# Learning Disentangled Representation for Fine-Grained Visual Categorization

Wenjie Dang[1], Shuiwang Li[2] , Qijun Zhao[1,2](✉), and Fang Liu[3]

[1] National Key Laboratory of Fundamental Science on Synthetic Vision,
Sichuan University, Chengdu, China
qjzhao@scu.edu.cn

[2] College of Computer Science, Sichuan University, Chengdu, China

[3] School of Information Science and Technology, Tibet University, Lhasa, China
liu1221@utibet.edu.cn

**Abstract.** Fine-grained visual categorization (FGVC) that aims to recognize objects from subcategories with very subtle differences remains a challenging task due to the large intra-class and small inter-class variation caused by, e.g., deformation, occlusion, illumination, background clutter, etc. A great deal of recent work tackles this problem by forcing the network to focus on partial discriminable features using attention mechanisms or part-based methods. However, these methods neglect the point that the network may learn to discriminate objects from identity-unrelated features, for instance, when backgrounds are discriminable in training samples, degrading the network's generalization ability. In this paper, for the first time, we use disentangled representation learning to disentangle the fine-grained visual feature into two parts: the identity-related feature and the identity-unrelated feature. Only the identity-related feature is used for the final classification. Since identity-unrelated information is neglected in classification, intra-class variation is reduced while inter-class variation is amplified through the disentanglement, improving the classification performance as a result. Experimental results on three standard fine-grained visual categorization datasets, i.e., CUB-200-2011 (CUB), Stanford Cars (CAR) and FGVC-Aircraft (AIR), demonstrate the effectiveness of our method and show that we achieve state-of-the-art performance on the benchmarks.

**Keywords:** Fine-grained visual categorization · Disentangled representation learning · Adversarial learning

## 1 Introduction

The tasks of fine-grained visual categorization (FGVC) are to classify object categories that are similar in appearance and subtle in differences, e.g., bird species [19], car models [8], aircraft [12] and retail commodity [20], etc. Such tasks are more challenging than generic object classification. For one thing, true

Y. Peng et al. (Eds.): ICIG 2021, LNCS 12888, pp. 327–339, 2021.
https://doi.org/10.1007/978-3-030-87355-4_28

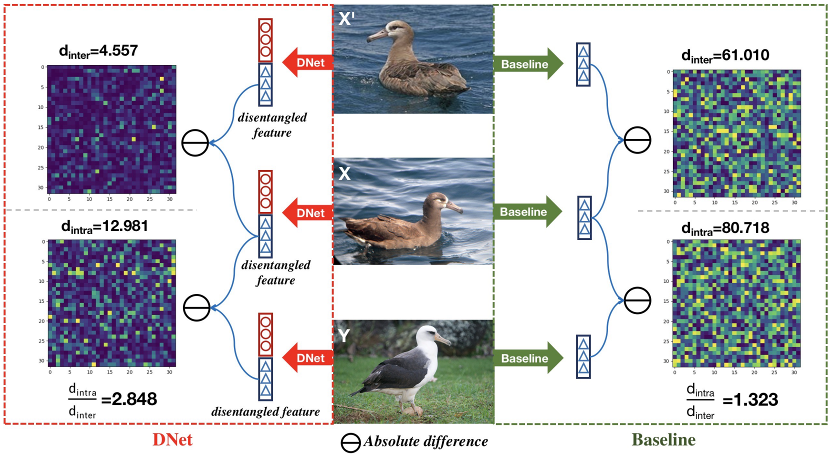

**Fig. 1.** Feature comparison between the proposed DNet and the Baseline. The former uses disentangle representation learning to get disentangled identity-related features for classification, whereas the latter does not. $X$ and $X'$ are of the same class label but $Y$ is a different class. The heat maps show the absolute differences between the feature pairs while $d_{inter}$ and $d_{intra}$ denote the Euclidian distances between the feature pairs. $\frac{d_{intra}}{d_{inter}}$ represents the ratio of inter-class distance to intra-class distance. Note that the ratio of inter-class distance to intra-class distance of the DNet is significantly larger than that of the Baseline.

discriminative information is much less due to the subtlety of inter-class differences. For another, annotated training data is very deficient due to the difficulties of data collection. Benefiting from the progress of deep learning, the recognition performance of FGVC has been significantly improved in the past years [2,5,6,11]. A great deal of recent work tackles this problem by forcing the network to focus on partial discriminative features using attention mechanisms or part-based methods. For instance, Fu et al. [4] proposed a reinforced attention proposal network to learn discriminative region attention and region-based feature representation at multiple scales. Sun et al. [17] proposed an attention-based convolutional neural network that first learns multiple attention region features of each input image through the one-squeeze multi-excitation (OSME) module and then applies a multi-attention multi-class constraint in a metric learning framework. Zheng et al. [26] proposed a part learning approach by using a channel grouping network to generate multiple parts by clustering and then classified these parts features to predict the categories of input images. To avoid costly annotations of parts or key areas for training some researchers used weakly supervised methods or tried to explicitly constraint the model to locate discriminative regions. For instance, Peng et al. [15] proposed the object part attention model (OPAM) for weakly supervised fine-grained image classification, in which part-level attention is exploited to select discriminative parts of objects and an object spatial constraint is used to ensure selected parts highly representative. Yang et al. [23] proposed a self-supervision mechanism to localize informative regions without the need for part annotations. Chen et al. [2] proposed a destruction and construction learning method to learn discriminative regions and features by first partitioning the input image into local regions, then shuffling them by a region confusion mechanism, and finally restoring the original spatial layout of local regions through a region alignment network.

All of the methods just mentioned intend to learn a model which is able to locate the discriminative information for fine-grained visual categorization, and either explicitly or implicitly guide the model with information such as part annotations, part-based constraints, and attention mechanisms. However, these methods neglect the point that the network may learn to discriminate objects from identity-unrelated features, for instance, when backgrounds are discriminative in training samples, degrading the network's generalization ability. In this paper, for the first time, we use disentangled representation learning to disentangle the fine-grained visual feature into two parts: the identity-related feature and the identity-unrelated feature. Only the identity-related feature is used for the final classification. Since the identity-unrelated information is neglected in classification, intra-class variation is reduced while inter-class variation is amplified through the disentanglement, improving the classification performance as a result. We visualize the pair-wise absolute differences of the features for classification of three images $X$, $X'$ and $Y$, as shown in Fig. 1. $X$ and $X'$ are of the same class label which is different from the label of $Y$. Obviously, regarding our DNet, the identity-related features of $X$ and $X'$ are visually more similar than those of $X$ and $Y$. And, quantitatively, the ratio of inter-class distance to intra-class distance of the DNet is significantly larger than that of the Baseline as well in these examples. It suggests the effectiveness of disentangled representation learning in reducing intra-class variation while amplifying inter-class variation for fine-grained visual categorization.

The contributions of this paper can be summarized as follows,

- We propose for the first time to use disentangled representation learning for fine-grained visual categorization, and we propose a disentanglement network (DNet) that combines two strategies for training.
- We evaluate the proposed method on three benchmarks for fine-grained visual categorization. Experimental results demonstrate the effectiveness of the proposed method and show that the proposed DNet achieves state-of-the-art performance.

The remainder of this paper is organized as follows: We review the related work in Sect. 2. In Sect. 3, we will elaborate on our methods. Experimental results are presented and analyzed in Sect. 4, and the paper is finally concluded in Sect. 5

## 2 Related Work

### 2.1 Fine-Grained Visual Categorization

There have been a variety of methods designed to distinguish fine-grained categories. Strong supervised fine-grained classification uses additional manual annotation information such as bounding box or part annotation. It can reduce the clutter of background and improve the accuracy of classification. And it has certain interpretability. The frameworks of early works [1,24] are similar to detection, which select regions and classify the pose-normalized objects. They use bounding- box/part annotations during the training and inference phase. Although this setting makes fine-grained classification more useful in practice, it is very expensive to obtain annotation information and has poor universality. Therefore, the research of fine-grained image classification is gradually replaced by weak supervision. This paper will mainly consider the last setting, where bounding-box/part annotations are not needed either at the training or inference phase.

In order to learn without fine-grained annotations, Xiao et al. [22] proposed a two-level attention algorithm, which uses the selective search algorithm [18] to detect and extract the foreground image from the original image to reduce the interference of the background and get the candidate region with local discrimination. Lin et al. [10] proposed the Bilinear CNN model, which used high-order images to capture the relationship between feature channels, and achieved 84.1% classification accuracy on CUB-200-2011. And Bilinear CNN has done a lot of work towards improvement and simplification in the later related research. Later, in order to obtain local feature information better-attention mechanism was introduced into fine-grained image classification. Fu et al. [4] develop a recurrent attention module to recursively learn discriminative region attention and region-based feature representation at multiple scales in a mutually reinforced way. Chen et al. [2] deconstructed and re-constructed input images to find discriminative regions and features. Zheng et al. [27] proposed a trilinear attention sampling network to learn features from different details.Vgg-16/Vgg-19, ResNet, DenseDet, and GoogleNet are often used as backbone networks for fine-grained classification, among which ResNet-50 is the one most used. In this paper, we also use ResNet-50 as the backbone to construct our DNet model.

### 2.2 Disentangled Representations

The disentangled representation is a kind of distributed feature representation that could separate the latent codes into disjoint explanatory factors. It can be used to learn not only representations that each factor corresponds to a single interpretable factor of variation in data sets of which the interpretable latent factors are not too many for applications such as image and video generation, but also representations of merely coarse interpretability in data sets of which the interpretable latent factors are too many or too hard to be correctly separated for applications such as image recognition and classification. There are two popular strategies in learning disentangled representations for image recognition and classification. One strategy is to use a pair of images to guide the representation learning from a generative perspective. It requires that the generated image decoded from a swapped representation, which is obtained by swapping the identity-unrelated factors of the two disentangled representations, should be close to the input image with the same identity-unrelated factors so that the identity-related factors are an invariant representation for recognition and classification. The other strategy is to use a single image to guide the representation learning from a discriminative perspective, in which methods of adversarial learning are usually used to squeeze out identity-unrelated factors from the latent codes. For instance, Zhang et al. [25] use an encoder-decoder network to disentangle the appearance feature and gait feature of the human body in the walking videos. Gait features are extracted by a similarity loss between two videos from the same person. DrNet [3] disentangles content and pose vectors with a two-encoders architecture, which removes content information in the pose vector by generative adversarial training. Peng et al. [14] proposes a pose independent feature representation method to find a rich embedding layer to encode identity-related features and identity-unrelated features. A new feature reconstruction metric is proposed to learn how to disentangle the features.

However, the first strategy does not work well in the case where the identity-unrelated factors are too complicated to encode. In especially the fine-grained birds' categorization concerted here, the backgrounds of a pair of images could be so different that the identity-unrelated factors of the two images may not even intersect semantically. So it is very hard to encode the identity-unrelated factors by which a swapped representation can be decoded validly. Therefore, in this paper, we adopt the first strategy to learn disentangling feature representation and decompose the depth feature into two parts: identity-related feature and identity-unrelated feature. They are then decoded under the constraints of minimizing the reconstruction error. And the Euclidian distance between the identity-related features of the image pair is minimized as usual to learn an invariant representation for classification. Since the swapping strick is abandoned in the adjusted first strategy, the identity-unrelated factors in the disentangled representations are learned with very weak constraints. To make up this, we use the second strategy to purposefully guide the learning of the identity-unrelated factors in an adversarial manner.

## 3 Method

In this section, we propose a disentanglement network (DNet) for fine-grained visual categorization, as shown in Fig. 2. It combines two strategies of disentangled representation learning to disentangle identity-related and identity-unrelated components in the feature obtained by ResNet-50. Our DNet consists of a backbone network, two encoders $E_1$ and $E_2$, two classifiers $C_{noid}$ and $C_{id}$, and one decoder $D$, in which two subnetworks could be identified, i.e., the single sample adversarial learning subnetwork and the paired samples disentanglement subnetwork, as shown in Fig. 3.

The former aims to train the encoder $E_1$ using adversarial learning with a single sample so that it outputs identity-unrelated feature only, corresponding to the first strategy of disentangled representation learning mentioned in Sect. 2.2, while the latter aims to jointly train both the encoders $E_1$ and $E_2$, the decoder $D$ and the classifier $C_{id}$ using a pair of samples of the same categorical label,

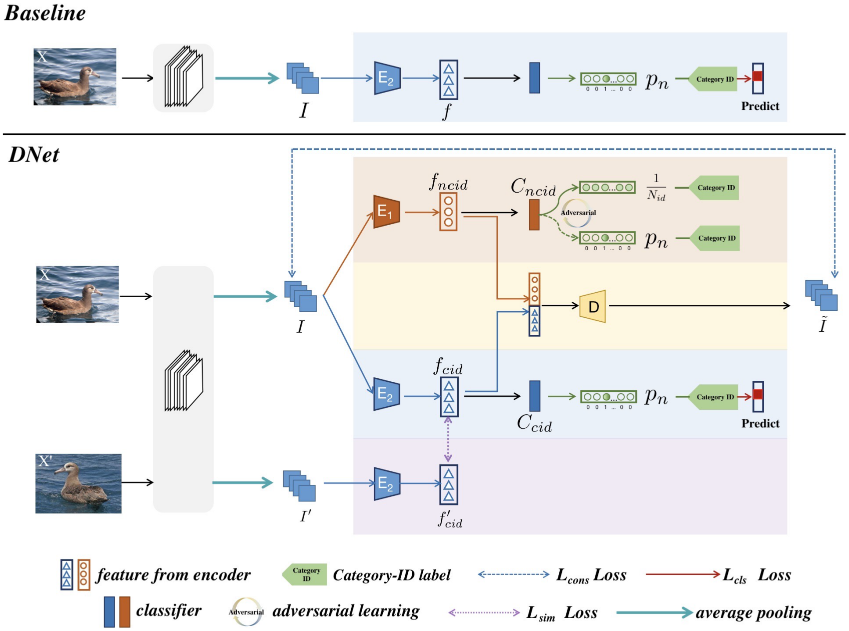

**Fig. 2.** Overview of the proposed DNet and the Baseline model. Note that the backbones of the two models are both ResNet-50. Also note that the disentangled representation learning is used in our DNet but not used in the Baseline.

corresponding to the second strategy of disentangled representation learning. Note that during inference time, only the backbone, the encoder $E_2$, and the classifier $C_{id}$ are needed. Our DNet is trained in two stages sequentially, which will be discussed in detail in the following.

### 3.1 The Single-Sample Adversarial Learning Subnetwork (SSALNet)

This subnetwork will be trained first. Given a sample $X$, let's denote by $I = B(X)$ the feature map obtained by compositing the ResNet-50 network and the average pooling layer. The encoder $E_1$ and the classifier $C_{ncid}$ are trained with adversarial learning by alternately performing the following two steps. In the first step, the encoder $E_1$ is fixed, the classifier $C_{ncid}$ is trained to minimize the following cross-entropy loss:

$$L_{adv1} = -\sum_{j=1}^{N_{id}} p[j] \log(softmax(C_{ncid}(f_{ncid}))[j]), \tag{1}$$

where $N_{id}$ is the number of classes, $p$ is the one-hot label corresponding to the input $X$, $p[j]$ denotes the $j$th entry of $p$, $f_{ncid} = E_1(I)$ represents the output

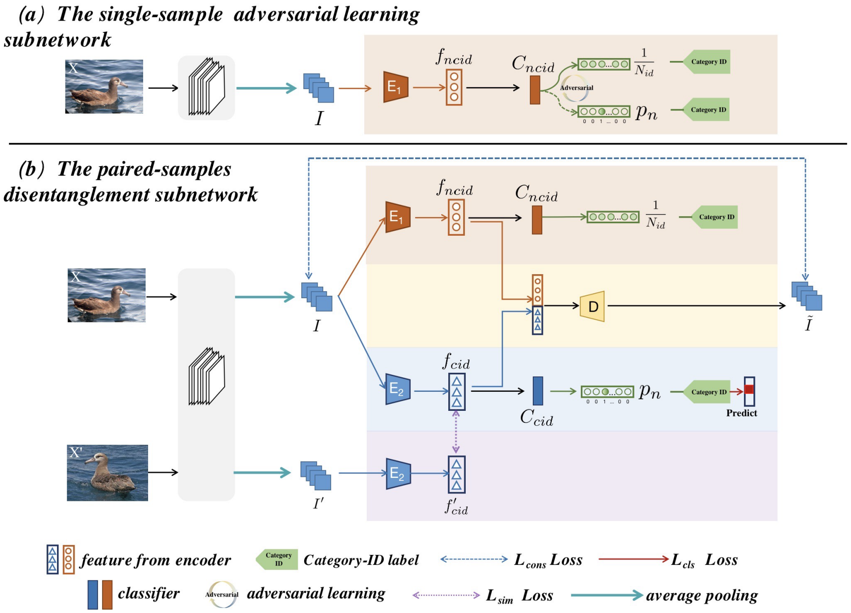

**Fig. 3.** Overview of the two subnetworks of the proposed DNet. (a) and (b) show, respectively, the single-sample adversarial learning subnetwork and the paired-samples disentanglement subnetwork.

codes of the encoder $E_1$ with $I$ as input, $softmax(\cdot)$ indicates the softmax function. In the second step, we update the encoder $E_1$ with the classifier $C_{ncid}$ being fixed. The way to ensure that the feature $f_{ncid}$ has lost all information about identity is that it produces the same prediction for all classes after being sent into the classifier $C_{ncid}$ [28]. One way to impose this constraint is to assign the probability of each id label to be $\frac{1}{N_{id}}$ in the softmax cross-entropy loss. The problem of this loss is that it would still backward gradient for updating parameters even if it reaches the minimum, so the Euclidean distance is used instead in [28], which is also utilized here. Thus, with fixed $C_{ncid}$, $E_1$ is trained with the following loss:

$$L_{adv2} = \sum_{j=1}^{N_{id}} \|softmax(C_{ncid}(f_{ncid}))[j] - \frac{1}{N_{id}}\|_2^2. \tag{2}$$

### 3.2 The Paired-Samples Disentanglement Subnetwork (PSDNet)

If the SSALNet is well trained, the encoder $E_1$ tends to output identity-unrelated features. Then, given two samples $X$ and $X'$ of the same class label, the PSDNet is trained to learn the final disentangled representation and the classifier $C_{cid}$ for fine-grained visual categorization under four losses, i.e., the $L_{adv2}$ loss, the reconstruction loss, the classification loss and the identity similarity loss. Hopefully, under the constraint of the $L_{adv2}$ loss, $E_1$ should repel identity-related information, under the constraint of the reconstruction loss the features $f_{cid}$ outputted by the encoder $E_2$ should be complementary to $f_{ncid}$, hence $f_{cid}$ are basically identity-related; under the constraint of the identity similarity loss, $f_{cid}$ and $f'_{cid}$ should be similar and therefore $E_2$ are supposed to be invariant to identity-unrelated changes of input images; under the constraint of the classification loss, the $E_2$ should produce discriminative feature and $C_{cid}$ should be discriminative. The $L_{adv2}$ loss has been defined in Sect. 3.1, the rest three losses are described as follows.

**Reconstruction Loss.** In PSDNet, the feature $I$ is first disentangled by the encoders $E_1$ and $E_2$ separately into $f_{ncid}$ and $f_{cid}$, which are then concatenated and decoded by the decoder $D$ to get a reconstructed feature $\tilde{I}$. The reconstruction loss is to punish the differences between $\tilde{I}$ and $I$, which is defined as follow,

$$L_{rcons} = \|\tilde{I} - I\|_2^2 = \|D(\{f_{ncid}, f_{cid}\}) - I\|_2^2, \tag{3}$$

where $\{f_{ncid}, f_{cid}\}$ denotes the concatenation of $f_{ncid}$ and $f_{cid}$.

**Classification Loss.** Only the identity-related feature $f_{cid}$ is used for classification. The classification loss is just the cross-entropy loss as usually defined, i.e.,

$$L_{cls} = -\sum_{j=1}^{N_{id}} p[j] \log(softmax(C_{cid}(f_{cid}))[j]). \tag{4}$$

**Identity Similarity Loss.** Since the input images $X$ and $X'$ are of the same class label, ideally the two identity-related features $f_{cid}$ and $f'_{cid}$ corresponding to $X$ and $X'$, respectively, should be identical for the purpose of classification. Note that the backbone is shared by $X$ and $X'$. So the identity similarity loss is used to punish the differences between $f_{cid}$ and $f'_{cid}$, which is defined by

$$L_{Idsim} = \|f_{cid} - f'_{cid}\|_2^2. \tag{5}$$

The overall loss $L_{PSDNet}$ for training the PSDNet is:

$$L_{PSDNet} = \alpha L_{adv2} + \beta L_{rcons} + \gamma L_{cls} + \theta L_{Idsim} \tag{6}$$

where $\alpha$, $\beta$, $\gamma$ and $\theta$ are coefficients to balance these losses.

## 4 Experiments

We evaluate the performance of our proposed DNet on three standard fine-grained object recognition datasets: CUB-200-2011 (CUB) [19], Stanford Cars (CAR) [8] and FGVC-Aircraft (AIR) [12]. We do not use any bounding box/part annotations in all our experiments. The category label of the image is the only annotation used for training.

### 4.1 Implementation Details

The input images are resized to a fixed size of 512-512 and randomly cropped into 448-448. Random rotation and random horizontal flip are applied for data augmentation. The average pooling layer connected to the backbone ResNet maps an output of the backbone to a feature of size 2048 × 1 × 1. The encoders $E_1$, $E_2$ both consist of 1024 convolution kernels of size $1 \times 1$. The classifiers $C_{ncid}$ and $C_{cid}$ are both a fully connected layer that maps a 1024 dimensional vector to a $N_{id}$ dimensional one. The decoder D consists of 2048 convolution kernels of size $1 \times 1$. $f_{ncid}$ and $f_{cid}$ are feature vectors of dimensions 1024. In addition, in view of that the bird images in the CUB-200-2011 dataset have very large variations in birds' posture, age, shooting angle and etc., we manually annotate these images into two classes, i.e., the normal and the extremal classes. The former are images of normal birds' posture, age, shooting angle and etc. while the latter corresponds to extremal cases. In the training phase, $X'$ is selected from those normal images only in this dataset.

**Table 1.** Comparison of the proposed DNet with state-of-the-art methods on CUB-200-2011 (CUB) [19], Stanford Cars (CAR) [8] and FGVC-Aircraft (AIR) [12]. The best results are in bold. Note that our DNet outperforms all of the competing methods.

| Method | Accuracy (%) | | |
|---|---|---|---|
| | CUB-200-2011 | Stanford Cars | FGVC-Aircraft |
| M-CNN (+BBox) (PR, 2018) [21] | 84.2% | - | - |
| HS-net (+BBox) (CVPR, 2017) [9] | 87.5% | - | - |
| lB-CNN (CVPR, 2017) [7] | 84.2% | 90.9% | 87.3% |
| MA-CNN (ICCV, 2017) [26] | 86.5% | 92.8% | 89.9% |
| NTS-net (ECCV, 2018) [23] | 87.5% | 93.9% | 91.4% |
| DCL (CVPR, 2019) [2] | 87.8% | 94.5% | **93.0%** |
| TASN (CVPR, 2019) [27] | 87.9% | 93.8% | - |
| Bi-modal PMA (IEEE TIP, 2020) [16] | 87.5% | 93.1% | 90.8% |
| Cross-X (CVPR, 2020) [11] | 87.7% | 94.6% | 92.6% |
| CIN (AAAI, 2020) [5] | 88.1% | 94.1% | 92.6% |
| ACNet (CVPR, 2020) [6] | 88.1% | 94.6% | 92.4% |
| Baseline | 84.0% | 92.4% | 89.7% |
| DNet | **88.3%** | **95.0%** | 92.7% |

Baseline and DNet are both trained for 200 epochs to obtain stable accuracies, and learning rates decay by a factor of 10 for every 40 epochs. We set $\alpha$=1, $\gamma = 1$, $\theta = 10$ for all experiments reported in this paper. It is worth mentioning that we use dynamic adjustment to train the reconstruction loss. From epoch 1 to 60, $\beta$ is set to 1 and from epoch 60 to 200 it is set to 0.1.

### 4.2 Performance Comparison

The results on CUB-200-2011, Stanford Cars, and FGVC-Aircraft are presented in Table 1. Considering that some of the compared methods use image-level labels or bounding box annotations, the information of extra annotations is also presented in parentheses for direct comparisons. As can be seen, our DNet significantly outperforms the Baseline on the three benchmarks, having gains of 4.3%, 2.6% and 3.0%, respectively, on CUB-200-2011, Stanford Cars, and FGVC-Aircraft, justifying the effectiveness of the proposed method. Moreover, DNet surpasses state-of-the-art methods on CUB-200-2011 and Stanford Cars and is also very competitive on FGVC-Aircraft. Considering the simple structures of the encoders and decoder used in DNet, it suggests that disentangled representation learning is promising in improving the performance of fine-grained visual categorization.

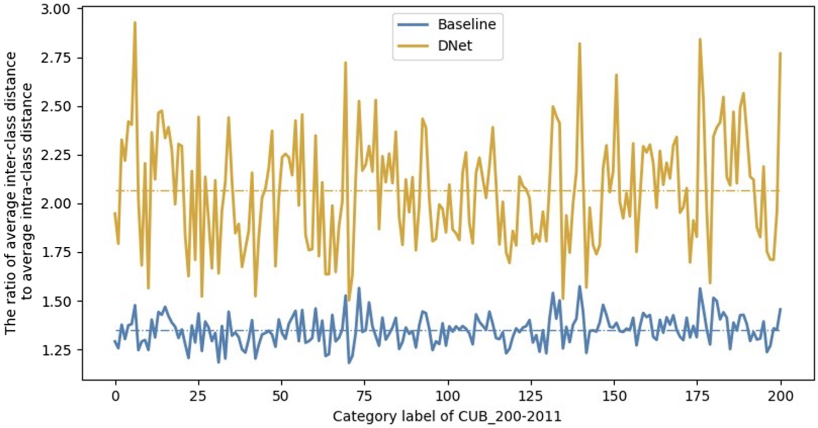

**Fig. 4.** Ratios of average inter-class distance to average intra-class distance of Our DNet and the Baseline model on CUB-200-2011 test data, computed in the one-vs-rest manner.

### 4.3 Ablation Study

We conduct ablation studies on CUB-200-2011 dataset to validate that the proposed DNet is effective in reducing inter-class variation meanwhile amplifying intra-class variation. We evaluate inter-class and intra-class distances of feature vectors of Baseline and DNet on the test data of CUB-200-2011. Since the number of inter-class distances is a big combinatorial number for CUB-200-2011 which has 200 categories, it is not suitable to use class distance matrices for evaluation. In view of that a multi-class classification can be splitted into multiple binary classification problems using the one-vs-rest method [13], we use this method to reduce the large amount of inter-class distances into 200 average inter-class distances, which is equal to the number of average intra-class distances. In doing this, we obtain 200 ratios of average inter-class distance to average intra-class distance for each model. By plotting these ratios, we can visually compare different models' performance in reducing inter-class variation meanwhile amplifying intra-class variation. Figure 4 shows the 200 ratios of average inter-class distance to average intra-class distance of the Baseline and our DNet, respectively. The average ratio of each model, specifically 2.07 and 1.35, is plotted as a dash line in corresponding color. As can be seen, our DNet is significantly better than the Baseline, justifying the effectiveness of the proposed method in reducing inter-class variation meanwhile amplifying intra-class variation.

## 5 Conclusion

We proposed the network DNet that uses disentangled representation learning to extract feature representation for FGVC. As far as we know, this is the first time to utilize feature disentanglement to solve the tasks of FGVC. In view of the difficulties in disentangling features of images with complicated backgrounds, we combined two strategies to train the DNet. Experimental results show that the proposed method achieves state-of-the-art performance. Considering the simple structures of the encoders and decoder used in DNet, we believe that disentangled representation learning is promising for FGVC.

**Acknowledgments.** This research is supported by the Science and Technology Department of Tibet (Grant No. XZ202102YD0018C).

## References

1. Branson, S., Beijbom, O., Belongie, S.: Efficient large-scale structured learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1806–1813 (2013)
2. Chen, Y., Bai, Y., Zhang, W., Mei, T.: Destruction and construction learning for fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5157–5166 (2019)
3. Denton, E., Birodkar, V.: Unsupervised learning of disentangled representations from video. arXiv preprint arXiv:1705.10915 (2017)
4. Fu, J., Zheng, H., Mei, T.: Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4438–4446 (2017)
5. Gao, Y., Han, X., Wang, X., Huang, W., Scott, M.: Channel interaction networks for fine-grained image categorization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 10818–10825 (2020)
6. Ji, R., et al.: Attention convolutional binary neural tree for fine-grained visual categorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10468–10477 (2020)
7. Kong, S., Fowlkes, C.: Low-rank bilinear pooling for fine-grained classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 365–374 (2017)

8. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3D object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 554–561 (2013)
9. Lam, M., Mahasseni, B., Todorovic, S.: Fine-grained recognition as HSnet search for informative image parts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2520–2529 (2017)
10. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear CNN models for fine-grained visual recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1449–1457 (2015)
11. Luo, W., et al.: Cross-x learning for fine-grained visual categorization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8242–8251 (2019)
12. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
13. Murphy, K.P.: Machine Learning: A Probabilistic Perspective (2012)
14. Peng, X., Yu, X., Sohn, K., Metaxas, D.N., Chandraker, M.: Reconstruction-based disentanglement for pose-invariant face recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1623–1632 (2017)
15. Peng, Y., He, X., Zhao, J.: Object-part attention model for fine-grained image classification. IEEE Trans. Image Process. **27**(3), 1487–1500 (2017)
16. Song, K., Wei, X.S., Shu, X., Song, R.J., Lu, J.: Bi-modal progressive mask attention for fine-grained recognition. IEEE Trans. Image Process. **29**, 7006–7018 (2020)
17. Sun, M., Yuan, Y., Zhou, F., Ding, E.: Multi-attention multi-class constraint for fine-grained image recognition. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11220, pp. 834–850. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01270-0_49
18. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. Int. J. Comput. Vision **104**(2), 154–171 (2013)
19. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 dataset (2011)
20. Wei, X.S., Cui, Q., Yang, L., Wang, P., Liu, L.: RPC: a large-scale retail product checkout dataset. arXiv preprint arXiv:1901.07249 (2019)
21. Wei, X.S., Xie, C.W., Wu, J., Shen, C.: Mask-CNN: localizing parts and selecting descriptors for fine-grained bird species categorization. Pattern Recogn. **76**, 704–714 (2018)
22. Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z.: The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 842–850 (2015)
23. Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., Wang, L.: Learning to navigate for fine-grained classification. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 438–454. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_26
24. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based R-CNNs for fine-grained category detection. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 834–849. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_54
25. Zhang, Z., et al.: Gait recognition via disentangled representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4710–4719 (2019)

26. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5209–5217 (2017)
27. Zheng, H., Fu, J., Zha, Z.J., Luo, J.: Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5012–5021 (2019)
28. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9299–9306 (2019)