IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 September 2023, accepted 25 September 2023, date of publication 2 October 2023,
date of current version 10 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3321111

RESEARCH ARTICLE

CMENet: A Cross-Modal Enhancement Network for Tobacco Leaf Grading

QINGLIN HE[1,*], XIAOBING ZHANG[2,*], JIANXIN HU[1], ZEHUA SHENG[1], QI LI[2], SI-YUAN CAO[1,3], AND HUI-LIANG SHEN[1], (Member, IEEE)

[1]College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China
[2]China Tobacco Zhejiang Industrial Company Ltd., Hangzhou 310008, China
[3]Ningbo Innovation Center, Zhejiang University, Ningbo 315100, China

Corresponding authors: Qi Li (liqi@zjtobacco.com) and Si-Yuan Cao (karlcao@hotmail.com)

This work was supported in part by the "Pioneer" and "Leading Goose" Research and Development Program of Zhejiang under Grant 2023C03136, and in part by the Ten Thousand Talents Program of Zhejiang Province under Grant 2020R52003.

[*]Qinglin He and Xiaobing Zhang contributed equally to this work.

ABSTRACT Tobacco leaf grading plays a crucial role in ensuring the quality of tobacco production. For a very long period, the grading process is manually determined by experienced experts. In recent years, some methods have been introduced to automate the grading process by utilizing the reflection images of tobacco leaves. However, the high visual similarity among reflection images at different grades renders a single reflection image insufficient for achieving accurate grading. Besides, the tobacco leaves with an identical grade may have inconsistent visual appearances due to their different planting locations. It is known that an expert integrates multimodal information such as visual, tactile, and planting location cues when performing grading. Inspired by this, we propose an end-to-end Cross-modal Enhancement Network, named CMENet, for automatic tobacco leaf grading. In addition to the common reflection image, the network also adopts a transmission image to incorporate the thickness information in manual grading. CMENet comprises a difference-aware fusion module and a meta self-attention module, enabling the extraction of multimodal information from the transmission image and the planting location, respectively. Experimental results demonstrate that our CMENet achieves a high grading accuracy (80.15%) when incorporating multimodal information, surpassing the performance of existing methods that rely solely on reflection images.

INDEX TERMS Tobacco leaf grading, image classification, convolutional neural network, cross-modal information fusion.

I. INTRODUCTION

Tobacco leaf grading is very important to the tobacco production process [1]. Because of its agricultural importance, tobacco grows in more than 100 countries, consumed as cigarettes, cigars, snuff, etc. [2], [3]. To maintain the quality of production, it is essential to categorize the tobacco leaves after re-drying. The grading is commonly achieved by assessing the visual appearance of the leaves, which is intricately related to the intrinsic quality of the tobacco leaves [1].

Manual grading has dominated the tobacco leaf assessment process for a very long period [4]. Experienced experts assess tobacco leaves based on visual, olfactory, and tactile cues, as well as information regarding their planting locations. Experts consider multiple factors during the grading process, contributing to a high degree of accuracy. However, there are several limitations in manual grading. First, the scarcity of experts fails to meet the increasing demand for tobacco leaf grading. Second, inexperienced staff demonstrate subjective discrepancies due to the reliance of grading on human sensory perception [5]. As shown in Fig. 1 and Fig. 2, the different grades of reflection images have visual similarities while the same grade from different planting locations displays inconsistent visual appearances. These inconsistencies

The associate editor coordinating the review of this manuscript and approving it for publication was Wenming Cao.

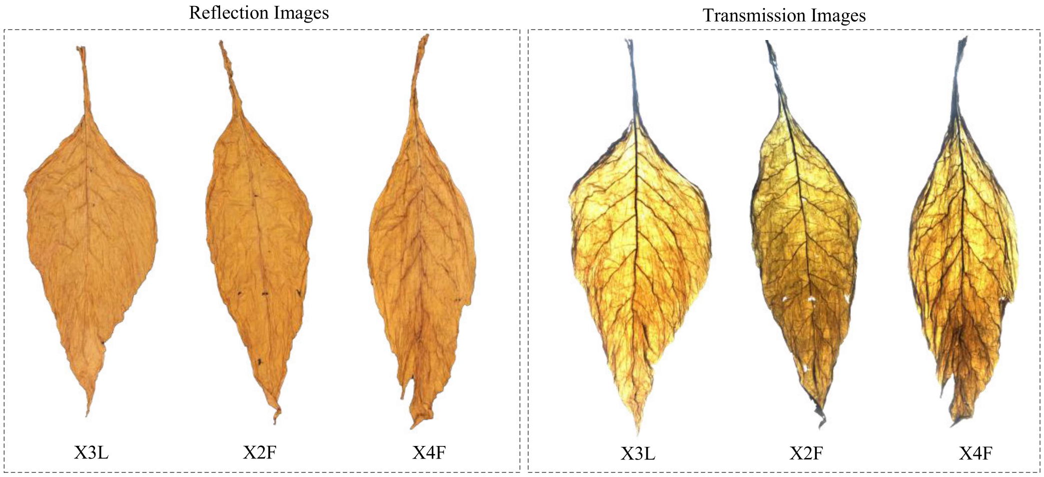


FIGURE 1. Reflection and transmission images of tobacco leaves at different grades. The reflection images of three grades ("X3L", "X2F", and "X4F") are visually similar, while the transmission images are perceptually quite different. These tobacco leaves are from the same planting location "FJ-NP" (see Table 2).

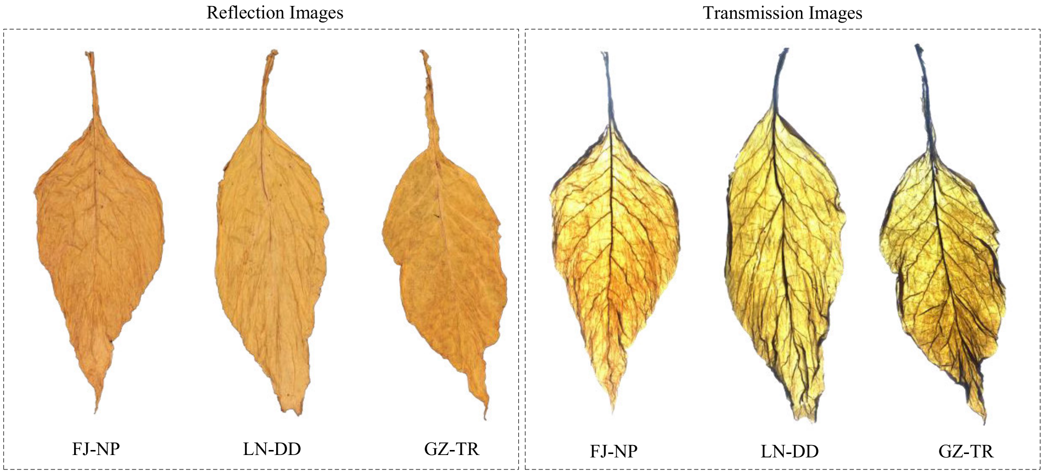


FIGURE 2. Reflection and transmission images of tobacco leaves from the planting locations "FJ-NP", "LN-DD", and "GZ-TR" illustrate distinct visual appearances, despite belonging to the same grade "X3L."

significantly increase the risk of misclassification during the grading process.

To overcome the limitations of manual grading, in recent years several methods have been employed to achieve automatic tobacco leaf grading. Previous works [6], [7], [8], [9] have focused on extracting handcrafted features and designing specific classifiers to perform tobacco leaf grading. With the rapid development of deep learning, some works have employed neural networks to improve agricultural production efficiency, such as fruit counting [10], leaf disease detection [11], plant recognition [12], etc. Deep learning techniques have been applied in the development of vision-based systems for automatic tobacco leaf grading [5], [13], [14]. These methods rely solely on the reflection image of the tobacco leaf as the input for their systems. However, due to the presence of high visual similarity among reflection images at different grades and the absence of additional information used by experienced experts in the manual grading process, the grading accuracy is affected.

In this work, inspired by the incorporation of additional cues during manual grading, we intend to integrate multi-modal information into our designed system. Compared to the reflection images, transmission images contain information about the thickness of tobacco leaves, the density of leaf tissue, and variations in intracellular substances. Regions with thinner tobacco leaves display increased brightness in transmission images. Therefore, they can be used as complementary information to the reflection image, providing the tactile information involved in manual grading. Furthermore, it is observed that reflection images of the same grade show visual variations attributed to dissimilarities in soil chemical composition across different planting locations [15],

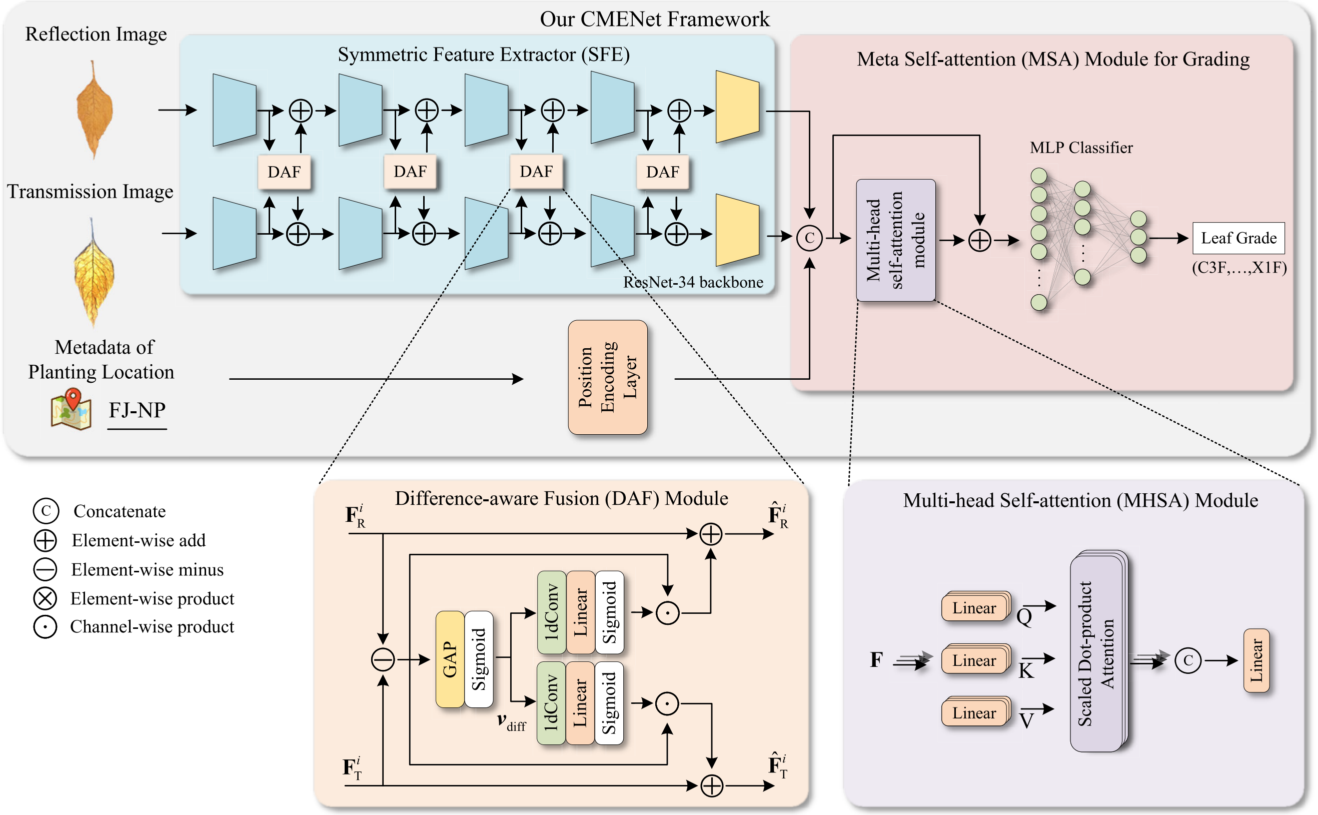


FIGURE 3. The framework of our cross-modal enhancement network (CMENet) for tobacco leaf grading.

as shown in Fig. 2. Consequently, relying solely on reflection images is inadequate for achieving automatic tobacco leaf grading. We aim to improve the performance by incorporating complementary information such as transmission images and metadata of planting location.

Based on the above analysis, we propose an end-to-end cross-modal enhancement network, named CMENet, for automatic tobacco leaf grading. The framework of CMENet is illustrated in Fig. 3. By incorporating the transmission image and the metadata of planting location, our method can overcome the visual inconsistency of sole reflection image input. Specifically, we design two novel modules for multimodal information fusion. The difference-aware fusion module effectively aggregates the extracted features from reflection images and transmission images at each layer of the backbone network. The meta self-attention module for grading further enhances feature representations by incorporating metadata of planting location. With these modules, CMENet can distinguish the inconsistencies in the visual appearance of tobacco leaves at the same grade caused by different planting locations, and capture the subtle differences among tobacco leaves of different grades. In summary, the main contributions of this work are as follows.

- We propose a novel **C**ross-**m**odal **E**nhancement **Net**work, named CMENet, for automatic tobacco leaf grading. CMENet has three inputs, including reflection image, transmission image, and metadata of planting location.
- We specially design two novel modules to fuse multimodal information. These modules are used to effectively integrate features related to leaf thickness and planting location differences.
- The evaluation on our self-collected dataset indicates that CMENet can achieve a high grading accuracy (80.15%) by incorporating multimodal information, outperforming the current methods using only reflection images.

The rest of the paper is organized as follows. Section II reviews the related work. Section III presents the framework and details of our CMENet. Section IV discusses the experimental results. Finally, Section V concludes this work.

II. RELATED WORK

This section provides a brief review of the work related to the application of deep learning in agriculture, automatic tobacco leaf grading, and multimodal information fusion.

A. APPLICATIONS OF DEEP LEARNING IN AGRICULTURE

Deep learning is a technical tool with broad application prospects and plays an important role in the field of image recognition [16]. Convolutional neural networks (CNNs), which constitute a specific class of deep learning, demonstrate excellent performance in extracting image features, rendering them well-suited for image classification tasks [17]. CNNs have been successfully applied in diverse domains [18], and their application has recently extended to the field of agriculture as well [19]. For example, Lin et al. [20] design a lightweight residual network to trace the quality of soybean from different origins. Zeng et al. [21] adopt LeNet to identify pears with different damage levels. Dyrmann et al. [22] adopt VGG16 to classify weeds from crop species based on 22 different species in total. Some works employ CNNs to conduct image classification and recognition tasks, such as plant species identification [23], leaf disease detection [11], and fruit counting [24]. Some works employ CNNs to predict future values of agricultural products, such as crop yield production [25] and the estimation of field soil moisture content [26]. Recently, transformer [27] is proposed in [28] to model sequential data in the field of NLP. Vision Transformer (ViT) [29] and many other ViT variants [30], [31], [32] are proposed from then, which achieve promising performance compared with its counterpart CNNs for image analysis tasks. Deep learning have demonstrated significant potentials in agricultural production such as productivity, environmental impact, food security, and sustainability [33].

B. AUTOMATIC TOBACCO LEAF GRADING

Many previous works [1], [7], [8], [9] conduct automatic tobacco leaf grading via image processing. For example, Thomas [8] first applies image processing techniques in tobacco leaf grading. Cho et al. [9] extract the shape, appearance, and other features of tobacco leaves for grading purposes. Zhang et al. [1] develop an automatic grading method for tobacco leaves based on the nearest neighbor algorithm. Zhang and Zhang [7] further improve the accuracy of automatic tobacco leaf grading by combining machine vision with fuzzy set theory. Because of the high visual similarity among reflection images at different grades, traditional approaches cannot achieve high grading accuracy. With the development of deep learning, recent works [5], [14], [34] have adopted various architectural neural networks to accomplish automatic tobacco leaf grading. For example, Luo et al. [14] adopt AlexNet and hard example mining to improve the grading accuracy based on the color, shape, texture, and other extracted features from reflection images. Li et al. [34] employ VGG16 and transfer learning to achieve automatic tobacco leaf grading. Lu et al. [5] adopt an improved ResNet for automatic tobacco leaf grading. We notice that almost all previous works use only the tobacco leaf image and neglect the additional cues (*e.g.*, leaf thickness and planting locations) that experienced experts always adopt in manual grading process [35].

C. MULTIMODAL INFORMATION FUSION

Multimodal information fusion refers to the integration of information from different sensors or modalities to obtain a more comprehensive, accurate, and reliable understanding. It has been widely applied in various fields. For example, the integration of patient images and clinical information can aid in the diagnosis of Alzheimer's disease within the medical domain [36]. Within the realm of autonomous driving, remarkable advancements have been made by combining data from LiDAR (Light Detection and Ranging) sensors and RGB cameras [37].

In the field of tobacco leaf grading, the quality of the leaves is mainly determined by their chemical composition that can be significantly influenced by the planting region. Considering that geographic factors have a significant influence on leaf quality, Jiang et al. [15] employ NIR spectroscopy to classify tobacco leaves originating from distinct growing areas. Tang et al. [38] conduct a detailed analysis to explore the relationship between the natural conditions of the tobacco leaf cultivation environment and its chemical composition. Li et al. [39] suggest that incorporating multimodal information such as hyperspectral data into tobacco leaf grading would be a future research trend. Although previous works have demonstrated the significant impact of different modalities of information on tobacco leaf quality, they have not been specifically used in automatic tobacco leaf grading.

III. METHOD

A. CMENET FRAMEWORK

Figure 3 shows the framework of our CMENet. CMENet takes reflection image $\mathbf{I}_{\mathrm{R}} \in \mathbb{R}^{H\times W\times 3}$, transmission image $\mathbf{I}_{\mathrm{T}} \in \mathbb{R}^{H\times W\times 3}$, and metadata of planting location $\mathrm{M}_{\mathrm{L}} \in \mathbb{R}^{1\times 1}$ as input. CMENet mainly consists of three modules, *i.e.*, symmetric feature extractor (SFE), difference-aware fusion (DAF) module, and meta self-attention (MSA) module for grading. SFE extracts the reflection image feature $\mathbf{F}_{\mathrm{R}} \in \mathbb{R}^{1\times c}$ and transmission image feature $\mathbf{F}_{\mathrm{T}} \in \mathbb{R}^{1\times c}$. The DAF module fuses the feature map extracted by each layer. The MSA module aggregates the image feature $\mathbf{F}_{\mathrm{R}}$, image feature $\mathbf{F}_{\mathrm{T}}$, and metadata of planting location M_{L}, and finally obtain grading prediction $\mathbf{Y} \in \mathbb{R}^{1\times N}$.

B. SYMMETRIC FEATURE EXTRACTOR (SFE)

Traditional CNNs framework on image classification typically only accommodates a single image input. Due to the significant visual distinctions between reflection and transmission images, adopting a shared-weight network may not be suitable for the feature extraction stage. Inspired by symmetrical architecture of other networks such as SEDRFuse [40] and Siamese CNN [41], we specially design a dual-branch feature extract network for feature extraction to mutually promote the learning for features of both reflection and transmission images. However, our SFE module is different to [40] and [41] in that the two pathways in SFE do not share weights.

The architectures of the two branches is symmetric, ensuring that the feature extracted from reflection and transmission images have a consistent size. We adopt ResNet-34 [42] as the backbone of this module, because ResNet-34 achieves the optimal balance between accuracy and inference latency among all well-known backbone network (see Table 4). The SFE module $\mathtt{E}_{\mathtt{F}}(\cdot)$ takes the tobacco leaf reflection image $\mathbf{I}_{\mathrm{R}}$ and the transmission image $\mathbf{I}_{\mathrm{T}}$ as inputs, to extract the reflection and the transmission feature maps, respectively. The whole process can be mathematically formulated as

$$\mathbf{F}_{\mathrm{R}}, \mathbf{F}_{\mathrm{T}} = \{\mathtt{E}_{\mathtt{F}}(\mathbf{I}_{\mathrm{R}}), \mathtt{E}_{\mathtt{F}}(\mathbf{I}_{\mathrm{T}})\}, \tag{1}$$

where $\mathbf{F}_{\mathrm{R}}$ and $\mathbf{F}_{\mathrm{T}}$ represent the reflection and the transmission feature maps yielded by SFE, respectively.

C. DIFFERENCE-AWARE FUSION (DAF) MODULE

Unlike the common strategy adopted in information fusion, in which no cross-modal information is exchanged in the model until after the classifier [43], our difference-aware fusion (DAF) module achieves the feature exchange between two pathways. The key idea of our DAF module is to leverage channel weighting to fully integrate global context for both reflection and transmission images because of their complementary nature.

It is known that ResNet-34 is composed of several stacked layers, including five convolutional layers (`conv1`, `conv2_x`, `conv3_x`, `conv4_x`, `conv5`), a pooling layer, and full connection layer. Specifically, we denote $\mathbf{F}_{\mathrm{R}}^{i} \in \mathbb{R}^{h_i \times w_i \times c_i}$ and $\mathbf{F}_{\mathrm{T}}^{i} \in \mathbb{R}^{h_i \times w_i \times c_i}$ as the feature tensors yielded from the i-th convolutional layer of SFE.

Since the reflection and transmission images are from the same tobacco leaf but different lighting conditions, the reflection image emphasizes color and texture features while the transmission image emphasizes leaf thickness and veins. Mathematically, we compute the common part as

$$\mathbf{F}_{\mathrm{comm}}^{i} = \frac{\mathbf{F}_{\mathrm{R}}^{i} + \mathbf{F}_{\mathrm{T}}^{i}}{2}, \tag{2}$$

and compute the complementary part as

$$\mathbf{F}_{\mathrm{comp}}^{i} = \frac{\mathbf{F}_{\mathrm{R}}^{i} - \mathbf{F}_{\mathrm{T}}^{i}}{2}. \tag{3}$$

In the DAF module, we adopt the element-wise minus to obtain the complementary part, thereby enhancing the common part feature in the subsequent procedures.

The goal of the DAF module is to fully integrate the common and complementary information from the features $\mathbf{F}_{\mathrm{R}}^{i}$ and $\mathbf{F}_{\mathrm{T}}^{i}$ extracted at different layers of the SFE. We extract the feature maps of the complementary part as

$$\mathbf{v}_{\mathrm{diff}} = \sigma(\mathtt{GAP}(\mathbf{F}_{\mathrm{R}}^{i} - \mathbf{F}_{\mathrm{T}}^{i})), \tag{4}$$

where the symbol $\sigma(\cdot)$ denotes the sigmoid activation function, $\mathtt{GAP}(\cdot)$ represents global average pooling, and $\mathbf{v}_{\mathrm{diff}} \in \mathbb{R}^{1\times C}$ refers to the complementary features. In (4), the global average pooling compresses the complementary features into a vector $\mathbf{v}_{\mathrm{diff}}$ that captures the discrepancy between the features of reflection and transmission images.

Then, we employ an one-dimensional convolution and a fully connected layer to generate two channel-weighted vectors that enhance the original features of both the reflection and transmission images. These weighted vectors are normalized to a range of $[0, 1]$ using the sigmoid activation function. Finally, the complementary features are multiplied by the normalized channel weights and added to the original features. The enhanced reflection image feature $\hat{\mathbf{F}}_{\mathrm{R}}^{i}$ can be formulated as

$$\hat{\mathbf{F}}_{\mathrm{R}}^{i} = \mathbf{F}_{\mathrm{R}}^{i} \oplus \sigma(f(\mathbf{W}_{\mathrm{R}} * \mathbf{v}_{\mathrm{diff}})) \odot (\mathbf{F}_{\mathrm{R}}^{i} - \mathbf{F}_{\mathrm{T}}^{i}), \tag{5}$$

where $\oplus$ and $\odot$ denotes element-wise addition and channel-wise multiplication, respectively. $f(\cdot)$ denotes fully connected layers, $\mathbf{W}_{\mathrm{R}}$ denotes the parameters of the one-dimensional convolution, and $*$ denotes the operation of convolution.

In the same way, the enhanced transmission image feature $\hat{\mathbf{F}}_{\mathrm{T}}^{i}$ can be mathematically formulated as

$$\hat{\mathbf{F}}_{\mathrm{T}}^{i} = \mathbf{F}_{\mathrm{T}}^{i} \oplus \sigma(f(\mathbf{W}_{\mathrm{T}} * \mathbf{v}_{\mathrm{diff}})) \odot (\mathbf{F}_{\mathrm{R}}^{i} - \mathbf{F}_{\mathrm{T}}^{i}), \tag{6}$$

where $\mathbf{W}_{\mathrm{T}}$ denotes the parameters of the one-dimensional convolution.

D. META SELF-ATTENTION (MSA) MODULE FOR GRADING

The meta self-attention (MSA) module for grading is designed to further enhances feature representations by incorporating metadata of planting location.

We obtain the reflection image feature $\mathbf{F}_{\mathrm{R}} \in \mathbb{R}^{1\times c}$, transmission image feature $\mathbf{F}_{\mathrm{T}} \in \mathbb{R}^{1\times c}$, and metadata of planting location feature $\mathbf{F}_{\mathrm{L}} \in \mathbb{R}^{1\times c}$ through the SFE and position encoding layer, respectively. The position encoding layer is a fully connected layer, encoding the metadata M_{L} into a feature vector $\mathbf{F}_{\mathrm{L}}$. These three features are from different modalities. We concatenate them together to obtain a multimodal feature $\mathbf{F} \in \mathbb{R}^{1\times 3c}$. We adopt the self-attention mechanisms [28], [44], which effectively capture non-local contextual information, in designing our meta self-attention (MSA) module based on the multi-head self-attention (MHSA) module.

The j-th head attention matrix $\mathbf{A}_j$ is computed as

$$\mathbf{A}_j = \mathtt{softmax}(\frac{\mathbf{W}_j^{\mathrm{Q}}\mathbf{F}(\mathbf{W}_j^{\mathrm{K}}\mathbf{F})^{\mathsf{T}}}{\sqrt{d_k}})\mathbf{W}_j^{\mathrm{V}}\mathbf{F}, \tag{7}$$

where $d_k \in \mathbb{R}^{1\times(c/H)}$ denotes the scaling parameter, $\mathbf{W}_j^{\mathrm{Q}}$, $\mathbf{W}_j^{\mathrm{K}}$, and $\mathbf{W}_j^{\mathrm{V}}$ are the j-th head learnable parameter matrixes. Then, we concatenate each head $\mathbf{A}_j$ as output,

$$\mathbf{A} = \mathtt{concat}(\mathbf{A}_1, \ldots, \mathbf{A}_j, \ldots, \mathbf{A}_H)\mathbf{W}^{\mathrm{O}} \tag{8}$$

where $\mathbf{W}^{\mathrm{O}}$ denotes the parameter matrix of the output linear layer, and H denotes the number of heads of MHSA.

Finally, we add the $\mathbf{A}$ to the multimodal feature $\mathbf{F}$ as output and pass it to the multilayer perceptron (MLP) to predict the grade of tobacco leaf. The output of the MSA module for grading can be described as

$$\mathbf{Y} = \mathtt{MLP}(\mathbf{A} + \mathbf{F}), \tag{9}$$

where $\mathbf{Y} \in \mathbb{R}^{1\times N}$ denotes the one-hot encoding of the grade, with N denoting the total number of grades.

E. LOSS FUNCTION

In practice, the numbers of tobacco leaves in different grades are imbalanced (see Table 1). We adopt the weighted cross entropy loss function that assigns higher weights to samples from the minority class during training,

$$\mathcal{L}_{\text{WCE}} = -\frac{1}{N}\sum_{k=1}^{N} w_k y_k \log \hat{y_k} + (1 - y_k)\log(1 - \hat{y_k}) + \gamma \|\mathbf{W}\|_2, \tag{10}$$

where y_k and $\hat{y}_k$ denote the actual and predicted grades, respectively. w_k is the weight assigned to the k-th class. N is the total number of grades of tobacco leaf. γ is the coefficient to control the trade-off between the cross-entropy loss and the regularization term. $\| \cdot \|_2$ denotes the L2 norm to prevent network overfitting. $\mathbf{W}$ represents all the learnable parameters of CMENet.

IV. EXPERIMENTS

In the experiments, we conduct detailed performance evaluations of our CMENet and other methods on the self-collected dataset. We first present our dataset construction, experimental configurations, and implementation details. Then, we compare and analyze the grading accuracy of our approach against current methods. Finally, we conduct ablation studies to validate the effectiveness of our network design.

A. DATASET

We note that there are no public datasets that can meet our tobacco leaf grading task conducted in this work. Therefore we establish a multimodal image data acquisition system specifically for tobacco leaves, which seamlessly integrates device control functions with automatic tobacco leaf grading software. As illustrated in Fig. 4, the system mainly comprises two cameras and two light sources. To prevent the interference of external ambient light, all images are captured within a sealed environment. The light source of the system is provided by LEDs and halogen lamps positioned within the black cabinet, ensuring consistent lighting conditions when acquiring images.

Bottom illumination is essential to capture transmission images. To ensure the smooth movement of tobacco leaves, the transmission area is kept relatively small. Consequently, the complete transmission image is obtained by stitching multiple images together.

Our tobacco leaf dataset consists of 9799 reflection images and 9799 transmission images. The grades of tobacco leaves follow the GB2635-94 standard of China [45]. These tobacco leaves are categorized into 29 grades. Table 1 provides a comprehensive list of the 29 grades along with their corresponding numbers of tobacco leaves. As in practical applications, the sample numbers of different grades are imbalanced. For example, the highest quantity grade (X3L) consists of 1880 samples, while the lowest quantity grade (X1F) has only 33 samples.

TABLE 1. **The 29 grades of tobacco leaves and their corresponding numbers of samples in our dataset.**

Index	Grade	#	Index	Grade	#	Index	Grade	#
1	C3F	1880	11	CX1K	361	21	GY2	109
2	B2F	915	12	B2K	283	22	CX2K	100
3	C4F	783	13	B3F	276	23	B1L	84
4	C3L	730	14	X3L	266	24	S1	79
5	B1K	514	15	C2F	242	25	B4F	65
6	B2L	502	16	C3V	219	26	S2	55
7	C2L	423	17	GY1	200	27	B2V	53
8	X2F	378	18	C4L	177	28	X1L	47
9	X3F	370	19	B3L	172	29	X1F	33
10	X2L	366	20	X4F	117			

Our dataset comprises tobacco leaves collected from 24 different planting locations across China. Table 2, presents a comprehensive overview of the planting locations, including their abbreviations, full names, and corresponding numbers of tobacco leaves. For instance, "FJ-NP" corresponds to the Nanping City in Fujian Province, where 1799 tobacco leaves have been collected.

B. EXPERIMENTAL SETUP

We implement our CMENet using Pytorch1.12.0. All the experiments are performed on the Ubuntu 20.04.1 LTS operating system, equipped with an Intel(R) Xeon(R) Gold 6226R CPU and NVIDIA GeForce RTX 3090 with 24 GB VRAM. Only the weight parameters of ResNet-34, which serves as the backbone network, are initialized with pre-trained ResNet-34 weights from ImageNet before start training. The parameters of other components of TCENet, such as SFE, DAF, and MSA, will be updated as the network undergoes training. The weights in (10) are set inversely proportional to the class frequencies, meaning that the less frequent classes are assigned higher weights. The coefficient γ is set to 0.0009. In the training stage, we use the Adam optimizer with an initial learning rate of 0.0003. The learning rate follows an exponential decay strategy, which is adjusted to 0.9 times the current value after 30 epochs. The batch size is set to 32. The network was trained for a total of 150 epochs, and as shown in Fig. 5, it has converged quite well.

We randomly select 7839 tobacco leaf samples from our dataset as the training set and 1960 samples as the test set, ensuring a similar class distribution between the training and testing sets. Both reflection and transmission images are of size 428 × 286 pixels.

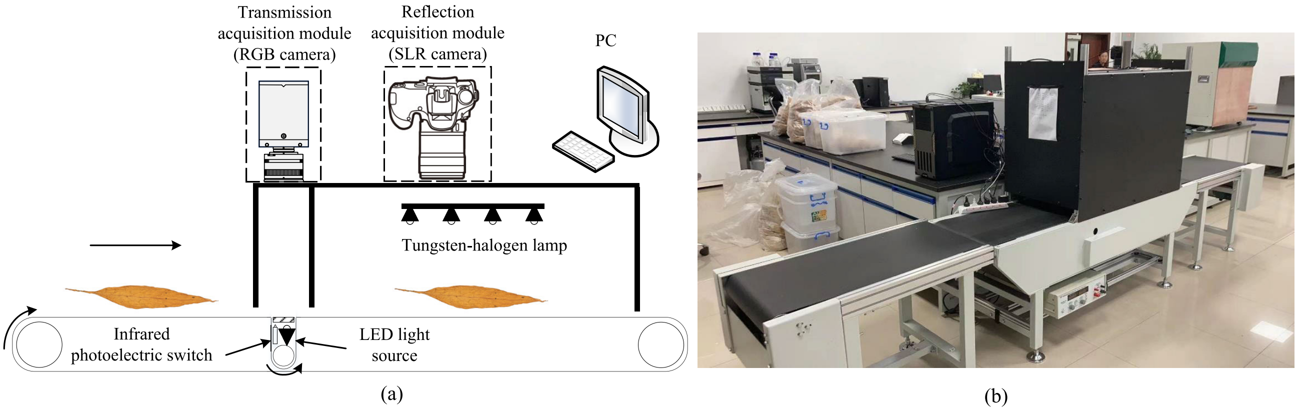

Transmission acquisition module (RGB camera)
Reflection acquisition module (SLR camera)
PC
Tungsten-halogen lamp
Infrared photoelectric switch
LED light source
(a)
(b)

FIGURE 4. The multimodal imaging data acquisition system for tobacco leaves. (a) Schematic diagram, (b) Real system.

TABLE 2. The 24 distinct planting locations and corresponding numbers of tobacco leaves in our dataset.

Abbr.	Full Address	#	Abbr.	Full Address	#	Abbr.	Full Address	#
FJ-NP	Nanping City, Fujian Province	1799	YN-YX	Yuxi City, Yunnan Province	482	HN-NY	Nanyang City, Henan Province	277
FJ-LY	Longyan City, Fujian Province	279	YN-QJ	Qujing City, Yunnan Province	377	HN-LY	Luoyang City, Henan Province	171
LN-DD	Dandong City, Liaoning Province	1275	YN-HH	Honghe City, Yunnan Province	352	HN-SMX	Sanmenxia City, Henan Province	140
GZ-BJ	Baijie City, Guizhou Province	702	YN-DL	Dali City, Yunnan Province	331	HN-YZ	Yongzhou City, Hunan Province	374
GZ-TR	Tongren City, Guizhou Province	282	YN-CX	Chuxiong City, Yunnan Province	276	HN-CZ	Chenzhou City, Hunan Province	178
SC-LS	Liangshan City, Sichuan Province	647	YN-PE	Pu'er City, Yunnan Province	216	HN-CS	Changsha City, Hunan Province	137
HB-ES	Enshi City, Hebei Province	467	AH-WN	Wannan City, Anhui Province	333	SD-LY	Linyi City, Shandong Province	185
CQ	Chongqing City	256	GX-BS	Baise City, Guangxi Province	157	JX-GZ	Ganzhou City, Jiangxi Province	106

C. EVALUATION METRICS

We evaluate the performance of our CMENet and other existing approaches using four indicators, including testing accuracy rate, recall rate, precision rate, and F1-score. The accuracy rate measures the ratio of correct predictions over the total number of instances evaluated. It is computed as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{11}$$

where TP and TN denote the number of positive and negative instances that are correctly classified, respectively. FP and FN denote the number of misclassified positive and negative instances, respectively. The recall rate measures the number of correctly graded samples selected by the classifier, computed as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{12}$$

The precision rate shows the percentage of all the corrected grading samples in all the selected samples by the classifier, computed as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{13}$$

F1-score aggregates precision and recall measures under the concept of harmonic mean, computed as

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{14}$$

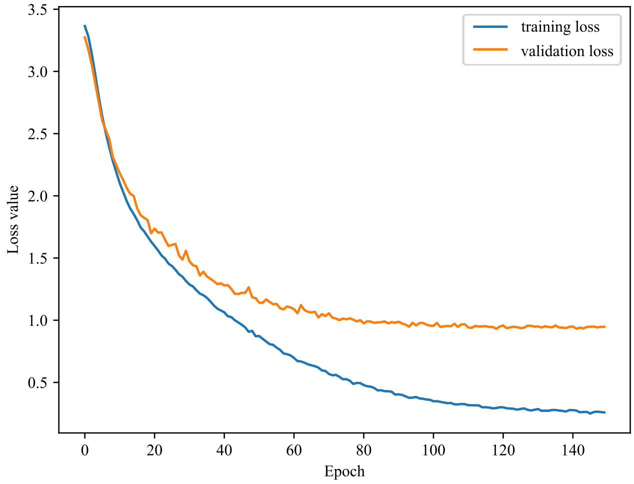


FIGURE 5. The trend of CMENet in loss for the training set and validation set.

D. EVALUATION OF FUSION STRATEGY

We explore the influence of using different fusion strategies on grading performance, including direct concatenation, element-wise multiplication, low-rank multimodal fusion [46], and autoencoder [47]. The low-rank multimodal fusion (LMF) approach adopts low-rank matrix decomposition of the weights. It transforms the combination of tensor outer product and fully connected layers into individual linear

TABLE 3. The accuracy (Acc), F1-score (F1), precision (Pre), and recall (Rec) values of our strategy (DAF+MSA) and other multimodal fusion strategies.

Fusion Strategy	Acc (%)	F1 (%)	Pre (%)	Rec (%)
Concat	76.38	74.92	**77.78**	73.41
Multiply	78.98	75.73	76.71	75.52
LMF [46]	71.58	66.55	67.33	66.79
Autoencoder [47]	72.35	67.45	67.50	68.01
DAF+MSA (Ours)	**80.15**	**76.51**	77.47	**76.45**

transformations for each modality, followed by the multidimensional dot product. LMF can be regarded as a summation of the results from multiple low-rank vectors, which can effectively reduce the number of parameters in the model. The autoencoder approach encodes all modality features and then decodes them for reconstruction. Autoencoder aims to capture the underlying latent representations shared across different modalities.

Table 3 demonstrates that directly using concatenation for multimodal information fusion leads to a decrease in classification accuracy. In the context of tobacco leaf grading, it is conjectured that fusion strategies employed in other works, such as LMF and Autoencoder specifically designed for multimodal feature fusion in domains like speech and video, may not be suitable for this particular scenario. In comparison, our fusion strategy (DAF+MSA) yields the best performance for tobacco leaf grading.

E. RESULTS OF TOBACCO LEAF GRADING

We compare our CMENet with the backbone network used for general image classification, including ResNet series [42], ConvNeXt_tiny [48], CSPDarkNet53 [49], EfficientNet-B0 [50], InceptionV3 [51], VGG-13 [52], Swin Transformer [30], Visformer [31] and MobileViT [32]. Considering the limited size of our dataset, we adopt pre-trained weights from the ImageNet [53] dataset to initialize all the models. The grading performance of different methods on our dataset is summarized in Table 4. In Table 4, apart from CMENet, the input for other backbone networks consists solely of reflection images. It can be observed that the performance of transformer-based models with larger parameter sizes is not satisfactory when fine-tuned on our dataset. ResNet-34, on the other hand, demonstrates a good balance between inference speed and accuracy. Thanks to the effective integration of multimodal information, CMENet achieves mediocre inference times but outstanding accuracy. Such results revealing that the latency caused by multimodal inputs is tolerable, and CMENet is feasible for designing a accurate and real-time automatic tobacco leaf grading system in the future.

We also compare our CMENet with the current deep learning approaches used for existing automatic tobacco leaf grading and plant classification, including Lu et al. [5], Tang et al. [54], and Nasiri et al. [55]. Table 5 indicates that our CMENet performs better than these competitors.

TABLE 4. The accuracy (Acc), F1-score (F1), precision (Pre), recall (Rec), latency (LAT), and GFLOPs values of our CMENet and other backbone networks.

Method	Acc (%)	F1 (%)	Pre (%)	Rec (%)	LAT (ms)	GFLOPs
ResNet-18 [42]	76.17	73.84	77.08	71.65	2.81	1.82
ResNet-34 [42]	76.68	75.68	**77.84**	74.53	4.90	3.68
ConvNeXt_tiny [48]	63.78	62.05	66.84	60.17	5.41	4.45
CSPDarkNet53 [49]	72.45	70.98	73.67	69.17	9.40	5.00
InceptionV3 [51]	72.50	70.54	70.54	68.71	11.91	2.85
EfficientNet-B0 [50]	75.46	74.34	75.75	73.70	8.76	**0.38**
VGG-13 [52]	75.97	72.84	76.60	71.22	**1.85**	11.35
Swin Transformer [30]	74.13	72.31	73.93	71.51	16.26	8.54
Visformer [31]	75.15	73.01	75.25	71.99	8.40	4.76
MobileViT [32]	75.61	72.63	73.88	72.55	11.47	1.42
CMENet (Ours)	**80.15**	**76.51**	77.47	**76.45**	10.58	7.37

TABLE 5. The accuracy (Acc), F1-score (F1), precision (Pre), recall (Rec), latency (LAT), and GFLOPs values of our CMENet, the current tobacco leaf grading and plant classification approaches.

Method	Acc (%)	F1 (%)	Pre (%)	Rec (%)	LAT (ms)	GFLOPs
Lu *et al.* [5]	71.28	69.76	71.99	68.44	-	-
Tang *et al.* [54]	49.29	47.26	54.17	45.10	-	-
Nasiri *et al.* [55]	75.10	72.69	74.91	71.56	2.38	15.52
CMENet (Ours)	**80.15**	**76.51**	**77.47**	**76.45**	10.58	7.73

TABLE 6. The grades and number of tobacco leaves with grading accuracy above 90%, between 80% and 90%, and below 60%.

Acc > 90%			80% ≤ Acc < 90%			Acc < 60%		
Grade	#	Acc (%)	Grade	#	Acc (%)	Grade	#	Acc (%)
B4F	65	92.31	B2F	915	85.25	B1L	84	52.94
X1L	47	100.00	B2K	283	87.72	B2V	53	45.45
X2L	366	91.78	B3F	276	87.27	CX1K	361	58.33
X3F	370	93.24	B3L	172	85.29	GY2	109	50.00
X3L	266	96.30	C2L	423	82.35	X1F	33	42.86
X4F	117	95.65	C3F	1880	86.44			

Figure 6 shows the confusion matrix of our CMENet. It is observed that CMENet achieves good grading performance. Table 6 further shows the grades and numbers of tobacco leaves of three accuracy ranges, *i.e.*, above 90%, between 80% and 90%, and below 60%. It is observed that, despite having a limited number of training samples, the grades "B4F" and "X1L" still achieve accuracy above 90%. CMENet have good classification capabilities for most grades with prefix "B" and "X", even though they have visually similar appearances. However, CMENet does not perform well on the grades "B1L", "B2V", "CX1K", "GY2", and "X1F". This indicates that our CMENet still have limitation in extracting tobacco leaf features for certain grades.

F. ABLATION STUDY

We conduct ablation experiments with different inputs to evaluate the impact of transmission images and metadata of planting location on grading accuracy.

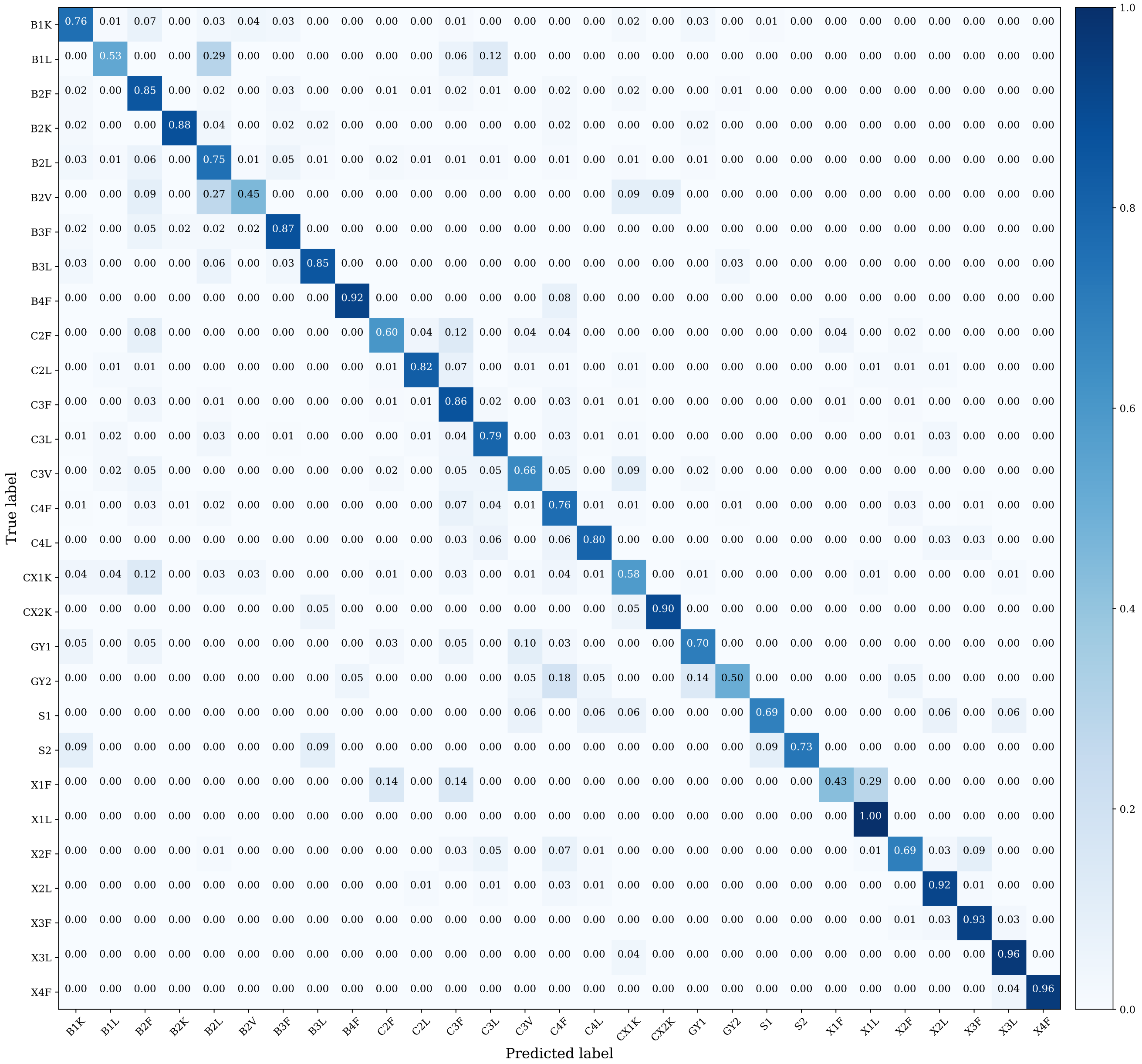


FIGURE 6. The confusion matrix of CMENet.

TABLE 7. Ablation study of reflection image $\mathbf{I_R}$, transmission image $\mathbf{I_T}$, and metadata of planting location $\mathbf{M_L}$ on the inputs of CMENet.

Model inputs			Acc (%)	F1 (%)	Pre (%)	Rec (%)
$\mathbf{I}_\mathrm{R}$	$\mathbf{I}_\mathrm{T}$	M_L				
✓			76.68	75.68	**77.84**	74.53
	✓		45.61	41.52	45.96	39.77
✓	✓		78.62	75.49	75.34	**76.73**
✓	✓	✓	**80.15**	**76.51**	77.47	76.45

Table 7 demonstrates the influence of incorporating multimodal information on the accuracy of tobacco leaf grading. Particularly, with the introduction of transmission image, the accuracy improves from 76.68% (when using only reflection images) to 78.62%. This highlights the importance of leveraging multimodal information to capture a comprehensive representation of the tobacco leaves. In addition, we conduct testing by training the classification model solely on transmission images, which results in a grading accuracy of 45.61%. This confirms that transmission image also carries grade information but itself is insufficient for grading. Furthermore, the incorporation of metadata of planting location leads to the best grading accuracy (80.15%).

We also conduct ablation experiments with coefficient γ in (10). The coefficient γ allows CMENet to control

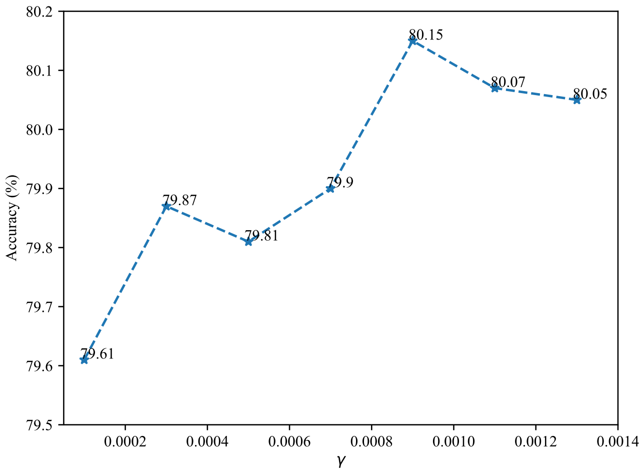


FIGURE 7. The impact of the coefficient γ on the accuracy of CMENet.

the relative importance of the weighted cross-entropy loss and the regularization term during training. By adjusting the value of γ, we can effectively balance the impact of the two components on the overall training process. As shown in Fig. 7, when the coefficient γ is set to 0.0009, the balance between weighted cross-entropy and regularization term is optimal, resulting in the highest model accuracy.

V. CONCLUSION

In this work, we have introduced a novel cross-modal enhancement network, named CMENet, for tobacco leaf grading. Inspired by the manual grading process, the network incorporates multimodal information as input, including reflection image, transmission image, and metadata of planting location. By introducing multimodal information, CMENet achieved an increase in grading accuracy from 76.68% to 80.15%. Extensive experiments show that it outperforms the state-of-the-art approaches that use only reflection images. As CMENet achieves a fast inference times (10.58 ms) and a high grading accuracy, it has potential in practical applications.

Limitations: The multimodal data employed in this work are still insufficient for tobacco leaf grading. It is known that the chemical composition of the tobacco leaf is also relevant to grading. In our future work, we will develop a near-infrared multispectral imaging system to acquire the chemical composition, aiming to further improve grading accuracy.

REFERENCES

[1] J. Zhang, S. Sokhansanj, S. Wu, R. Fang, and W. Yang, “A trainable grading system for tobacco leaves,” *Comput. Electron. Agricult.*, vol. 16, no. 3, pp. 231–244, Feb. 1997.

[2] F. M. A. Ahmed and S. W. Peeran, “Significance and determinants of tobacco use: A brief review,” *Dentistry Med. Res.*, vol. 4, no. 2, pp. 33–38, 2016.

[3] S. Soneji, J. Sargent, and S. Tanski, “Multiple tobacco product use among U.S. adolescents and young adults,” *Tobacco Control*, vol. 25, no. 2, pp. 174–180, Mar. 2016.

[4] J. Liu, J. Shen, Z. Shen, and R. Liu, “Grading tobacco leaves based on image processing and generalized regression neural network,” in *Proc. IEEE Int. Conf. Intell. Control, Autom. Detection High-End Equip.*, Jul. 2012, pp. 89–93.

[5] M. Lu, S. Jiang, C. Wang, D. Chen, and T. Chen, “Tobacco leaf grading based on deep convolutional neural networks and machine vision,” *J. ASABE*, vol. 65, no. 1, pp. 11–22, 2022.

[6] L. Han, “Recognition of the part of growth of flue-cured tobacco leaves based on support vector machine,” in *Proc. 7th World Congr. Intell. Control Autom.*, 2008, pp. 3624–3627.

[7] F. Zhang and X. Zhang, “Classification and quality evaluation of tobacco leaves based on image processing and fuzzy comprehensive evaluation,” *Sensors*, vol. 11, no. 3, pp. 2369–2384, Feb. 2011.

[8] C. Thomas, “Techniques of image analysis applied to the measurement of tobacco and related products,” in *Proc. Tobacco Chemists Res. Conf.*, 1988.

[9] H. Cho and K. Paek, “Feasibility of grading dried burley tobacco leaves using machine vision,” *J. Korean Soc. Agricult. Machinery*, vol. 22, no. 1, pp. 30–40, 1997.

[10] M. Rahnemoonfar and C. Sheppard, “Deep count: Fruit counting based on deep simulated learning,” *Sensors*, vol. 17, no. 4, p. 905, Apr. 2017.

[11] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, “Deep neural networks based recognition of plant diseases by leaf image classification,” *Comput. Intell. Neurosci.*, vol. 2016, pp. 1–11, Jun. 2016.

[12] G. L. Grinblat, L. C. Uzal, M. G. Larese, and P. M. Granitto, “Deep learning for plant identification using vein morphological patterns,” *Comput. Electron. Agricult.*, vol. 127, pp. 418–424, Sep. 2016.

[13] J. Li, H. Zhao, S. P. Zhu, H. Huang, Y. Miao, and Z. Jiang, “An improved lightweight network architecture for identifying tobacco leaf maturity based on deep learning,” *J. Intell. Fuzzy Syst.*, vol. 41, no. 2, pp. 4149–4158, Sep. 2021.

[14] H. Luo and C. Zhang, “Features representation for flue-cured tobacco grading based on transfer learning to hard sample,” in *Proc. 14th IEEE Int. Conf. Signal Process. (ICSP)*, Aug. 2018, pp. 591–595.

[15] D. Jiang, G. Qi, G. Hu, N. Mazur, Z. Zhu, and D. Wang, “A residual neural network based method for the classification of tobacco cultivation regions using near-infrared spectroscopy sensors,” *Infr. Phys. Technol.*, vol. 111, Dec. 2020, Art. no. 103494.

[16] Y. Li, “Research and application of deep learning in image recognition,” in *Proc. IEEE 2nd Int. Conf. Power, Electron. Comput. Appl. (ICPECA)*, Jan. 2022, pp. 994–999.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[18] W. Rawat and Z. Wang, “Deep convolutional neural networks for image classification: A comprehensive review,” *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017.

[19] A. Kamilaris and F. X. Prenafeta-Boldú, “A review of the use of convolutional neural networks in agriculture,” *J. Agricult. Sci.*, vol. 156, no. 3, pp. 312–322, Apr. 2018.

[20] H. Lin, H. Chen, C. Yin, Q. Zhang, Z. Li, Y. Shi, and H. Men, “Lightweight residual convolutional neural network for soybean classification combined with electronic nose,” *IEEE Sensors J.*, vol. 22, no. 12, pp. 11463–11473, Jun. 2022.

[21] X. Zeng, Y. Miao, S. Ubaid, X. Gao, and S. Zhuang., “Detection and classification of bruises of pears based on thermal images,” *Postharvest Biol. Technol.*, vol. 161, Mar. 2020, Art. no. 111090.

[22] M. Dyrmann, H. Karstoft, and H. S. Midtiby, “Plant species classification using deep convolutional neural network,” *Biosyst. Eng.*, vol. 151, pp. 72–80, Nov. 2016.

[23] S. H. Lee, C. S. Chan, P. Wilkin, and P. Remagnino, “Deep-plant: Plant identification with convolutional neural networks,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 452–456.

[24] S. Bargoti and J. Underwood, “Deep fruit detection in orchards,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 3626–3633.

[25] K. Kuwata and R. Shibasaki, “Estimating crop yields with deep learning and remotely sensed data,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 858–861.

[26] X. Song, G. Zhang, F. Liu, D. Li, Y. Zhao, and J. Yang, ''Modeling spatio-temporal distribution of soil moisture by deep learning-based cellular automata model,'' *J. Arid Land*, vol. 8, no. 5, pp. 734–748, Oct. 2016.

[27] T. Lin, Y. Wang, X. Liu, and X. Qiu, ''A survey of transformers,'' *AI Open*, vol. 3, pp. 111–132, Nov. 2022.

[28] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ''An image is worth 16×16 words: Transformers for image recognition at scale,'' 2020, *arXiv:2010.11929*.

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, ''Swin transformer: Hierarchical vision transformer using shifted windows,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[31] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, and Q. Tian, ''Visformer: The vision-friendly transformer,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 569–578.

[32] S. Mehta and M. Rastegari, ''MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer,'' 2021, *arXiv:2110.02178*.

[33] R. Gebbers and V. I. Adamchuk, ''Precision agriculture and food security,'' *Science*, vol. 327, no. 5967, pp. 828–831, Feb. 2010.

[34] G. Li, H. Zhen, F. Jiao, T. Hao, D. Wang, and K. Ni, ''Research on tobacco leaf grading algorithm based on transfer learning,'' in *Proc. IEEE Int. Conf. Artif. Intell. Comput. Appl. (ICAICA)*, Jun. 2021, pp. 32–35.

[35] D. Kurt, ''Impacts of environmental variations on quality and chemical contents of oriental tobacco,'' *Contrib. Tobacco Nicotine Res.*, vol. 30, no. 1, pp. 50–62, Mar. 2021.

[36] T. N. Wolf, S. Pölsterl, and C. Wachinger, ''DAFT: A universal module to interweave tabular data and 3D images in CNNs,'' *NeuroImage*, vol. 260, Oct. 2022, Art. no. 119505.

[37] A. Prakash, K. Chitta, and A. Geiger, ''Multi-modal fusion transformer for end-to-end autonomous driving,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7073–7083.

[38] Z. Tang, L. Chen, Z. Chen, Y. Fu, X. Sun, B. Wang, and T. Xia, ''Climatic factors determine the yield and quality of honghe flue-cured tobacco,'' *Sci. Rep.*, vol. 10, no. 1, Nov. 2020.

[39] G. Li, H. Zhen, D. Wang, and C. Wang, ''Review of tobacco leaf classification research based on artificial intelligence,'' in *Proc. Int. Conf. Culture-Oriented Sci. Technol. (ICCST)*, Oct. 2020, pp. 413–416.

[40] L. Jian, X. Yang, Z. Liu, G. Jeon, M. Gao, and D. Chisholm, ''SEDRFuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion,'' *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–15, 2021.

[41] L. Huang and Y. Chen, ''Dual-path Siamese CNN for hyperspectral image classification with limited training samples,'' *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 518–522, Mar. 2021.

[42] 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.

[43] I. Sobh, L. Amin, S. Abdelkarim, K. Elmadawy, M. Saeed, O. Abdeltawab, M. E. Gamal, and A. E. Sallab, ''End-to-end multi-modal sensors fusion system for urban automated driving,'' in *Proc. Neural Inf. Process. Syst. Workshops*, 2018, pp. 1–9.

[44] 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.

[45] *Flue-Cured Tobacco*, Standards Press China, Nat. Tobacco Monopoly Bureau, Beijing, China, 1992.

[46] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, ''Efficient low-rank multimodal fusion with modality-specific factors,'' in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 2247–2256.

[47] G. Sahu and O. Vechtomova, ''Dynamic fusion for multimodal data,'' 2019, *arXiv:1911.03821*.

[48] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, ''A ConvNet for the 2020s,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11966–11976.

[49] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, ''YOLOv4: Optimal speed and accuracy of object detection,'' 2020, *arXiv:2004.10934*.

[50] M. Tan and Q. Le, ''EfficientNet: Rethinking model scaling for convolutional neural networks,'' in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, ''Rethinking the inception architecture for computer vision,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[52] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*.

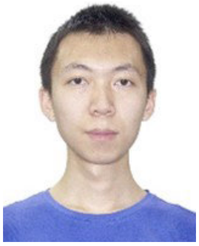
[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ''ImageNet: A large-scale hierarchical image database,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[54] Z. Tang, J. Yang, Z. Li, and F. Qi, ''Grape disease image classification based on lightweight convolution neural networks and channelwise attention,'' *Comput. Electron. Agricult.*, vol. 178, Nov. 2020, Art. no. 105735.

[55] A. Nasiri, A. Taheri-Garavand, and Y.-D. Zhang, ''Image-based deep learning automated sorting of date fruit,'' *Postharvest Biol. Technol.*, vol. 153, pp. 133–141, Jul. 2019.

QINGLIN HE received the B.E. degree from South China Normal University, in 2022. He is currently pursuing the master's degree with the College of Information Science and Electronic Engineering, Zhejiang University, China. His research interests include deep learning and image processing.

XIAOBING ZHANG received the master's degree from Henan Agricultural University, in 2008. He is currently with the Technology Center of China Tobacco Zhejiang Industry Company Ltd. His research interests include tobacco leaf evaluation and cultivation techniques.



JIANXIN HU received the B.E. degree from Jilin University, in 2021. He is currently pursuing the master's degree with the College of Information Science and Electronic Engineering, Zhejiang University, China. His research interests include deep learning and computer vision.

ZEHUA SHENG received the B.E. degree from Zhejiang University, China, in 2017, where he is currently pursuing the Ph.D. degree with the College of Information Science and Electronic Engineering. His research interests include image denoising and multimodal image processing.

QI LI received the master's degree from South China Agricultural University, in 2009. He is currently with the Technology Center of China Tobacco Zhejiang Industry Company Ltd. His research interests include agriculture material and product development.

SI-YUAN CAO received the B.E. degree in electronic information engineering from Tianjin University, in 2016, and the Ph.D. degree in electronic science and technology from Zhejiang University, China, in 2022. He is currently an Assistant Researcher with the Ningbo Innovation Center, Zhejiang University. His research interests include multispectral/multimodal image registration, homography estimation, place recognition, and image processing.

HUI-LIANG SHEN (Member, IEEE) received the B.E. and Ph.D. degrees in electronic engineering from Zhejiang University, Hangzhou, China, in 1996 and 2002, respectively. From 2001 to 2005, he was a Research Associate and a Research Fellow with The Hong Kong Polytechnic University, Hong Kong. He is currently a Professor with the College of Information Science and Electronic Engineering, Zhejiang University. His research interests include multispectral imaging, image processing, computer vision, and machine learning.