IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 6, December, 2023, pp. 2353-2365
Journal homepage: http://iieta.org/journals/ts

Enhanced Cigarette Pack Counting via Image Enhancement Techniques and Advanced SAFECount Methodology

Yanghua Gao*, Zhenzhen Xu, Xue Xu

Information Center, China Tobacco Zhejiang Industrial Co., Ltd., Hangzhou 310008, China

Corresponding Author Email: yhgao@zju.edu.cn

https://doi.org/10.18280/ts.400603

Received: 15 July 2023
Revised: 2 October 2023
Accepted: 27 October 2023
Available online: 30 December 2023

Keywords:
image enhancement techniques, cigarette pack counting, Similarity-Aware Feature Enhancement block for object Counting (SAFECount), Few-shot learning (FSC), warehouse inventory management

ABSTRACT

In the realm of cigarette pack counting systems, prevalent challenges persist, notably the low accuracy in count, limited adaptability to intricate scenes and varying environments, and a lack of responsiveness to diverse pack types and shapes. This study introduces an advanced method for cigarette pack counting, leveraging a combination of various image enhancement techniques and an improved Similarity-Aware Feature Enhancement block for object Counting (SAFECount) approach. The methodology comprises three integral modules: an image enhancement module, a feature extraction module, and a counting module. The image enhancement module, tasked with noise reduction and deblurring, ensures targeted enhancement effects on cigarette box images. To contend with the rapid shifts in cigarette pack appearances, this research integrates specialized color and boundary feature extraction networks with the SAFECount method. This integration facilitates the fusion of multi-scale, key semantic information, thus amplifying the model's detection efficacy. Addressing the scalability limitations prevalent in general models, the study employs a few-shot counting (FSC) approach, which endows the model with essential generalization and flexibility, requisite for practical applications, even with a minimal training dataset. Empirical analyses, conducted using actual data from the Zhongyan Corporation's cigarette pack dataset, substantiate the superiority of the proposed method in real-world warehouse environments. The method demonstrates a marked improvement in counting performance, evidenced by a Mean Absolute Error (MAE) of 1.71 and a Root Mean Square Error (RMSE) of 1.95.

1. INTRODUCTION

The evolution of AI technology has ushered in a new era of applications and advancements, particularly in the domain of counting. Deep learning methods for counting have been effectively utilized across various fields. In urban areas, these methods have been instrumental for crowd monitoring and counting, as observed in busy streets, stations, and stadiums [1-6]. Such applications provide critical insights into crowd dynamics and behaviors. In agriculture, deep learning has been employed for crop quantity estimation, facilitating precision in agricultural management [7-12].

The realm of cigarette pack counting has not been exempt from the influence of these advanced methodologies. Accurate counting of cigarette packs is a pivotal component of industrial production, ensuring the efficiency of production processes. The autonomous nature of deep learning methods enables automated counting in high-activity production environments, thereby obviating the need for manual counting. This automation significantly augments both accuracy and efficiency. Furthermore, these methods exhibit remarkable adaptability in complex scenarios, such as dealing with stacked or obscured cigarette boxes, thereby bolstering production line management.

In the broader context of image-based counting, deep neural networks have become a cornerstone for object counting. Predominantly, these methods are categorized into two types [13]: detection and segmentation-based counting methods, and density estimation-based counting methods. Detection and segmentation-based methods incorporate object detectors and segmentation modules within counting systems. These methods identify objects in an image and proceed to count them, with Mask-RCNN [14] and RetinaNet [15] being notable examples. While offering high accuracy and comprehensive information, these methods are often hampered by higher computational demands and sensitivity to variations in target objects. Conversely, density estimation-based methods approach counting as a regression task, generating a heat map correlated to the image to estimate target object quantity, followed by count regression from the heatmap. LC-DenseFCN [16] and CentroidNet [17] exemplify this category. Characterized by simpler structures and higher computational efficiency, these methods demonstrate greater adaptability to changes in target objects compared to their detection and segmentation-based counterparts. However, applying these methods to cigarette pack counting in warehouse environments presents unique challenges. Firstly, images from cigarette warehouses often suffer from noise interference due to suboptimal lighting conditions or sensor

noise. Secondly, motion blur may occur due to vehicle movement during loading processes. The variable appearance of cigarette packs, influenced by different batches or design changes, further complicates the counting task. Moreover, conventional methods, typically trained on fixed datasets, struggle to adapt to evolving cigarette pack designs or promotional variations, leading to diminished counting performance when applied directly to real-world cigarette pack counting scenarios.

To confront these challenges, a plethora of researchers have embarked on refining counting methodologies. A prevalent strategy, data enhancement, has been extensively employed to mitigate image noise and blur [18-25]. In this vein, Zhang et al. [26] introduced a novel approach, Noise2Noise, which interprets annotations as noise. This method establishes a self-supervised pre-training task, employing image enhancement within the Noise2Noise framework. Additionally, Waqas Zamir et al. [27] developed the Restormer model, a robust solution for image denoising. This model integrates a Multi-DCONV Transposed Attention (MDTA) module with a Gate DCONV Feed-forward Network (GDFN), thereby accentuating local spatial contexts and implicitly modeling global pixel relationships. In response to the variable appearances of tobacco packs, the FSC method has gained traction [28-32]. This innovative approach enables users to identify target objects using merely one or a few support images. Significantly, this method does not necessitate an extensive array of training samples for the test objects. Providing the support images are available, a proficiently trained model can adeptly conduct inference on new classes. Within this context, Lu et al. [29] proposed the Graph Matching Network (GMN) model. This model synergizes support features with query features, followed by the learning of regression heads for point-wise feature comparison. However, this approach's comparative effectiveness is relatively diminished when contrasted with using similarity for comparison. Yang et al. [31] introduced the CFOCNet (Coupled Feature-Offset Comparison Network) model, which initiates feature comparison with point generation, subsequently regressing density maps from the derived similarity maps. Ranjan et al. [30] developed the FamNet (Few-shot Adaptive Metric Network) model, which augments the reliability of similarity maps through multi-scale augmentation and test-time adaptation. However, the informational capacity of these similarity maps is somewhat limited. Consequently, regression from these maps may not yield optimal results, particularly in scenarios where objects are densely arranged. Numerous studies have endeavored to enhance general methods from various angles, aiming to tailor them more effectively to the task of cigarette pack counting in real-world settings. Despite these efforts, a comprehensive solution framework that effectively addresses all the aforementioned challenges in their entirety remains elusive.

To address the challenges identified, an innovative two-step solution is proposed. Initially, original images undergo processing through an image enhancement module. Subsequently, an enhanced version of the SAFECount method [32] is employed for the counting process. The first challenge, involving denoising and deblurring, is tackled through the application of advanced image enhancement techniques. In dealing with the frequent variations in the appearance of cigarette packs, it is suggested to amalgamate specialized color and boundary feature extraction networks with the SAFECount method. This combination is key to effectively managing these variations, thereby improving the accuracy of the counting process. To overcome the scalability issues inherent in general counting models, the adoption of a FSC approach is proposed. This method facilitates rapid learning with a minimal number of samples, enabling the model to accurately predict unknown classes based on the knowledge of known classes. This approach significantly enhances the model's generalization and scalability capabilities. The selected representative method for this purpose is the improved SAFECount method. This method incorporates a multi-scale feature aggregation mechanism, enhancing its flexibility and adaptability. By updating and adjusting the model with a small number of support samples, it can swiftly adapt to changes in the content of cigarette packs and perform accurate counting. Empirical studies conducted using the Zhongyan cigarette pack dataset reveal that the proposed method excels in practical engineering applications. It demonstrates superior counting accuracy and scalability compared to existing methods. These experimental findings corroborate the efficacy and reliability of the proposed approach in addressing the challenges of cigarette pack counting tasks.

The contributions of this research are multifaceted and significant:

(1) Image Enhancement: Central to this study is the development of an image enhancement module, comprising two key components: denoising and deblurring. The denoising component substantially reduces image noise, thereby augmenting clarity and detail visibility. Concurrently, the deblurring component addresses image blur, a consequence of camera or object movement, resulting in significantly clearer images. These enhancements collectively bolster object detection capabilities and consequently, improve counting accuracy.

(2) FSC: A novel FSC approach is introduced, enabling swift learning and generalization with a limited number of samples. This methodology allows for efficient adaptation to various tasks and domains. It offers remarkable flexibility and adaptability, significantly reducing manpower requirements and enhancing learning efficiency.

(3) Enhanced SAFECount Method: This research extends the SAFECount method by integrating color semantic information and boundary semantic information into the model. This integration markedly improves the model's detection efficiency and accuracy, positioning it at the forefront of counting technologies.

(4) End-to-End Optimization Approach: A unique aspect of this study is the development of an end-to-end optimization approach. This approach aligns the objectives of image enhancement and counting, facilitating a cohesive and synergistic improvement in detection performance.

2. RELATED WORKS

The SAFECount method [32] represents a significant advancement in feature enhancement for object counting. It primarily focuses on enhancing counting features by analyzing the similarity between query images and reference images.

Historically, FSC solutions have revolved around the representation of sample objects, namely support images, and query images using expressive features. These features are then scrutinized for feature correlations to determine candidate objects. In the context of few-shot learning, 'support' and

'query' are pivotal concepts. During training, datasets are bifurcated into support sets and query sets. The support set comprises a limited number of labeled samples, while the query set contains unlabeled samples from the same class space as the support set. Within this framework, 'support images' denote samples within the support set, and 'support features' are the extracted features from these images. Similarly, 'query images' and 'query features' pertain to the samples and their respective features in the query set.

FSC methods are generally divided into two categories. The first is feature-based, where pooled support features are concatenated with query features. A regression head is then utilized to ascertain if the combined features closely align. A critical drawback of this approach is the pooling step, which tends to overlook spatial information from the support images, leading to the acquisition of less reliable features. The second category is similarity-based, involving the use of original features to create similarity maps for regression analysis. However, the informational content conveyed by similarity maps is considerably lower compared to direct features, posing challenges in accurately discerning clear object boundaries. This limitation becomes particularly pronounced in scenarios where target objects are densely packed, as often encountered in practical applications like the dense arrangement of cigarette boxes. In such cases, the counting performance of these methods can be severely compromised, illustrated in Figure 1.

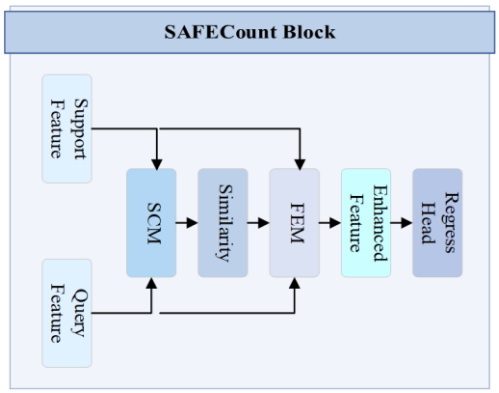


Figure 1. Schematic diagram of the SAFECount module

The SAFECount module, a pivotal component in the proposed methodology, is composed of two integral parts: the Similarity Comparison Module (SCM) and the Feature Enhancement Module (FEM). The SCM initiates the process by projecting the features of both support and query images into a unified contrastive space. This is achieved using shared convolutions and layer normalization techniques. Following this, the SCM employs the support feature as a convolutional kernel, executing a sliding operation over the query feature to compute a score map. This score map undergoes normalization through two distinct processes: exemplar norm and spatial norm. Exemplar norm is responsible for normalization along the support dimension, while spatial norm focuses on the spatial dimensions, collectively resulting in a comprehensive similarity map. In the subsequent stage, the Feature Enhancement Module takes precedence. Here, the similarity map serves as a set of weights for conducting weighted computations on the support features. To preserve spatial consistency, the support features are inverted and then utilized as a convolutional kernel in conjunction with the similarity map. This convolution produces a feature that is enhanced by similarity weighting. The next step involves merging this similarity-weighted feature with the query feature, culminating in an enriched feature representation. The culmination of this process is the employment of a regression head. This head is tasked with predicting the density map derived from the enhanced feature representation. The intricate design of the SAFECount module, encompassing the SCM and FEM, exemplifies the sophisticated approach taken to enhance feature representation and accuracy in object counting tasks, particularly in challenging scenarios such as counting densely packed objects. This module represents a significant stride forward in addressing the complexities associated with feature enhancement in counting methodologies.

SAFECount, with its innovative approach, offers several key advantages:

(1) Integration of feature extraction and similarity analysis: SAFECount uniquely combines the strengths of feature extraction and similarity analysis through the Similarity-Aware Feature Enhancement (SAFE) block. This integration ensures that the enhanced features not only encapsulate rich semantic information but also accurately identify areas in the query image analogous to the support image. This dual capability significantly boosts the accuracy of counting.

(2) Few-shot learning approach: The model adopts a few-shot learning strategy, enabling the characterization of objects of interest with just a few support images, without the necessity of prior knowledge about the object types. This feature imparts remarkable flexibility and scalability to SAFECount, making it adept at adapting to diverse and rapidly evolving counting scenarios.

(3) Multi-Block structure: SAFECount incorporates a Multi-Block structure, allowing for the reintegration of enhanced features as query features back into the module. Adjusting the number of these blocks can lead to further enhancements in model accuracy.

(4) End-to-end framework: The framework of SAFECount is designed to be end-to-end, efficiently extracting features directly from raw images and conducting counting predictions. This eliminates the requirement for any additional prior knowledge or manual preprocessing, streamlining the counting process and improving efficiency.

While SAFECount has exhibited commendable performance across various domains such as crowd monitoring, product counting, and crop estimation, its direct application to cigarette pack counting presents certain challenges. Primarily, images pertinent to cigarette pack counting frequently exhibit issues such as noise, blurriness, and complex backgrounds. These factors can significantly impede the accuracy of counting. Additionally, the extensive variety of cigarette pack types, each with its unique appearance, coupled with the rapid changes occurring in cigarette warehouses, poses a challenge to the model's ability to accurately recognize and count packs. Consequently, it becomes imperative to tailor and refine the SAFECount model, ensuring it is suitably adapted to meet the specific demands and nuances of cigarette pack counting.

3. METHOD

3.1 Overall framework

The methodology proposed in this study encompasses a comprehensive framework for estimating cigarette pack

quantities, incorporating three primary modules: a data enhancement module, a feature extraction network, and a counting module, detailed in Figure 2.

Initially, the data enhancement module is deployed to address potential noise interference and motion blur encountered during the capture of cigarette pack images. This is achieved through the integration of two sub-modules dedicated to the sequential performance of denoising and deblurring tasks on the images.

Subsequently, the feature extraction network focuses on extracting color semantic information and boundary semantic information from the target objects. This network comprises two sub-modules responsible for generating feature maps containing color and boundary features. These maps are then fused and subsequently utilized as input for the counting module.

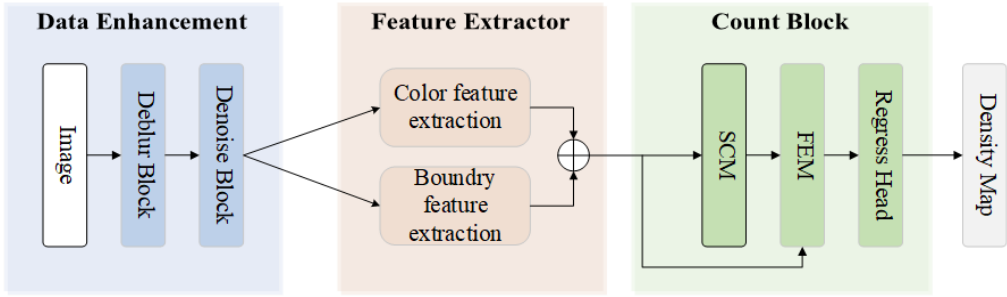


Figure 2. Schematic diagram of cigarette pack counting based on various image enhancement techniques and improved SAFECount

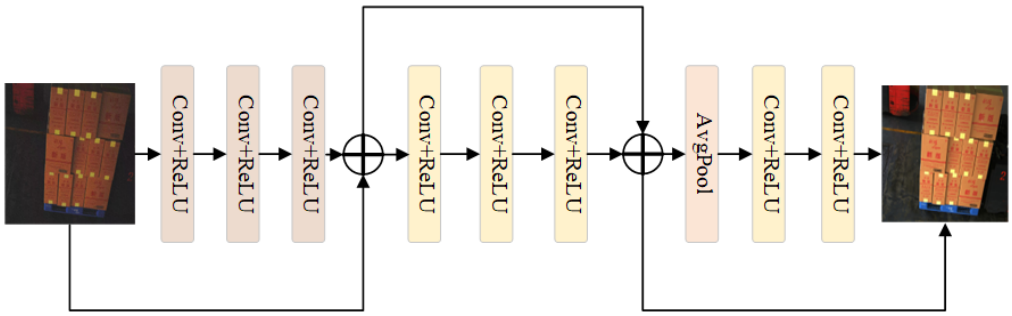


Figure 3. Schematic diagram of the denoising module

The counting module, employing the SAFECount method, is pivotal in enhancing regression features by using similarity as a guiding principle. The module integrates the SCM and the FEM. These modules enable the refined query features to concentrate more effectively on areas similar to exemplar objects, as defined by support images, thus facilitating enhanced accuracy in counting.

Further sections of this paper will delve into detailed descriptions of each component within the proposed model, including the image enhancement module, the feature extraction network, and the counting module.

3.2 Data enhancement

Within the data enhancement module, two distinct sub-modules are operational. The initial sub-module is dedicated to denoising, and this paper introduces a single-stage blind denoising network algorithm, underpinned by a feature attention mechanism, designed to refine noisy cigarette pack images. The architecture of this network comprises a feature extraction module, a feature learning module based on a residual structure, and an image reconstruction module, with a total of eight convolutional layers, as shown in Figure 3.

The process begins with feature extraction, executed through three dilated convolutional layers. This approach expands the receptive field, thereby capturing a more extensive range of information. Following this, feature learning is facilitated using residual blocks comprising three convolutional layers, adept at detecting nuanced differences within the image. This step significantly enhances the feature representation capability of the network.

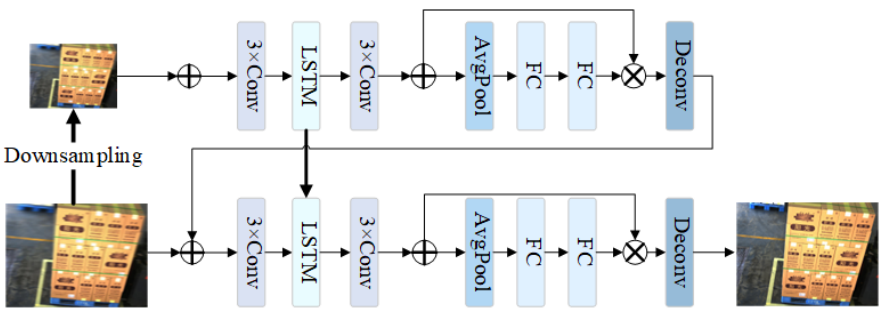


Figure 4. Schematic diagram of the deblurring module

The final stage involves image reconstruction, employing a residual network augmented with a feature attention mechanism, which is generated through convolutional processes. This mechanism is instrumental in automatically identifying and focusing on pivotal image features, thereby diminishing the impact of noise and augmenting the quality of the denoising results. Crucially, this is achieved while maintaining the integrity of image details.

The meticulously structured network design is thus proficient in accomplishing blind denoising and enhancement of cigarette pack images. It is characterized by a reduced parameter count and efficient computational capacity, rendering it suitable for real-time denoising applications. This innovative approach represents a significant advancement in the field of image processing, particularly in contexts demanding high-quality denoising under stringent computational constraints, as shown in Figure 4.

The second one is the deblurring module. This module is inspired by a successful network structure frequently utilized in image deblurring, known as the image pyramid structure. Utilizing this structure, the feature extraction layers progressively restore a clear image, thereby achieving blind image deblurring. The architecture of this network adheres to a top-down encoder-decoder model, wherein the reconstruction of the clear image is progressively realized by integrating outputs from different scales of network layers. Comprising the network are two sub-networks, each mirroring the same structural composition: seven convolutional layers, two fully connected layers, and one Long Short-Term Memory (LSTM) layer. The process initiates with feature extraction executed through convolutional layers, followed by image restoration via deconvolutional layers. A "coarse-to-fine" strategy is central to this process. This approach involves a progressive reduction in the size of the source image, with the clear image being restored based on inputs at varying resolutions:

$$I_i, h_i = \mathrm{Net}_i(B_i, I_{i+1}\uparrow, h_{i+1}\uparrow) \tag{1}$$

where, i denotes the resolution index, i=1 corresponds to the original image resolution. $B_i$and C_i represent the blurry image and the restored clear image at the i-th resolution, respectively. Net_i signifies the i-th layer sub-network, while h_i indicates hidden features. The upsampling of parameters to match the network layers is denoted by the operator. The transformation of the input image into feature maps involves a reduction in dimensions but an increase in the number of channels, facilitated by a CNN network. To capture hidden features from low-resolution inputs, an LSTM network is integrated, thereby enhancing the expressive capacity of the feature maps. Additionally, an attention mechanism is implemented, dynamically adjusting the weights of the feature maps. This adjustment is responsive to variations in image content, further improving the deblurring effects. This sophisticated approach, combining the strengths of the image pyramid structure, LSTM networks, and attention mechanisms, significantly advances the capability of image deblurring, particularly in challenging scenarios such as those presented by cigarette pack images.

3.3 Feature extractor

The feature extraction network consists of two specialized subnetworks: one for color feature extraction and another for edge feature extraction. Support images and query images are processed through these subnetworks to extract pertinent features. The color feature extraction network isolates features rich in color information from the support images, while the edge feature extraction network focuses on extracting boundary-related features. These extracted features are then fused with the query features using an additive fusion technique. This process entails the pixel-wise addition of corresponding positions in the feature maps of both the support and query features. The resultant combined features, encompassing both color and boundary information, are subsequently utilized as inputs for the counting module, illustrated in Figure 5.

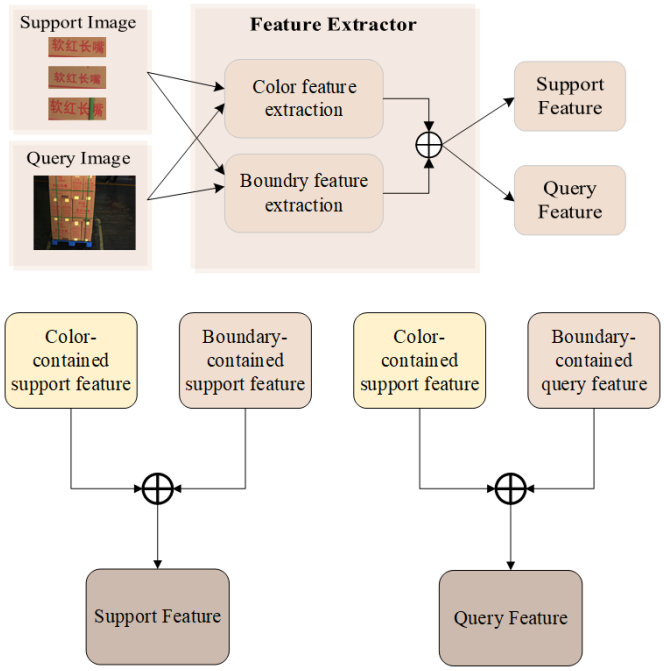


Figure 5. Schematic diagram of feature extraction

In the Color Feature Extraction process, feature extraction is executed in two distinct steps. The first step involves initially extracting color features through a convolutional network, while the second step utilizes self-attention mechanisms to enhance the focus on color features, thereby improving cigarette pack recognition and overall performance, detailed in Figure 6.

During the first step, the input image is processed through two convolutional layers, followed by a ReLU activation function and a max-pooling layer, to extract low-level features. A deconvolutional layer is then employed to upscale the low-level feature map back to the original input image size. These low-level features undergo normalization using a SoftMax function, ensuring the feature values are confined within a specific range. The normalized low-level features are further processed through color mapping via a fully connected layer, aligning the feature values with the color space. The resultant mapped feature map is then element-wise multiplied with the original image, infusing color information into the image.

In the second step, the color-weighted image from the first step is further refined. Initially, it is passed through two convolutional layers to generate a feature map rich in color semantic information. This is followed by the application of three 3x3 convolutional kernels with ReLU activation functions, mapping the input feature map into spatial representations for Query, Key, and Value features. Point-wise

multiplication between the Query and Key features generates a matrix, indicative of the similarity between queries and keys. This similarity matrix undergoes a Softmax operation to derive attention weights, where each element reflects the relative importance of its position. Finally, these attention weights are matrix-multiplied with the Value features, producing a feature representation attentively weighted at each position. This methodical approach ensures a detailed and focused enhancement of color features, crucial for the accurate recognition of cigarette packs.

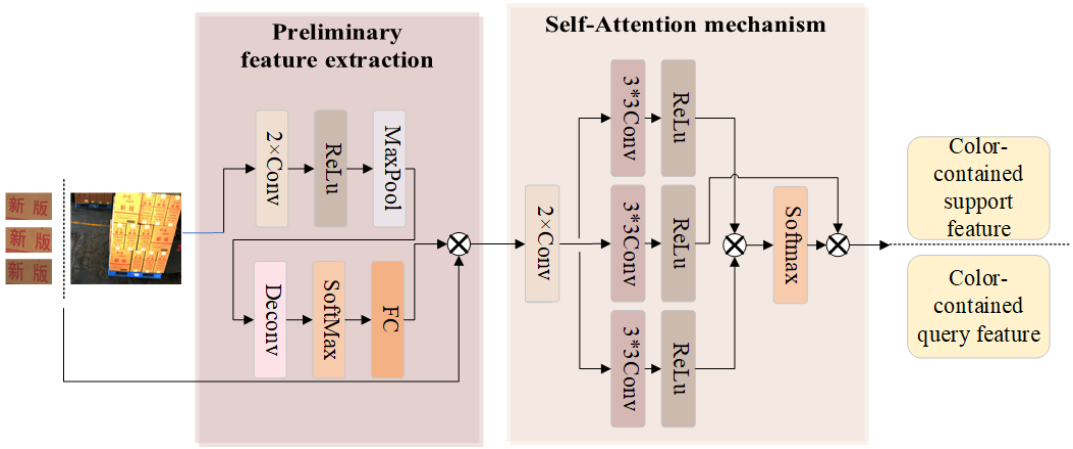


Figure 6. Schematic diagram of the color feature extraction

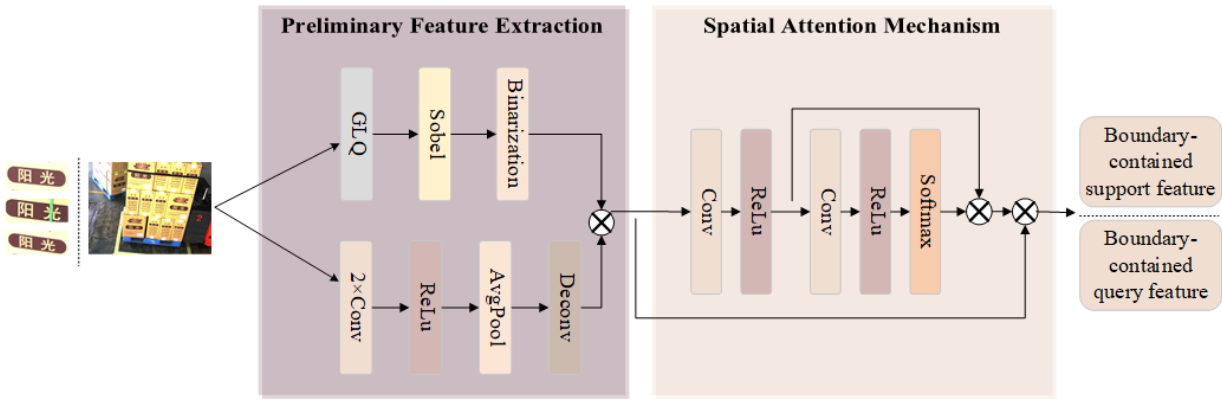


Figure 7. Schematic diagram of the boundary feature extraction

In the Boundary Feature Extraction process, feature extraction is conducted in two stages: the initial stage involves the preliminary extraction of boundary features using convolutional networks and the Sobel operator, while the second stage employs spatial attention mechanisms to intensify the focus on boundary features, thus enhancing the recognition capabilities for cigarette packs and overall technical performance, detailed in Figure 7.

The first stage commences with the input image being processed through two convolutional layers, a ReLU activation function, and an average pooling layer to extract low-level features. A deconvolutional layer then restores the size of the low-level feature map to that of the original input image. In parallel, the input image is subjected to grayscale quantization, converting it into a grayscale image. The Sobel operator is subsequently applied via convolution operations in both horizontal and vertical directions to obtain response values for these directions. By calculating the gradient magnitude and angle for each pixel, edge intensity and direction are determined. The resulting edge intensity is then thresholded, producing a binary edge map that contains crucial edge information.

In the second stage, the newly formed feature map first undergoes processing through a convolutional layer and a ReLU function for advanced feature mapping, leading to a more abstract representation of boundary features. Following this, convolutional attention mechanisms are applied, resulting in an attention weight map. This map is computed using a two-dimensional convolution function, further refined with a ReLU function to introduce non-linear transformations, and finally normalized with a Softmax function to maintain attention weights between 0 and 1. These attention weights are then element-wise multiplied with the feature map, yielding a feature representation attentively weighted at each position. Matrix multiplication is the final step, fusing the attention-weighted feature map with the original feature map, thereby enhancing the emphasis on boundary features in the final feature map.

3.4 Count block

In the Counting Block, the SAFECount method, specifically designed for object counting, is utilized. Both support and query features are information-rich, and the utilization of similarity offers a more effective means of capturing the support-query relationship. This module capitalizes on this aspect to enhance regression features, integrating the advantages of both support and query features. Moreover, color semantic information and boundary semantic information are amalgamated into the Support and Query

Features. This integration is crucial, as it enables the extraction of each cigarette pack's color through the color information and determination of their positions through boundary information, thereby facilitating a more precise extraction of target cigarette packs from the image. Ultimately, the enhanced features not only embody rich semantic information extracted from the image but also accurately identify regions in the query image analogous to the exemplar objects, as shown in Figure 8.

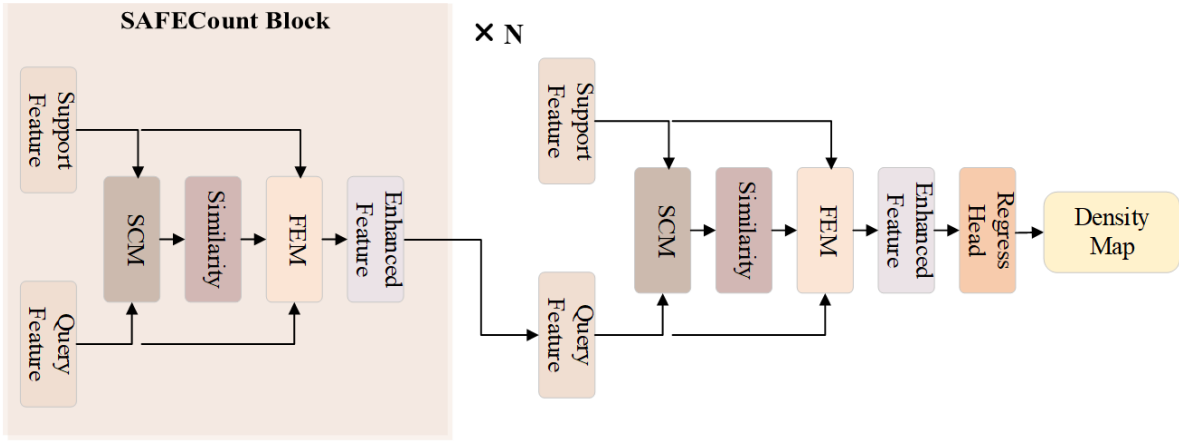


Figure 8. Schematic diagram of the counting block

To realize this, a SCM and a FEM are proposed, as depicted in the accompanying figure. The SCM extends beyond basic feature comparisons by learning feature projections and executing comparisons on these projected features to derive a score map. This process is formulated as follows:

$$\boldsymbol{R_0} = \text{conv}\left(g(\boldsymbol{f_Q}), \text{kernel } = g(\boldsymbol{f_S})\right) \tag{2}$$

where, $\boldsymbol{R_0}$ represents the score map, $\boldsymbol{f_Q}$ stands for the query feature, $\boldsymbol{f_S}$ represents the support feature, and $g()$ denotes the feature projection, achieved through a 1×1 convolutional layer followed by layer normalization. This design aids in selecting the most relevant information for object counting from the features. After the comparison step, the score maps associated with all support images (few samples) are aggregated. This aggregation is followed by normalization along two dimensions: the exemplar dimension and the spatial dimension. The normalization process is essential for generating a reliable similarity map, which is instrumental in the counting process. The computation of the similarity map is defined by the following formula:

$$\boldsymbol{R} = \boldsymbol{R_{EN}} \otimes \boldsymbol{R_{SN}} \tag{3}$$

where, $\boldsymbol{R}$ represents the similarity map, $\boldsymbol{R_{EN}}$ denotes the exemplar normalization result, and $\boldsymbol{R_{SN}}$ signifies the spatial normalization result. The formula for the exemplar normalization result, $\boldsymbol{R_{EN}}$, is as follows:

$$\boldsymbol{R_{EN}} = \text{softmax}\, dim = 0\left(\frac{\boldsymbol{R_0}}{\sqrt{H_S W_S C}}\right) \tag{4}$$

The formula for the spatial normalization result, $\boldsymbol{R_{SN}}$, is as follows:

$$\boldsymbol{R_{SN}} = \frac{\exp\left(\frac{\boldsymbol{R_0}}{\sqrt{(H_S W_S C)}}\right)}{\underset{\text{dim=(2,3)}}{max}\left(\exp\left(\frac{\boldsymbol{R_0}}{\sqrt{H_S W_S C}}\right)\right)} \tag{5}$$

The $\underset{\text{dim=()}}{max}$ function in the equation represents finding the maximum value along the specified dimension. Furthermore, FEM utilizes point-to-point similarity as weighting coefficients to fuse support features into query features. The implementation steps are:

In the first step, aggregation of $\boldsymbol{f_S}$ into $\boldsymbol{f'_R}$ is achieved through the similarity $\boldsymbol{R}$:

$$\boldsymbol{f'_R} = \text{conv}(\boldsymbol{R}, \text{kernel } = \text{flip}(\boldsymbol{f_S})) \tag{6}$$

In the second step, the sum_{dim} function is employed to accumulate $\boldsymbol{f'_R}$ along a specific dimension, resulting in the similarity-weighted feature $\boldsymbol{f_R}$:

$$\boldsymbol{f_R} = \text{sum}_{\text{dim=0}}(\boldsymbol{f'_R}) \tag{7}$$

Finally, $\boldsymbol{f_R}$ is efficiently fused into $\boldsymbol{f_Q}$ through a network, yielding the enhanced feature $\boldsymbol{f'_Q}$:

$$\boldsymbol{f'_Q} = \text{layer\_norm}\left(\boldsymbol{f_Q} + h(\boldsymbol{f_R})\right) \tag{8}$$

The fusion process within the SAFECount method ensures that the enhanced query feature intensifies its focus on areas resembling the exemplar objects as defined by the support images, thus facilitating more accurate counting. The integration of the SCM and FEM allows for the optimal utilization of similarity information to refine the accuracy of feature representation. Consequently, the enhanced features are enriched with more profound semantic information and are adept at prioritizing areas analogous to exemplar objects, guided by the support images. This leads to a notable improvement in counting precision.

In this study, the number of SAFECount blocks employed is three. Experiments were conducted to determine the optimal number of basic blocks. It was observed that as the number of SAFECount blocks increased from one to three, there was a corresponding enhancement in the model's technical performance. However, upon increasing the number of blocks to four, a decline in performance was noted. This decrease in performance can be attributed to the model becoming

excessively complex with the additional blocks, leading to overfitting issues. The specific outcomes of these experiments are detailed in Table 1.

Table 1. Number of SAFECount block

Block Number	MAE	RMSE
1	1.83	2.20
2	1.79	2.11
3	1.71	1.95
4	1.78	2.08

4. LOSS FUNCTION

The Loss Function of the entire cigarette pack counting system consists of three parts: $L_{Denoise}$ stands for the denoising module, L_{Deblur} stands for the deblurring module, and L_{Count} stands for the counting module. The overall Loss is formulated as follows:

$$L_{Total} = L_{Denoise} + L_{Deblur} + L_{Count} \tag{9}$$

The following are the Loss Functions for each individual component:

4.1 Denoising module

The denoising submodule employs a pre-training dataset where certain cigarette pack images are artificially augmented with three types of noise: Gaussian noise, impulse noise, and uniform noise. This approach creates complex noise patterns in cigarette pack images, each having distinct characteristics.

Gaussian Noise: A prevalent type of random noise, is characterized by amplitudes that follow a Gaussian distribution. Its implementation is relatively straightforward, and the probability density function of Gaussian noise can be mathematically represented. This type of noise typically contributes a 'grainy' appearance to images, and it can be written as:

$$P(z) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{10}$$

Impulse Noise: Impulse noise, also known as salt-and-pepper noise, is a type of interference that appears abruptly in data. It is akin to sudden, burst-like spikes of high-energy noise and often manifests as discrete white or black spots in images. The probability density function of impulse noise is:

$$P(z) = \begin{cases} P_a, z = a \\ P_b, z = b \\ 1 - P_a - P_b, \text{else} \end{cases} \tag{11}$$

Uniform Noise: Uniform noise is a type of random interference wherein noise values at each sampling point in the dataset follow a uniform distribution. This noise often emerges due to systematic issues or imperfections during the image capture or generation process. The probability density function of uniform noise is:

$$P(z) = \begin{cases} \frac{1}{b-a}, a \le z \le b \\ 0, \text{else} \end{cases} \tag{12}$$

In the pre-training phase, the L1 loss function is employed to quantify the deviation between the network's output image and the original noise-free image. The L1 loss function is advantageous for this purpose due to its ability to effectively measure absolute differences. The minimization loss function required in this process is formulated as:

$$L_{Denoise} = \frac{1}{N} \sum_{i=1}^{N} \| y(x_i) - y_i^* \|_1 \tag{13}$$

where:

-x_i is the input noisy image.

-$y(x_i)$ is the output of the blind denoising network.

-y_i^* is the ground truth for x_i, i.e., the noise-free image.

-N is the number of paired noisy images and their corresponding noise-free images within a batch during training.

The L1 loss measures the absolute difference between the network's output image and the noise-free image, allowing the network to better capture noise characteristics and distribution. This improves the accuracy and robustness of the denoising effect.

4.2 Deblurring module

For the pretraining of the deblurring submodule, the dataset includes cigarette pack images ('smokebox images') that have been artificially subjected to motion blur. This blur is created by averaging consecutive short-exposure frames taken with a high-speed camera. Such a method is adept at producing images that closely mimic real-world scenarios, replicating the complex effects of camera shake and object motion often encountered in practical photography. This level of realism in the training dataset is crucial for ensuring that the deblurring module is well-equipped to handle real-world image degradation.

In the training process of the motion deblurring network, the L2 norm loss is employed to measure the difference between the network's output image and the ground truth image, i.e.:

$$L_{Deblur} = \sum_{i=1}^{n} \frac{1}{N_i} \| y^i - y_*^i \|_2^2 \tag{14}$$

where:

-y^i is the image after motion deblurring.

-y_i^* is the ground truth image.

-N_i is the batch size.

4.3 Counting block

In the pre-training phase for the counting module, the process involves a comparative analysis of the generated images against their respective ground truth counterparts. The MSE loss is utilized to quantify the differences between these sets of images. This loss function is:

$$L_{Count} = \frac{1}{H \times W} \| D - D_{GT} \|_2^2 \tag{15}$$

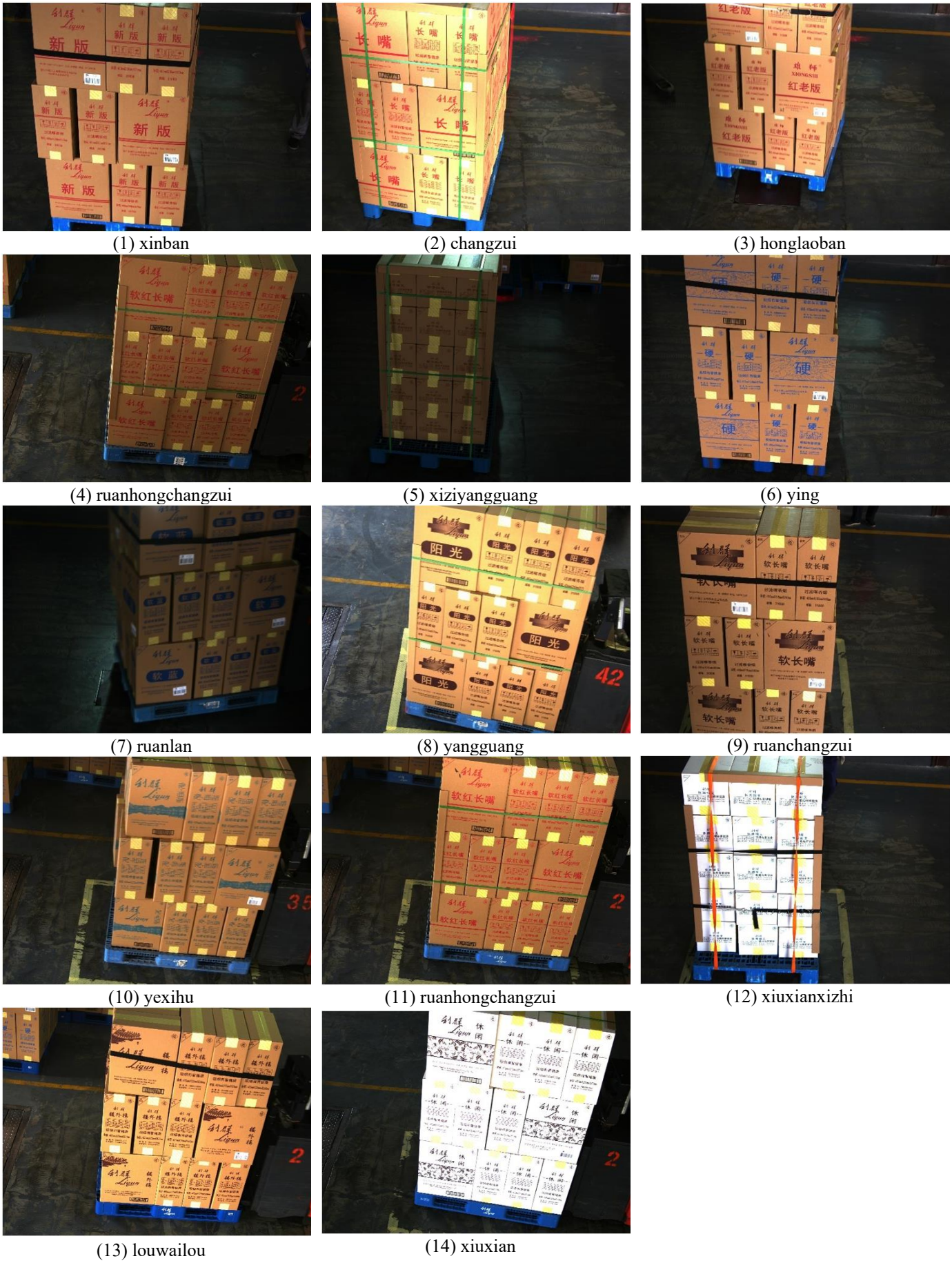
where:

-H is the height of the image.

-W is the width of the image.

-D is the real image.

-D_{GT} is the generated image.



(1) xinban (2) changzui (3) honglaoban

(4) ruanhongchangzui (5) xiziyangguang (6) ying

(7) ruanlan (8) yangguang (9) ruanchangzui

(10) yexihu (11) ruanhongchangzui (12) xiuxianxizhi

(13) louwailou (14) xiuxian

Figure 9. Actual images of various types of cigarette packs in our dataset

The task of counting cigarette packs demands high precision in accuracy. By employing the MSE loss, the model is enabled to precisely learn the variations and subtle differences in the quantities of the target objects. This approach ensures that the

model is finely tuned to recognize and accurately count the number of cigarette packs, which is critical for achieving reliable and accurate results in practical applications.

5. EXPERIMENTS

5.1 Dataset

The dataset utilized in this research is designated as the "Zhongyan cigarette pack dataset". This dataset categorizes cigarette packs into 14 distinct classes, named "xinban,""changzui,""honglaoban,""ruanhongchangzui,""xizi yangguang,""ying,""ruanlan,""yangguang,""yexihu,""ruanch angzui,""xiuxianxizhi,""xiuxian,""xihulian," and "louwailou." The dataset is characterized by a variation in the number of cigarette packs per image, ranging from 6 to 12, with an average of 10 packs per image.

For the purpose of this experiment, 150 images from each cigarette pack category were selected, ensuring a balanced representation of target objects across all classes. The dataset is divided into a training set, comprising 1200 images, and a validation set, consisting of 450 images. Additionally, there is a test set which also includes 450 images. Each image in the dataset is accompanied by 3 support images, which are used to describe the target object in detail.

The dataset's categorization for training, validation, and testing is distinct. The training set encompasses 8 classes, whereas the validation and test sets each contain 3 unique classes that are disjoint from each other, as shown in Figure 9.

In the Zhongyan cigarette pack dataset as originally compiled, the dataset splits and the number of support images are predetermined. However, typical FSC tasks often incorporate multiple dataset splits and a variable number of support images. To address this and to provide a comprehensive evaluation of the model proposed in this study, the Zhongyan cigarette pack dataset was augmented using cross-validation techniques. For this purpose, the various categories within the Zhongyan cigarette pack dataset were sequentially numbered from 0 to 13. Subsequently, all the images were divided into three distinct groups, with careful consideration to ensure that the categories within these groups did not overlap. Table 2 in the paper details the class indices, class labels, and the number of images assigned to each group. In the experimental setup, when a particular fold, designated as fold-i (where i=0, 1, or 2), is selected as the test set, the other two groups are combined to form the training set. This approach allows for a thorough and varied testing of the model's performance. Furthermore, the proposed method was evaluated under two different scenarios: 1-shot and 3-shot. In the 3-shot scenario, the original three support images from the dataset were utilized as per the original dataset configuration. Conversely, in the 1-shot scenario, to mimic a more challenging few-shot learning environment, a single support image was randomly chosen from the three available support images for each instance in the original dataset.

Table 2. Classification of datasets used for cross-validation experiments

Fold	Class Indices	Classes	Images
0	0-4	5	750
1	5-9	5	750
2	10-13	4	600

5.2 Metrics

MAE and RMSE are selected as the primary evaluation metrics for the model, and are calculated as follows:

$$MAE = \frac{1}{N_Q}\sum_{i=1}^{N_Q}\left|C^i - C_{GT}^i\right| \tag{16}$$

$$RMSE = \sqrt{\frac{1}{N_Q}\sum_{i=1}^{N_Q}\left(C^i - C_{GT}^i\right)^2} \tag{17}$$

where:

-N_Q is the number of query images.

-C^i is the predicted count for the i-th query image, which is the sum of densities of all pixels in the predicted density map.

-C_{GT}^i is the ground truth count for the i-th query image, which is the sum of densities of all pixels in the density map.

5.3 Comparative experiments

To further substantiate the efficacy of the proposed model in the context of cigarette pack counting tasks, a comparative experiment was conducted. This experiment juxtaposed our model against other established counting models, using the Zhongyan cigarette pack dataset as the testing ground. The results of various counting methods on this dataset are compiled in the following table:

Table 3. Test results of different methods on Zhongyan cigarette pack dataset

Method	*MAE*	*RMSE*
HLCNN	1.92	2.36
BMNet	2.15	3.08
CounTR	2.03	2.59
FamNet	3.15	4.71
OurModule	1.71	1.95

Upon examining Table 3, it becomes evident that the proposed counting model demonstrates superior performance overall. Notably, when compared to the current state-of-the-art CNN-based object counting model, HLCNN, our model shows a marked improvement, with a decrease of 0.21 in MAE and 0.41 in RMSE. Furthermore, in comparison to the leading Transformer-based object counting model, CounTR, our model exhibits enhanced performance, with a reduction of 0.32 in MAE and 0.64 in RMSE. Additionally, when pitted against the FamNet model, which also falls under the FSC (Feature Selection and Classification) category, our model achieves significant improvements, with reductions of 1.44 in MAE and 2.76 in RMSE. These results unequivocally demonstrate the robustness and accuracy of our proposed model, particularly in the challenging and complex task of counting cigarette packs. The improvements in both MAE and RMSE across different comparative models underscore the model's superior capability in handling intricate counting tasks under varied and complex environmental conditions.

5.4 Cross-validation experiments

To comprehensively evaluate the performance of the

proposed model in cigarette pack counting, cross-validation experiments were carried out using the Zhongyan cigarette pack dataset. These experiments incorporated three distinct data splits and two different quantities of support images. The model's performance was compared with that of the FamNet model, and the results are detailed in the following table:

Table 4. Test results of cross-validation experiments

Metric	Method	1-Shot				3-Shot			
		Fold-0	Fold-1	Fold-2	Mean	Fold-0	Fold-1	Fold-2	Mean
MAE	GMN	4.79	4.66	4.60	4.68	4.32	4.17	4.98	4.49
	FamNet	3.57	3.41	3.39	3.46	3.18	2.99	2.80	2.99
	Ours	2.13	2.29	2.41	2.28	1.84	1.47	1.35	1.55
RMSE	GMN	9.05	8.93	8.52	8.83	8.06	7.49	7.24	7.60
	FamNet	6.68	6.23	5.89	6.27	5.45	4.71	4.57	4.91
	Ours	2.44	2.86	3.27	2.86	2.27	2.09	1.93	2.10

An analysis of Table 4 reveals that, across all dataset splits and irrespective of the number of support images used, the proposed model consistently outperforms FamNet. In the 1-shot scenario, our model demonstrates an average improvement of 1.18 in MAE and 3.41 in RMSE when compared to FamNet. In the more challenging 3-shot scenario, the proposed model shows an even more pronounced improvement, with a 1.44 increase in MAE and a 2.81 increase in RMSE relative to FamNet. These results suggest that the proposed model is more adept at learning the similarity relationship between support and query features, thereby yielding superior counting accuracy.

When comparing the 1-shot and 3-shot scenarios as outlined in the table, it is apparent that the proposed model attains more significant performance gains compared to the FamNet method. In the case of our model, the 3-shot scenario shows a 32% enhancement in MAE and a 57% improvement in RMSE over the 1-shot scenario. Conversely, FamNet exhibits only a 14% increase in MAE and a 22% increase in RMSE in the 3-shot scenario compared to the 1-shot scenario. This disparity highlights the effectiveness of the proposed model in leveraging the advantages of multiple support images. The ability to utilize additional support images more efficiently translates into notably improved performance in practical cigarette pack counting tasks, emphasizing the robustness and adaptability of the proposed counting approach.

5.5 Cross-validation experiments

To substantiate the enhancements brought by the image enhancement module, along with the color and edge feature extraction modules in addressing cigarette pack counting challenges, ablation experiments were conducted on various components of our proposed approach. These included the denoising module, brightness enhancement module, deblurring module, and feature extraction module. The objective was to demonstrate the individual and collective impact of each module on the final detection outcomes. The results of these experiments are summarized in the accompanying table. "Baseline" in this context refers to the SAFECount model devoid of these enhancement modules. "M1,""M2," and "M3" correspond to the inclusion of the denoising module, deblurring module, and our specifically developed feature extraction module, respectively. "M4" indicates the combined application of the denoising and deblurring modules, "M5" denotes the integration of the denoising module with the feature extraction module, "M6" represents the combination of the deblurring module and the feature extraction module, and "M7" encapsulates the complete methodology as proposed in this paper.

Table 5. The test results of the ablation experiments conducted on the Zhongyan cigarette pack dataset using our proposed method

Method	Denoise	Deblur	Feature Extractor	MAE	RMSE
Baseline				1.90	2.41
M1	√			1.81	2.19
M2		√		1.84	2.24
M3			√	1.86	2.30
M4	√	√		1.79	2.13
M5	√		√	1.79	2.15
M6		√	√	1.80	2.19
M7	√	√	√	1.71	1.95

The analysis of Table 5 reveals notable improvements in performance metrics when comparing various model configurations with the baseline SAFECount model. M1, which incorporates the denoising module, demonstrates superior performance over the baseline, indicating its efficacy in enhancing the model's recognition of cigarette packs in noisy environments. This is quantified by a 4.7% increase in MAE and a 9.1% increase in RMSE.

Similarly, M2, which includes the deblurring module, effectively counters the effects of image blurriness, leading to a 3.2% rise in MAE and a 7.1% rise in RMSE compared to the baseline. This underscores the module's contribution to improving the model's technical performance in complex environments.

The comparison between M3 and the baseline further demonstrates the effectiveness of the proposed feature extraction module, particularly in extracting individual cigarette packs in complex scenarios, resulting in a 2.1% improvement in MAE and a 4.6% improvement in RMSE.

When examining M4, which combines the denoising and deblurring modules, there is a more significant improvement across all performance metrics, with increases of 5.8% in MAE and 11.6% in RMSE over the baseline. Likewise, M5, which merges the denoising module with the feature extraction module, shows enhanced improvements, with a 5.8% rise in MAE and a 10.8% rise in RMSE. M6, combining the deblurring module and the feature extraction module, also shows superior performance, with a 5.3% increase in MAE and a 9.1% increase in RMSE.

Notably, the results of M4, M5, and M6 surpass those of M1, M2, or M3 when used individually, suggesting that the combined usage of these two modules can further boost the model's performance.

Finally, M7, which represents the full integration of the three modules as proposed in this paper, shows the most significant improvements. The MAE and RMSE values

increase by 10.0% and 19.1%, respectively, compared to the baseline. This comprehensive integration demonstrates that each module contributes uniquely and synergistically, greatly enhancing the overall model's performance in the precise task of cigarette pack counting.

6. CONCLUSION

The primary objective of this study was to enhance counting accuracy in cigarette pack counting tasks, particularly in complex environments, while addressing challenges such as noise reduction, deblurring, variations in cigarette pack appearance, and improving model scalability. To achieve this, we introduced a novel method named "cigarette pack counting based on various image enhancement techniques and improved SAFECount". This method encompasses three integral components: an image enhancement module, a feature extraction network, and a counting module. Each component is meticulously designed to effectively tackle specific issues—noise reduction, image blurriness, changes in the appearance of cigarette packs, and limitations in model scalability.

We conducted comprehensive comparative experiments to evaluate our proposed method against the state-of-the-art CNN-based object counting model HLCNN and the leading Transformer-based object counting model CounTR. In addition, we performed ablation studies to assess the impact of the image enhancement module and the feature extraction network developed in this study. The outcomes of these experiments unequivocally demonstrated that our proposed model excels in counting cigarette packs within complex environments, outperforming existing models in accuracy and robustness. Notably, it achieves a MAE of 1.71 and a RMSE of 1.95, validating its effectiveness in real-world cigarette warehouse scenarios.

Looking ahead, our research will continue to concentrate on overcoming the challenges associated with counting cigarette boxes in intricate environments. The focus will be on further refining the model, exploring new techniques and approaches to enhance its accuracy and adaptability, and ensuring its applicability and efficiency in practical settings. The goal remains to develop a model that not only addresses current challenges but is also versatile enough to adapt to future advancements and changes in this field.

REFERENCES

[1] Gao, H., Zhang, W., Zhang, D., Deng, M. (2023). Application of improved transformer based on weakly supervised in crowd localization and crowd counting. Scientific Reports, 13(1): 1144. https://doi.org/10.1038/s41598-023-00000-1

[2] Patel, A., Dalal, M., Thakkar, V., Singh, P. (2023). Crowd counting analysis using deep learning: A critical review. Procedia Computer Science, 218: 2448-2458. https://doi.org/10.1016/j.procs.2023.01.220

[3] Biliavska, V., Castanho, R.A., Vulevic, A. (2022). Analysis of the impact of artificial intelligence in enhancing the human resource practices, Journal of Intelligent Management Decision, 1(2): 128-136. https://doi.org/10.56578/jimd010206.

[4] Zhang, S., Wang, W., Zhao, W., Wang, L., Li, Q. (2023). A cross-modal crowd counting method combining CNN and cross-modal transformer. Image and Vision Computing, 129: 104592. https://doi.org/10.1016/j.imavis.2022.104592

[5] Nasrullah, I., Shahzad, A., Kim, K. (2019). Convolutional-neural network-based image crowd counting: Review, categorization, analysis, and performance evaluation. Sensors, 20(1): 43. https://doi.org/10.3390/s20010043

[6] Vatambeti, R., Mamidisetti, G. (2023). Routing attack detection using ensemble deep learning model for IIoT. Information Dynamics and Applications, 2(1): 31-41. https://doi.org/10.56578/ida020104

[7] Zhang, Y., Zhang, W., Yu, J., He, L., Chen, J., He, Y. (2022). Complete and accurate holly fruits counting using YOLOX object detection. Computers and Electronics in Agriculture, 198: 107062. https://doi.org/10.1016/j.compag.2022.107062

[8] Bhattarai, U., Karkee, M. (2022). A weakly-supervised approach for flower/fruit counting in apple orchards. Computers in Industry, 138: 103635. https://doi.org/10.1016/j.compind.2022.103635

[9] Wu, Z., Sun, X., Jiang, H., Mao, W., Li, R., Andriyanov, N., Fu, L. (2023). NDMFCS: An automatic fruit counting system in modern apple orchard using abatement of abnormal fruit detection. Computers and Electronics in Agriculture, 211: 108036. https://doi.org/10.1016/j.compag.2023.108036

[10] Saddik, A., Latif, R., Abualkishik, A.Z., El Ouardi, A., Elhoseny, M. (2023). Sustainable yield prediction in agricultural areas based on fruit counting approach. Sustainability, 15(3): 2707. https://doi.org/10.3390/su15032707

[11] Iqbal, S.H., Mansoor, M.A., Irfan, Y.M.Z. (2022). Rice crop counting using aerial imagery and GIS for the assessment of soil health to increase crop yield. Sensors, 22(21): 8567. https://doi.org/10.3390/s22218567

[12] Bian, X., Wang, P., Cao, Z., Lu, H., Xiong, H., Yang, A., Yao, J. (2023). Rice plant counting, locating, and sizing method based on high-throughput UAV RGB images. Plant Phenomics, 5: 0020. https://doi.org/0.34133/plantphenomics.0020

[13] Farjon, G., Huijun, L., Edan, Y. (2023). Deep-learning-based counting methods, datasets, and applications in agriculture: A review. Precision Agriculture, 24: 1683-1711. https://doi.org/10.1007/s11119-023-10034-8

[14] Gao, G.H., Wang, S.Y., Shuai, C.Y., Zhang, Z.H., Zhang, S., Feng, Y.B. (2022). Recognition and detection of greenhouse tomatoes in complex environment. Traitement du Signal, 39(1): 291-298. https://doi.org/10.18280/ts.390130

[15] Subramani, M., Rajaduari, K., Choudhury, S.D., Topkar, A., Ponnusamy, V. (2020). Evaluating one stage detector architecture of convolutional neural network for threat object detection using X-ray baggage security imaging. Revue d'Intelligence Artificielle, 34(4): 495-500. https://doi.org/10.18280/ria.340415

[16] Cao, L., Xiao, Z., Liao, X., Yao, Y., Wu, K., Mu, J., Li, J., Pu, H. (2021). Automated chicken counting in surveillance camera environments based on the point supervision algorithm: LC-DenseFCN. Agriculture, 11(6): 493. https://doi.org/10.3390/agriculture11060493

[17] Dijkstra, K., van de Loosdrecht, J., Schomaker, L.R., Wiering, M.A. (2019). Centroidnet: A deep neural network for joint object localization and counting. In

Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, Part III 18, pp. 585-601. https://doi.org/10.1007/978-3-030-10928-8_27

[18] Chow, Y.T., Lee, H.K., Chan, L.K. (2023). Detection of targets in road scene images enhanced using conditional GAN-based dehazing model. Applied Sciences, 13(9): 5326. https://doi.org/10.3390/app13095326

[19] Nagireddy, V.R.S.R., Shaik, K.S. (2023). Advanced hybrid segmentation model leveraging AlexNet architecture for enhanced liver cancer detection. Acadlore Transactions on AI and Machine Learning, 2(3): 116-128. https://doi.org/10.56578/ataiml020301

[20] Veregin, H., Lanter, D.P. (1995). Data-quality enhancement techniques in layer-based geographic information systems. Computers, Environment and Urban Systems, 19(1): 23-36. https://doi.org/10.1016/0198-9715(94)00032-8

[21] Russo, F. (2002). An image enhancement technique combining sharpening and noise reduction. IEEE Transactions on Instrumentation and Measurement, 51(4): 824-828. https://doi.org/10.1109/TIM.2002.803394

[22] Huang, Y.L., Li, N., Liu, Z.B. (2023). An enhanced convolutional neural network for accurate classification of grape leaf diseases. Information Dynamics and Applications, 2(1): 8-18. https://doi.org/10.56578/ida020102

[23] Buades, A., Coll, B., Morel, J.M. (2005). A review of image denoising algorithms, with a new one. Multiscale Modeling & Simulation, 4(2): 490-530. https://doi.org/10.1137/040616024

[24] Nah, S., Son, S., Lee, S., Timofte, R., Lee, K.M. (2021). NTIRE 2021 challenge on image deblurring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, USA, pp. 149-165. https://doi.org/10.1109/CVPRW53098.2021.00025

[25] Zhang, K., Ren, W., Luo, W., Lai, W.S., Stenger, B., Yang, M.H., Li, H. (2022). Deep image deblurring: A survey. International Journal of Computer Vision, 130(9): 2103-2130. https://doi.org/10.1007/s11263-022-01633-5

[26] Zhang, Y., Jiang, N., Xie, Z., Cao, J., Teng, Y. (2023). Ultrasonic image's annotation removal: A self-supervised Noise2Noise approach. arXiv preprint arXiv:2307.04133. https://doi.org/10.48550/arXiv.2307.04133

[27] Waqas Zamir, S., Arora, A., Khan, S., Hayat, M., Shahbaz Khan, F., Yang, M.H. (2021). Restormer: Efficient transformer for high-resolution image restoration. arXiv E-Prints, arXiv-2111. https://doi.org/10.48550/arXiv.2111.09881

[28] You, Z., Yang, K., Luo, W., Lu, X., Cui, L., Le, X. (2023). Few-shot object counting with similarity-aware feature enhancement. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, USA, pp. 6315-6324. https://doi.org/10.48550/arXiv.2201.08959

[29] Lu, E., Xie, W., Zisserman, A. (2018). Class-agnostic counting. arXiv E-Prints. https://doi.org/10.48550/arXiv.1811.00472

[30] Ranjan, V., Sharma, U., Nguyen, T., Hoai, M. (2021). Learning to count everything. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, USA, pp. 3394-3403. https://doi.org/10.48550/arXiv.2104.08391

[31] Yang, S.D., Su, H.T., Hsu, W.H., Chen, W.C. (2021). Class-agnostic few-shot object counting. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, USA, pp. 870-878. https://doi.org/10.1109/WACV48630.2021.00091

[32] You, Z., Yang, K., Luo, W., Lu, X., Cui, L., Le, X. (2023). Few-shot object counting with similarity-aware feature enhancement. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, USA, pp. 6315-6324. https://doi.org/10.48550/arXiv.2201.08959