2020

DPSA: Dense pixelwise spatial attention network for hatching egg fertility detection

Lei Geng

Yunyun Xu

Zhitao Xiao

Jun Tong
University of Wollongong, jtong@uow.edu.au

DPSA: Dense pixelwise spatial attention network for hatching egg fertility detection

Abstract

© 2020 SPIE and IS & T. Deep convolutional neural networks show a good prospect in the fertility detection and classification of specific pathogen-free hatching egg embryos in the production of avian influenza vaccine, and our previous work has mainly investigated three factors of networks to push performance: depth, width, and cardinality. However, an important problem that feeble embryos with weak blood vessels interfering with the classification of resilient fertile ones remains. Inspired by fine-grained classification, we introduce the attention mechanism into our model by proposing a dense pixelwise spatial attention module combined with the existing channel attention through depthwise separable convolutions to further enhance the network class-discriminative ability. In our fused attention module, depthwise convolutions are used for channel-specific features learning, and dilated convolutions with different sampling rates are adopted to capture spatial multiscale context and preserve rich detail, which can maintain high resolution and increase receptive fields simultaneously. The attention mask with strong semantic information generated by aggregating outputs of the spatial pyramid dilated convolution is broadcasted to low-level features via elementwise multiplications, serving as a feature selector to emphasize informative features and suppress less useful ones. A series of experiments conducted on our hatching egg dataset show that our attention network achieves a lower misjudgment rate on weak embryos and a more stable accuracy, which is up to 98.3% and 99.1% on 5-day and 9-day old eggs, respectively.

Disciplines

Engineering | Science and Technology Studies

Publication Details

DPSA: dense pixelwise spatial attention network for hatching egg fertility detection

Lei Geng
Yunyun Xu
Zhitao Xiao
Jun Tong

DPSA: dense pixelwise spatial attention network for hatching egg fertility detection

Lei Geng,[a,b] Yunyun Xu,[a,b] Zhitao Xiao,[a,b,*] and Jun Tong[c]
[a]Tianjin Polytechnic University, School of Electronics and Information Engineering, Tianjin, China
[b]Tianjin Key Laboratory of Optoelectronic Detection Technology and Systems, Tianjin, China
[c]University of Wollongong, School of Electrical, Computer and Telecommunications Engineering, Wollongong, Australia

Abstract. Deep convolutional neural networks show a good prospect in the fertility detection and classification of specific pathogen-free hatching egg embryos in the production of avian influenza vaccine, and our previous work has mainly investigated three factors of networks to push performance: depth, width, and cardinality. However, an important problem that feeble embryos with weak blood vessels interfering with the classification of resilient fertile ones remains. Inspired by fine-grained classification, we introduce the attention mechanism into our model by proposing a dense pixelwise spatial attention module combined with the existing channel attention through depthwise separable convolutions to further enhance the network class-discriminative ability. In our fused attention module, depthwise convolutions are used for channel-specific features learning, and dilated convolutions with different sampling rates are adopted to capture spatial multiscale context and preserve rich detail, which can maintain high resolution and increase receptive fields simultaneously. The attention mask with strong semantic information generated by aggregating outputs of the spatial pyramid dilated convolution is broadcasted to low-level features via elementwise multiplications, serving as a feature selector to emphasize informative features and suppress less useful ones. A series of experiments conducted on our hatching egg dataset show that our attention network achieves a lower misjudgment rate on weak embryos and a more stable accuracy, which is up to 98.3% and 99.1% on 5-day and 9-day old eggs, respectively. © 2020 SPIE and IS&T [DOI: 10.1117/1.JEI.29.2.023011]

Keywords: hatching eggs; fertility detection; convolutional neural network; classification; spatial attention; depthwise separable convolution; dilated convolution.

1 Introduction

The most mature and safe avian influenza vaccine cultivation method, recognized by academia and industry, is the chicken embryo method. Currently, avian influenza vaccines are usually produced by brewing live influenza strains from pathogen-free eggs. During the live hatching process, dead embryos can easily breed bacteria and contaminate other embryos. The cost and damage caused by dead embryos can be great. Therefore, the embryo activity detection and classification is a significant research goal for the production of avian influenza vaccine. Fertility detection of hatching eggs can usually be divided into four periods: 5-day, 9-day, 14-day, and 16-day, whereas hatching eggs have different features during different hatching periods. Currently, the detection and classification of egg embryo fertility use traditional methods, e.g., by manually determining whether embryonic vascular characteristics of eggs are viable. This approach requires a large amount of labor and time. The results are also susceptible to bias based on subjective factors. In addition, due to the high-intensity work pressure, workers experience visual fatigue and low detection efficiency, resulting in a high rate of false detections

---

*Address all correspondence to Zhitao Xiao, E-mail: xiaozhitao@tjpu.edu.cn

and missed inspections, which is difficult to meet the high standard requirements of the modern embryo detection and classification industries.

There are many traditional methods of detecting embryo activity in eggs. Bioelectrical detection[1] began in the 20th century. Ultrasonic image-based detection[2] soon followed, which led to hyperspectral imaging technology.[3–5] Finally, multi-information fusion technology[6,7] developed from the abundance of technological imaging and detection methodology. Romanoff and Frank[1] designed a radio-frequency-based measurement circuit to determine the electrical conductivity and dielectric constants of the embryos; from these features, they were able to determine the activity of the embryos. Mcquinn et al.[2] introduced ultrasonic imaging technology to detect the embryonic activity; they successfully solved the problem of poor visibility of embryos after a 5-day incubation. A hyperspectral imaging system that measured egg activity was proposed by Smith et al.[3]; this method was the first to use the hyperspectral images and data to detect the hatching eggs. Jones et al.[4] developed an artificial network algorithm to detect embryonic from hyperspectral images, but the method had low accuracy due to a lack of samples. Liu and Ngadi[5] developed a near-infrared hyperspectral imaging system to detect the activity of young embryos via textural information that was extracted from egg hyperspectral images. Then, Wei et al.[6] proposed a method that fuses a computer vision technique and an impact excitation technique, where the computer vision model adopts a learning vector quantization artificial neural network. Xu et al.[7] also established a back propagation neural network by fusing the images that contained the egg embryo blood vessels, black spots extracted from RGB space, mean and standard of each component in the Lab color space, temperature, and transmittance. In recent years, some innovative approaches adopting machine vision technology based on blood vessel processing have been proposed to improve the detection efficiency of hatching eggs in industrial production. In 2014, the SUSAN operator, a multilayer feature extraction method, was employed to remove high-brightness speckle noise to more accurately extract the blood vessel information. Then, the percentage of the vascular region was calculated to determine the activity of the embryonic eggs.[8] A weight fuzzy C-means algorithm was also used by Shan[9] for adaptive segmentation and to extract the major vascular information. Despite the success of these methods mentioned above, these technologies are either destructive or based on traditionally complicated image processing, such as image enhancement, image segmentation, etc. Due to the low efficiency of feature extraction, these approaches cannot be applied in actual production.

With the development of frameworks based on deep learning, many modern convolution networks[10–13] have been developed for image classification tasks. In previous work, we have modified several convolutional neural networks (CNNs) based on existing popular models for specific period hatching-egg activity detection. In 2017, the TB-CNN,[14] a CNN-based structure that was divided into two branches, was raised to realize the 5-day old hatching-egg classification. The feature extraction, by adopting a series of convolutional layers based on deep learning, achieved a commendable detection accuracy. Later, a hatching-egg classification method, based on CNN with a channel weighting method and joint supervision model,[15] was proposed for 9-day eggs. We also tried predicting embryos viability by detecting heartbeat signals based on fully convolutional networks and a gated recurrent unit method.[16] Even though we have proposed several CNN-based models to solve classification of our hatching eggs via vascular information, most of them internally treat all types of information equally and may not efficiently distinguish the most discriminative characteristic. The remaining problem is that weak embryos with local thin blood vessels, which are similar to the fertile embryos, interfere with the classification accuracy.

Recently, the benefits of neural networks combined with attention mechanisms have been shown across a range of tasks in the vision field. One work[17] introduced a novel attention mechanism in language understanding and processing and achieved the best accuracy among all sentence encoding methods at that time. The authors in another work,[18] which was related to a recurrent neural network (RNN), considered the attention problem as the sequential decision of a goal-directed agent interacting with a visual environment. Long short-term memory network[19] (LSTM), which is a special type of RNNs, can capture the long-term dependencies information of the sequential inputs. It is the attention capability that makes it popular in processing the dataset with spatial–temporal features like video sequences for action recognition. For example, the existing works[20,21] use RNNs and CNNs combined with LSTM attention module to enhance the network to focus selectively on informative parts of the video frames using the memory cell

and obtain promising results. Similarly, another work[22] develops the cross-link layers that embed the attention to guide the spatial-stream to pay more attention to the human foreground areas and be less affected by background clutter. Meanwhile, attention mechanisms are increasingly applied in image recognition.

But image recognition is essentially different from the above because image classification tasks aim to explore and capture the semantic information and pixels correlation in a single image instead of sequential inputs. Several works[23,24] presented attention-based models for recognizing multiple objects and image captioning, which were capable of learning to both adaptively localize and recognize the most relevant regions of the input images.

Inspired by these attention mechanisms, in our work, we introduce effective encoding layers as the attention module and attempt to use soft-attention mechanisms of deep convolutional neural networks (DCNNs) to guide the network to the most discriminative features learning and the most relevant regions localizing. We take both channel and spatial relationships into consideration and propose an end-to-end deep convolution neural attention network. It can enhance feature representations with large receptive fields and enlarge the feature scope for decision making.

To summarize, our main contributions of this work are threefold:

1. We propose a fused attention network to enhance the feature discriminative ability and achieve more stable and superior classification performance both on 5-day-old and 9-day-old egg embryos.
2. We validate the effectiveness of our module by integrating our attention into the existing CNNs.
3. Furthermore, we conduct extensive ablation experiments, and our results indicate that our method has a higher confidence coefficient for the final prediction compared with previous methods and reduces the error rate of weak embryo classification.

2 Related Works

2.1 Attention Mechanisms

Spatial attention can be interpreted as a pixelwise weighting operation and a learning mechanism that can help capture spatial correlations. The algorithm learns the most informative features and assigns more available computational resources to the focus area. DCNNs have their own function of attention mechanism; for example, in classification tasks, the pixels learned and activated in deep-level feature maps are concentrated on the discriminant region of an image naturally.

Several prior attempts[25–27] to strengthen the representation of CNNs by attention mechanisms have been made in classification tasks. Wang et al.[27] proposed the residual attention network that performed large-scale classification well and was also robust to noisy labels. The attention module cascades a bottom-up and a top-down structure to explore fine-grained feature maps. The bottom-up feedforward structure produces low-resolution feature maps with strong semantic information. Then, the top-down architecture, which aims to generate the weight mask, employs deconvolution[28] to recover the resolution.

In our work, by contrast, we disassemble the process and compute the channel and spatial attention, respectively, rather than directly learning the mixed 3D spatial attention map. First, we generate the channelwise attention map by utilizing the existing squeeze-and-excitation (SENet),[25] which has been proved to perform well. In particular, we tactfully adopt the depthwise separable convolution[13,29] to connect the channelwise attention and continue to learn the spatial attention. We find that the depthwise convolution operation effectively increases representation efficiency. Unlike the method adopted by Ref. 27, we also argue that the approach of adopting downsampling and then upsampling will result in loss of spatial information, and motivated by dense prediction issues like semantic segmentation, we exploit multiscale features by adopting multiple parallel dilated filters[30–32] with different sampling rates to maintain spatial resolution and produce the spatially dense attention mask. The dilated convolution provides an efficient mechanism for controlling the receptive field size and seeking the best balance between accurate location (small field-of-view) and context assimilation (large field-of-view).[32]
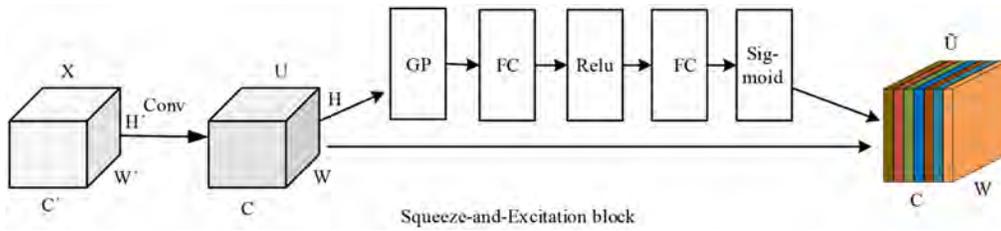
Fig. 1 Architecture of SE block.

2.2 *Squeeze-and-Excitation Networks*

The SENet, which we call the SE block, focus on the channel relationships with the goal of improving the quality of representations. The architecture comprises a lightweight gating mechanism and explicitly models channel interdependencies in a computationally efficient manner.[25] The network can make dynamic channelwise feature recalibration and boost the feature discriminability. The structure of the SE building is depicted in Fig. 1.

In Fig. 1, we can see that the structure of the SE block is simple and can be integrated directly with existing state-of-the-art architectures. First, the structure adopts a global pooling (called GP in Fig. 1) to shrink the 3D feature maps through spatial dimension to a 1D vector. Second, a fully connected (FC) layer is used to reduce parameters; the reduction ratio r is set to 16. Then, the output is sent to a RELU function to increase nonlinearity and another FC layer to restore the feature size to its original dimension. Finally, the values are normalized to [0, 1] via a sigmoid activation as a set of per-channel modulation weights, which is used to rescale the transformation output U. Formally, we assume that $U = [u_1, u_2, \ldots, u_n]$ with $u_i \in \mathbb{R}^{W \times H}$, $i = 1, 2, \ldots, n$ denoting a set of n channel features, so we can illustrate the outputs as $\hat{U} = [\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n]$ with $\hat{u}_i \in \mathbb{R}^{W \times H}$ $i = 1, 2, \ldots, n$, where

$$\hat{u}_i = \alpha_i \cdot u_i \qquad i = 1, 2, \ldots, n, \tag{1}$$

Here, α is the weight for the channelwise attention. The final output is the channel-refined feature maps shown in Fig. 1, where each color represents a specific channel.

3 Methods

Our attention module is a sub-branch splitting from the trunk. We employ residual units as our backbone to perform downsampling and feature processing, then the attention structures (DPSA) make deeper and more specific feature extraction for particular object categories (class discriminative features), as layers going deeper. The resulting fine-grained saliency maps with normalized weight superimposed on the output of the backbone help to make the informative features more highlighted and suppress noises at the same time. The deeper the layers, the more selectively the attention model will activate and focus on object-specific goals that are helpful for classification. Therefore, integration of our DPSA modules into the backbone at different stages gradually enhances the class-specific features representation (Fig. 2).
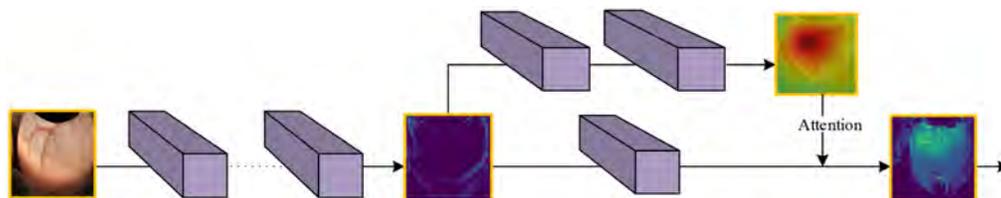

Fig. 2 Overview of our attention mechanism.

3.1 *Dense Pixelwise Spatial Attention Module*

The SE block applies unequal weight to each channel by dynamic learning and recalibrates the channelwise features adaptively. Although the method decides "which" channel to concentrate on, we also need to explore and focus on the most informative components in spatial locations. It is difficult for a classification task when key features are not spatially dominant. Therefore, we argue that it is equally important to fully extract spatial information from each channel further and explore spatial attention. We describe our fused attention module in detail below.

Our proposed DPSA architecture can be divided into three parts. First, to take better advantage of already generated channelwise attention maps and enhance spatial encodings, we adopt a depthwise convolution layer to split the channels that are already weighted. As shown in Fig. 3, we can see the features, which are reweighted along channels with different channelwise importance output through the SE block. Then, we aim to make spatial feature learning along channels; a single 3×3 filter with a stride of 1 is applied to each channel. However, it only filters input channels and ignores the semantic hierarchy between different channels. Therefore, a 1×1 (pointwise) convolution layer needs to be used to fuse the output and generate new features. Equations (3)–(5) ($\odot$ denotes the elementwise product) have illustrated the mathematical formulation of standard convolutions and depthwise separable convolutions, where K denotes the kernels with channels of N and x denotes the input features:

$$\text{Conv}(K, x)_{(i,j)} = \sum_{l,m,n}^{L,M,N} K_{(l,m,n)} \cdot x_{(i+l,j+m,n)}, \tag{2}$$

$$\text{Depthwise Conv}(K, x)_{(i,j)} = \sum_{l,m}^{L,M} K_{(l,m)} \odot x_{(i+l,j+m)}, \tag{3}$$

$$\text{Pointwise Conv}(K, x)_{(i,j)} = \sum_{n}^{N} K_n \cdot x_{(i,j,n)}. \tag{4}$$

The factorized convolution has indicated that our features have both fairly independent channels and highly correlated spatial locations. In our work, depthwise separable convolutions help us make full use of weighted channels and extract the features completely. Meanwhile, this kind of convolution increases computational efficiency by converting high-dimensional features into low-dimension ones. For example, the kernel size is $K_{\text{conv}} \times K_{\text{conv}} \times C_{\text{in}}$ and the size of the output is $N_{\text{out}} \times N_{\text{out}}$ with channels C_{out}. The computational cost of the depthwise separable convolution, which is the total of the depthwise and pointwise (1×1) convolutions, is as follows:

$$K_{\text{out}} \times K_{\text{out}} \times C_{\text{in}} \times N_{\text{out}} \times N_{\text{out}} + 1 \times 1 \times C_{\text{in}} \times N_{\text{out}} \times N_{\text{out}} \times C_{\text{out}}. \tag{5}$$
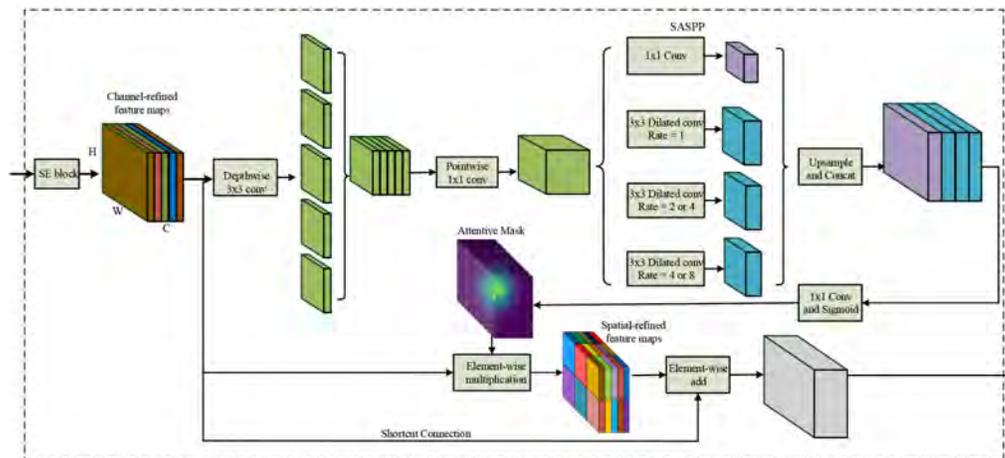

Fig. 3 The architecture of our proposed DPSA module.

Compared with the standard convolutions, the ratio in calculation consumption is as follows:

$$\frac{K_{\text{conv}} \times K_{\text{conv}} \times C_{\text{in}} \times N_{\text{out}} \times N_{\text{out}} + 1 \times 1 \times C_{\text{in}} \times N_{\text{out}} \times N_{\text{out}} \times C_{\text{out}}}{K_{\text{conv}} \times K_{\text{conv}} \times C_{\text{in}} \times N_{\text{out}} \times N_{\text{out}} \times C_{\text{out}}} = \frac{1}{C_{\text{out}}} + \frac{1}{K_{\text{conv}}^2}. \tag{6}$$

Second, we consider that the quality of attention mask with salient features is related to the diversity of features, so we use dilated convolutions with various rates for dense feature extraction, which is based on the fact that dilated convolutions support exponential expansion of receptive fields without loss of resolution or coverage.[30] For our egg embryos, the blood vessels are discriminant features for the network to make prediction, so the detailed information gained for small targets is especially significant, which requires denser resolution and multiscale information. Even though DCNNs have shown to be successful for classification tasks, the repeated combination of network pooling and striding at consecutive layers remarkably reduces the spatial resolution and loses local detail information of the resulting features maps. Deconvolutional layers have been employed to recover the spatial resolution[33,34] but it it difficult to restore the lost detail. Therefore, we advocate the use of "dilated convolution," which not only obtains resolution enhancement but also enlarges the receptive fields to incorporate larger semantic context. We have clarified the algorithm's operation in 1D with a simple example illustrated in Fig. 4 (modified from Ref. 32), and the mathematic formulation is as follows:

$$P[i] = \sum_{m=1}^{M} s[i + r \cdot m]f[m], \tag{7}$$

where $s[i]$ represents the input signals, $f[m]$ is the filter of length M, and r denotes the dilation rates we use to sample the input.

Our approach is inspired by the atrous spatial pyramid pooling in Refs. 32 and 35 for the task of semantic segmentation. In our work, we simplify the architecture further and call this revised method simplified atrous spatial pyramid pooling (SASPP), which is illustrated in Fig. 3. In general, classification networks are able to identify one or small discriminative parts with a high response naturally for correctly recognizing images. Meanwhile, we also argue that dilated
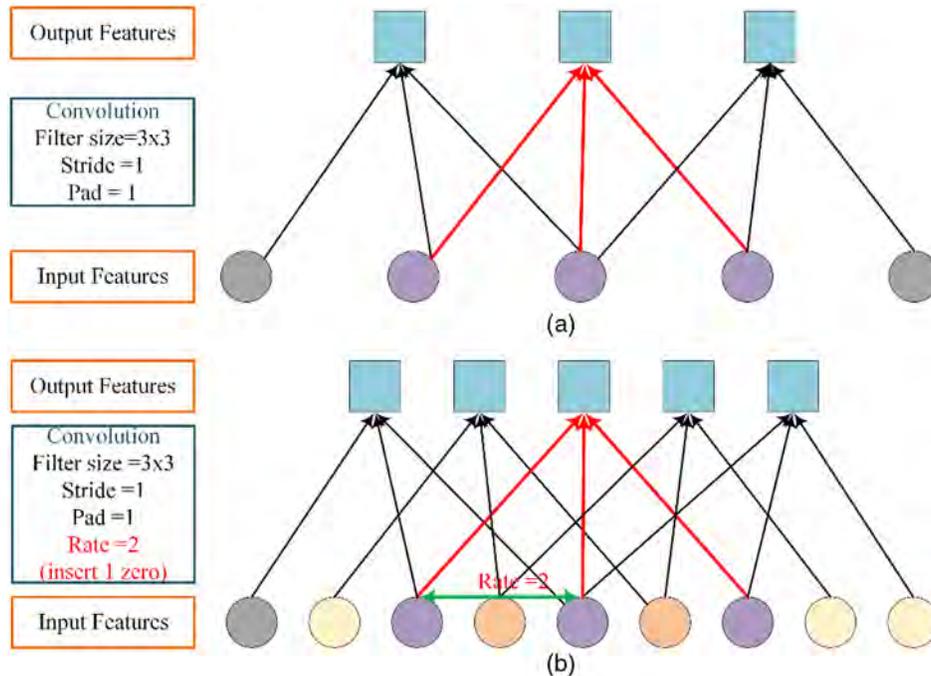

Fig. 4 Illustration of dilated convolution in 1D (modified from Ref. 29): (a) sparse feature extraction with standard convolution and (b) dense feature extraction with dilated convolution with a rate $r = 2$.

convolution with multiple rates can help to capture richer contextual detail and produce dense and reliable target object localization effectively. We adopt spatially small convolution kernels 3×3 to resample features with various dilated rates 2^{k-1}, $k = \{1, 2, \ldots, k\}$. In our attention module, the largest resolution is only 56×56, so the maximum value of k is set to 4. We should avoid the condition that the receptive fields are too large to preserve local detailed information. We also add a batch normalization (BN)[36] layer after each dilated convolution, which can accelerate deep network training by reducing internal covariate shifts. However, as the layers go deeper, if we use the same sampling rate r, the valid region over which the filter weights are applied becomes smaller because there are $r - 1$ padded zeros. So, we stack our DPSA module in four stages of our design and the sampling rate varies as the layer deepens. The complete setting of the parameters will be provided and discussed in Sec. 4.6.

Third, to produce the final attention maps, the features from the parallel dilated convolution branches are interpolated bilinearly to the original features' resolution. These resulting features are then concatenated and passed through another 1×1 convolution to reduce the channel dimension. A sigmoid activation function is applied to normalize the weights to the interval [0, 1] and generate the attentive weight mask. During the elementwise multiplication step, the weight values of the 3D spatial attention mask are broadcasted to each pixel for each channel of previous features. As illustrated in Fig. 3, each small square with a different color represents a pixel with a different weight in each channel. We can obtain the spatial-domain-refined features, which achieve global emphasis across spatial dimension. In addition, we also add a shortcut to connect the channel-refined features' output by SENet, as in the deep residual network algorithm.[11] We argue that the practice enables the information from low and high levels to fuse better while making the significant features more emphasized. In conclusion, the final output of the dense pixelwise attention module is

$$H_{i,c}(x) = [1 + S_{i,c}(x)] \odot C_{i,c}(x), \tag{8}$$

where $S_{i,c}(x)$ and $C_{i,c}(x)$ represent the dense spatial attention mask and the channel-refined features, respectively, i denotes the index of the pixel, which ranges over all spatial positions, and c refers to the index of the channels.

3.2 Network Design

We have adopted the widely used residual units to construct our basic architectures, according to the size and characteristics of the dataset. The input image size is 224×224, and our network begins with a 7×7 convolution layer, followed by four stages, which are made of bottleneck templates with different numbers. The template is composed of two 1×1 convolution layers and a 3×3 convolution layer. We add a BN layer and RELU activation after each convolution layer. The first stage contains two bottleneck blocks and maintains a 56×56 resolution. The feature size of the subsequent stages is halved, and the number of the blocks following is three, four, two, respectively. The last layer ends with a global average pooling layer and a two-way FC layer with a softmax activation function (see Table 1 below for the network design).

The overall design of our network and the related hyperparameters setting can be found above. Our architecture is built by inserting our attention modules into different resolution stages of the basic network, which is illustrated in Table 1. Now, we list related parameters of the SASPP (see Fig. 3 for illustration) submodule in the DPSA structure (Table 2).

4 Experiments

4.1 Data Preparing and Preprocessing

Data acquisition is the first step of deep learning. To obtain sufficient image data of egg embryos, we have set up an image acquisition system. The system is composed of a sterile dark box, an LED light source, industrial cameras, and automation equipment. In our work, according to the structural design and the embryo size characteristics, the HIKVISION (model MV-CE013-50GC) is a color camera that was selected to collect data with a resolution of 1.3 million. The lens is an MVL-HF0828M-6MP model with a focal length of 8 mm. There is a light source under

Table 1 The table depicts the layers of our network integrated with DPSA modules. Downsampling is conducted by Conv2_1, Conv3_1, and Conv4_1 layers with a stride of 2.

Layer name	Layer type	Related parameters	Output size
Conv1	Convolution	7×7, 64, stride 2	112×112
Pool	Max pooling	3×3, 64, stride 2	56×56
Conv1_x	Convolution	$\begin{bmatrix} 1 \times 1, 64, \text{stride}1 \\ 3 \times 3, 64, \text{stride}1 \\ 1 \times 1, 128, \text{stride}1 \end{bmatrix} \times 2$	56×56
DPSA	Attention	—	56×56
Conv2_x	Convolution	$\begin{bmatrix} 1 \times 1, 128, \text{stride}1 \text{ or } 2 \\ 3 \times 3, 128, \text{stride}1 \\ 1 \times 1, 256, \text{stride}1 \end{bmatrix} \times 3$	28×28
DPSA	Attention	—	28×28
Conv3_x	Convolution	$\begin{bmatrix} 1 \times 1, 256, \text{stride}1 \text{ or } 2 \\ 3 \times 3, 256, \text{stride}1 \\ 1 \times 1, 512, \text{stride}1 \end{bmatrix} \times 4$	14×14
DPSA	Attention	—	14×14
Conv4_x	Convolution	$\begin{bmatrix} 1 \times 1, 512, \text{stride}1 \text{ or } 2 \\ 3 \times 3, 512, \text{stride}1 \\ 1 \times 1, 1024, \text{stride}1 \end{bmatrix} \times 2$	7×7
DPSA	Attention	—	7×7
Pool	Average pooling	7×7, stride 1	1×1
FC	Inner product	2D	1×1

Table 2 The SASPP structure details in our DPSA module. The parameter r denotes the $n \times n$ dilated convolutional kernel with a sampling rate of r, which specifies the number of zeros (or holes) between pixels and then the kernel size is $[n + (n - 1)(r - 1)]$. The $r = (1, 4, 8)$ means employing the rates $= 1$, 4, and 8 for the three parallel branches.

SASPP structure	Parameters
Stage I	1×1 conv; 3×3 dilated conv, $r = (1, 4, 8)$, stride 1
Stage II	1×1 conv; 3×3 dilated conv, $r = (1, 2, 4)$, stride 1
Stage III	1×1 conv; 3×3 dilated conv, $r = (1, 2, 4)$, stride 1
Stage IV	1×1 conv; 3×3 dilated conv, $r = (1, 2)$, stride 1

each egg and a rubber above. When an egg is photographed, the light source of the other eggs is off. A plate of 72 eggs is run through the conveyor belt and is sent to the dark box, which triggers industrial cameras. An embryo can be captured from both sides to increase the number and feature diversity of samples. The image size generated by the system is 1280×960, which contains all of the regions of a single egg. In the actual production process, the egg embryo formation activity detection is divided into several stages. Therefore, our dataset has different kinds. The 5-day embryo images, which can be categorized as either fertile or infertile as early as the embryonic stage, have the least obvious characteristics. This is why it is difficult to determine viability in industrial production. Nine-day embryos, which are the first batch of embryos to be inoculated, can be classified as living and dead. Our dataset is shown below.

Figures 5 and 6 are the original samples. It can be seen that the embryo images obtained directly from the data acquisition system have information about adjacent eggs, so the data need to be simply processed. To train our model better, first, we set the threshold value and carry out binarization. Then, we find the edge information of the rubber from top to bottom. In addition, we set the lowest point of the rubber border as the center and set a constant width. We cropped
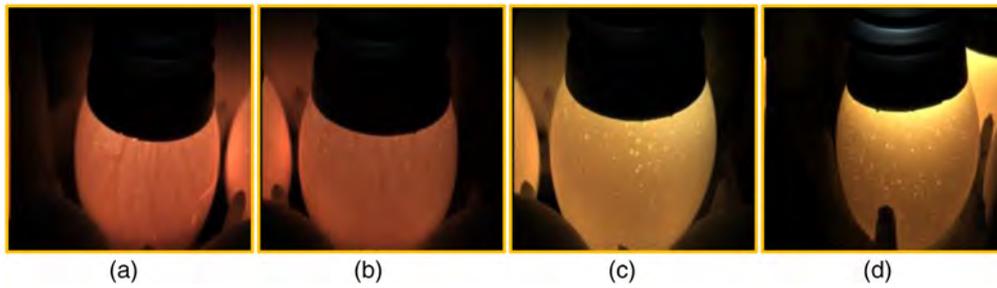
Fig. 5 The dataset consisting of 5-day-old egg embryos. (a, b) The samples shown are the fertile egg embryos, while (c, d) the samples shown are the infertile.
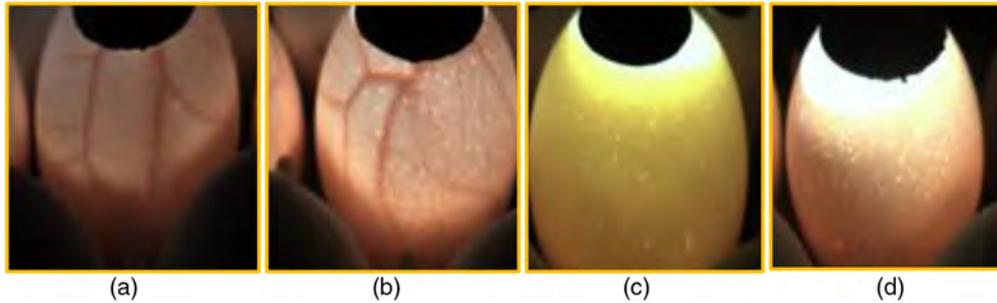

Fig. 6 The dataset consists of 9-day-egg embryos. (a, b) The samples shown are living samples and (c, d) the samples shown are the dead.
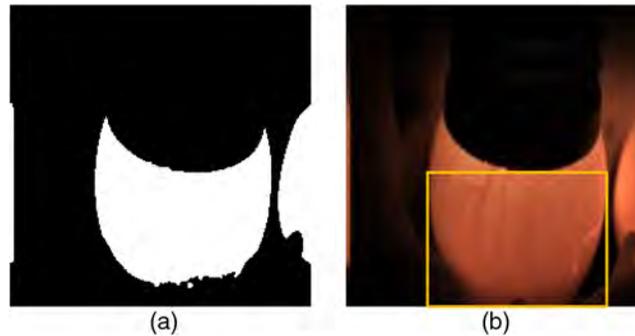

Fig. 7 Data processing: (a) binary image and (b) cropped feature regions.

the feature regions, which are labeled by yellow lines, and then resized images to 227×227 to remove the adjacent interference as much as possible (Fig. 7).

4.2 Implementation Details

Our implementation is based on the Caffe[37] framework. In the input layer, we follow standard practices and perform data augmentation by randomly cropping an image to down to the size of 224×224 pixels or conducting horizontal flip and random mirror. Our input training and testing batch sizes are set to 64 and 16, respectively. Each input image is normalized via mean RGB-channel subtraction by the training dataset mean file. Optimization is performed using stochastic gradient descent with a momentum of 0.9 and a weight decay of 0.0005. Our base learning rate is set to 0.001, and the update strategy follows the multistep policy with the gamma ("γ") of 0.1:

$$lr = \text{base\_}lr * \gamma^{\left|\frac{\text{iters}}{\text{step value}}\right|}. \tag{9}$$

We set the step value parameters to 20,000, 35,000, 50,000, and 60,000. When the iteration reaches one of these values, the learning rate is decreased according to the equation lr. As

reported in a work[38] and observed in our experiments, we find that the method of "MSRA"[38] filter weights initialization better accommodates RELU activation than "Gaussian"[39] or "Xavier"[40] in our network. We also have made analysis theoretically and assumed that the response of a convolution layer can be expressed as follows:

$$y_l = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b = W_l X_l + b, \tag{10}$$

where w is the weight of filters and x denotes the inputs. Let us assume random variables w and x are independent and each of their elements shares the same distribution. In particular, w has zero mean. Then, we can obtain variance:

$$\mathrm{Var}[y_l] = n\mathrm{Var}[w_n x_n] \quad = n\mathrm{Var}[w_n]E[x_n^2]. \tag{11}$$

We use l to denote the index of a layer. For the RELU activation, we have $X_l = f(Y_{l-1})$; then we can calculate $E[x_l^2] = \frac{1}{2}\mathrm{Var}[y_{l-1}]$, and putting this into Eq. (12), then we have

$$\mathrm{Var}[y_l] = \frac{1}{2} n\mathrm{Var}[w_l]\mathrm{Var}[y_{l-1}]. \tag{12}$$

According to the above equation, to keep the variance of data at each layer consistent, the weight should meet the following:

$$\frac{1}{2} n\mathrm{Var}[w_l] = 1, \quad \forall\, l. \tag{13}$$

In the final, we get a zero-mean Gaussian distribution whose standard deviation (std) is $\sqrt{\frac{2}{n}}$. This is our way of "MSRA" initialization $w \sim G[0, \sqrt{\frac{2}{n}}]$.

4.3 Image Classification on Hatching Eggs

In this section, to evaluate our proposed DPSA block, we first perform an ablation experiment on our 5-day-old and 9-day-old egg embryo datasets, respectively. Our 5-day embryo dataset comprises 8200 training images and 2725 validation images. We also obtain a final result from the 2680 testing images. Meanwhile, we further perform experiments on the 9-day old egg dataset, which comprises 20,000 images. We randomly select 12,000 samples for the training set and 4000 for validation; the final accuracy is gained on the 4000 testing images. All of the datasets are from two classes. We train the dataset on the original basic network, the basic network integrated with SE blocks (basic network + SE), and the basic network integrated with DPSA modules (basic network + DPSA). Each experiment is trained for 100,000 iterations from scratch.

Figure 8 has depicted the training curves on the 5-day embryos; the green line is the result of our fused attention model and the red one is the result of only a single channel attention.
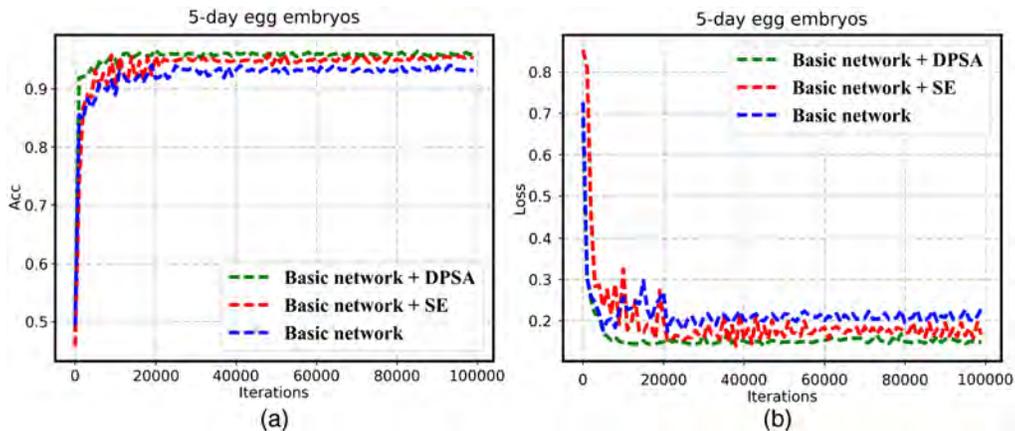

Fig. 8 (a) The accuracy and (b) loss curves during 5-day-old embryos training.
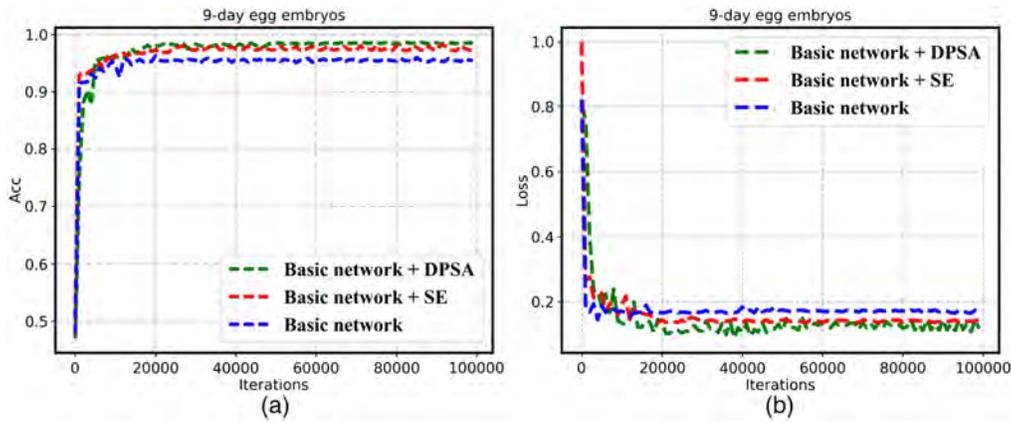
Fig. 9 (a) The accuracy and (b) loss curves during 9-day-old embryos training.

We can observe that our proposed method has achieved the highest accuracy. Meanwhile, we evaluate our method on 9-day-old embryos; the performance in Fig. 9 has verified the effectiveness of our attention mechanism as well. We also can conclude that our dense pixelwise spatial attention (DPSA) combined with SENet can push the performance of SE blocks. We argue that the accuracy gains are due to the self-recalibration on features, which guides the network to localize the most class-specific and relevant targets better.

4.4 Integration with Modern Architectures

We find that the light-weight and efficient network MobileNet[13] is a streamlined architecture, which is based on depthwise separable convolutions instead of standard convolutions. This
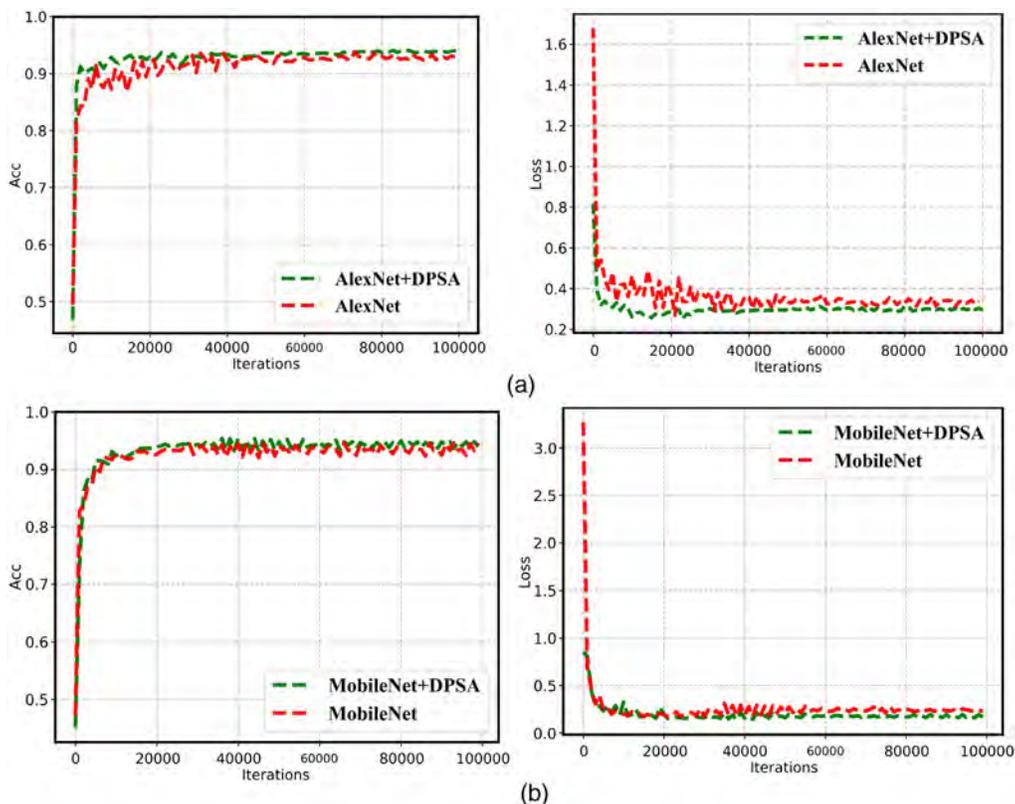

Fig. 10 Training curve comparisons between different baseline architectures with their DPSA module counterparts. (a) AlexNet and AlexNet + DPSA and (b) MobileNet and MobileNet + DPSA.

motivates us to integrate our fused attention modules with the MobileNet and further evaluate the effectiveness of our attention mechanism. In our experiments, we insert our DPSA module into the model once at each different resolution stage. In addition, we also attempt to integrate our fused attention modules with the classic and shallow AlexNet[39] for performance exploration (Fig. 10).

Through a series of ablation experiments, we have validated the effectiveness of our dense pixelwise attention module. However, it cannot be denied that the improvement of performance is at the cost of increasing training time. To explore the impact of our attention module to the practical runtime, we make further study to compare the added time when integrating it into existing architectures. We print and display results every 1000 iterations during training for a total of 100,000 iterations. According to the saved logs, the results are reported in Fig. 11 (Table 3).

We can observe that for different baseline networks, the increase in time is different. For lightweight MobileNet, our attention modules brought a little extra time. Meanwhile, even though the increased training time, which is about 1.8 h (6700 s), induced by our attention modules in the AlexNet is obvious, it can also be accepted. In conclusion, the relationship between performance improvement and runtime increase is reasonable. The results are consistent with the fact that the SE block and depthwise separable convolution are low computational overhead operations.
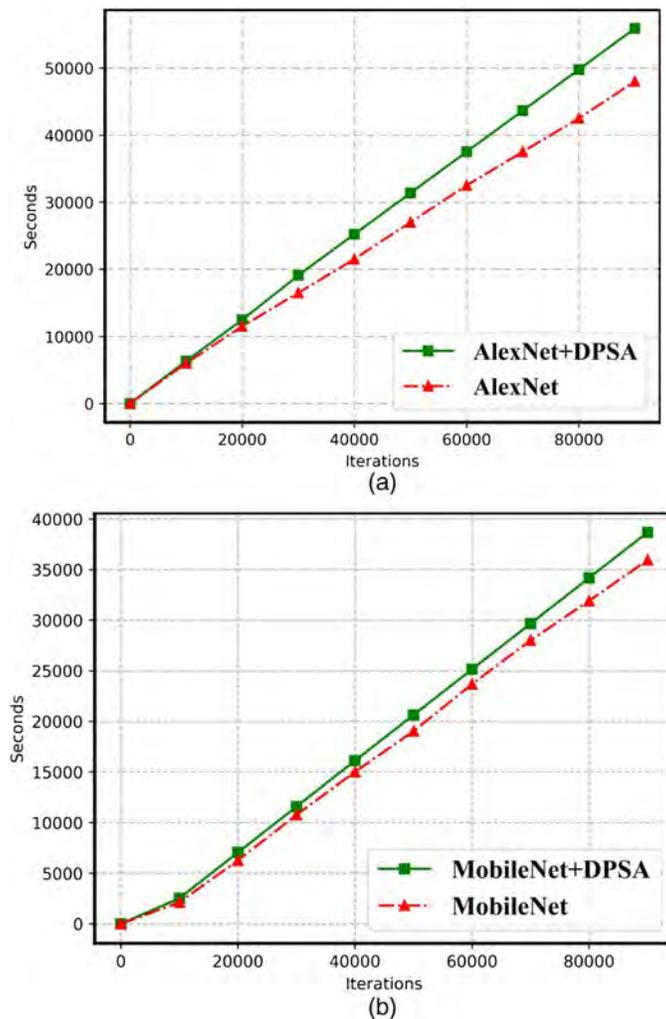

Fig. 11 Training time of (a) baseline architectures and (b) their DPSA counterparts.

Table 3 The performance comparison of several architectures equipped with DPSA modules.

Networks	Top-1 accuracy (%)	Time/h (100k iteration)
AlexNet	93.2	13.5
AlexNet (fused attention)	94.6	15.3
MobileNet	95.1	10.0
MobileNet (fused attention)	95.8	10.6

4.5 Comparisons with State-of-the-Art Models

Currently, there are many existing networks achieving good performance in classification tasks. In this section, we conduct comparative experiments with several state-of-the-art models on our dataset. In our work, all of the models are trained from scratch. The setting of training strategies and other hyperparameters (like batch size, initial learning rate) follows the same principle. But due to different conditions, we could not reproduce the results in the same way that the original papers demonstrate. We report the final results on the test set in Table 4.

Compared with the state-of-the-art methods and our previously proposed structure SJ-CNN (SE module and joint supervision based on a convolution neural network), our attention network achieves good and stable performance both on the 5-day and 9-day old embryos. Despite the success of DenseNet and residual attention network in the original papers, the deep networks are difficult to train well on our small-scale and simple datasets and are prone to overfitting. SJ-CNN (in 2018), the model trained specially for 9-day old eggs detection, has a poor generalization on 5-day eggs. Our DPSA-integrated structure has an increased accuracy of 5.1% and 0.7% compared with SJ-CNN. Although the improvement of our network performance is not very huge, it still makes sense in the classification of embryos because we require as high an accuracy rate as possible to prevent the weak and dead embryos from being misjudged and contaminating the living.

For further qualitative analysis, we adopt the gradient-weighted class activation mapping (Grad-CAM)[41] to make "visual explanations" for decisions from CNN-based models. The method can use the gradient information of the final convolutional layer to produce a coarse localization map highlighting the important regions and spatial locations in the image for predicting the concept, which is able to evaluate the effectiveness of our proposed DPSA. In our work, we randomly select six egg embryo images from two classes (the weak embryos belong to the dead) with different characteristics to evaluate our model. The results are shown below.

From the results shown above, we can see the area that is enhanced and emphasized by our attention mechanism. The red zone is the field where the network learns discriminative features

Table 4 Comparison with other state-of-the-art methods on our 5-day-old and 9-day-old egg embryo datasets.

Method	5-day/9-day	Accuracy rate (%)
DenseNet[10]	5-day	94.3
	9-day	96.2
SJ-CNN[15]	5-day	93.2
	9-day	98.4
Residual attention network[27]	5-day	96.8
	9-day	98.7
Proposed method	5-day	98.3
	9-day	99.1

and applies additional weight on the selective parts. Thus, Grad-CAM tells us, in the form of a heat map, which pixels the model focuses on to determine whether the image is of a living or dead embryo. We can clearly see that the red zone in the last column, which is learned by our proposed architecture, is larger than in others. More importantly, we observe in the second row that the blood vessels are not exactly in the center of the image. Our structure has learned all of the fields according to blood vessels rather than only a corner, which is learned by other networks. The third row shows that our previous proposed SJ-CNN performs poorly when the detected egg is interfered with by surrounding eggs. Though each of the four models has made the correct decision for the images of the first four rows, the confidence score c (where c denotes the softmax score of each network for the truth label) of our proposed architecture is higher than its counterparts, as illustrated in Fig. 12.
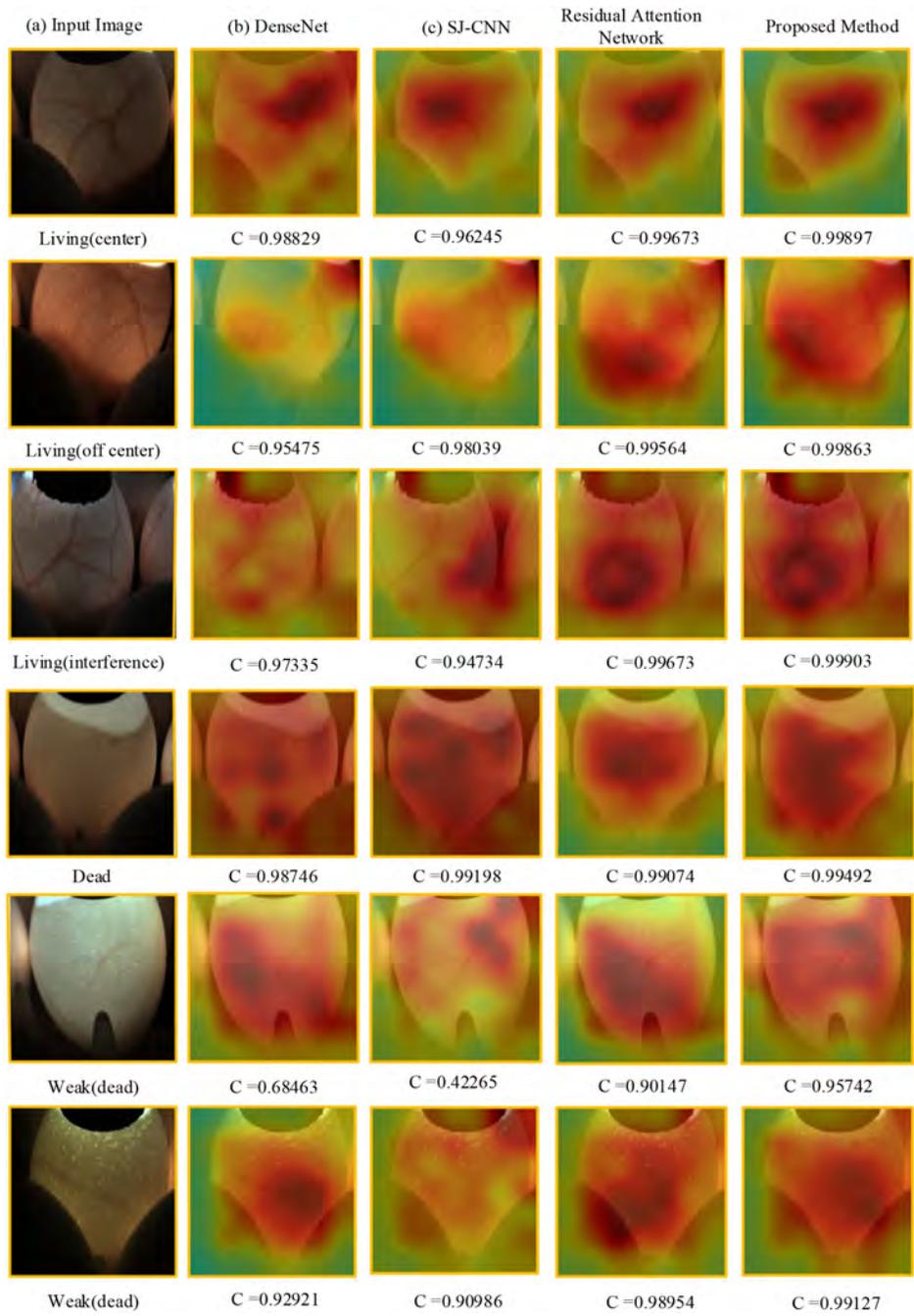

Fig. 12 Grad-CAM visualization results on 9-day-old embryos.

Likewise, as we mentioned before, the remaining trouble we face is that weak embryos sometimes are judged to be living ones. Our model has alleviated the problem. We can see from the last two rows in Fig. 12 that weak embryos have several locally thin vessels. Unfortunately, when the model only locates and learns this small region, the final prediction tends to be that it is a living embryo (in fact, it belongs to the class of dead while the probability of being judged dead is only 0.42265). Conversely, when the region of class-discriminative feature localization is larger, the final decision perhaps is different. To put it in a simpler way, it is hard to make a prediction when we put our eyes close to the back of animals to distinguish between a donkey and a horse. Our DPSA integrated residual network covers a much larger area of the target object regions and is more precise; therefore, it reduces the error rates for the weak embryos. The proposed DPSA maps, which aggregate multiscale contextual information, help boost the ability to explore the highly class-discriminative features (the blood vessels in living embryo) with larger field-of-view and accurately and densely localize object regions. In conclusion, our figures above have evaluated the model credibility and provided reasonable explanations for why the network embedded with our DPSA modules has a better performance in the experiments.

To validate that our network mitigates the problem of weak embryo misjudgment, we have selected 4000 images for the test set, which comprises 2000 living embryos and 2000 dead embryos. In particular, among the 2000 dead samples, there are 123 weak embryo images. We have evaluated our attention network model and previous SJ-CNN on the test set. Figure 13 shows the confusion matrix for the classification results of the two models. It is obvious that our attention network has achieved better performance in recognition accuracy, and the number of weak embryos being misjudged in dead embryo samples is reduced by half.

4.6 Effect of Different Multirates for SASPP

In this section, we conduct experiments to analyze the effects of different dilation rate groups for the four network transformation stages. SASPP with various sampling rates helps us capture multiscale information and expand spatial density. Yet, in Ref. 35, researchers noticed that as sampling rates increase, in other words, the zeros inserted in the pixels of the feature map are more, the amount of valid filter weights (the weights applied to the informative feature
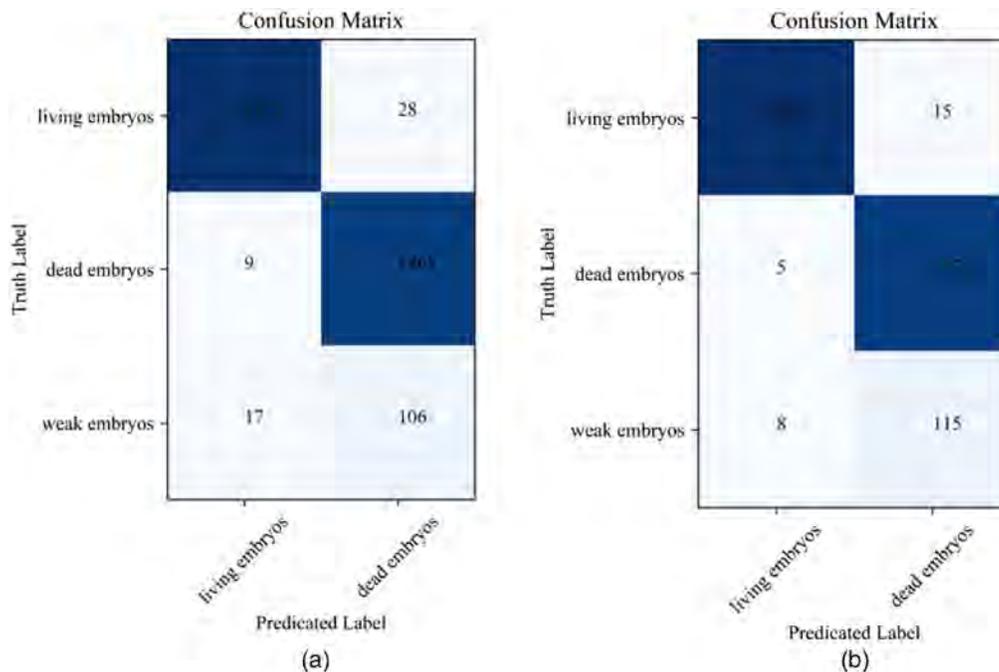

Fig. 13 Confusion matrix for the experimental results: (a) SJ-CNN and (b) the proposed method.

Table 5 Different strategies on employing multiple rates for parallel branches of SASPP.

Method	Stage I	Stage II	Stage III	Stage IV
Strategy 1	(2,4)	(2,4)	(2,4)	(1,2)
Strategy 2	(1,2,4)	(1,2,4)	(1,2,4)	(1,2)
Strategy 3	(1,4,8)	(1,2,4)	(1,2,4)	(1,2)
Strategy 4	(1,4,8)	(1,4,8)	(1,2,4)	(1,2)
Strategy 5	(1,4,8)	(1,4,8)	(1,4,8)	(1,2)

regions, rather than the filled zeros) decreases. In our work, our kernel size is 3×3 with rates associated with the values in set $\{1, 2, 4, 8\}$. Therefore, we attempt to employ dilation rates depending on the resolution of the feature map in our network, e.g., the lower layers have larger sampling rates than the upper layers. We conduct five group experiments by testing a combination of different factors and measure the effects on 9-day embryo classification accuracy. We list the five strategies in Table 5.

As illustrated in Table 5, the SASPP structure at stage 4 in our network has only two parallel dilated convolution layers, which employs smaller rates $r = \{1, 2\}$ because the features' resolution is only 7×7. We attempt to set the maximum rate to be 8 because the largest resolution is 56×56 during the four stages transformation (introduced in Table 1). When the rate $= 2$, the 3×3 filter is enlarged to 5×5 and it equals the standard convolution when the sampling rate $= 1$. In our experiment, the maximum number of parallel dilated convolution layers in SASPP architecture is set to be 3; we have not experimented with more branches (Fig. 14).

From the histogram above, we can obviously find that the performance of three parallel branches is better than that of two. We adopt the same sampling rates $= \{1, 2, 4\}$ at the first three stages and yield 1.2% better than employing two parallel branches. Furthermore, we also attempt to use larger rates $= \{1, 4, 8\}$ at different stages, and the results show that the first stage employing the rates $= \{1, 4, 8\}$ outperforms that employing rates $= \{1, 2, 4\}$. The performance improves from 98.4% to 99.1%. Therefore, we have experimented with the next two strategies by employing rates $= \{1, 4, 8\}$ at the second or the third stage. However, when we train the model employing larger rates at more stages simultaneously, the performance has a consistent drop. We observe that strategy 3 yields the best performance; thus, strategy 3 is selected as our final choice.
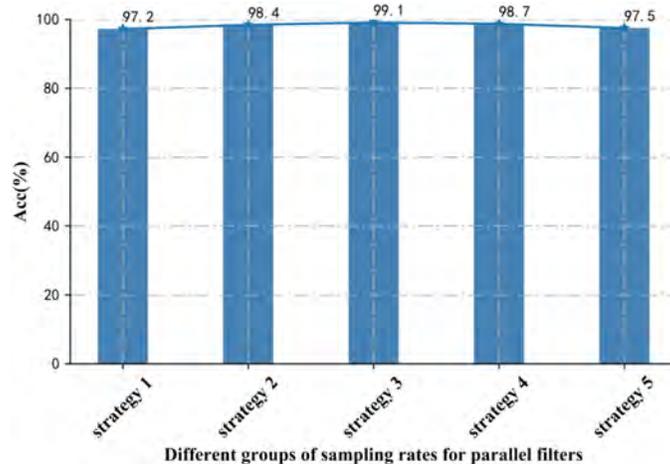
Fig. 14 Results of different strategies on 9-day-old embryo classification.

5 Conclusions

In this paper, instead of a single attention mechanism design, we introduce an attention-based feature refinement along both dimensions: channel attention and spatial attention. To generate the attention mask, our key idea is to employ parallel dilated convolutions with different sampling rates to achieve denser feature extraction and field-of-view enlargement. Similar to channelwise attention, which is based on an adaptive recalibration of features between channels, our spatial attention is pixelwise weighting and the refinement process. The weight mask with strong semantic information can help emphasize useful features and dismiss unimportant ones. Our network achieves a steady and superior accuracy, which is up to 98.3% and 99.1% on 5-day and 9-day embryos, respectively, through the attention optimization procedure, providing evidence that the feature refinement process of our attention modules is effective. Nevertheless, there are still some limitations in our attention network, and we still need to do a lot. In the future, we will research compatibility with other models, and as the samples of weak embryos increase, we will attempt to classify weak embryos into a single class. In addition, the multimodel fusion method that combines a sequence of embryonic heartbeat signals with images will be considered to optimize our detection task.

Acknowledgments

References

1. A. L. Romanoff and K. Frank, "High frequency conductivity and dielectric effect of fresh fertile and infertile hens' eggs," *Exp. Biol. Med.* **47**(2), 527–530 (1941).
2. T. C. Mcquinn et al., "High-frequency ultrasonographic imaging of avian cardiovascular development," *Dev. Dyn.* **236**(12), 3503–3513 (2007).
3. D. P. Smith, J. M. Mauldin, and K. C. Lawrence, "Detection of fertility and early development of hatching eggs with hyperspectral imaging," in *Proc. 11th Eur. Symp. Quality Eggs and Egg Products*, Doorwerth, The Netherlands, pp. 176–180 (2005).
4. S. T. Jones, R. E. Shattuck, and A. I. Center, "Detection of early embryonic development in hatching eggs: a hyperspectral imaging systems and neural network approach," Johns Hopkins APL, Technical Digest 1, pp. 67–73 (2005).
5. L. Liu and M. O. Ngadi, "Detecting fertility and early embryo development of chicken eggs using near-infrared hyperspectral imaging," *Food Bioprocess Technol.* **6**(9), 2503–2513 (2013).
6. W. Zhang et al., "Early fertility detection of hatching duck egg based on fusion between computer vision and impact excitation," *Trans. Chin. Soc. Agric. Mach.* **43**(2), 140–145 (2012).
7. Y. W. Xu et al., "Automatic sorting system of egg embryo in biological vaccines production based on multi-information fusion," *Trans. Chin. Soc. Agric. Mach.* **46**(2), 20–26 (2015).
8. Q. L. Xu and F. Y. Cui, "Non-destructive detection on the fertility of injected SPF eggs in vaccine manufacture," in *Proc. 26th Chin. Control and Decis. Conf.*, Changsha, China, pp. 1574–1579 (2014).
9. B. Shan, "Fertility detection of middle-stage hatching egg in vaccine production using machine vision," in *Proc. Second Int. Workshop Educ. Technol. and Comput. Sci.*, Wuhan, China, pp. 95–98 (2010).
10. G. Huang, Z. Liu, and L. Van Der Maaten, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 4700–4708 (2017).
11. K. He, X. Zhang, and S. Ren, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778 (2016).
12. S. Xie, R. Girshick, and P. Dollár, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1492–1500 (2017).

13. A. G. Howard, M. Zhu, and B. Chen, "MobileNets: efficient convolutional neural networks for mobile vision applications," arXiv:1704.04861 (2017).

14. L. Geng et al., "Hatching eggs classification based on deep learning," *Multimedia Tools Appl.* **77**(17), 22071–22082 (2018).

15. L. Geng et al., "Hatching egg classification based on CNN with channel weighting and joint supervision," *Multimedia Tools Appl.* 1–16 (2018).

16. L. Geng, H. Wang, and Z. Xiao, "Fully convolutional network with gated recurrent unit for hatching egg activity classification," *IEEE Access* **7**, 92378–92387 (2019).

17. T. Shen, T. Zhou, and G. Long, "DiSAN: directional self-attention network for RNN/CNN-free language understanding," in *Proc. Thirty-Second AAAI Conf. Artif. Intell.* (2018).

18. V. Mnih et al., "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 2204–2212 (2014).

19. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.* **9**(8), 1735–1780 (1997).

20. S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," in *Neural Inf. Process. Syst. Time Series Workshop* (2015).

21. J. Liu et al., "Skeleton-based human action recognition with global context-aware attention LSTM networks," *IEEE Trans. Image Process.* **27**(4), 1586–1599 (2018).

22. A. Tran and L. F. Cheong, "Two-stream flow-guided convolutional attention networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 3110–3119 (2017).

23. J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," arXiv:1412.7755 (2014).

24. L. Chen et al., "SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 5659–5667 (2017).

25. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 7132–7141 (2018).

26. B. Zhao, X. Wu, and J. Feng, "Diversified visual attention networks for fine-grained object classification," *IEEE Trans. Multimedia* **19**(6), 1245–1256 (2017).

27. F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3156–3164 (2017).

28. M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. Int. Conf. Comput. Vision*, Barcelona, Spain, pp. 6–13 (2012).

29. L. Sifre and S. Mallat, "Rigid-motion scattering for image classification," PhD Thesis, Vol. 1 (2014).

30. F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv:1511.07122 (2015).

31. M. Holschneider, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets, Inverse Problems and Theoretical Imaging*, J. M. Combes, A. Grossmann, and P. Tchamitchian, Eds., pp. 286–297, Springer, Berlin, Heidelberg (1990).

32. L. C. Chen et al., "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(4), 834–848 (2018).

33. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3431–3440 (2015).

34. H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 1520–1528 (2015).

35. L. C. Chen, G. Papandreou, and F. Schroff, "Rethinking atrous convolution for semantic image segmentation," arXiv:1706.05587 (2017).

36. S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Int. Conf. Mach. Learn.*, pp. 1–11 (2015).

37. Y. Q. Jia et al., "Caffe: convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, Orlando, Florida, pp. 675–678 (2014).

38. K. M. He et al., "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *IEEE Int. Conf. Comput. Vision* (2015).

39. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 1097–1105 (2012).
40. X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. and Stat.*, pp. 249–256 (2010).
41. R. R. Selvaraju et al., "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 618–626 (2017).

Lei Geng is an associate professor at the School of Electronics and Information Engineering, Tianjin Polytechnic University. He received his PhD from the School of Precision Instrument and Opto-Electronics Engineering, Tianjin University, in 2012. His research interests cover image processing and pattern recognition, intelligent signal processing technology and systems, and DSP system research and development.

Zhitao Xiao is a professor at the School of Electronics and Information Engineering, Tianjin Polytechnic University. He received his PhD from the School of Electronics and Information Engineering, Tianjin University, in 2003. His research interest covers intelligent signal processing, image processing and pattern recognition.

Biographies of the other authors are not available.