Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Research on the Review and Processing of Evidence for Public Interest Litigation Based on Computer Vision Technology

Yongjun Wang[1,] ✉

[1] *Law School, Henan University of Urban Construction, Pingdingshan, Henan, 467036, China*

ABSTRACT

This paper analyzes public interest litigation and its salient features, and organizes the audit rules for the electronic transformation of litigation evidence. Aiming at the phenomenon of varying text length in litigation evidence, a joint CTC-Attention decoding model (HCADecoder) based on bigram hybrid labeling is proposed. Based on the existing research on computer vision technology for target number prediction, the stacked object occlusion problem existing in special scenes is proposed, and an algorithm for predicting the number of stacked objects combining planar density map and depth map is proposed. Combined with the public interest litigation evidence document corpus dataset, we analyze the recognition of basic elements of litigation evidence by text label recognition algorithm, and select the commonly used precision rate P, recall rate R and F1 value to evaluate the recognition results of basic elements. Subdivide the text length of litigation evidence and analyze the recognition accuracy of each algorithm on different text lengths. Bring the text label recognition algorithms into real cases to analyze the element extraction. For this paper, we propose monocular image target counting algorithm, which is brought into different scenarios for performance testing. This paper proposes text label recognition algorithm with evidence image target counting algorithm for litigation evidence text image recognition with mean value at 80%.

Keywords: bigram hybrid labeling, CTC-Attention module, computer vision techniques, image target counting, litigation evidence review

1. Introduction

Evidence review elements, refers to the litigation activities, according to the litigation law and judicial interpretation of the norms, the litigation parties to cross-examine the evidence, the judicial organs to review the determination of the evidence of the basic elements of consideration [1-2]. Over the years,

✉ Corresponding author.
E-mail address: wyj30150603@163.com (Y. Wang).
Received 25 February 2024; Revised 05 April 2024; Accepted 15 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-282

in China's litigation activities, the basic elements of evidence cross-examination and review and determination are the so-called "three characteristics", that is, the relevance (relevance), objectivity (authenticity) and legitimacy of evidence, and the review method is also formed based on the "three characteristics" review, that is, through the cross-examination and review of the "three characteristics", it is determined whether the evidence can be the basis for determining the facts of the case and its probative role [3-6]. Due to the introduction of new concepts of proof, the concept of elements of evidence review, and the method of determining facts, as well as the development of China's litigation system and evidence system itself [7-8], the "three genders" review system of evidence has been questioned, including the theoretical questioning of individual elements (such as objectivity and legitimacy) as the basic attributes of evidence, and the overall challenge to the "three genders" review system (such as the review system with the dichotomy of evidence qualification and proof evaluation as an alternative) [9-11].

Computer vision is one of the important branches of computer science, which refers to the use of cameras and computers, etc. to recognize, track and measure targets and process them into images that are more suitable for the human eye to observe or transmit to instruments for detection [12-13]. The goal is to enable computers to understand the world through visual observation like a human being and have the ability to adapt to the environment autonomously, and its core technology includes visual perception and visual generation, in which visual perception involves recognition and classification, target detection, image segmentation, representation learning, etc., and visual generation involves the generation of images and videos, and the combination of vision and text, etc. [14-16]. Under the support of policy planning and standard system, computer vision architecture takes characters, graphics, video and other contents as the main analyzing object, and is logically divided into foundation layer, technology layer, product layer and application layer [17-19]. The downstream commercial value of China's computer vision industry continues to broaden and deepen, and the future landing areas will be more diverse [20-21]. Computer vision technology introduced into the evidence review link, whether it can enhance the judicial efficiency and quality of evidence review is very worth thinking about.

In this paper, for the different text lengths existing in public interest litigation evidence, an effective deep learning-based text recognition model is designed, which mainly consists of the CTC module and Attention module jointly constituting the CTC-Attention decoding model, and is coupled with a hybrid annotation method based on bigram. Aiming at the real scene exists in the object stacking affects the number of target counting, and affects the authenticity of the litigation evidence situation, this paper is based on computer vision counting, and proposes a combination of planar density map and depth map of the number of stacked objects prediction algorithm. Using public interest litigation evidence data for this paper's text label recognition algorithm test, including litigation evidence of element identification and litigation case causal identification. Analyze the performance of the image target counting algorithm proposed in this paper in different scenarios.

2. Computer vision technology

2.1. Text label recognition algorithm

As the word frequency of Chinese text varies greatly. The effect of word frequency will create some preferences for the model. At the same time, the confusion brought to the recognition model by Chinese characters with complex structures makes the model prone to shape-sharing errors. Moreover, the longer the text length, the easier it is to produce recognition errors.

In this paper, we design an effective deep learning-based text recognition model for indeterminate-length Chinese text image scenarios. From the perspective of utilizing Chinese word frequency differences, a joint CTC-Attention decoding model (HCADecoder) based on bigram hybrid labels is proposed.

2.1.1. Hybrid annotation approach based on bigram. This chapter uses single-character and double-character subwords to construct a mixed-label dictionary. Based on the mixed-label dictionary, the label sequence is then reconstructed for each sample. Let $\{I,Y\}$ represent an input sample in the training set, where I is a text image and $Y=\left\{\left(y_1,y_2,\cdots y_l\right)\mid y_i\in D,i\in[1,l]\right\}$ is the original single-word label sequence.

The construction of the mixed label sequence is done by replacing all neighboring unigram labels matching $\left\langle y_i,y_{i+1}\right\rangle\in S$ with a bigram label. If a unigram tag does not find a suitable match, it is kept in the tag sequence without any change.

The final hybrid tag sequence is denoted as $Y'=\left\{\left(y_1',y_2',\cdots,y_{l'}'\right)\mid y_j'\in D',j\in\left[1,l'\right]\right\}$, which replaces the original single-word tag Y, and is used with the proposed joint CTC-Attention decoder for Chinese text recognition.

2.1.2. Joint CTC-Attention Decoding Models:

1) Shared Encoder Module. The use of ResNet residual modules allows the deeper parts of the network to have complete access to the previous features, and saves the gradient of the parameters, facilitating the network to be able to return better gradient values and avoid gradient dispersion.

In this paper, a 32-layer ResNet-based CNN is used to extract deeper text features, and all ResNet blocks have the following format:

$$\left\{\left[\ker nel\ size,num-ber\ of\ channels\right]\times\right\} \tag{1}$$

where the default is $\{stride,pad\}=\{0,0\}$. The other convolutional layers have the following format:

$$\left\{kernel_W\times kernel_H\times stride_W\times stride_H,pad_W\times pad_H,channels\right\} \tag{2}$$

The maxpooling layer has the following format:

$$pool:\left\{kernel_W\times kerne_H\times stride_W\times stride_H,pad_W\times pad_H\right\} \tag{3}$$

where subscripts W and H denote the width and height of the feature map, respectively. The feature sequences extracted from the CNN are then fed into a bidirectional long short-term memory network (BiLSTM).

In the shared encoder, the recurrent neural network layer is immediately followed by the convolutional layer (ResNet32). The selected recurrent neural network layer (RNN) is two superimposed bidirectional long short-term memory networks (BiLSTM).

The LSTM is a directional neural network unit and the information it utilizes is also limited to past contextual information. The combined utilization of contextual information from both directions to form a bi-directional LSTM (BiLSTM) can make the recognition performance of the neural network much better.

LSTM is composed of an input gate, a forgetting gate, and an output gate.

Oblivion gate: the oblivion parameter equation is:

$$f_t=\sigma\left(W_f\cdot\left[h_{t-1},x_t\right]+b_f\right) \tag{4}$$

Enter the gate:

This consists of two steps:

Step1: A sigmoid function called “input gate layer” decides what information needs to be updated.

The input gate layer accepts the input $\left(x_t\right)$ at the current moment and the hidden state $\left(h_{t-1}\right)$ at the previous moment, which is processed by a linear transformation and a sigmoid function to obtain an input gate vector $\left(i_t\right)$ with values between 0 and 1, indicating how much new information should be input at the current moment. The update parameter formula is:

$$i_t=\sigma\left(W_t\cdot\left[h_{t-1},x_t\right]+b_t\right) \tag{5}$$

Step2: Generate an alternative cell state $\tilde{c}_c$ by a layer $\tanh$, indicating how much new information will be added to the cell state.

The candidate state formula is:

$$\tilde{c}_t = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \tag{6}$$

The final alternative new state vectors $\left(\tilde{c}_t\right)$ and i_t are multiplied by elements to determine the degree of updating of the new information, and then added to the retained information from the previous moment to obtain the new unit state vector.

The update state formula is:

$$c_t = f_t * c_{t-1} + i_t * c_t \tag{7}$$

Output Gate: A sigmoid function called an "output gate" is used to compute which part of the unit state needs to be output.

The output gate takes the current input $\left(x_t\right)$ and the hidden state $\left(h_{t-1}\right)$ of the previous moment as current inputs, and then uses a linear transformation and a sigmoid function to generate an output gate vector $\left(o_t\right)$ between 0 and 1, which is then multiplied by the elements to get the final output vector $\left(h_t\right)$.

The output parameter equation is:

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{8}$$

The output implicit layer state formula is:

$$h_t = o_t * \tanh\left(c_t\right) \tag{9}$$

The implicit layer state for backward propagation is:

$$\overrightarrow{h_t} = f\left(\vec{W}x_t + \vec{V}\overrightarrow{h}_{t-1} + \vec{b}\right) \tag{10}$$

The implicit layer state for forward propagation is:

$$\overleftarrow{h_t} = f\left(\vec{W}x_t + \vec{V}\overleftarrow{h}_{t+1} + \vec{b}\right) \tag{11}$$

2) CTC module. In the decoding phase, the CTC module first decodes H to obtain a CTC sequence $Z = \left(z_1, z_2, \cdots, z_T\right)$, where $z_t \in D' \cup \langle b \rangle$ represents a one-word or two-word subword. ' $\langle b \rangle$ ' represents a space character.

Then it modifies Z to obtain Z' according to the mapping relationship between double-word subword and single-word. $t_i \in [1, T]$ represents the ith position in the sequence.

Suppose that the prediction result corresponding to positions t_3, z_3 is s_1, where $s_1 = \left\{\left\langle d_1, d_2\right\rangle \mid d_1, d_2 \in D, \left\langle d_1, d_2\right\rangle \in S\right\}$. Then it is necessary to find position t_m in Z from position t_3 forward to satisfy $z_m = d_1 \wedge z_{m-1} \neq d_1$, and backward to find position t_n to satisfy $z_n = d_2 \wedge z_{n+1} \neq d_2$. And every prediction result between $z_m \sim z_n$ belongs to $\left\{d_1, d_2, s_1, \langle b \rangle\right\}$. Then all the prediction results between $z_m \sim z_n$ are corrected to s_1.

When the repeated labels in Z' are collapsed to a single occurrence, Z' can be expressed as:

$$Z' = \left(\langle b \rangle^*, \left(z'\right)_1^{\varepsilon_1}, \langle b \rangle^*, \left(z'\right)_2^{\varepsilon_2}, \langle b \rangle^*, \cdots \left(z'\right)_L^{\varepsilon_L}, \langle b \rangle^*\right) \tag{12}$$

where * denotes zero or more repetitions and ε denotes the number of repetitions of Z'. Let function B denote the mapping of Z' into $\left(\left(\left(z'\right)_1, \left(z'\right)_2, \cdots \left(z'\right)_L\right), z' \in D'\right)$ by collapsing and deleting blank symbol operations. Conditional probability is then defined as the sum of the probabilities of all

paths π mapped by B to the truth label Y':

$$P_{CTC}\left(Y'|Z'\right)=\sum_{\pi:B(\pi)-Y'}p\left(\pi|Z'\right) \tag{13}$$

where $p\left(\pi|Z'\right)=\prod_{t=1}^{T}\left(z'\right)_{\pi t}^{t}$ denotes the probability of π and $\left(z'\right)_{\pi t}^{t}$ denotes the probability of labeling π_t at position t.

3) Attention module. In the CTC module, the obfuscated predictions are converted into bigram-level final results and collapsed together. Based on these results, the Attention module decodes each collapsed subsequence, which is equivalent to narrowing down the attention to distinguish the confusing predictions [22]. To wit:

$$P_{Atten}\left(Y'|Z'^{*},H\right)=\prod_{l=1}^{L}p\left(y_l'|y_1',\ldots,y_{l-1}',Z'^{*},H\right) \tag{14}$$

The alignment information provided by the modified sequence Z'^{*} is only used to help the attention module classify H as a subsequence:

$$P_{Atten}\left(Y'|Z'^{*},H\right)=\prod_{l=1}^{L}p\left(y_l'|y_1',\ldots,y_{l-1}',h_1,\ldots,h_{\tau_\tau}\right) \tag{15}$$

where $\tau_l=t_l'+\varepsilon_l$, t_l' and ε_l are the position index number and the number of repetitions of Z'^{*}, respectively.

The conditional probability $p\left(y_l'|y_1',\ldots,y_{l-1}',h_1,\ldots,h_{T_l}\right)$, obtained through the textual attention mechanism, is computed as shown below:

$$\begin{gathered}a_{l,t}=Context\ \ Attention\left(q_{l-1},h_t\right)\\ r_l=\sum_{t-1}^{T_l}a_{l,t}h_t\\ p\left(y_l'|y_1',\ldots,y_{l-1}',h_1,\ldots,h_{T_l}\right)=Decoder\left(r_l,q_{l-1},y_{l-1}'\right)\end{gathered} \tag{16}$$

where q_{l-1} represents the hidden state of the decoder at the previous position. At each step l, the decoder generates a context vector based on the first tau_l feature h and the attention weight a.

4) Loss function. Joint CTC-Attention is co-trained using a multi-task loss function defined as the weighted sum of the CTC loss and the Attention-based loss:

$$\begin{gathered}\mathrm{L}_{Total}=\lambda\mathrm{L}_{CTC}+(1-\lambda)\mathrm{L}_{Attention}\\ \mathrm{L}_{CTC}=-\sum_{(I,Y')\in\mathrm{X}}\log P_{CTC}\left(Y'|Z'\right)\\ \mathrm{L}_{Attention}=-\sum_{(I,Y')\in\mathrm{X}}\log P_{Atten}\left(Y'|Z'^{*},H\right)\end{gathered} \tag{17}$$

where the regulation parameter λ controls the weights between the two objective functions.

Both the CTC module and the Attention module share the same hybrid labeling dictionary. In HCADecoder, the output of the CTC module is only the preliminary result that guides the training of the Attention module. Therefore. The output of the Attention module will be used as the final unique result in the inference process. It is only in the training phase that both the outputs of the CTC module and the attention module are used to compute the loss.

2.2. Monocular Image Target Counting Algorithm

Computer vision-based target number prediction, which firstly performs feature extraction on the

input image to get the number density of the counted objects, and finally integrates the output statistics, shows superior performance in terms of 2D target counting accuracy. However, in some special scenes, the single target shape contour is irregular, and the shape presents diversity due to stacked occlusion. At the same time, the single image is not visible to the occlusion location, and the planar counting method cannot be directly used for three-dimensional counting, which makes the automated acquisition of the number challenging.

In order to solve the problem of counting the number of stacked targets in single-view images, this paper proposes an algorithm for estimating the number of stacked objects by combining planar density map and depth map based on the existing computer vision counting research.

2.2.1. Plane density map generation module:

1) In this paper, DM-count counting method is used as a benchmark.DM-count distribution matching network is an algorithm applied to planar crowd counting.Crowd counting uses a dataset whose true value (GT) is described in terms of the coordinates of the centroid of the head in the image, i.e., point labeling.

2) The material counted in the scene will not appear the same type of target scale changes too much, this paper borrows the idea of distribution matching and optimal transmission, using a single column network, in addition to increase the GAM attention mechanism, so that the network automatically focus on the features important for counting, to improve the performance of the network in the counting of the real material.

Before entering the density regression computation, this paper adds self-attention mechanism layer to the new network structure to enhance the weight of the attributes that are important for counting in the input data of the neural network. In this paper, the best optimized GAM attention mechanism is chosen.GAM uses channel-space attention mechanism and the redesigned CBAM sub-module is able to amplify the global dimensional interaction features while reducing the information dispersion.Given the input feature mapping, the intermediate and final outputs are defined as F_2 and F_3 respectively as shown in Eqs. (18) and Eq. (19).

$$F_2 = M_c\left(F_1\right) \otimes F_1 \tag{18}$$

$$F_3 = M_s\left(F_2\right) \otimes F_2 \tag{19}$$

M_c and M_s are channel attention maps and spatial attention maps, respectively, and $\otimes$ represents the multiplication operation by element.

2.2.2. Monocular depth prediction module. In this paper, we consider depth information, which can provide more adequate feature support for stacked object counting, assisting stacked object counting in monocular images.

S2R-DepthNet is a monocular depth prediction algorithm, where the network structure is learned to obtain a depth-specific structural representation, from which the key information in depth estimation is captured, and some irrelevant stylistic information in the data is excluded, so that the depth network focuses on the structural information of the scene, and the model trained on synthetic data has a good generalization ability even. The whole network structure contains the following three parts:

1) Structural information extraction module (STE)
2) Deep feature attention module (DSA)
3) Deep Prediction Module (DP)

2.2.3. Monocular stereo counting module. After the planar density matrix and depth matrix are acquired by the Optimized Density Map Generation Module and the Monocular Depth Prediction

Module, respectively, the aim of this section is to project the number of stacked objects using the 3D information acquired on the monocular image.

The flow of the whole monocular stacked object counting algorithm is shown below, which presents the number of objects contained in each pixel point in the monocular image from the planar density matrix acquired by the planar density counting network. Through experimental testing, a threshold of 0.001 is selected to divide the region where the target is located and the irrelevant region according to the continuity of the planar map, and the target region on the one hand assists the depth map to confirm the region. On the other hand, by integrating the planar density map, the total number of planes of the objects in the picture can be obtained, and the total pixel value occupied by the same target region can be used to estimate the number of pixels occupied by a single object.

3. Recognition review of text/images for litigation evidence

3.1. Public interest litigation and its distinctive features

Public interest litigation has a number of distinctive features, the identification of which is the basis for the special requirements of the evidentiary approach. When discussing the characteristics of public interest litigation, it is in fact in contrast to private interest litigation. The distinctive features of public interest litigation related to evidence investigation activities are as follows:

1) The public interest purpose of litigation. As far as the litigation system is concerned, the purpose of litigation is the ultimate pursuit of judicial dispute resolution, guiding the whole process of litigation, determining the construction of litigation and the construction of procedural rules. In terms of civil public interest litigation, its main purpose is to safeguard the public interest of society. The interest is a kind of many people enjoy but indivisible entity interests, this is the public interest litigation is different from the private interest litigation of the biggest characteristics.

Whether it is a typical public interest litigation for the purpose of purely safeguarding the public interest of society, or a hybrid public interest litigation for the purpose of safeguarding the public interest of society and taking into account private interests, the purpose of the litigation has a clear public interest.

“Evidence is the king of litigation” and is at the heart of the entire litigation. In litigation, whoever holds the evidence holds the direction of the litigation, and public interest litigation is no exception.

2) Implementing a Moderated Dispositional Doctrine and Limited Argumentation Doctrine. As mentioned above, the object of public interest litigation is to protect the public interest of many unspecified subjects, rather than the private interest of the plaintiff who filed the lawsuit. Therefore, the traditional civil litigation follows the dispositive and argumentation, in the application of the need to be amended and limited, in order to prevent the occurrence of behavior that harms the public interest.

Dispositionalism in civil law countries, also known as the principle of disposition, its core meaning is the parties in the litigation of their substantive rights and procedural rights to enjoy complete control, can choose whether to exercise these rights according to free will. Argumentation is the dispositive right of the doctrine outside the civil litigation structure to support another core principle, which is based on the two parties have equal rights to litigation.

3) The Need for Broad Discretionary Powers for Judges. Discretion is an important judicial power of the judge, can effectively make up for the written law of the lag and the principle of legal rules brought about by the shortcomings. At the same time, the results of public interest litigation decisions have a certain outlook, and can play a modeling role for subsequent cases of the same kind. In litigation, the key for a judge to accurately determine the facts of a case lies in the evidence supporting the facts.

3.2. Rules for the submission and review of electronic evidence

3.2.1. Mode of submission. Evidentiary material submitted in online proceedings should be electronic. In practice, electronic materials can be divided into litigation documents and evidence materials in terms of content, and can be divided into three types in terms of presentation:

The first is the parties in the court of jurisdiction required by the aforementioned online litigation platform and other litigation service platform, the litigation materials they have already written offline, including the original of the indictment, defense, representation, etc. through scanning, shooting, transcription and other ways of electronic processing, uploaded to the aforementioned online litigation platform or other litigation service platform.

The second type is that the parties directly fill in the indictment, defense, agency opinions, case information and other materials online on the aforementioned online litigation platform or other litigation service platform required by the competent court to form electronic materials.

The third type is that the materials submitted by the parties are in the form of electronic data. In this case, the parties may directly upload such electronic materials on the online litigation platform or other litigation service platform, such as electronic contracts signed through online means, electronic payment vouchers and other evidentiary materials in the form of electronic data.

3.2.2. Audit Methods. In online litigation, the evidentiary materials submitted by the party bearing the burden of proof shall meet the requirements of the rules for the submission of electronic evidence in online litigation.

In the entire online litigation, litigation activities basically revolve around electronic evidence, "who claims, who gives evidence", the court for both parties to the dispute to determine the facts, must be based on the evidence of both parties to the determination of the material, therefore, electronic evidence for the entire litigation activities are very important.

This provision of the Rules stems from the fact that the electronic materials required for online litigation are susceptible to tampering, and is therefore designed to emphasize the strict examination of the consistency of the electronic evidence with the original documents, to prevent the occurrence of false litigation, malicious litigation and other acts of interference with litigation activities, and to ensure that online litigation develops in a healthy and orderly manner.

3.3. Identification of elements of litigation evidence

In order to validate the effectiveness of the methodology of this paper, experiments were conducted using Pytorch to implement the model structure of this paper, comparing the corresponding benchmark models.

3.3.1. Experimental setup. The dataset of this paper is derived from the corpus of public interest litigation evidence documents provided by legal companies, which contains 6,000 labeled sentences, and the labeling includes three basic categories of evidence name, substantiation content, and file number, and the statistics of the dataset are shown in Table 1. This experiment takes 80% of the dataset as the training set, 10% as the development set, and 10% as the test set. The training set, development set and test set are 4800, 600 and 600 respectively.

Table 1. Data set statistics

Basic elements	Training set	Verification set	Test set	Mean length	Mean length
Evidence	3375	430	430	6.69	12.75
Confirmed content	2190	280	285	33.25	55.31
File number	775	100	100	4.25	9.06

An example of the basic elements is shown in Table 2. The table shows the basic elements of

litigation evidence, including the name of the evidence, what it confirms, and the file number.

Table 2. Examples of basic elements

Basic elements	Basic description	Example
Evidence	There are several following: The documents, the evidence, the audio-visual materials, the witness testimony, the statement of the parties, the conclusion of the identification, and the examination of the record.	1-Related invoice. 2-Field information, population basic information, etc.
Confirmed content	The fact that the evidence is used.	1-Related invoice. 2-The above evidence is lawful and valid, and the evidence is linked to the ring and the whole proof is according to the chain.
File number	Here is the registration and registration of the contents.	1-Witness note. 2-Judgment, certificate, etc

In order to verify the validity of the methods in this paper, the following eight benchmark experiments are set up:

1) B-LSTM-CRF. The model first vectorizes the representation of each word in the sentence, and then inputs it into the bidirectional LSTM to obtain each word hidden layer representation which contains the contextual information of the current word, and finally inputs the obtained hidden layer representation into the CRF layer to jointly decode the labeled sequences at the sentence level.

2) CNN-LSTM-CRF. CNN is firstly used to get character-level word representation, and then the obtained word representation and the trained word embedding vectors are jointly input to bidirectional LSTM, after which the operation is consistent with the B-LSTM-CRF method.

3) LSTM-LSTM-CRF. This method is similar to the CNN-LSTM- CRF method, only the character-level CNN is replaced with LSTM, and the rest remains unchanged.

4) Lattice-LSTM. The essence of this model is character-based B-LSTM-CRF, and the focus of the model is how to design the Cell structure.

5) Att-LSTM-CRF. Firstly, self-attention is computed for each word in the sentence to get the internal representation of the sentence, and the obtained internal representation of the sentence and word vectors are spliced and inputted into bidirectional LSTM, which is also decoded by CRF.

6) LM-LSTM-CRF. This method is similar to the Att-LSTM-CRF method, except that the self-attention module is replaced with an LSTM-based contextual word representation module, and other things remain unchanged.

7) WLA-LSTM-CRF. This method splices the sentence internal representations, LSTM-based context word representations and word vectors obtained from the self-attention module and feeds them into the LSTM-CRF.

8) JCWA-DLSTM. The parameter settings for the implementation of the three methods for B-LSTM-CRF, CNN-LSTM-CRF and LSTM-LSTM-CRF are shown in Table 3. For the implementation of the Lattice-LSTM method, the original code is used and the SGD optimizer is changed to the Adam optimizer, and the rest of the parameter settings are the same as in this paper.

Table 3. Parameter setting

Model	**Parameter**	**Value**
Char CNN	**embedding size**	**80**
	kernel size	**5**
Char LSTM	**embedding size**	**100**
	char hidden dim	**100**

Word LSTM	**batch size**	**20**
	embedding size	**500**
	word hidden dim	**300**
	optimizer	**Adam**
	learn rate	**0.003**
	dropout	**0.5**

The specific parameter settings for the remaining four methods are shown in Table 4.

Table 4. The concrete parameter Settings of the other four methods

Parameter	Value	Parameter	Value
batch size	**30**	**optimizer**	**Adam**
embedding size	**500**	**learn rate**	**0.022**
word hidden dim	**500**	**dropout**	**0.55**

3.3.2. Experimental results and analysis. The commonly used precision rate P, recall rate R and F1 value are selected to evaluate the results of basic element recognition. Experiments were conducted on the eight proposed methods, and the experimental results of different methods are shown in Fig. 1.

1) The HCADecoder method proposed in this paper outperforms the eight compared methods, indicating that the method proposed in this paper is useful for recognizing the basic elements of public interest litigation evidence documents.

2) The experimental results of Lattice-LSTM are inferior to all the methods, mainly because some sentences in the dataset are too long and matched to too many words, and a large portion of these words are noisy, which makes the vocabulary quality poor and affects the experimental results.

3) The results of Att-LSTM-CRF are better than the first four methods, which indicates that the inter-word dependencies within the sentences are learned as well as the sentence internal information is captured through the self-attention mechanism, and these are helpful for the recognition of the basic elements of public interest litigation evidence documents.

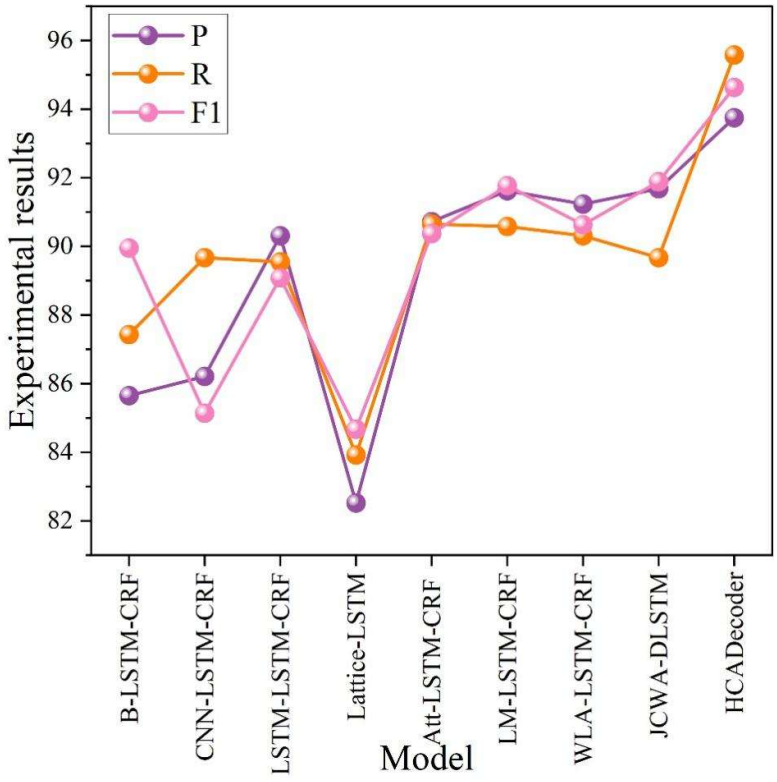


Fig. 1. Different methods of experimental results

In order to better analyze the textual image dataset of public interest litigation evidence provided by law firms, this study carefully classified images with different text lengths. The text was categorized into seven intervals: [0,10), [10,20), [20,30), [30,40), [40,50), [50,100), and [100,150), and the distribution of the number of samples within each interval was recorded in detail. The recognition accuracy of this paper's methods for different text lengths is shown in Fig. 2. The figure shows the recognition accuracy of each algorithm for different text lengths, and this paper's algorithm always maintains a stable accuracy rate for different text lengths.

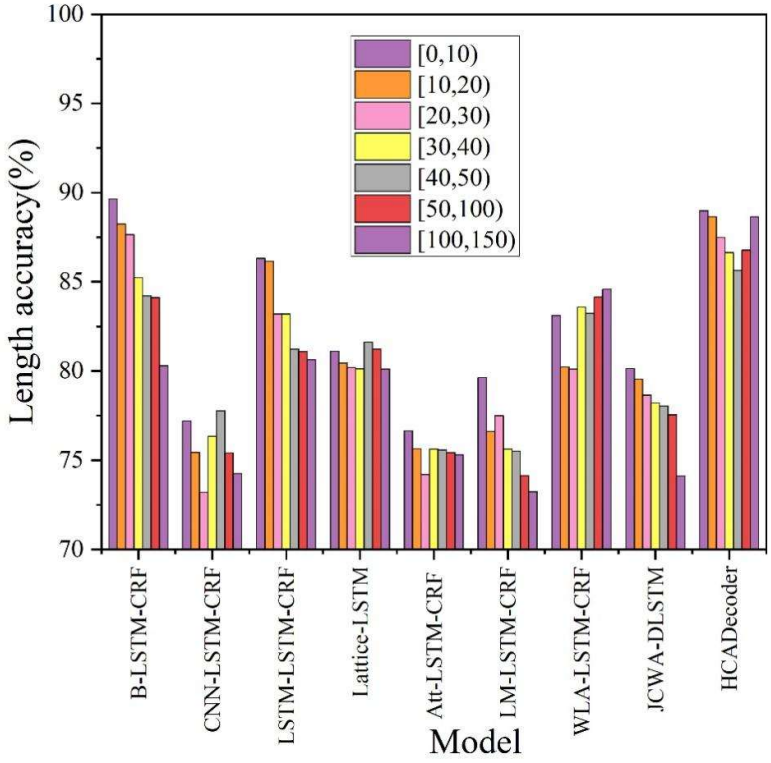


Fig.2. The accuracy of the identification of different text lengths in this paper

3.3.3. Case studies. To illustrate the effectiveness of the method proposed in this paper, the following two sentences are analyzed in this section as examples.

1) Expense invoice. Proof: the plaintiff's injuries and the fact of treatment and the costs incurred as a result of treatment.

2) 1 certificate of 105 house register of a certain park, owner of a certain house, with a floor area of 189.21 square meters.

For sentence 1), the correct essential element is "the plaintiff's injuries and the facts of the treatment and the expenses incurred as a result of the treatment", "invoice" and "expenses" are related, and both methods of CNN-LSTM-CRF and LSTM-LSTM-CRF are missing "expenses". HCADecoder can correctly identify the substantiated content, which means that the CTC-Attention module set in this article is helpful for identifying the basic elements.

In sentence 2), the basic element of the correct is "the owner of such and such a house", and the two methods of CNN-LSTM-CRF and LSTM-LSTM-CRF are influenced by the participle to obtain the basic element of error. However, HCADecoder can correctly identify the basic elements, indicating that the hybrid annotation method based on bigram proposed in this paper can help reduce the impact of word segmentation error on the recognition of basic elements.

In summary, the HCADecoder method proposed in this paper is effective for determining the

boundaries of basic elements such as the name of evidence and the content of substantiation.

3.4. Causal identification of the merits of the proceedings

3.4.1. Data sets. The dataset for the experiments in this chapter comes from the labeled public interest litigation adjudication paper case causality dataset. The model is trained on the training set, the optimal model architecture is selected with the development set, and the model performance is tested on the test set. In this paper, the dataset is constructed by random sampling, and the ratio of 8:1:1 is used to divide the training set, test set, and development set.

The distribution of each relationship in the case description inter-sentence causality dataset is shown in Table 5, including event motivation, event action, loss situation, financial compensation, behavioral compensation, and no causality.

Table 5. The distribution of the relationship of the case data

Classification	Event motive	Event action	Loss condition	Economic compensation	Behavior compensation	Free fruit
Train	156	67	120	60	30	200
Test	52	4	36	10	10	30
Dev	10	9	12	12	8	30
Total	218	80	168	82	48	260

3.4.2. Analysis of results. The same eight comparative methods mentioned above were chosen and brought into the actual public interest litigation evidence review session to analyze the actual performance of each method.

The performance of each method in each classification on the data set of causal relationship between case description sentences is shown in Figure 3. The classification performance of this paper's algorithm in each aspect of event motivation, event action, loss situation, economic compensation, behavioral compensation, and no causation is 0.7993, 0.3749, 0.9091, 0.8649, 0.9101, and 0.7057, respectively. The classification performance in all aspects was higher than the other compared methods.

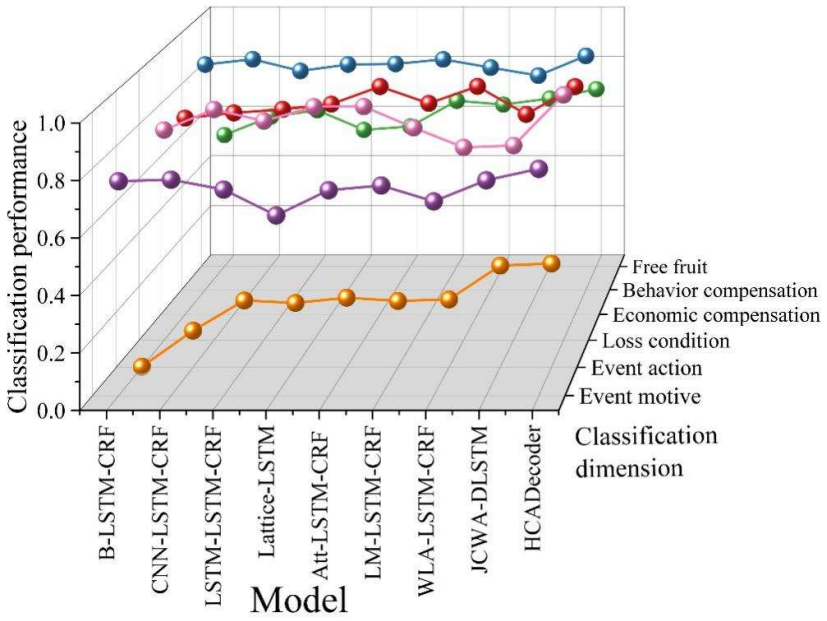


Fig. 3. Performance of each classification in the data set of the parties

3.5. Evidence image review and target counting

3.5.1. Target counting in different scenarios:

1) Experimental dataset. The ShanghaiTech dataset is used for training and testing in this experiment. The dataset is divided into two parts, part A and part B, which contains 1300 annotated images, randomly selected images in part A and randomly selected images in part B. Part A contains 600 images randomly grabbed from the Internet, mostly from places with a large number of people and dense distribution, and part B contains 770 images. Part A has 600 randomly grabbed images from the internet, mostly from places with large number of people and densely distributed.Part B has 762 images from busy streets in urban areas, which are more sparsely distributed than A. However, because they are derived from public places, they are more sparsely distributed than A. Part B has 762 images from busy streets in urban areas, which are more sparsely distributed. However, because it comes from the actual surveillance of public places, this sample is more practical for the training and testing of the algorithm model.

In this paper, both datasets are used for metrics testing to prove that the model in this paper has better performance in both dense and sparse scenarios.

2) Evaluation Metrics. The evaluation metrics used in the counting algorithm are mainly the mean absolute error (MAE) and mean square error (MSE), which measure the difference between the number of predictions obtained and the true number, as shown in Equation (20) and Equation (21):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - f\left(x_i\right)\right| \tag{20}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - f\left(x_i\right)\right)^2 \tag{21}$$

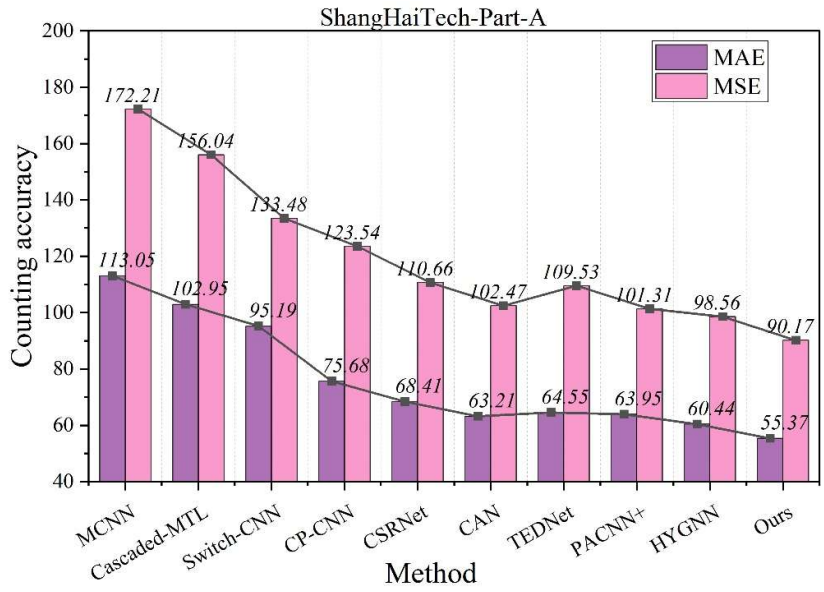
Due to the differences in the target number dimensions between the pictures, the mean absolute percentage error (MAPE), which represents the ratio of the error between the predicted value and the true value in a single test picture to the true value, is used for easy comparison, as shown in Equation (22):

$$MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - f\left(x_i\right)}{y_i}\right| \tag{22}$$

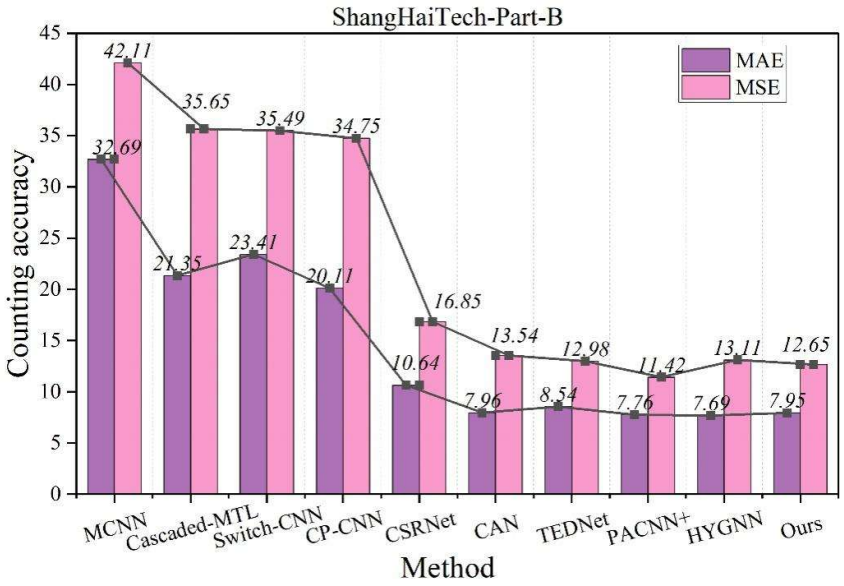
where N represents the total number of pictures used in the experiment, $f\left(x_i\right)$ represents the number of targets predicted in the ird picture, and y_i represents the number of accurately labeled targets in the ith picture.

3) Experimental environment and result analysis. The algorithm model designed in this paper is compared with nine more advanced crowd counting methods, such as MCNN, Cascaded-MTL, Switch-CNN, CP-CNN, CSRNet, etc., in terms of counting accuracy on ShanghaiTech dataset, in order to examine the counting performance of this paper's algorithm.

The counting accuracy comparison of each algorithm on ShanghaiTech dataset is shown in Figure 4. The MAE and MSE metrics of this paper's algorithm and the above methods on this dataset are shown. This paper's algorithm decreases the MAE metric by 23.55% and the MSE metric by 22.72% on the Part-A dataset compared to the algorithm CSRNet. The values of the corresponding metrics on the Part-B dataset correspondingly decreased by 33.84% and 33.20%, respectively. The present algorithm shows considerable counting accuracy and robustness both in the scenarios with dense distribution of crowd targets and in the scenarios with sparse distribution of crowd targets.



(a) ShangHaiTech-Part-A



(b) ShangHaiTech-Part-B

Fig. 4. Algorithm is compared to the counting accuracy in the shanghaitech data set

3.5.2. Counting Accuracy and Reasoning Time. The dataset selected for this section was obtained from the publicly available global tree detection dataset with additions to that dataset collected and secondary processed. The dataset labeling was reprocessed using Matlab R2018b to relabel the detection frames in the original dataset in the form of coordinate points, generating a label file in .mat format, with the number of calibrated points being the true number of trees in the image, and the coordinate positions of the calibrated points reflecting the true spatial distribution information of the trees in the image.The number of images from the GWHD and this paper's dataset is shown in Table 6 The number of images in GWHD and this paper's dataset is shown in Table 6.

Table 6. GWHD and the number of data collection figures in this article

Data set	Image	Gross	The maximum number of	Average number of

	quantity	number	single images	images per image
GWHD2020	3625	156421	120	45.21
GWHD2021	3200	137987	110	42.36
Training Set	2894	132005	125	47.49
Testing Set	756	37065	95	50.85

In order to verify the effectiveness of the algorithm, the present algorithm is compared with other four algorithms in a comparison experiment, and the experimental results are shown in Table 7. Compared with the counting algorithms, the MAE and MSE values of this paper's algorithm are the lowest, i.e., this algorithm has the highest counting accuracy for different tree images. Compared to CSRNet algorithm, the counting accuracy of this paper's algorithm is slightly improved with basically the same inference time. Compared with CAN, it has some advantages in terms of counting accuracy and inference time. Therefore, the overall performance of this paper's algorithm is better than other counting algorithms as seen from the comparison of both counting accuracy and inference time.

Table 7. Experimental results

Model	MAE	MSE	Average single image reasoning time /s
MCNN	13.52	21.53	0.068
Cascaded-MTL	8.95	17.43	0.072
Switch-CNN	4.36	15.66	0.098
CP-CNN	8.67	14.49	0.049
CSRNet	4.62	10.20	0.053
CAN	9.15	6.41	0.071
TEDNet	7.37	9.63	0.039
PACNN+	5.28	10.45	0.045
HYGNN	9.64	11.23	0.067
Ours	4.00	5.57	0.058

3.6. Evidence Text Image Recognition

In addition, in order to comprehensively evaluate the performance of the proposed model in different scenarios, this study was tested on scene text image data from different types of public interest litigation evidence materials. The quantitative distribution of the datasets includes 25613 for signage, 526 for scanning, 1500 for handwriting, 4000 for small tickets, and 5000 for maps.The string-level accuracy of the different scene text image sub-datasets is shown in Table 8. It is sought that the computer vision text image processing technique proposed in this paper can reach more than 80% for the classification and recognition of different types of public interest litigation evidence, which can be further deepened.

Table 8. The string level accuracy of the different scene text image subdata sets

Method	Map	Small ticket	Hand in hand	Scanning	Sign plate
HCADecoder	78.95	81.21	89.45	79.15	83.29
Single image target counting algorithm	89.45	85.43	79.59	83.62	86.74
Total	84.2	83.32	84.52	81.385	85.015

4. Conclusion

This paper proposes a text label recognition algorithm and a monocular image target counting algorithm for the text image recognition and classification needs of public interest litigation evidence review. The feasibility of the proposed methods based on computer vision technology is analyzed by comparing different data sets.

Combined with the public interest litigation evidence electronic data conversion submission and review methods, the use of public interest litigation evidence data for litigation evidence element identification and litigation case causal identification. Public interest litigation evidence for the review of the image is particularly important, there are some images due to the stacking of objects produced by the counting of the unknown situation, affecting the legal proceedings to prove. Combining the image data of dense scene and sparse scene in the dataset, the stacked object number estimation algorithm proposed in this paper combines the planar density map and depth map to carry out MAE, MSE test, counting accuracy and reasoning time judgment. In summary, the text label recognition algorithm and image target counting algorithm proposed in this paper based on computer vision technology can be applied to the review of public interest litigation evidence to assist in public interest litigation.

About the Author

Yongjun Wang was born in Yuzhou, Henan, P.R. China, in 1973. I obtained a master's degree from Henan University in China. I am currently working at the School of law, Henan University of Urban Construction. My main research direction is procedural law and judicial system.

References

[1] Kendall, S. (2021). Reconceptualising reforms to cross-examination: Extending the reliability revolution beyond the forensic sciences. Canberra L. Rev., 18, 36.

[2] Caldwell, H. M., & Elliot, D. S. (2018). Avoiding the Wrecking Ball of a Disastrous Cross Examination: Nine Principles for Effective Cross Examinations with Supporting Empirical Evidence. SCL Rev., 70, 119.

[3] Crozier, W. E., Kukucka, J., & Garrett, B. L. (2020). Juror appraisals of forensic evidence: Effects of blind proficiency and cross-examination. Forensic Science International, 315, 110433.

[4] Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. Psychological Science in the Public Interest, 18(1), 10-65.

[5] Olaleye, O. (2024). CROSS-EXAMINATION; the greatest legal engine for the discovery of the Truth?. Available at SSRN.

[6] McQuiston-Surrett, D., & Saks, M. J. (2017). Communicating opinion evidence in the forensic identification sciences: Accuracy and impact. In Expert Evidence and Scientific Proof in Criminal Trials (pp. 421-451). Routledge.

[7] Griffin, L. K. (2023). False Accuracy in Criminal Trials: The Limits and Costs of Cross-Examination. Tex. L.

Rev., 102, 1011.

[8] Fikiet, M. A., Khandasammy, S. R., Mistek, E., Ahmed, Y., Halámková, L., Bueno, J., & Lednev, I. K. (2019). Forensics: evidence examination via Raman spectroscopy. Physical Sciences Reviews, 4(2), 20170049.

[9] Thompson, W. C., & Scurich, N. (2019). How cross-examination on subjectivity and bias affects jurors' evaluations of forensic science evidence. Journal of forensic sciences, 64(5), 1379-1388.

[10] Bodziak, W. J. (2017). Footwear impression evidence: detection, recovery and examination. CRC Press.

[11] Edmond, G., Cunliffe, E., Martire, K., & San Roque, M. (2018). Forensic science evidence and the limits of cross-examination. Melb. UL Rev., 42, 858.

[12] Szeliski, R. (2022). Computer vision: algorithms and applications. Springer Nature.

[13] Khan, A. A., Laghari, A. A., & Awan, S. A. (2021). Machine learning in computer vision: a review. EAI Endorsed Transactions on Scalable Information Systems, 8(32), e4-e4.

[14] Davies, E. R. (2017). Computer vision: principles, algorithms, applications, learning. Academic Press.

[15] Zhang, X., & Xu, S. (2020, July). Research on image processing technology of computer vision algorithm. In 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL) (pp. 122-124). IEEE.

[16] Nixon, M., & Aguado, A. (2019). Feature extraction and image processing for computer vision. Academic press.

[17] Li, X., & Shi, Y. (2018, August). Computer vision imaging based on artificial intelligence. In 2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS) (pp. 22-25). IEEE.

[18] Patel, R., & Patel, S. (2020). A comprehensive study of applying convolutional neural network for computer vision. International Journal of Advanced Science and Technology, 29(6s), 2161-2174.

[19] Murino, V., Cristani, M., Shah, S., & Savarese, S. (2017). Group and crowd behavior for computer vision. Academic Press.

[20] Bayoudh, K., Knani, R., Hamdaoui, F., & Mtibaa, A. (2022). A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. The Visual Computer, 38(8), 2939-2970.

[21] Feng, X., Jiang, Y., Yang, X., Du, M., & Li, X. (2019). Computer vision algorithms and hardware implementations: A survey. Integration, 69, 309-320.

[22] Nabila Zrira,Anwar Jimi,Mario Di Nardo,Issam Elafi,Maryam Gallab & Redouan Chahdi El Ouazzani. (2024). GCBAM-UNet: Sun Glare Segmentation Using Convolutional Block Attention Module. Applied System Innovation(6),128-128.