IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 11, 2020, accepted December 29, 2020, date of publication January 5, 2021, date of current version January 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3049305

Localization-Aware Adaptive Pairwise Margin Loss for Fine-Grained Image Recognition

TAEHUNG KIM[1], HOSEONG KIM[2], (Member, IEEE), AND HYERAN BYUN[1], (Member, IEEE)
[1]Department of Computer Science, Yonsei University, Seoul 03722, South Korea
[2]Agency for Defense Development, Daejeon 34186, South Korea

Corresponding author: Hyeran Byun (hrbyun@yonsei.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) under Grant 2019R1A2C2003760, and in part by the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) (Artificial Intelligence Graduate School Program, Yonsei University) under Grant 2020-0-01361.

ABSTRACT Fine-grained image recognition is a highly challenging problem due to subtle differences between images. There are many attempts to solve fine-grained image recognition problems using data augmentation, jointly optimizing deep metric learning. CutMix is one of the excellent data augmentation strategies which crops and merges to generate new images. However, it sometimes generates meaningless and obscured object images that degrade recognition performance. We propose a novel framework that solves the above problem and expands the CutMix leveraging localizing method. Also, we improve the recognition accuracy to joint optimizing with a pairwise margin loss using generated images from the improved CutMix. There are some images similar to the reference image among the generated images. They are generated by replacing similar parts from the reference image. Those generated images should not be located much farther than the margin value in embedding space because those generated images and a reference image have similar semantic meaning. However, the conventional margin loss can not consider those images which are located much farther than the margin. To solve this problem, we propose an additional margin loss to consider those generated images. The proposed framework consists of two stages: the part localization-aware CutMix and an adaptive pairwise margin loss. The proposed method achieves state-of-the-art performance on the CUB-200-2011, FGVC-Aircraft, Stanford Cars, and DeepFashion datasets. Furthermore, extensive experiments demonstrate that each stage improves the final performance.

INDEX TERMS Adaptive margin, deep neural networks, fine-grained image recognition, metric learning, image augmentation, image generation.

I. INTRODUCTION

Fine-grained image recognition is a highly challenging problem. In the coarse-level image recognition task, the trained network distinguishes between coarse-level objects (such as a tree and a car). Coarse-level image recognition has been studied for a long time and has now achieved a very low recognition error than a human being [1], [2]. However, the fine-grained image recognition task attempts to classify a specific level among 200 subcategories (for the CUB-200-2011 dataset), using only subtle differences between images. Another problem is that some images in the different subcategories look similar (having low inter-variance, as shown in Fig. 1(a)), and others within the same subcategory look different (having high intra-variance, as shown in Fig. 1(b)). Furthermore, Fig. 1(a) shows that the fine-grained image recognition is a challenging task for different subcategories, with subtle differences that can only be classified by bird experts.

There are many ways to solve these problems, e.g., augmentation and deep metric learning. Many data augmentation methods [3]–[5] have focused on the generalization of deep neural network models. Devries and Taylor [3] erased random areas to prevent a CNN from focusing on specific areas or on small areas to too great an extent. Zhang *et al.* [4] proposed Mixup, which mixes two images and two labels using interpolation. However, the output image of Mixup is unnatural. Yun *et al.* [5] addressed this weakness of Mixup and proposed CutMix, to improve the model generalization by replacing each image with random patch regions and

The associate editor coordinating the review of this manuscript and approving it for publication was Jin-Liang Wang.

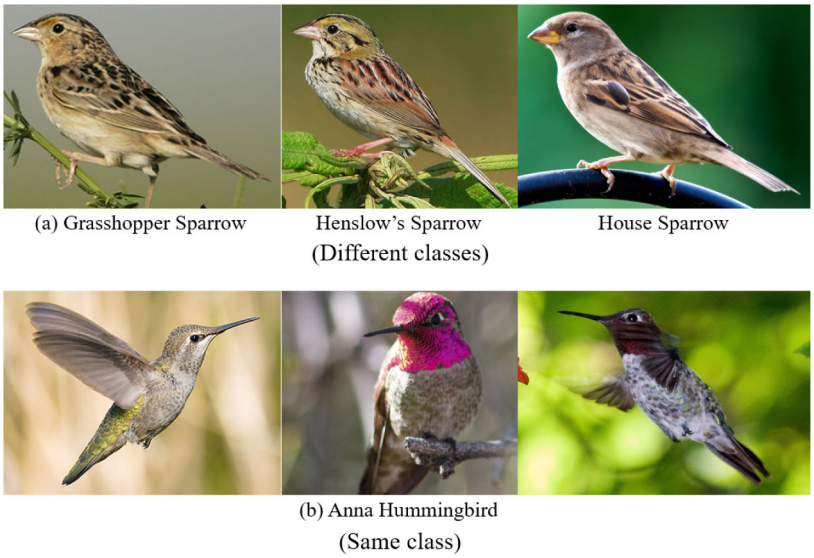

FIGURE 1. **(a) Small variance among different classes, (b) Large variance in the same class.**

labels. However, CutMix sometimes produces meaningless and obscured object images during image generation (as shown in Fig. 2(b)). In extreme cases of CutMix, most areas may be composed of the background. Since the object region plays a most important role in the image recognition task, it is inevitable that failure cases cause degrading recognition performance.

Deep metric learning fits well with the fine-grained image recognition task since it captures the semantic similarity between images. Previous researches [6], [7] have shown improved recognition accuracy with jointly optimizing classification loss with deep metric learning.

We propose a novel framework that solves the above problem and expands the CutMix to obtain better accuracy than existing jointly optimizing methods. The proposed method consists of two stages. In the first stage, we overcome the limitation of CutMix algorithm, combined with the existing part localization method [8], for fine-grained image recognition. The existing part localization method localizes three parts from the input image using a channel grouping network. To prevent failure cases, we replace parts images with other images' discriminative parts which correspond to the same part instead of cropping random areas when merging two different images.

In the conventional pairwise loss, one of the famous deep metric learning, the margin value plays a role in determining which different samples push away from the reference image. Since the margin value determines the samples to be pushed away, selecting margin is important and hard to determine. However, most deep metric learning methods select the margin value by empirical experiments. To improve this issue, in the second stage of our framework, we propose an adaptive margin using the distribution of the generated negative samples from the improved CutMix outputs.

Next, to improve better recognition accuracy than the previous jointly optimizing method, we leverage the generated images from the first stage with pairwise margin loss. As shown in Fig. 4, the parts localization-aware CutMix module generates many images. There are **some** similar images with the reference image among the generated images because they are generated by replacing similar parts from the reference image. The similarity leads to a similar semantic meaning with the reference image even though those generated images belong to a different class with the reference image. Those images should not be located much farther than the margin value in embedding space because the generated images and a reference image have similar semantic meaning. However, the conventional margin loss only pushes the samples out within the margin and does not consider those generated images which have similar semantic meaning with the reference image. To solve this problem, we propose an additional loss term in pairwise margin loss which pulls the samples to the reference image. (see the red boxes in Fig. 5).

To summarize, the contributions of this work are as follows:

- To overcome the disadvantages of the CutMix method and improve the fine-grained image recognition performance, we enhance CutMix with our proposed "part localization-aware CutMix module" using an existing weakly supervised part localization method.
- Beyond conventional pairwise loss, we propose an adaptive margin and additional pairwise loss term to improve fine-grained image recognition accuracy using generated images from the first stage.
- We achieve the state-of-the-art performance on four standard benchmark datasets (CUB-200-2011, FGVC-Aircraft, Stanford Cars, and DeepFashion), where the proposed framework consistently outperforms existing

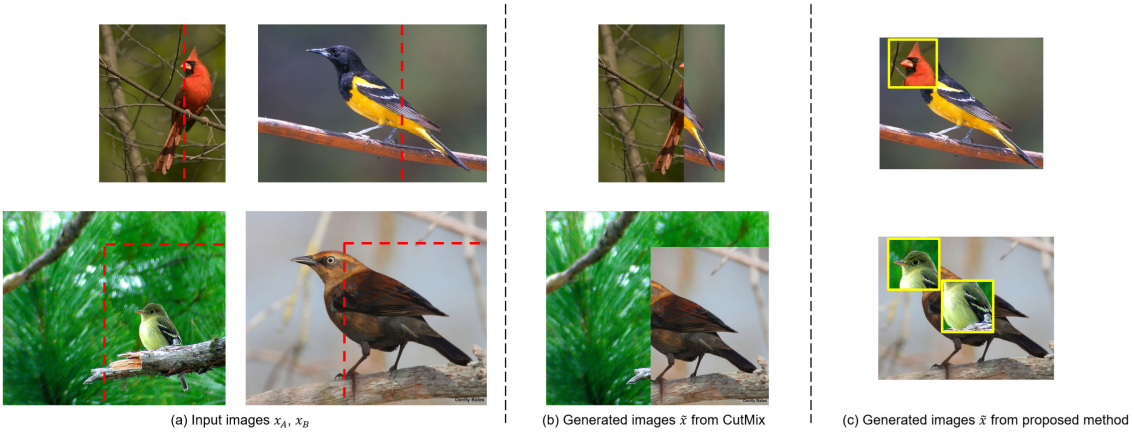

FIGURE 2. A visual comparison of CutMix and the proposed method: (a) Two images to be augmented by cropping and merging. (b) Examples of failure cases for CutMix. (c) Examples of the proposed method leveraging the existing part localization method.

methods. The importance and necessity of each objective function are also demonstrated.

In Section II, a brief review of related works on fine-grained image recognition is presented. Section III introduces the proposed method, "part localization-aware CutMix" and an "adaptive pairwise margin loss." The experimental results and discussion are presented in Section IV, and finally, Section V presents the conclusions of this paper.

II. RELATED WORKS

A. FINE-GRAINED IMAGE RECOGNITION

Deep learning has shown potential for feature learning and has achieved substantial progress on fine-grained image recognition tasks. Early works on fine-grained image recognition used a general coarse-level recognition approach. However, existing approaches using Convolutional Neural Networks (CNNs) are not apt for fine-grained image recognition due to subtle differences that are hard to classify. Previous works [9]–[14] have directly exploited parts annotation information to enhance the object recognition performance by using classifiers for every part. However, human-annotated training data is highly expensive to obtain since experts must concentrate on the data manually. Subsequent works based on weakly supervised learning [15]–[17] have attempted to find distinct parts without part annotation, using selective search [18] or a Region-Convolutional Neural Network (R-CNN) [19]. In particular, Peng *et al.* [20] have proposed a framework that selects the discriminative parts of images from a selective search result using "parts-object" constraints. So too, Zheng *et al.* [8] have induced part attention by activating feature channels and have trained networks utilizing this attention information.

Other approaches include the method of Chen *et al.* [21], which has proposed leveraging additional puzzle images with an adversarial loss to capture subtle local differences between the original images and the puzzle images. Zhuang *et al.* [22] have also proposed an attentive pairwise interaction framework, inspired by the human mechanism, to identify contrastive clues by comparing two images. Recently, Ji *et al.* [23] have proposed an attention convolutional binary neural tree that characterizes a coarse-to-fine hierarchical model.

On the other hand, we propose a novel framework that improves CutMix and leverages the discriminative parts of images obtained using the weakly supervised learning method, achieving state-of-the-art recognition performance.

B. DEEP METRIC LEARNING

Metric learning is designed to measure the similarity among samples while using the optimal distance metric for learning tasks. The seminal works in the field of deep metric learning mainly concern facial recognition, person re-identification, the ranking system, and fine-grained image recognition. Koch *et al.* [24] have proposed a pairwise objective function that trains a model with two shared networks by distinguishing between samples which are and samples which are not in the same class. So too, Schroff *et al.* [25] have proposed a triplet objective function that learns a network using the relationship between three samples (a reference, a positive, and a negative). Wang *et al.* [26] proposed a feature embedding method using the semantic similarity between images after carrying out patch clustering.

Other approaches [6], [7], [27] have solved the subtle difference problems between the images using deep metric learning by capturing the semantic similarity. Zhang *et al.* [6] has shown that combining classification loss with triplet loss had better recognition accuracy than using classification loss alone. Cui *et al.* [7] have incorporated humans in the loop for the training step, obtaining additional hard negative samples from the web crawling. Sun *et al.* [27] have also proposed a multi-attention multi-class constraint with a squeeze-excitation module. The constraints used in the literature have produced multiple different negative sample groups. It performs better recognition accuracy than using one negative sample.

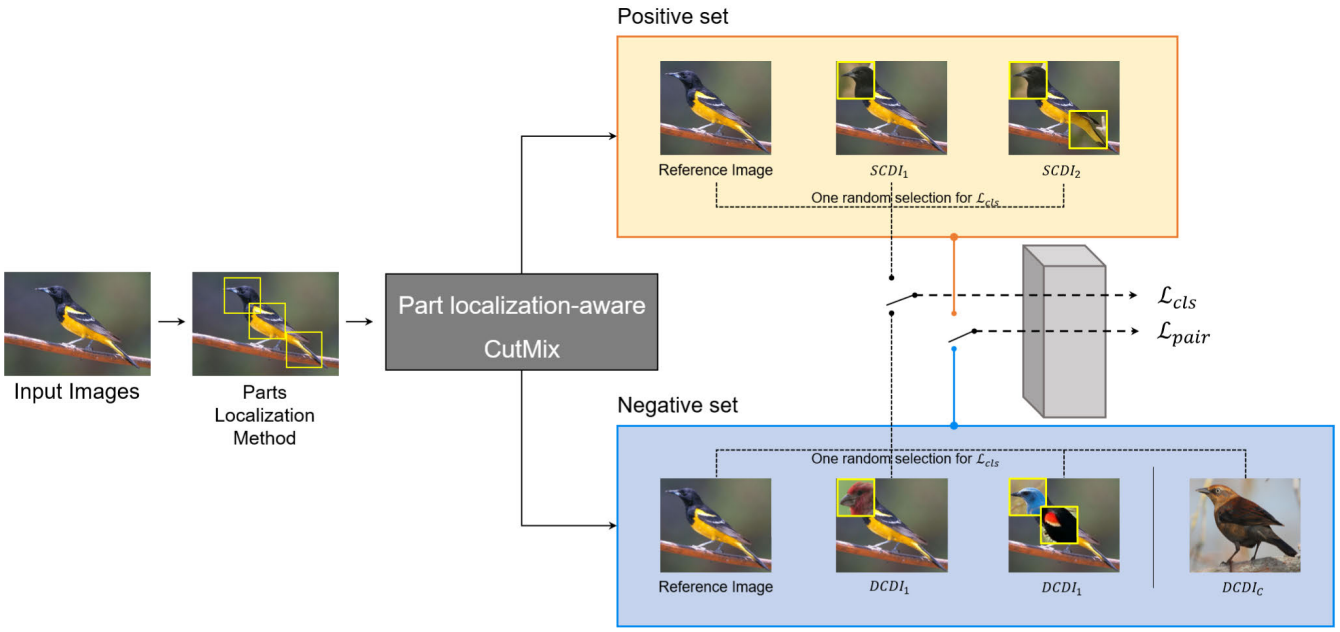

FIGURE 3. An overview of the proposed framework. We obtain the various images from the part localization-aware CutMix module using the existing localization method. The generated images are included in a positive set or negative set, depending on the replaced images. The indicator value selects each set for training to jointly optimizing classification loss and proposed adaptive pairwise loss.

We propose an adaptive pairwise margin loss that considers some generated images which have similar semantic meaning with the reference image not previously considered, using conventional pairwise margin loss with proposed an additional constraint. The results demonstrate state-of-the-art performance by jointly optimizing the classification loss and the proposed adaptive pairwise margin loss.

III. PROPOSED METHOD

This paper proposes a new adaptive pairwise margin loss, improving on the CutMix method, as shown in Fig. 3. The proposed method utilizes the existing part localization method with CutMix, to improve the fine-grained image recognition task performance. The overall framework consists of two stages: part localization-aware CutMix and an adaptive pairwise margin loss. The part localization-aware CutMix stage improves on CutMix by leveraging the existing part localization method. In the adaptive pairwise margin loss stage, a novel loss is proposed which improves the conventional margin loss using the augmented images from the first stage. We show the final result by jointly optimizing the proposed adaptive pairwise margin loss with classification loss in the last step.

A. PART LOCALIZATION-AWARE CutMix

CutMix is an augmentation strategy proposed by Yun *et al.* [5], with the goal of generating new images $(\tilde{x}, \tilde{y})$ for training. The corresponding algorithm is described as follows:

$$\tilde{x} = M \odot x_A + (1 - M) \odot x_B, \tag{1}$$

$$\tilde{y} = \theta y_A + (1 - \theta) y_B, \tag{2}$$

$$M \in \{0, 1\}^{W \times H}, \quad 0 < \theta < 1, \tag{3}$$

where x_A and x_B are training images, and y_A and y_B are their labels, respectively. In the foregoing, $\tilde{x}$ and $\tilde{y}$ are generated images, $\odot$ is element-wise multiplication, and M denotes a binary mask indicating where to drop out and where to fill in using two images. CutMix decides the bounding box coordinates and size of the cropping area by uniform sampling with a random value θ.

However, as shown in Fig. 2(a) and (b), randomly selecting box coordinates using the CutMix algorithm and cropping the image often generates inappropriate images. In some cases, important parts of the object are partly or fully occluded, and some parts of the object may not even be visible. In extreme cases of CutMix, most areas may be composed of the background. The inappropriate images influence the training step negatively, causing a reduction in accuracy.

We propose a method that generates various images $(\tilde{x})$ from one reference image by leveraging the methods from existing research [8] to solve this problem. As shown in Fig. 4, based on one reference image, many images are generated, along with other images. Three images (x_A, x_B, and x_C) applied object localization method CAM [28] to remove unnecessary background areas before processing. Three parts of the images (x_A, x_B, and x_C) are obtained (M_{part1}, M_{part2}, and M_{part3}), and information about these parts (the size and coordinates) are derived using the method from the existing work [8] (see the yellow and white boxes in Fig. 4). We change parts of the reference image compared with other images, replacing images in the same part deterministically,

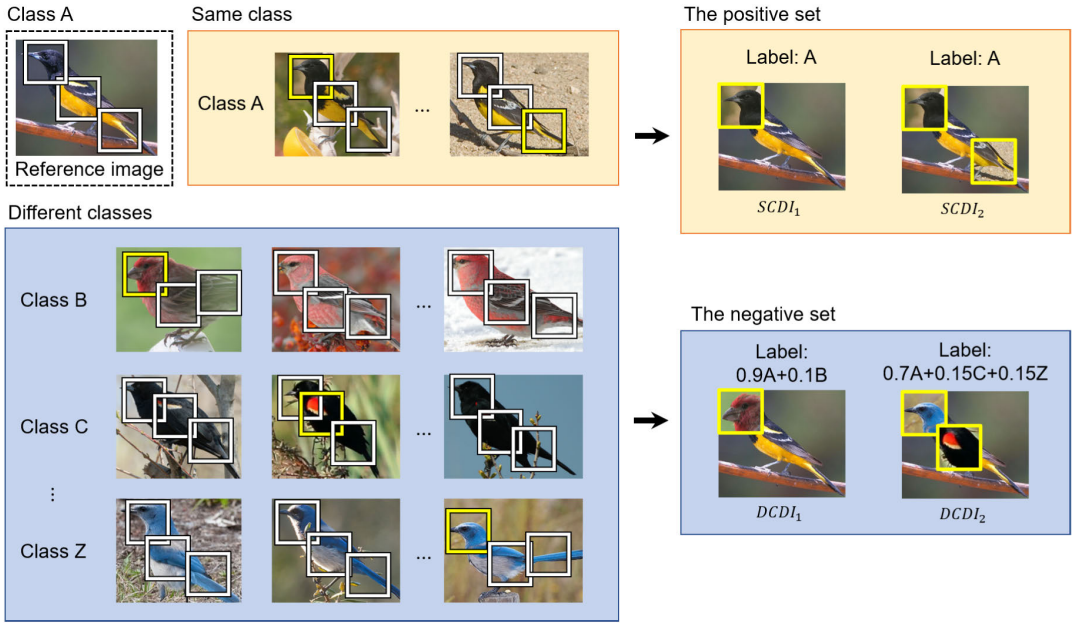

FIGURE 4. Parts localization-aware CutMix module generates images with the reference image and several different class images. Each image has three discriminative parts (white box) from the existing method. The white box is a candidate for replacing. The reference image is replaced with one or two parts of the three candidates. When the reference image is replaced with images from the same class, it becomes a positive set. On the other hand, when the reference image is replaced with images from the different classes, it belongs to the negative set. The label of the generated image is determined in proportion to the area of the image.

rather than randomly. There are two types of generated images, as follows. Same Class Different Image (*SCDI*) involves one or two parts in the reference image being replaced with corresponding parts from the same image class. However, Different Class Different Image (*DCDI*) involves one or two parts in the reference image being replaced by images from the different classes. CutMix is expanded and modified, with the expanding and combining operation being defined as follows:

$$\tilde{x} = M_{part1} \odot x_A + M_{part2} \odot x_B + M_{part3} \odot x_C, \quad (4)$$

$$\tilde{y} = B_A y_A + B_B y_B + B_C y_C, \quad (5)$$

$$s.t\ B_A + B_B + B_C = 1. \quad (6)$$

Each label (y_A, y_B, and y_C) is given using the ratio of the images x_A, x_B, and x_C to the reference image. Since the reference image is replaced with a discriminative part, instead of random crop, failure cases do not occur in the generated images when using the method described above (as shown in Fig. 2(c)). The proposed method goes on to use the generated images by separating them into positive and negative sets, as characterized in the following sections.

B. ADAPTIVE PAIRWISE MARGIN LOSS

Deep metric learning includes a pairwise loss, a triplet loss, and a quadruplet loss. Unlike conventional classification loss, deep metric learning is widely used in ranking systems, face recognition, and person re-identification, due to having characteristics that capture semantic similarity adequately. The conventional pairwise margin loss L_{conv} is characterized by training either two images from the same class (r, p) or images from different classes (n), according to an indicator, as follows:

$$L_{conv} = \begin{cases} d(f(r), f(p)), & if\ Positive\ Set, \\ max(0, m - d(f(r), f(n))), & if\ Negative\ Set, \end{cases} \quad (7)$$

where the r, p, and n are the reference image, the positive image, and the negative image, respectively. $f(\cdot)$ is the feature vector. $d(\cdot)$ is the Euclidean distance between two feature vectors in the positive set. m is the margin value for the negative set. Each loss term is improved and expanded using a positive set and a negative set for the fine-grained image recognition task. We also propose an adaptive margin and an additional pairwise loss in the negative set for improving manual margin and considering the samples which have similar semantic meaning not considered in the conventional pairwise loss, respectively.

1) THE POSITIVE SET

As discussed in Section III-A, this involves enhancing CutMix, to crop and merge parts from two portions of reference images to generate a new image. As shown in Fig. 4, the new image ($SCDI_1$) is obtained by randomly changing one discriminative part of other image which is the same class as the reference image. Another image ($SCDI_2$) is also obtained by changing two such parts. Since the two images consist of the same class parts, the generated images can be considered the same class as the reference image. Furthermore, since the

generated two images have the same semantic meaning as the reference image, the proposed loss is designed such that the three samples are embedded in one point in the embedding space. Any two images (i, j) are defined in the three-sample relationship as $sim(i, j)$. From the three images, three similar relationships are defined and considered in the loss term: $sim(reference\ image, SCDI_1)$, $sim(reference\ image, SCDI_2)$, and $sim(SCDI_1, SCDI_2)$. Here, $sim(i, j)$ denotes the Euclidean distance between the two samples i and j, as follows:

$$sim(i, j) = \|f(i) - f(j)\|_2 . \tag{8}$$

We define the positive relationship between the three images (*reference image*, $SCDI_1$, and $SCDI_2$) in this stage as follows:

$$\begin{aligned} positive\ relationship = {} & sim(ref, SCDI_1) \\ & + sim(ref, SCDI_2) \\ & + sim(SCDI_1, SCDI_2). \end{aligned} \tag{9}$$

2) THE NEGATIVE SET

The negative set consists of four images. The first stage generates two additional images from the one sample-based reference and one completely negative image ($DCDI_C$) (from the different class). Two images from different classes are selected randomly for ($DCDI_1$ and $DCDI_2$). One of the three parts of the selected image is used to generate an image ($DCDI_1$), which is replaced with the same part of the reference image. $DCDI_2$ is generated by randomly selecting two of the three parts from each image. The negative image ($DCDI_C$) is a random image in different classes from the reference image.

Equation (7) shows that when the input is from the negative set, the conventional pairwise margin loss selects only for samples within a certain margin, and the network is updated to minimize the loss. The margin value plays an important role in deep metric learning. Specifically, if the margin value is too large, the network will overfit, and if the margin value is too small, the computational cost of the training process will increase. In this case, the margin value m has the disadvantage of being chosen manually. Since the margin value depends on the samples' distribution, m is found experimentally for each new data set or new network. An existing work [29] has proposed a similar adaptive margin loss, this being valid only for triplet or quadruplet loss and not applicable to tasks comparing pairs in the "negative set." Hence, the proposed margin value improves and rectifies the adaptive margin loss, considering the relationship between the samples in the negative set.

There are four pairs in a negative set. Three pairs are the three relationships between the samples (reference, $DCDI_1$, and $DCDI_2$) in a similar way to the positive set. The other pair consists of a reference image and a negative image ($DCDI_C$). To determine the adaptive margin values of each pair, their numerous samples (n_i) are required, barring the reference image.

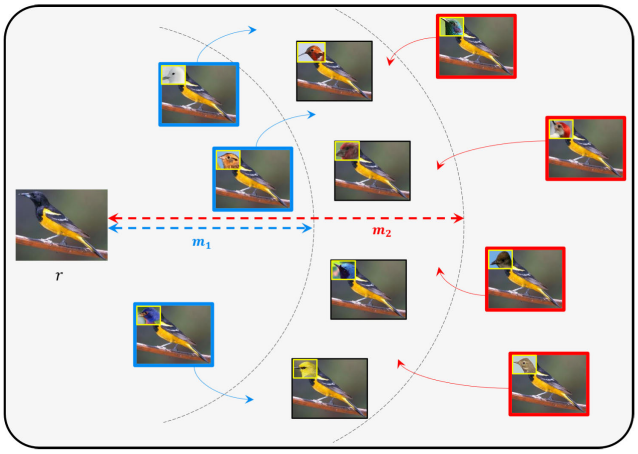

FIGURE 5. Example of the $dissim(ref, DCDI_1)$ with adaptive margin m_1 and m_2. The conventional margin loss only pushes away the samples located within the m_1 margin (blue boxes). On the other hand, our proposed additional pairwise loss pulls the samples located much farther (red boxes) than the m_1 margin into the m_2 margin.

In each pair, we define m_1 using the distribution of n_i images, without manually setting the margin value. In this process, many $DCDI$ images (n_i) are generated by randomly replacing the merged part with other images. In the reference image and $DCDI_1$ case, numerous samples (n_i) are generated by replacing one part of the random images from different classes. The m_1 value is defined as follows, using the distribution of the n_i samples:

$$m_1 = \frac{1}{N} \sum_{i=1} \|f(r) - f(n_i)\|_2 , \tag{10}$$

where the N is the number of the n_i. m_1 value is updated every epoch in the training step.

As shown is Fig. 5, there are some similar images with the reference image among the generated images because they are generated by replacing similar parts from the reference image or replacing only a tiny proportion of the image. The similarity leads to a similar semantic meaning with the reference image even though those generated images belong to a different class with the reference image. Those images should not be located much farther than the margin value in embedding space because the generated images and a reference image have similar semantic meaning. However, the conventional margin loss only pushes the samples out within the margin and does not consider those generated images which have similar semantic meaning with the reference image. Therefore, the samples not reflected in the conventional margin value are considered using an additional loss term as follows:

$$max\ (0, \|f(r) - f(n)\| - m_2) . \tag{11}$$

Since most samples are located near the centroid of the $DCDI$ images in the embedding space, we set the gap σ from the centriod. We define a second margin value, m_2, as follows:

$$m_2 = m_1 + \alpha\sigma. \tag{12}$$

The dissimilar term to which the proposed additional loss with the conventional pairwise loss is applied is as follows:

$$dissim(r, n) = max\,(0, m_1 - \|f(r) - f(n)\|) + max\,(0, \|f(r) - f(n)\| - m_2)\,. \quad (13)$$

The three dissimilar terms and one conventional loss term are defined by one negative relationship as follows:

$$\begin{aligned} &negative\ relationship \\ &= dissim(ref, DCDI_1) + dissim(ref, DCDI_2) \\ &\quad + dissim(DCDI_1, DCDI_2) \\ &\quad + max\,(0, m_1 - \|f(ref) - f(DCDI_C)\|)\,. \end{aligned} \quad (14)$$

A positive set and a negative set are characterized as follows:

$$L_{pair} = \begin{cases} positive\ relationship,\ if\ Positive\ set, \\ negative\ relationship,\ if\ Negative\ set. \end{cases} \quad (15)$$

The proposed loss L_{pair} can be written equivalently as follows:

$$L_{pair} = S(positive\ relationship) + (1 - S)(negative\ relationship), \quad (16)$$

where S is an indicator equal to 0 for a negative set and 1 for a positive set. Our proposed method jointly optimizes the classification loss and the pairwise margin loss as follows:

$$L = \lambda L_{cls} + (1 - \lambda) L_{pair}. \quad (17)$$

IV. EXPERIMENTS

A. DATASETS

1) CUB-200-2011 [53]

It is the most popular dataset for fine-grained image recognition. CUB-200-2011 contains many bird species all around the world. The dataset is approximately twice as large as the CUB-200 dataset. The number of images in the dataset is 11,788, with 200 different subcategories. The dataset is split into 5,994 images for training and 5,794 images for testing. The annotation of the dataset consists of bounding box information about the object and 312 attributes of each bird (including the wing color and the length of the beak).

2) STANFORD CARS [54]

It contains 16,185 images of cars. The dataset is divided into 8,144 images for training and 8,041 for tests. Each subcategory consists of 24–84 images for training and 24–84 images for testing. This dataset has one bounding box and one label.

3) FGVC-AIRCRAFT [55]

It is an aircraft image dataset with 102 subcategories. The dataset is composed of 10,200 images and is equally split into training, testing, and validation. Each subset has 33–34 images. All images are annotated with the model, family, variant, and manufacturer information.

4) DeepFashion [48]

It is used for image retrieval, recognition, and detection. It contains many kinds of clothes (such as T-shirts, dresses). We experimented with the first subset of the DeepFashion dataset, the "Category and Attribute Prediction Benchmark" dataset. DeepFashion consists of 289,222 images in 50 subcategories, all of which are annotated by a bounding box and information on the type of clothing.

B. IMPLEMENTATION

All experiments in this paper were trained with and tested using a computer with 192GM RAM, Cascade Lake 24C processors of 2.5GHz, and 4x T4 NVIDIA GPUs. For a fair comparison, the framework was designed to depend on VGG-16 and ResNet-50. The SGD optimizer was established, and the initial learning rate was set to 0.001, decaying by 0.09 every 50 epochs. The α values are set to 1.0, 1.2, 1.0, and 1.0 for CUB-200-2011, Stanford Cars, FGVC-Aircraft, and DeepFashion, respectively. Since classification loss involves more information than pairwise loss, the weightings were set as follows: $\lambda = 0.75$, 0.8, 0.8, and 0.75 for CUB-200-2011, Stanford Cars, FGVC-Aircraft, and DeepFashion, respectively.

C. COMPARISON WITH STATE-OF-THE-ART METHOD

The "Backbone" column in tables denotes which CNN model was used as the backbone network. The results of fine-grained image recognition tasks with the CUB-200-2011, FGVC-Aircraft, Stanford Cars, and DeepFashion datasets are described in Tables 1, 2, 3, and 4, respectively. The columns in every table show the method, the "Backbone," and the accuracy of each method. For a fair comparison, the results were compared with those of studies which used VGG-16 and ResNet-50. Additionally, all of the results were obtained fairly, without external information such as annotations or a bounding box. As shown in Table 1, The proposed framework outperforms MGE-CNN [43], which includes many experts' input and a gating network, by 0.73% on the same ResNet-50 network. The result of the proposed method have been confirmed to be 1.13% higher than the second-highest result, ISQRT-COV [36]. The proposed framework also outperforms GSFL-Net [37], which shares significant features of interest and divides existing classes into groups, by 0.94% on the VGG-16 backbone network. The second-highest performance for the VGG-16 backbone network is 87.2%, exhibited by ISQRT-COV [36], which utilizes sandwiching Newton-Schulz iteration to relieve the computational burden of an improved CNN. The result of the proposed method is 1.34% higher than that for ISQRT-COV [36]. The results for the proposed method show state-of-the-art performance not only on the CUB-200-2011 dataset but also on the Stanford Cars, FGVC-Aircraft, and DeepFashion datasets. The results for each dataset are 0.92%, 1.43%, and 5.73% higher than the second-highest results, respectively, on ResNet-50. The VGG-16 results are

TABLE 1. Comparison of our approach to recent results on CUB-200-2011.

Method	Backbone	Acc(%)
Bilinear-CNN [30]	VGGNet	84.1
RA-CNN [31]	VGG-19	85.3
Improved B-CNN [32]	VGG-16	85.8
OPAM [20]	VGG-16	85.83
Kernel-Pooling [33]	VGG-16	86.2
Refined-CNN [34]	VGG-16	86.4
MA-CNN [8]	VGG-19	86.5
DFL-CNN(2-scale) [35]	VGG-16	86.7
DCL [21]	VGG-16	86.9
iSQRT-COV [36]	VGG-16	87.2
GSFL-Net([33]based) [37]	VGG-16	87.60
Ours	VGG-16	**88.54**
Kernel-Pooling [33]	ResNet-50	84.7
MAMC [27]	ResNet-101	86.5
HBPASM [38]	ResNet-34	86.8
DFL-CNN(1-scale) [35]	ResNet-50	87.4
DBTNet-50 [39]	ResNet-50	87.5
Cross-X [40]	ResNet-50	87.7
DCL [21]	ResNet-50	87.8
TASN [41]	ResNet-50	87.9
iSQRT-COV [36]	ResNet-50	88.1
S3N [42]	ResNet-50	88.5
MGE-CNN [43]	ResNet-50	88.5
MGE-CNN [43]	ResNet-101	**89.4**
Ours (w/ Bounding box)	ResNet-50	89.35
Ours	ResNet-50	89.23

TABLE 2. Comparison of our approach to recent results on Stanford Cars.

Method	Backbone	Acc(%)
Bilinear-CNN [30]	VGGNet	91.3
Improved B-CNN [32]	VGG-16	92.0
OPAM [20]	VGG-16	92.19
Kernel-Pooling [33]	VGG-16	92.4
Refined-CNN [34]	VGG-16	92.4
RA-CNN [31]	VGG-19	92.5
iSQRT-COV [36]	VGG-16	92.5
MA-CNN [8]	VGG-19	92.8
TASN [41]	VGG-19	93.2
DFL-CNN(2-scale) [35]	VGG-16	93.8
GSFL-Net([33]based) [37]	VGG-16	93.92
DCL [21]	VGG-16	94.1
Ours	VGG-16	**94.27**
iSQRT-COV [36]	ResNet-50	92.8
MAMC [27]	ResNet-101	93.0
DFL-CNN(1-scale) [35]	ResNet-50	93.1
MGE-CNN [43]	ResNet-101	93.6
TASN [41]	ResNet-50	93.8
HBPASM [38]	ResNet-34	93.8
MaxEnt [44]	ResNet-50	93.85
MGE-CNN [43]	ResNet-50	93.9
DBTNet-50 [39]	ResNet-50	94.1
DCL [21]	ResNet-50	94.5
Cross-X [40]	ResNet-50	94.6
Ours	ResNet-50	**95.52**

TABLE 3. Comparison of our approach to recent results on FGVC-Aircraft.

Method	Backbone	Acc(%)
Bilinear-CNN [30]	VGGNet	84.1
Kernel-Pooling [33]	VGG-16	86.9
Refined-CNN [34]	VGG-16	87.7
RA-CNN [31]	VGG-19	88.2
HIHCA [45]	VGG-16	88.3
Improved B-CNN [32]	VGG-16	88.5
GSFL-Net([30]based) [37]	VGG-16	89.26
MA-CNN [8]	VGG-19	89.9
iSQRT-COV [36]	VGG-16	90.0
DFL-CNN(1-scale) [35]	VGG-16	91.1
DCL [21]	VGG-16	91.2
DFL-CNN(2-scale) [35]	VGG-16	92.0
Ours	VGG-16	**93.08**
Kernel-Pooling [33]	ResNet-50	85.7
iSQRT-COV [36]	ResNet-50	90.0
DBTNet-50 [39]	ResNet-50	91.2
HBPASM [38]	ResNet-34	91.3
DFL-CNN(1-scale) [35]	ResNet-50	91.7
Cross-X [40]	SENet-50	92.7
S3N [42]	ResNet-50	92.8
DCL [21]	ResNet-50	93.0
Ours	ResNet-50	**94.43**

0.17%, 1.08%, and 0.42% higher than the second-highest result, respectively.

D. ABLATION STUDY

1) THE EFFECT OF OBJECTIVE FUNCTIONS

This paper has studied how the adaptive margin and the proposed additional loss term affect the entire performance. A variety of experiments were designed and evaluations were performed on the CUB-200-2011 dataset. As shown in Table 5, results were obtained for the four cases: whether the adaptive margin was applied or not and whether the proposed loss was applied or not. For the cases where the adaptive margin was applied or not, there were differences of 1.14% and 0.31% respectively between the conventional pairwise margin loss and the proposed pairwise margin loss. The margin value was set manually in the cases ''without adaptive margin.'' When the proposed pairwise loss term was added to the conventional pairwise margin loss, the results for the cases of the applied adaptive margin being applied or not showed differences of 2.17% and 1.34%, respectively. Each adaptive margin and additional pairwise margin affects the entire recognition performance. The results also show that the proposed additional pairwise margin loss is more critical than the adaptive margin in determining the overall recognition performance.

E. DISCUSSION

1) THE EFFECT OF THE NUMBER OF IMAGES OF m_1

Several images were generated based on the reference sample by changing one or two image parts in the first stage. The next stage provides the m_1 value based on the various images generated in the first stage, and the m_2 value is dependent on the m_1 value. In the second stage, the number of images is set to 100–400. As shown in Table 6, as the number of images increases, the performance increases significantly. The optimal result is achieved for 250 images on CUB-200-2011, but improvements in the performance become less prominent when more than 400 extra images are used.

2) RATIO OF THE POSITIVE AND NEGATIVE SETS

The default ratio of the positive set to the negative set is fixed at 1:1. Table 7 shows the recognition accuracy on

TABLE 4. Comparison of our approach to recent results on Deepfashion.

Method	Backbone	Top3(%)	Top5(%)
WTBI [46]	Custom	43.73	66.26
DARN [47]	Custom	59.48	79.58
FashionNet [48]	Custom	82.58	90.17
FAFS [49]	VGG-16	86.72	92.51
AFGN [50]	VGG-16	90.99	95.78
FAN [51]	VGG-16	91.16	96.12
Ours	VGG-16	**91.25**	**96.54**
LWAD [52]	ResNet-50	86.30	92.8
Ours	ResNet-50	**92.03**	**96.75**

TABLE 5. Ablation performance on each loss function on CUB-200-2011.

Loss term	Acc(%)
Conventional pairwise loss w/o adaptive margin	86.75
Conventional pairwise loss w/ adaptive margin	87.89
Proposed pairwise loss w/o adaptive margin	88.92
Proposed pairwise loss w/ adaptive margin	89.23

TABLE 6. Comparison of our approach to recent on different of the number of samples for mean value m_1.

The number of	100	200	250	300	400
Acc(%)	88.54	89.12	89.23	89.1	89.16

TABLE 7. Comparison of our approach to the different ratio between positive pairs and negative pairs.

Ratio	3:1	2:1	1:1	1:2	1:3	0:1	1:0
Acc(%)	88.12	88.79	89.23	88.92	88.60	86.73	86.05

CUB-200-2011, with various ratio settings. The results show that the negative set is much more significant than the positive set, since the results for the 1:2 and 1:3 ratios are superior to those for the 2:1 and 3:1 ratios, respectively.

3) PART LOCALIZATION

The proposed framework has a double-edged sword effect and, as detailed above, leverages an existing approach [8]. The results are influenced by the number of parts and the part localizing quality used in an existing approach. If a superior part localization approach is implemented, the proposed framework shows superior results.

4) OBJECT LOCALIZATION

As shown in Fig. 6, some failure cases (namely, occlusion) occur in the object localization method [28]. Localizing failures cause the failure of the images generated in the first stage of the experiment. As shown in Table 1, additional experiments were carried out on the CUB-200-2011 dataset, using bounding box (BB) information instead of CAM [28]. Since there were no localizing failures, the last results were enhanced by 0.12%. These results give the upper bounds of the performance. If object localizing approaches superior to CAM [28] are implemented in the proposed framework, the performance may increase yet further.

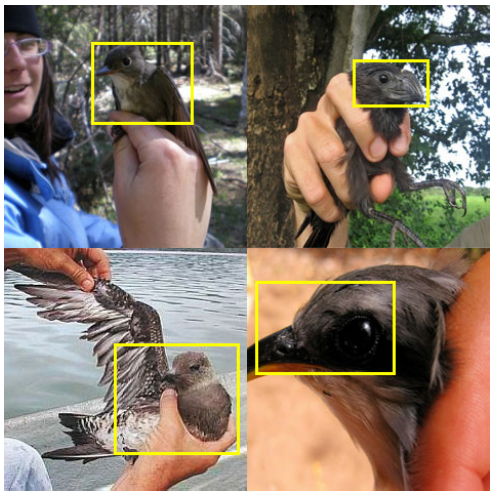

FIGURE 6. There are some failure cases of localizing in CUB-200-2011.

V. CONCLUSION

There have been many methods to improve image recognition performance (such as Cutmix and optimizing deep metric learning). Cutmix, one of the augmentation methods, generated new images by random cropping and merging. However, some generated images were meaningless images that degraded recognition performance. Because conventional pairwise loss updated the network using the samples only within the fixed margin value m, that loss could not consider the images generated by improved Cutmix. To overcome these limitations, we proposed an improved Cutmix method and localization-aware adaptive pairwise margin loss. The first stage of the proposed method improved CutMix by leveraging an existing part localization method and generating images. In the second stage, a novel adaptive pairwise margin loss was proposed, using the generated images from the first stage. The proposed additional loss considers the samples which have a similar semantic meaning with the reference image and are located much farther than the margin not considered in the existing pairwise margin loss. The limitation was that the first and second stage were affected by the result of the parts localization and first stage, respectively. Therefore, our future work will focus on improving the end-to-end framework using the part attention method to avoid dependencies on the results of each stage. In this paper, extensive experiments were conducted on the CUB-200-2011, Stanford Cars, FGVC-Aircraft, and DeepFashion datasets, and state-of-the-art performance has been achieved using the proposed framework. Additionally, the need for each proposed loss and each stage of the ablation study has been verified.

REFERENCES

[1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran, 2012, pp. 1097–1105.

[3] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*. [Online]. Available: http://arxiv.org/abs/1708.04552

[4] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. 6th Int. Conf. Learn. Represent., ICLR*, Vancouver, BC, Canada, Apr. 2018, pp. 1–13. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[5] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031, doi: 10.1109/ICCV.2019.00612.

[6] X. Zhang, F. Zhou, Y. Lin, and S. Zhang, "Embedding label structures for fine-grained feature representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1114–1123.

[7] Y. Cui, F. Zhou, Y. Lin, and S. Belongie, "Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1153–1162.

[8] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5219–5227.

[9] R. Farrell, O. Oza, N. Zhang, V. I. Morariu, T. Darrell, and L. S. Davis, "Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 161–168, doi: 10.1109/ICCV.2011.6126238.

[10] T. Berg and P. N. Belhumeur, "POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 955–962.

[11] L. Xie, Q. Tian, R. Hong, S. Yan, and B. Zhang, "Hierarchical part matching for fine-grained visual categorization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1641–1648.

[12] S. Branson, G. Van Horn, P. Perona, and S. Belongie, "Improved bird species recognition using pose normalized deep convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, 2014, p. 7.

[13] S. Branson, G. Van Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," 2014, *arXiv:1406.2952*. [Online]. Available: http://arxiv.org/abs/1406.2952

[14] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "PANDA: Pose aligned networks for deep attribute modeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1637–1644, doi: 10.1109/CVPR.2014.212.

[15] M. Simon and E. Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1143–1151.

[16] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, "Picking deep filter responses for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1134–1142.

[17] Y. Zhang, X.-S. Wei, J. Wu, J. Cai, J. Lu, V.-A. Nguyen, and M. N. Do, "Weakly supervised fine-grained categorization with part-based image representation," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1713–1725, Apr. 2016.

[18] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[20] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1487–1500, Mar. 2018.

[21] Y. Chen, Y. Bai, W. Zhang, and T. Mei, "Destruction and construction learning for fine-grained image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5152–5161.

[22] P. Zhuang, Y. Wang, and Y. Qiao, "Learning attentive pairwise interaction for fine-grained classification," in *Proc. 34th AAAI Conf. Artif. Intell., AAAI, 32nd Innov. Appl. Artif. Intell. Conf., IAAI, 10th AAAI Symp. Educ. Adv. Artif. Intell., EAAI*. New York, NY, USA: AAAI Press, Feb. 2020, pp. 13130–13137. [Online]. Available: https://aaai.org/ojs/index.php/AAAI/article/view/7016

[23] R. Ji, L. Wen, L. Zhang, D. Du, Y. Wu, C. Zhao, X. Liu, and F. Huang, "Attention convolutional binary neural tree for fine-grained visual categorization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10465–10474, doi: 10.1109/CVPR42600.2020.01048.

[24] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, Jul. 2015. [Online]. Available: https://www.bibsonomy.org/bibtex/26f83b8c4cf316e77e6f6ce1e97411b30/bsc

[25] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[26] Y. Wang, J. Choi, V. I. Morariu, and L. S. Davis, "Mining discriminative triplets of patches for fine-grained classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1163–1172, doi: 10.1109/CVPR.2016.131.

[27] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-attention multi-class constraint for fine-grained image recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 805–821.

[28] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[29] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1320–1329, doi: 10.1109/CVPR.2017.145.

[30] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1449–1457.

[31] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4476–4484.

[32] T.-Y. Lin and S. Maji, "Improved bilinear pooling with CNNs," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–12. [Online]. Available: https://www.dropbox.com/s/fc6qtzvno7ln684/0395.pdf?dl=1

[33] Y. Cui, F. Zhou, J. Wang, X. Liu, Y. Lin, and S. Belongie, "Kernel pooling for convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3049–3058.

[34] W. Zhang, J. Yan, W. Shi, T. Feng, and D. Deng, "Refining deep convolutional features for improving fine-grained image recognition," *EURASIP J. Image Video Process.*, vol. 2017, no. 1, p. 27, Dec. 2017, doi: 10.1186/s13640-017-0176-3.

[35] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a CNN for fine-grained recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4148–4157.

[36] P. Li, J. Xie, Q. Wang, and Z. Gao, "Towards faster training of global covariance pooling networks by iterative matrix square root normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 947–955. [Online]. Available: http://openaccess.thecvf.com

[37] X. Li and V. Monga, "Group based deep shared feature learning for fine-grained image classification," in *Proc. 30th Brit. Mach. Vis. Conf. BMVC*, Cardiff, U.K.: BMVA Press, Sep. 2019, p. 143. [Online]. Available: https://bmvc2019.org/wp-content/uploads/papers/0885-paper.pdf

[38] M. Tan, G. Wang, J. Zhou, Z. Peng, and M. Zheng, "Fine-grained classification via hierarchical bilinear pooling with aggregated slack mask," *IEEE Access*, vol. 7, pp. 117944–117953, 2019, doi: 10.1109/ACCESS.2019.2936118.

[39] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Learning deep bilinear transformation for fine-grained image representation," in *Proc. 32nd Annu. Conf. Adv. Neural Inf. Process. Syst. (NeurIPS)*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds. Vancouver, BC, Canada: Curran, Dec. 2019, pp. 4279–4288. [Online]. Available: http://papers.nips.cc/paper/8680-learning-deep-bilinear-transformation-for-fine-grained-image-representation

[40] W. Luo, X. Yang, X. Mo, Y. Lu, L. Davis, J. Li, J. Yang, and S.-N. Lim, "Cross-X learning for fine-grained visual categorization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8241–8250.
[41] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5007–5016.
[42] Y. Ding, Y. Zhou, Y. Zhu, Q. Ye, and J. Jiao, "Selective sparse sampling for fine-grained image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6598–6607.
[43] L. Zhang, S. Huang, W. Liu, and D. Tao, "Learning a mixture of granularity-specific experts for fine-grained categorization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8330–8339.
[44] A. Dubey, O. Gupta, R. Raskar, and N. Naik, "Maximum-entropy fine-grained classification," in *Proc. NeurIPS*, 2018, pp. 637–647.
[45] S. Cai, W. Zuo, and L. Zhang, "Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 511–520.
[46] H. Chen, A. C. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proc. ECCV*, 2012, pp. 609–623.
[47] J. Huang, R. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1062–1070.
[48] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1096–1104.
[49] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. Feris, "Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1131–1140.
[50] W. Wang, W. Wang, Y. Xu, J. Shen, and S.-C. Zhu, "Attentive fashion grammar network for fashion landmark detection and clothing category classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4271–4280.
[51] J. Liu and H. Lu, "Deep fashion analysis with feature map upsampling and landmark-driven attention," in *ECCV Workshops*, Sep. 2018, pp. 30–36.
[52] C. Corbiere, H. Ben-Younes, A. Rame, and C. Ollion, "Leveraging weakly annotated data for fashion image retrieval and label prediction," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2268–2274.
[53] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.
[54] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 554–561.
[55] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," 2013, *arXiv:1306.5151*. [Online]. Available: http://arxiv.org/abs/1306.5151

TAEHUNG KIM received the B.S. degree from Korea Aerospace University, Seoul, South Korea, in 2012. He is currently pursuing the Ph.D. degree with Yonsei University. His research interests include machine learning, computer vision, fine-grained image recognition, object detection, and deep neural networks.

HOSEONG KIM (Member, IEEE) received the B.S. and Ph.D. degrees in computer science from Yonsei University, Seoul, South Korea, in 2013 and 2020, respectively. He is currently a Senior Researcher with the Agency for Defense Development, Daejeon, South Korea. His research interests include computer vision, machine learning, deep learning, zero-shot learning, object detection, video highlight detection, explainable artificial intelligence, generative adversarial networks, and video analysis.

HYERAN BYUN (Member, IEEE) received the B.S. and M.S. degrees in mathematics from Yonsei University, Seoul, South Korea, and the Ph.D. degree in computer science from Purdue University, West Lafayette, IN, USA. She is currently a Professor of computer science with Yonsei University. Her research interests include computer vision, artificial intelligence, and pattern recognition.