Category-Specific Nuance Exploration Network for Fine-Grained Object Retrieval

Shijie Wang[1], **Zhihui Wang**[1,2], **Haojie Li** [1,2*], **Wanli Ouyang**[3]

[1] International School of Information Science and Engineering, Dalian University of Technology, China
[2]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, China
[3]Sense Time Computer Vision Research Group, The University of Sydney, Australia
wangsj@mail.dlut.edu.cn, {zhwang, hjli}@dlut.edu.cn, wanli.ouyang@sydney.edu.au

Abstract

Employing additional prior knowledge to model local features as a final fine-grained object representation has become a trend for fine-grained object retrieval (FGOR). A potential limitation of these methods is that they only focus on common parts across the dataset (e.g., head, body, or even leg) by introducing additional prior knowledge, but the retrieval of a fine-grained object may rely on category-specific nuances that contribute to category prediction. To handle this limitation, we propose an end-to-end Category-specific Nuance Exploration Network (CNENet) that elaborately discovers category-specific nuances that contribute to category prediction, and semantically aligns these nuances grouped by subcategory without any additional prior knowledge, to directly emphasize the discrepancy among subcategories. Specifically, we design a Nuance Modelling Module that adaptively predicts a group of category-specific response (CARE) maps via implicitly digging into category-specific nuances, specifying the locations and scales for category-specific nuances. Upon this, two nuance regularizations are proposed: 1) semantic discrete loss that forces each CARE map to attend to different spatial regions to capture diverse nuances; 2) semantic alignment loss that constructs a consistent semantic correspondence for each CARE map of the same order with the same subcategory via guaranteeing each instance and its transformed counterpart to be spatially aligned. Moreover, we propose a Nuance Expansion Module, which exploits context appearance information of discovered nuances and refines the prediction of current nuance by its similar neighbors, leading to further improvement on nuance consistency and completeness. Extensive experiments validate that our CNENet consistently yields the best performance under the same settings against most competitive approaches on CUB Birds, Stanford Cars, and FGVC Aircraft datasets.

Introduction

Fine-grained object retrieval (FGOR) aims at retrieving images belonging to various subcategories of a certain meta-category and returning images with the same subcategory as the query image. It is a more challenging problem than general image retrieval due to the inherently subtle inter-class object variances among subcategories. As a result, the key to FGOR lies in picking out nuances buried in the local regions to address the aforementioned challenge of FGOR.

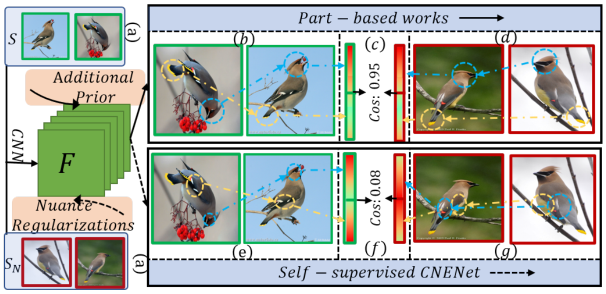

Figure 1: Motivation of CNENet. (a) are visually similar images (S, S_N) from different subcategories as inputs. The yellow and blue circles in (b)(d)(e)(g) denote the selected nuances, and the values in (c)(f) represent the cosine similarity (Cos) between (S, S_N). We can see that although part-based works discover common parts of head and tail across subcategories by introducing additional prior knowledge (*i.e.*, key points, etc.), the similarity (c) between the extracted features from selected parts with different subcategories is large. Our CNENet discards additional prior knowledge and designs the nuance regularizations to discover category-specific nuances which do not have to be common parts, thus emphasizing the discrepancy among subcategories.

Recently, quite a few approaches (Zheng et al. 2018; Shen et al. 2017; Moskvyak et al. 2021) have been proposed for exploring nuances, which primarily brings together representation learning, and fine-grained object auxiliary information into a framework to consider the nuances of a fine-grained object. However, these works require extra prior knowledge (i.e., object location, key points, or object parsing information) for discovering and aligning common parts across all subcategories, while neglecting the fact that these common parts are not always discriminative, and accordingly degrades the retrieval performance. For example, visually similar birds can be retrieved using their category-specific nuances that contribute to category prediction in Fig. 1, but previous part-based works select the common parts (e.g., head and tail) across all subcategories by the guidance of additional prior knowledge, making the selected

[*] Corresponding author.

parts not always inclusive of category-specific nuances. Therefore, we argue that additional prior knowledge can only provide the location of common parts across the dataset but cannot explicitly point out the category-specific nuances among subcategories. When additional prior knowledge is useless, how to effectively extract category-specific nuances and how to semantically align these nuances grouped by category are worthy of investigation for FGOR.

To this end, we propose an end-to-end Category-specific Nuance Exploration Network (CNENet) to elaborately discover category-specific nuances and semantically align these nuances of the same subcategory in order by introducing additional nuance regularizations, which directly emphasizes the discrepancy among subcategories. The CNENet consists of Nuance Modelling Module (NMM) and Nuance Expansion Module (NEM). NMM predicts a set of category-specific response (CARE) maps by implicitly digging into nuances relevant to categories under two nuance regularizations, which specifies the location and scale for category-specific nuances. Concretely, the nuance regularizations are: 1) semantic discrete loss that forces each CARE map to attend to different spatial regions to discover diverse nuances; 2) semantic alignment loss that constructs a consistent semantic correspondence for each CARE map of the same order with the same subcategory via guaranteeing each instance and its transformed counterpart to be spatially aligned. The multiple nuances generated by NMM are expected to be spatially discrete as much as possible to achieve semantic diversity of category-specific nuances. However, some vital nuances may cover the entire object or overlap with the others, resulting in some nuances being shrunk. Therefore, NEM exploits context appearance information of discovered nuances and refines the prediction of current nuance by its similar neighbors, leading to further improvement on nuance consistency and completeness. Finally, these two modules without any pairwise metric losses are cascaded and jointly optimized, to learn the category-specific nuances, which have the property of benefiting FGOR performance.

Main contributions of this paper can be summarized:

- To the best of our knowledge, we are the first to dig into and align category-specific nuances grouped by category rather than focus on common parts across the dataset in FGOR.
- We design two nuance regularizations: semantic discrete loss to explore diverse category-specific nuances and semantic alignment loss to semantically align nuances of the same subcategory, thus achieving category-specific nuance exploration in a self-supervised manner.
- We evaluate the proposed method on three datasets (CUB Birds, Stanford Cars, and FGVC Aircraft), and the results demonstrate that our CNENet achieves the state-of-the-art.

Related Work

Fine-grained Object Retrieval: Existing FGOR methods can be roughly divided into three groups. The first group, ***metric-based schemes***, is learning an embedding space where similar examples are attracted, and dissimilar examples are repelled (Teh et al. 2020; Wang et al. 2019a; Boudiaf et al. 2020). PNCA++ (Teh et al. 2020) proposes a proxy-based deep metric learning (DML) solution to embed image-level features and thus represent class distribution. The shortcoming of metric-based schemes is that they focus on the optimization of image-level features which contain many noisy and non-discriminative information. Therefore, the second group, ***object-based schemes***, focuses on localizing the objects from images via exploring the activation of features (Wei et al. 2017; Zheng et al. 2018). SCDA (Wei et al. 2017) only localizes the objects while discards the noisy background for extracting informative descriptors for FGOR. CRL (Zheng et al. 2018) designs an attractive object feature extraction strategy to facilitate the retrieval task. Instead of localizing object-level features, the third group, ***part-based schemes*** tends to dig into common parts across the dataset via the guidance of additional prior knowledge (Zheng et al. 2018; Shen et al. 2017; Moskvyak et al. 2021). LFE (Shen et al. 2017) selects the *specific filters* to localize the semantically coherent parts, which achieves the goal of encoding common regions. KAE-Net (Moskvyak et al. 2021) learns features corresponding to each *keypoint position* to construct a representation. However, these approaches are difficult to guarantee the learnt features are discriminative enough. Different from these works, we propose CNENet to dig into category-specific nuances that contribute to category prediction, and thus explicitly emphasize discrepancies among subcategories.

Nuance exploration: Recently, nuance exploration is mainly applied to fine-grained image recognition, and has made great progress (Ding et al. 2019; Yang et al. 2018; Zhang et al. 2016; Zheng et al. 2019a; Wang et al. 2020c; Zhou et al. 2020; Wang et al. 2021, 2020a, 2019b). S3Ns (Ding et al. 2019) produces sparse attention to localize object and discriminative nuances by collecting local maximums of class response maps. ACNet (Ji et al. 2020) introduces the attention transformer to facilitate coarse-to-fine hierarchical feature learning to grab discriminative nuances. CGP (Wang et al. 2020b) establishes correlation between regions by graph propagation to discover the more discriminative nuance groups. The recognition task maps the learned nuances to the category space while not considering other samples, and thus is not sensitive to the order of the learned nuances. In contrast, the category-specific nuances in retrieval require matching with nuances from other samples in the dataset and thus are sequentially sensitive. Upon this, we design two nuance regularizations to adaptively discover and semantically align category-specific nuances guided by category to address the problem of sequence sensitivity. To our best knowledge, this is the first work to explore category-specific nuances in a self-supervised manner for FGOR.

Proposed Method

We aim to explore the nuances that contribute to category prediction for emphasizing discrepancies among subcategories in FGOR. To this end, we propose the Category-specific Nuance Exploration Network (CNENet). It introduces two new components: the nuance modeling module

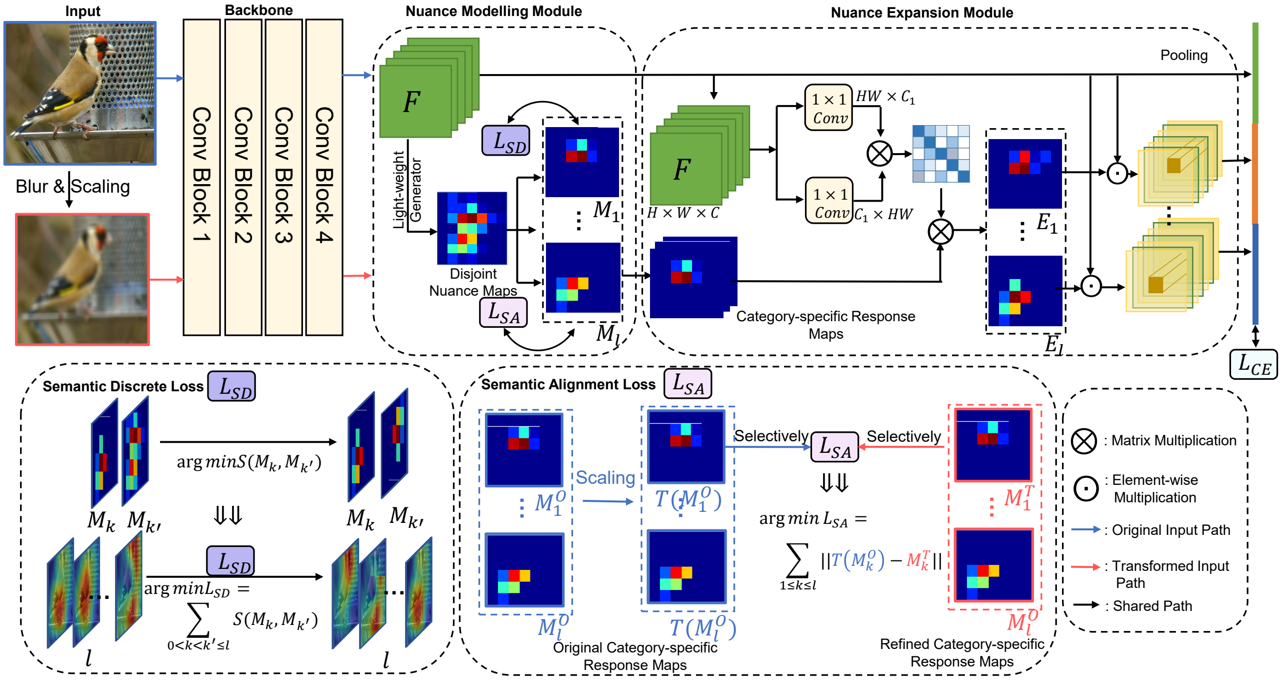

Figure 2: Framework of CNENet. The Nuance Modelling Module receives feature maps from the backbone network to discover and align category-specific response (CARE) maps with two nuance regularizations, which are the Semantic Discrete Loss $\mathcal{L}_{SD}$ to force CARE maps to capture diverse nuances and Semantic Alignment Loss $\mathcal{L}_{SA}$ to semantically align CARE maps guided by category. Subsequently, the Nuance Expansion Module exploits context appearance information of discovered nuances and refines the prediction of current nuance by its similar neighbors, which can restores those missing details due to the constraint of the semantic discrete loss. Finally, we extract the aligned category-specific nuances and concatenate them as retrieval features.

(NMM) to discover the category-specific nuances and align them guided by category, and the nuance expansion module (NEM) to refine the prediction of current nuance by its similar neighbors. Our framework is illustrated in Fig. 2.

Nuance Modelling Module

Understanding discriminative semantics among subcategories is a prerequisite for retrieving visually similar images. A typical approach is to introduce the additional prior knowledge, i.e., bounding boxes or key points, to capture common parts across the dataset. However, these common parts cannot explicitly point out discrepancies among subcategories, and thus are useless. To handle this issue, we propose a Nuance Modelling Module (NMM) to help the network simultaneously discover the category-specific nuances and align them guided by category in a self-supervised manner.

For an input image X, we denote its feature maps $F \in \mathbb{R}^{c\times h\times w}$ extracted by the convolutional blocks as the input of NMM, where c, h, w are the dimension, height, and width of the feature maps. To obtain category-specific nuances, NMM aims to discover and align the nuances of fine-grained objects with the same subcategory. Specifically, NMM consists of three sub-modules: category-specific response generation, semantic discrete loss, and semantic alignment loss. They are explained in detail as below.

Category-specific response generation. NMM first splits the feature maps F into l category-specific response (CARE) maps $M = [M_1, M_2, ..., M_l] \in \mathbb{R}^{h\times w\times l}$. Concretely, these maps are generated by a light-weight generator $G(\cdot)$ followed by a normalization operation as follows:

$$\hat{M} = ReLU(G(F)), \tag{1}$$

where $ReLU(\cdot)$ denotes the rectified linear unit (ReLU) activation function, and $G(\cdot)$ is a convolutional operation with kernel size $C \times 1 \times 1 \times l$. Then $\hat{M}$ is passed through a min-max layer to normalize the nuanced response coefficients M, which forces M into $[0, 1]$:

$$M = \frac{\hat{M} - min(\hat{M})}{max(\hat{M}) - min(\hat{M}) + \varepsilon}, \tag{2}$$

where ε is a protection item to avoid dividing-by-zero, and is set to 10^{-5} in our experiments.

Note that by the operation of lightweight generation, the only goal of learning CARE maps is to capture and represent the scales and locations of category-specific nuances between input images and corresponding class information. Since the class information can implicitly determine the relevant and irrelevant features in F, optimal features would capture the relevant features while compressing F by suppressing the irrelevant visual patterns which do not contribute to the prediction of categories. Considering the corresponding relationship between compressed F and CARE maps M, F produces category-specific M, which thus indicates the spatial locations of category-specific nuances.

Semantic discrete loss. The category-specific response generation tends to activate category-specific nuances by utilizing the correlation between features and category information, but does not consider the fact that CARE maps should cover diverse nuances of a fine-grained object. To ensure the CARE maps can capture diverse nuances, we thus design the semantic discrete loss $\mathcal{L}_{SD}$ as a nuance regularization to force each CARE map to attend to a different spatial region.

Specifically, we introduce $\mathcal{L}_{SD}$ to make the l CARE maps in M as discrepant with each other much as possible. Therefore, this is equivalent to minimize the similarity among CARE maps, as

$$argmin \quad \mathcal{L}_{SD} = \frac{2}{l(l-1)} \sum_{1 \leqslant k < k' \leqslant l} S(M_k, M_{k'}), \qquad (3)$$

where $M_k, M_{k'}$ denote the k-th and $k^{'}$-th CARE maps respectively. $S(M_k, M_{k\prime}) = \frac{M_k \cdot M_{k'}}{||M_k|| \cdot ||M_{k'}||}$ is the cosine similarity between M_k and $M_{k'}$.

Once Eq. 3 is optimized, the CARE maps are obviously discrepant with each other. By this means that if a CARE map discovers one nuanced region, the other maps will be forced to activate other spatially exclusive nuances.

Semantic alignment loss. The semantic discrete loss only aims to force learned CARE maps to be discrepant in space, capturing diverse nuances for fine-grained objects. Nonetheless, it can not guarantee that the activated CARE maps with the same subcategory are semantically corresponding in order, which leads to the problem of feature incoherency for images with the same subcategory and decreases the retrieval performance accordingly.

Inspired by the data augmentation stage of fully supervised object detection or semantic segmentation (Wu et al. 2020; Li et al. 2020), the spatial annotations should be applied with the same affine transformation as input images. It introduces an implicit equivariant regularization for the network to enforce spatial alignment between transformed images and corresponding annotations. Therefore, we design a semantic alignment loss as an implicit equivariant regularization to imitate the contribution of full supervision, making selected nuances semantically correspond to the ones from other samples in the same subcategory. Concretely, we expand the network into a shared-weight siamese structure to integrate the semantic alignment loss $\mathcal{L}_{SA}$ into the original network, thus being able to semantically align category-specific nuances guided by category:

$$\mathcal{L}_{SA} = ||T(G(B(I))) - G(B(T(I)))||, \qquad (4)$$

where $G(B(\cdot))$ represents the backbone network $B(\cdot)$ followed by category-specific response generation operation $G(\cdot)$, $T(\cdot)$ is any spatial affine transformation, e.g. rescaling, rotation, flip, and so on. One branch $T(G(B(I)))$ applies the transformation on the CARE map to output $T(M_k^O)$, the other branch $G(B(T(I)))$ warps the input samples by the same affine transformation before the feed-forward of the network to output transformed CARE maps M_k^T. Therefore, according to Eq. 4, regularizing the CARE maps from two branches to guarantee the spatially corresponding can be rewritten as:

$$\mathcal{L}_{SA} = \frac{1}{l} \sum_{i=1}^{l} ||T(M_k^O) - M_k^T||. \qquad (5)$$

Moreover, to further improve the ability of network for semantically aligning nuances selected from images with the same subcategory, we change the data distribution of the input images by utilizing content augmentation manners (e.g., Gaussian blur, saturation adjustment) in addition to the spatial affine transformations. By this means that it can enlarge the distance between original samples and transformed samples, which further narrows the supervision gap between fully and weakly supervised signals. By encouraging spatial correspondence between CARE maps of the same instance but with different affine transformations, the effective category-specific nuance generator is learned to match the discrete but semantically consistent nuances of the same subcategory in order.

Nuance Expansion Module

The nuances generated by NMM are spatially discrete as much as possible to ensure semantic diversity of discovering nuances. However, some vital nuances could cover the entire objects or overlap with other nuances, thus resulting in some nuances being shrunk due to the constraint of semantic discrete loss. To handle this limitation, we propose a Nuance Expansion Module (NEM) to exploit context appearance information of discovering nuances and refine the prediction of current nuance by its similar neighbors.

NEM works as a reinforcement operation by capturing context feature dependency to revise category-specific nuances. Therefore, we refer to the core part of the self-attention mechanism (Wang et al. 2018) with some modifications to achieve the key structure of NEM. NEM consists of two steps: 1) pixel correlation prediction and 2) nuance reassembly. Before taking a look at two steps, let's review the self-attention mechanism.

Revisiting self-attention. Self-attention mechanism (Wang et al. 2018) meets the ideas of most methods using the similarity of pixels to refine the original activation regions. Following the denotation (Wang et al. 2018), the general self-attention mechanism can be integrated into NEM to refine CARE maps M:

$$E_i = \frac{1}{\mathcal{N}(M_i)} \sum_{\forall j} f(M_i, M_j)\eta(M_j) + M_i, \qquad (6)$$

where

$$f(M_i, M_j) = e^{\vartheta(M_i)^T \delta(M_j)}, \qquad (7)$$

and three embedding functions ϑ, δ, η can be implemented by individual 1×1 convolution operations. Here M_i and E_j respectively denote the original and refined CARE maps with the spatial position index i and j, and function $\eta(M_j)$ provides a feature vector of input M_j at each position and all of them are integrated into position i based on the correlation coefficient given by $f(M_i, M_j)$, which calculates the dot-product feature affinity in an embedding space. The output value E is normalized by $\mathcal{N}(M_i) = \sum_{\forall j} f(M_i, M_j)$. However, since CARE maps M are constrained by the semantic discrete loss and are thereby orthogonal to each other, the $f(M_i, M_j)$ equals to 0, further failing to refine M.

Pixel correlation prediction. To handle this problem, we select the feature vectors in high-level features F rather than orthogonal CARE maps M to learn the pixel correlation. More importantly, since features contain more visual clues compared to CARE maps, we can obtain more accurate correlation coefficients. Specifically, the feature projection layer can be implemented by an individual convolution operation as follows:

$$\hat{F} = \vartheta(W_\theta \cdot F + b) \tag{8}$$

where $W_\theta \in \mathbb{R}^{C\times 1\times 1\times C_1}$ and b are the learned weight parameters and bias vector of a convolution layer ϑ, respectively. 1×1 is the size of convolution kernel. $\hat{F}$ denote the new feature maps. Unlike classical self-attention in object detection (Wang et al. 2018), since our network only provides image-level supervision and two nuance regularizations, which is not as accurate as full supervision, we reduce parameters by removing two embedding functions δ, η to avoid overfitting on inaccurate supervision.

Let's take only a single correlation of two positions as an example. The correlation of two positions at p_1 and p_2 in $\hat{F}$ is then defined as

$$f(\hat{F}_{p_1}, \hat{F}_{p_2}) = ReLU(\frac{\hat{F}_{p_1}^T \cdot \hat{F}_{p_2}}{||\hat{F}_{p_1}|| \cdot ||\hat{F}_{p_2}||}). \tag{9}$$

Here we take the inner-product $\cdot$ in normalized feature space to calculate the reassembly correlation coefficient $f(\hat{F}_{p_1}, \hat{F}_{p_2})$ between current pixel $\hat{F}_{p_1}$ and others. Compared to Eq. 7, we use ReLU activation function with $L1$ normalization to mask out irrelevant pixels and generate an correlation map which is smoother in relevant regions.

Nuance reassembly. With the reassembly correlation coefficients $f(\hat{F}_{p_1}, \hat{F}_{p_2})$, Eq. 6 can be rewritten as:

$$E^{P_1} = \frac{1}{\sum_{\forall p_2} f(\hat{F}_{p_1}, \hat{F}_{p_2})} \sum_{\forall p_2} f(\hat{F}_{p_1}, \hat{F}_{p_2}) M^{P_1}, \tag{10}$$

where refined CARE maps $E \in \mathbb{R}^{l\times H\times W}$ are the weighted sum of the original CARE maps M with the normalized $f(\hat{F}_{p_1}, \hat{F}_{p_2})$. Moreover, we remove the residual connection to keep the same activation intensity of the original CARE maps.

With these refined CARE maps, we can split the feature maps F into l nuances as follows:

$$U_k = E_k \odot F, \quad k = 1, 2, ..., l \tag{11}$$

where $\odot$ denotes element-wise multiplication.

Once the feature maps are split into l nuances according to refined CARE maps, the features of the k-th nuances $u_k = g(U_k) \in \mathbb{R}^C$ are extracted by global average pooling $g(\cdot)$. Finally, the output features $f \in \mathbb{R}^{(l+1)\times C}$ for retrieval can be represented by:

$$f_c = [u_1^T, u_2^T, ..., u_l^T, g(F)^T]^T. \tag{12}$$

Method	Recall@1
BL (He et al. 2016)	66.3%
BL + NMM (w/o $\mathcal{L}_{SD}$&$\mathcal{L}_{SA}$)	67.8%(1.5% ↑)
BL + NMM (w/o $\mathcal{L}_{SD}$)	69.2%(2.9% ↑)
BL + NMM (w/o $\mathcal{L}_{SA}$)	66.9%(0.6% ↑)
BL + NMM	71.2%(4.9% ↑)
BL + NMM + Self-attention	70.3%(0.9% ↓)
BL + NMM + NEM	74.5%(3.3% ↑)
BL + NMM + NEM + Triplet loss	73.4%(1.1% ↓)

Table 1: The ablative retrieval results of different variants of our method. We test the models on CUB-200-2011.

Loss function

The full multi-task loss $\mathcal{L}$ can be denoted as below:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{SD} + \mathcal{L}_{SA}, \tag{13}$$

where $\mathcal{L}_{CE}$ represents the classification cross-entropy loss.

Experiments

Experimental Setting

Datasets. CUB-200-2011 (Branson et al. 2014) contains 200 bird subcategories with 11,788 images. We utilize the first 100 classes (5,864 images) in training and the rest (5,924 images) in testing. The spilt in Stanford Cars (Krause et al. 2013) is also similar to CUB, which contains 196 classes with 16,185 images, i.e. with the first 98 classes (8,045 images) for training and the remaining class (8,131 images) for testing. FGVC Aircraft (Maji et al. 2013) is divided into first 50 classes (5,000 images) for training and the rest 50 classes (5,000 images) for testing.

Evaluation protocols. We evaluate the retrieval performance by *Recall@K* with cosine distance, which is average recall scores over all query images in the test set and strictly follows the setting in (Song et al. 2016). Specifically, for each query, our model returns the top K similar images. In the top K returning images, the score will be 1 if there exists at least one positive image, and 0 otherwise.

Implementation details. We apply the widely-used Resnet (He et al. 2016) in our experiments with the pre-trained parameters. The input raw images are resized to 256×256 and cropped into 224×224. We train our models through using Stochastic Gradient Descent (SGD) optimizer with weight decay of 0.0001, momentum of 0.9, epochs of 90, and batch size of 32 on one GTX 2080ti GPU. The initial learning rate is set to 10^{-5}, with the exponential decay of 0.9 after every 5 epochs.

Ablation Experiments

We conduct some ablation experiments to illustrate the effectiveness of proposed modules, including the Nuance Modelling Module (NMM) and the Nuance Expansion Module (NEM). The baseline method uses ResNet-50 as the backbone network, followed by an FC layer as the classifier and trained with $\mathcal{L}_{CE}$ in the same setting.

As shown in Tab. 1, the contribution of each component is revealed. Compared with baseline, the NMM improves the

Method	Arch	CUB-200-2011				Stanford Cars 196				FGVC Aircraft			
Recall@k=		1	2	4	8	1	2	4	8	1	2	4	8
EPSHN (Xuan et al.)	R50	64.9	75.3	83.5	-	82.7	89.3	93.0	-	-	-	-	-
NSM (Zhai and Wu)	R50	65.3	76.7	85.4	91.8	89.3	94.1	96.4	98.0	-	-	-	-
MS (Wang et al.)	In3	65.7	77.0	86.3	91.2	84.1	90.4	94.0	96.1	-	-	-	-
HORDE (Jacob et al.)	In3	66.8	77.4	85.1	91.0	86.2	91.9	95.1	97.2	-	-	-	-
DGCRL (Zheng et al.)	R50	67.9	79.1	86.2	91.8	75.9	83.9	89.7	94.0	70.1	79.6	88.0	93.0
DCML (Zheng et al.)	R50	68.4	77.9	86.1	91.7	85.2	91.8	96.0	98.0	-	-	-	-
CEP (Boudiaf et al.)	R50	69.2	79.2	86.9	91.6	89.3	93.9	96.6	98.1	81.3	84.3	90.1	92.3
ETLR (Kim et al.)	R50	72.1	81.3	87.6	-	89.6	94.0	96.5	-	-	-	-	-
PNCA++ (Teh et al.)	R50	72.2	82.0	89.2	93.5	90.1	94.5	97.0	98.4	-	-	-	-
DAM (Xu et al.)	In3	72.3	81.2	87.8	92.7	88.9	93.4	96.0	97.7	-	-	-	-
SCDA (Wei et al.)	R50	57.3	70.2	81.0	88.4	48.3	60.2	71.8	81.8	56.5	67.7	77.6	85.7
PDDM (Bell and Bala)	R50	58.3	69.2	79.0	88.4	57.4	68.6	80.1	89.4	-	-	-	-
CRL (Zheng et al.)	R50	62.5	74.2	82.9	89.7	57.8	69.1	78.6	86.6	61.1	71.6	80.9	88.2
HDCL (Zeng et al.)	R50	69.5	79.6	86.8	92.4	84.4	90.1	94.1	96.5	71.1	81.0	88.3	93.3
Our CNENet	R50	**74.5**	**83.1**	**89.2**	**93.8**	**94.2**	**96.9**	**98.2**	**98.8**	**85.6**	**91.5**	**94.8**	**96.8**

Table 2: Comparison of different methods on CUB-200-2011, Stanford Cars 196 and FGVC Aircraft datasets. ”Arch” denotes the architecture of using backbone network. ”R50” and ”In3” represent Resnet50 (He et al. 2016) and Inception V3 (Szegedy et al. 2016), respectively.

Recall@1 accuracy by 4.9% due to discovering category-specific nuances and semantically aligning them guided by category. Moreover, we also verify the effectiveness of the semantic discrete loss $\mathcal{L}_{SD}$ and semantic alignment loss $\mathcal{L}_{SA}$, and find that $\mathcal{L}_{SA}$ plays a more vital role in FGOR. Based on the above results, we apply the original CARE maps generated by NMM to refine the selected nuances (Self-attention), while the performance drops by 0.9%. The result verifies that directly using self-attention can not refine CARE maps while introducing more learnable parameters, further making the network overfit on them. Therefore, NEM learns the correlation based on the features for refining the CARE maps, and outperforms BL + NMM by 3.3%. For existing metric-based methods, they use or design the pair-wise loss (i.e., Triplet loss) to perform the retrieval task. Therefore, we add Triplet loss to further constrain the learned features more compactly, but the accuracy drops by 1.1%. By this means that the pair-wise constraint limits the discriminative ability of feature representation, and our model can directly emphasize category-specific discrepancy to minimize the intra-class variances and maximize the inter-class differences. These results demonstrate that each module plays a role in effectively discovering category-specific nuances and semantically aligning them guided by the category information.

Comparison with the State-of-the-Art Methods

We compare our CNENet with state-of-the-art (SOTA) fine-grained object retrieval approaches. In Tab. 2, the performance of different methods on CUB-200-2011, Stanford Cars-196, and FGVC Aircraft datasets is reported, respectively. In the table from top to bottom, the methods are separated into three groups, which are (1) metric-based frameworks, (2) localization-based networks, and (3) our CNENet.

The success behind these models based on deep metric learning can be largely attributed to being able to precisely identify the negative/positive pairs via enlarging/shrinking their distances, which indirectly explores the discriminative ability of features. Despite the encouraging achievement, the existing works still have limited ability in learning discriminative features across different subcategories due to only paying more attention to the optimization of global features while overlooking nuances buried in the local regions. Existing works tend to localize regions to directly improve the discriminative ability of feature representation. Although the localization-based networks work well on various datasets, they are difficult to guarantee that the learned features are discriminative enough. Unlike these works, we propose CNENet to dig into category-specific nuances that contribute to category prediction, and thus explicitly emphasize discrepancies among subcategories. Therefore, our CNENet approach achieves new SOTA without any extra annotations and enjoys consistent improvement on various datasets.

As shown in Tab. 3, our approach outperforms these deep metric learning-based methods in the first group, which indicates that the proposed method can better minimize the intra-class variances and maximize the inter-class distances by directly exploring the category-specific nuances. Compared with recent localization-based works, they demonstrate the importance of localizing objects/parts. We run CNENet to directly learn category-specific nuances from images for emphasizing discrepancies among subcategories and achieve the new state-of-the-art.

Discussions

Response to Nuances. One of the keys to fine-grained images is to pick out discriminative nuances for improving the discriminability of features. To further illustrate the

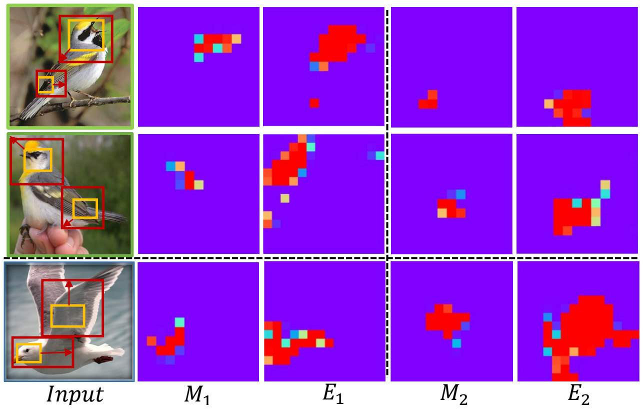

Figure 3: Visualization results of category-specific nuances on CUB-200-2011. In the first column, we show an input image with the red and yellow boxes respectively projected by NMM and NEM. The second and forth columns are CARE maps generated by NMM. The third and fifth columns are the refined category-specific response maps. The first and second rows have the same subcategory, and the third row has a different subcategory.

Number l	Performance (*Recall@k*)			
	1	2	4	8
2	74.2%	82.5%	88.8%	93.1%
4	**74.5%**	**83.1%**	**89.2%**	**93.8%**
5	72.2%	81.8%	88.3%	92.6%

Table 3: The retrieval accuracy on CUB-200-2011 of model trained with different number l of nuances in CNENet.

effectiveness of our proposed CNENet, which can attend category-specific nuances for discovering category-specific nuances and semantically align these nuances guided by category, we visualize the category-specific response maps (CARE) learned by NMM and NEM, respectively. Fig. 3 illustrates individual response nuances for three bird images of two subcategories. We can observe that each CARE map M_1, M_2 generated by NMM focuses on a certain nuance different from the others without the effect of pose or viewpoint. Moreover, CARE maps of the same order emphasize the same semantic information in images of the same subcategory, whereas this relationship does not exist for ones with different subcategories. To verify the effectiveness of NEM, we also visualize the refined CARE maps E to expand the shrank nuances by utilizing the correlation between feature vectors. Compared with the original maps M_i, the corresponding E_i can pay attention to the entire nuances rather than shrank ones caused by the constraint of semantic discrete loss, which resumes some discriminative nuances and further improves the discriminative ability of feature representation. To more intuitively display the contribution of NMM, we roughly project the localization of nuances generated by NMM and NEM into the yellow and red bounding boxes in the images.

Visualized Distributions. To illustrate the impact of CNENet on exploring subcategory discrepancy, we carefully select 10 subcategories with small discrepancy from the testing set to visualize the distributions of learned features in Fig. 4, where each distinct color denotes a fine-grained subcategory. As shown in Fig. 4(a), the features extracted by baseline network have difficulty in alleviating the large intraclass variances, and using these features thus degrades the retrieval performance. In Fig. 4(b), the learnt features with CNENet are well clustered by subcategory. Besides, the distance between the features of different subcategories is farther, and the features of the same subcategory are more compact. Furthermore, improving the discriminative ability of features by discovering and aligning category-specific nuances achieves a vital improvement.

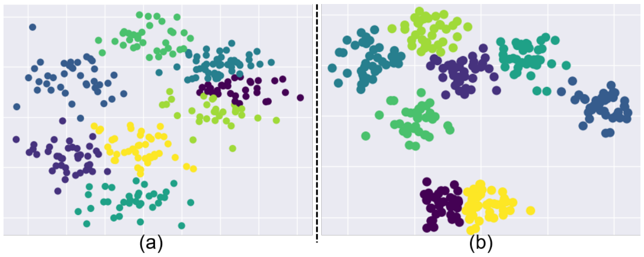

Figure 4: Visualization of learned features, where each color represents an subcategory in the testing set. The triangles indicate the features extracted from input samples. (a) is features extracted by baseline; and (b) denotes features extracted by our CNENet.

The more, the better? We show the retrieval performance with the different number of category-specific nuances, as shown in Tab. 3. The performance of CNENet drops when the number of nuances increases to 4. The result means that an excessive number of category-specific nuances can introduce more useless features, while fewer details can miss informative features. It should be clarified that the number of nuances is explicitly divided into l groups from the spacial perspective for emphasizing the discrepancies among subcategories. Nevertheless, since each category-specific response map may contain different semantic nuances, the number of nuances could be different from a semantic perspective.

Conclusion

In this paper, we propose a novel method called category-specific nuance exploration network (CNENet) for FGOR, which solves the problem of how to effectively extract category-specific nuances and how to semantically align these nuances grouped by category. The exploration strategy can be considered as a self-supervised scheme that enables the network to adaptively dig into category-specific nuances by category. Extensive experiments show that the retrieval performance can be improved significantly by discovering the nuances. The last but the most important, our algorithm is end-to-end trainable, and achieves state-the-of-the-art in CUB-200-2011, Stanford Cars and FGVC Aircraft datasets.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Grant NO. 61976038, NO. 61932020, and NO. 61772108.

References

Bell, S.; and Bala, K. 2015. Learning visual similarity for product design with convolutional neural networks. *ACM Trans. Graph.*, 34(4): 98:1–98:10.

Boudiaf, M.; Rony, J.; Ziko, I. M.; Granger, E.; Pedersoli, M.; Piantanida, P.; and Ayed, I. B. 2020. A Unifying Mutual Information View of Metric Learning: Cross-Entropy vs. Pairwise Losses. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, volume 12351 of *Lecture Notes in Computer Science*, 548–564. Springer.

Branson, S.; Horn, G. V.; Belongie, S. J.; and Perona, P. 2014. Bird Species Categorization Using Pose Normalized Deep Convolutional Nets. *CoRR*, abs/1406.2952.

Ding, Y.; Zhou, Y.; Zhu, Y.; Ye, Q.; and Jiao, J. 2019. Selective Sparse Sampling for Fine-Grained Image Recognition. In *The IEEE International Conference on Computer Vision (ICCV)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778.

Jacob, P.; Picard, D.; Histace, A.; and Klein, E. 2019. Metric Learning With HORDE: High-Order Regularizer for Deep Embeddings. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 6538–6547. IEEE.

Ji, R.; Wen, L.; Zhang, L.; Du, D.; Wu, Y.; Zhao, C.; Liu, X.; and Huang, F. 2020. Attention Convolutional Binary Neural Tree for Fine-Grained Visual Categorization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 10465–10474. IEEE.

Kim, S.; Kim, D.; Cho, M.; and Kwak, S. 2021. Embedding Transfer With Label Relaxation for Improved Metric Learning. In *CVPR*, 3967–3976. Computer Vision Foundation / IEEE.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3D Object Representations for Fine-Grained Categorization. In *ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, 554–561.

Li, X.; Yang, Y.; Zhao, Q.; Shen, T.; Lin, Z.; and Liu, H. 2020. Spatial Pyramid Based Graph Reasoning for Semantic Segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 8947–8956. IEEE.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M. B.; and Vedaldi, A. 2013. Fine-Grained Visual Classification of Aircraft. *CoRR*, abs/1306.5151.

Moskvyak, O.; Maire, F.; Dayoub, F.; and Baktashmotlagh, M. 2021. Keypoint-Aligned Embeddings for Image Retrieval and Re-identification. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, 676–685. IEEE.

Shen, C.; Zhou, C.; Jin, Z.; Chu, W.; Jiang, R.; Chen, Y.; and Hua, X. 2017. Learning Feature Embedding with Strong Neural Activations for Fine-Grained Retrieval. In Wu, W.; Yang, J.; Tian, Q.; and Zimmermann, R., eds., *Proceedings of the on Thematic Workshops of ACM Multimedia 2017, Mountain View, CA, USA, October 23 - 27, 2017*, 424–432. ACM.

Song, H. O.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep Metric Learning via Lifted Structured Feature Embedding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 4004–4012. IEEE Computer Society.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2818–2826. IEEE Computer Society.

Teh, E. W.; DeVries, T.; Taylor, G. W.; and Graham. 2020. ProxyNCA++: Revisiting and Revitalizing Proxy Neighborhood Component Analysis. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIV*, volume 12369 of *Lecture Notes in Computer Science*, 448–464. Springer.

Wang, S.; Li, H.; Wang, Z.; and Ouyang, W. 2021. Dynamic Position-aware Network for Fine-grained Image Recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 2791–2799. AAAI Press.

Wang, S.; Wang, Z.; Li, H.; and Ouyang, W. 2020a. Category-specific Semantic Coherency Learning for Fine-grained Image Recognition. In Chen, C. W.; Cucchiara, R.; Hua, X.; Qi, G.; Ricci, E.; Zhang, Z.; and Zimmermann, R., eds., *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, 174–183. ACM.

Wang, X.; Girshick, R. B.; Gupta, A.; and He, K. 2018. Non-Local Neural Networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 7794–7803. IEEE Computer Society.

Wang, X.; Han, X.; Huang, W.; Dong, D.; and Scott, M. R. 2019a. Multi-Similarity Loss With General Pair Weighting for Deep Metric Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 5022–5030. Computer Vision Foundation / IEEE.

Wang, Z.; Wang, S.; Li, H.; Dou, Z.; and Li, J. 2020b. Graph-Propagation Based Correlation Learning for Weakly

Supervised Fine-Grained Image Classification. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 12289–12296. AAAI Press.

Wang, Z.; Wang, S.; Yang, S.; Li, H.; Li, J.; and Li, Z. 2020c. Weakly Supervised Fine-Grained Image Classification via Guassian Mixture Model Oriented Discriminative Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Z.; Wang, S.; Zhang, P.; Li, H.; Zhong, W.; and Li, J. 2019b. Weakly Supervised Fine-grained Image Classification via Correlation-guided Discriminative Learning. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, 1851–1860.

Wei, X.; Luo, J.; Wu, J.; and Zhou, Z. 2017. Selective Convolutional Descriptor Aggregation for Fine-Grained Image Retrieval. *IEEE Trans. Image Process.*, 26(6): 2868–2881.

Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; and Fu, Y. 2020. Rethinking Classification and Localization for Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 10183–10192. IEEE.

Xu, F.; Wang, M.; Zhang, W.; Cheng, Y.; and Chu, W. 2021. Discrimination-Aware Mechanism for Fine-Grained Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 813–822. Computer Vision Foundation / IEEE.

Xuan, H.; Stylianou, A.; Pless, R.; and Pless, R. 2020. Improved Embeddings with Easy Positive Triplet Mining. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, 2463–2471. IEEE.

Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; and Wang, L. 2018. Learning to Navigate for Fine-Grained Classification. In *ECCV, Germany, September 8-14, 2018, Proceedings, Part XIV*, 438–454.

Zeng, X.; Liu, S.; Wang, X.; Zhang, Y.; Chen, K.; and Li, D. 2021. Hard Decorrelated Centralized Loss for fine-grained image retrieval. *Neurocomputing*, 453: 26–37.

Zhai, A.; and Wu, H. 2019. Classification is a Strong Baseline for Deep Metric Learning. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, 91. BMVA Press.

Zhang, Y.; Wei, X.; Wu, J.; Cai, J.; Lu, J.; Nguyen, V. A.; and Do, M. N. 2016. Weakly Supervised Fine-Grained Categorization With Part-Based Image Representation. *TIP*, 25(4): 1713–1725.

Zheng, H.; Fu, J.; Zha, Z.; and Luo, J. 2019a. Looking for the Devil in the Details: Learning Trilinear Attention Sampling Network for Fine-Grained Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 5012–5021.

Zheng, W.; Wang, C.; Lu, J.; and Zhou, J. 2021. Deep Compositional Metric Learning. In *CVPR*, 9320–9329. Computer Vision Foundation / IEEE.

Zheng, X.; Ji, R.; Sun, X.; Wu, Y.; Huang, F.; and Yang, Y. 2018. Centralized Ranking Loss with Weakly Supervised Localization for Fine-Grained Object Retrieval. In Lang, J., ed., *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 1226–1233. ijcai.org.

Zheng, X.; Ji, R.; Sun, X.; Zhang, B.; Wu, Y.; and Huang, F. 2019b. Towards Optimal Fine Grained Retrieval via Decorrelated Centralized Loss with Normalize-Scale Layer. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 9291–9298. AAAI Press.

Zhou, M.; Bai, Y.; Zhang, W.; Zhao, T.; and Mei, T. 2020. Look-Into-Object: Self-Supervised Structure Modeling for Object Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 11771–11780. IEEE.