Computer Systems Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.032047
Article

Learning Noise-Assisted Robust Image Features for Fine-Grained Image Retrieval

Vidit Kumar[1,*], Hemant Petwal[2], Ajay Krishan Gairola[1] and Pareshwar Prasad Barmola[1]

[1]Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, 248002, India
[2]School of Computer Science Engineering and Technology, Bennett University, Greater Noida, 201310, Uttar Pradesh, India
*Corresponding Author: Vidit Kumar. Email: viditkumar.cse@geu.ac.in
Received: 05 May 2022; Accepted: 08 December 2022

Abstract: Fine-grained image search is one of the most challenging tasks in computer vision that aims to retrieve similar images at the fine-grained level for a given query image. The key objective is to learn discriminative fine-grained features by training deep models such that similar images are clustered, and dissimilar images are separated in the low embedding space. Previous works primarily focused on defining local structure loss functions like triplet loss, pairwise loss, etc. However, training via these approaches takes a long training time, and they have poor accuracy. Additionally, representations learned through it tend to tighten up in the embedded space and lose generalizability to unseen classes. This paper proposes a noise-assisted representation learning method for fine-grained image retrieval to mitigate these issues. In the proposed work, class manifold learning is performed in which positive pairs are created with noise insertion operation instead of tightening class clusters. And other instances are treated as negatives within the same cluster. Then a loss function is defined to penalize when the distance between instances of the same class becomes too small relative to the noise pair in that class in embedded space. The proposed approach is validated on CARS-196 and CUB-200 datasets and achieved better retrieval results (85.38% recall@1 for CARS-196% and 70.13% recall@1 for CUB-200) compared to other existing methods.

Keywords: Convolutional network; zero-shot learning; fine-grained image retrieval; image representation; image retrieval; intra-class diversity; feature learning

1 Introduction

For the past two decades, extensive research has been done on image retrieval, which has proven useful [1]. In applications where classes have more significant inter-class variance, prior image retrieval methods work better, but not for classes with sizeable intra-class variance compared to inter-class

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

variance. Conceptually, but in the real World, the search must be meticulous, that is, the relevant subcategory to be found by its definition, as a concept, in general, may not contain every single piece of data. For instance, a query for images like birds or dogs, or cars categories must have the same fine-grain subset of images returned. It becomes complex and challenging to search out. This is because, e.g., when looking at a single component of the bike or bird, only its textural details, such as the bird's feathers and the bike's headlight, are evident; only in detail comparison is anything identifiable, and therefore the problem is challenging to solve [2]. Learning robust descriptions is crucial in fine-grained image retrieval. If an efficient retrieval method is used, it enables images to be associated with similar features at the front and dissimilar ones at the end. In this regard, most of the prior works were based on deep metric learning (DML) paradigms. Several methods were built upon contrastive loss [3], triplet loss [4,5] and quadruplet loss [6,7]. The contrastive loss is usually based on the pair, i.e., a positive pair consists of images from the same class, while the negative pair consists of images of different classes. Contrastive loss aims to penalize the case when the negative pair distance is smaller than some margin and when the positive pair distance is greater. The disadvantage of this loss is that it does not take relative distances but only focuses on the pair. This was tackled by triplet loss, which is based on the triplets rather than pairs. A triplet is formed by choosing an anchor image, a positive image from the same anchor class, and a negative image from a different class. The Triplet loss aims to penalize the case when a positive pair (anchor-positive) distance is greater than the negative pair (anchor-negative) distance by some margin. Many previous works employed the triplet loss and improved it by including more negatives in the loss function [8–10]. These strategies were based on the common goal of DML, which aims to cluster all similar class images in the embedding space as closely as possible, resulting in more tightly clustered classes in the embedded space. However, this is not true for large intra-class variance databases since no image does not necessarily look like all the images within the same class. In addition, existing efforts [8–12] had limited efficiency since they hugely depended on sampling strategies. In contrast, this paper deals with instances sampled from the minibatch without requiring hard-mining strategies. This paper proposes an approach to create positive pairs inside the class cluster to diversify class instances while separating them from other class instances. The basic idea is illustrated in Fig. 1.

Contribution:

- A noise-assisted feature learning approach for fine-grained image search is proposed, which aims to learn diverse features and reduce the costly sampling process of triple loss.
- The main goal of the proposed work is to learn a manifold of each class instead of just class discriminative learning in the triplet-based loss. We do this by contrasting positive pairs (created by noise) in each class cluster in the embedded space.
- The importance of intra-class diversity in the embedded space is studied as it helps in better generalization to unseen classes.
- To include more diversity in the learned features, this paper exploits self-supervised constraint, which further helps to improve retrieval performance.
- Finally, the proposed approach is validated with experiments conducted on well-known fine-grain datasets, which show improved performance compared to existing techniques.

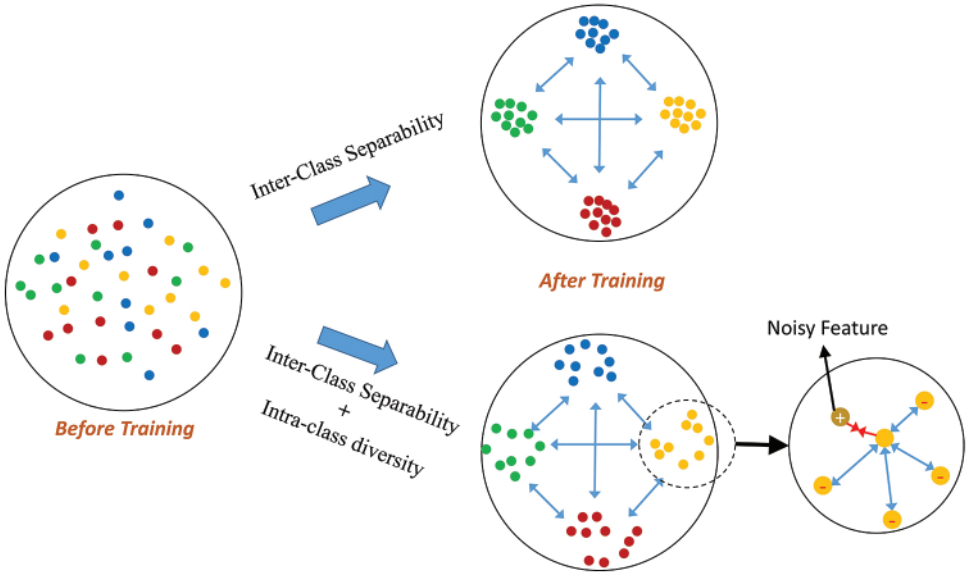

Figure 1: The proposed approach focuses on inter-class separability and intra-class diversity learning. The proposed work lies in the inclusion of intra-class diversity in embedded space. We aim to learn manifolds for each class because no query image necessarily looks identical to all instances of that class but only to the most similar instances

2 Related Work

Research into image retrieval was also spurred by the success of the convolutional neural network (CNN) [13] and other deep learning techniques. For example, a pretrained CNN was used by Babenko et al. [14] for image retrieval and representation, and the CNN's responses were used to fine-tune it on the target images. Using sum pooling on deep features, the feature aggregation method was proposed in [15,16]. According to Shakarami et al. [17], a descriptor for image retrieval was built on the fusion of three different features: local binary pattern (LBP), histogram of gradients (HOG), and CNN. To encode many locations with the convolutional layers' encodings, Tolias et al. [18] described a method for producing compact features. A feature learning technique with a cross-batch reference strategy for picture retrieval was suggested by Yang et al. [19]. The bag-of-words approach was used by Mohedano et al. [20] to exploit CNN features, while CNN features with vector locally aggregated descriptors (VLAD) are utilized in [21]. These techniques were good at a coarse level, but fine-grained images require the detection of subtle locations.

Recently, research on fine-grained image tasks has been done using the deep learning paradigm. The challenge is how to locate and represent subtle details. An approach to fine-grained vehicle classification by Watkins et al. [22] is to first locate the item with a trained classifier and then deploy this detected border-box for vehicle classification. For object localization, the authors in [23] used pre-trained VGG-16 [24] and removed noise or background to pick its deep descriptors. The author of [25] made use of convolutional kernels to pick and represent the object's sections. ResNet18 [26] was investigated by Kumar et al. [27] for the fine-grained image retrieval (FGIR) task, and its activations were employed for retrieval. For fine-grained categorization, Zhou et al. [28] investigated the label hierarchy to exploit rich associations through bipartite graphs and VGG-nets [24]. The centralized

ranking loss was proposed by Zheng et al. [29] for weakly supervised object localization. The contours of the CNN response map were then used to extract the features.

In addition, some efforts were made to embed learning. Like, [4] utilized pairwise loss, and [5] utilized triplet loss on top of CNN for image embedding learning. Other extensions of these works are [8–12]. Most of these methods ignored inter-class diversity and were based on costly hard mining strategies. Further, Vasudeva et al. [30] looked for optimal hard mining. In addition, Xuan et al. [31] showed the importance of intra-class variance in learning embedding. To this end, we focus on incorporating intra-class diversity into the embedding space while minimizing the sampling costs.

3 Methodology

The main objective of our work is illustrated in Fig. 2. Here, the goal is to learn class manifold by including both Inter-Class separability and Intra-class diversity. To enhance a network's potential for feature representation, noise can aid in the deep CNN's learning of more accurate representations. Prior efforts such as [32,33] commonly used noisy labels to train feature representation networks, which requires a large dataset with noisy labels. Rather than of utilizing noisy labels, we train the network in this study by inserting randomness (noise) at the input and last layer. At the input layer, a noisy image is formed, and at the output layer, a noisy feature is created, both of which act as randomness to the network, which helps in achieving the model's generalizability.

Let training images $\{x_1, x_2, \ldots, x_m\} \in X$ with associated labels $\{y_1, y_2, \ldots, y_m\} \in Y$ in the given minibatch for training the network. The goal is to learn low embedding for each image with the following constraint.

$$\begin{cases} D_y\left(f_\theta\left(x_i\right), f_\theta\left(x_j\right)\right) \to 0, \text{ if } y_i = y_j; \\ D_y\left(f_\theta\left(x_i\right), f_\theta\left(x_j\right)\right) > \alpha, \text{ if } y_i \neq y_j, \end{cases} \tag{1}$$

where, $f_\theta\left(x_j\right)$ is the feature embedding of image x_j, f_θ is the feature extraction network, and D_y may be Euclidean or cosine distance.

Let $f_i = f_\theta\left(x_i\right)$, I be the set of indices corresponding to the number of instances in a minibatch, A(i) be the set of indices of all positives to the i^{th} image instance, and C(i) be the set of indices of all negatives to i^{th} image instance.

For inter-class variance, the positives from simple transformation operations like rotation, cropping, flipping, etc., are considered. The metric constraint for $\left(X_i, X_p, X_n\right) : Y_i = Y_p, Y_i \neq Y_n$ in terms of cosine similarity could be defined as:

$$\left(f_i \odot f_p\right) > \left(f_i \odot f_n\right), \forall i, p, n \in I, \; y_i = y_p, y_i \neq y_n \tag{2}$$

and the loss for constraint (2) can be defined as:

$$L_{inter} = \sum_{i \in I} \frac{1}{|A(i)|} \sum_{p \in A(i)} \left(-\log \frac{\exp(f_i \odot f_p / \tau)}{\exp(f_i \odot f_p / \tau) + \sum_{n \in C(i)} \exp(f_i \odot f_n / \tau)} \right) \tag{3}$$

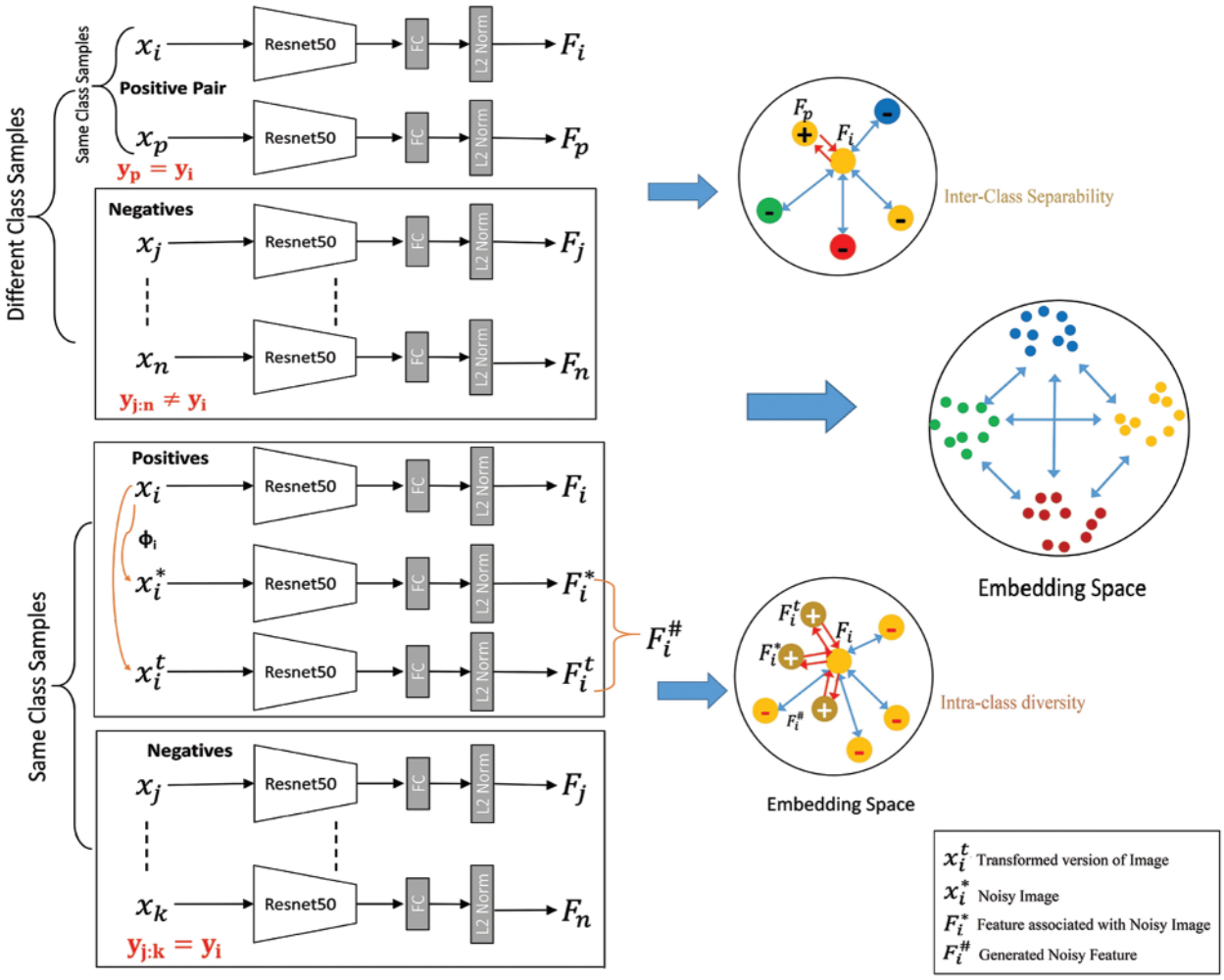

Figure 2: Overview of the proposed approach under FGIR

3.1 Noise-Assisted Feature Learning (NAFL)

Fine-grain images generally possess a large intra-class variance than an inter-class variance. Hence, the same should also be true in the embedding space. Prior works focused on separating class clusters as much as possible, neglecting intra-class variance and resulting in the tightening of class clusters.

Thus, to incorporate intra-class variance, class manifold learning should be utilized. For this, the positive pairs are formed as: given an input image x_i, the positive candidate for x_i computed as (4)

$$x_i^*[a{:}a+k,\ b{:}b+k] = \text{shuffle}(x_i[a{:}a+k,\ b{:}b+k] \tag{4}$$

Eq. (4) states that for each non-overlapping window of size $k \times k$ in the image x_i, do shuffle the pixel values.

Fig. 3 depicts the noisy images with various-sized windows. On comparing to Gaussian noise, which adds noise all over the image, this work creates a noisy image by simply shuffling image intensities in the neighbor. This can be clearly visible in Fig. 3.

Now, consider the f_i^* be the L_2 normalized feature embedding of the noisy image x_i^*. Consider the cosine similarity $(\cdot, \odot, \cdot)$ between f_i and f_i^* as $f_i \odot f_i^*$, where $\odot$ is the dot product. The goal is to maximize $f_i \odot f_i^*$ for all instances $x_i \in X$. Similar to (1), the metric constraint for (x_i, x_i^*, x_j) : $y_i = y(x_i^*) = y_j$ could be defined as:

$$(f_i \odot f_i^*) > (f_i \odot f_j), \forall x_i, x_j \in X, \quad y_i = y(x_i^*) = y_j \tag{5}$$

where, x_i^* is a noised version of x_i which is produced by applying (4).

For constraint (5), the loss can be defined as:

$$L_{intra} = \sum_{i \in I} \left[-\log \frac{\exp(f_i \odot f_i^* / \tau)}{\exp(f_i \odot f_i^* / \tau) + \sum_{n \in A(i)} \exp(f_i \odot f_n / \tau)} \right] \tag{6}$$

In addition to the noised image, noisy feature $f_i^{\#} = f_i^* + f_i^t$ is computed to further include it as the positive sample, where f_i^t is the embedding of the transformed version. The loss L_{intra} can be rewritten as:

$$L_{intra} = \sum_{i \in I} \sum_{p1 \in (f_i^*, f_i^{\#})} \left(-\log \frac{\exp(f_i \odot p1/\tau)}{\exp(f_i \odot p1/\tau) + \sum_{n \in A(i)} \exp(f_i \odot f_n/\tau)} \right) \tag{7}$$

The network is trained with backpropagation to jointly minimize the two losses (3) and (7) over each minibatch m sampled from the database.

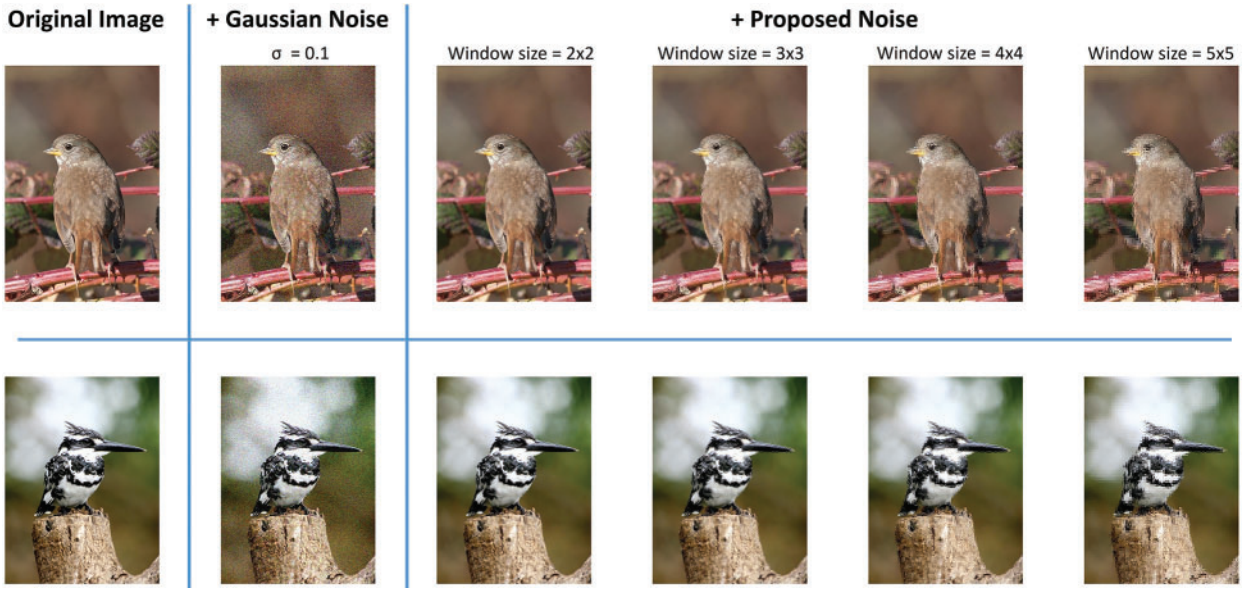

Figure 3: Noisy images under different window sizes (Best view during zoom and in color)

4 Implementation and Dataset Settings

The Imagenet pre-trained model Resnet-50 [26] is used as a base network. The embedding size is set to 1024. The learning rate is set to 10e-4, and the minibatch size is set to 64 images. For data augmentation, random rotation, reflection, crop, and color augmentation [34] are used. All the experiments are performed on MATLAB 2019b with NVIDIA Tesla K40c. Two widely used datasets,

CARS-196 [35] and CUB-200 [36], are selected to evaluate the proposed method. The CUB-200 dataset consists of 11,788 images of 200 bird species. Following the state-of-the-art evaluation protocol [37], the first 100 species are used for training and the next 100 for testing. Similarly, the CARS-196 dataset consists of 198 car models and 16,185 images. The first 98 models are used for training, and the next 98 for testing.

5 Results

In this section, the FGIR results are reported, which includes an analysis of the effect of noise and the importance of intra-class diversity in retrieval performance.

5.1 Effect of Noise on the Network

To evaluate the influence of the noise on retrieval performance, the experiments are performed on the CARS-196 and CUB-200, and the results in the form of recall@k are depicted in Figs. 4 and 5. It can be observed from the figures that using a gaussian noise image improves retrieval performance compared to not using noise. Consequently, with the noisy image formed with the proposed method, performance is further improved, which is reflected in recall@k. This demonstrates that noise helps in learning the model generalization better.

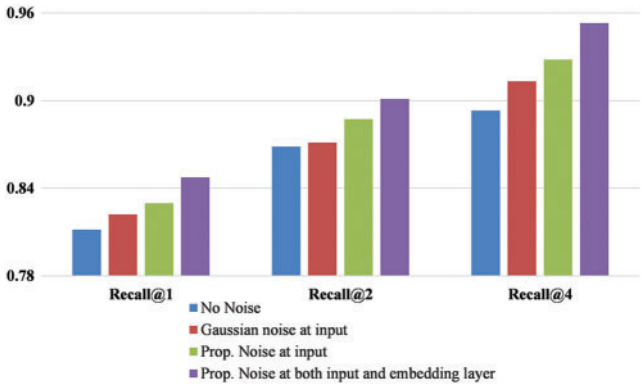

Figure 4: Recall@k for CARS-196 under different settings of noise

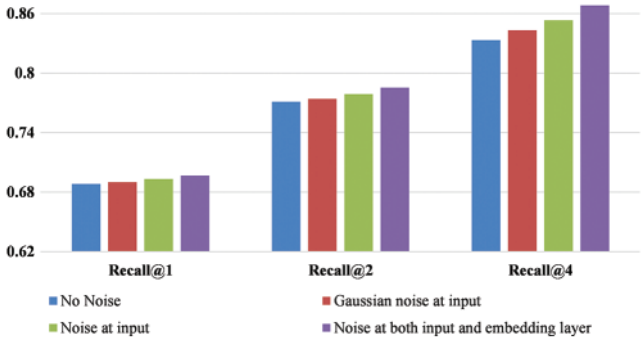

Figure 5: Recall@k for CUB-200 under different settings of noise

5.2 Importance of Intra-Class Diversity

Next, the importance of intra-class diversity in the embedding space is examined. The fine-grained classification results by following standard training and testing protocol [35] are depicted in Table 1. In addition, the retrieval results are depicted in Table 2. As shown in Tables 1 and 2, when the inter-separability is included as a constraint in the embedding space, the classification accuracy improves for both datasets. For example, as in Table 1, for CARS-196, the accuracy of 90.2% is improved to 91.72% by training the model with the inter-separability loss compared to the cross entropy loss. Further, this too improved to 93.68% accuracy when the intra-diversity loss was also added. Therefore, this demonstrates the importance of intra-class diversity in the embedded space. Table 2 reports the retrieval results for both datasets where top-1 and top-2 retrieval accuracy improves with the inclusion of the intra-class diversity loss. This further confirms the importance of intra-class diversity.

Table 1: Top-1 classification accuracy

Base network	Loss	CARS-196	CUB-200
Resnet50	Standard cross entropy	90.23%	80.12%
	L_{inter}	91.72%	81.53%
	$L_{inter} + L_{intra}$	93.68%	82.44%

Table 2: Top-1 and Top-2 retrieval accuracy

Base network	Loss	CARS-196		CUB-200	
		Top-1	Top-2	Top-1	Top-2
Resnet50	Triplet loss	79.03%	84.01%	67.15%	76.23%
	L_{inter}	81.16%	86.86%	68.83%	77.11%
	$L_{inter} + L_{intra}$	84.75%	90.12%	69.67%	78.53%

5.3 Performance Enhancement with Self-Supervision Constraint

To further improve the model's performance, the self-supervised learning approach can be explored. For this, the rotation [38], exemplar [39], and Jigsaw-puzzle [40] are tested. For the rotation task, the rotation angles in the range (0°, 90°, 180°, 270°) are considered. For the exemplar task, six transformations (translation, scaling, rotation, contrast, adding color intensity) are considered for each surrogate class. The results are depicted in Figs. 6 and 7, where it can be seen that one with the Jigsaw-puzzle as additional self-supervised learning (SSL) can improve retrieval performance compared to others. This appears to be because the jigsaw-puzzle task forces the model to focus on sub-parts or fine-grained parts of the image. However, rotation and exemplar tasks give the network a small amount of regularization.

5.4 Comparison to State-of-the-Arts

Next, the comparative analysis of the proposed method with other state-of-the-art are reported in Table 3, where the proposed method achieves better results with 84.75% Recall@1, 90.12% Recall@2, 93.27% Recall@4, 95.32% Recall@8, 97.21% Recall@16 for CARS-196. Further, this is improved

by including self-supervision task (Jigsaw-puzzle), and the results are 85.38% Recall@1, 91.90% Recall@2, 95.1% Recall@4, 97.14% Recall@8, 98.99% Recall@16.

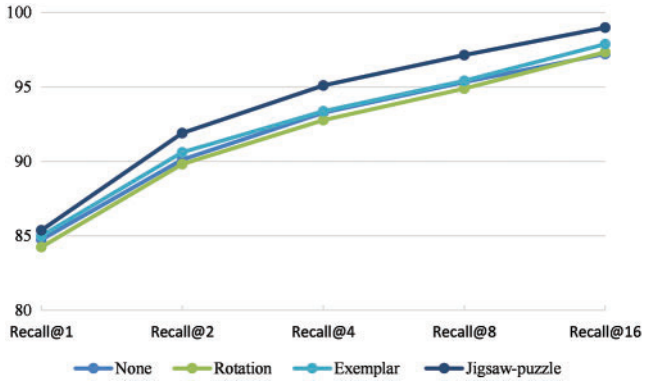

Figure 6: Recall@k for CARS-196 under different self-supervision tasks

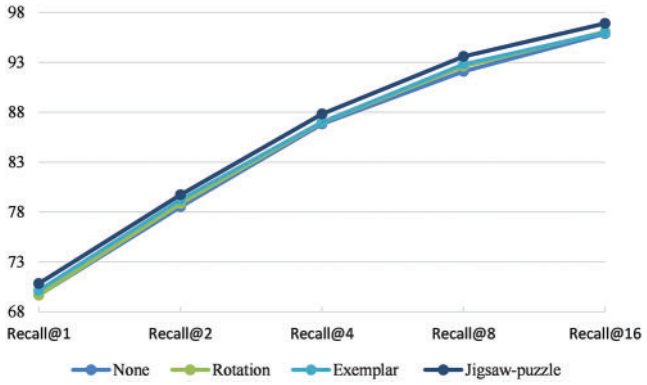

Figure 7: Recall@k for CUB-200 under different self-supervision tasks

Table 3: Performance (Recall@k) comparison under zero-shot setting

Method	CARS-196					CUB-200				
	k = 1	k = 2	k = 4	k = 8	k = 16	k = 1	k = 2	k = 4	k = 8	k = 16
Contrastive [3]	21.7	32.3	46.1	58.9	72.2	26.4	37.7	49.8	62.3	76.4
Triplet [4]	39.1	50.4	63.3	74.5	84.1	36.1	48.6	59.3	70.0	80.2
LiftedStruct [9]	49.0	60.3	72.1	81.5	89.2	47.2	58.9	70.2	80.2	89.3
N-pairs [8]	53.9	66.8	77.7	86.3	-	45.4	58.4	69.5	79.4	-
Facility location [10]	58.1	70.6	80.3	87.8	-	48.2	61.4	71.8	81.9	-
PDDM+Quadruplet [7]	57.4	68.6	80.1	89.4	92.3	58.3	69.2	79.0	88.4	93.1
SCDA [23]	58.5	69.8	79.1	86.2	91.8	62.2	74.2	83.2	90.1	94.3
CRL-WSL [29]	63.9	73.7	82.1	89.2	93.7	65.9	76.5	85.3	90.3	94.4
DGCRL [37]	75.9	83.9	89.7	94.0	96.6	67.9	79.1	86.2	91.8	94.8
EPSHN [31]	82.7	89.3	93.0	-	-	64.9	75.3	83.5	-	-

(Continued)

Table 3: Continued

Method	CARS-196					CUB-200				
	k = 1	k = 2	k = 4	k = 8	k = 16	k = 1	k = 2	k = 4	k = 8	k = 16
(NAFL) (R50)	**84.75**	**90.12**	**93.27**	**95.32**	**97.21**	**69.67**	**78.53**	**86.82**	**92.1**	**95.87**
(NAFL + SSL) (R50)	**85.38**	**91.90**	**95.1**	**97.14**	**98.99**	**70.13**	**78.88**	**87.17**	**92.4**	**96.8**

Similarly, for CUB-200, the proposed method NAFL achieves 69.67% Recall@1, 78.53% Recall@2, 86.82% Recall@4, 92.1% Recall@8, 95.87% Recall@16. In addition, this is further improved to 70.13% Recall@1, 78.88% Recall@2, 87.17% Recall@4, 92.4% Recall@8, and 96.8% Recall@16 by SSL. We also plot the retrieval results for the randomly sampled query from both datasets in Figs. 8 and 9.

Figure 8: Retrieval results on CARS-196 dataset. The green boundary box indicates the correct retrieved instance, and the red boundary box indicates the wrong retrieved instance

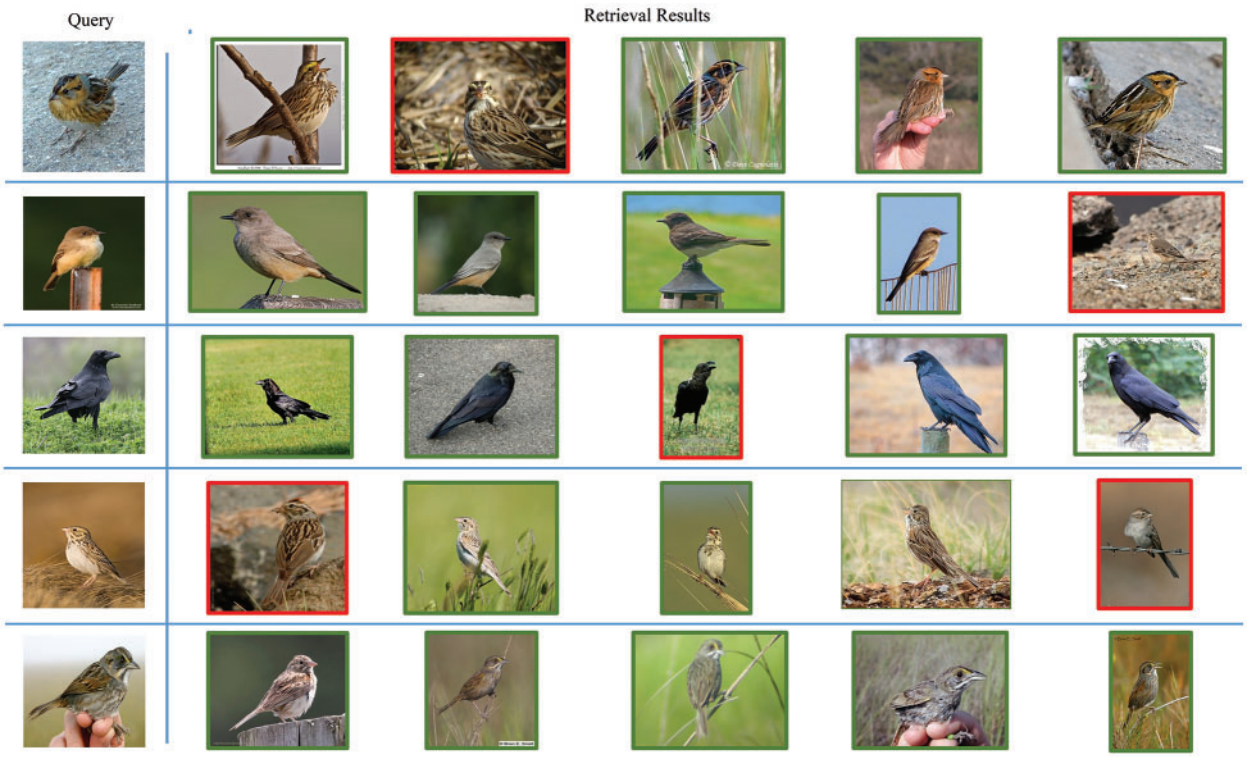

Figure 9: Retrieval results on the CUB-200 dataset. The green boundary box indicates the correct retrieved instance, and the red boundary box indicates the wrong retrieved instance

6 Conclusion

This paper presents a noise-assisted feature learning strategy for FGIR, which avoids the costly sampling process (as in triplet-based learning). This is accomplished by inserting noise into the input sample as well as the deep CNN's embedding. The Resnet-50 architecture is used as the backbone network and trained jointly with a multi-loss objective, which deals with both class discriminative and intra-class diversity via noise-assisted learning. The CUB-200 and CARS-196 datasets are considered to validate our approach. Moreover, it showed that the proposed approach is able to improve over existing schemes. Future work will test the proposed method for large-scale datasets with the use of other deeper variants of CNN. In addition, it can be extended to deal with video analysis tasks like action classification and video retrieval. It can also be tested in the medical domain using both supervised and unsupervised learning strategies. In addition, it can be tested in self-supervised learning [41–45] mode.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] I. M. Hameed, S. H. Abdulhussain and B. M. Mahmmod, "Content-based image retrieval: A review of recent trends," *Cogent Engineering*, vol. 8, no. 1, pp. 1–37, 2021.

[2] L. Xie, J. Wang, B. Zhang and Q. Tian, "Fine-grained image search," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 636–647, 2015.

[3] S. Bell and K. Bala, "Learning visual similarity for product design with convolutional neural networks," *ACM Transactions on Graphics*, vol. 34, no. 4, pp. 1–10, 2015.

[4] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang *et al.,* "Learning fine-grained image similarity with deep ranking," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 1386–1393, 2014.

[5] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, USA, pp. 815–823, 2015.

[6] W. Chen, X. Chen, J. Zhang and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc.—30th IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, HI, USA, pp. 403–412, 2017.

[7] C. Huang, C. C. Loy and X. Tang, "Local similarity-aware deep feature embedding," *Advances in Neural Information Processing Systems*, vol. 29, pp. 1270–1278, 2016.

[8] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," *Advances in Neural Information Processing Systems*, vol. 29, pp. 1857–1865, 2016.

[9] H. O. Song, Y. Xiang, S. Jegelka and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 4004–4012, 2016.

[10] H. O. Song, S. Jegelka, V. Rathod and K. Murphy, "Deep metric learning via facility location," in *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 2206–2214, 2017.

[11] B. Yu, T. Liu, M. Gong, C. Ding and D. Tao, "Correcting the triplet selection bias for triplet loss," in *European Conf. in Computer Vision, ECCV*, Munich, Germany, pp. 71–86, 2018.

[12] C. Y. Wu, R. Manmatha, A. J. Smola and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *2017 IEEE Int. Conf. on Computer Vision (ICCV)*, Venice, Italy, pp. 2859–2867, 2017.

[13] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 1, pp. 1097–1105, 2012.

[14] A. Babenko, A. Slesarev, A. Chigorin and V. Lempitsky, "Neural codes for image retrieval," in *European Conf. on Computer Vision*, Zurich, Switzerland, pp. 584–599, 2014.

[15] A. B. Yandex and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 1269–1277, 2015.

[16] Y. Kalantidis, C. Mellina and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *European Conf. on Computer Vision*, Amsterdam, The Netherlands, pp. 685–701, 2016.

[17] A. Shakarami and H. Tarrah, "An efficient image descriptor for image classification and CBIR," *Optik (Stuttg)*, vol. 214, pp. 164833, 2020.

[18] G. Tolias, R. Sicre and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," in *ICLR 2016-Int. Conf. on Learning Representations*, San Juan, Puerto Rico, pp. 1–12, 2016.

[19] H. F. Yang, K. Lin and C. S. Chen, "Cross-batch reference learning for deep classification and retrieval," in *Proc. of the 24th ACM Int. Conf. on Multimedia*, Ottawa, ON, Canada, pp. 1237–1246, 2016.

[20] E. Mohedano, K. Mcguinness, N. E. O'Connor, A. Salvador, F. Marqués *et al.,* "Bags of local convolutional features for scalable instance search," in *Proc. of the 2016 ACM on Int. Conf. on Multimedia Retrieval*, New York, USA, pp. 327–331, 2016.

[21] J. Y. H. Ng, F. Yang and L. S. Davis, "Exploiting local features from deep networks for image retrieval," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, Boston, MA, USA, pp. 53–61, 2015.

[22] R. Watkins, N. Pears and S. Manandhar, "Vehicle classification using ResNets, localisation and spatially-weighted pooling," ArXiv., 2018.

[23] X. S. Wei, J. H. Luo, J. Wu and Z. H. Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2868–2881, 2017.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd Int. Conf. on Learning Representations, ICLR 2015*, San Diego, CA, USA, 2015.

[25] X. Zhang, H. Xiong, W. Zhou, W. Lin and Q. Tian, "Picking deep filter responses for finegrained image recognition," in *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 1134–114, 2016.

[26] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016.

[27] V. Kumar, V. Tripathi and B. Pant, "Content based fine-grained image retrieval using convolutional neural network," in *2020 7th Int. Conf. on Signal Processing and Integrated Networks (SPIN)*, Noida, India, pp. 1120–1125, 2020.

[28] F. Zhou and Y. Lin, "Fine-grained image classification by exploring bipartite-graph labels," in *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 1124–1133, 2016.

[29] X. Zheng, R. Ji, X. Sun, Y. Wu, F. Huang *et al.,* "Centralized ranking loss with weakly supervised localization for fine-grained object retrieval," in *Proc. of the 27th Int. Joint Conf. on Artificial Intelligence*, Stockholm, Sweden, pp. 1226–1233, 2018.

[30] B. Vasudeva, P. Deora, S. Bhattacharya, U. Pal and S. Chanda, "Loop: Looking for optimal hard negative embeddings for deep metric learning," in *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 10614–10623, 2021.

[31] H. Xuan, A. Stylianou and R. Pless, "Improved embeddings with easy positive triplet mining," in *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision*, Snowmass, CO, USA, pp. 2474–2482, 2020.

[32] E. Rodner, M. Simon, R. B. Fisher and J. Denzler, "Fine-grained recognition in the noisy wild: Sensitivity analysis of convolutional neural networks approaches," in *British Machine Vision Conf. 2016, BMVC*, York, UK, pp. 60.1–60.13, 2016.

[33] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev *et al.,* "The unreasonable effectiveness of noisy data for fine-grained recognition," in *European Conf. on Computer Vision*, Amsterdam, The Netherlands, pp. 301–320, 2016.

[34] M. Afifi and M. Brown, "What else can fool deep learning? addressing color constancy errors on deep neural network performance," in *2019 IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 243–252, 2019.

[35] J. Krause, M. Stark, J. Deng and L. Fei-Fei, "3D object representations for fine-grained categorization," in *2013 IEEE Int. Conf. on Computer Vision Workshops*, Sydney, NSW, Australia, pp. 554–561, 2013.

[36] P. Welinder, S. Branson, T. Mita, C. Wah and F. Schroff, "Caltech-ucsd birds 200, caltech-UCSD," *Technical Report*, California Institute of Technology, pp. 1–15, 2010.

[37] X. Zheng, R. Ji, X. Sun, B. Zhang, Y. Wu *et al.,* "Towards optimal fine grained retrieval via decorrelated centralized loss with normalize-scale layer," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 33, no. 1, Honolulu Hawaii USA, pp. 9291–9298, 2019.

[38] N. Komodakis and S. Gidaris, "Unsupervised representation learning by predicting image rotations," in *Int. Conf. on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

[39] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1734–1747, 2016.

[40] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conf. on Computer Vision*, Amsterdam, Netherlands, pp. 69–84, 2016.

[41] V. Kumar, V. Tripathi and B. Pant, "Unsupervised learning of visual representations via rotation and future frame prediction for video retrieval," in *Int. Conf. on Advances in Computing and Data Sciences (ICACDS 2021)*, Nashik, India, pp. 701–710, 2021.

[42] V. Kumar, "Unsupervised learning of spatio-temporal representation with multi-task learning for video retrieval," in *2022 National Conf. on Communications (NCC)*, Mumbai, India, pp. 118–123, 2022.

[43] V. Kumar, V. Tripathi and B. Pant, "Enhancing unsupervised video representation learning by temporal contrastive modelling using 2d CNN," in *Int. Conf. on Computer Vision and Image Processing (CVIP-2021)*, Rupnagar, India, pp. 494–503, 2022.

[44] V. Kumar, V. Tripathi, B. Pant, S. S. Alshamrani, A. Dumka *et al.,* "Hybrid spatiotemporal contrastive representation learning for content-based surgical video retrieval," *Electronics*, vol. 11, no. 9, pp. 1353, 2022.

[45] V. Kumar, V. Tripathi and B. Pant, "Learning unsupervised visual representations using 3d convolutional autoencoder with temporal contrastive modeling for video retrieval," *International Journal of Mathematical, Engineering and Management Sciences*, vol. 7, no. 2, pp. 272–287, 2022.