Giant Panda Identification

Le Wang, *Senior Member, IEEE*, Rizhi Ding, *Student Member, IEEE*, Yuanhao Zhai, *Member, IEEE*, Qilin Zhang, *Member, IEEE*, Wei Tang, *Member, IEEE*, Nanning Zheng, *Fellow, IEEE*, and Gang Hua, *Fellow, IEEE*

***Abstract*— The lack of automatic tools to identify giant panda makes it hard to keep track of and manage giant pandas in wildlife conservation missions. In this paper, we introduce a new Giant Panda Identification (GPID) task, which aims to identify each individual panda based on an image. Though related to the human re-identification and animal classification problem, GPID is extraordinarily challenging due to subtle visual differences between pandas and cluttered global information. In this paper, we propose a new benchmark dataset iPanda-50 for GPID. The iPanda-50 consists of 6, 874 images from 50 giant panda individuals, and is collected from panda streaming videos. We also introduce a new Feature-Fusion Network with Patch Detector (FFN-PD) for GPID. The proposed FFN-PD exploits the patch detector to detect discriminative local patches without using any part annotations or extra location sub-networks, and builds a hierarchical representation by fusing both global and local features to enhance the inter-layer patch feature interactions. Specifically, an attentional cross-channel pooling is embedded in the proposed FFN-PD to improve the identify-specific patch detectors. Experiments performed on the iPanda-50 datasets demonstrate the proposed FFN-PD significantly outperforms competing methods. Besides, experiments on other fine-grained recognition datasets (*i.e.*, CUB-200-2011, Stanford Cars, and FGVC-Aircraft) demonstrate that the proposed FFN-PD outperforms existing state-of-the-art methods.**

***Index Terms*— Giant panda identification, feature fusion, patch detector, fine-grained recognition.**

I. Introduction

AUTOMATIC identifying giant panda is an important task in panda management and interpretation of images captured by motion activated trail cameras in wildlife research. However, the subtle differences between panda individuals and the appearance variations due to posture/viewpoint make it challenging to correctly identify each giant panda. Even though some biometric trait-based methods (*e.g.*, DNA-based assessment [1]) or RFID (Radio Frequency Identification) tags can help address such problem, these time-consuming data collection procedures makes them expensive, inconvenient or even impractical.

(a) Images of the same panda (named as "yingying").

(b) Images of different individual pandas.

Fig. 1. Examples from the proposed iPanda-50 dataset. (a) Images of the same panda could exhibit dramatic appearance variations due to different illumination, viewpoint, posture and occlusion conditions. (b) Images of different individual pandas could have only subtle appearance differences.

In this paper, we propose to address this problem with computer vision techniques based purely on an input image. We formulate the new Giant Panda Identification (GPID) task and propose a new benchmark dataset iPanda-50 for GPID. We assume an image database of giant pandas of known identities are available and the unidentified giant panda belongs to one of these identities. Thus, GPID is a closed-set identification problem. While this paper focuses on identifying giant pandas, we expect the technique could be extended to solve other animal identification tasks.

Image-based giant panda identification is very challenging due to large intra-identity variations and small inter-identity distances. Image examples shown in Figure 1 illustrate the challenges for the GPID task, *i.e.*, images of the same panda exhibit dramatic appearance variations caused by varying illumination conditions, viewpoints, postures, and occlusions, while images of different individual pandas could look very similar to untrained human eyes.

Furthermore, as illustrated in Figure 2, GPID is distinct from other visual identification tasks such as face recognition [2]–[4], person re-identification [5]–[7] and fine-grained recognition [8]–[12]. Compared with the face recognition task, where the human face is a rigid object with small local deformations, giant pandas in GPID exhibit postures with

Manuscript received December 2, 2020; revised January 25, 2021; accepted January 25, 2021. Date of publication February 4, 2021; date of current version February 12, 2021. This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0101400, in part by the NSFC under Grant 62088102 and Grant 61976171, in part by the Young Elite Scientists Sponsorship Program by CAST under Grant 2018QNRC001, and in part by the Natural Science Foundation of Shaanxi under Grant 2020JQ-069. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tao Mei. *(Corresponding author: Gang Hua.)*

Le Wang, Rizhi Ding, Yuanhao Zhai, and Nanning Zheng are with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: lewang@mail.xjtu.edu.cn; nnzheng@mail.xjtu.edu.cn; drz123@stu.xjtu.edu.cn; yuanhaozhai@gmail.com).

Qilin Zhang is with the ABB Corporate Research Center, Raleigh, NC 27606 USA (e-mail: samqzhang@gmail.com).

Wei Tang is with the Department of Computer Science, University of Illinois, Chicago, IL 60607 USA (e-mail: tangw@uic.edu).

Gang Hua is with Wormpex AI Research, Bellevue, WA 98004 USA (e-mail: ganghua@gmail.com).

Digital Object Identifier 10.1109/TIP.2021.3055627

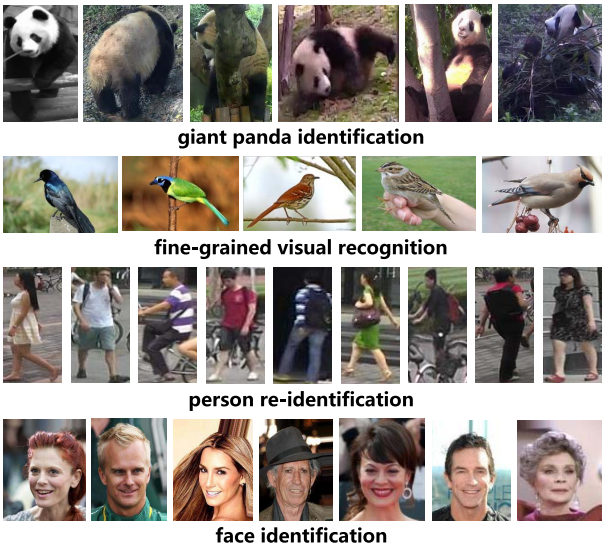


Fig. 2. Comparison between the giant panda identification task and similar computer vision tasks. Four rows above show image examples from the proposed iPanda-50 dataset for giant panda identification, the CUB-200-2011 dataset [13] for fine-grained visual recognition, the Market-1501 dataset [14] for person re-identification, and the MS-Celeb-1M dataset [15] for face identification, respectively.

huge difference, thus it is almost impossible to align the the giant panda images before comparing their discriminative features. Compared with the person re-identification task, the giant panda is not only an articulated object having a large degrees of freedom (thus various postures and occlusions) but also lacks discriminative attributes like clothing, while the pedestrains often show distinctive appearances, *e.g.*, different clothing, backpacks or hats. Specifically, GPID is highly related to the fine-grained visual recognition (FGVR) [8]–[12], where both aim to distinguish subtle differences between visually similar entities, especially in their local parts. However, the individual-level appearance differences in GPID are arguably more challenging to detect than the category-level differences in FGVR. Therefore, GPID is generally more challenging than the aforementioned tasks.

Since the differences among inter-category images occur on subtle parts, both GPID and FGVR should be capable of identifying discriminative local regions and learning features that capture their visual differences. However, most existing FGVR methods conduct part localization and feature learning independently. For instance, some part-based methods [16]–[19] train a part detection sub-network using part annotations and extract features from each part region, which are subsequently combined with the global feature for recognition. Despite the promising results, they rely heavily on manual part labeling, which could be time-consuming and expensive. Worse still, the pre-defined parts may not necessarily correspond to the most discriminative regions and thus result in inferior recognition results.

Recently, some region-attention methods [20]–[23] introduce multi-attention modules as sub-networks to learn the discriminative regions/features in a weakly-supervised way and do not need manual part annotations. However, such multi-stage methods rely on a complicated procedure of network training, and sometimes accumulate errors and result in degraded performance when prior stages focus on false attention regions. Alternatively, end-to-end methods [24], [25] require only image-level annotations. Some recent methods [26], [27] implement end-to-end training based on bilinear pooling frameworks, but most of them only use features from the last convolutional layer, which might be suboptimal for fine-grained recognition tasks.

To address the above challenges, we propose a Feature-Fusion Network with Patch Detector (FFN-PD) for GPID. The proposed FFN-PD does not rely on any location sub-networks and can be simply trained end-to-end without additional part annotations. Inspired by [28], we adopt an asymmetric multi-stream structure to capture both local and global features, and employ 1×1 convolution filters (*i.e.*, patch detectors) to automatically detect most discriminative local patches, which could be the key to identify giant pandas. Thanks to this design, the proposed method does not require additional part annotations, and the local patches of each giant panda are self-excavated by the network. In this way, we avoid fixed types of parts and these learned parts are not artificially constrained to be shared among different pandas. Furthermore, we propose a novel fusion stream to fuse global and local features, and generate a hierarchical representation. This mid-level representation embodies inter-layer patch feature interactions and allows the network to further focus on more commonly discriminative patch features. To facilitate the learning of identity-specific patch detectors, we further introduce a novel attentional cross-channel pooling to achieve convolution filter supervision.

To sum up, the key contributions of this paper are as follows:

- To the best of our knowledge, this is the first work addressing the important yet challenging task of Giant Panda Identification (GPID) in images. We build a new benchmark dataset called iPanda-50, which exhibits extreme similarity between different individual-level pandas (small inter-identity distances) and dramatic variations of appearances, illuminations, viewpoints, postures, and occlusions within each identity (large intra-identity variations).
- We propose a novel Feature-Fusion Network with Patch Detector (FFN-PD) for GPID with several technical innovations. First, we embed patch detectors across layers to generate more significant representations for local parts. Second, we apply a new hierarchical representation to capture inter-layer patch feature interactions. Third, a new attentional cross-channel pooling serves as the convolution filter supervision to enhance class-specific patch detectors. The proposed FFN-PD can be trained end-to-end and does not require any extra part annotations.
- We evaluate the proposed FFN-PD on the challenging iPanda-50 dataset as well as other fine-grained recognition datasets. The results show the proposed FFN-PD achieves a significant performance advantage against competing methods on the iPanda-50 dataset, and also achieve state-of-the-art performance on other fine-grained recognition datasets (*i.e.*, CUB-200-2011, Stanford Cars, and FGVC-Aircraft). Besides, extensive ablation studies

are carried out to validate the contribution of each component.
- We find that covering the eyes of pandas via Gaussian blur will significantly degrade the identification performance. This indicates that the panda's eyes play a critical role in panda identification.

This paper extends our previous conference paper [29] in four aspects. (1) More comprehensive review is included in Section II. (2) More details on problem formulation and implementation details are provided. (3) We augment the panda dataset and add extra annotations of the locations of pandas' eyes, which we found to be a critical factor for panda identification. (4) We include more experimental details and more competitive methods for comparison.

The rest of the paper is organized as follows. In Section II we review related work. In Section III we present the details of the proposed Feature-Fusion Network with Patch Detector. In Section IV, we introduce the new benchmark iPanda-50. In Section V, we introduce implementation details, along with detailed evaluation results and discussions. Finally, we conclude the paper in Section VI.

II. Related Work

In this section, we briefly review related work on fine-grained visual recognition, discriminative region excavation, feature fusion and multi-scale feature representation.

A. Fine-Grained Visual Recognition

Popular Fine-grained Visual Recognition (FGVR) tasks [30]–[33] often involve the classification of subspecies/breeds, where appearance variances among different categories differ only in slight parts. Profited by the evolvement of deep Convolutional Neural Networks (CNN), FGVR has transitioned from the strongly-supervised way [13], [16]–[18] to the weakly-supervised manner [10], [20], [34], [35] in the past few years. For strongly-supervised methods [17], [19], they generally utilize localization networks to locate discriminative part regions with part labels, and conduct "hard" crop operation on these regions to further extract region features. Despite their efficacy, their heavy reliance on part annotations restricts practical applications.

Recently, the attention mechanism introduced in several part annotation-free methods [20], [35] have been developed. Fu *et al.* [20] propose a recurrent attention CNN, which recursively learns discriminative region attention and region-based feature representation at multiple scales in an alternating optimization scheme. However, the alternate training of multiple sub-networks needs to be adjusted manually, which could limit their practicality. Sun *et al.* [35] employ a one-squeeze multi-excitation module to obtain multiple attention region features of each input image and then utilize a metric learning framework with a multi-attention multi-class constraint, but this work contains a non-trivial sample selection procedure.

Alternatively, to achieve the ease of training in an end-to-end manner, some neural network designs resort to the bilinear pooling [36] and its variants [24], [27]. However, most of them only exploit feature representation from the last convolutional layer, which are typically too coarse for fine-grained tasks and incur high computational cost due to the typical large depth/channel dimension.

The most relevant work to ours is DFL-CNN [28] as it also leverages patch detectors to discover the discriminative local patches. However, our work differs from [28] in three main ways. (1) DFL-CNN simply embeds patch detectors after multiple convolutional layers to achieve multi-scale patch learning, while we extend patch detectors in another layer to construct a fusion stream, which makes the network further focus on high-response discriminative local patches. As a result, our network promotes the interaction of local patches across layers while [28] neglects such interaction. (2) We introduce a new filter supervision (*i.e.*, attentional cross-channel pooling) to respond reasonably to the roles of different local patches, and it works better than the mediocre cross-channel pooling in [28]. (3) We discard the complicated method of non-random initialization which is designed to avoid bad local minima while learning the patch detectors in [28].

B. Discriminative Region Excavation

Due to differences of small object parts, discriminative region excavation plays an important role for fine-grained object recognition. A straightforward way to represent parts is to find where the discriminative regions are with location networks. One prior work [37] uses a volumetric poselet scheme to establish bird pose-normalized part appearance. Another work [38] trains two deformable part descriptors with object part annotations to localize the semantic parts. Furthermore, some methods locate object parts with key part point labels to regress region bounding boxes by fully convolutional network [16], [17] or Mask-CNN [19]. With image-level supervision, recent methods [20], [21], [23] learn discriminative region features by generating region attention maps in a multi-stage optimal manner. For example, Zhang *et al.* [23] learn multiple experts focused on the diversity of regions by combining a gradually-enhanced learning strategy. Ding *et al.* [39] collect local maximums to estimate informative regions and learn a set of sparse attention for capturing fine-detailed visual evidence.

In contrast, our design is an end-to-end network, which automatically learns discriminative regions without any expensive bounding box/part annotations or complex location subnetworks. We design our method so that it directly optimizes local region search with the patch-level classification loss.

C. Feature Fusion

CNNs have also become popular for instance-level classification, but the feature map out of a single convolution layer is often insufficient to distinguish subtle differences among very similar objects or categories, *e.g.*, feline/canine/avian subspecies. Recently, efforts of combining feature maps from multiple convolution layers have been proposed [40], [41]. Long *et al.* [42] combine coarse features and fine features from different convolution layers for image segmentation. Hariharan *et al.* [43] present a hyper-column representation for object segmentation and fine-grained localization,

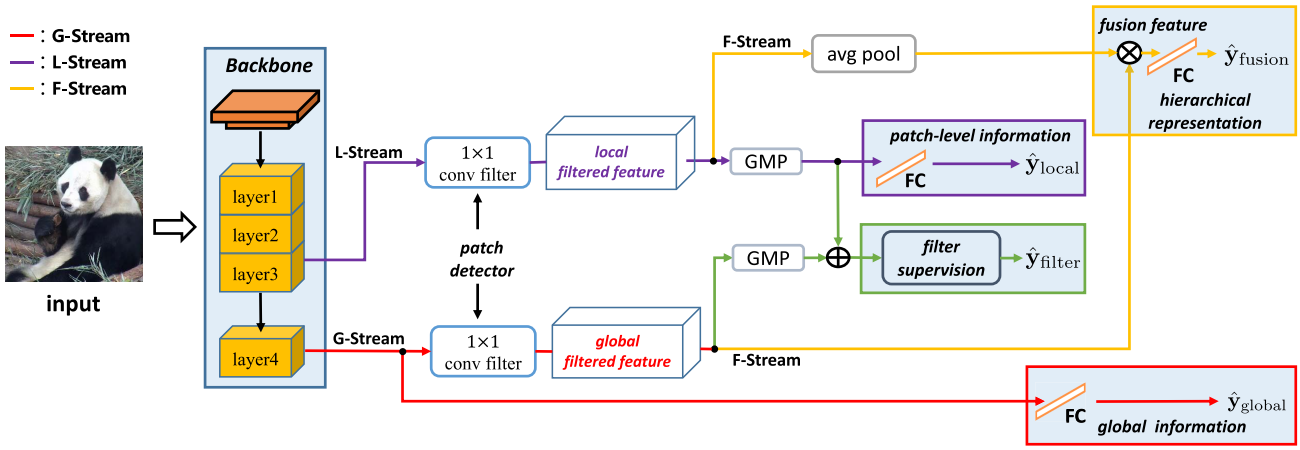


Fig. 3. An overview of the proposed Feature-Fusion Network with Patch Detector, which consists of a global stream (G-Stream), a local stream (L-Stream) and a fusion stream (F-Stream) to exploit global information, patch-level information and hierarchical information, respectively. Besides, an attentional convolution filter supervision is proposed to facilitate identity-specific discriminative patch learning.

which concatenates pixel-level activations from all CNN units. Cai *et al.* [27] concatenate multiple feature maps from different layers to exploit the intra-layer and inter-layer interactions. More recently, Yu *et al.* [25] claim that hierarchical bilinear pooling could enhance both inter-layer patch feature interaction and fine-grained feature representation.

D. Multi-Scale Feature Representation

In order to further improve the performance, many CNNs of computer visual tasks have exploited multi-scale features. Some methods [42], [44] concatenate feature vectors inferred from multiple layers and obtain the final result to include informative features of low-level spatial resolution and high-level semantic properties. Besides, Lin *et al.* [45] propose a Feature Pyramid Network (FPN) to build high-level semantic feature maps at all scales by a top-down architecture with lateral connections. Wang *et al.* [46] use a multiple granularity framework to encode informative and discriminative features covering all the grain levels. More recently, the deep layer aggregation structure studied in [47] produces better accuracy by iteratively and hierarchically merging the feature hierarchy.

In this paper, we perform multi-scale feature representation by independently computing four various losses, where each one accounts for a meaningful semantic representation, and they jointly contribute to the final classification.

III. Feature-Fusion Network with Patch Detector

In this section, we first formulate the task of Giant Panda Identification (GPID), then describe the framework of the proposed Feature-Fusion Network with Patch Detector (FFN-PD) in detail, and finally present the network training and testing process. As described in Section I, it is necessary to exploit both global and fine-grained discriminative information to identify pandas. To this end, the proposed FFN-PD adopts an asymmetric multi-stream structure. As illustrated in Figure 3, the proposed FFN-PD consists of a global stream to learn global features, a local stream with patch detectors to learn local discriminative features, and a fusion stream to learn hierarchical representation. Besides, a novel attentional cross-channel pooling is employed to force identity-specific patch features learning.

A. Problem Formulation

The training set contains a set of training tuples, where each tuple $(\mathbf{X}, \mathbf{y})$ consists of one RGB giant panda image $\mathbf{X} \in \mathbb{R}^{3\times H\times W}$ and its corresponding ground truth label $\mathbf{y} \in \mathbb{R}^{N}$, where H and W respectively indicate the height and width of the image, $\mathbf{y}$ is a one-hot label vector, and N is the number of panda individuals. The goal of GPID is to correctly map the testing giant panda image to its label vector.

B. Global Stream

Similar to generic image classification methods, the global stream (*i.e.*, G-Stream) consists of a feature-extraction backbone (*e.g.*, ResNet50 [48]) for image feature extraction, and a fully-connected classification layer with softmax to output the classification prediction $\hat{\mathbf{y}}_{\text{global}}$.

However, using the G-Stream alone is far from enough to clearly distinguish panda identities, as panda identities exhibit very similar global patterns. Therefore, we introduce the following local stream to identify pandas in a fine-grained level.

C. Local Stream

The appearance differences among panda individuals usually occur at subtle parts, which are crucial for panda identification. To this end, we propose a local stream (*i.e.*, L-Stream) to capture the local discriminative information. Naturally, early layers of a feature-extraction backbone have smaller receptive field than its deep counterpart, thus early layers are able to detect more fine-grained information [45]. Therefore, the L-Stream takes as input an intermediate feature map from the feature-extraction backbone (*e.g.*, feature map at the layer3 of ResNet50 [48]).

Denote the intermediate feature map is of size $C' \times H_l \times W_l$, where C' is the number of channels, H_l and W_l are respectively the height and width of the feature map. The feature map can

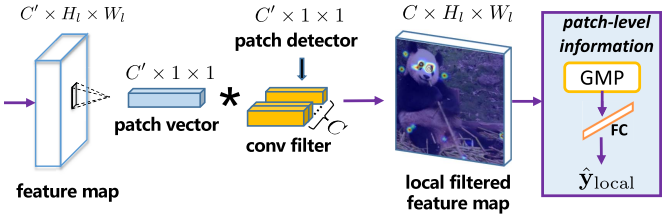


Fig. 4. An overview of the patch detector. A patch detector finds discriminative local patches with patch-level information, which is the L-stream in Figure 3.

be further interpreted as $H_l \times W_l$ patch vectors of size $C' \times 1 \times 1$, where each patch vector represents the local feature within its receptive field. Inspired by [28], we exploit 1×1 convolutional filters as patch detectors to detect local patches so that high responses represent subtle discriminative characteristics of a panda identity. Thanks to this design, the learned patch detectors will not be limited to manually pre-defined locations, thus free us from additional manual location annotations. As shown in Figure 4, C patch detectors are used to learn the most discriminative image patches. The output feature map is filterd with a global max pooling (GMP) to keep the most discriminative features, and passed through a fully-connected softmax layer to get the local stream prediction result $\hat{\mathbf{y}}_{\text{local}}$.

D. Fusion Stream

Despite using patch detectors in the L-Stream, as illustrated in Figure 3, we further embed patch detectors into the G-Stream to exploit various semantic patches in feature maps with different receptive fields. To guide the network to emphasize more on high-response patches, we introduce a fusion stream (*i.e.*, F-Stream) to fuse local and global features via element-wise multiplication. These inter-layer patch interactions can further activate common high-response regions and produce a local-global hierarchical representation. Figure 5 shows an overview of the F-Stream.

After passing deep feature maps through patch detectors of the L-Stream and G-Stream, we obtain the local and global filtered feature maps $\mathbf{M}_l \in \mathbb{R}^{C \times H_l \times W_l}$ and $\mathbf{M}_g \in \mathbb{R}^{C \times H_g \times W_g}$ respectively, where C_g, H_g, and W_g are the number of output channels, height and width of the global filtered feature map, respectively. Typically, $H_l > H_g$ and $W_l > W_g$, because the size of feature map generally decreases as the depth of layers increases. An additional average pooling layer is included to reduce the spatial dimension of the local filtered feature map and match the size of global filtered feature map, such that $\bar{\mathbf{M}}_l = \text{AvgPool}(\mathbf{M}_l) \in \mathbb{R}^{C \times H_g \times W_g}$. Subsequently, the feature fusion is implemented via element-wise multiplication:

$$\mathbf{M}_f = \mathbf{M}_g \odot \bar{\mathbf{M}}_l, \tag{1}$$

where $\odot$ denotes the element-wise multiplication and $\mathbf{M}_f \in \mathbb{R}^{C \times H_g \times W_g}$. As shown in Figure 5, the L-Stream may falsely exploit background noise as discriminative patches, while the G-Stream can only detect coarse regions. With the proposed element-wise multiplication fusion method, the hierarchical representation can guide the network to focus on the commonly interesting regions across layers, which could effectively promote identification accuracy. Then, global average pooling is implemented on the fusion feature map to reduce the spatial dimension: $\mathbf{m}_f = \text{GlobalAvgPool}(\mathbf{M}_f) \in \mathbb{R}^C$. Finally, ℓ_2 normalization is carried out to obtain the final hierarchical representation:

$$\bar{\mathbf{m}}_f = \frac{\mathbf{m}_f}{||\mathbf{m}_f||_2}, \tag{2}$$

where $||\cdot||_2$ denotes the ℓ_2 norm. $\bar{\mathbf{m}}_f$ is subsequently fed into a fully-connected layer with softmax to get the fusion prediction $\hat{\mathbf{y}}_{\text{fusion}}$.

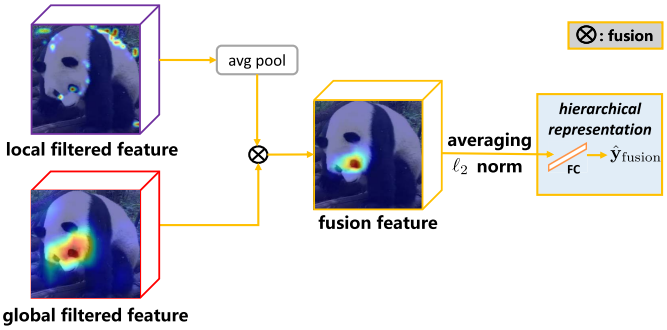


Fig. 5. An overview of the fusion stream. The feature fusion process fuses local filtered features and global filtered features via element-wise multiplication and results in a hierarchical representation. The learned heat maps of each feature map are illustrated.

E. Attentional Convolution Filter Supervision

As illustrated in Figure 5, the fully-connected layer used to compute $\hat{\mathbf{y}}_{\text{local}}$ inevitably mixes all discriminative patches together. Moreover, there is no guarantee that the 1×1 convolutional filters (designated as patch detectors) will focus on specific discriminative patches of a certain identity. Therefore, an additional supervision is needed to encourage patch detectors to emphasize on identity-specific discriminative patches. To this end, as illustrated in Figure 6, we propose a novel attentional cross-channel pooling module to address the aforementioned problem.

Specifically, let the number of patch detectors be $C = k \cdot N$, where k is a pre-defined hyperparameter indicating the top-k most discriminative local patches for each identity. Given local and global filtered feature maps $\mathbf{M}_l$ and $\mathbf{M}_g$, we first use global max pooling to obtain two $(k \cdot N)$-dimensional feature vectors $\mathbf{v}_l$ and $\mathbf{v}_g \in \mathbb{R}^{k \cdot N}$, respectively. As mentioned above, each identity is assigned with k patch detectors to detect its discriminative features, thus the k values in the feature vectors are expected to have high activations, while others are expected to have low activations. Therefore, we propose to directly generate identification predictions from the feature vectors. Specifically, we first combine $\mathbf{v}_g$ and $\mathbf{v}_l$ via element-wise addition as $\mathbf{v} = \mathbf{v}_l + \mathbf{v}_g$. Then, for convenience, we reshape the $(k \cdot N)$-dimensional feature vector as a matrix $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_N] \in \mathbb{R}^{k \times N}$.

In [28], this combined $\mathbf{V}$ is simply average-pooled to generate a N-dimensional vector $\mathbf{a} \in \mathbb{R}^N$ for identification prediction:

$$\mathbf{a} = \frac{1}{k}\mathbf{V}^{\mathrm{T}}\mathbf{1}_{k \times 1}, \tag{3}$$

where $\mathbf{1}_{k \times 1}$ denotes an all-one vector of size $k \times 1$.

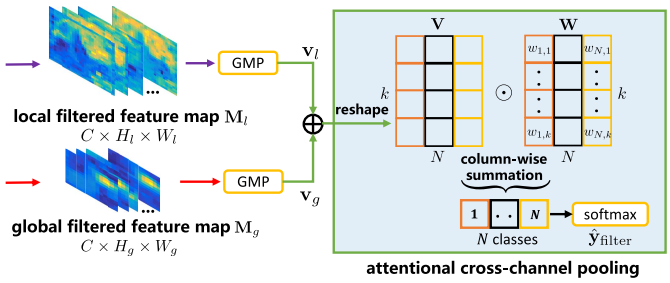


Fig. 6. An overview of the proposed attentional cross-channel pooling. The local and global filtered feature vectors are weighted sumed via an attention weight matrix $\mathbf{W}$, and generate a filter identification prediction. This figure provides details of the "filter supervision $\rightarrow \hat{\mathbf{y}}_{\text{filter}}$" green rectangle in Figure 3.

However, such simple averaging strategies are reported to induce balanced responses in these k patches during back-propagation [49]. Besides, the k discriminative features may not necessarily appear in each image due to different panda poses. Therefore, we concern that the simple averaging in Eq. (3) might assign equal weights to patches with different semantic significance and falsely detect irrelevant patches, thus adversely affect the performance. To address this problem, we propose a new attention mechanism, which automatically learns the weights assigned to the k local patches. Specifically, let $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_N] \in \mathbb{R}^{k \times N}$ denote the attention weights, with each column $\mathbf{w}_i = \left[w_{i,1}, \cdots, w_{i,k}\right]^T \in \mathbb{R}^k$, $i = 1, \cdots, N$. All elements in $\mathbf{W}$ are initialized to $1/k$ and automatically updated during the training process via back-propagation. Then, we weighted sum the feature matrix $\mathbf{V}$ via the attention weights $\mathbf{W}$, and generate the filter prediction $\hat{\mathbf{y}}_{\text{filter}}$ via class-wise softmax:

$$\hat{\mathbf{y}}_{\text{filter}} = \text{softmax}\left((\mathbf{V} \odot \mathbf{W})^{\text{T}} \mathbf{1}_{k \times 1}\right), \tag{4}$$

where softmax$(\cdot)$ denotes a class-wise softmax.

F. Network Training and Testing

By exploiting information from different streams, we obtain different semantic identification predictions from features with different receptive fields. Specifically, three predictions $\hat{\mathbf{y}}_{\text{global}}$, $\hat{\mathbf{y}}_{\text{local}}$, and $\hat{\mathbf{y}}_{\text{fusion}}$ are obtained from the global stream, the local stream, and the fusion stream, respectively. Besides, an additional filter supervision prediction $\hat{\mathbf{y}}_{\text{filter}}$ is obtained directly from the feature vectors. During training, the four predictions are supervised by the standard cross entropy loss, and a weighted sum of the four losses forms the total loss $\mathcal{L}$:

$$\mathcal{L} = \sum_{i \subset P} \mathcal{L}_{\text{ce}}(\hat{\mathbf{y}}_i, \mathbf{y}) + \lambda \mathcal{L}_{\text{ce}}(\hat{\mathbf{y}}_{\text{filter}}, \mathbf{y}), \tag{5}$$

where $P = \{\text{global}, \text{local}, \text{fusion}\}$, $\mathcal{L}_{\text{ce}}$ is the standard cross entropy loss, and λ is a weight parameter.

During testing, we also use a weighted average to generate the final prediction $\hat{\mathbf{y}}$:

$$\hat{\mathbf{y}} = \frac{\sum_{i \in P} \hat{\mathbf{y}}_i + \lambda \hat{\mathbf{y}}_{\text{filter}}}{3 + \lambda}. \tag{6}$$

IV. The iPanda-50 Dataset

We present a new iPanda-50 dataset for the giant panda identification task. We first collect giant panda streaming videos from the Panda Channel,[1] which contains daily routine videos of pandas at different ages (cubs, juveniles, and adults). The identity annotations are provided by professional zookeepers and breeders. To extract panda images from videos, we compute the similarities between adjacent video frames with the structural similarity index measure (SSIM) [50], which is defined as

$$\text{SSIM}_{x,y} = \frac{1}{M} \sum_{i=1}^{N} \frac{\left(2\mu_{x_i}\mu_{y_i} + c_1\right)\left(2\sigma_{x_i y_i} + c_2\right)}{\left(\mu_{x_i}^2 + \mu_{y_i}^2 + c_1\right)\left(\sigma_{x_i}^2 + \sigma_{y_i}^2 + c_2\right)}, \tag{7}$$

where μ_{x_i}, μ_{y_i} are the averages of the i_{th} pair of patches in the images x, y respectively, σ_x, σ_y are their variances, σ_{xy} is their covariance, and M is the total number of image patches. c_1 and c_2 are constants to prevent the denominator of Eq. (7) from being zero. In this way, only key frames that are different from their previous ones are retained.

We further manually select images with various illuminations, viewpoints, postures, and occlusions. In addition, we manually crop out each individual panda with a tight bounding box of varying aspect ratios. The iPanda-50 dataset consists of 6,874 images of 50 giant panda identities with $49 \sim 292$ images per panda identity. The split ratio of the training set and testing set is 2:1. Some sample images and statistics of this dataset are shown in Figure 7. This iPanda-50 dataset is available online,[2] and we are further expanding the dataset.

V. Experiments and Discussions

This section contains four parts. First, we introduce the implementation details of the proposed Feature-Fusion Network with Patch Detector (FFN-PD). Second, we evaluate the proposed method on the challenging iPanda-50 dataset. Third, we apply our method to conventional fine-grained visual recognition (FGVR) datasets, *i.e.*, CUB-200-2011 [13], Stanford Cars [32], and FGVC-Aircraft [51], which are most related to our task. Finally, we demonstrate ablation studies to prove the contribution of each component in the proposed FFN-PD.

A. Implementation Details

The proposed FFN-PD is implemented via PyTorch [52] and trained with 4 NVIDIA 1080Ti GPUs. We apply ImageNet [53] pre-trained ResNet50 [48] as the backbone in our experiments, and extract feature map at the layer3 for local stream. It is worth noting that our backbone is not tied to any specific networks. The regular stochastic gradient descent optimizer is used with a momentum of 0.9, a weight decay of 5×10^{-4}, and a batch size of 32. The initial learning rates are 0.01 and 0.1 for the pre-trained layers and the newly added layers, respectively. The learning rates are decayed by a factor

[1] Video streaming website, http://www.ipanda.com
[2] https://github.com/iPandaDateset/iPanda-50

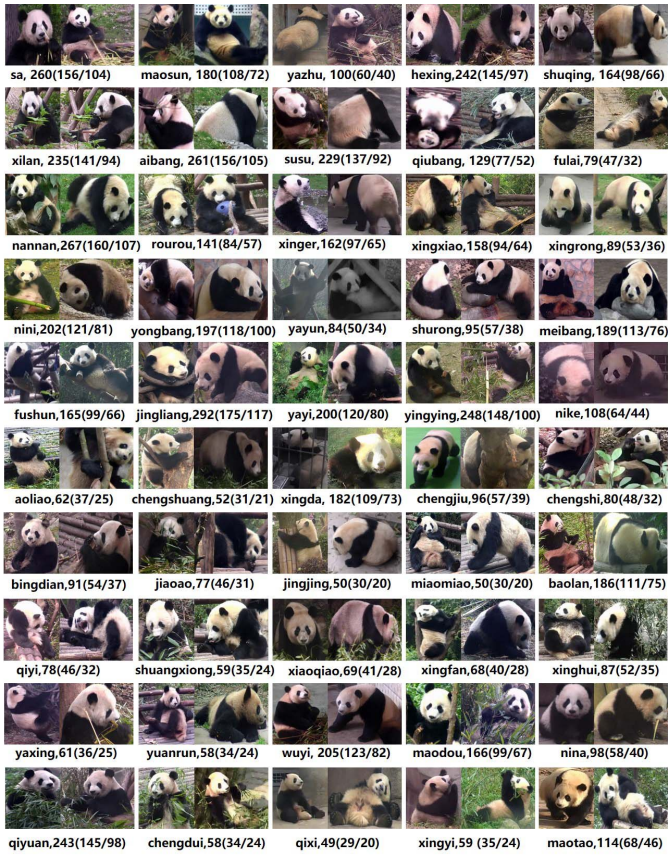


Fig. 7. Sample images and statistics of all the panda identities from the iPanda-50 dataset. Statistics are formatted as *identity*, #*images*(#*training-images*/#*testing-images*).

of 0.1 every 20 epochs. The number of the most discriminative patches per identity, *i.e.*, k, is empirically fixed to 10. Both the kernel size and the stride of the average pooling in the F-stream are 2. The weight parameter λ is set to 0.1 according to grid search.

As mentioned in previous sections, the receptive field of the deep layer is too large to locate object parts. To tackle this issue, we zoom in the input image, and thus the receptive fields of the subtle parts become larger compared to the original image at the same convolution layers. Specifically, for the giant panda identification task, all input images are resized to 448×448 regardless of aspect ratios, and augmented with random horizontal flipping during training, while they are only resized during testing. The FGVR settings on the CUB-200-2011, Stanford Cars, and FGVC-Aircraft datasets are highly similar, except that (1) all images are resized so that the shorter edge is 448 pixels wide (keeping aspect ratio unchanged), (2) training images are randomly cropped so that their sizes are 448×448, and (3) testing images are cropped at the center so that their sizes are 448×448. We use the multi-class classification accuracy as the evaluation metric.

B. Evaluation on the iPanda-50 Dataset

The performance comparison between the proposed FFN-PD and competing methods on the iPanda-50 dataset is summarized in Table I. All statistics in Table I are obtained

TABLE I

PERFORMANCE COMPARISON ON THE IPANDA-50 DATASET VIA 5 RANDOM TRIALS. THE MAXIMUM, MEAN, AND STANDARD DEVIATION OF THE ACCURACY ARE REPORTED AS MAX., MEAN, AND δ, RESPECTIVELY. P-VALUES AND ALTERNATIVE HYPOTHESIS H_1 CONFIDENCE IN A SERIES OF ONE-TAILED STUDENT'S T-TESTS (WITH NULL HYPOTHESIS H_0 BEING THERE IS NO EFFECTIVE ADVANTAGE OF THE PROPOSED FFN-PD OVER OTHERS) ARE REPORTED AS P-VALUE AND CONF., RESPECTIVELY

Method	Max. (%)	Mean (%)	δ	P-value	Conf.
Baseline	76.3	75.7	0.70	2.0e-6	>99%
MFC	80.8	79.9	1.00	1.2e-4	>99%
HBP [25]	78.4	77.8	0.48	4.1e-7	>99%
B-CNN [36]	77.3	76.9	0.30	1.3e-9	>99%
DFL-CNN [28]	84.8	84.1	0.49	2.0e-4	>99%
FFN-PD (ours)	**86.3**	**86.1**	**0.16**	-	-

with 5 independent random training/testing splits (4, 106 and 2, 768 images in training and testing splits, respectively). The maximum (Max.), mean (Mean), and standard deviation (δ) of the accuracy (%) over these 5 trials are reported. Additionally, to account for coincidental fluctuations and reveal statistical significance, a series of 5 one-tailed student's t-tests are carried out. The null hypothesis H_0 is there is no obvious advantage of the proposed FFN-PD over other competing methods. P-values and the confidence intervals (Conf.) of the alternative hypothesis H_1 being true are also reported for each competing methods. As presented in Section III, the "Baseline" in Table I denotes a generic classification network with a single G-Stream branch.

We also implement a multi-feature concatenation ("MFC" in Table I) method, which directly concatenates features from multiple convolution layers (*i.e.*, layer3 and layer4 of ResNet50). It outperforms the baseline, indicating the value of incorporating cross-layer features.

More importantly, we re-implement three FGVR methods, *i.e.*, (1) a classical bilinear pooling method BCNN [36], (2) a hierarchical bilinear pooling framework HBP [25], which concatenates multiple cross-layer bilinear features, (3) DFL-CNN [28], which learns a mid-level representation to capture identity-specific discriminative patches. The proposed FFN-PD outperforms all these competing ones with confidence intervals of $> 99\%$.

Figure 8 provides Grad-CAM [54] visualizations of the local filtered features for four different identities. The second column in Figure 8 is the feature visualization before training, and it shows that the high activations are scattered all over the image, and most of the patches with high response are located on the background instead of the target panda.

After training, such discriminative patches focus on the pandas, especially on their faces, as shown in the third column of Figure 8. Besides, for different panda identities, the patches may also focus on different locations (*e.g.*, eyes of "wuyi", nose of "sa", back of "qiyuan", and ears of "susu"), which demonstrates patch detectors can detect identity-specific discriminative parts. We further remap the local patches with the top-3 activations back to the original image in the fourth

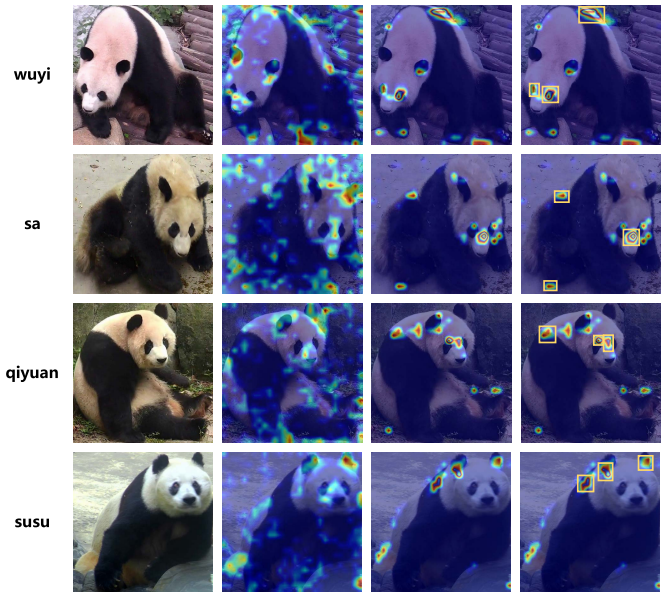


Fig. 8. The Grad-CAM [54] visualization of the feature maps of different panda identities at the end of the local L-Stream branch. The first column is the original images, the second column is the feature visualization before training, the third column shows the active regions after training, and the last column shows the top-3 patches remapped onto the original image.

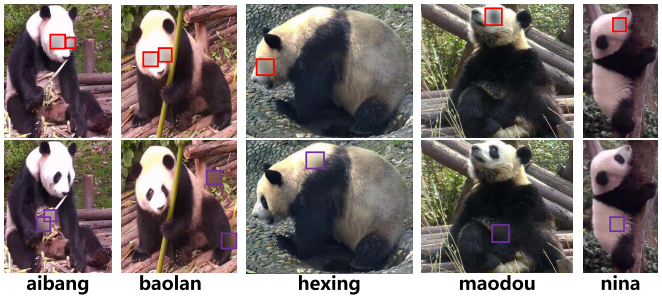


Fig. 9. The visualization of covering different regions of the giant panda with Gaussian blur. We cover the giant panda' eyes on the image as shown in the first row, and randomly cover other regions except the eyes shown in the second row.

column. We are surprised that these visualizations agree well with [55], which claims that such black eye patches may help pandas recognize one another.

In order to verify the aforementioned conjecture that the features around the eye regions of the giant panda may promote the accuracy of panda identification, we further perform the following experiments. We use Gaussian blur to cover the panda eyes in the iPanda-50 dataset (as shown in the first row of Figure 9). Since occlusion based on Gaussian blur also brings noise, it is difficult to directly judge whether the factor affecting the accuracy is the Gaussian blur itself or the occlusion of the panda's eyes. Therefore, we also use Gaussian blur occlusion on other regions besides eyes (as shown in the second row of Figure 9), and the number of Gaussian blur occlusions is set to the same as the number of eyes exposed in the image (due to panda's pose, some image may show one eye or two eyes). In addition, random occlusion also brings random errors. The best way to eliminate random errors is to repeat the similar process and use the mean to offset the random errors. Therefore, we randomly conduct five independent Gaussian

TABLE II

PERFORMANCE COMPARISON BETWEEN COVERING PANDA EYES AND RANDOMLY COVER OTHER REGIONS ON THE IPANDA-50 DATASET

iPanda-50	cover eyes	randomly cover other parts					
		r1	r2	r3	r4	r5	Mean.
86.3 %	82.2	84.4	84.6	84.0	84.9	84.7	84.5

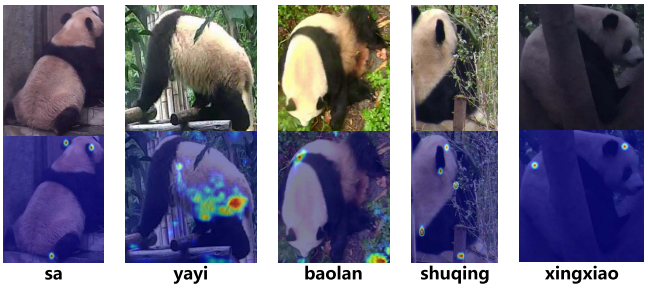


Fig. 10. Failure cases of the proposed FFN-PD due to the bad illumination condition and the back posture that does not show obvious discriminative characteristics.

blur occlusions on other regions. Finally, we evaluate with our proposed algorithm on the iPanda-50 dataset after Gaussian blur processing.

The results are summarized in Table II. We obtain an accuracy of 82.2% on the same iPanda-50 dataset split by only covering the giant panda's eyes, which is 4.1% lower than 86.3% without any cover processing. Meanwhile, the mean accuracy of the five independent random occlusions is 84.5%, which slightly degrades the performance. It indicates that Gaussian blur does partly reduce the identification performance, but the impact it brings is far worse when covering the eyes. We believe that the features around eyes could affect panda identification. The experimental results indicate that our method can well learn the features around the eyes without any part-level supervision. It further illustrates the effectiveness and feasibility of the proposed method.

Figure 10 shows failure cases with the proposed FFN-PD. Specifically, we speculate that the back view of giant pandas "sa" and "yayi" and the poor illumination of the giant panda "xingxiao" account for such identification failures.

Additionally, a closer look reveals that a small portion of the learned discriminative image regions are unfortunately related to background objects, which indicates the network might rely on panda's background/habitat to distinguish their identities. We speculate that this phenomenon results from the limited dataset size and diversity.

C. Evaluation on Other FGVR Datasets

1) FGVR Datasets: To further demonstrate the effectiveness of the proposed method, we conduct experiments on three FGVR datasets (*i.e.*, CUB-200-2011 [13], Stanford Cars [32], and FGVC-Aircraft [51]).

- CUB-200-2011 [13] is an avian species classification dataset which contains 11, 788 images of 200 categories. The ratio of training images and test images is about 1 : 1. We use the publicly available split [31] in our experiments.
- Stanford Cars [32] contains 16,185 images of 196 car categories with a roughly 50% – 50% split of each

category. The car images of this dataset usually show various angles and sizes, and the categories are typically divided according to the production year and car model.
- FGVC-Aircraft [51] contains 10,000 images of 100 categories. The ratio of training and testing images is 2 : 1. The airplane images of this dataset are assigned in four levels from finer to coarser, *i.e.*, Model, Variant, Family, and Manufacturer.

2) Baselines: We compare the proposed FFN-PD with 15 existing state-of-the-art FGVR methods, including four part-based methods [16]–[19] with bounding box/part annotations, four region-attention methods [20], [21], [23], [35] using attention maps with image-level labels, and seven end-to-end (*i.e.*, one-stage) methods [11], [24], [25], [28], [36], [56], [57] with image-level labels. All baselines are listed as follows.

- **Part-RCNN** [17]: a part-based model that extends R-CNN [58] to extract features based on bottom-up region proposals with part annotations.
- **DeepLAC** [16]: a deep location, alignment, and classification architecture that forms a valve linkage function for simple back-propagation and recognizes in pose-aligned part images.
- **PS-CNN** [18]: a part-stacked CNN architecture that performs object part localization with a fully convolutional network and simultaneously encodes object-level and part-level features.
- **Mask-CNN** [19]: a mask-CNN model that contains a fully convolutional network to learn the discriminative part masks and uses these masks to select deep descriptors.
- **RA-CNN** [20]: a recurrent attention CNN that begins with whole images and combines the previous result to iteratively generate region areas from coarse to fine stages.
- **MA-CNN** [21]: a multi-attention CNN that clusters the spatially-correlated channels to generate multiple parts and learns fine-grained features based on these parts in a mutual reinforced way.
- **MAMC** [35]: a multi-attention multi-class constraint method that learns attention maps in the one-squeeze multi-excitation module and then regularizes features in a metric learning manner.
- **MGE-CNN** [23]: a mixture of granularity-specific experts approach that learns experts with former experts to focus on finer regions and guides each expert to produce diverse prediction distribution via a Kullback-Leibler constraint.
- **B-CNN** [36]: a bilinear CNN model that extracts pairwise feature interactions for fine-grained recognition in an end-to-end training.
- **Compact B-CNN** [56]: a compact bilinear CNN that reduces feature dimensions with the same discriminative power compared with B-CNN [36].
- **Kernel-Pooling** [24]: a kernel pooling method that uses the form of kernels to capture higher order feature interactions for fine-grained recognition.
- **Low-rank B-CNN** [57]: a low-rank bilinear pooling method that proposes the covariance feature representation with a low-rank bilinear classifier to reduce compute time.
- **HBP** [25]: a hierarchical bilinear pooling approach that captures the inter-layer part feature relations and integrates multiple cross-layer bilinear features.
- **DFL-CNN** [28]: a discriminative filter bank model that exploits mid-level representation and learns identity-specific patches by the filter bank.
- **DCL** [11]: a destruction and construction learning model that enhances the difficulty of recognition by destructing images and then reconstructs images to learn fine-grained features.

3) Results on the CUB-Birds Dataset: We first conduct the experiment on the CUB-200-2011 dataset, which not only provides the class label but also provides additional bird part annotations including beak, eyes, nape, wing, and tail *etc*. We compare the proposed FFN-PD with 15 exiting FGVR methods on this dataset, and the detailed discussion is as below.

Of the 15 competing FGVR methods, Part-CNN [17], DeepLAC [16], PS-CNN [18], and Mask-CNN [19] learn fine-grained features based on various parts by using location networks to locate object parts with additional part annotations. Specifically, the located parts should be shared across categories, which means such part representations are similar, but the later fine-grained learning encourages these subtle parts to be different. Thus, it should balance the localization and classification networks, which is hard to achieve in practice. B-CNN [36] performs classification via high-dimensional feature representation with supplementary object bounding boxes. Due to their privileged access with additional bounding box/part annotations, their performances are not fairly comparable with others. Nevertheless, our method surpasses these methods with 12.2%, 8.3%, 12.0%, 1.3%, and 3.5% relative accuracy gains, respectively.

RA-CNN [20], MA-CNN [21], MAMC [35] and MGE-CNN [23] rely on attention maps to facilitate the fine-grained feature learning with only image-level labels. However, such methods almost require additional architectures such as the attention network to locate discriminative parts or encode region features, which leads to more computation both in training and testing. For example, MA-CNN [21] consists of the convolution, channel grouping, and part classification sub-networks, which requires alternative optimization of each sub-network. The most recent method MGE-CNN [23] achieves a high classification accuracy that is closest to ours on this dataset. But it iteratively generates region-specific experts in multiple stages, which will be in an extreme predicament when the previous stage focuses on error attention regions.

Recent methods that can be trained end-to-end (*i.e.*, one-stage) are also included, such as B-CNN [36], Compact B-CNN [56], Low-rank B-CNN [57], and Kernel-Pooling [24], which typically exploit very high-dimensional features compared to the other two groups. For instance, Kernel-Pooling [24] encodes higher order interaction representation for fine-grained feature learning, but its dimension is still

TABLE III

PERFORMANCE COMPARISON BETWEEN THE PROPOSED FFN-PD AND 15 EXITING FGVR METHODS ON THE CUB-200-2011 DATASET. THE FIRST GROUP USES LOCATION SUB-NETWORK WITH BOUNDING BOX/PART ANNOTATIONS. THE SECOND GROUP LEVERAGES REGION ATTENTION MAPS WITH ONLY IMAGE-LEVEL LABELS. THE THIRD GROUP PERFORMS IN AN END-TO-END MANNER (*i.e.*, ONE-STAGE) WITH IMAGE-LEVEL LABELS

Method	Backbone	BBox/Parts	1-Stage	Accuracy
Part-RCNN [17]	AlexNet	✓		76.4
DeepLAC [16]	AlexNet	✓		80.3
PS-CNN [18]	AlexNet	✓		76.6
Mask-CNN [19]	ResNet-50	✓		87.3
B-CNN [36]	VGG-16	✓	✓	85.1
RA-CNN [20]	VGG-19			85.3
MA-CNN [21]	VGG-19		✓	86.5
MAMC [35]	ResNet-50		✓	86.5
MGE-CNN [23]	ResNet-50			88.5
B-CNN [36]	VGG-16		✓	84.1
Compact B-CNN [56]	VGG-16		✓	84.0
Kernel-Pooling [24]	VGG-16		✓	86.2
Low-rank B-CNN [57]	VGG-16		✓	84.2
HBP [25]	VGG-16		✓	87.1
DFL-CNN [28]	ResNet-50		✓	87.4
DCL [11]	ResNet-50		✓	87.8
FFN-PD (ours)	ResNet-50		✓	**88.6**

TABLE IV

PERFORMANCE COMPARISON BETWEEN THE PROPOSED FFN-PD AND 10 EXITING FGVR METHODS ON THE STANFORD CARS DATASET. THE FIRST GROUP LEVERAGES REGION ATTENTION MAPS WITH ONLY IMAGE-LEVEL LABELS. THE SECOND GROUP PERFORMS IN AN END-TO-END MANNER (*i.e.*, ONE-STAGE) WITH IMAGE-LEVEL LABELS

Method	Backbone	1-Stage	Accuracy
RA-CNN [20]	VGG-19		92.5
MA-CNN [21]	VGG-19	✓	92.8
MAMC [35]	ResNet-50	✓	93.0
MGE-CNN [23]	ResNet-50		93.9
B-CNN [36]	VGG-16	✓	91.3
Kernel-Pooling [24]	VGG-16	✓	92.4
Low-rank B-CNN [57]	VGG-16	✓	90.9
HBP [25]	VGG-16	✓	93.7
DFL-CNN [28]	ResNet-50	✓	93.1
DCL [11]	ResNet-50	✓	94.5
FFN-PD (ours)	ResNet-50	✓	**94.7**

too large. Here the weakly-supervised B-CNN [36] is reported only with image-level labels, whose accuracy (84.1%) is lower than the counterpart with bounding box annotations (85.1%). In addition, DCL [11] proposes a "Destruction and Construction Learning" method to increase recognition difficulty and make the network learn expert knowledge, however, its training process is complicated. HBP [25] and DFL-CNN [28] have been discussed in the previous section.

As summarized in Table III, the proposed FFN-PD nevertheless outperforms all 15 competing algorithms, even including those with privileged access to additional bounding box/part annotations.

4) Results on the Stanford Cars Dataset: The classification accuracy on the Stanford Cars dataset is present in Table IV. Stanford Cars dataset does not have part annotations, thus the part-based methods are not reported in Table IV. Our method achieves the best performance against other state-of-the-art methods. Compared to the region-attention method MGE-CNN [23], which learns a mixture of granularity-specific experts in multiple stages, the proposed FFN-PD outperforms it by 0.8%. We can also observe that our method surpasses end-to-end methods. Although DCL [11] attains a high accuracy on this dataset, it needs a special stage for destruction initialization. Our method is much simpler and can surpass it.

TABLE V

PERFORMANCE COMPARISON BETWEEN THE PROPOSED FFN-PD AND 8 EXITING FGVR METHODS ON THE FGVC-AIRCRAFT DATASET. ALL METHODS PERFORM IN AN END-TO-END MANNER (*i.e.*, ONE-STAGE) WITH IMAGE-LEVEL LABELS, AND THE FIRST GROUP LEVERAGES REGION ATTENTION MAPS WITH IMAGE-LEVEL LABELS

Method	Backbone	1-Stage	Accuracy
RA-CNN [20]	VGG-19		88.2
MA-CNN [21]	VGG-19	✓	89.9
B-CNN [36]	VGG-16	✓	84.1
Kernel-Pooling [24]	VGG-16	✓	86.9
Low-rank B-CNN [57]	VGG-16	✓	87.3
HBP [25]	VGG-16	✓	90.3
DFL-CNN [28]	ResNet-50	✓	91.7
DCL [11]	ResNet-50	✓	93.0
FFN-PD (ours)	ResNet-50	✓	**93.2**

5) Results on the FGVC-Aircraft Dataset: The classification results on the FGVC-Aircraft dataset are shown in Table V. The FGVC-Aircraft dataset also does not contain part annotations, thus we compare our method with two groups of methods including region-attention methods and end-to-end methods. Obviously, the proposed FFN-PD obtains the best classification performance among these methods. Due to our multiple representations, we surpass DFL-CNN [28] by 1.5% relative accuracy gains, which also exploits discriminative identity-specific patches. Furthermore, we still outperform DCL [11] and region-attention methods (*e.g.*, RA-CNN [20] and MA-CNN [21]), which further demonstrates the significance of the proposed FFN-PD.

D. Ablation Studies

To validate the contribution of each component in the proposed FFN-PD, we conduct a set of ablation experiments on the iPanda-50 dataset.

1) Different Stream Combinations: To validate the effectiveness of each stream in the proposed FFN-PD, we conduct experiments on the iPanda-50 dataset with different stream combinations. The results are summarized in Table VI. With a single stream, the performance deteriorates obviously. The combination of G-Stream with any other streams boosts their performance obviously, indicating the necessity of including global image features as visualized in Figure 11 (b). Moreover, methods including F-Stream always perform better than those without it (*e.g.*, L-Stream + F-Stream versus L-Straem), which agrees with our speculation that the hierarchical representation is beneficial. Especially, the fusion feature (visualized in Figure 11 (e)) is fused by the local filtered feature (visualized in Figure 11 (c)) and the global filtered feature (visualized

TABLE VI

IDENTIFICATION PERFORMANCE COMPARISON OF DIFFERENT COMBINATIONS ON THE IPANDA-50 DATASET. MAXIMUM, MEAN, AND STANDARD DEVIATION OF THE ACCURACY (%) ARE REPORTED FROM 5 RANDOM TRIALS. AVG AND ATT DENOTE THE AVERAGE CONVOLUTION FILTER SUPERVISION PROPOSED IN [28] AND OUR PROPOSED ATTENTIONAL CONVOLUTION FILTER SUPERVISION, RESPECTIVELY

Method					Max. Acc.	Mean Acc.	STD Dev.
Global	Local	Fusion	Avg	Att			
✓					76.3	75.7	0.70
	✓				73.8	73.4	0.38
		✓			60.3	58.7	1.10
	✓	✓			74.8	74.6	0.15
✓	✓				84.2	83.8	0.13
✓		✓			84.7	84.4	0.24
	✓	✓		✓	75.4	75.1	0.16
✓	✓			✓	85.8	85.6	0.17
✓	✓	✓			85.9	85.5	0.27
✓	✓	✓	✓		86.0	85.4	0.65
✓	✓	✓		✓	**86.3**	**86.1**	0.16

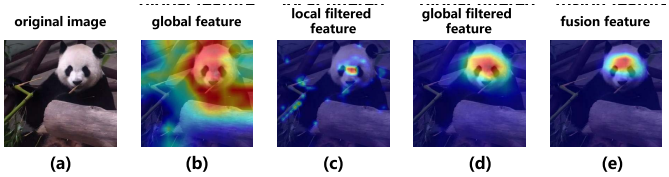


Fig. 11. Grad-CAM [54] visualizations of feature maps of (a) the original image, (b) the global feature map at the G-Stream, (c) the local filtered feature map at the L-Stream, (d) the global filtered feature map, and (e) the fusion feature map at the F-Stream.

in Figure 11 (d)), which further activates regions of high responses and assists the network to emphasize important regions and suppress background noise. Note that in our proposed method, filter supervision (*i.e.*, Avg and Att columns) must be evaluated in conjunction with the L-Stream, where the patch detectors are used to detect discriminative patches.

The visualization of the local GMP features $\mathbf{v}_l$ is shown in Figure 12. Thanks to the design of the attentional filter supervision, the network has learned a set of supervised patch detectors, where patch-level representations could present the highest activation value in the corresponding identity.

2) Effect of Attentional Filter Supervision: We conduct experiments on the iPanda-50 dataset with different convolution filter supervisions in the proposed FFN-PD. The last three rows of Table VI present their identification accuracies, which are characterized by columns "Avg" (as in Eq. (3), described in [28]) and "Att" (proposed by us in Eq. (4)). They are all combined with all three streams. The results show that both average pooling and attentional pooling are effective, with our proposed attentional pooling outperforms the average pooling, which indicates the effectiveness of the attentional pooling.

Finally, the combination of three streams and our proposed attentional pooling achieves the best result, which supports our claim that the patch-level information, the global information, and the hierarchical representation jointly contribute to the overall performance.

3) Size of the iPanda Dataset: The number of identities in the iPanda-50 dataset is relatively small (*i.e.*, 50 panda

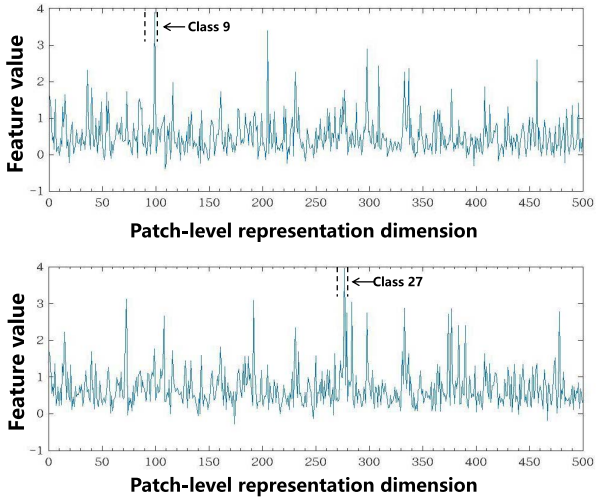


Fig. 12. The visualization of the local GMP features $\mathbf{v}_l$. For the samples in a given testing class (*e.g.*, class 9 and class 27) in iPanda-50, the peak feature value produced by the discriminative patch detector locates at the corresponding class donated by the dash line.

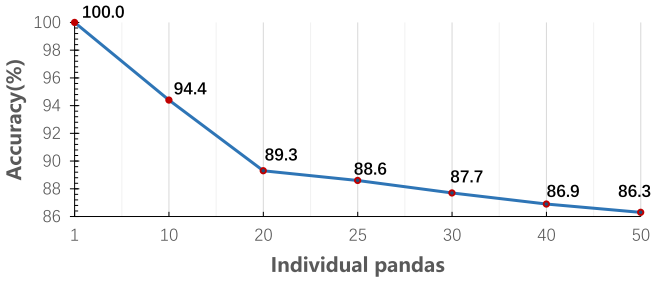


Fig. 13. The identification accuracy versus the number of panda identities curve on the iPanda-50 dataset.

identities), and we speculate that the increase/decrease of the panda dataset size should also increase/decrease its identification difficulty. To verify this hypothesis, we conduct an experiment on different subsets of the iPanda-50 dataset with the proposed FFN-PD, and the categories in each subset are randomly selected as shown in Figure 13. For example, the accuracy improves significantly if the number of individual pandas is largely reduced to 10. This also indicates that a more challenging dataset could potentially be built by including more individual pandas.

VI. CONCLUSION

We propose a Feature-Fusion Network with Patch Detector (FFN-PD) to address the important yet challenging Giant Panda Identification (GPID) problem. The proposed FFN-PD exploits discriminative local image patches in each image via the patch detector without any bounding box/part annotations, and fuses both global and local features to generate a hierarchical representation, which effectively improves identification performance. Specifically, a new attentional cross-channel pooling module is proposed to provide more effective training supervision of the convolution filters in the patch detectors. These multiple feature representation simultaneously facilitate the recognition performance. Moreover, we propose a new iPanda-50 dataset to evaluate the proposed FFN-PD and existing FGVR algorithms on the GPID task, where the proposed

FFN-PD outperforms other methods by a large margin, thus verify the effectiveness of the proposed method. In addition, the eye-covering experiment indicates that visual features around eyes play a significant role in panda identification.

REFERENCES

[1] W. M. Matkowski, A. W. K. Kong, H. Su, P. Chen, R. Hou, and Z. Zhang, "Giant panda face recognition using small dataset," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1680–1684.

[2] L. He, H. Li, Q. Zhang, and Z. Sun, "Dynamic feature matching for partial face recognition," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 791–802, Feb. 2019.

[3] S. Ge, S. Zhao, C. Li, and J. Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2051–2062, Apr. 2019.

[4] C. Peng, N. Wang, J. Li, and X. Gao, "Re-ranking high-dimensional deep local representation for NIR-VIS face recognition," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4553–4565, Sep. 2019.

[5] Y.-J. Cho and K.-J. Yoon, "PaMM: Pose-aware multi-shot matching for improving person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3739–3752, Aug. 2018.

[6] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.

[7] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, "Dynamic graph co-matching for unsupervised video-based person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2976–2990, Jun. 2019.

[8] H. Zheng, J. Fu, Z.-J. Zha, J. Luo, and T. Mei, "Learning rich part hierarchies with progressive attention networks for fine-grained image recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 476–488, 2020.

[9] 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.

[10] W. Luo *et al.*, "Cross-X learning for fine-grained visual categorization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8242–8251.

[11] 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. 5157–5166.

[12] H. Yao, S. Zhang, C. Yan, Y. Zhang, J. Li, and Q. Tian, "AutoBD: Automated bi-level description for scalable fine-grained visual categorization," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 10–23, Jan. 2018.

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

[14] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[15] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 87–102.

[16] D. Lin, X. Shen, C. Lu, and J. Jia, "Deep LAC: Deep localization, alignment and classification for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1666–1674.

[17] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 834–849.

[18] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked CNN for fine-grained visual categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1173–1182.

[19] X.-S. Wei, C.-W. Xie, J. Wu, and C. Shen, "Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization," *Pattern Recognit.*, vol. 76, pp. 704–714, Apr. 2018.

[20] 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. 4438–4446.

[21] 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. 5209–5217.

[22] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 842–850.

[23] 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. 8331–8340.

[24] 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. 2921–2930.

[25] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical bilinear pooling for fine-grained visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 574–589.

[26] X. Wei, Y. Zhang, Y. Gong, J. Zhang, and N. Zheng, "Grassmann pooling as compact homogeneous bilinear pooling for fine-grained visual classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 355–370.

[27] 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.

[28] 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.

[29] R. Ding, L. Wang, Q. Zhang, Z. Niu, N. Zheng, and G. Hud, "Fine-grained giant panda identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 2108–2112.

[30] E. Rodner, M. Simon, G. Brehm, S. Pietsch, J. W. Wägele, and J. Denzler, "Fine-grained recognition datasets for biodiversity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2015, pp. 1–7.

[31] 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.

[32] 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.

[33] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2011, pp. 1–2.

[34] X. He and Y. Peng, "Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification," in *Proc. AAAI. Conf. Artif. Intell.*, 2017, pp. 4075–4081.

[35] 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.*, 2018, pp. 805–821.

[36] 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.

[37] 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.

[38] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 729–736.

[39] 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. 6599–6608.

[40] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.

[41] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2129–2137.

[42] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[43] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 447–456.

[44] T. Kong, A. Yao, Y. Chen, and F. Sun, "HyperNet: Towards accurate region proposal generation and joint object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 845–853.

[45] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

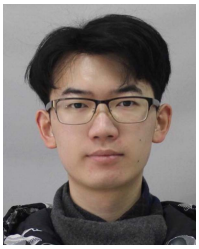
[46] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, and Z. Zhang, "Multiple granularity descriptors for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2399–2406.

[47] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[49] 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.
[50] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[51] 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
[52] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.
[53] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[54] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[55] V. Morell, "How pandas got their patches," in *Science*. Washington, DC, USA: American Association for the Advancement of Science, 2017.
[56] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 317–326.
[57] S. Kong and C. Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 365–374.
[58] 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.

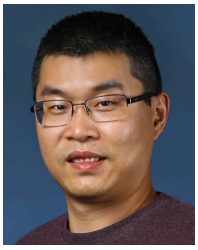
Le Wang (Senior Member, IEEE) received the B.S. and Ph.D. degrees in control science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2008 and 2014, respectively. From 2013 to 2014, he was a Visiting Ph.D. Student with the Stevens Institute of Technology, Hoboken, New Jersey, USA. From 2016 to 2017, he was a Visiting Scholar with Northwestern University, Evanston, Illinois, USA. He is currently an Associate Professor with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China. His research interests include computer vision, pattern recognition, and machine learning. He is the author of more than 50 peer reviewed publications in prestigious international journals and conferences.

Rizhi Ding (Student Member, IEEE) received the B.S. degree in the measurement and control technology and instrumentation from Chang'an University, Xi'an, China, in 2017. He is currently pursuing the M.S. degree with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests include image processing and computer vision.



Yuanhao Zhai (Member, IEEE) received the B.Eng. degree in computer science from Xi'an Jiaotong University, Xi'an, China, in 2020. He is currently a Research Intern with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China. His research interests lie in computer vision and machine learning.

Qilin Zhang (Member, IEEE) received the B.E. degree in electrical information engineering from the University of Science and Technology of China, Hefei, China, in 2009, the M.S. degree in electrical and computer engineering from University of Florida, Gainesville, Florida, USA, in 2011, and the Ph.D. degree in computer science from the Stevens Institute of Technology, Hoboken, New Jersey, USA, in 2016. He is currently a Senior Research Scientist with ABB Corporate Research Center, Raleigh, NC, USA. Before that, he was a Senior Research Engineer from 2016 to 2018 and then a Lead Research Engineer from 2018 to 2020 with HERE Technologies, Chicago, IL, USA. His research interests include computer vision and signal processing.



Wei Tang (Member, IEEE) received the B.E. and M.E. degrees from Beihang University, Beijing, China, in 2012 and 2015 respectively, and the Ph.D. degree in electrical engineering from Northwestern University, Evanston, Illinois, USA, in 2019. He is currently an Assistant Professor with the Department of Computer Science, University of Illinois at Chicago. His research interests include computer vision, pattern recognition, and machine learning.

Nanning Zheng (Fellow, IEEE) graduated from the Department of Electrical Engineering, Xi'an Jiaotong University (XJTU), Xi'an, China, in 1975. He received the M.E. degree in information and control engineering from Xi'an Jiaotong University, Xi'an, China, in 1981, and the Ph.D. degree in electrical engineering from Keio University, Keio, Japan, in 1985. He is currently a Professor and the Director with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China. His research interests include computer vision, pattern recognition, computational intelligence, and hardware implementation of intelligent systems. Since 2000, he has been the Chinese representative on the Governing Board of the International Association of Pattern Recognition. He became a member of the Chinese Academy Engineering in 1999.

Gang Hua (Fellow, IEEE) received the B.S. degree in automatic control engineering and the M.S. degree in control science and engineering from XJTU in 1999 and 2002, respectively, and the Ph.D. degree in electrical engineering and computer science from Northwestern University, Evanston, Illinois, USA, in 2006. He was enrolled in the Special Class for the Gifted Young of Xi'an Jiaotong University (XJTU), Xi'an, China, in 1994. He is currently the Vice President and a Chief Scientist of Wormpex AI Research. Before that, he served in various roles at Microsoft from 2015 to 2018 as the Science/Technical Adviser to the CVP of the Computer Vision Group, the Director of Computer Vision Science Team, Redmond and Taipei ATL, and a Principal Researcher/Research Manager at Microsoft Research. He was an Associate Professor at the Stevens Institute of Technology from 2011 to 2015. From 2014 to 2015, he took an on leave and worked on the Amazon-Go project. He was an Visiting Researcher from 2011 to 2014 and a Research Staff Member from 2010 to 2011 at the IBM Research T. J. Watson Center, a Senior Researcher from 2009 to 2010 at the Nokia Research Center Hollywood, and a Scientist from 2006 to 2009 at Microsoft Live Labs Research. He is an Associate Editor of TIP, TCSVT, CVIU, IEEE MULTIMEDIA, TVCJ, and MVA. He also served as the Lead Guest Editor on two special issues in TPAMI and IJCV, respectively. He is a General Chair of ICCV'2025. He is a Program Chair of CVPR'2019&2022. He is an Area Chair of CVPR'2015&2017, ICCV'2011&2017, ICIP'2012&2013&2016, ICASSP'2012&2013, and ACM MM 2011&2012&2015&2017. He is the author of more than 190 peer reviewed publications in prestigious international journals and conferences. He holds 19 U.S. patents and has 15 more U.S. patents pending. He was a recipient of the 2015 IAPR Young Biometrics Investigator Award for his contribution on Unconstrained Face Recognition from Images and Videos, and a recipient of the 2013 Google Research Faculty Award. He is an IAPR Fellow and an ACM Distinguished Scientist.