Complex Scene Understanding and Recognition based on the Explainable Machine Learning in AI models

Bin Wu[a], Yuhong Fan[b]*, Yeh-Cheng Chen[c] , Tao Zhang[d]*
[a] *School of Internet of Things Engineering, Jiangnan University, Wuxi, 214122, China*
zsn827@163.com
[b] *Department of Computer Engineering, LangFang YanJing Vocational Technical College,Sanhe,065200,China*
fanyuhong8406@163.com
[c] *Department of computer science, University of California, Davis, CA, USA*
ycch@ucdavis.edu
[d] *Jiangsu Key Laboratory of Image and Video Understanding for Social Safety, Nanjing University of Science and Technology, Nanjing, 210094, P.R. China.*
taozhang@jiangnan.edu.au
**Corresponding author: ()*

Abstract—**Information fusion is an important part of numerous neural network systems and other machine learning models. However, there exist some problems about fusion in scene understanding and recognition of complex environment, such as difficulty in feature extraction, small sample size and interpretability of the model. Deep reinforcement learning can combine the perception ability of deep learning with the decision-making ability of reinforcement learning to learn control strategies directly from high-dimensional original data. However, It faces these challenges, such as low optimization efficiency, poor generality of network model, small labeled samples, explainable decisions for users without a strong background on Artificial Intelligence (AI). Therefore, at the level of application and theoretical research, this paper aims to solve the above problems,the main contributions include: (1)optimize the feature representation methods based on spatial-temporal feature of the behavior characteristics in the scene, deep metric learning between adjacent layers and cross-layer learning theory, and then propose a lightweight reinforcement learning network model to solve these problems of the complexity of the model to be explained, the difficulty of extracting feature and the difficulty of tuning parameter; (2)construct the self-paced learning strategy of the deep reinforcement learning model, introduce transfer learning mechanism in the optimization process, and solve the problem of low optimization efficiency and small labeled samples; (3)design the behavior recognition framework of the multi-perspective deep knowledge transfer learning model, construct a explainable behavior descriptor, and solve the problems of poor network generality and weak 1explanation of network. Our research is of great theoretical and practical significance in the fields of artificial intelligence and public security.**

***Keywords*: Explainable Machine Learning, Scene Understanding and Recognition, Reinforcement Learning, Self-paced Learning, Artificial Intelligence.**

1. INTRODUCTION

In recent years, skynet and other widely deployed video surveillance systems have provided an efficient means of data acquisition for scene analysis and recognition, and are also an important part of smart city and urban security monitoring. The understanding and recognition of complex scenes is one of the important technologies to realize artificial intelligence [1-2]. This paper mainly makes an in-depth study of crowd density estimation in surveillance

scenes, face occlusion recognition in front of ATM and the recognition of group abnormal behavior in video scenes. Its purpose is to determine what targets and behaviors exist in a scene, so it can be widely used in banks, stations, office buildings and other public places. On the other hand, in the last few years, the interest in deriving complex AI models capable of achieving unprecedented levels of performance has been progressively displaced by a growing concern with alternative design factors, aimed at making such models more usable in practice.

Considering that these behaviors seriously threaten the safety of society and individuals, it will undoubtedly provide more protection for social security if these data can be automatically analyzed and processed and an early warning is issued. Given the multimodal nature of the data and the complexity of the scene, we expect to use popular machine learning algorithms to solve this complex scene understanding and recognition problem. Compared with other scene analysis [3-5], the scene analysis and recognition of this project is more difficult, and the difficulty in extracting target and behavior features, small sample size and interpretability of the model are the bottlenecks and keys to solve this problem.

Daniel Wolpert believed that the reason for the evolution of the brain is not for thinking and feeling, but for controlling motion, which is the core idea of the Deep Reinforcement Learning. The problem model studied by it includes an environment and an agent interacting with the environment [6-9]. The goal of reinforcement learning is to design a behavioral strategy for the agent that maximizes its benefits in interacting with the environment. Google DeepMind used this strategy in 2016 to get computers to go beyond the level of top professionals. However, the development of object detection and behavior recognition in complex environment is slow, that is mainly because :(1) the efficiency of model optimization algorithm is not high; (2) Lack of interpretability of constructed model; (3) Small sample with label; (4) Complex and unrestricted scenes; (5) High-dimensional video data leads to difficulty in parameter adjustment. Curriculum Learning and self-paced Learning represent the recently proposed Learning strategy [10-11]. Their core idea is to simulate the cognitive mechanism of human beings, they first learn simple and general knowledge structure, then gradually increase the difficulty degree and transition to more complex and professional knowledge. These two methods have similar conceptual learning paradigms, but differ in their specific learning schemes.

In curriculum learning, the curriculum is predetermined by prior knowledge and remains fixed after that, this kind of approach relies heavily on the quality of prior knowledge and ignores feedback about the learner. In self-paced learning, the course is dynamically determined to adapt to the learners' Learning rhythm. However, the self-paced Learning was unable to process previous knowledge, making it prone to over-fitting. For many computer application problems, it is necessary to build very accurate and easy proofs to understand machine learning models. Especially in the field of scene analysis and recognition, there is a growing demand for artificial intelligence (AI) methods that not only perform well, but are reliable, transparent and interpretable. This would allow non-professional persons to have possibilities to understand how and why a artificial intelligent algorithm makes that kind of decision, which will increase reliability of machine learning models in AI systems.

The main problem for explainability is to show enough justification for a ML model so that non-professionals know why a conclusion was drawn, and tell the non-professionals know when a model will perform well or not, and Under

what conditions this model can be trusted. Mainstream machine learning algorithms, especially deep learning models, are now ubiquitous and they has been widely used in many fields: visual recognition, bioinformatics, scene analysis, etc. The good performance of these models is largely due to their strong approximation and estimation properties, and a large number of training samples. An important obstacle, however, is interpretability: these models are often seen as black boxes, allowing little insight into how to make predictions. There does seem to be a trade-off between performance and interpretability, because the best-performing methods are often the least interpretable, while the most interpretable methods has long been regarded as not the best. Figure 1 qualitatively illustrates the tradeoff between the performance and interpretability of a machine learning algorithm. When a researcher wants to ensure the accuracy of the model, model control is often considered at this time, which may come from security, regulatory, or fairness considerations. In this case, the model is interpreted against these properties to check that they meet the requirements and to ensure that the model can be safely designed.

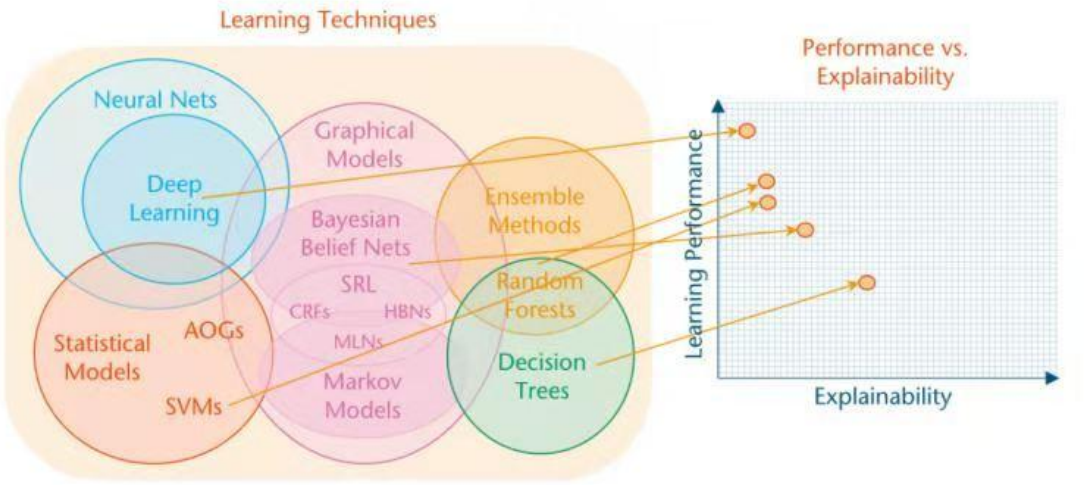


Fig. 1. The tradeoff between learning performance and interpretability for different algorithms.

In recent years, with the rapid development of deep learning, great progress has been made in solving various scene analysis tasks. However, the deep neural network used in the deep learning model is often used as a "black box", that is, the model only gives the final classification results, without making an understandable explanation for the classification results and decisions of the model [12].

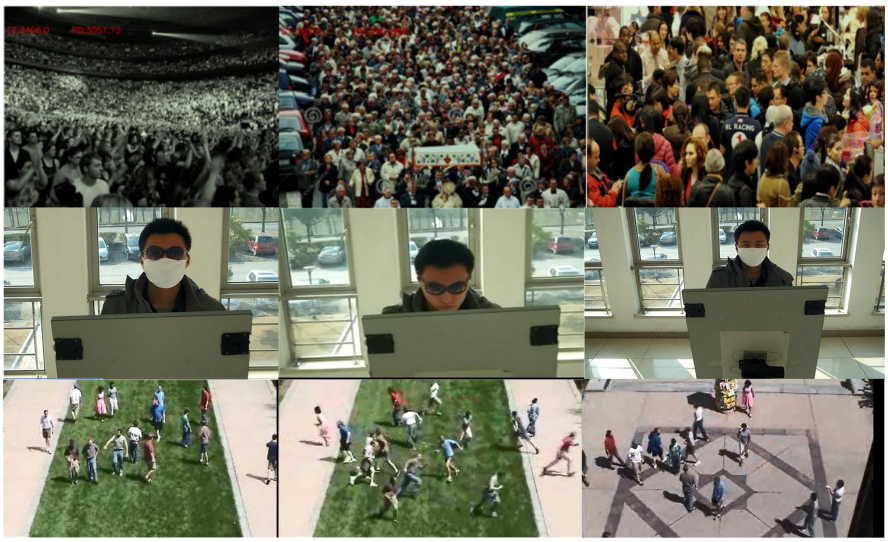


Fig. 2. Some sample sequence images on the crowd density database (the first row), the simulated ATM monitoring scene (the second row), and the crowd behavior analysis database (the third row).

Although many models with respect to machine learning have appeared continuously [7-12], robust scene analysis under explainable model is still very difficult issue. A big problem is the uninterpretability of deep model. As shown in Fig.2, Based on the interpretability of the model, we consider the recognition problem in a variety of scenarios, constructing a model that accommodates different scene is not an easy task.

So, in this paper, our research starts from the problems of difficult feature extraction of targets and behaviors in complex scenes, small labeled samples, non-restrictive scenes and interpretability of the model, By constructing the deep reinforcement network model based on transfer learning and the self-step learning mechanism, the traditional deep feature learning is further exploited from the explainable view, and the classification effect is improved by constructing the deep reinforcement network based on discriminative Fisher vector, from the frontier of application and algorithm of innovation, it is clearly a challenge and a new attempt.

Main contributions of this paper are described as follows:

First, our work focuses on the robust construction of smart scene analysis system, this system can provide explainable decisions for users without a strong background on Artificial Intelligence. We propose novel methods that aim at solving the model interpretability problem, mainly for neural networks models by enhancing features importance and adding type of explanations;

Second, based on the traditional manifold learning algorithm, a new feature learning process is constructed by means of multi-model modeling and simple cascading deep reinforcement learning network, so as to solve the problem of multi-modal data classification in complex scenes. The measurement mechanism and cross-layer learning method of deep learning adjacency layer are studied, the theoretical model and framework of sample selection and labeling based on self-step learning strategy are proposed, and reinforcement learning combined with transfer learning strategy are also proposed;

In the detection of targets and behaviors, the existing deep network model is fully used to mine the feature information and deep knowledge of scene targets and behaviors. Hierarchical rough and fine expression strategy

cascade are adopted, and the multi-scale lightweight learning method is coordinated to complete the efficient detection of targets and behaviors in the scene. Based on SPP algorithm, a kind of supervised dimensionality reduction algorithm based on sparse representation and non-parameter discriminant analysis is proposed, interpretability of Model is improved;

Last, In order to effectively recognize targets and behaviors in a complex environment, this paper proposes a behavior recognition framework based on multi-perspective deep transfer learning. Dense trajectory is used to describe behavior characteristics, the self-paced learning strategy is adopted to solve the problems of small number of labeled samples and model interpretation. A nonlinear model based on deep transfer learning is proposed to solve the problem that it is difficult to distinguish the perspective-related features from behavior-related features, the interpretability of the model is further improved.

The remainder is designed as follows. In Section 2, some related works are summarized. The constructed explainable machine learning models and corresponding learning algorithms are given in Section 3. Designed experiments and discussion are depicted in Section 4. At last, we conclude our contribution and give future research in Section 5.

2. Related work

Explainable AI (XAI) refers to those Artificial Intelligence techniques aimed at explaining, to a given audience, the details or reasons by which a model produces its output [13]. To this end, XAI borrows concepts from philosophy, cognitive sciences and social psychology to yield a spectrum of methodological approaches that can provide explainable decisions for users without a strong background on Artificial Intelligence. Therefore, XAI targets at bridging the gap between the complexity of the model to be explained, and the cognitive skills of the audience for which explainability is sought. Interdisciplinary XAI methods have so far embraced assorted elements from multiple disciplines, including signal processing, adversarial learning, visual analytics or cognitive modeling, to mention a few. Although reported XAI advances have risen sharply in recent times, there is global consensus around the need for further studies around the explainability of ML models. A major focus has been placed on XAI developments that involve the human in the loop and thereby, become human-centric. This includes interpretable reasoning of models, neurosymbolic reasoning or systems based on fuzzy rules, etc.

This paper mainly verifies the interpretability of the model through three scenarios, namely crowd density estimation in the monitoring scene, face occlusion detection in front of ATM, and crowd abnormal behavior recognition in the video scene.

2.1 crowd density estimation

At present, the main methods for crowd density estimation are detection-based approaches and regression-based approaches [14-18]. We found that both methods have their advantages. Detection-based methods work better when the crowd density is low. When the crowd density is high, the effect of regression method is better. It is challenging to generate an accurate crowd distribution diagram, and one of the major difficulties is discretization. People do not occupy only one pixel in the image, and the density diagram needs to maintain the continuity of local neighborhood. Other difficulties include the variety of scenes and camera angles. The main reason for the small size of the crowd

density estimation database is the large amount of image marking, which requires marking each head of the dense crowd in the image. Popular crowd density estimation based on CNN network [15-18] mainly adopts multi-scale structural network. Although good performance has been achieved, there are two problems: when the network gets deeper, it is easy to fall into local optimal, and the decision is is short of interpretability.

The application of deep learning scheme in crowd counting has achieved substantial progress [18,19,20,22]. In order to accommodate the multi-scale changes, most of the previous methods adopted multi-column CNN [26,27,29,33] or multi-branch blocks to extract multiple features under different scenes, and then finally fused the obtained features to obtain the final density estimate map, but there are still many problems that have not been perfected solved. 1) Existing methods use multi-column stacked convolutional networks to generate density maps, and do not consider the relationship between the feature layer channels, only the accumulation in space. 2) Existing methods only use the traditional Euclidean Loss optimization of the proposed model, the traditional Euclidean loss itself presents some shortcomings, it is difficult to deal with the blurred images that indicate high sensitivity of outliers. Therefore, it should be noted that convolution kernels of different sizes multi-scale spatial and temporal characteristics, each sub-convolution module independently minimizes the regression loss of its current scale, thus, the final fusion density map is predicted. However, when designing multiple subnets, there is no collaborative regression itself, which results in that the final fusion density map is not optimal in a certain sense, the resulted density map is of low quality, resulting in fuzzy results , 3) The existing methods do not focus on the consistency of generated estimated density map at multiple scales, that is to say, the sum of the crowd number of local patches does not necessarily correspond to the total number of their original patches .

Early crowd density detection methods were mostly based on image processing and computer vision application technologies [23,24,25,37,39]. With the development of related fields, people can use computer technology to analyze image data and obtain valuable information from the images. As a result, crowd density estimation based on computer vision is gradually emerging [44,45,46,47].

Traditional methods mainly refer to methods based on image processing and some shallow classification or regression learning models [13,14,15,17,21]. Main ideas of such methods generally include image acquisition, image preprocessing, feature extraction, feature analysis and classification, and calculation of the final crowd density map. Among them, the most critical step is the process of image feature extraction and analysis [67,68,69]. According to the extracted feature category, it can be divided into direct method and indirect method [70,71,72]. Direct method: refers to the method based on detecting human features and counting, common features have direction Gradient histogram, Hear wavelet features, etc. After successfully obtaining features, some mainstream classifiers are then adopted, such as: Radial Basis Function , Support Vector Machine, Ada-boost, etc. Feature analysis scheme is used to get a completed detection model. The indirect method is a regression-based method, which generally constructs a function mapping relationship between population density-related features and population density levels through regression models. Common regression models include Gaussian process regression, linear regression, and ridge regression.

Traditional methods often require proper image preprocessing according to different application scenarios. The advantages and disadvantages of such methods are all very obvious. The advantage is that the processing speed of

regression methods is faster, and it is easier to conduct reasonable and effective algorithm research and improvement on the model. The disadvantage is that such methods require a large number of positive and negative sample training sets to get a good training model [65,66], and the scope of application is relatively narrow and only be used in low-density scenes with sparse crowds, also it is not suitable for scenes with dense crowds and severe occlusion, as well as scenes with complex and changeable backgrounds.

However, it can be seen in the relevant literature that the crowd density estimation detection method based on deep learning also presents some defects in network structure. They did not grasp the relationship between sub-block fusion and the original image density map, nor did they consider the inherent connection between layers, the main problem is that the short of interpretability.

2.2 face occlusion detection in front of ATM

With the rise of ATM-related crime, enhancing security through surveillance technology has been at the top of the agenda in academia and industry. Although cameras are usually installed in ATMs to capture images of the user's face, the feature is limited to recording subsequent criminal investigations. Therefore, facial occlusion detection becomes very important to prevent ATM-related crimes. Traditional methods of solving this problem usually include location, segmentation, feature extraction and recognition [16-27]. At present, there are also some studies at home and abroad on intelligent monitoring of ATM, but most of them are used to detect unconventional behaviors such as running, squatting, jumping, fighting, prying ATM and installing illegal equipment before ATM [16]. Xia et al. [21] proposed a robust and effective facial occlusion detection method based on convolutional neural network and multi-task learning. Compared with the previous method, the multi-task learning strategy is added to improve the performance of the learning algorithm by learning multiple task classifiers jointly. However, different tasks in this algorithm need to be trained separately, so the implementation is complicated and the training takes a long time. Chen et al. [26] introduced an anti-occlusion face perception detector that detects the occlusion face and divides the occlusion area at the same time, and introduced a countermeasure training strategy to detect the blocked face area. However, this algorithm has a high requirement for background segmentation and high computational complexity, so it is easy to get into local optimization. Zhao et al. [27] proposed a robust automatic decoder model based on the LSTM model, which can effectively detect partially obscured faces even in the field. However, this algorithm is sensitive to complex backgrounds and is greatly influenced by the head posture and movement.

2.3 crowd abnormal behavior recognition

In recent years, the detection of abnormal behavior in the population has attracted considerable attention in the field of public safety [32,35]. Real-time monitoring of abnormal behaviors in the scene can not only reduce the cost of human monitoring, but also deal with emergencies in a timely manner. One of the biggest challenges in computer vision, given the difficulty of defining group anomalies, is to define and analyze groups of people. Early researchers tried to do these efforts, such as providing and judgment of mathematical analysis model of social force model [28], for example, the abnormal behavior prediction model based on trajectory judgment [31], the conventional detection method is limited to block out crowd behavior (covered) between pedestrians, crowded, low resolution, ignore the interaction of people.

Zhang et al. [29] proposed a population abnormal behavior detection algorithm based on the characteristics of motion prospect effect map for abnormal behavior in the population. This algorithm uses the adaptive mixed Gaussian model to block the video frame image, and combines the obtained foreground region to calculate the motion effect diagram of the moving foreground object block, but the algorithm is very dependent on the segmentation of the foreground, so it is not robust. In view of the low accuracy and real-time performance of crowd monitoring in public places, Hu et al. [30] proposed a detection method of crowd abnormal behavior based on motion saliency graph. Because light flow is easily affected by noise, the false alarm rate of this algorithm is very high. Mousavi et al. [33] proposed a novel video descriptor, called directional trajectory histogram, to identify abnormal conditions in crowded scenes. Instead of using the standard method of estimating the motion vector from just two consecutive frames, they divided the video sequence into space-time cuboids and collected the tracks of the crowd for statistical data. Similarly, the algorithm is based on optical flow characteristics, and the false alarm rate is relatively high. Rojas et al. [34] proposed the Gaussian mixture model (GMM) to simulate the behavior of abnormal population and fully consider the characteristics of abnormal population behavior. This algorithm can automatically adapt to environmental changes and online learning without the need to track the population and large-scale training data. However, this method only computes feature points in the region of interest, and general interest points are difficult to determine.

2.4 Explainable machine learning scheme

As machine learning becomes more widely used in more areas, understanding the reasons behind model decisions has become the trend for future model development, which is more convenient for government legislation and project security.

At present, explanations provided by different algorithms are fragmented and independent, which makes it difficult to determine reasonable decisions and explain model structures. In addition, in the design of interpretable classifier, the selection of optimal training set, correlation selection of heat graph, semantic analysis, model visual interpretation and error analysis cannot be combined compulsively. Moreover, the text interpretation can't match the characteristics of a certain layer of deep learning network, and it lacks of continuous interpretibility. In general, the text explanation generated for classification comes from training data based on model annotations. Up to now,data labels are set manually and are very subjective, and it doesn't take into account the differences between the different elements. Therefore, it is not possible to determine the relevant region of the image that is most useful for classification. In most cases, experts are encouraged to use their attribute label data as interpretable evidence. Existing interpretable artifical intelligence models can provide the basis behind the classification. However, in the existing classification model,there is no mechanism for identifying potential misclassification of classifiers. Warning users about misclassification will help prevent errors from entering the system.

One of the reasons for misclassification is the reduction of distance between classes. Some outliers or edge elements of a class can share the common characteristics of adjacent classes. However, there is no mechanism to ensure the number of subclasses of a given class and whether it makes sense to merge closely related subclasses of two adjacent classes into a new class and implement the correct classification. Hagras [51] proposes a solution based on database

transaction model interpretation, whose explanation is on the basic of logical structure or reasoning. The static structure makes it unsuitable for the deep network classifier. In the constructed model, the system dynamically give appropriate explanations from stored vocabularies, which in turn are generated based on model learning. It provides a consistent view of models and interpretations beyond the scope of existing technology.

Reference [52] proposes an interpretable model in which they are interactively validated through visual features and similarity. Moreover, k-means clustering was used to analyze the similar features, so that the average features obtained had greater robustness and relatively low time complexity. In addition, it does not consider the importance or relevance of model, nor does it cluster with respect to output classes. Samek et al. [53] conducts the explanatory demonstration of the model mainly from the aspect of correlation calculation. By observing the change of the connection mode of the network, the hidden layer is explained visually. However, the feature learned by the network layer are not described in detail. Mao et al. [54] proposed an image interpretation and generation method based on visual features. Take an image signature with a fixed length of 8000 to generate a caption. In this model, the correlation of features is firstly determined and the signature generation is carried out on this basis. Since eigenvalues can be of any length, strategies that follow highly correlated features are interpretable. As the power of interpretation becomes more important in intelligent decision-making, AI systems are no longer there to serve as black boxes. Decision makers of AI services have the right to know the reasons behind their decisions so that they can better play to their strengths.

3. Proposed model

This paper makes an deep study of scene understanding and recognition from the most cutting-edge technical perspectives such as deep reinforcement learning, self-paced learning and transfer learning. In view of the existing problems in the field of scene recognition in complex environments, this paper considers the solutions based on deep network, and further studies popular algorithms such as deep reinforcement learning, self-paced learning and transfer learning, aiming at the urgent model explanation problems in the framework of deep network. On this basis, object detection technology and behavior recognition technology in complex environments are specifically studied.The overall technical framework of the paper and the relationship between the research contents of each part are shown in Figure 3, and the research of this paper is also carried out according to these contents one by one.

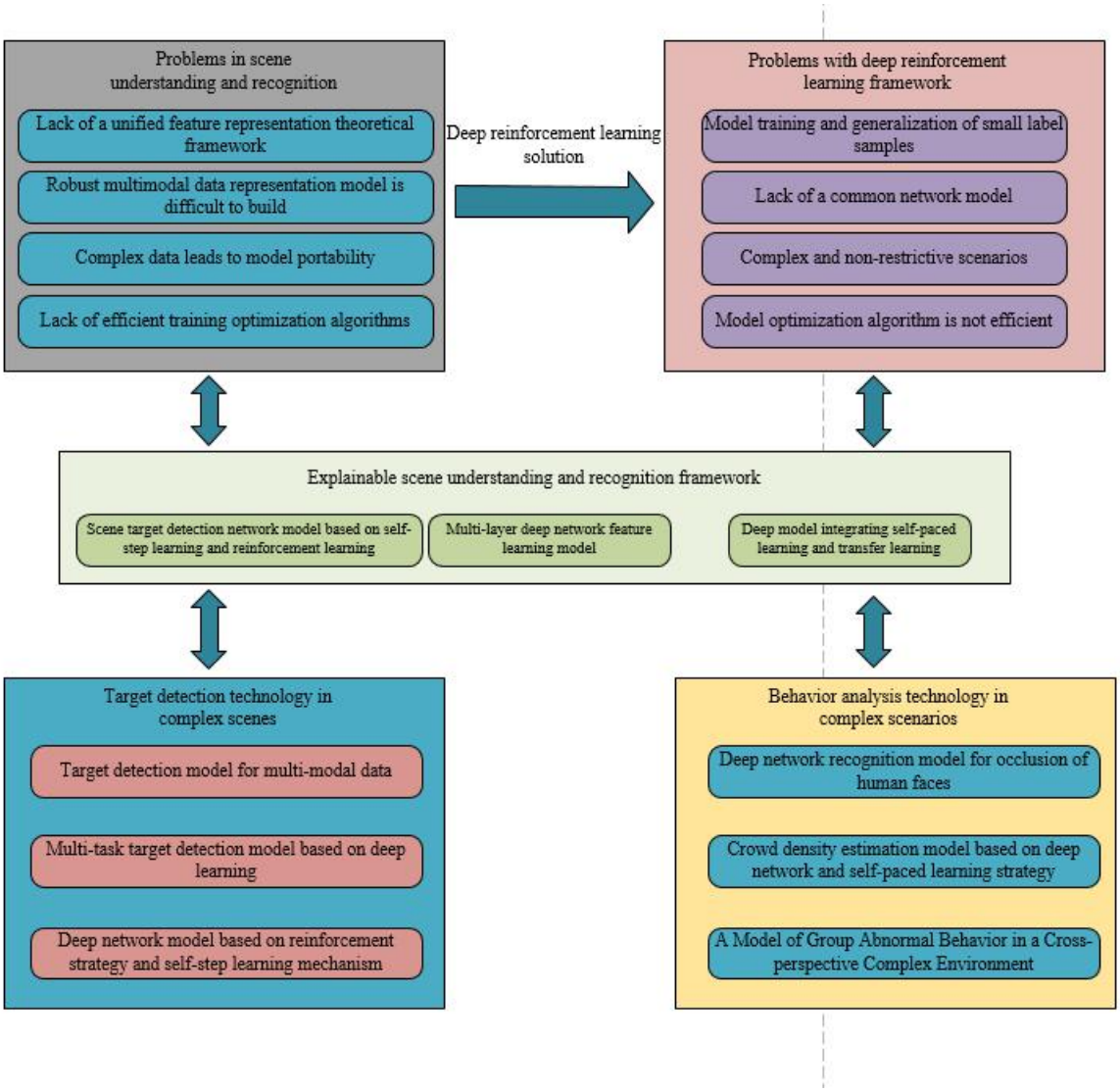


Fig. 3. Proposed technical framework diagram

3.1 Explainable scene object detection model of combing self-paced learning and deep reinforcement learning

The deep network model is designed for learning across model characteristics.This paper proposes a new application in deep network, using deep network to learn multi-mode. In particular, this paper demonstrates that cross-modal feature learning - if multimodal features are present during feature learning, better features can be learned for a modal (multimodal learning, monomodular-tested). In addition, the paper designed how to learn a shared feature between multiple modes and evaluate it on a particular task -- the classifier was trained with only audio data but tested on video-only data. Multi-mode explainable deep network model can be seen in Figure 4. This model consists of two streams, one for video information and the other for audio information. The structure of the two streams is identical, each consisting of eight layers (including the input layer).

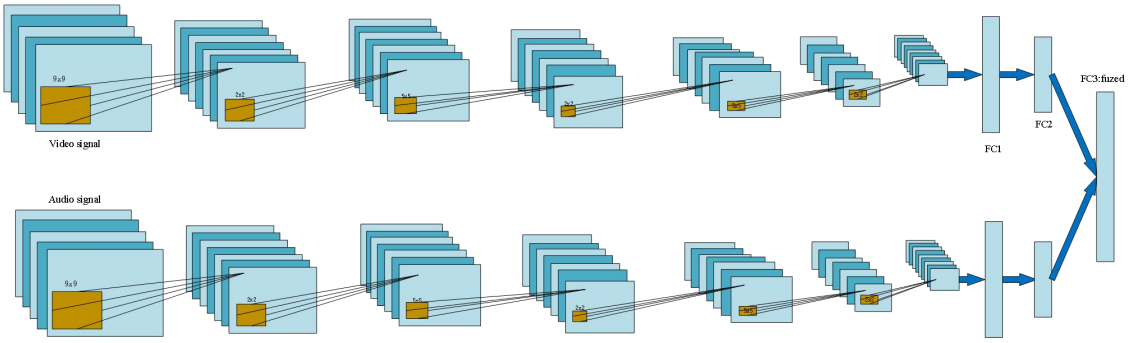


Fig. 4. Proposed multi-mode explainable deep network model

There are two problems with the traditional multimodal model. First, there is no clear goal for the model to find correlations across modes. Some hidden layer units adjust the parameters only for voice, and others adjust the parameters only for video, so that it is possible for the model to find the desired feature. Second, there is only one mode for supervised training and testing in the cross-modal learning arrangement, which makes the model unexplainable. If there is only one modal representation, it is necessary to integrate observable variables that are not observed. Therefore, this paper proposes a deep self-coding model to solve the above problems.Inspired by the noise-reducing self-coding model, this paper proposes a training two-mode deep self-coding model (Figure 4), which uses an extended (extended single-mode input) but noisy data set. In fact, the model is still required to reconstruct the two modes when one mode uses zero as input and the other uses the original value as input when expanding.Therefore, one-third of the training data is input only by video, one-third of the training data is input only by voice, and the last third has both video and voice. This model can be viewed as an example of multitasking learning.

When designing the intensification strategy, the paper uses the Q network to interact with its environment during the data generation phase. The system looks at the current scene, which consists of audio and video frames, and takes actions using the -greedy strategy. This environment in turn provides scalar rewards.Interaction experiences are stored in replay memory M. Replaying M preserves N recent experiences, which are then used to update the network parameters during the training phase. In the training stage, the network structure will use the data stored in replay memory M to train the network.Assume that the superparameter n represents the number of experiences replay, and for each experience replay, a mini-cache B containing several interactions is randomly sampled from the finite size replay memory M. The model will be trained by sampling from cache B, and the parameters of the network will be updated iteratively in the direction of The Behrman target.The algorithm is divided into two phases to avoid latency. Therefore, this paper divides the algorithm into two stages: in the first stage, the robot collects data through limited time interaction with human beings; In the second stage, it enters the stage. During this rest phase, the training phase is activated to train the multimodal depth Q network.

In this paper, for the sake of the regularization model and make it thin, it is needed to make each unit has a hidden layer using the regularized the expected activation function of punishment, the form of the regularized punishment is the need to focus on research, it determines the cell activation function of hidden layers on the sparse sex (whether the function of the activation of the hidden layer unit is activated or not).

In order to avoid non-convex optimization problems from falling into poor local solutions, the proposed network optimization method adopts multiple random initializations to train the model, and then chooses the initialization network with the best effect to construct the model. However, this method is too adhoc and the calculation cost is too high. Self-learning is just the best solution to non-convex optimization problems.The curriculum learning is to simulate the cognitive mechanism of human beings by first learning simple and universal knowledge structure and then gradually increasing the difficulty to learn more complex and specialized knowledge. However, self-learning has been improved in course learning. Instead of assigning prior knowledge to sample learning sequence in advance, the learning algorithm itself determines the next learning sample in each iteration.

3.2 Explainable occlusion face detection model

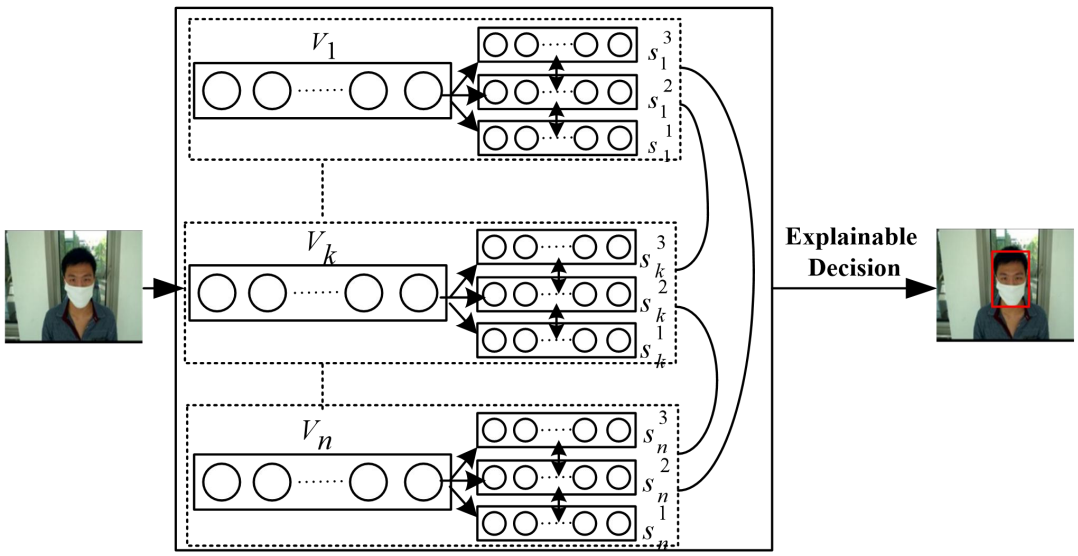


Fig. 5. Proposed occlusion face detection model

As shown in Figure 5, in our built model, $v=[v_1,v_2,\ldots,v_k,\ldots,v_n]$ represents the voting values for the n*th* part. $s=[s_1^i,s_2^i,\ldots,s_k^i,\ldots,s_n^i],(i=1,2,3.)$ represents the I layer voting value of the n*th* part, which is an invisible random vector. It should be noted that during the training, the values of V and S will be adjusted according to the feedback learning between layers, and these two values are the part that needs to be studied in the model design. In the above designed deep model, in order to increase the interpretability of the model, we assume that Y represents the detected window. Then, from the perspective of probability, we can get the data distribution p(y) of Y, which can be expressed by the following formula:

$$p(y)=\sum_v p(y|v)p(v)=\sum_v p(y|v)\sum_x p(v|x)q(x) \tag{1}$$

Where q(x) is the empirical distribution on the data. The model is degraded to the distribution $p(y|v)$ of the bottom layer and the distribution $p(v)$ of the top layer. We noticed that if we want to stratify, by fixing $p(v|y)$, we could only

learn a priori $p(v)$. Therefore, this prior value $p(v|y)$ will not be optimal when all characteristics of the data are not retained. Therefore, we plan to set $p(y|v,s)=e^{\sum_i yvisi}$ to achieve $p(y|v)$:

$$p(y|v)=\sum_s e^{\sum_i yvisi}\, p(s|v) \tag{2}$$

Where formula (2) can be calculated using the average field theory. In addition, we plan to design an optimization algorithm for adjacent layers, which requires training parameters layer by layer. The probability distribution is as follows:

$$p(s_k^i \left| s^{i+1}, v\right.)=\alpha(w_k^i s^{i+1}+q_k^i+\beta_k^i v_k^i) \tag{3}$$

$$p(s_j^{i+1} \left| s^{i}, v\right.)=\alpha(w_j^i s^{i}+q_j^{i+1}+\beta_j^{i+1} v_j^{i+1}) \tag{4}$$

Where $k, j \in n$, $k \neq j$, and w_k^i denotes the correlation between the representation layer s^{i+1} and s^i , β_k^i is the weight of the correlation between the voting function and other parts, q_k^i and q_j^{i+1} represents the bias term.

3.3 Density regression

As mentioned in the introduction, most previous methods use L2 loss to optimize their networks and generate predicted density maps. Most methods use multi-column convolution kernels to generate density maps, and then join multiple columns of density maps to form the final density map. It is assumed that a path i is calculated forwardly as Si, so the loss of its full path can be defined as follows:

$$L = Min_F \left|\left| F(S_1 S_2 S_3 \ldots - M) \right|\right|_2^2 . \tag{5}$$

Where M represents the ground-truth heat map, and $F(S_1 S_2 S_3)$ denotes the fused heat map generated using different convolution kernels.

The definition of such loss will have the following main problems:

1. The first concern is that most of the methods only use multi-column convolution kernels to generate density maps, without taking into account the internal relationship of the network feature layer, only to extract features in the spatial stacking network of the sub-network. We have noticed that some of the information extracted inside the feature layer is useful, and some is basically useless , so if you just blindly pass the extracted feature layer to the next layer, in a certain sense, it makes the semantic information extracted in the subsequent deep layers worse, thus the generated density map will be fuzzy and unreliable in comparison.

2. Secondly, it can be noted that even if different convolution kernels are used to extract semantic information at different depths, the method of using multiple columns of convolution kernels does not achieve a real collaborative way, but in a competitive manner. Because each sub-network is implementing one thing: the density map generated by the sub-network under the convolution kernel is closer to the real density map, which leads to the competition of

multiple convolution kernels since the features extracted under different convolution kernels are not suitable for another convolution kernel.

3. Thirdly, the population density estimation with multi-scale information lacks the constraint of cross-scale. Since the input scale of each sub-network is different, previous methods do not consider relationship of the generated density map and previous original size, and find that there is a certain loss in the fusion density map generated by the subnet and the density map of the original size.

3.4 Network architecture diagram

Fig. 6 shows the structure of our adversarial generation network. In our method, the generator network G learns from input images of different scales and generates corresponding density maps. This is an end-to-end mapping. Specifically, we used U-net as the encoder-decoder structure to construct the generator. In order to deal with the scale changes, we use two different size, namely G-large and G-small. Above two generators of different sizes cooperate with each other. G-larger extracts large-scale feature of target, and G-small focuses on small-scale information of target. For G-large generator, eight network layers are used in the encoder part, each layer contains a batch normalization layer and corresponding Leaky-Relu activation layer, and among the eight convolutional layers in each encoder part add feature self-learning layer for better feature extraction, and deconvolution in the subsequent decoder, each layer also has batch normalization layer and relu activation layer (the last layer not), skip connections are added during the deconvolution operation, these layers are connected after the self-learning of the corresponding convolution layer features. This is also a process to obtain a clearer contour map. G-small and G-large have similar structures. The structures of G-large and G-small generators could be shown in Table 1. The size of the input picture is 720*720 and 200*200 respectively, and the ratio of the input and output part is consistent.

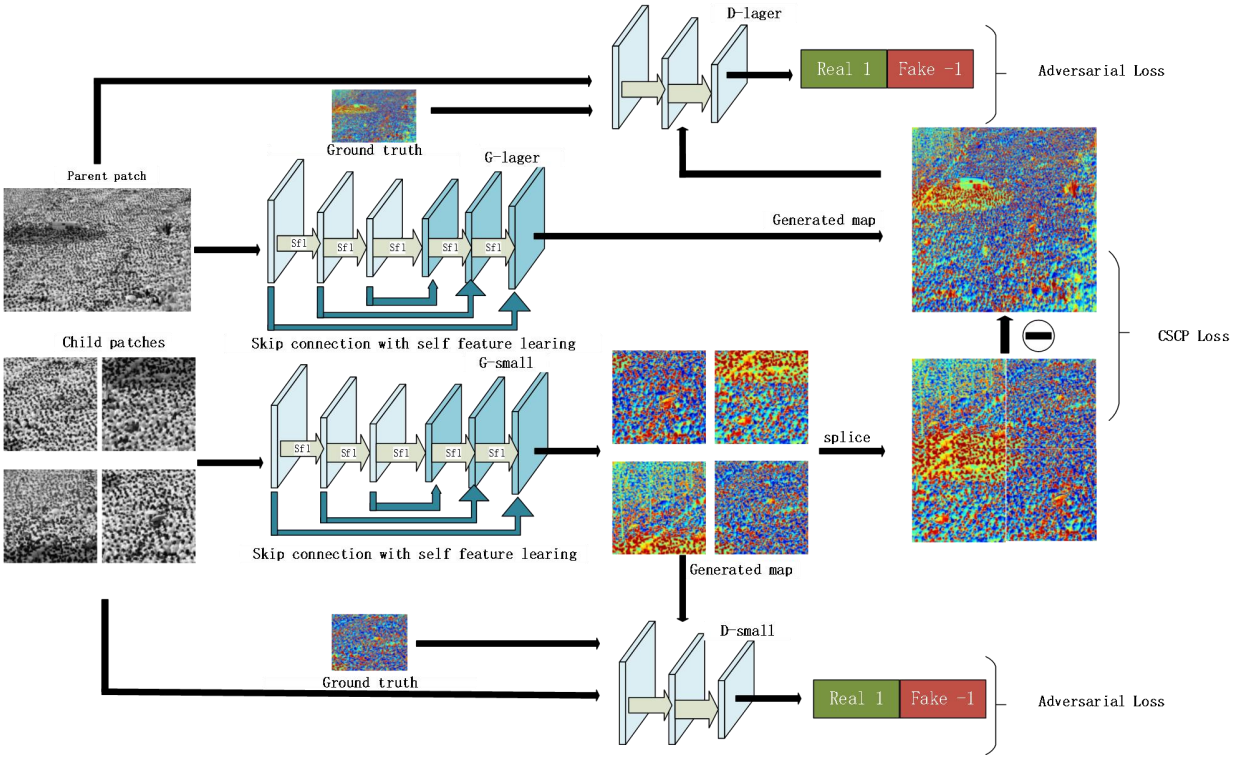


Fig. 6. Feature self-learning generation adversarial network.

TABLE I
STRUCTURE OF THE NETWORK GENERATORS G-LARGE AND G-SMALL

Layer	G-larger	Layer	G-small
1	6*6*64conv, stride 2+ SFL (64)	1	4*4*64conv, stride 2 + SFL (64)
2-7	4*4*64conv, stride 2+ SFL (64)	2-6	4*4*64conv, stride 2+ SFL (64)
8	4*4*64conv, stride 1+ SFL (64)	7	3*3*64conv, stride 1+ SFL (64)
9	3*3*64decv, stride 1	8	3*3*64decv, stride 1
10-15	4*4*64decv, stride 2	9-13	4*4*64decv, stride 2
16	6*6*1decv, stride 2	14	4*4*1decv, stride 2

As the previous Fig. 6 also illustrates our discriminator structure, input the corresponding patch density map (both generated and ground truth), for patches of different sizes, the density map they generate is consistent with its patch size. The discriminator contains five convolutional layers, a batch normalization layer and corresponding Leakrelu layer (the last layer not), which serves as feature extraction layer. And the tanh function is placed at the end of deep network structure to return the probability score to -1.0~1.0, which means that if it is True (it is closer to 1.0),otherwise it is False (it will be closer -1.0), it can be seen from Fig. 5 that the D-large and D-small have the same network structure.

3.5 Feature self-learning

Convolutional neural network is composed of several convolutional layers, non-linear layers and down-sampling layers, so it can describe the image by capturing the feature of the image from the whole cognitive field. However, it is very hard to construct a powerful network to capture better features. The difficulty comes from many aspects. Therefore, we hope to select better features between different channels, that is, redirect the features to subsequent convolutional layers to obtain better features.

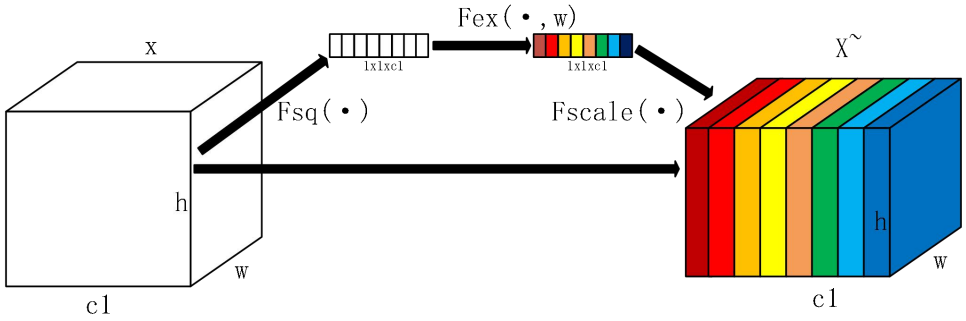


Fig. 7. Feature self-learning network structure

As shown in Fig. 7, the feature self-learning module mainly includes two steps [48]. First, the crowd density map is computed based on the spatial dimension, and then The two-dimensional eigenmatrix is transformed into a one-dimensional number, each real number has the global receptive field of the current channel in a sense. And it has input features and output dimensions. The number of channels is matched, which is useful for characterizing the global probability distribution of input sample and obtaining the global receptiveness field that is close to network layer. The specific definitions are as follows:

$$z_c = F_{sq}(\mu_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mu_c(i, j), z \in R^c. \tag{6}$$

Where μ_c represents the c-th convolution kernel, H and W are the image resolution, and i and j denotes the pixels in the image that are long and wide.

After obtaining the global description feature after the first step, the next step is to grab the relationship between the channels. In the grabbing stage, it is necessary to ensure that it is flexible at first, and the nonlinearity relationship between the various channels is needed to determine. The second point is that the learned relationship is not mutually exclusive, allowing multi-channel features instead of one-hot form, so a sigmoid-style gate mechanism is used to generate a weight for each network channel, where the learned model parameter w is used to reflect the correlation between the modeling layers, which makes the effective channel information be enlarged. At the end, the weight output through the gate mechanism is regarded as the optimal channel selection scheme, and then each channel are re-weighted to the other layers to complete the channel size restoration, so that the function before the calibration channel selection is defined as follows:

$$S_c = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 RELU\ (W_1 z)). \tag{7}$$

Where $W_1, W_2 \in R^{\frac{c}{r}}$, In order to further optimize the model and improve the transformation ability of the model, the bottleneck structure with two fully connected layers is adopted, the first fully connected layer plays a role of dimensionality reduction operation could be finished by one of the two fully connected layers, the dimensionality reduction coefficient r is a hyperparameter, and then activated by ReLu. The last fully connected layer restores the original dimensions.

3.6 Cross-scale consistency estimation

As mentioned in the introduction, the traditional method uses L2-based regression to train a multi-scale network and finally forms a fused prediction density map. It also mentions why it is only based on L2 regression will make the final density map fuzzy estimation. In order to solve such problems and make the final density clearer, we use adversarial loss. It comes from the generative adversarial network. The adversarial generative network involves the model generator G and model discriminator D. The two are like playing a minmax game: the image genera-ted by the training generator G deceives model D, while the aim of training D is to distinguish the synthesized image from the actual image, if it is inconsistent, it is Fake, and if it is consistent, it is True [41]. In our method, this adversarial loss is defined as follows:

$$\begin{aligned} L_A(G, D) = E_{x,y \sim P_{data}(x,y)}[LogD\ (x, y)] + \\ E_{x \sim P_{data}(x)}[Log\ (1 - D(x, G(x)))]. \end{aligned} \tag{8}$$

Where x indicates the training patch and y indicates the corresponding ground heat map. The aim of G is to minimize this goal, and D tries to maximize it.

Therefore, compared with the traditional regression loss, the advantages of using adversarial loss are as follows. The traditional pixel-by-pixel Euclidean loss is based on large deviations between pixels, so when facing sharp edges or outliers, it will make the feature map fuzzy, and thus the generated density map will become fuzzy. But the adversarial loss discards the large deviation between the existing pixels. It is a binary judgment for each pixel, either true or false.

Using this adversarial loss will stimulate the distribution of true values. In other words, if the generated picture tends to be blurred, the discriminator D will tend to -1.0 to avoid the blurred picture and generate a clearer picture because of its excitation.

Because of the lack of punishment based directly on the ground real image, only using adversarial losses can sometimes lead to abnormal spatial structure. As suggested in previous work, we also use two common losses to smooth the solution. Details are as follows:

• Euclidean loss: In our model, the L2 loss is also adopted to change the estimated density map G into the discriminator model D, and to approximate the basic facts in the sense of L2. Assuming that W×H resolution image with c channels is given, we design the following rule about defining the pixel-by-pixel loss:

$$L_E(G) = \frac{1}{c}\sum_{c=1}^{c} \| P^G(c) - P^{GT}(c) \|_2^2 . \quad (9)$$

Where $P^G(c)$ represents the pixel generating the heat map and $P^{GT}(c)$ indicates the pixel of the ground truth map, here we set C = 3.

• Perceptual loss: This kind of loss function was originally added into the image task by Johnson. For image conversion and super-resolution tasks, it compares the features obtained by convolution of the real picture with the features obtained by the convolution of the generated picture, making the high-level information (content and global structure) are more similar, which means trying to minimize the perceived difference between the two. The definition of perceived loss is as follows:

$$L_p(G) = \frac{1}{c}\sum_{c=1}^{c} \| f^G(c) - f^{GT}(c) \|_2^2 . \quad (10)$$

Where $f^G(c)$ represents the pixels in the advanced receptiveness features of the previous heat map and $f^{GT}(c)$ denotes the pixels in the advanced receptiveness features of the ground truth, it should be noted that C = 128.

So the overall first-stage loss could be defined as follows:

$$L_I = ArgMin_G Max_D L_A(G, D) + \lambda_e L_E(G) + \lambda_p L_p(G). \quad (11)$$

Among them, λe and λp are the weights that redefine Euclidean loss and perceptual loss. After previous work, we set λe =λp = 150.

3.7 Cross-scale consistency constraints

As mentioned earlier, the basis of the L2 loss is adopted, and the perceptual loss is added to make the generated density map better consistent with the ground real density map, but the problem that needs to be solved is the cross-scale, so we used the cross-scale consistency regulator to improve the robustness and generalization between the child patch and the parent patch density map, that is to say, this constraint is to reduce the residual error generated by the child and father patch in the population density estimation. As mentioned, the original method did not notice that each word network works in a specific way, and above sub-networks could not perform well in a cooperative way, making the resulting density maps prone to inconsistent results. More specifically, as can be seen from Fig. 5, during

our model training process, the patches are sent to G-large and G-small, respectively, to obtain the estimated heat map P-parent and generated by G-small At the end of the subnetwork, the four sub-pictures P-child are spliced to form P-concat. The constraint of cross-scale consistency is between P-concat and P-parent. In general, given the W×H density map with c channels, the cross-scale consistency constraint based on L2 loss is defined as follows:

$$L_c(G) = \frac{1}{c}\sum_{c=1}^{c} \| P^{prt}(c) - P^{cnt}(c) \|_2^2 . \tag{12}$$

Where $P^{prt}(c)$ denotes the pixels in the parent block heat map, and $P^{cnt}(c)$ denotes corresponding pixels after the child block density map is spliced, C=3. By continuously optimizing this constraint, the gap between the density map of the parent block and the child block will be forced to decrease. We only pay attention to the total number of people who obtain the entire image, and this constraint is to deal with multi-scale problems, so we can see that this constraint can be more general applied.

The ultimate goal is to combine the four loss functions mentioned above to achieve our final loss function:

$$L_{II} = L_1 + \lambda_c L_c(G). \tag{13}$$

Among them, λc is a predefined weight to achieve cross-scale consistency with respect to constructed loss function. It's worth noting that if λc is 0, the two normal forms will be generated independently.

4 Experiments results

4.1 Data set

Our captured data. We set up cameras in the hallway of a teaching building on the Campus of Shanghai Jiao Tong University, and recorded 12 video clips, each one contains 2 minutes clip. Students and teachers are coming in and out of the building, sometimes carrying a schoolbag, sometimes in groups. This dataset is made up of 28800 frames, and a cycle is set to 300 frames.

UCF CC 50. The UCF CC 50 dataset [11], is a very representative testing samples that contains 50 annotated pedestrians samples, which contain various characters and scenes. The character is between 94 and 4543. Sliding window scheme is used (the size of the sliding step can be set, which will cause the sample size obtained to change) for data expansion, and split into 50% for cross-validation to evaluate the proposed method.

Shanghai-Tech. The Shanghai-Tech dataset is created by Zhang [14], and contains 1198 annotated pedestrians samples, which are captured by street View cameras and webcams. Our proposed model is trained and tested on this dataset. In order to increase the training samples, we adjust the samples images to 720*720 resolution and shape 200*200 sub-blocks from the picture.

4.2 Experiment details

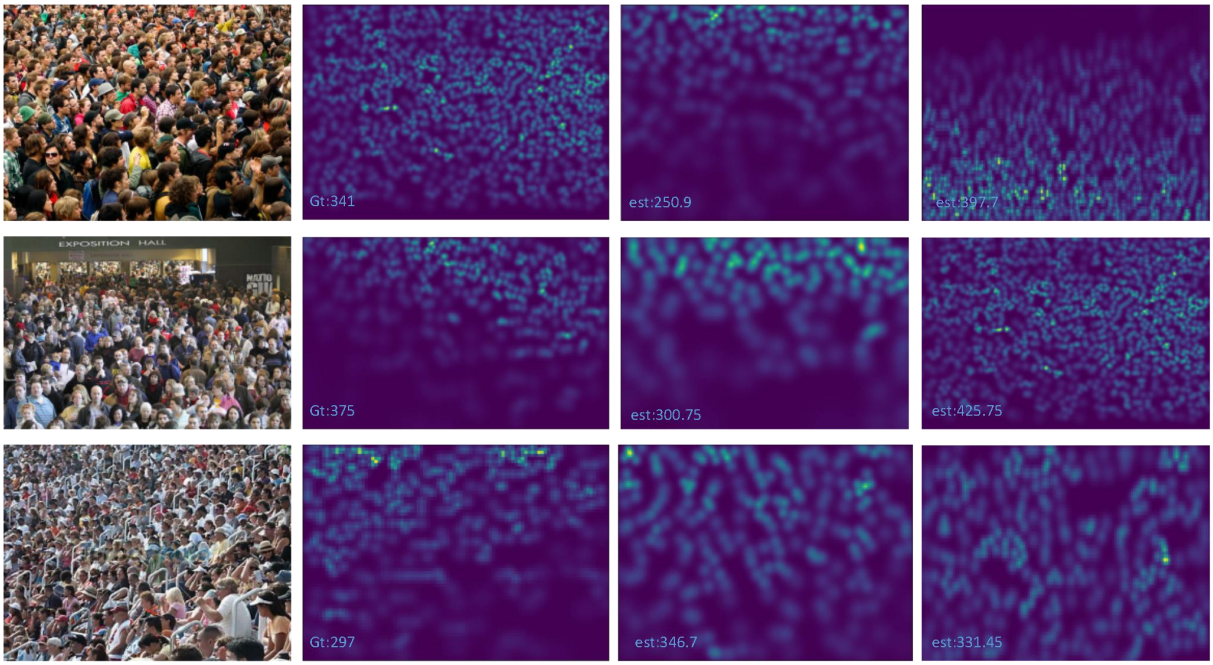


Fig. 8. The comparison between the predicted value and the real value during the training process. The first column denotes the real images, the second column denotes the real heat maps, the third column denotes the prediction heat maps generated by the adversarial network without feature self-learning, and the fourth column the prediction density map generated for the adversarial network generated after the feature self-learning is added

In the algorithm module, the input part is the image pair composed of the image and the corresponding density map. For G-large, the original image is input, while for G-small, the corresponding quarter small image is used to set the overall network. The initial learning rate is 0.00005 to update the parameters in our network. For the population suppression rate -people-thr, it is set to generate more data samples during the data expansion phase. The people-thr set in UCF CC 50 is 15. The corresponding setting in shanghai_partA is 0, and then a sliding window 200*200 is used to randomly crop an image block of a specific size to expand our current data set. The number of iterations of our model in UCF CC 50 and Shanghai_partA is 500 epochs, and the training and testing of our network are implemented on torch, as shown in Fig. 8, we can view the predicted heat map and the real heat map during the training process.

As can be seen from Fig. 8., the prediction density map we get after adding feature self-learning is closer to the true value, so we can see from Fig.9 and Fig.10 below that we can more accurately describe the feature after adding feature self-learning Loss reduction and corresponding changes in MAE accuracy, we also give the performance of net similarity, as is shown in Fig.11.

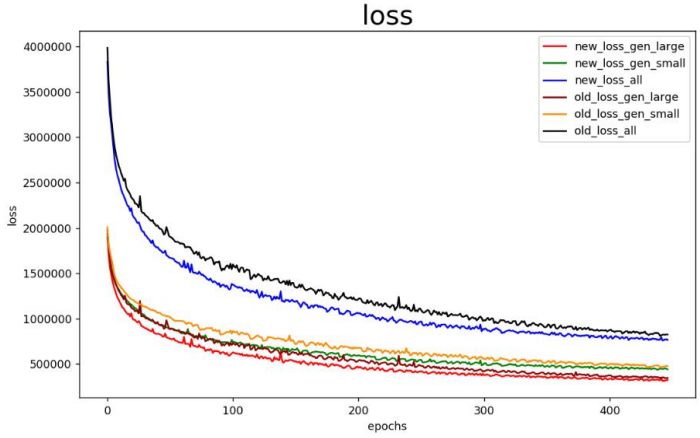


Fig. 9. Comparison of loss_gen_large, loss_gen_small, loss_all after adding SFL during Fold_1 training

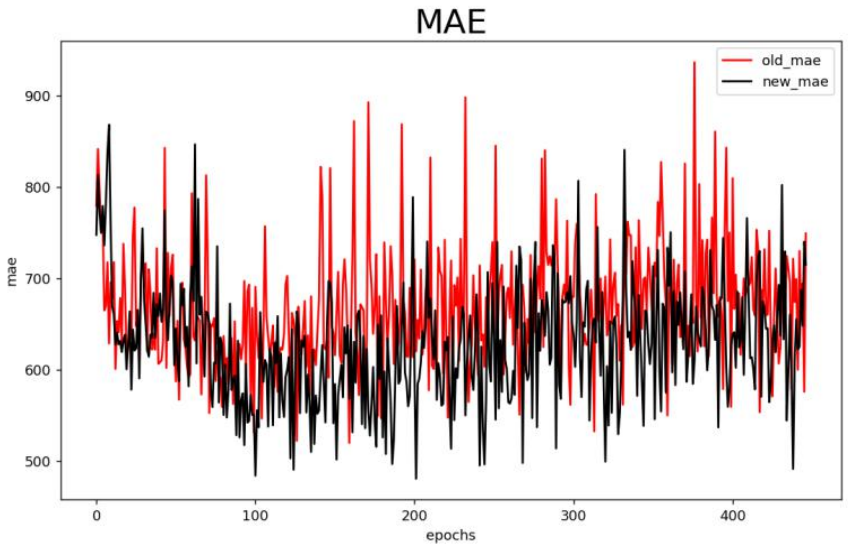


Fig. 10. Fold1_MAE overall convergence after joining SFL

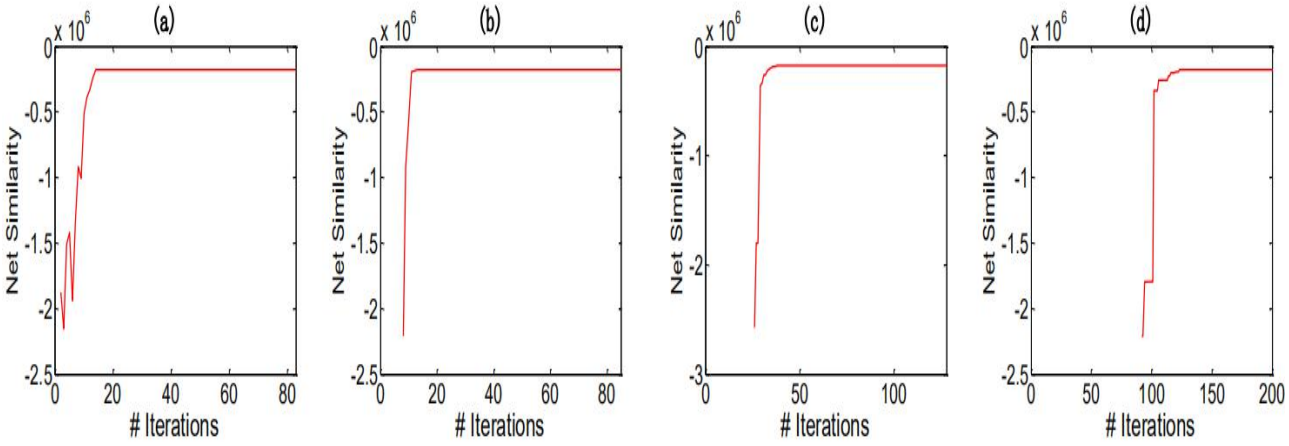


Fig. 11. Training of net similarity

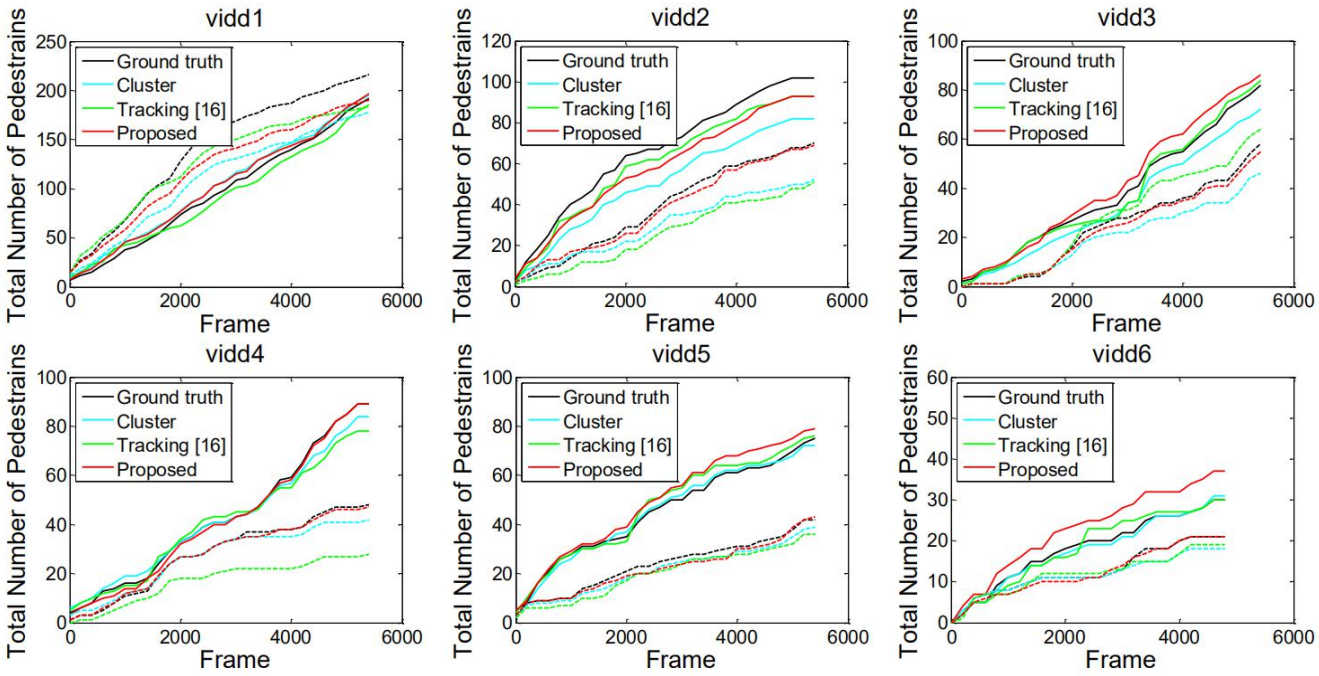


Fig. 12. Comparison in the Our constructed data set

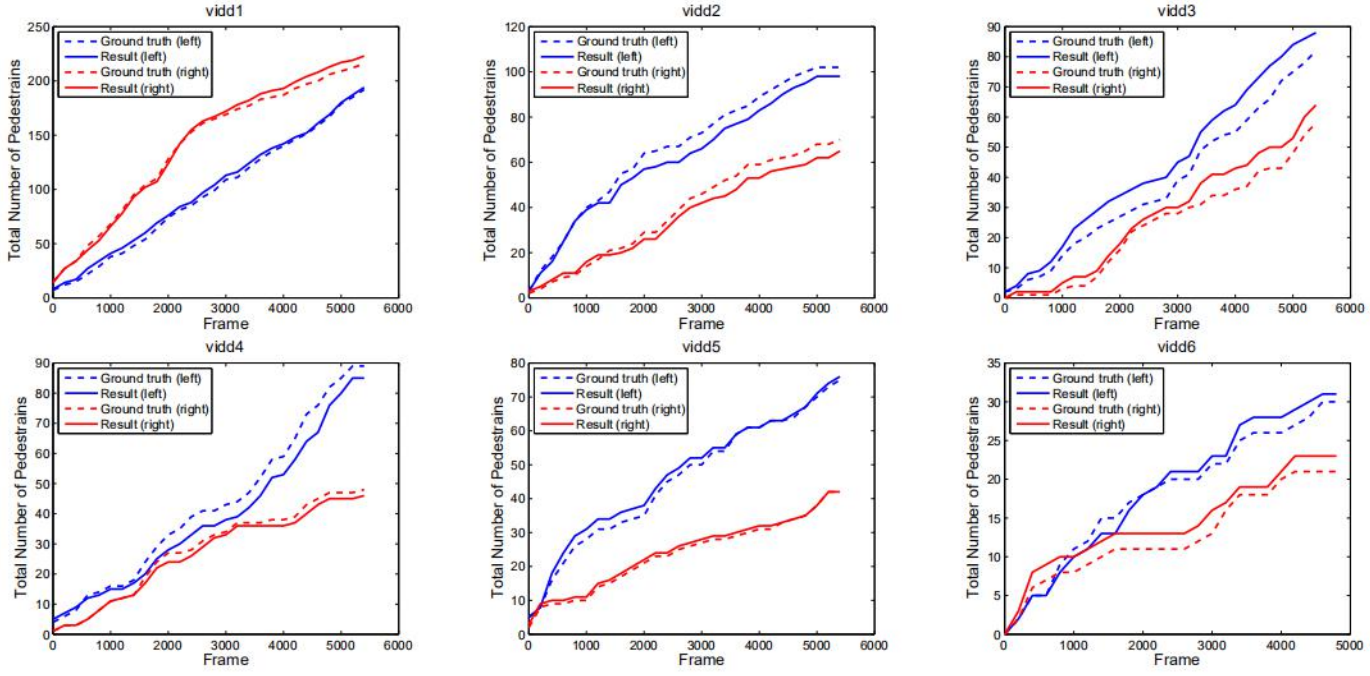


Fig.13 Comparison in the Our constructed data set

Also we give the performance of the whole system in Fig. 13, through observing performance of our constructed system with the ground truth, we can draw that the performance of our designed system is very stable. Some problems mainly derive from occlusions and the interaction of other goals.

4.3 Comparison with the latest technology

The proposed model is evaluated with several latest models on our constructed dataset and two benchmarks, and the results are given in Fig.12, Tables 2 and 3. From all the tables, we notice that our method is always much better than the previous method. Table 3 gives a comparison of the Shanghai-Tech Part-A data set, and their images are closer to the real monitoring screen than other data sets. Our proposed SFL-GAN has achieved considerable improvement

compared to the existing technology. Table 2 indicates that proposed method obtained the best MAE and competitive MSE among the five latest methods in the UCF-CC-50 dataset.

TABLE II
COMPARISON IN THE UCF CC 50 DATA SET

Methods	MAE	MSE
Idress [11]	419	541
C.Zhang [14]	467	498
Y.Zhang [27]	377	509
Vishwanath A [37]	322	341
D.B.Sam [33]	318	439
SFL-GAN(ours)	**290**	443

TABLE III
COMPARISON IN THE SHANGHAI_PARTA DATA SET

Methods	MAE	MSE
Zhang [14]	181	277
Deepak Babu Sam [55]	154	229
Y.Zhang [27]	110	173
Vishwanath A [37]	101	152
D.B Sam[43]	97	145
SFL-GAN(ours)	**93**	137

5 Conclusion

In this work, we propose to construct a crowd density prediction machines via self-learning generative adversarial network in the framework of soft computing, the proposed model is used to obtain more accurate data collection develop an intelligent prediction machines for crowd detection and counting. Considering the decentralized, secure and trusted features of the IoT, an IoT generation algorithm for data under surveillance scenarios is designed. Moreover, based on the serious occlusion problem of the crowd, aiming at different crowd densities, a novel self-learning generative adversarial networks model is first proposed, then a feature self-learning layer to the generator, to avoid the blur of the generated image, the anti-loss is constructed in the following step. Finally, in order to avoid excessive cross-scale loss, the cross-scale consistency criterion is used to optimize the final fusion density map. The proposed model can processes the data from real video scene, the modules we designed can effectively save and improve the speed of the system, so all the steps can be operated in an effective way and the effectiveness of the system is very high.

In the following research, we will pay more attention to the improvement of real-time algorithm and effectiveness of IoT-based hardware, also be able to generate density maps closer to the ground truth during the density map generation phase based on more effective machine learning approaches.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NO. 61702226); the 111 Project (B12018); open Fund of Jiangsu Key Laboratory of Image and Video Understanding for Social Safety, Nanjing University of Science and Technology, Nanjing (J2021-7).

Conflict of interest: The authors declare that they have no conflict of interest.

Data availability: The data that support the findings of this paper are available from the corresponding author.

Authorship contributions: **Bin Wu**: Investigation, Methodology, Writing- original draft, Supervision.

Yuhong Fan, Tao Zhang: Writing- Reviewing and Editing, Methodology, Visualization.

Yeh-Cheng Chen: Validation, Visualization, Data curation.

REFERENCES

[1] M.Tironi, M.Valderrama, The militarization of the urban sky in Santiago de Chile: the vision multiple of a video-surveillance system of aerostatic balloons[J]. Urban Geography, 14:1-20 (2019).

[2] Y.Zhang, J.f. Wan，T. Wang，Y.h Zhang, Physically-Based Rendering for Indoor Scene Understanding Using Convolutional Neural Networks, In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition , (2018).

[3] C.Sakaridis, D.Dai, G.L.Van, Semantic Foggy Scene Understanding with Synthetic Data[J]. International Journal of Computer Vision, (2017).

[4] Z.Qiu, Y.Zhuang, H.Hu, et al. Using Stacked Sparse Auto-Encoder and Superpixel CRF for Long-Term Visual Scene Understanding of UGVs[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 50(4):1331-1342, (2020).

[5] M.Tironi, M.Valderrama.,The militarization of the urban sky in Santiago de Chile: the vision multiple of a video-surveillance system of aerostatic balloons[J]. Urban Geography, (14):1-20, (2019).

[6] K.Arulkumaran, M.P.Deisenroth, M.Brundage, et al. A Brief Survey of Deep Reinforcement Learning[J]. IEEE Signal Processing Magazine, 34(6), (2016).

[7] S. An, W. Liu, and S. Venkatesh. Face recognition using kernel ridge regression. 2007 IEEE Conference on Computer Vision and Pattern Recognition, pages. 1–7, (2007).

[8] M. Mirza, S. Osindero, Conditional Generative Adversarial Nets, arXiv:1411.1784,(2014).

[9] H. Idrees, I. Saleemi, C. Seibert, and M. Shah. Multi-source multi-scale counting in extremely dense crowd images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages. 2547–2554, (2013).

[10]A.Pentina, V.Sharmanska, and C.H.Lampert. Curriculum learning of multiple tasks. In Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition, pages. 2547–2554, (2015).

[11]L.Lin, K.Wang, D.Meng, et al. Active Self-Paced Learning for Cost-Effective and Progressive Face Identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, PP(99):7-19, (2017).

[12]A.Holzinger, C.Biemann, S.P.Constantinos, and B.K.Douglas, What do we need to build explainable ai systems for the medical domain?, arXiv:1411.1784,(2017).

[13]A. Barredo Arrieta, N. Diaz-Rodriguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”, Information Fusion, vol. 58, pp. 82-115, June (2020).

[14]A. Bansal and K. S. Venkatesh. People counting in high density crowds from still images. CoRR, abs/1507.08445, (2015).

[15]C. Zhang, H. Li, X. Wang, and X. Yang. Cross-scene crowd counting via deep convolutional neural networks. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages. 833–841, (2015).

[16]Y.Z. Xia, B,L. Zhang. Face Occlusion Detection Using Deep Convolutional Neural Networks. International Journal of Pattern Recognition and Artificial Intelligence, 30(09):1-24, (2016).

[17]G. Fernández, Á. F, S. L, M. M. R. Multiple target tracking based on sets of trajectories[J]. IEEE Transactions on Aerospace and Electronic Systems, 56(3):1685-1707, (2020).

[18]M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440, (2015).

[19]C. Zhang, H. s. Li, X. g. Wang, and X. k. Yang. Cross-scene crowd counting via deep convolutional neural networks. In CVPR, pages. 833–841, (2015).

[20]A. Radford, L. Metz, S. Chintala, Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, Computer Science, (2015).

[21]J.L. Xing, Z.H. Niu, J.S. Huang, W.M. Hu, X. Zhou, S.C. Yan. Towards Robust and Accurate Multi-view and Partially-occluded Face Alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence, PP:987-1001, (2018).

[22]T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan. Crowded scene analysis: A survey. IEEE transactions on circuits and systems for video technology, 25(3):367–386, (2015).

[23]P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image- to-image translation with conditional adversarial networks. arXiv preprint arXiv:1611.07004, (2016).

[24]J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In European Conference on Computer Vision, pages 694–711. Springer, (2016).

[25]C. Li and M. Wand. Precomputed real-time texture syn- thesis with markovian generative adversarial networks. In European Conference on Computer Vision, pages. 702–716. Springer, (2016).

[26]Y.J. Chen, L.X. Song, R. He. Adversarial Occlusion-aware Face Detection. 4th Asian Conference on Pattern Recognition, pages 354–361, (2018).

[27]F.Zhao, J.S.Feng, J.Zhao, W.H.Yang, S.CYan. Robust LSTM-Autoencoders for Face De-Occlusion in the Wild. IEEE Transactions on Image Processing, 27(2):778-790, (2018).

[28]Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Singleimage crowd counting via multi-column convolutional neural network. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages. 589–597, (2016).

[29]C. Ledig, L. Theis, F. Husza r, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint arXiv:1609.04802, (2016).

[30]A. Dosovitskiy, T. Brox, Generating Images with Perceptual Similarity Metrics based on Deep Networks, arXiv:1602.02644, (2016).

[31]Y. Li, S. Liu, J. Yang, and M.-H. Yang. Generative face completion. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2017).

[32]L. Zeng, X. m. Xu, B. l. Cai, S. Qiu, and T. Zhang. Multi-scale convolutional neural networks for crowd counting. In ICIP, pages. 465–469. IEEE, (2017).

[33]M. Arjovsky, S. Chintala, L. Bottou, Wasserstein GAN, arXiv:1701.07875, (2017).

[34]D. B. Sam, S. Surya, and R. V. Babu. Switching convolutional neural network for crowd counting. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2017).

[35]V. A. Sindagi and V. M. Patel. Cnn-based cascaded multitask learning of high-level prior and density estimation for crowd counting. In Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on, pages. 1–6. IEEE, (2017).

[36]S. Zhang, G. Wu, J. P. Costeira, and J. M. Moura. Understanding traffic density from large-scale web camera data. arXiv preprint arXiv:1703.05868, (2017).

[37]H. Zhang, V. Sindagi, and V. M. Patel. Image de-raining using a conditional generative adversarial network. arXiv preprint arXiv:1701.05957, (2017).

[38]V. A. Sindagi and V. M. Patel. Generating high-quality crowd density maps using contextual pyramid cnns. 2017 IEEE International Conference on Computer Vision, (2017).

[39]D. Xu, W. l. O. Yang, Xavier. Alameda-Pineda, E. Ricci, X. g. Wang, and N.cu. Sebe. Learning deep structured multi-scale features using attention-gated crfs for contour prediction. In NIPS, (2017), pages. 3961–3970.

[40]L. Zhang, J. Dai, H. c. Lu, Y. He, and G. Wang. A bi-directional message passing model for salient object detection. In CVPR, (2018), pages. 1741–1750.

[41]O. Sbai, M. Elhoseiny, A. Bordes, Y. Lecun, C. Couprie, DeSIGN: Design Inspiration from Generative Networks, (2018), arXiv:1804.00921.

[42]Z. Shen, Y. Xu, B. b. Ni, Min.si. Wang, J. g. Hu, X. k. Yang. Crowd Counting via Adversarial Cross-Scale Consistency Pursuit. 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2018), pages. 5245-5254.

[43]Y. Li, X. Chen, Z. Zhu, L. Xie, G. Huang, D. Du, X. Wang, Attention-guided Unified Network for Panoptic Segmentation (CVPR), (2018),1812.0394.

[44]D. B. Sam, R. V. Babu, Top-Down Feedback for Crowd Counting Convolutional Neural Network, (2018), In AAAI.

[45]W. z. Liu, M. Salzmann, and P. Fua. Contextaware crowd counting. arXiv preprint , (2018),arXiv:1811.10452.

[46]L. Zhang, M. j. Shi, and Qiao.bo. Chen. Crowd counting via scale-adaptive convolutional neural network, (2018), In WACV. IEEE.

[47]N. Liu, Y. c. Long, C. q. Zou, Q. Niu, L. Pan, and H. f. Wu. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding, (2018), arXiv preprint arXiv:1811.11968.

[48]Z. l. Qiu, L. bo. Liu, G. b. Li, Q. Wang, N. Xiao, and L. Lin. Crowd counting via multi-view scale aggregation networks, (2019), In ICME.

[49]J. Hu, L. Shen, S. Albanie, G. Sun, E. hu. Wu. Squeeze-and-Excitation Networks, (2017), In arXiv:1709.01507

[50]I. J. Goodfellow, J. P. Abadie⋆ , M. Mirza, B. Xu, D. W. Farley, S. Ozair†, A. Courville, Y. s. Bengio. GenerativeAdversarialNets, (2017), In arXiv:1406.2661v1.

[51]H. Hagras, ”Toward Human-Understandable, Explainable AI” in Computer, vol. 51, no. 09, pp. 28-36, (2018).

[52]A. Punjabi and A. K. Katsaggelos, ”Visualization of feature evolution during convolutional neural network training,” 2017 25th European Signal Processing Conference (EUSIPCO), Kos, 2017, pp. 311-315,(2017).

[53]W.Samek, T.Wiegand, K.R.Mller, Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. ITU J ICT Discov Special Issue 1 Impact Artif Intell (AI) Commun Netw Serv 1(1):3948 (2018).

[54]J. Mao, J.Huang, A.Toshev, O.Camburu, A.Yuille, K.Murphy,Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

[55]C. Y, Ji. H, W. F, et al. Using High-Bandwidth Networks Efficiently for Fast Graph Computation[J]. IEEE Transactions on Parallel and Distributed Systems, (2018), 1-2.

[56]T. Zhang, W. j. Jia, X. j. He, J. Yang, Discriminative Dictionary Learning with Motion Weber Local Descriptor for Violence Detection, IEEE Transactions on Circuits and Systems for Video Technology, (2017), 27(3):696~709.

[57]C. J, L. Y, Z. Z, et al. An adaptive hybrid memetic algorithm for thermal-aware non-slicing VLSI floor planning[J]. Integration, the VLSI Journal, (2017), 58: 245-252.

[58]Y. Yang, X. Zheng, V. Chang, et al. Semantic keyword searchable proxy re-encryption for postquantum secure cloud storage[J]. Concurrency and Computation: Practice and Experience, (2017), 29(19): e4211.

[59]T. Zhang, Z. j. Yang, W. j. Jia, Q. Wu, J. Yang, X. j. He, Fast and robust head detection with arbitrary pose and occlusion, Multimedia Tools and Applications,(2015)，74(21):9365~9385.

[60]T. Zhang, Z. j. Yang, W. j. Jia, B. q. Yang, J. Yang, X. j. He, A new method for violence detection in surveillance scenes, Multimedia Tools and Applications, (2016),74(12):7327~7349.

[61]Y. Cheng, F. Wang, H. Jiang, et al. A communication-reduced and computation-balanced framework for fast graph computation[J]. Frontiers of Computer Science, (2018), 12(5).

[62]Y. D, L. X, S. H, et al. Relative influence maximization in competitive social networks[J]. Science China Information Sciences, (2017), 60(10): 108101.

[63]T. Zhang, W. j. Jia, J. j. Li, J. Sun, H. h. Yang, Fast and Robust Occluded Face Detection in ATM Surveillance, Pattern Recognition Letters, (2018), 107:33~40.

[64]G. L, S. H, Z. W. Efficient approximation algorithms for multi-antennae largest weight data retrieval[J]. IEEE Transactions on Mobile Computing, (2017), 16(12): 3320-3333.

[65]Y. Y, Z. X, T. C. Lightweight distributed secure data management system for health internet of things[J]. Journal of Network and Computer Applications, (2017), 89: 26-37.

[66]W. J, Z. X. M, L. Y, et al. Event-triggered dissipative control for networked stochastic systems under non-uniform sampling[J]. Information Sciences,(2018), S0020025518301749.

[67]T. Zhang, W. j. Jia, C. Gong, J. Sun, X. n. Song, Semi-supervised Dictionary Learning via Local Sparse Constraints for Violence Detection, Pattern Recognition Letters, (2018),107:98~104.

[68]Y. z. N, W. q. L, X. K. CF-based optimisation for saliency detection[J]. IET Computer Vision, (2018), 12(4):365-376.

[69]Z. Tao, J. Zou, J. W. Fast and robust road sign detection in color images. Applied Intelligence, (2019), 48:4113~4127.

[70]T. Zhang, W. j. Jia, B. q. Yang, J. Yang, X. j. He, Z. l. Zheng, MoWLD: A Robust Motion Image Descriptor for Violence Detection, Multimedia Tools and Applications, (2017), 76(1): 1419~1438.

[71]W. S, Guo. W. Sparse multi-graph embedding for multimodal feature representation[J]. IEEE Trans. Multimedia PP, (2017), 99: 1-1.

[72]N. Y, Chen. J, G. W. Meta-metric for saliency detection evaluation metrics based on application preference[J]. Multimedia Tools and Applications, (2018), doi:10.1007/s11042-018-5863-2.

[73]Z, Jian & Dong, Le & Wu, L. Wen. New Algorithms for the Unbalanced Generalized Birthday Problem. IET Information Security, (2017), 12. 10.1049/iet-ifs.2017.0495.

[74]L. B, G. W, X. N, et al. A pretreatment workflow scheduling approach for big data applications in multicloud environments[J]. IEEE Transactions on Network and Service Management, (2016), 13(3): 581-594.

[75]G. Liu, Z. Chen, Z. Zhuang, W. Guo, G. Chen A unified algorithm based on HTS and self-adapting PSO for the construction of octagonal and rectilinear SMT." Soft Computing, (2020), 24(6): 3943–3961. doi: 10.1007/s00500-019-04165-2.

[76]G. Liu, W. Guo, R. Li, et al. XGRouter: high-quality global router in X-architecture with particle swarm optimization[J]. Frontiers of Computer Science, (2015a), 9(4): 576-594.

[77]G. Liu, W. Guo, R. Li, Y. Niu, G. Chen XGRouter: high-quality global router in X-architecture with particle swarm optimization[J]. Frontiers of Computer Science, (2015b), 9(4): 576-594.

[78]G. Liu, W. Guo, Y. Niu, G. Chen, X. Huang A PSO-based-timing-driven Octilinear Steiner Tree Algorithm for VLSI Routing considering Bend Reduction." Soft Computing,（2015c）,19 (5): 1153–1169. doi:10.1007/s00500-014-1329-2.

[79]G. Liu, X. Huang, W. Guo, Y. Niu, G. Chen Multilayer Obstacle-Avoiding X-Architecture Steiner Minimal Tree Construction Based on Particle Swarm Optimization." IEEE Transactions on Cybernetics, (2015d), 45 (5): 989–1002. doi:10.1109/TCYB.2014.2342713.

[80]T. Ma, Q. Liu, J. Cao, Y. Tian, A. D. A, A.-R. M LGIEM: Global and local node influence based community detection, Future Generation Computer Systems, (2020), 105, 533–546.

[81]Q. Ye, Z. Li, L. Fu, Z. Zhang, W. Yang, G. w. Yang. G Nonpeaked Discriminant Analysis for Data Representation. IEEE Trans. Neural Networks Learn. (2019), Syst. 30(12): 3818-3832.

[82]Y. Dongyi, Chen, et al A novel and better fitness evaluation for rough set based minimum; attribute reduction problem[J]. Information Sciences, (2013), 222(3):413-423.

[83]Z. Yu, Z. Yu., Y. Chen. Multi-hop mobility prediction[J]. Mobile Networks and Applications, (2016), 21(2): 367-374.

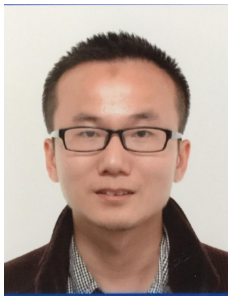
[84]K. Gai, Y. Wu, L. Zhu, et al. Permissioned Blockchain and Edge Computing Empowered Privacy-Preserving Smart Grid Networks[J]. IEEE Internet of Things Journal, (2019),7992-8004.

BIN WU received the B.Sc. degree from Jiangnan University, Wuxi, China, in 1996, and the M.S. degree from Jiangnan University, Wuxi, China, in 2005. He is currently a lecturer with the School of Internet of Things Engineering, Jiangnan University. His major research interests include visual surveillance, object detection, integrated circuit design and application of embedded system.

Yuhong Fan graduated from Shandong Agricultural University, Tai'an, China in 2008. In 2011, she obtained a master's degree from Xihua University, Chengdu, China.

She is currently a lecturer in the Department of Computer Engineering，LangFang YanJing Vocational Technical College, Sanhe, China. She has studied many topics and written several high-quality journal papers and conference papers. His current research interests include data mining, information systems, wireless networks, artificial intelligence, Internet of things and security, medical data analysis, visual monitoring, scene understanding, behavior analysis, target detection and pattern analysis.

Yeh-Cheng Chen is a PhD at the Department of Computer Science, University of California, Davis, CA, USA. His research interests are radio frequency identification (RFID), data mining, social network, information systems, wireless network artificial intelligence, IoT and security.



Tao Zhang received the bachelor's degree from Henan Polytechnic University, Jiaozuo, China, in 2008,and the Ph.D.degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 2016.

He is currently a associate professor with the Jiangsu Provincial Engineering Laboratory for Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi, China. He has led many research projects (e.g.,the National Science Foundation and the National Joint Fund), He has authored over thirty quality journal articles and conference papers. His current research interests include data mining, information systems, wireless network, artifificial intelligence, IoT and security, medical data analysis, visual surveillance, scene understanding, behavior analysis, object detection,and pattern analysis.