

Received January 30, 2021, accepted February 9, 2021, date of publication February 22, 2021, date of current version April 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3061084

Multi-Region Two-Stream Deep Architecture for Visual Power Monitoring Systems

JINRUI GAN¹, WEI JIANG², TING ZHAO¹, PENG WU¹, GUOLIANG ZHANG¹,
AND ZIWEN ZHANG³

¹Artificial Intelligence on Electric Power System State Grid Corporation Joint Laboratory (GEIRI), Global Energy Interconnection Research Institute Company Ltd., Beijing 102209, China

²State Grid Corporation of China, Beijing 100031, China

³Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China

Corresponding author: Jinrui Gan (ganjinrui@geiri.sgcc.com.cn)


This work was supported by the State Grid Corporation of China Headquarters Project (Basic Technology Research on Video Behavior Analysis of Power Grid Operations Using Artificial Intelligence) under Grant 5500-202058318A-0-0-00.

ABSTRACT Judging imaging quality is an important part of the maintenance of visual intelligent monitoring systems for electrical power scenes. However, accurate and efficient identification of possible abnormalities in imaging quality remains challenging. This paper proposes a novel multi-region two-stream deep architecture to improve judging abnormalities. The proposed architecture incorporates two-stream scheme and multi-region strategy to identify relevant information and explore hidden details. More specifically, in addition to color and intensity in the original images, the two-stream scheme uses high-frequency structure information from gradient images to enhance its performance. The multi-region strategy employs spatial pyramid random cropping and region fusion to handle locally non-uniform changes among categories: spatial pyramid random cropping characterizes images at different spatial pyramid levels, while region fusion focuses attention on cropped regions relevant to quality perception by using adaptive learning weights in a fully connected layer. In this way, the proposed strategy guides the framework to adequately and adaptively explore the discriminative regions hidden in the input images, and provides an end-to-end learning procedure. Experimental results demonstrate its strong performance for judging abnormalities, and the proposed method can be easily extended to the entire surveillance system.

INDEX TERMS Abnormal judgement, power systems, deep learning, two-stream scheme, region fusion.

I. INTRODUCTION

Intelligent monitoring systems (IMSs) are playing increasingly significant roles in the rapid development of smart electrical grids and unmanned substations. They can aid the efficiency of normal operation by providing visual technical support and intelligent decisions for power production and management. They also enhance safety by controlling the dispatch of emergency services during an emergency [1]. However, any abnormality in imaging quality (*e.g.*, distortion or jittering of output images) will severely affect the reliability, availability, and stability of an IMS. Therefore, understanding the operational status of the monitoring equipment in real time and further judging any abnormality in its imaging quality are essential parts of daily maintenance.

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea F. Abate .

Generally speaking, the quality of captured video images is directly linked to the status of the monitoring equipment. Old or failed cameras or lenses can lead to abnormal focus, brightness, gain, or color cast, or even complete video signal loss. Transmission failure, poor contact, or electromagnetic interference can cause noise to overlay the image; *e.g.*, streaks and snowflakes. Fig. 1 shows a variety of typical imaging abnormalities. Historically, judging any abnormality in imaging quality has mostly been a manual task, making it expensive, time-consuming, somewhat subjective, and also decreasingly practicable as schemes for monitoring power system increase in scale. Therefore, this paper focuses on the automatic analysis of captured video images to develop a system to accurately judge abnormalities in imaging quality, thereby facilitating the timely and efficient maintenance of the entire surveillance system.

Generally, abnormalities can be automatically identified or located through various computer algorithms or machine

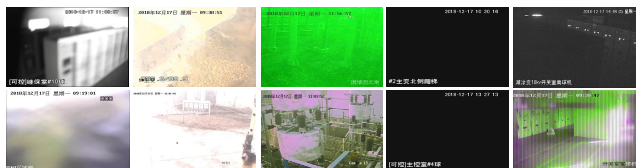


FIGURE 1. Example of abnormal images captured by an intelligent monitoring system (IMS). From left to right: abnormal focus, abnormal brightness or gain, color cast, no video signal, and overlay noise.

learning models [2]–[8]. In computer vision, most of techniques are built upon specific highly discriminative and local invariant characteristics with the help of certain classifiers (*e.g.*, support vector machines). Once the classification model is trained from extracted features, the consequent judgment task can be automatically performed. Therefore, extracting discriminative and robust features from images is key to satisfactory judgment performance. Existing feature-extraction methods can be roughly categorized as hand-crafted [9]–[11] or deep learnt [12]. Given the superiority of automatic feature training over manual design, deep learning frameworks have the potential to extract more relevant features with higher efficiency. During the last decade, representative works [13], [14] have revolutionized the development of deep learning methods in the field of machine learning and pattern recognition. For instance, AlexNet (8 layers) [15], VGG-Net (16–19 layers) [16], GoogLeNet (22 layers) [17], and residual net (152 layers) [18] have demonstrated great potential in classifying natural images.

Designing deep neural network architecture for the optimal trade-off between accuracy and efficiency has been an active research area in recent years. Deep ConvNets are often overparameterized. Model compression [19]–[21] is a common way to reduce model size by trading accuracy for efficiency. SqueezeNets [22], [23], MobileNets [24], [25], and ShuffleNets [26] *et al.* are another alternatives to handcraft efficient mobile-size ConvNets. Recently, neural architecture search (NAS) becomes increasingly popular in designing efficient mobile-size ConvNets [27]–[30], and achieves even better efficiency than hand-crafted mobile ConvNets by extensively tuning the network width, depth, convolution kernel types and sizes.

Although these previous methods have achieved good results, certain issues remain in the specified task. For example, these classic deep networks take advantage of the original image intensity information, but do not effectively use other useful information from other sources or representations (*e.g.*, gradient information). Two further points are noted for the task of judging abnormalities, as follows.

- Generally, deep networks require extensive training data to learn the parameters. However, the collection of suitably labeled abnormal images for training a deep network is laborious and expensive, and the use of only a limited set of abnormal training data results in overfitting. Therefore, a shallow and small network would

be more suited to the task of judging abnormalities in an IMS.

- Compared with natural images classification, the abnormal images encountered by an IMS show much greater variety in appearance; *i.e.*, the captured abnormal images are characterized by arbitrary noise or dynamic saturation, although they have the same content as normal images. Furthermore, the IMS should identify normal images with non-uniform changes (such as local noise and blur) as abnormal, resulting in small inter-class variations. As the categories of abnormal images appear confusing, their judgement becomes challenging, and a more discriminative deep architecture should be carefully designed.

In fact, the topic of no-reference image quality assessment (NR-IQA) [31], [32] can provide broad opportunities for the task of judging abnormalities in imaging quality. NR-IQA works through the use of observation statistics or learned features to map input images to corresponding subjective perception scores [33]–[35]. Recently, significant progresses have been achieved by exploring deep neural networks for better feature representation [36]–[38]. For instance, the most recent approach is the two-stream convolutional network (TSCN) [39], for which development was motivated by action recognition work [40]. Despite great achievements in NR-IQA, the relevant previous research is not well suited to abnormality judgment in an IMS, as the specified task involves large local changes among categories. Overall, a successful deep architecture must deal with these aforementioned difficulties.

Accordingly, we propose a novel multi-region two-stream deep architecture for judging abnormalities in an IMS. Specifically, the proposed method adopts a two-stream scheme containing the gradient stream and the RGB stream. The underlying idea is that the two-stream techniques generally boost performance by combining information from multiple sources (*e.g.*, action recognition [40]). The RGB stream considers variations of the image intensities and colors, while the gradient stream focuses on extracting structure features in detail, as the gradient of the image plays an important role in many vision tasks [41]. The two-stream scheme can capture much of the higher-level meaning of the image, making it more suited to judging abnormalities in an IMS. Actually, the advantages are further demonstrated in NR-IQA [39].

However, note that the categories of abnormal and normal share certain similarities as well as some differences. Hence, a multi-region strategy is proposed to adequately and adaptively explore the discriminative regions hidden in the images. More specifically, spatial pyramid random cropping is adopted when characterizing images. This sampling, which randomly crops regions at different spatial pyramid levels, provides adequate performance while lowering computational cost. Furthermore, it is worth noting that a commonly used method of aggregating these cropped regions is simply a simple average in the literature. The independent consideration of individual regions brings regression restrictions

and limitations. In contrast, a simple and effective region fusion strategy is proposed and reformulated as a region fusion layer to adaptively identify discriminative regions and provide end-to-end learning for the whole image. The experimental results demonstrate the proposed method's effectiveness and superiority over previous methods for judging abnormalities in an IMS.

Compared to the existing literatures, this paper makes three main contributions as follows.

- The first is the proposed novel multi-region two-stream deep architecture for the task of judging abnormalities in the field of power systems and its improved performance in that task. Although the proposed framework is general, it is considered here in the specific context of a complete surveillance system.
- Instead of patch based methods [37], [39], [42], spatial pyramid random cropping is adopted to sample regions at different spatial pyramid levels, and is demonstrated to achieve good balance between accuracy and efficiency.
- These patch based methods roughly assign image's label to its sampling numerous of patches for training. As a result, it leads to limited performance and gives rise to other issues, e.g., how to reasonably assign labels for these patches in training. Instead, the proposed region fusion strategy defined as a fully connected layer can guide the framework to adequately and adaptively explore the discriminative regions hidden in images, and provide end-to-end learning for the whole image rather than individual patches; this end-to-end learning framework enables our method to avoid confusion in assigning labels and independent consideration for individual patches, and makes our architecture more suitable for this specified task.

The remainder of this paper is organized as follows. Section II reviews related works with emphasis on their limitations. Section III describes the proposed method in detail. The experimental setup and results are presented in Section IV, and conclusions follow in Section V.

II. RELATED WORK

In this section, we provide a brief review of the closely related topic of no-reference image quality assessment (NR-IQA) [31], [32]. For a more comprehensive treatment of general NR-IQA, please refer to [43], [44].

Traditional NR-IQA methods generally follow a two-stage processing framework including feature extraction and quality regression. Related works have shown that the performance of these NR-IQA models heavily depends on their carefully designed quality aware features based on the domain knowledge of natural scene statistics (NSS) [45], [46] and human visual properties [47], [48]. For instance, BRISQUE [45] was proposed to capture the statistics of locally normalized illumination coefficients to evaluate image quality. This type of similar methods rely on various local invariant features related to presence changes in domains; it performs well under previously known and

precisely modeled abnormalities. These approaches that largely rely on the extracted global or local spatial patterns of structures perform well on uniform images. However, the NSS-based NR-IQA methods heavily depend on the domain knowledge on NSS modeling which is too complex to achieve a sufficient understanding.

Recently, convolutional neural networks (CNNs) have significantly pushed the performance of vision tasks based on their rich representation power. Thus, another type of methods are based on deep learning, which has drawn much attention in the field NR-IQA. Multi-layer-structural neural networks have been extensively investigated for NR-IQA, and have demonstrated strong representation ability [49]–[51]. They utilize multi-task learning, rank learning or adversarial learning, and the following fully connected layers map the features to a quality score that is consistent with human perception. Kang *et al.* [36] firstly implemented a shallow CNN model to extract features on small image patches for NR-IQA. The final quality score of an input image was computed by averaging the predictions of all patches cropped from it. Further, multi-task CNN was applied to extend to estimate image quality and distortion type simultaneously [38]. A novel multi-task end-to-end optimized deep neural network (MEON) [38] for NR-IQA was proposed, which consists of two sub-networks a distortion identification network and a quality prediction network sharing layers. DB-CNN (deep bilinear CNN) [52] models synthetic and authentic distortions as two-factor variations, and bilinearly pooled the two pre-trained feature sets into a unified representation. Instead, SGDNet (Saliency-guided deep neural network) [53] additionally considers a sub-task of visual saliency prediction with a shared feature extractor. However, it is worth noting that these training requires ground truths of additional sub-task and subjective quality to be both available, which largely limits the total number of valid training samples. Also, remind again that dataset captured in the electrical power scenes is limited and characterizes local non-uniform degradation. Therefore, multi-region scheme is adopted in the work to adequately and adaptively explore the discriminative regions hidden in images.

Closely related works to ours are sampling regions strategy based convolutional networks [37], [39], [42]. These CNNs methods first sample numerous of patches, and then fed these patches into a convolutional network to get their corresponding classification scores, and finally applies simply average to get score of image. During the training, these methods roughly assign image's label to its sampling patches. As a result, it leads to limited performance and gives rise to other issues, e.g., how to reasonably assign labels for these patches in training. Instead, the proposed method takes a step further by taking not only two-stream scheme but also multi-region strategy into account, resulting in an end-to-end framework. Especially, the proposed multi-region strategy considers spatial pyramid random cropping to sample regions at different spatial pyramid levels; on the other hand, it also uses region fusion strategy defined as a fully connected layer which can

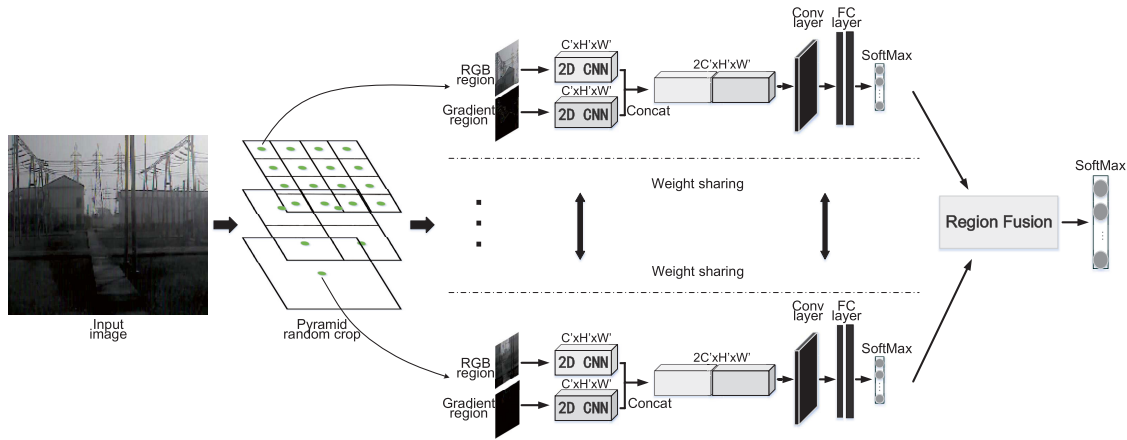


FIGURE 2. Overview of the proposed multi-region two-stream deep architecture. The inputs for the two-stream deep network are the sampled RGB and gradient image regions obtained by spatial pyramid random cropping.

guide the framework to adequately and adaptively explore the discriminative regions hidden in images, providing end-to-end learning for the whole image. Consequently, our method results in better judgement performance.

III. METHOD

This section describes the overall architecture of the proposed framework. As shown in Fig. 2, the two-stream and multi-region schemes are two key features. The two-stream scheme extends general deep networks to the task of judging abnormalities: the two streams for simultaneous analysis are the RGB image and the gradient image information. The two extracted feature maps are subsequently concatenated, and then fed into a sequence of convolutional and fully connected layers that finally branch into a softmax layer for which a classification score is obtained. To handle non-uniform changes among categories, the multi-region scheme adopts spatial pyramid random cropping to randomly sample regions at different spatial pyramid levels. Finally, the classification scores of these regions are combined via region fusion strategy, thus improving performance. Details of the proposed framework are described in the following subsections.

A. TWO-STREAM SCHEME

Intuitively, sufficient and diversity of image representation is beneficial for the abnormal judgement task; *i.e.*, maximize the use of color image to a certain extent is to be encouraged. The colors and intensities of pixels are the original and primary information, and reflect the image quality. In addition, structure patterns among the pixels yield useful information to be explored. As image gradients explicitly reflect the high-frequency information of an image, considering them in image analysis can significantly enhance performance. To this end, the streams of the proposed scheme analyze RGB and gradient information simultaneously, similar to other two-stream works [39], [40].

As shown in Fig. 2, both streams employ the same two-dimensional (2D) CNN structure. The RGB stream

captures color or saturation abnormalities, while the gradient stream highlights high-frequency changes. In theory, the basic 2D CNN used here can be designed using any classic deep architecture such as ResNet18 or VGG-16. Note that the basic 2D CNNs formed here by employing these classic deep architectures retain all of the layers before the last average pooling layer. In the study, we experimented with different 2D CNN models and reported comparison results.

The input for the 2D CNN has a shape of $C \times H \times W$, while the last layer outputs a feature map of shape $C' \times H' \times W'$, where $C = 3$, H and W are the channel, height and width of the input images, C' is the number of output channels, $H' = H/32$, and $W' = W/32$. The network structure for the gradient stream is similar to that for the RGB stream. Subsequently, a concat layer simply concatenates the two feature maps to obtain a shape $2C' \times H' \times W'$, and a following 1×1 conv layer adjusts the channel number to 512 for the concatenated feature map. Two fully connected layers with the nodes size of 1024 and categories number, respectively, come after global average pooling. Finally, a softmax layer outputs a classification score.

Two-stream scheme includes two subcomponents for an image and a gradient image to capture input information at different levels and facilitate the extraction of features from the individual streams. Instead of considering RGB intensities, the two-stream scheme considers high-frequency information in the gradient images. It can improve discrimination between normal and abnormal categories, and the following experiments demonstrate its good performances. Another important characteristic of the scheme is that the 2D CNNs can be replaced by any arbitrary classic deep architecture, making it highly flexible. Overall, the two-stream scheme is designed to allow simple and easy model switching.

B. MULTI-REGION STRATEGY

The deep architecture is assumed to be able to encode locally non-uniform characteristics among categories, as mentioned above. A simple approach is to generate numerous

overlapping patches, and then aggregate them by simple averaging, as employed by the TSCN. As a result, it is computationally expensive, and the independent consideration of sampling patches limits performance and gives rise to other issues, e.g., how to reasonably assign labels for these patches in training. A multi-region strategy is proposed here to avoid these obstacles, further endowing the whole architecture with self-adaptive learning for discriminating regions hidden in images.

Instead of overlapping sampling or other strategies, we propose a new spatial sampling strategy named as spatial pyramid random cropping. An input image is first partitioned into increasingly fine spatial subregions. Typically, $2^l \times 2^l$ subregions ($l = 0, 1, 2$) are used here, as shown in Fig. 2. We then apply random cropping within each subregion to produce $R = 21$ randomly cropped regions. Deep architectures based on these cropped regions are not only robust to variations but also more discriminating in identifying categories for which local changes are dominant.

All these randomly cropped regions should be considered in the final judgement. However, it is noted that these cropped regions should not contribute to perception equally. That is to say, the regions, appearing the local changes relevant to quality perception, deserve more attention. Therefore, a region fusion layer is defined and added:

$$\mathcal{P} = \sum_{r=1}^R \mathbf{P}_r \odot \mathbf{w}_r, \quad (1)$$

where $\mathbf{P}_r \in \mathbb{R}^C$ denotes the probability vectors of the r th region obtained from the two-stream scheme, $\mathbf{w}_r \in \mathbb{R}^C$ denotes the corresponding fusion weight, C is the number of categories, and \odot represents the element-wise product. Rather than use the traditional uniform weight fusion, we attempt to learn the fusion weights of the different regions adaptively when training the whole framework. Further, we can express equation (1) in matrix form:

$$\begin{aligned} \mathcal{P} &= \mathbf{P}\mathbf{w} \\ &= [\mathbf{P}_1^T, \dots, \mathbf{P}_R^T] \begin{bmatrix} \text{diag}(\mathbf{w}_1) \\ \vdots \\ \text{diag}(\mathbf{w}_R) \end{bmatrix}, \end{aligned} \quad (2)$$

where $\mathbf{P} \in \mathbb{R}^{1 \times (C * R)}$ denotes the stacked probability vectors of all regions, and $\mathbf{w} \in \mathbb{R}^{(C * R) \times C}$ denotes the corresponding fusion weight matrix. The aim of region fusion is to learn the optimal region fusion weight values, which is equivalent to the formula for learning the weights of a fully connected layer. Hence, a fully connected layer is applied here to build the region fusion architecture. A softmax layer is subsequently added. In this way, the region fusion weights, \mathbf{w} , can be learnt and adjusted to adaptively find more relevant regions by minimizing the loss function of the softmax layer. Instead of random weights of the fully connected layer in the initialization, the fusion weights are initialized with only the corresponding diagonal elements, while other elements are

simultaneously set to zero. This initial constraint is loosened during training, which means that correlations among different regions can also be considered during fusion. To reduce the risk of overfitting in terms of the region fusion weights, a ℓ_2 regularizer is integrated into the loss function, which can be expressed as

$$\mathcal{L} = \text{loss}(\mathbf{P}, \mathbf{Y}; \mathbf{w}) + \lambda \|\mathbf{w}\|_2, \quad (3)$$

where $\text{loss}(\cdot)$ is the cross entropy loss function of the softmax layer, \mathbf{Y} represents the ground-truth labels, $\|\cdot\|_2$ is the ℓ_2 norm, and the regularization parameter λ enforces smooth of weights.

Conventional region fusion via simple averaging is always independent of the learning process. In contrast, the proposed region fusion strategy is integrated as a layer into the whole learning pipeline. Therefore, the fusion weights can be learned adaptively by referring to some discriminative regions, and the discriminative regions can also be tuned by referring to the fusion weights. In this way, the joint training of the overall architecture can improve its performance. Moreover, it is worth mentioning that the region fusion involves category-specific fusion, which adaptively considers correlations among different categories.

C. IMPLEMENTATION DETAILS

Take VGG-16 based architecture for example, the training parameters configurations are illustrated in Fig. 3. We train our deep architecture using mini-batch stochastic gradient descent with Nesterov momentum, and utilize dropout in each fully connected layer. In addition, we apply data augmentation techniques (e.g., scale jittering with horizontal flipping [54]) and each random crop sampling to the same location of both the RGB and gradient images. We train the network with a momentum of 0.9, a weight decay of 0.0005, and mini-batches of size 16. To avoid overfitting the CNN models in our experiments, we initialize the 2D CNNs with pre-trained weights obtained from a large-scale dataset (i.e., ImageNet [55]). Due to the small dataset size,

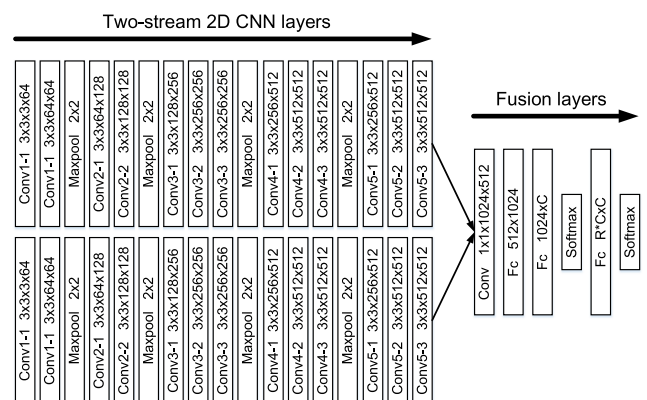


FIGURE 3. Illustration of training parameters configurations for the proposed architecture based on VGG-16. We denote the parameterization of the convolutional layer as “height × width × input channel × output channel”.

we initialize the learning rate as 0.0001 and run the stochastic gradient descent algorithm for 10,000 iterations, but reduce the learning rate by a factor of 0.1 at 4,000 iterations and again at 8,000 iterations. Note that abnormal judgement is essentially a two-category classification problem such that each softmax layer in Fig. 2 outputs a 2 dimensional vector. The complete architecture is implemented and trained end-to-end in PyTorch.

Although our two-stream architecture consists of two 2D CNN branches, the parameters are able to be updated jointly. For training the two-stream networks, we fine-tune the two-stream networks separately for each stream due to the small number of training samples. In particular, we freeze two 2D CNN networks parameters, and tune the conv layer, the two fully connected layers and the softmax layer, ensuring that the convergence is fast so as to reduce the risk of over-fitting. After training the two-stream architecture, we freeze it by further training the topped region fusion layer and softmax layer. Note that spatial pyramid random cropping splits each image into the equivalent of $R = 21$ cropped regions of equal size. These cropped regions share the same weight parameters, as shown in Fig. 2, and thus can be stacked together and fed into the two-stream architecture. This kind of processing makes the architecture straightforward and ensures it is efficiently trained and tested.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This section details the experimental setup in our experiments, and reports the performance of the proposed method.

A. EXPERIMENTAL SETUP

The performance of our method is tested using the IMS outlined in Fig. 4. The overall architecture of the system can be divided into three parts: video server, transmission channel, and functional server. The functional server includes a decoding terminal and a data analysis terminal. The video server stores the video data and device information of each monitoring device, and responds to requests from various clients. The data source can be either real-time streams from cameras or monitoring data stored in a database. Generally, these video data are in YUV format before processing. Therefore, an intermediate video decoder receives the video data and device information through the transmission channel, converts the YUV format into the required format, and then sends the preprocessed video data to the data analysis server for further work. Note that this study focuses on diagnosing the quality of video images in the data analysis server.

The proposed method is based on the constructed IMS. The extensive evaluation uses a dataset containing 4137 images compiled in our experiments. These captured images (all 720×480 pixels) are from different cameras and various environments relevant to power systems. The images are labeled manually via crowdsourcing as either abnormal (1034 images) or normal (3103 images). The training set comprises a random selection of 80% of the images

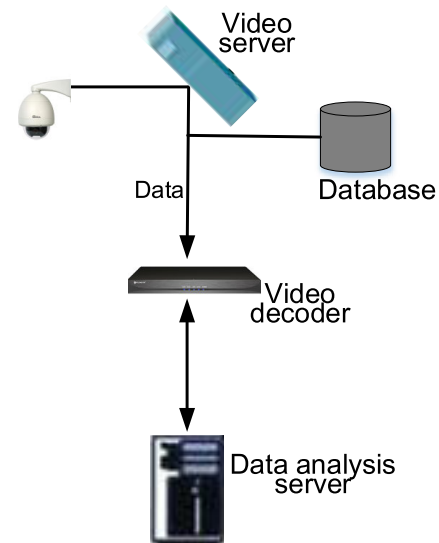


FIGURE 4. Flow diagram of the IMS.

of each category; the remainder forms the testing set to measure performance in the following experiments. All experiments are repeated five times under the fixed ratio of training set to testing set, and the average classification performances are reported. The regularization parameter (λ) of the region fusion layer is determined by five-fold cross-validation. Grid searching using the arrangement of $[10^{-4}, 10^{-3.8}, \dots, 1, \dots, 10^1]$ for λ is used in the following experiments. The proposed architecture is implemented and trained end-to-end in PyTorch on a computer with 40 processors of 2.2 GHz Intel Xeon CPU Silver 4114, 64 GB of random access memory and a single NVIDIA Titan Xp GPU. Using ResNet18 as the 2D CNNs, our method takes approximately average of 58.39 ms to judge an input image (720×480 pixels) on a single NVIDIA Titan Xp GPU.

In an imbalanced binary classification problem, alternative ways to think about predictions are *Precision*, *Recall* and *F1-measure*, which are widely used in pattern recognition and information retrieval communities [56]. *Precision* (*Pre*), *Recall* (*Rec*) and *F1-measure* (*F1*) are defined in equation (4):

$$\begin{aligned}
 Pre &= TP/(TP + FP), \\
 Rec &= TP/(TP + FN), \\
 F1 &= 2 * Pre * Rec/(Pre + Rec), \quad (4)
 \end{aligned}$$

where TP denotes the number of correctly classified abnormal samples, FP is the number of samples falsely classified as abnormal, and FN is the number of samples falsely classified as normal. There is an inverse relationship between precision and recall. That is, it is possible to increase one at the cost of reducing the other. Hence, the F1-measure is introduced to simply combine them, with a higher $F1$ value indicating better predictions.

B. ABLATION STUDY

The operation status of the monitoring equipment has a generally constant effect on the IMS, and any failure would complicate and hinder the subsequent intelligent decision modules. The performance of the whole system therefore relies heavily on the algorithm for judging imaging quality. Complete evaluation of the proposed method is undertaken as follows by assessing its performance from different perspectives.

As mentioned above, our deep architecture is designed to be flexible and allow any arbitrary CNN architecture to replace the 2D CNNs. The following experiments evaluate different 2D CNNs (VGG-16, ResNet18, MobileNet V3-small [29], and EfficientNet-B0 [30]), and assess performance with respect to experimental setup.

1) ANALYSIS OF THE TWO-STREAM SCHEME

To verify the effectiveness of our two-stream scheme, we compare it with a one-stream scheme (RGB only). Specifically, the input for the 2D CNNs in our deep architecture is only RGB information rather than RGB and gradient information simultaneously. Fig. 5 presents comparative experimental results based on the four 2D CNNs mentioned above. The *F1-measure* of the two-stream scheme is higher than that of the one-stream scheme in all cases, demonstrating its effectiveness. Interestingly, VGG-16 and ResNet18 show much greater improvement when using the two-stream scheme than either MobileNet V3-small or EfficientNet-B0. It can be concluded that the additional gradient information capturing the images' high-frequency structural information brings greater benefit to a simpler 2D CNN's ability to handle classification problem.

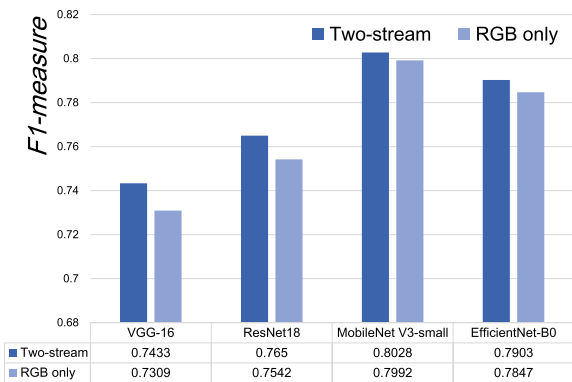


FIGURE 5. Comparison of our architecture equipped with different 2D CNNs (VGG-16, ResNet18, MobileNet V3-small, and EfficientNet-B0) using either an RGB-only scheme or our two-stream scheme.

2) ANALYSIS OF THE MULTI-REGION STRATEGY

We now consider the multi-region strategy. In this study, an input image is treated as a collection of sampling regions to encode locally non-uniform characteristics among categories, as suggested elsewhere [39]. In general, while increasing the sampling region size leads to better information, it increases the computational burden; therefore, we set the sampling region size to 64×64 experimentally, considering

the trade-off between accuracy and efficiency. To further demonstrate the advantage of our spatial pyramid random crop strategy, we compare it with other commonly used sampling strategies (random cropping, non-overlapping sampling, and overlapping sampling). We adopt random cropping to sample both $R = 1$ and $R = 21$ regions. Overlapping sampling uses a stride set to 1 in the experiments. Note that all the sampling region sizes in the experiment are set to 64×64 . Fig. 6 compares the performances of these strategies as implemented using the four different 2D CNNs of VGG-16, ResNet18, MobileNet V3-small and EfficientNet-B0. The proposed spatial pyramid random cropping strategy is superior to random cropping (with R also set to 21), indicating that our strategy can capture complete information and avoid missing important details. Our proposed strategy can achieve competitive results as compared with overlapping sampling, which (when using a stride of 1) is more computational expensive due to its numerous sampling regions.

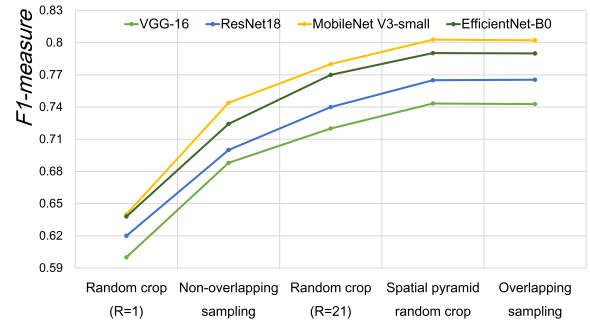


FIGURE 6. Comparison of spatial pyramid random cropping with other commonly used sampling strategies as implemented using different 2D CNNs.

Another important concern is the effectiveness of the proposed region fusion layer. This is assessed by comparing our region fusion layer with the simple average of multi-regions obtained by spatial pyramid random cropping. Fig. 7 compares the results for the different 2D CNNs of VGG-16,

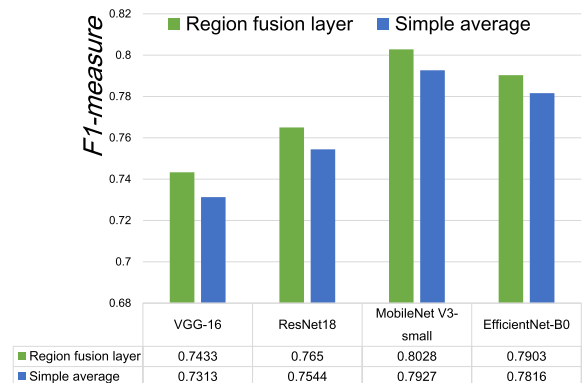


FIGURE 7. Comparative results from simple average of multi-regions and our region fusion layer as implemented with different 2D CNNs.

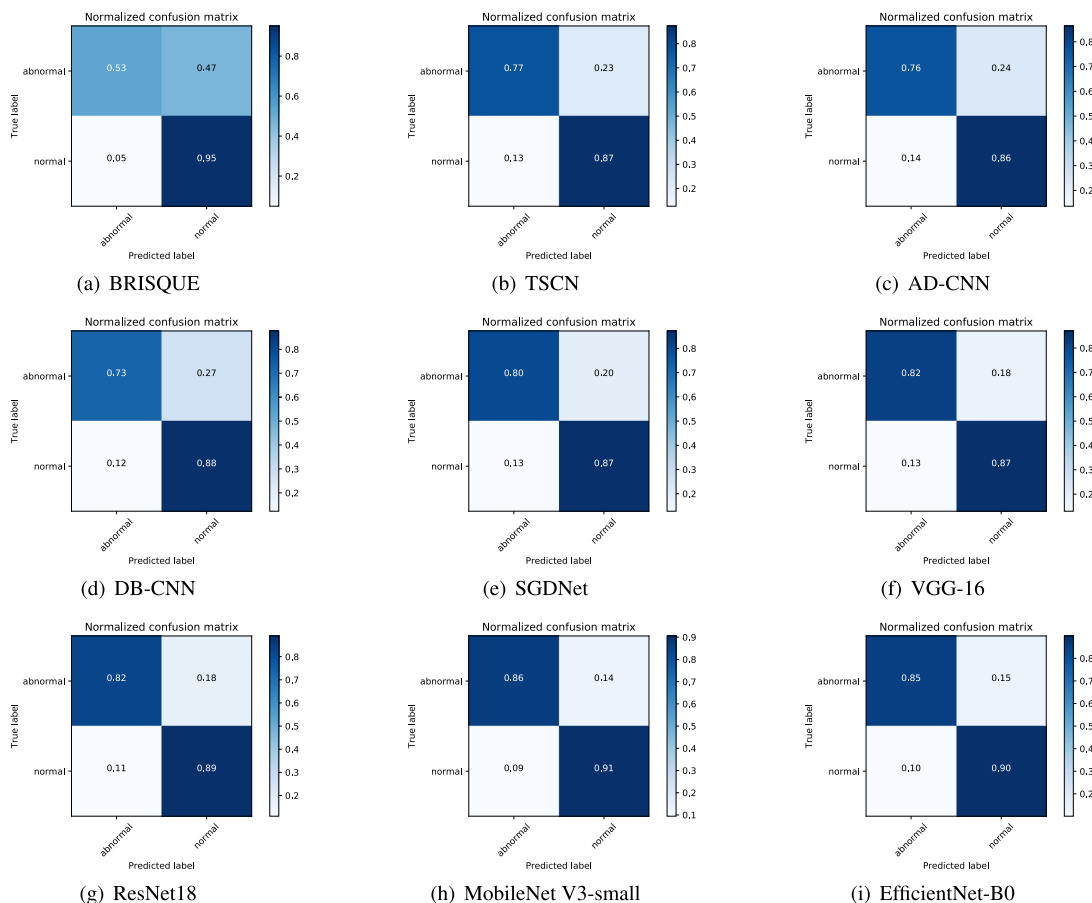


FIGURE 8. Confusion matrix results of our deep architecture based on different 2D CNN networks (VGG-16, ResNet18, MobileNet V3-small, and EfficientNet-B0) and other methods.

ResNet18, MobileNet V3-small and EfficientNet-B0. The *F1-measures* for our region fusion layer are consistently around 1% higher than those achieved using simple averaging, demonstrating that the proposed region fusion layer can adaptively learn the most discriminative regions hidden in the images. This result further verifies that the proposed region fusion layer is an attractive strategy when compared with the simple average adopted in TSCN.

C. COMPARISON AND ANALYSIS

1) COMPARISON WITH ESTABLISHED METHODS

To verify the performance of the proposed method, we compare our deep architecture based on the four different 2D CNNs with BRISQUE [45], accurate deep CNN (AD-CNN) [42], SGDNet [53], DB-CNN [52] and TSCN [39] conducted under the same experimental setup (*i.e.*, repeated five times under the fixed ratio of training to testing data). In DB-CNN, two branches of pre-trained CNNs are fine-tuned on our target databases with a variant of the stochastic gradient descent method. As for the auxiliary sub-task of saliency prediction in SGDNet, saliency map is obtained also by a teacher saliency model [57], which is also pre-trained on large-scale saliency datasets; then we train the network on our target

databases from scratch. It is worth noting that these image quality assessment networks output a 2-dimensional vector to achieve comparison, since the specified judgement task is essentially a classification problem. The hyper-parameters of BRISQUE, AD-CNN, SGDNet, DB-CNN and TSCN are the default values. Note a fixed threshold *T* should be needed to yield the category label of abnormal or normal, since BRISQUE outputs the quality score ranging from 0 to 100. In the following experiments, we set *T* to 42, which results in a good balance between *Precision* and *Recall*. Table 1 lists the comparative experimental results, and Fig. 8 gives the corresponding confusion matrix results. As expected, BRISQUE performs poorly in terms of *F1-measure*, as hand-crafted features are less suitable for the classification task than adaptively trained deep features. BRISQUE has low recall because it mis-classifies most abnormal images as normal, as shown in Fig. 8(a). Despite its good accuracy, the inferior recall contributes to a low *F1-measure*. TSCN performs better than AD-CNN, since it additionally considers a region-based fully convolutional layer for using the information of input image patches. As compared with our architecture, TSCN has inferior performance in terms of *Precision*, *Recall* and *F1-measure*, since the independent consideration of regions

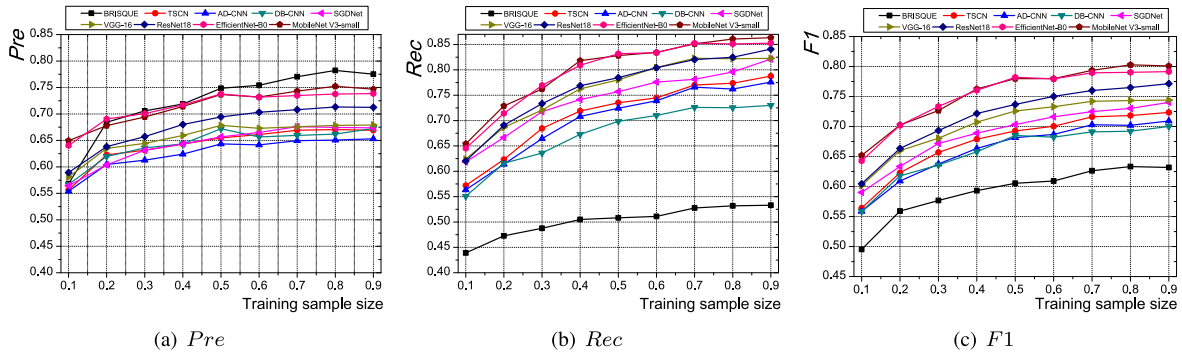


FIGURE 9. Comparison of our deep architecture based on different 2D CNNs (VGG-16, ResNet18, MobileNet V3-small, and EfficientNet-B0) and other methods with respect to the ratio of training to testing data.

TABLE 1. Comparison of our deep architecture based on different 2D CNNs with BRISQUE, TSCN, AD-CNN, DB-CNN, SGDNet.

Approaches	Pre	Rec	F1
BRISQUE	0.7823	0.5319	0.6332
TSCN	0.6705	0.7736	0.7184
AD-CNN	0.6507	0.7621	0.7020
DB-CNN	0.6619	0.7253	0.6922
SGDNet	0.6745	0.7959	0.7303
VGG-16 (ours)	0.6783	0.8220	0.7433
ResNet18 (ours)	0.7132	0.8249	0.7650
MobileNet V3-small (ours)	0.7523	0.8607	0.8028
EfficientNet-B0 (ours)	0.7376	0.8510	0.7903

brings a lot of confusion in assigning label for regions in training procedure. This result further verifies that our region fusion strategy is an attractive and promising strategy. It can be seen that SGDNet performs better than TSCN, indicating that saliency strategy locating discriminative patches is superior to uniform considering sampling patches in TSCN. The *F1-measure* of the proposed deep architecture based on MobileNet V3-small is around 7% and 9% higher than that of SGDNet and TSCN, respectively, and is the best performance among these baseline methods.

Fig. 10 visualizes some example images and classification results using BRISQUE, TSCN, AD-CNN, SGDNet, DB-CNN and our method based on MobileNet V3-small. It can be seen that the images in first row characterise local non-uniform changes. As a result, BRISQUE and DB-CNN, which considering whole image patterns and structures, performs inferior. Although SGDNet based on pre-trained saliency model can be regarded as a region based method, it captures saliency parts well but local changes. It indicates that the learned saliency information on the limited dataset is inferior to patch-based training strategy. As compared with TSCN considering patches uniformly, our method can achieve better classification results, since the region fusion layer ensures the scheme’s adaptability in learning discriminative regions hidden in the images. As compared with first row, the second row has uniform characteristics. Consequently, these comparative methods have good classification results. DB-CNN, pre-trained from large scale datasets

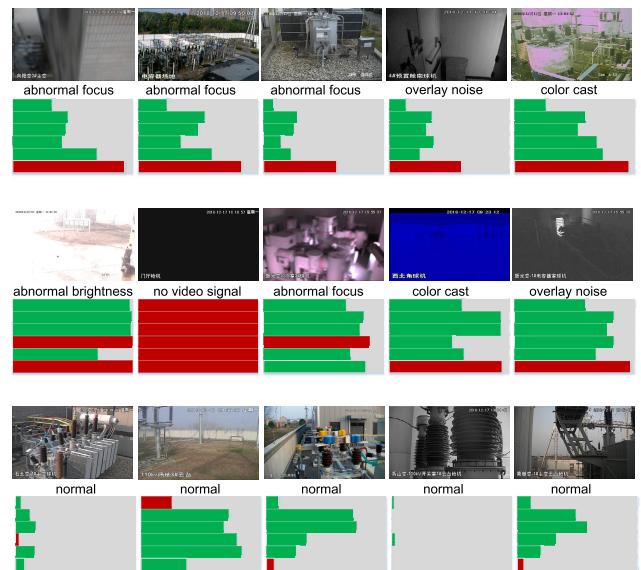


FIGURE 10. Sample example images and classification results using BRISQUE, TSCN, AD-CNN, SGDNet, DB-CNN and our method based on MobileNet V3-small (from top to bottom). (Note: The probability ranges from 0 to 1, and the marked green area denotes the probability of judging abnormal.).

with synthetic distortions, conducts good performance in terms of uniform overlay noise and abnormal focus. It is worth mentioning that DB-CNN learned from the auxiliary synthetic distortions cannot accurately identify the complex and unknown abnormal for the specified characteristics in dataset with limited training samples. As presented in Table 1, DB-CNN has low recall over the dataset. By contrast, our model is more universal. For the normal images classification, our method still can achieve good performance. Two factors must be credited for the proposed deep-learning architecture. The first is the two-stream scheme, which considers high-frequency information in the gradient image to enhance performance. The second originates from our multi-region strategy behavior. Specifically, spatial pyramid random cropping encodes locally non-uniform characteristics among categories, and the region fusion layer ensures the scheme’s adaptability in learning discriminative regions hidden in the

TABLE 2. Inference speeds (seconds) of various methods: our deep architecture based on different 2D CNNs, BRISQUE and TSCN.

	BRISQUE	TSCN	VGG-16 (ours)	ResNet18 (ours)	MobileNet V3-small (ours)	EfficientNet-B0 (ours)
Time	0.1731	1.4893	35.0871	5.1361	1.4516	1.6426

images and makes the learning procedure an end-to-end operation for a whole image.

To further demonstrate the advantages of the proposed deep architecture, we evaluate *precision*, *recall*, and *F1-measure* for the comparative methods described above in terms of different ratios of training to testing data. As presented in Fig. 9, the *precision*, *recall*, and *F1-measure* generally improve as the number of training samples grows. As expected, BRISQUE with a small training set also performs poorly. It can be seen that our deep architecture based on different 2D CNNs performs better in terms of *precision*, *recall*, and *F1-measure* than TSCN under different ratios of training to test data. Note that even if the shallower convolutional network in TSCN is replaced with our adopted 2D CNNs, our method is still better than TSCN as shown in Figs. 6 and 7, because of advantages in our region fusion strategy. Interestingly, MobileNet V3-small seems to be slightly more suitable to the data distribution in the field of power systems than EfficientNet-B0. In general, our deep architecture coupled with these high-level 2D CNNs achieves the best overall performance in terms of *precision*, *recall*, and *F1-measure*, showing its great potential in application to image quality judgement.

2) SPEED COMPARISON

We compare the runtimes of our deep architecture based on the different 2D CNNs with those of other state-of-the-art methods, and list the average execution times in Table 2. In order to achieve fair comparison, all the methods were implemented on a single core of a 2.2 GHz Intel Xeon CPU Silver 4114. As expected, BRISQUE using hand-crafted features has the fastest processing speed. It can be seen that our deep architecture based on MobileNet V3-small and EfficientNet-B0 can achieve competition level on execution time, as compared with TSCN which has much shallower network structure. However, the classification performances puts TSCN at a disadvantage compared with our well-established architectures. It is worth mentioning that these deep architectures are much faster when implemented on a GPU. For instance, our method based on ResNet18 implemented on GPU is nearly 90 times faster than CPU.

V. CONCLUSION

This paper presents a novel multi-region two-stream deep architecture for judging abnormalities in IMSs. The two-stream scheme explicitly considers the high-frequency structure patterns of gradient images, which provides information complementary to the color information from the original images. In contrast to previous works on sampling non-overlapping patches, here we consider spatial pyramid

random cropping to characterize images at different spatial pyramid levels. The method is demonstrated to be complete, robust, and with low computational cost. Instead of considering patches independently as in previous works, we propose a simple and effective region fusion strategy. Specifically, we reformulate the weighted term for regions and further consider it as a fully connected layer attempting to learn the fusion weight. Consequently, the proposed region fusion architecture as a layer is integrated into the whole learning pipeline. The region fusion strategy can adaptively find discriminative regions hidden in the image and provide end-to-end learning for the whole image. The experimental results show promising judgment performance by our method, which is capable of handling the effects of locally non-uniform changes, such as local noise and blur, to obtain high accuracy. Although the developed method is designed for systems monitoring electrical power scenes, it can be easily extended to other surveillance systems in a timely and efficient manner.

Our future research will extend our work in the following aspects. Our two-stream scheme uses gradient images, without considering other useful structure patterns or high-frequency information in other domains. Therefore, we will investigate and combine various other types of information into the network architecture, which should further improve the judgment accuracy. Furthermore, although off-line analysis is possible in IMSs, more effective network designs and other parallel computing techniques would be worth developing to further increase the judgment speed.

REFERENCES

- [1] R. Aguilar-Ponce, A. Kumar, J. L. TecpanecatI-Xihuitl, and M. Bayoumi, "A network of sensor-based framework for automated visual surveillance," *J. Netw. Comput. Appl.*, vol. 30, no. 3, pp. 1244–1271, Aug. 2007.
- [2] S. Mei, H. Yang, and Z. Yin, "An unsupervised-learning-based approach for automated defect inspection on textured surfaces," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 6, pp. 1266–1277, Jun. 2018.
- [3] N. S. M. Zamani, W. M. D. W. Zaki, A. B. Huddin, A. Hussain, H. A. Mutalib, and A. Ali, "Automated pterygium detection using deep neural network," *IEEE Access*, vol. 8, pp. 191659–191672, 2020.
- [4] Y. He, K. Song, Q. Meng, and Y. Yan, "An end-to-end steel surface defect detection approach via fusing multiple hierarchical features," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1493–1504, Apr. 2020.
- [5] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, "Deep residual shrinkage networks for fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4681–4690, Jul. 2020.
- [6] G. Xu, M. Liu, Z. Jiang, W. Shen, and C. Huang, "Online fault diagnosis method based on transfer convolutional neural networks," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 2, pp. 509–520, Feb. 2020.
- [7] J. Chen, Z. Liu, H. Wang, A. Nunez, and Z. Han, "Automatic defect detection of fasteners on the catenary support device using deep convolutional neural network," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 2, pp. 257–269, Feb. 2018.
- [8] D. Long, X. Wen, W. Zhang, and J. Wang, "Recurrent neural network based robust actuator and sensor fault estimation for satellite attitude control system," *IEEE Access*, vol. 8, pp. 183165–183174, 2020.
- [9] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

- [10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
- [11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.
- [12] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.
- [13] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.
- [14] B. Schölkopf, J. Platt, and T. Hofmann, *Greedy Layer-Wise Training of Deep Networks*. Mainz, Germany: MITP, 2007, pp. 153–160.
- [15] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, no. 2, 2012, pp. 1097–1105.
- [16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [19] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–14.
- [20] Y. He, J. Lin, Z. Liu, H. Wang, L. Li, and S. Han, "AMC: AutoML for model compression and acceleration on mobile devices," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 784–800.
- [21] T. Yang, A. Howard, B. Chen, X. Zhang, A. Go, M. Sandler, V. Sze, and H. Adam, "Netadapt: Platform-aware neural network adaptation for mobile applications," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 289–304.
- [22] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2016, *arXiv:1602.07360*. [Online]. Available: <http://arxiv.org/abs/1602.07360>
- [23] A. Gholami, K. Kwon, B. Wu, Z. Tai, X. Yue, P. Jin, S. Zhao, and K. Keutzer, "SqueezeNext: hardware-aware neural network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1638–1647.
- [24] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: <http://arxiv.org/abs/1704.04861>
- [25] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
- [26] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.
- [27] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "MnasNet: Platform-aware neural architecture search for mobile," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2820–2828.
- [28] H. Cai, L. Zhu, and S. Han, "Proxylessnas: Direct neural architecture search on target task and hardware," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–13.
- [29] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.
- [30] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 6105–6114.
- [31] A. Conrad Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, Sep. 2013.
- [32] G. Ciocca, S. Corchs, F. Gasparini, and R. Schettini, "How to assess image quality within a workflow chain: An overview," *Int. J. Digit. Libraries*, vol. 15, no. 1, pp. 1–25, Nov. 2014.
- [33] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.
- [34] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.
- [35] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1098–1105.
- [36] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1733–1740.
- [37] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.
- [38] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.
- [39] Q. Yan, D. Gong, and Y. Zhang, "Two-stream convolutional networks for blind image quality assessment," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2200–2211, May 2019.
- [40] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 1, 2014, pp. 1–11.
- [41] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.
- [42] B. Bare, K. Li, and B. Yan, "An accurate deep convolutional neural networks model for no-reference image quality assessment," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 1356–1361.
- [43] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. San Rafael, CA, USA: Morgan & Claypool, 2006.
- [44] P. Ye, "Feature learning and active learning for image quality assessment," Ph.D. dissertation, Dept. Elect. Comput. Eng., Univ. Maryland, College Park, MD, USA, 2014.
- [45] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
- [46] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.
- [47] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 50–63, Jan. 2015.
- [48] Q. Li, W. Lin, J. Xu, and Y. Fang, "Blind image quality assessment using statistical structural and luminance features," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2457–2469, Dec. 2016.
- [49] L. Kang, P. Ye, Y. Li, and D. Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2791–2795.
- [50] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "DiplQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.
- [51] K.-Y. Lin and G. Wang, "Hallucinated-IQA: No-reference image quality assessment via adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 732–741.
- [52] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020.
- [53] S. Yang, Q. Jiang, W. Lin, and Y. Wang, "SGDNet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1383–1391.
- [54] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 20–36.
- [55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [56] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
- [57] S. Yang, G. Lin, Q. Jiang, and W. Lin, "A dilated inception network for visual saliency prediction," *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 2163–2176, Aug. 2020.



JINRUI GAN received the Ph.D. degree in computer science and technology from Beijing Jiaotong University, Beijing, China, in 2019. He is currently an Engineer with Global Energy Interconnection Research Institute Company Ltd., Beijing. His research interests include computer vision, data mining, and machine learning.



PENG WU received the M.Sc. degree from the Beijing Institute of Technology, Beijing, China, in 2006. He is currently the Director of the AI Team, Global Energy Interconnection Research Institute Company Ltd., Beijing. His research interests include computer vision, machine learning, and pattern recognition.



WEI JIANG received the Ph.D. degree in engineering instrument science and technology from Zhejiang University, Hangzhou, China, in 2013. He is currently the Manager of the Department of Internet, State Grid Corporation of China, Beijing, China. His research interests include machine learning and artificial intelligence.



GUOLIANG ZHANG received the Ph.D. degree in information technology from the Beijing University of Technology, Beijing, China, in 2019. He is currently an Engineer with Global Energy Interconnection Research Institute Company Ltd., Beijing. His research interests include action recognition, machine learning, and man-machine interaction system for robots.



TING ZHAO received the M.Sc. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2004. She is currently a Professorate Senior Engineer with Global Energy Interconnection Research Institute Company Ltd., Beijing. Her research interests include computer vision and artificial intelligence.



ZIWEN ZHANG is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China. His current research interests include computer vision, pattern recognition, and image processing.

...