IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 7, 2022, accepted March 2, 2022, date of publication March 11, 2022, date of current version March 30, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3158952

A Novel Attention-Based Multi-Modal Modeling Technique on Mixed Type Data for Improving TFT-LCD Repair Process

YI LIU[1], **HSUEH-PING LU**[2], **AND CHING-HAO LAI**[1]
[1]Division of Digital Technology, Department of Data Science Analysis, AU Optronics Corporation, Taichung 40763, Taiwan
[2]Division of Digital Technology, AU Optronics Corporation, Hsinchu 40763, Taiwan

Corresponding author: Hsueh-Ping Lu (lu.hsueh.ping@gmail.com)

ABSTRACT In Thin-Film Transistor Liquid-Crystal Display (TFT-LCD) manufacturing, conducting a machine learning based system with multiple data types has become actively desired to solve complicated problems. This paper proposes a multi-modal learning approach: *TabVisionNet*, which is modeled by utilizing the information from both tabular data and image data. A novel attention mechanism called *Sequential Decision Attention* was integrated into the multi-modal modeling framework that improves the comprehension of the information from two modalities. This cross-modal attention mechanism can capture the complex relationship between modalities then gain better generalization and faster convergence in the training process. Conducting an experiment, the performance of our novel approach was significantly better than single-modal and other multi-modal learning approaches in our real case scenario.

INDEX TERMS Smart manufacturing, TFT-LCD, deep learning, multi-modality machine learning, tabular data, image data and attention mechanism.

I. INTRODUCTION

In thin-film transistor liquid-crystal display (TFT-LCD) industry, computer vision techniques have been widely used to help manufacturers monitor abnormalities, identify potential process bottlenecks, and swiftly respond to process problems to reduce yield loss. In the earlier years, automatic optical inspection (AOI) machines have achieved satisfactory functional defect identification [1], [2]. However, it is difficult to identify irregular defect such as Mura by using typical AOI [3]. Therefore, human intervention is required in Mura defect classification, and successful classification typically depends on the experience and skill of engineers.

After 2010, the application of artificial intelligence into many domains have been successful. Deep learning models are being gradually applied in the TFT-LCD industry for detecting and identifying panel defects [4]–[7]. The salvation of high-level products (e.g. 85 inch panel) with defects becomes possible with the inclusion of a repair station.

Due to the high cost of scrapping, the repair process has been widely applied [8]–[10] for high-level products, which prevents the wasting of expensive material and process time when defective panels are repaired successfully. Traditionally, all defective panels identified in the inspection step are sent to the repair station, but in the final test process, some repaired panels are still treated as defective. The repair process can be further improved by identifying which panels are not worth repairing. The information collected from previous manufacturing processes can be used to predict the repair status to gain competitive advantage. However, it is a challenge to develop a learning based repair status prediction system by using only image data because it is difficult to determine the status of repair without other critical information such as process recipes, defect information or the measured parameter values collected from the IoT sensors. The key for solving this complex problem is to leverage all useful information from collected data then model the different modalities by considering data with multiple types from different process steps.

In recent years, multi-model machine learning (MMML) techniques [11]–[14] with different combination of modalities are widely used in audio-visual speech recognition (AVSR), event detection, media description, multimedia retrieval and several medical applications [15]–[25].

The associate editor coordinating the review of this manuscript and approving it for publication was Chi-Tsun Cheng.

However, multi-modal approaches for tabular and image data are rare, and the state-of-the-art methods still suffer from several training issues such as over-parameterized and over-fitting. This sparked an opportunity to construct a better learning framework to model tabular and image data simultaneously, as these data types are commonly collected during TFT-LCD manufacturing processes.

Additionally, the repair process offers further opportunities to improve the manufacturing process by identifying which panels are not worth repairing. In this paper, a novel deep learning framework is proposed for the repair rate improvement by utilizing a MMML technique with the tabular and image data. The main contributions of this paper are shown as follows:

- A novel deep neural network architecture which can be trained with multi-modal data is proposed to improve the repair rate in the real TFT-LCD manufacturing cases. It is hard to find any other previous MMML or similar techniques which are designed or implemented for TFT-LCD applications.
- The proposed model, *TabVisionNet* extracts a fused feature map from the tabular and image information and can be simultaneously trained by an end-to-end training strategy.
- Moreover, a novel cross-modal attention mechanism called *Sequential Decision Attention* is proposed to enhance the information exchange between two modalities then gain more stability and better performance in the real multi-modal modeling tasks.

The remainder of this paper is organized as follows: Several related studies are described and discussed in Section II. Section III illustrates the proposed multi-modal modeling method. Section IV demonstrates the experimental results and analysis. Finally, Section V presents the conclusion of this paper.

II. RELATED WORKS

A. ATTENTION MECHANISM

The human perception process [26] can prove that the attention mechanism is a very important part in recognizing objects. The attention mechanism biases the allocation of the most informative feature expression while suppressing the less useful ones. Recently, many studies have been made towards applying attention techniques into deep neural networks to improve performances. The deep Boltzmann machine (DBM) [27] contains the top-down attention by its reconstruction process in the training stage. For the sequential decision processing task, attention mechanism has also been widely applied in various recurrent neural networks [26], [28]. The top information is gathered sequentially and decided to attend for the next feature learning steps. Vaswani *et al.* [29] proposed a notable multi-headed self-attention mechanism to extract rich semantic representation from input data. Moreover, several transformer-like approaches have been proposed and have achieved great successful performance in many natural language processing (NLP) tasks [30], [31]. The attention mechanism has also been widely applied to computer vision tasks. Itti *et al.*. [32] proposed a bottom-up visual attention computing model based on biological principles. Antoni *et al.*. [33] proposed an efficient de-noising method by replacing the color of a pixel with an average of the colors of similar pixels. Since the most similar pixels to a given pixel have no reason to be close at all, the kernel-based similarity function were used to measure the relationship of each couple pixels. M. Jaderberg *et al.*. [34] proposed a spatial attention mechanism to enhance rotational-invariance and scaling-invariance explicitly during training by using an additional learnable module. Different from the spatial attention, J. Hu *et al.*. [35] proposed a channel-wise attention architectural unit called Squeeze-and-Excitation (SE) block, which adaptively recalibrates the channel-wise feature responses by explicitly modeling interdependencies between channels. Based on this work, several approaches have been proposed to improve the channel-wised attention. Gao *et al.*. [36] modeled the channel interdependencies by using the second-order information from input feature maps. Wang *et al.* [37] proposed an efficient channel-wised attention module to reduce the computation and gain better performance by using different CNN architectures. The mixed approach combines different attention mechanisms into one framework, which can model a more complex relationship from the input features and gain better performance. Sanghyun W. *et al.*. [38] firstly combined the mechanisms of the spatial and channel-wise attention by a well-designed trainable module with two sequential sub-modules for the channel-wise and spatial attention. Different from sequential processes for two attention operations, Zhang *et al.*. [39] proposed an efficient building block approach to model the spatial and channel-wise attention in parallel, which adopts shuffle operation [40] to combine two types of attention mechanisms effectively. Q. Yang *et al.* [41] proposed a novel efficient attention pyramid network (EAPNet) for semantic segmentation task by combining channel-wise and spatial attention. The proposed EAPNet can captures multi-scale context and integrates the low-level feature maps and spatial location information, which is suitable for the semantic segmentation application and gains better performance than other attention mechanism in their experiment. Self-attention mechanisms [29] are also applied in many computer vision tasks to gain better performance [42]–[44]. Based on [29] and [33], Wang *et al.*. [45] introduced a general non-local operation framework for capturing long-range dependencies. This non-local building block can be plugged into many computer vision architectures, such as object detection/segmentation and pose estimation. Cao *et al.*. [46] analyzed [45] empirically and found that the global contexts modeled by the non-local network are almost the same for the different query positions. In order to solve this problem, they simplified the attention modules based on a query-independent formulation, which maintains the accuracy of non-local network with significantly less computation.

In this paper, we modified the Global Context (GC) block described in [46] to a multi-modal version by adding a lightweight, learnable sub-module. Our cross-modal GC block can capture the complex interdependencies between two modalities and gain increased performance in multi-modal learning tasks. The detailed descriptions of the proposed attention module are presented in the next section.

B. MULTI-MODAL MACHINE LEARNING

Since significant progress have been made in deep neural networks (DNNs) techniques, a growing number of applications via deep learning models have been proposed and shown notable success in computer vision [47]–[49] and natural language processing (NLP) [29]–[31] domains. However, many applications in the artificial intelligence field involve more than one input data type, such as audio-visual speech recognition (AVSR), image and video captioning, visual question answering (VQA), text-to-image generation and several medical applications. In order to solve these complex problems, the studies of multi-modal machine learning (MMML) techniques aim to build models that can process and relate the information from different modalities. The key challenges for constructing a multi-modal learning algorithm are 1) learning representations from different types of the input data simultaneously and 2) fusing the representations of different modalities efficiently [11], [13].

Representations: Good representations (or features) are important for the performance of machine learning models. Bengio *et al.*. [50] identified the properties for good representations such as smoothness, natural clustering and temporal/spatial coherence. Srivastava *et al.* [51], [52] identified the additional properties for the multi-modal representations such as the consistency and complementarity. The multimodal deep belief networks (DBNs) extracts information from the modalities and then combines them into a joint representation. Some researchers [15], [53], [54] proposed end-to-end approaches by using individual neural layers for each modality followed by the number of shared layers that project the modalities into a common space. The joint multi-modal representation can be directly passed through the subsequent process or prediction. Since the deep learning approaches require large amounts of labeled data, some approaches were proposed [55], [56] to encode the multiple input modalities and represent the extracted features into a common feature space by using an auto-encoder approach. The major advantage of the deep joint representation model is the ability to pre-train a model on unlabeled data but usually gain superior performance. However, it is difficult to use joint representation model when there are any modalities not present during the training or inference phase. The alternative way is to learn separate representations for each modality but coordinate them through a constraint [11]. Instead of projecting all modalities into a joint space, L. Wang *et al.* [57] used a novel large margin objective to coordinate two representations for image/text retrieval task. For unsupervised pre-training, Humam Alwassel *et al.* [58] proposed clustering-based self-supervised method to learn coordinated representation for the downstream task, which is the first approach that outperforms the large-scale fully-supervised pre-training for action recognition on the same model architecture.

Fusion strategies: Fusion is another key challenge of multi-modal learning algorithm, which aims to integrate the information from different modalities. Fusion strategies in the multi-modal machine learning field can be divided into different levels and different ways, corresponding to where the information is integrated, and how they are combined and projected into a single compact representation.

Zhou *et al.* introduced three types of fusion levels in medical image segmentation: the input-level, layer-level and decision-level fusion [12]. In the input-level fusion, multi-modal input data are fused by the channel concatenation, which directly integrates modalities in the original input space. Tseng *et al.* [59] proposed a join sequence learning and cross-modality approach for biomedical segmentation. The slices of different MRI images are combined together as the input data then feed forward to the proposed networks to learn the multi-modal representation. Kuniaki Noda *et al.* [60] combined multiple signal as the input data then constructed a deep neural networks model for robot behavior learning. In the layer-level fusion, each modality first pass through the individual layers to extract modality-specific features then be integrated in subsequent shared-weights layers to project into a joint space. The connections among the different fusion layers can capture complex relationships between the modalities. Jose Dolz *et al.* [61] proposed a hyper-densely connected architecture for the multi-modal image segmentation task and yielded significant improvement over state-of-the-art approaches. H. Prabowo *et al.* [25] proposed a novel GPA prediction algorithm by aggregating time series and tabular data. The information from two modalities was fused by adding the output of each layer in the LSTM branch to the output of MLP layer at the same level. In the decision-level fusion, which is similar to layer-level fusion, each modality is used as a single input of the single encoder to extract the modality-specific information, the outputs of each encoder will be integrated by a final output layer. The main advantage of decision-level fusion strategy comes from the ability to learn the complementary information from the different modalities independently since the latent information in different modalities can be diverse. The works of [21], [22], [62] developed multi-modal diagnosis algorithms by leveraging the representations from both structured and unstructured data.

Chao Zhang *et al.* introduced different types of fusion ways such as the simple operation-based and attention-based fusion [13]. The simple operation-based fusion ways most commonly integrates information from multiple modalities because they usually required few or no additional parameters. Some previous works [15], [59], [62] used the concatenation operation to combine modalities. An element-wise multiplication operation [54] was proposed to integrate

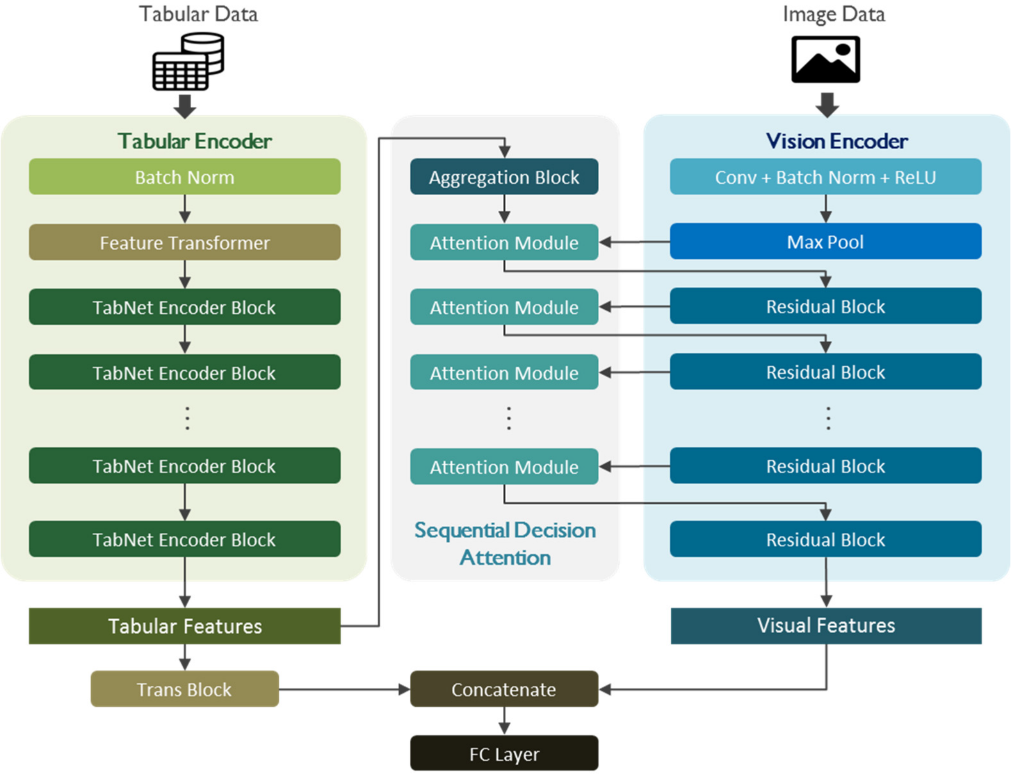


FIGURE 1. Overall architecture of proposed TabVisionNet.

vision and language information for the VQA application. Xitong Yang *et al.* [16] designed a temporal fusion model for the temporal multimodal learning (TML) task. The proposed model has the ability to learn the joint representation and temporal dependencies between modalities by adding a correlate fusion function during the model training. Attention mechanisms are widely applied for fusion of features, which aims to generate several suitable dynamic weights for a fusion operation. RNN-based models were widely used on attention mechanism. For VQA applications, the hidden states of a RNN-based model were used to find out the regions in an image that relate to the question or answer [63], [64]. Different from generating attention weights on one modality based on information extracted from the others, some studies [65], [66] used symmetric structures to generate attention weights on all modalities.

The proposed approach of this paper contains two fusion strategies, a layer-level fusion was designed with a novel cross-modal attention mechanism to capture complex relationship from two modalities. Additionally, a decision-level fusion with a concatenation operation was applied to project two modalities into a joint representation space. The detailed descriptions of the fusion strategies are presented in the next section.

III. PROPOSED APPROACH

A novel deep neural networks architecture is proposed to combine the information from two modalities based on a cross modal attention mechanism. Figure. 1 illustrates the overall architecture of the proposed approach. The architecture of the proposed approach contains three main components: 1) TabNet [67] encoder is used as the tabular encoder to extract features from the structure data, e.g. the defect log or parameters collected from the previous process; 2) Convolutional neural networks (CNNs) [68], [47], [48] is used as the vision encoder to extract the visual features from the defective images; 3) Finally, a novel attention mechanism called *Sequential Decision Attention* is integrated to control the vision encoder to focus on the suitable visual features by the given global context information extracted from both the image and tabular data. The two encoders are jointly and simultaneously trained with the two modal data. The details of the descriptions about the proposed approach are presented in the following subsections.

A. TABULAR ENCODER

First, TabNet, used as a tabular encoder, extracts features from the tabular data, which had shown competitive performance compared with the state-of-the-art models in several tabular datasets [69]–[72]. The main building blocks in the TabNet encoder are: 1) the Feature Transformer for the feature transformation task; 2) the Attentive Transformer for the soft-mask generation task. In each decision step, a soft-mask is generated by the attentive transformer for the salient feature selection, similar to the decision trees (DTs) [73]. Then, the feature transformer extracts sematic information from the selected features and divides the processed representations into the subsequent decision step to be used by the attentive transformer, as well as the outputs of this step. In the initial step, the input data is fed to a batch normalization [74] and a feature transformer to extract the initial tabular features. The forward computation can be expressed as:

$$F_{tabular}^{0} = FeatsTransformer\left(BN\left(X_{tabular}\right)\right) \quad (1)$$

where *FeatsTransformer* is the feature transformer module that projects the input features into a representation space, *BN* is the batch normalization layer and $X_{tabular}$ is the input tabular data. Next, the main encoding computations can be expressed as:

$$F_{tabular}^{d}{=}TabNetEncoderBlock^{d}\left(F_{tabular}^{d-1}\right),\ \ \mathrm{d}=1,2\ldots,K \quad (2)$$

where*TabNetEncoderBlock* is the TabNet encoder block that contains both feature transformer and attentive transformer, $F_{tabular}^{d}$ is the extracted tabular features in the decision step d.

B. VISION ENCODER

The vision encoder can be any canonical convolution neural network architectures with several stages. In the end of each stage, down-sampling operations are applied to reduce the spatial dimensions of the image, either mathematical or learnable operations. The fully-connected layers at the end of the CNN architecture are removed, because only the visual features are needed for the subsequent fusion. The forward computation can be expressed as follows:

$$F_{visual}^{s+1} = CNNStageBlock^{s+1}\left(F_{visual}^{s}\right), \quad \mathrm{s}=0,1,2,\ldots,\mathrm{K} \quad (3)$$

where F_{visual}^{s} is the visual features in stages, and $CNNStageBlock^{s}$ is a set of layers in stages. The original input image F_{visual}^{0} is sent to the first stage block $CNNStageBlock^{1}$ to extract the features F_{visual}^{1}. And then, the extracted features can be the input of the subsequent stage block to get the next-stage features.

C. SEQUENTIAL DECISION ATTENTION

In order to integrate the tabular features $F_{tabular}$ and the visual features F_{visual}, an efficient fusion method is designed by combining two fusion strategies, based on the attention mechanism which is called *Sequential Decision Attention*. A novel cross-modal attention mechanism for the layer-level fusion is applied in each CNN stage for modeling the relationship between two modalities. There are two main components in layer-level fusion: 1) The Decision Aggregation Block is used to stabilize the feature distribution from the TabNet encoder. 2) The cross-modal GC block is used to fuse the global information from two modal data. Finally, a decision-level fusion is applied to be the final fusion module. The outputs of the final fusion module is connected to the task-specified head.

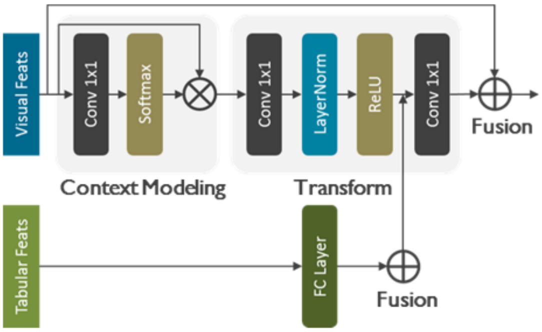


FIGURE 2. Architecture of our proposed attention module.

1) DECISION AGGREGATION

An aggregation module is proposed to integrate all tabular features extracted from different decision steps of the TabNet encoder. The module contains a transformation block for the tabular features in each decision step and a concatenate operation. The forward computation is expressed as:

$$R^{d} = TransBlock_{agg}^{d}\left(F_{tabular}^{d}\right),\ d{=}1,2,\ldots,\mathrm{K} \quad (4)$$

$$Agg_{reprs} = Cat\left(R^{1}, R^{2}, \ldots, \mathrm{K}\right) \quad (5)$$

where $F_{tabular}^{d}$ is the extracted tabular features from (1), and $TransBlock_{agg}^{d}$ is the transformation block for converting $F_{tabular}^{d}$ to the decision representation in step d, denotes as R^{d}. The transformation block contains a full-connected layer, batch normalization layer and Swish [75] activation. Finally, Agg_{reprs} is the aggregated decision representation for the subsequent attention module and $Cat\left(.\right)$ is the concatenate operation

2) CROSS-MODAL GLOBAL CONTEXT BLOCK

Following the Global Context modeling framework for capturing the long-range dependency of input feature maps, which is described in [46], denoted as:

$$z_i = f\left(x_j, \delta\left(\sum\nolimits_{j=1}^{N_p} \alpha_j x\right)\right) \quad (6)$$

where a) $\sum_j \alpha_j x_j$ denotes the visual **context modeling** module which groups the visual features of all positions together to obtain the global context information; b) $\delta\left(\cdot\right)$ denotes the feature **transform** to model the channel-wise dependency; c) f denotes the **fusion** function to aggregate the cross-modal context information to the visual features of each position.

Figure 2 illustrates the proposed cross-modal GC block for multi-modal learning, the forward computation can be expressed as follows:

$$z_i = f\left(F_{visual,i}, \delta\left(\sum\nolimits_{j=1}^{N_p} \alpha_j F_{visual,j} + g\left(Agg_{reprs}\right)_i\right)\right) \quad (7)$$

where a) $\alpha_j = \frac{e^{W_k F_{visual,j}}}{\sum_m e^{W_k F_{visual,m}}}$ is the weights for global attention pooling, and $g\left(\cdot\right)$ denotes the linear projection for changing the dimension of Agg_{reprs} to the channel dimension of F_{visual}. Then an element-wise summation is applied to fuse the context information from two modalities. b) The same as [46], $\delta(\cdot) = W_{v1}ReLu(LN(W_{v1}(\cdot)))$ denotes the bottleneck transform. c) $f(\cdot, \cdot)$ denoted the broadcast element-wise summation for fusion.

The traditional GC blocks intrinsically produce dynamic weights based on the input feature maps, which can be regarded as a self-attention mechanism. The proposed attention mechanism generates dynamic weights depending on cross-modal global context information, which can be regarded as a cross-modality attention mechanism, then the model can learn more complex interdependencies between two modalities.

3) FINAL FUSION

A decision-level fusion strategy is proposed to combine features extracted from two modalities. The proposed approach first sums tabular features over each decision step with ReLU activation to obtain the aggregated decision contribution of tabular features, denote as:

$$D_{contrib} = \sum_{d=1}^{N_{steps}} ReLU(F_{tabular}^{d}) \tag{8}$$

where N_{steps} is the number of decision steps and D_{out} is the aggregated decision contribution of the TabNet encoder.

After the feature combination, a non-linear transform is applied to ensure the output distribution is not skewed for the stable training, denote as:

$$D_{out} = TransBlock_{final}(D_{contrib}) \tag{9}$$

where D_{out} is the final decision features from tabular data and $TransBlock_{final}$ is the non-linear transformation which contains fully-connected (FC) layer, batch normalization layer and Swish activation.

For visual features, the global average pooling (GAP) [76] is used to squeeze the 2D feature maps to a 1D vector. The operation can be formulized by:

$$v_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{visual}^{S}(i,j) \tag{10}$$

where H and W is the height and width of the final feature maps F_{visual}^{S}, and v_c is the global statistic of the channel c.

Finally, a concatenate operation is used to integrate two modalities and apply a linear mapping as the task-specified head. The forward computations can be expressed as follows:

$$F_{fusion} = Cat(D_{out}, V_{out}) \tag{11}$$

$$Output = W_{out}F_{fusion} \tag{12}$$

where D_{out} is the final decision features extracted from the tabular data, $V_{out} = [v_1, v_2, \ldots, \text{K}, v_c]$ is the final visual features and, $Cat(.)$ is the concatenate operation and W_{out} is the weights of the task-specified layer.

IV. CASE STUDY

The proposed approach was used in the Polyimide (PI) process which is a sub-process of the cell process in TFT-LCD manufacturing. An additional inspection by using AOI machines was applied for measuring the quality of processed panels after the PI process. All defective panels are sent to the repair station and will be re-inspected by another AOI machine in the final cell test after a few days. Our goal is to conduct a learning algorithm to identify the repair status before the repair process. If the repair process is predicted to be unsuccessful, the process can be preemptively stopped, then the wasting of material and the process time can be reduced. The detailed description are presented in the following subsection.

A. DATA DESCRIPTION

The experimental data that contains tabular data and the defective images was collected from the online PI process and the inspection station. The tabular data contains the information from the manufacturing and repair process such as chamber temperature and the defect measured size. The defective images are taken at the inspection station. Both tabular data and image data are collected automatically by processing machines and IOT sensors.

Figure 3 illustrates the collected data in the experiment. A total of 10726 paired samples mixed with different product and defect types, and only 4% samples are unsuccessful repairs. Each paired sample consist of one record of a table and one defective image, denoted as:

$$X^{i} = \left(X_{tabular}^{i}, X_{visual}^{i}\right), \quad i = 1, 2, \ldots, K$$

where $i \in \mathbb{N}$ are the sample indices.

The collected data has been modified in order to avoid exposing company secrets. However, to ensure the repair status is clearly identified, the defect area of defective image has been left unmodified.

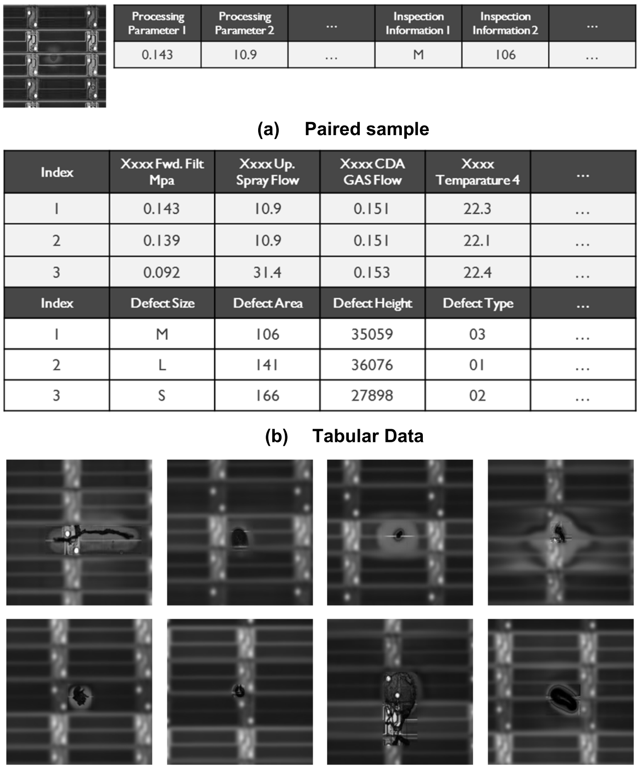
B. EXPERIMENTAL SETTING

In the training process, we implemented deep learning experiments with two NVIDIA GV100 GPUs. The primary deep learning framework is PyTorch 1.6, and the software environment was created by Anaconda with Python runtime.

In order to deal with the imbalance data, stratified sampling [77] was applied to split training, validation and testing data, which can ensure the same positive rate in each split set. 80% of the data were used for training and cross-validation, and the remaining 20% were used for testing. A 5-fold cross-validation was applied for hyper-parameter tuning. The input size of the images was 256 × 256 × 3 and the number of feature columns of tabular data was 148.

1) BASELINE APPROACHES

In our experiment, four different baseline models were trained for comparison. For single modal approaches, SVM and XGBoost [69] were trained with the tabular data $X_{tabular}$, and



Processing Parameter 1	Processing Parameter 2	...	Inspection Information 1	Inspection Information 2	...
0.143	10.9	...	M	106	...

(a) Paired sample

Index	Xxxx Fwd. Filt Mpa	Xxxx Up. Spray Flow	Xxxx CDA GAS Flow	Xxxx Temparature 4	...
1	0.143	10.9	0.151	22.3	...
2	0.139	10.9	0.151	22.1	...
3	0.092	31.4	0.153	22.4	...
Index	**Defect Size**	**Defect Area**	**Defect Height**	**Defect Type**	**...**
1	M	106	35059	03	...
2	L	141	36076	01	...
3	S	166	27898	02	...

(b) Tabular Data

(c) Defective Images

FIGURE 3. (a) Paired sample, contains one image which is taken from camera before the repair process, and one record of the tabular data, which is related to the information of product, defect and measured process parameters. (b) Samples of tabular data collected from processing machines and IOT sensors are shown in the above table with two sections. The upper section are process parameters of the PI cleaner, such as the pressure of the forward-direction filter and the chamber temperature; the lower section are the measured values of the AOI machines such as the defect size and the measured height (thickness). (c) Defective images took before repair process.

a canonical convolutional neural network (CNN) was trained with the image data X_{visual}. For multi-modal modeling, a common alternative approach was trained by both the tabular data and image data. For all experiments with deep learning techniques, the training batch size was set to 512, exponential scheduler was applied to reduce learning rate in every epoch, and the optimal early stopping epoch was determined by cross validation.

First, SVM and XGBoost were trained with tabular data. One-hot encoding was used for extracting the categorical features, and the grid search technique with cross validation was applied to find the best hyper-parameters.

Next, CNN model was trained for image data. ResNet 18 [37] backbone was used as the feature extractor and the final fully-connected (FC) layer's dimension was altered to satisfy our case. In the training process, Adam optimizer [78] was used with the initial learning rate=1e-3 and the decay rate=1e-6 to avoid over-fitting.

Finally, an alternative approach denoted as ***MLP Fusion***, which has been used in other application domains [20]–[22], was trained with two modalities for the multi-modal modeling baseline. For the vision encoder, ResNet 18 was used as the backbone and the global average pooling (GAP) was applied to obtain a 1D vector for subsequent fusion. Embedding encoding [79] was applied for categorical features to stabilize the training process and gain better performance. In the training process, Adam optimizer was used with the initial learning rate=5e-4 and the decay rate=1e-6 and the dropout rate=0.3 to obtain the generalized results.

TABLE 1. Architecture setting for tabnet encoder.

Parameter	Value	Description
d_{reprs}	16	Dimension of the decision representations.
d_{atten}	16	Dimension of the attention representation.
N_{step}	4	Number of the decision steps.
N_{indep}	2	Number of step-specified GLU Block.
N_{shared}	1	Number of shared fc layers cross all steps.
γ	1.3	The relaxation parameter.

2) PROPOSED APPROACH

For our proposed approach, the TabNet encoder architecture setting is illustrated in Table 1 with hyper-parameters tuned to follow the experimental setting. ResNet 18 was used to be the CNN backbone with Adam optimizer and the embedding encoding was applied to extract the categorical features, learning rates were set to 3e-3 and 3e-4 for the tabular encoder and vision encoder respectively. The loss function to be optimized by back-propagation algorithm can be formulized as:

$$L_{total} = L_{cls} + \lambda \cdot L_{sparse} \tag{13}$$

where L_{cls} is the cross entropy loss for the binary classification task, denote as:

$$L_{cls} = -\frac{1}{N}\sum\nolimits_{i}^{N} y^i \log \hat{y}^i + (1 - y^i)\log(1 - \hat{y}^i) \tag{14}$$

where $y^i$and $\hat{y}^i$are the ground-truth label and the output probability from the model respectively, and N is the sample size of data.

L_{sparse} is the regularization penalty to control the sparsity of the selected features in TabNet encoder [60], denote as:

$$L_{sparse} = \sum\nolimits_{i=1}^{N_{step}} \sum\nolimits_{b=1}^{B} \sum\nolimits_{j=1}^{D} \frac{-M_{b,j}[i]\log(M_{b,j}[i] + \epsilon)}{N_{step} \cdot B} \tag{15}$$

where ϵ is a small number for numerical stability, N_{step} is number of decision step of TabNet encoder, B and D are batch size and the feature dimension of the tabular data respectively. Coefficient λ is a hyper-parameter of the penalty weight.

C. EXPERIMENTAL RESULTS

Different random seeds were used for splitting data and random initialization of the model weights in the 15 runs of our experiments. Area under the receiver operating

TABLE 2. Quantitative evaluation results.

Method	AUC (Mean)	AUC (Std.)
SVM	0.729041	0.012653
XGBoost	0.738567	0.041910
CNN	0.752486	0.016006
MLP Fusion	0.791517	0.014722
TabVisionNet	***0.819157**	0.017419

TABLE 3. The statistical test results.

Method	T-Test (p-value)	KS-Test (p-value)
SVM	***2.209e-15	***1.005e-07
XGBoost	***3.341e-07	***8.771e-07
CNN	***1.061e-10	***1.888e-05
MLP Fusion	***3.390e-04	**4.7150e-02

*: P < 0.05 (significant), **: P < 0.01 (highly significant), ***: P < 0.001 (extremely significant)

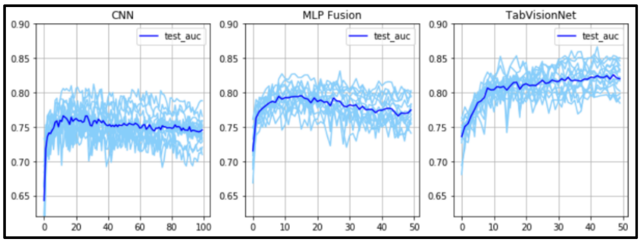
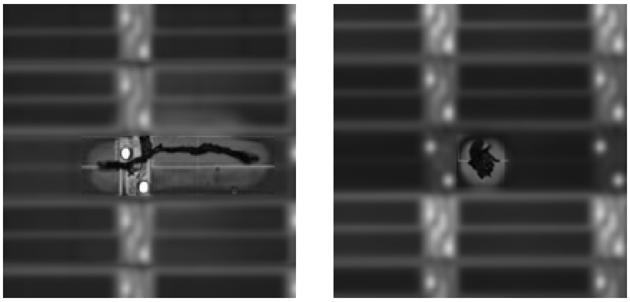


FIGURE 4. Learning curves of different approaches on testing set.

characteristic curve (ROC-AUC) was used as the evaluation metric. Table 2 shows the proposed approach outperforms the other three methods with regard to the AUC score. Furthermore, statistical tests were conducted in order to evaluate the reliability of the experiment. Table 3 shows the statistical evaluation of the XGBoost, CNN, MLP Fusion models and the proposed model. These comparison results show that the proposed model significantly gains more statistical improvement than the other methods.

Figure 4 shows the learning curves of three deep learning approaches in the experiment (with ResNet 18 backbone) on the test data set. This result shows that the proposed approach can gain more robustness than the others. For single-modal learning, both of the two models performed badly with the test set because some critical information is contained in the different data sources. Figure 5 shows that the panel contains a large defective area but it had passed the final test inspection.

The panel with the small defect area is still treated as defective because the defect is too thick so that the repair process requires more operations, and it also increases the fail rate. Additionally, the information of defect thickness is recorded in tabular data. The experimental results show that it is reasonable to combine the information of the tabular data and the image data. In multi-modal learning, we investigated the MLP Fusion-like approaches which are the main network architectures trained with the tabular data and the image data simultaneously. The MLP Fusion model had superior performance than the single-modal methods by leveraging more meaningful information from two modalities. However, the large amount of trainable weights led the model to be over-fitted on the training set so regularization techniques is required to be applied carefully during the model training.



(a) Repair successful

(b) Repair failed

FIGURE 5. The left (a) panel has larger defect area but pass the final test inspection, the right panel (b) has smaller defect area but still be identified as defective, which means the repair process failed to fix the defect.

The architecture of the proposed model is a bi-branch structure, the TabNet and CNN encoder are used to extract the information from the tabular data and the image data respectively. Additionally, the cross-modal attention mechanism is applied to integrate the information from two modalities efficiently, which stabilizes the training process and gains more generalized results than the multi-modal approach, MLP Fusion. The experimental results indicated that the proposed approach can solve complex problems that were difficult to solve in the past. Due to past works only considered single-modal data while our approach utilizes both tabular data and the image data at the same time, gaining better performances and feature extraction benefits on in the real TFT-LCD manufacturing cases.

D. EXTEND TO MEDICAL APPLICATIONS

In order to evaluate the generalization ability to other applications, another experiment was implemented on the medical open dataset: SIIM-ISIC Melanoma Classification challenge on Kaggle [80]. In this competition, the competitors need to identify melanoma by using images and patient-level contextual information.

In the experimental setting, we only used the training set of the competition and stratified sampling was applied for imbalance data. 80% of the data were used for training and cross-validation, remaining 20% of the data were used for testing. A 5-fold cross-validation was applied for hyper-parameter tuning. The input size of the images was 256 × 256 × 3 and the number of feature columns of tabular data was 5. Four different models were trained for the comparison. For single modal learning, XGBoost was trained

TABLE 4. The results of siim-isic dataset.

Method	Backbone	AUC
CNN	*ResNet*18	0.9201
MLP Fusion	*ResNet*18	0.9282
TabVisionNet	*ResNet*18	***0.9474**
CNN	*ResNet*34	0.9062
MLP Fusion	*ResNet*34	0.9479
TabVisionNet	*ResNet*34	***0.9639**
XGBoost	*NA*	0.8510

with only tabular data and the CNN models were trained with only image data. For multi-modal modeling, MLP Fusion and proposed *TabVisionNet* were trained with image and tabular data.

Table 4 illustrates the comparison results of models trained from scratch with the same dataset. The experimental results shows that the proposed approach outperforms the other models, which indicates that the proposed *TabVisionNet* has better learning ability without prior information, so that it is more suitable to be used in the scenarios with less related or labeled data. This scenario is very common in many domains such as the manufacturing or medical applications.

V. CONCLUSION

Deep learning techniques are gradually being developed in the TFT-LCD industry for solving complex problems in various manufacturing scenarios. However, conducting a learning based system by using only one modality is challenging because the discriminated information may be from multiple data sources. To leverage all useful information from collected data more efficiently, a novel deep multi-modal learning approach based on a well-designed attention mechanism was introduced, which utilizes the information from both tabular data and image data. The proposed *Sequential Decision Attention* captures the complex relationships between two modalities by a composite modeling strategy, which produces better joint representations for the multi-modal learning task. The experiment results proved the effectiveness of the proposed approach in the real-world multi-modal task for TFT-LCD repair process. Moreover, the proposed approach can also be applied in other domain such as medical applications and get competitive performances. In future studies, the proposed method can be extended to other combination of different modalities, such as time-series or text data, which are the common data sources in manufacturing scenarios.

REFERENCES

[1] J. Z. Tsai, R.-S. Chang, and T.-Y. Li, "Detection of gap Mura in TFT LCDs by the interference pattern and image sensing method," *IEEE Trans. Instrum. Meas.*, vol. 62, no. 11, pp. 3087–3092, Nov. 2013.

[2] A. Y. Jazi, J. J. Liu, and H. Lee, "Automatic inspection of TFT-LCD glass substrates using optimized support vector machines," in *Proc. 8th IFAC Symp. ADCHM, Int. Fed. Autom. Control*, Singapore, Jul. 2012.

[3] S. Mei, H. Yang, and Z. Yin, "Unsupervised-learning-based feature-level fusion method for Mura defect recognition," *IEEE Trans. Semicond. Manuf.*, vol. 30, no. 1, pp. 105–113, Feb. 2017.

[4] L. F. Chen, C. T. Su, and M. H. Chen, "A neural-network approach for defect recognition in TFT-LCD photolithography process," *IEEE Trans. Electron. Packag. Manuf.*, vol. 32, no. 1, pp. 1–8, Jan. 2009.

[5] T. Nakazawa and D. V. Kulkarni, "Wafer map defect pattern classification and image retrieval using convolutional neural network," *IEEE Trans. Semicond. Manuf.*, vol. 31, no. 2, pp. 309–314, May 2018.

[6] H.-P. Lu, C.-T. Su, S.-Y. Yang, and Y.-P. Lin, "Combination of convolutional and generative adversarial networks for defect image demoiréing of thin-film transistor liquid-crystal display image," *IEEE Trans. Semicond. Manuf.*, vol. 33, no. 3, pp. 413–423, Aug. 2020, doi: 10.1109/TSM.2020.3005164.

[7] Y. Wang, L. Ma, M. Jiu, and H. Jiang, "Detection of conductive particles in TFT-LCD circuit using generative adversarial networks," *IEEE Access*, vol. 8, pp. 101338–101350, 2020, doi: 10.1109/ACCESS.2020.2997807.

[8] L. W. Chul, S. J. Bong, K. B. Yeol, P. S. Hyuk, L. S. Muk, and L. W. Jeong, "Auto defect repair algorithm for lcd panel review & repair machine," in *Proc. SICE Annu. Conf.*, Aug. 2008, pp. 2200–2203, doi: 10.1109/SICE.2008.4655029.

[9] H. Honoki, N. Nakasu, T. Arai, K. Yoshimura, and T. Edamura, "In-line automatic defect inspection and repair method for a high yield TFT array production," *ECS Trans.*, vol. 8, no. 1, pp. 267–272, Jul. 2007, doi: 10.1149/1.2767319.

[10] L. Peng, L. Tingting, F. Meng, L. Jian, L. Pingfu, K. Heewoong, and L. Yanchun, "P-70: Improving defect repair rate in automatic repair process of color filter manufacturing," in *Proc. SID Symp. Dig. Tech. Papers*, vol. 49, 2018, pp. 1452–1455, doi: 10.1002/sdtp.12185.

[11] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

[12] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vols. 3–4, Sep. 2019, Art. no. 100004.

[13] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 478–493, Mar. 2020, doi: 10.1109/JSTSP.2020.2987728.

[14] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019, doi: 10.1109/ACCESS.2019.2916887.

[15] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 2130–2134, doi: 10.1109/ICASSP.2015.7178347.

[16] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. A. Bernal, and J. Luo, "Deep multimodal representation learning from temporal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5066–5074, doi: 10.1109/CVPR.2017.538.

[17] Z. Z. Lan, L. Bao, S. I. Yu, W. Liu, and A. G. Hauptmann, "Multimedia classification and event detection using double fusion," *Multimedia Tools Appl.*, vol. 71, pp. 333–347, 2014.

[18] Y.-R. Yeh, T.-C. Lin, Y.-Y. Chung, and Y.-C.-F. Wang, "A novel multiple kernel learning framework for heterogeneous feature fusion and variable selection," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 563–574, Jun. 2012.

[19] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, 2010, doi: 10.1007/s00530-010-0182-0.

[20] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *NPJ Digit. Med.*, vol. 3, no. 1, p. 136, Dec. 2020, doi: 10.1038/s41746-020-00341-z.

[21] T. Xu, H. Zhang, X. Huang, S. Zhang, and D. N. Metaxas, "Multimodal deep learning for cervical dysplasia diagnosis," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Lecture Notes in Computer Science), vol. 9901, S. Ourselin, L. Joskowicz, M. Sabuncu, G. Unal, and W. Wells, Eds. Cham, Switzerland: Springer, 2016, pp. 115–123, doi: 10.1007/978-3-319-46723-8_14.

[22] Z. Lan, L. Bao, S. I. Yu, W. Liu, and A. G. Hauptmann, "Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data," *Neurocomputing*, vol. 412, pp. 197–215, Oct. 2020, doi: 10.1016/j.neucom.2020.05.087.

[23] M. Artzi, E. Redmard, O. Tzemach, J. Zeltser, O. Gropper, J. Roth, B. Shofty, D. A. Kozyrev, S. Constantini, and L. Ben-Sira, "Classification of pediatric posterior fossa tumors using convolutional neural network and tabular data," *IEEE Access*, vol. 9, pp. 91966–91973, 2021, doi: 10.1109/ACCESS.2021.3085771.

[24] P. Tupe-Waghmare, P. Malpure, K. Kotecha, M. Beniwal, V. Santosh, J. Saini, and M. Ingalhalikar, "Comprehensive genomic subtyping of glioma using semi-supervised multi-task deep learning on multimodal MRI," *IEEE Access*, vol. 9, pp. 167900–167910, 2021, doi: 10.1109/ACCESS.2021.3136293.

[25] H. Prabowo, A. A. Hidayat, T. W. Cenggoro, R. Rahutomo, K. Purwandari, and B. Pardamean, "Aggregating time series and tabular data in deep learning model for university students' GPA prediction," *IEEE Access*, vol. 9, pp. 87370–87377, 2021, doi: 10.1109/ACCESS.2021.3088152.

[26] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2204–2212.

[27] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order Boltzmann machine," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2010, pp. 1243–1251.

[28] J. H. Kim, S. W. Lee, D. Kwak, M. O. Heo, J. Kim, J. W. Ha, and B. T. Zhang, "Multimodal residual learning for visual Qa," in *Adv. Neural Inf. Process. Syst.*, 2016, pp. 361–369.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[31] T. B. Brown, B. Mann, and N. Ryder, "Language models are few-shot learners," 2020, *arXiv:2005.14165*.

[32] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[33] A. Buades, B. Coll, and J.-M. Morel, "Non-local means denoising," *Image Process. On Line*, vol. 1, pp. 208–212, Mar. 2011, doi: 10.5201/ipol.2011.bcm_nlm.

[34] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[36] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3024–3033.

[37] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539, doi: 10.1109/CVPR42600.2020.01155.

[38] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19.

[39] Q.-L. Zhang and Y.-B. Yang, "SA-net: Shuffle attention for deep convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 2235–2239, doi: 10.1109/ICASSP39728.2021.9414568.

[40] 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, doi: 10.1109/CVPR.2018.00716.

[41] Y. Pei, L. Mu, Y. Fu, K. He, H. Li, S. Guo, X. Liu, M. Li, H. Zhang, and X. Li, "Colorectal tumor segmentation of CT scans based on a convolutional neural network with an attention mechanism," *IEEE Access*, vol. 8, pp. 64131–64138, 2020, doi: 10.1109/ACCESS.2020.2982543.

[42] Q. Yang, T. Ku, and K. Hu, "Efficient attention pyramid network for semantic segmentation," *IEEE Access*, vol. 9, pp. 18867–18875, 2021, doi: 10.1109/ACCESS.2021.3053316.

[43] W. Gaihua, Z. Tianlun, D. Yingying, L. Jinheng, and C. Lei, "A serial-parallel self-attention network joint with multi-scale dilated convolution," *IEEE Access*, vol. 9, pp. 71909–71919, 2021, doi: 10.1109/ACCESS.2021.3079243.

[44] X. Wang, Z. Cao, R. Wang, Z. Liu, and X. Zhu, "Improving human pose estimation with self-attention generative adversarial networks," *IEEE Access*, vol. 7, pp. 119668–119680, 2019, doi: 10.1109/ACCESS.2019.2936709.

[45] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local Neural Networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803, doi: 10.1109/CVPR.2018.00813.

[46] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Global context networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 24, 2020, doi: 10.1109/TPAMI.2020.3047209.

[47] 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, doi: 10.1109/CVPR.2016.90.

[48] 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.

[49] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[50] 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, doi: 10.1109/TPAMI.2013.50.

[51] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. NIPS*, 2012.

[52] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3687–3691.

[53] Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, and X. Xue, "Exploring inter-feature and inter-class relationships with deep neural networks for video classification," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 167–176.

[54] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

[55] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. ICML*, 2011.

[56] C. Silberer and M. Lapata, "Learning grounded meaning representations with autoencoders," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2014.

[57] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5005–5013, doi: 10.1109/CVPR.2016.541.

[58] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, and D. Tran, "Self-supervised learning by cross-modal audio-video clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020.

[59] K.-L. Tseng, Y.-L. Lin, W. Hsu, and C.-Y. Huang, "Joint sequence learning and cross-modality convolution for 3D biomedical segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6393–6400.

[60] K. Noda, H. Arie, Y. Suga, and T. Ogata, "Multimodal integration learning of robot behavior using deep neural networks," *Robot. Auton. Syst.*, vol. 62, no. 6, pp. 721–736, 2014, doi: 10.1016/j.robot.2014.03.003.

[61] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "HyperDense-net: A hyper-densely connected CNN for multi-modal image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 5, pp. 1116–1126, May 2019.

[62] T. Chen, X. Ma, X. Ying, and W. Wang, "Multi-modal fusion learning for cervical dysplasia diagnosis," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1505–1509, doi: 10.1109/ISBI.2019.8759303.

[63] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4613–4621.

[64] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[65] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. NIPS*, 2016.

[66] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 299–307.

[67] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. 30th International Conf. Neural Inf. Process. Syst. (NIPS)*, Dec. 2016, pp. 289–297.
[68] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.
[69] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. KDD*, 2016, pp. 785–794.
[70] J. Yoon, J. Jordon, and M. van der Schaar, "INVASE: Instance-wise variable selection using neural networks," in *Proc. ICLR*, 2019.
[71] D. C. Mocanu, E. Mocanu, P. Stone, P. H. Nguyen, M. Gibescu, and A. Liotta, "Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science," *Nature Commun.*, vol. 9, no. 1, p. 2383, Dec. 2018.
[72] R. Tanno, K. Arulkumaran, D. C. Alexander, A. Criminisi, and A. Nori, "Adaptive neural trees," 2018, *arXiv:1807.06699*.
[73] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, CA, USA: Wadsworth and Brooks, 1984.
[74] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.
[75] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2017, *arXiv:1710.05941*.
[76] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*.
[77] R. Singh and N. S. Mangat, "Stratified sampling," in *Elements of Survey Sampling* (Kluwer Texts in the Mathematical Sciences), vol. 15, Dordrecht, The Netherlands: Springer, 1996, pp. 102–144, doi: 10.1007/978-94-017-1404-4_5.
[78] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[79] C. Guo and F. Berkhahn, "Entity embeddings of categorical variables," 2016, *arXiv:1604.06737*.
[80] [Online]. Available: https://www.kaggle.com/c/siim-isic-melanoma-classification

YI LIU was born in Taichung, Taiwan. He received the M.S. degree from the Department of Statistic, Tunghai University, in December 2016. He is currently a Senior Engineer with the Data Science Analysis Department, AU Optronics Corporation, Taiwan. His research interests include pattern recognition, computer vision, deep learning, statistical computing, and survival analysis.

HSUEH-PING LU received the Ph.D. degree in industrial engineering and engineering management from the National Tsing Hua University, Hsinchu, Taiwan. He is currently the Director of the Division of Digital Technology, AU Optronics Croporation, Taiwan.

CHING-HAO LAI was born in Kaohsiung, Taiwan. He received the Ph.D. degree from the Department of Computer Science and Engineering, National Chung-Hsing University, in January 2010. During he works toward his Ph.D. degree, he was also a Lecturer with the National Taichung Institute of Technology and the Central Taiwan University of Science and Technology. From 2010 to 2021, he was a Research and Development Manager and a Senior Researcher with the Institute for Information Industry and the Industrial Technology Research Institute, Taiwan. He was also a Deep Learning Researcher at Kneron Inc., USA, in 2020. He is currently a Senior Engineering Manager with the Data Science Analysis Department, AU Optronics Corporation, Taiwan. His research interests include pattern recognition, image processing, computer vision, deep machine learning, bioinformatics, data mining, and cloud computing.