ (i.e., labeled data from at least two domains). We represent an image sampled from a class $c \in \mathcal{C}_{\text{seen}}$ and domain $d \in \mathcal{D}_{\text{seen}}$ as $\mathbf{x}^{c,d}$. Now, during testing, we are given a query image $\mathbf{q}$ that belongs to an unseen class and unseen domain. The problem is to match this query image to pre-defined search set of images.

3.2. Semantic Contrastive Concept Network

As noted above, we are interested in image retrieval in the specific scenario where test-time queries come from classes and domains unseen during training. To achieve this, our proposed framework, called Semantic Contrastive Concept Network (SCNNet, Fig. 2), trains a model to learn local visual concepts that be used to generate local feature representations for unseen-class images. Note that this is significantly different from our closest baseline [22] that relies on global object representations that have been shown to have relatively (w.r.t. their local counterparts) poor generalization to zero-shot settings in the context of other problems, e.g., attribute localization [28]. SCNNet achieves this with a novel local concept extraction and representation module (CEM) discussed under Section 3.2.1. While these local features address unseen class generalization, this is insufficient to handle test-time domain shift. To address this issue, SCNNet generates and exploits supervision from natural language as part of the semantic ordinal distance module (SOM), discussed under Section 3.2.2.

As can be noted from Fig. 2, during training, the input to our model is a cross-domain triplet where the anchor $\mathbf{a}^{c_i,d_i}$ and positive $\mathbf{p}^{c_i,d_j}$ belong to the same class (*flower*) but different domains (*quickdraw* and *real*). The negative image $\mathbf{n}^{c_j,d_j}$ is randomly sampled from a different class (*dog*). Note that both $(c_i, c_j) \in \mathcal{C}_{\text{seen}}$ and $(d_i, d_j) \in \mathcal{D}_{\text{seen}}$. Given an image $\mathbf{x}$ in the triplet, our model first computes, with its base image encoder, a convolutional feature map $\mathbf{F}_{\mathbf{x}} \in \mathbb{R}^{z \times w \times h}$, where z is the number of channels in the feature map and (w, h) represents the spatial dimensions of each channel. SCNNet operates on these features $\mathbf{F}_{\mathbf{x}}$ as part of the CEM and SOM modules discussed next.

3.2.1 Concept Extraction Module (CEM) for Learning Class Representations

As noted above, existing work [20, 22] relies on global object representations for handling test query images from unseen classes. However, such global representations tend to be brittle under distribution shift and local features have been shown to generalize better under zero-shot settings in the context of other problems [32, 28]. Furthermore, recent work in model interpretability and explainability [12, 9] shows how one can use the notion of local visual concepts for explaining images under a wide variety of classes and

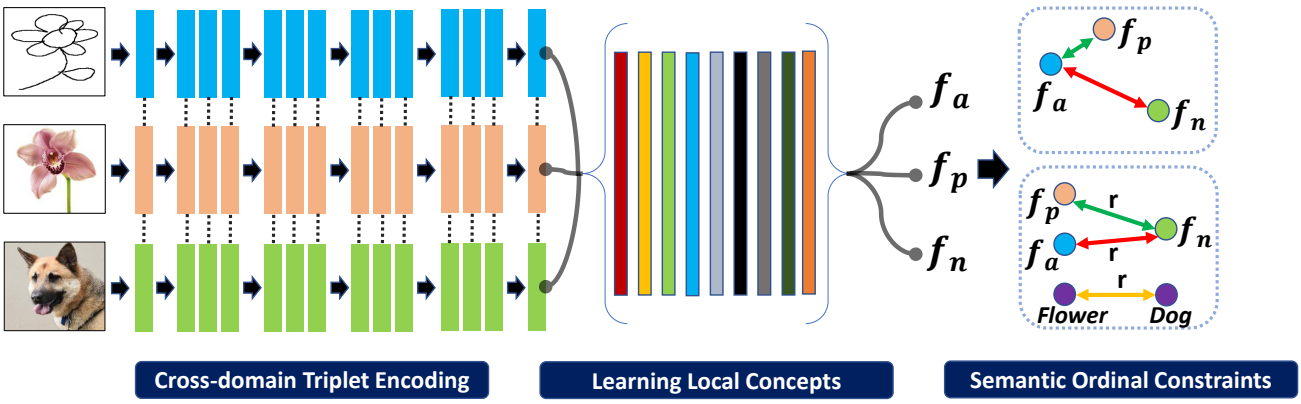

Figure 2: Architecture of our proposed semantic contrastive concept network.

domains, suggesting such concepts can be exploited to compute features that generalize better than global representations. However, these methods [9] do not currently support end-to-end training for learning local features contrastively while also needing additional external models (e.g., pixel-wise segmentation). To this end, SCNNet proposes a trainable concept extraction and representation (CEM) module.

Our conjecture is that there exists a set of local concepts represented by a set of orthonormal basis vectors $\mathcal{V}$. Given such a basis, we are inspired by basic concepts from linear algebra in representing any new image (e.g., an image from an unseen class) as a linear combination of the learned concept vectors present in $\mathcal{V}$. Our intuition here is simple: these local concepts form primitives that are common across classes, e.g., a concept vector corresponding to a wheel will be used to build the feature representation for multiple classes such as *car, truck, motorbike* etc. and that these repeating concepts across semantically similar classes help determine the local features for unseen-class images.

Given the convolutional feature maps $\mathbf{F_x} \in \mathbb{R}^{z\times w\times h}$ computed from the base encoder for each image $\mathbf{x} \in \{\mathbf{a}^{c_i,d_i}, \mathbf{p}^{c_i,d_j}, \mathbf{n}^{c_j,d_j}\}$, we first reshape it to obtain a two-dimensional matrix representation as $\mathbf{F_x} \in \mathbb{R}^{z\times n}$ where $n = w \times h$. This enables running a singular value decomposition (SVD) operation on the matrix that helps give an orthonormal set of basis vectors sorted according to the inferred singular values. Formally, the SVD operation decomposes the $\mathbf{F_x}$ matrix as $\mathbf{F_x} = \mathbf{USV}^T$. We only retain the top-k entries in these matrices to give the feature map $\mathbf{F}^k_\mathbf{x} = \mathbf{U}_{z\times k}\mathbf{S}_{k\times k}\mathbf{V}^T_{k\times n}$. The goal of CEM is to learn a set $\mathcal{V} = \{\mathbf{v}_j \in \mathbb{R}^z, j = 1, 2, 3, \ldots, C\}$ of C concepts that can be used to represent any sample in the $z-$dimensional space spanned by the concept vectors $\{\mathbf{v}_j\}$. During training, this is achieved by "assigning" entries in the feature matrix to each concept $\mathbf{v}_j$ with appropriate weights, and during inference, an unseen-class image's feature matrix is input to the learned set $\mathcal{V}$ to automatically give a vector corresponding to a linear combination operation over the concepts.

To determine the assignment of entries in $\mathbf{F}^k_\mathbf{x}$ to $\mathcal{V}$, we follow standard dictionary assignment practices [8, 40]. Specifically, given each column $\mathbf{f}_i \in \mathbf{F}^k_\mathbf{x}$, we compute a residual vector $\mathbf{r}_{ij} = \mathbf{f}_i - \mathbf{v}_j$ for all values of i and j. Given this, the assignment weight a_{ij} associated with the concept vector $\mathbf{v}_j$ is calculated as:

$$a_{ij} = \frac{exp(-b_j||\mathbf{r}_{ij}||^2)}{\sum_{j=1}^{K} exp(-b_j||\mathbf{r}_{ij}||^2)} \quad (1)$$

where b_j is a learnable parameter corresponding to the concept vector $\mathbf{v}_j$. These parameters are learned during the optimization of the overall loss function discussed in Section 3.2.3. Given these assignment weights, the final feature vector for the image $\mathbf{x}$ is determined as: $\mathbf{f_x} = \sum_{i=1}^{n}\sum_{j=1}^{C} a_{ij}r_{ij}$, where $n = w \times h$ and C is the number of learned concepts. This process gives feature vectors $\mathbf{f_a}$, $\mathbf{f_p}$, and $\mathbf{f_n}$ for the anchor, positive, and negative images in the current training triplet. We apply a triplet loss on these feature vectors to help learn a discriminative feature space for retrieval. In particular, the loss function is:

$$\mathcal{L}_{\text{CEM}} = \frac{1}{B}\sum_{i=1}^{B} \max(0, ||\mathbf{f}^i_\mathbf{a} - \mathbf{f}^i_\mathbf{p}||^2 - ||\mathbf{f}^i_\mathbf{a} - \mathbf{f}^i_\mathbf{n}||^2 + m) \quad (2)$$

where B is the number of triplets sampled in the current training batch and m is a margin parameter.

3.2.2 Semantic Ordinal Module (SOM) for Learning Domain Representations

The local features $\mathbf{f_x}$ from CEM above help handle class shift during testing but they are insufficient when there is an additional domain gap. While existing work [22] used

the mixup operation [39] to learn domain-agnostic features, this requires careful finetuning, via multiple rounds of trial and error, of the mixup parameter while also being prone to manifold intrusion where mixed-up samples collide with real samples [10]. In such a case, a model trained with classification-type losses of Paul et al. [22] for mixture prediction results in model confusion and underfitting.

To address the aforementioned issues, we are motivated by the availability of large amounts of labeled data in adjacent data modalities. For instance, most labeled classification datasets have text labels associated with every image. To mine supervision signals from this vast trove of data, we propose an ordinal contrastive training strategy as part of the semantic ordinal distance module (SOM). Unlike existing work, e.g., Radford et al. [25], that seek to make image and text features map to the same point, SOM proposes to rely on relative distance constraints to ensure inter-domain consistency. Such an approach is easier to train and generalizes better since we are only imposing relative constraints as opposed to hard equality constraints that tend to be domain specific, leading to out-of-domain underperformance.

Our key insight is that we can sample triplets just like in Section 3.2.1 to calculate relative pairwise cross-domain feature distances and make them consistent with those calculated from the text modality. Concretely, given the training triplet $\{\mathbf{a}^{c_i,d_i}, \mathbf{p}^{c_i,d_j}, \mathbf{n}^{c_j,d_j}\}$, we first obtain their CEM embeddings $\mathbf{f_a}$, $\mathbf{f_p}$, and $\mathbf{f_n}$. Next, given the textual class labels c_i and c_j, we use an off-the-shelf word embedding model to obtain the respective semantic features $\mathbf{w}_i$ and $\mathbf{w}_j$. Now, if we sample different-class pairs from this triplet, i.e., $(\mathbf{a}^{c_i,d_i}, \mathbf{n}^{c_j,d_j})$ and $(\mathbf{p}^{c_i,d_j}, \mathbf{n}^{c_j,d_j})$, their distance in the image feature space must be similar to the semantic distance between $\mathbf{w}_i$ and $\mathbf{w}_j$. Consequently, since SOM explicitly enforces this during training with our proposed SOM loss (see below), the model will have learned to map features of same-class-different-domain features close. This is because the relative semantic distance in the case of a same class pair would in principle be zero and the model is trained to be consistent with this distance. To achieve this, our proposed training objective is:

$$\begin{aligned}\mathcal{L}_{\text{SOM}} = \|\mathcal{D}(\mathbf{f_a}, \mathbf{f_n}) - \mathcal{D}(\mathbf{w}_i, \mathbf{w}_j)\|^2 + \\ \|\mathcal{D}(\mathbf{f_p}, \mathbf{f_n}) - \mathcal{D}(\mathbf{w}_i, \mathbf{w}_j)\|^2\end{aligned} \tag{3}$$

where $\mathcal{D}(\mathbf{p}, \mathbf{q})$ is the Euclidean distance metric.

3.2.3 Overall Training Objective

We optimize the parameters of both the encoder and the concept extraction module with following overall loss:

$$\mathcal{L}_{\text{ALL}} = \lambda \mathcal{L}_{\text{CEM}} + (1 - \lambda)\mathcal{L}_{\text{SOM}} \tag{4}$$

where λ denotes the loss weights.

4. Experiments and Results

4.1. Datasets and Implementation Details

We use three standard benchmark datasets in our experiments. First, we train models and evaluate them on the DomainNet [23] benchmark dataset. Images in DomainNet span 345 classes and add up to $\sim$ 0.6 million. These images also span six different domains- *Clip-art*, *Sketch*, *Real*, *Quickdraw*, *Infograph*, and *Painting*. To ensure experimental and evaluation consistency, we follow Paul et al. [22] and split images in the DomainNet dataset into three disjoint train, validation, and testing sets comprising 245, 55 and 45 classes respectively. For test-time retrieval, we construct the search set with images from the *Real* domain. In the evaluation tables below, we refer to this as *Unseen-class* search set. As done in Paul et al. [22], we also consider the scenario where images in the search set can come from both seen and unseen classes. We refer to this evaluation as *Seen+Unseen-class* search set.

Next, we also evaluate our trained models on the PACS [14] and Sketchy Extended [26] datasets, again following the protocol in Paul *et al*. [22]. While PACS consists of $\sim$10,000 images from 7 classes across 4 domains- *Photo*, *Art Painting*, *Cartoon*, and *Sketch*, the Sketchy dataset comprises $\sim$75,000 sketches and $\sim$75,000 real images. For PACS, we construct the search set with images from the *Photo* domain and perform retrieval with query images from the *Art Painting*, *Cartoon*, and *Sketch* domains, whereas for Sketchy, we follow the splits provided in prior work [36, 3]. Please see supplementary material for full implementation details as well as additional results.

4.2. Evaluation Results

We first present and discuss qualitative results. For a query from an unseen class and an unseen domain, we retrieve images from the search set and rank them based on similarity to the query image. We collect the top-10 ranked images and visualize them in Fig. 3. Here, we see three examples, one each for images from the *tooth*, *rainbow*, and *ladder* classes. For each example, we show two rows of ranked retrieval images. The first row corresponds to results with our proposed method and the second row corresponds to the baseline SnMpNet method [22]. One can note from Fig. 3 that there are more images retrieved with our method that show up at lower ranks (e.g., ranks 1, 2, 3) as desired.

Next, we present quantitative performance on all the benchmark datasets and compare to both the state-of-the-art method of Paul et al. [22] and the baselines therein. We show the numbers in Table 1 for both scenarios: *Unseen-class* and *Seen+Unseen-class* search sets. In each case, we show both mAP@200 and Prec@200 numbers. The first two columns in the table show the set of seen domains $\mathcal{D}_{\text{seen}}$ that constitute the training images and the unseen query

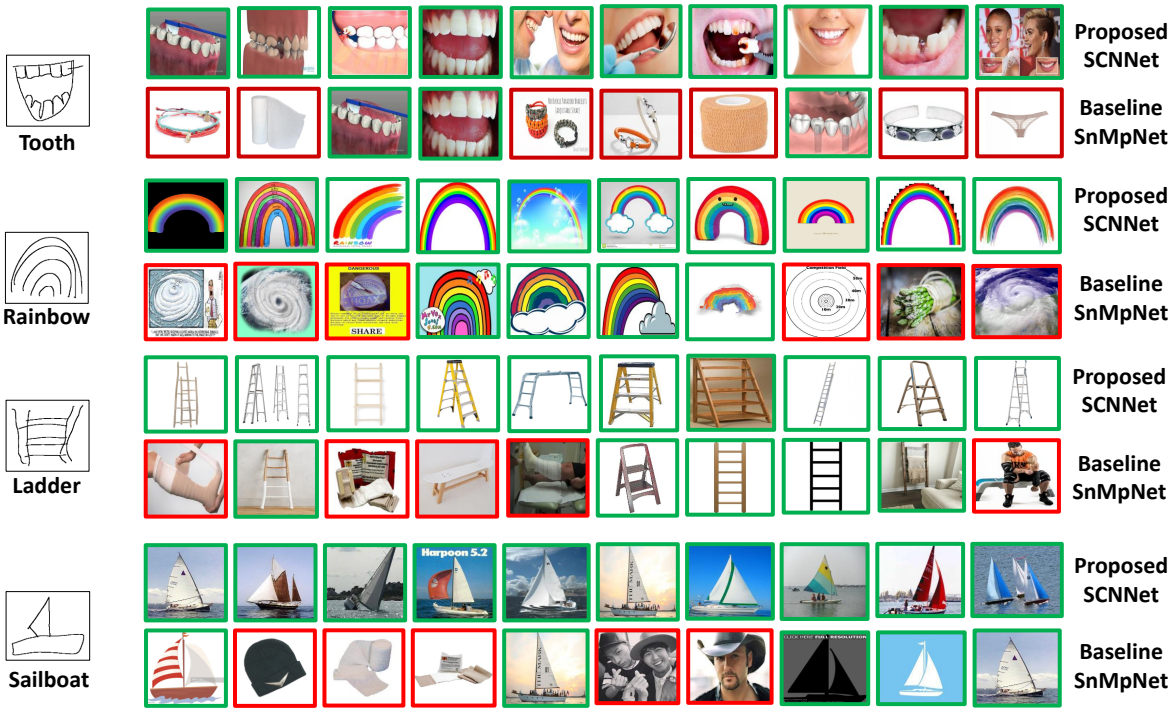

Figure 3: Illustrative examples for improved visual retrieval results with the proposed SCNNet vs. the baseline SnMpNet [22]. We show query images in the left column and the corresponding retrieval results on the right. Best viewed in color.

Training Domains	Query Domain	Method	*Unseen*-class Search		*Seen*+*Unseen*-class Search	
			mAP@200	Prec@200	mAP@200	Prec@200
Real, Quickdraw, Infograph Painting, Clip-art	Sketch	EISNet-retrieval [30, 22]	0.2611	0.2061	0.2286	0.1805
		CuMix-retrieval [20, 22]	0.2736	0.2168	0.2428	0.1935
		SnMpNet [22]	0.3007	0.2432	0.2624	0.2134
		SCNNet	**0.4075**	**0.4120**	**0.3422**	**0.3534**
Real, Sketch, Infograph Painting, Clip-art	Quickdraw	EISNet-retrieval [30, 22]	0.1273	0.1016	0.1101	0.0870
		CuMix-retrieval [20, 22]	0.1304	0.1006	0.1118	0.0852
		SnMpNet [22]	0.1736	0.1284	0.1512	0.1111
		SCNNet	**0.1998**	**0.1580**	**0.1698**	**0.1411**
Real, Sketch, Infograph Quickdraw, Clip-art	Painting	EISNet-retrieval [30, 22]	0.3599	0.2913	0.3280	0.2653
		CuMix-retrieval [20, 22]	0.3710	0.3001	0.3400	0.2751
		SnMpNet [22]	0.4031	0.3332	0.3635	0.3019
		SCNNet	**0.4242**	**0.4409**	**0.3731**	**0.3964**
Real, Sketch, Painting Quickdraw, Clip-art	Infograph	EISNet-retrieval [30, 22]	0.1878	0.1512	0.1658	0.1323
		CuMix-retrieval [20, 22]	0.1931	0.1543	0.1711	0.1361
		SnMpNet [22]	0.2079	0.1717	0.1800	0.1496
		SCNNet	**0.2737**	**0.2476**	**0.2369**	**0.1983**
Real, Sketch, Painting Quickdraw, Infograph	Clip-art	EISNet-retrieval [30, 22]	0.3585	0.2792	0.3251	0.2496
		CuMix-retrieval [20, 22]	0.3764	0.2911	0.3428	0.2627
		SnMpNet [22]	0.4198	0.3323	0.3765	0.2959
		SCNNet	**0.4843**	**0.4664**	**0.4322**	**0.4016**
Average		EISNet-retrieval [30, 22]	0.2589	0.2059	0.2315	0.1829
		CuMix-retrieval [20, 22]	0.2689	0.2126	0.2417	0.1905
		SnMpNet [22]	0.3010	0.2418	0.2667	0.2144
		SCNNet	**0.3579**	**0.3449**	**0.3108**	**0.2981**

Table 1: Evaluation results on the DomainNet benchmark dataset [23].

Query Domain	mAP@200		Prec@200	
	SnMpNet [22]	SCNNet	SnMpNet [22]	SCNNet
Art Painting	0.7403	**0.7615**	0.6507	**0.7353**
Cartoon	0.7167	**0.7410**	0.6462	**0.6691**
Sketch	0.5925	**0.6045**	0.6225	**0.6395**
Average	0.6832	**0.7023**	0.6398	**0.6813**

Table 2: Evaluation results on the PACS dataset [14].

Method	mAP@200	Prec@200
Doodle-SingleNet [3, 22]	0.3980	0.3508
SAKE [19]	0.5484	0.4880
SnMpNet [22]	0.5781	0.5155
SCNNet	**0.6122**	**0.5854**

Table 3: Evaluation results on the Sketchy dataset [26].

domain. Apart from the SnMpNet baseline [22], we also compare to results obtained with the encoders of EISNet-retrieval [30] and CuMix-retrieval [20]. As can be noted from Table 1, our proposed SCNNet results in substantial performance gains; in particular, we establish a new state of the art for this problem on DomainNet [23] with average mAP@200 and Prec@200 values of 0.3579 and 0.3449 respectively (for unseen class search). These numbers represent respective relative performance improvements of 18.9% and 42.6% over SnMpNet. Furthermore, these results represent substantial gains over the CuMix-retrieval method [20, 22] (18.9% and 62.6%). Both these baselines use the mixup [39] technique to generate training samples, helping validate the efficacy of SCNNet's local features.

In Table 2, we report results on PACS [14] where images from the *photos* domain constitute the search set and the other three domains are used in leave-one-out fashion to test. One can note that SCNNet outperforms SnMpNet with average mAP@200 and Prec@200 improvements of 2.8% and 6.5% respectively. Finally, in Table 3, we report results on Sketchy Extended [26] where as above, SCNNet substantially outperforms Paul et al. [22] with mAP@200 and Prec@200 improvements of 5.9% and 13.6%.

4.3. Ablation Study, Insights, and Discussion

We next extensively analyze the contribution of the various components in the proposed SCNNet architecture towards the final perfomance, and provide associated insights and discussion. In Table 4, we provide results of an ablation study to understand the impact of each of the proposed modules. First, as expected, with only the base encoder not supported by either CEM or SOM, the performance is low. Next, on training with CEM (Equation 2), the Prec@200 number jumps to 0.1369 (from baseline 0.0924). It is important to put this in the context of the corresponding numbers for CuMix [20] and SnMpNet [22] from Table 1 (they are 0.1006 and 0.1284 for *Quickdraw* query). Note that the Prec@200 value of 0.1284 for the SnMpNet [22] method is with their full model (i.e., with handling for both unseen classes *and* domains), whereas the 0.1369 Prec@200 value for our method here is *only* with the CEM module (that handles only unseen-class images, not the unseen domain ones). These results clearly show the local features learned with our CEM outperform the mixup-based global features learned in these methods [20, 22].

Network Variant	Quickdraw	Sketch
	Prec@200	Prec@200
Base Encoder	0.0924	0.2621
Base Encoder + CEM	0.1369	0.3652
Base Encoder + SOM	0.1380	0.3704
SCNNet Full Model	**0.1580**	**0.4120**

Table 4: Evaluating the impact of CEM and SOM.

Network Variant	Quickdraw	Sketch
	Prec@200	Prec@200
Radford et al. [25]	0.0590	0.3040
SOM	**0.1380**	**0.3704**

Table 5: Proposed SOM module vs. baseline CLIP [25].

Embedding	mAP@200	Prec@200
Word2Vec [21]	**0.1998**	**0.1580**
fastText [1]	0.1923	0.1547
GloVe [24]	0.1906	0.1533

Table 6: Comparing various semantic knowledge sources.

Next, adding the SOM module also improves the baseline encoder's performance from a Prec@200 of 0.0924 to 0.1380. Note that whereas SnMpNet [22] used the mixup strategy to learn domain-agnostic features, our SOM outperforms this method (Prec@200 of 0.1284 as noted above), helping show one need not use mixup (which leads to issues like manifold collision as discussed above) to handle unseen-domain queries. Furthermore, in Table 5, we compare these results to that of Radford et al. [25]. As discussed in Section 3.2.2, the use of relative ordinal constraints where SOM proposes to ensure pairwise distance consistency is easier to train models as opposed to related work [25] that seeks to make the image and text features map to the same point. From Table 5, this gives substantially better results on both *Quickdraw* and *Sketch*. Finally, from Table 4, the full model, with CEM and SOM, outperforms both the components, helping demonstrate their complementary impact.

In Table 6, we compare various sources of semantic knowledge (see Equation 3). We use queries from the *Quickdraw* domain and follow the same protocol as above. One can note Word2Vec [21] (which we use for all results reported above) outperforms fastText [1] and GloVe [24].

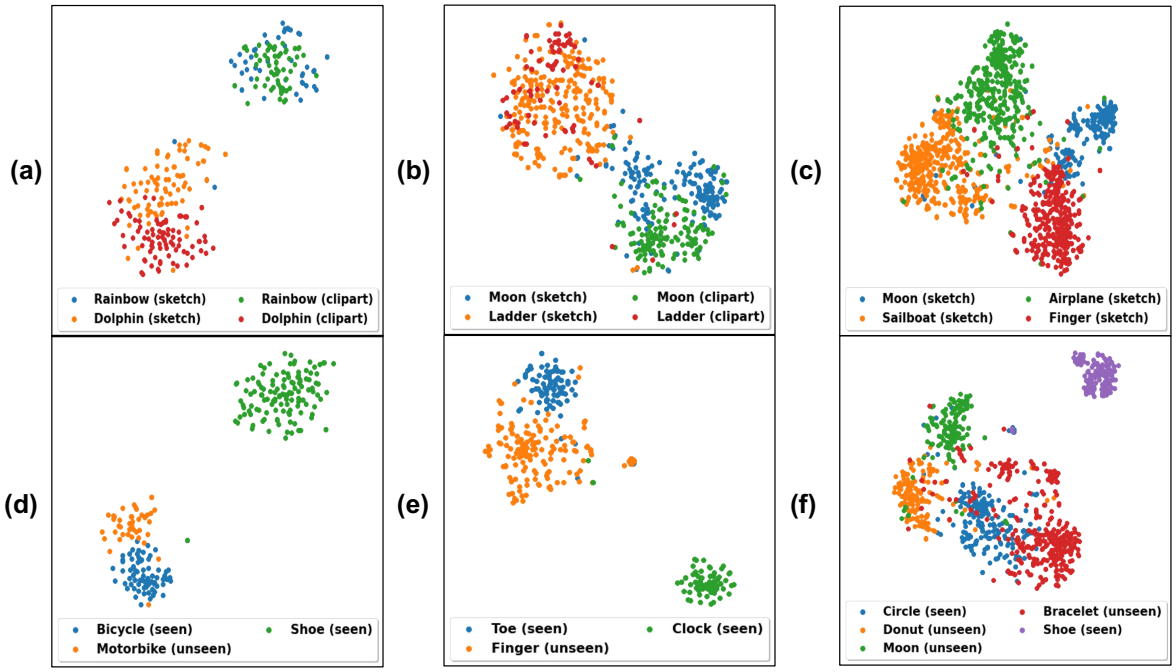

Figure 4: t-SNE plots to demonstrate CEM and SOM's impact. Best viewed in color.

In Fig. 4, we provide additional results to further substantiate the impact of the CEM and SOM components. In Fig. 4(a), we take images from two unseen classes (*Rainbow* and *Dolphin*) and unseen domains (*Sketch* and *Clipart*), compute their features from our CEM module, and generate the t-SNE plot. One can note a clear clustering of the cross-domain features from the same class (e.g., *Rainbow* from *Sketch* and *Clipart* are grouped together and *Dolphin* from *Sketch* and *Clipart* are grouped together and separate from the first group). Fig. 4(b) shows another example along the same lines (*Moon* and *Ladder* classes from *Sketch* and *Clipart* domains). Both these pictures clearly show the discriminative power of the representations- features of different-class images cluster separately and features of same-class-different-domain images cluster together. Fig. 4(c) further demonstrates this for four unseen classes, where we see clear clustering.

To show how CEM learns transferable local features, in Fig. 4(d), we take images from two classes that share some semantic attributes (*Bicycle* and *Motorbike* share notions of wheel) and one very different class (*Shoe*). Further, here, *Bicycle* is a *seen* class and *Motorbike* is an *unseen* class. As expected, from Fig. 4(d), there is clear discrimination between *Shoe* and the other two. The features for *Bicycle* and *Motorbike*, while separated themselves, are closer to each other than *Shoe*. This shows while CEM representations are discriminative, they generate are some concepts that are shared across these classes, helping generate unseen-class representations (*Motorbike* here). Fig. 4(e) shows another similar example. In Fig. 4(f), we take images from four related but different classes (*Circle*, *Donut*, *Moon*, and *Bracelet*) that share some semantic concepts (e.g., the *circle* theme) and a completely different class (*Shoe*). We can see all the *circle*-themed classes are clustered in the same larger region (although each class has its own separate subregion) when compared to *Shoe*, providing further evidence that CEM is indeed learning transferable local features.

5. Summary

We considered the relatively underexplored problem of generalized unseen-class *and* unseen-domain image retrieval and proposed a new framework, Semantic Contrastive Concept Network (SCNNet), comprising two key innovations. First, unlike existing work that used global object representations, we proposed a trainable local concept extraction and representation module that uses the learned concepts to produce local representations for unseen-class images. Next, to help generalize across domains, SCNNet proposed to leverage freely available textual data from the natural language modality to mine supervision signals. Here, our key novetly was to use relative semantic ordinal distance constraints as opposed to mapping image-text features to the same point. We conducted extensive experiments on standard datasets to demonstrate both state-of-the-art performance.

References

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[2] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019.

[3] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2179–2188, 2019.

[4] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32, 2019.

[5] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5089–5098, 2019.

[6] Titir Dutta, Anurag Singh, and Soma Biswas. Adaptive margin diversity regularizer for handling data imbalance in zero-shot sbir. In *European Conference on Computer Vision*, pages 349–364. Springer, 2020.

[7] Titir Dutta, Anurag Singh, and Soma Biswas. Styleguide: Zero-shot sketch-based image retrieval using style-guided image generation. *IEEE Transactions on Multimedia*, 23:2833–2842, 2020.

[8] Jan C van Gemert, Jan-Mark Geusebroek, Cor J Veenman, and Arnold WM Smeulders. Kernel codebooks for scene categorization. In *European conference on computer vision*, pages 696–709. Springer, 2008.

[9] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

[10] Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3714–3722, 2019.

[11] Houdong Hu, Yan Wang, Linjun Yang, Pavel Komlev, Li Huang, Xi Chen, Jiapei Huang, Ye Wu, Meenaz Merchant, and Arun Sacheti. Web-scale responsive visual search at bing. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 359–367, 2018.

[12] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[13] Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A Plummer, Stan Sclaroff, and Kate Saenko. Cds: Cross-domain self-supervised pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9123–9132, 2021.

[14] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[15] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[16] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455, 2019.

[17] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018.

[18] Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8886–8895, 2021.

[19] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan L Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3662–3671, 2019.

[20] Massimiliano Mancini, Zeynep Akata, Elisa Ricci, and Barbara Caputo. Towards recognizing unseen categories in unseen domains. In *European Conference on Computer Vision*, pages 466–483. Springer, 2020.

[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[22] Soumava Paul, Titir Dutta, and Soma Biswas. Universal cross-domain retrieval: Generalizing across classes and domains. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12056–12064, 2021.

[23] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

[24] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[26] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn

bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016.

[27] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3598–3607, 2018.

[28] Chaoqun Wang, Shaobo Min, Xuejin Chen, Xiaoyan Sun, and Houqiang Li. Dual progressive prototype network for generalized zero-shot learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[29] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

[30] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. In *European Conference on Computer Vision*, pages 159–176. Springer, 2020.

[31] Wenjie Wang, Yufeng Shi, Shiming Chen, Qinmu Peng, Feng Zheng, and Xinge You. Norm-guided adaptive visual embedding for zero-shot sketch-based image retrieval. In *IJCAI*, pages 1106–1112, 2021.

[32] Guo-Sen Xie, Li Liu, Xiaobo Jin, Fan Zhu, Zheng Zhang, Jie Qin, Yazhou Yao, and Ling Shao. Attentive region embedding network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9384–9393, 2019.

[33] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. *Advances in Neural Information Processing Systems*, 33:21969–21980, 2020.

[34] Xinxun Xu, Cheng Deng, Muli Yang, and Hao Wang. Progressive domain-independent feature decomposition network for zero-shot sketch-based image retrieval. *arXiv preprint arXiv:2003.09869*, 2020.

[35] Fan Yang, Ajinkya Kale, Yury Bubnov, Leon Stein, Qiaosong Wang, Hadi Kiapour, and Robinson Piramuthu. Visual search at ebay. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2101–2110, 2017.

[36] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 300–317, 2018.

[37] Hao Yu, Huanyu Wang, and Jianxin Wu. Mixup without hesitation. In *International Conference on Image and Graphics*, pages 143–154. Springer, 2021.

[38] Andrew Zhai, Hao-Yu Wu, Eric Tzeng, Dong Huk Park, and Charles Rosenberg. Learning a unified embedding for visual search at pinterest. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2412–2420, 2019.

[39] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[40] Hang Zhang, Jia Xue, and Kristin Dana. Deep ten: Texture encoding network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 708–717, 2017.