Towards Robust Visual Information Extraction in Real World: New Dataset and Novel Solution

Jiapeng Wang[1], Chongyu Liu[1], Lianwen Jin[1,3,*], Guozhi Tang[1], Jiaxin Zhang[1], Shuaitao Zhang[1], Qianying Wang[2], Yaqiang Wu[2,4], Mingxiang Cai[2]

[1]South China University of Technology
[2]Lenovo Research
[3]SCUT-Zhuhai Institute of Modern Industrial Innovation
[4]Xi'an Jiaotong University
{eejpwang, eechongyu.liu}@mail.scut.edu.cn, eelwjin@scut.edu.cn, {eetanggz, eejxzhang, eestzhang}@mail.scut.edu.cn, {wangqya, wuyqe, caimx}@lenovo.com

Abstract

Visual information extraction (VIE) has attracted considerable attention recently owing to its various advanced applications such as document understanding, automatic marking and intelligent education. Most existing works decoupled this problem into several independent sub-tasks of text spotting (text detection and recognition) and information extraction, which completely ignored the high correlation among them during optimization. In this paper, we propose a robust visual information extraction system (VIES) towards real-world scenarios, which is an unified end-to-end trainable framework for simultaneous text detection, recognition and information extraction by taking a single document image as input and outputting the structured information. Specifically, the information extraction branch collects abundant visual and semantic representations from text spotting for multimodal feature fusion and conversely, provides higher-level semantic clues to contribute to the optimization of text spotting. Moreover, regarding the shortage of public benchmarks, we construct a fully-annotated dataset called EPHOIE (https://github.com/HCIILAB/EPHOIE), which is the first Chinese benchmark for both text spotting and visual information extraction. EPHOIE consists of 1,494 images of examination paper head with complex layouts and background, including a total of 15,771 Chinese handwritten or printed text instances. Compared with the state-of-the-art methods, our VIES shows significant superior performance on the EPHOIE dataset and achieves a 9.01% F-score gain on the widely used SROIE dataset under the end-to-end scenario.

Introduction

Recently, visual information extraction (VIE) has attracted considerable research interest owing to its various advanced applications, such as document understanding (Wong, Casey, and Wahl 1982), automatic marking (Tremblay and Labonté 2003), and intelligent education (Kahraman, Sagiroglu, and Colak 2010).

Most existing works for VIE mainly comprise two independent stages, namely text spotting and information extraction. The former aims to locate and recognize the texts, while the latter extracts specific entities based on previous results. Recent studies (Liu et al. 2019; Yu et al. 2020; Xu et al. 2020) revealed that in addition to semantic features, the visual and spatial characteristics of documents also provided abundant clues. Although achieved encouraging results, these approaches still suffered from the following limitations: (1) Although their text spotting models had learned effective representations for detection and recognition, their information extraction modules discarded and then retrieved them again from the OCR results. This resulted in redundant computation, and the discarded features might be more effective than the newly learned ones. (2) The training processes of independent parts were irrelevant, leading to the lack of clues obtained by the information extraction module, while the text spotting module cannot be optimized adaptively according to the fundamental objective. Continuous stages were usually combined to accomplish a common task, while they did not collaborate with each other. To address the limitations mentioned above, in this paper, we propose a robust visual information extraction system towards real-world scenarios called VIES, which is an unified end-to-end trainable framework for simultaneous text detection, recognition and information extraction. VIES introduces vision coordination mechanism (VCM) and semantics coordination mechanism (SCM) to gather rich visual and semantic features from text detection and recognition branches respectively for subsequent information extraction branch and conversely, provides higher-level semantic clues to contribute to the optimization of text spotting. Concurrently, a novel adaptive feature fusion module (AFFM) is designed to integrate the features from different sources (vision, semantics and location) and levels (segment-level and token-level) in information extraction branch to generate more effective representations.

With the development of learning-based algorithms, a comprehensive benchmark conducted for a specific task is a prerequisite to motivate more advanced works. In VIE,

*Corresponding author

SROIE (Huang et al. 2019) is the most widely used one, which concentrates both on the optical character recognition (OCR) and VIE tasks for scanned receipts in printed English. However, it's difficult to satisfy the demand of real-world applications for documents with complex layouts and handwritten texts. To address this issue and promote the development of the field of VIE, we furthermore establish a challenging dataset called Examination Paper Head Dataset for OCR and Information Extraction (EPHOIE), which contains 1,494 images which are collected and scanned from real examination papers of various schools in China, and we crop the paper head regions which contains all key information. The texts are composed of handwritten and printed Chinese characters in horizontal and arbitrary quadrilateral shape. Complex layouts and noisy background also enhance the generalization of EPHOIE dataset. Typical examples are shown in Figure 1.

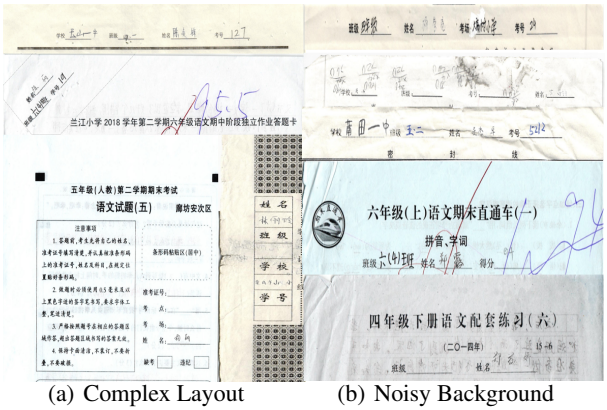
Our main contributions can be summarized as follows:

- We propose a robust visual information extraction system towards real-world scenarios called VIES, which is an unified end-to-end trainable framework for simultaneous text detection, recognition and information extraction.
- We introduce VCM and SCM to enable the independent modules to benefit from joint optimization. AFFM is also designed to integrate features from different sources and levels to boost the entire framework.
- We propose a fully-annotated dataset called EPHOIE, which is the first Chinese benchmark for applications of both text spotting and visual information extraction.
- Our method achieves state-of-the-art performance on both the EPHOIE and widely used benchmarks, which fully demonstrated the effectiveness of the proposed VIES.

Ralated Work

Datasets for Visual Information Extraction For VIE, SROIE (Huang et al. 2019) is the most widely used public dataset that has brought great impetus to this fields. It concentrates on scanned receipts in printed English, and contains complete OCR annotations and key-value pair information labels for each image. (Guo et al. 2019) proposed a Chinese benchmark with fixed layouts including train tickets, passports and business cards. However, the overwhelming majority of images were totally synthetic and only annotated with key-value pair labels without any OCR annotations. In this regard, for the development of both OCR and VIE tasks in Chinese documents and handwritten information, a comprehensive dataset with complex background, changeable layouts and diverse text styles towards real-world scenarios is in great demand.

Methods for Visual Information Extraction In recent years, VIE methods have achieved encouraging improvement. Early works mainly used rule-based (Esser et al. 2012; MUSLEA 1999) or template matching (Huffman 1995) methods, which might led to the poor generalization. With the development of deep learning, more researchers converted the results obtained by text spotting into plain texts, and then extracted feature embeddings for a subsequent sequence labeling model such as BiLSTM-CRF (Lample et al. 2016) to obtain the final entities. However, the lack of visual and location information often led to poor performance.



(a) Complex Layout (b) Noisy Background

Figure 1. Some typical and challenging instances in EPHOIE. (a) Complex layout. (b) Noisy background.

The fact that visual and spatial features of documents also play a vital role in information extraction has been recognized by recent works. Typical methods such as Post-OCR parsing (Hwang et al. 2019) took bounding box coordinates into consideration. LayoutLM (Xu et al. 2020) modeled the layout structure and visual clues of documents based on the pre-training process of a BERT-like model. GraphIE (Qian et al. 2019), PICK (Yu et al. 2020) and (Liu et al. 2019) tried to use Graph Neural Networks (GNNs) to extract global graph embeddings for further improvement. CharGrid (Katti et al. 2018) used CNNs to integrate semantic clues contained in input matrices and the layout information simultaneously. However, these existing traditional methods only focused on the performance related to the information extraction stage, but ignored the preconditioned OCR module.

At present, more related works of VIE were gradually developing towards end-to-end manner. (Guo et al. 2019) generated feature maps directly from input image and used several entity-aware decoders to decode all the entities. However, it could only process documents with fixed layout and its efficiency could be significantly reduced as the number of entities increases. (Carbonell et al. 2020) localized, recognized and classified each text segment in image, which was difficult to handle the situation where a text segment was composed of characters with different categories. (Zhang et al. 2020) proposed an end-to-end trainable framework to solve VIE task. However, it focused more on the performance of entity extraction and can only be applied to the scenarios where the OCR task was relatively simple.

Examination Paper Head Dataset for OCR and Information Extraction

In this section, we introduce the new Examination Paper Head Dataset for OCR and Information Extraction (EPHOIE) benchmark and its characteristics.

Dataset	Year	Scenario	Language	Image Number	Text Shape	Script	Entities
SROIE	2019	Scanned receipts	English	973	H	Printed	4
EPHOIE	2020	Scanned paper head	Chinese	1494	H, Q	Printed/Handwritten	10

Table 1. Comparison between EPHOIE and SROIE. 'H' or 'Q' denotes horizontal or arbitrary quadrilateral text.

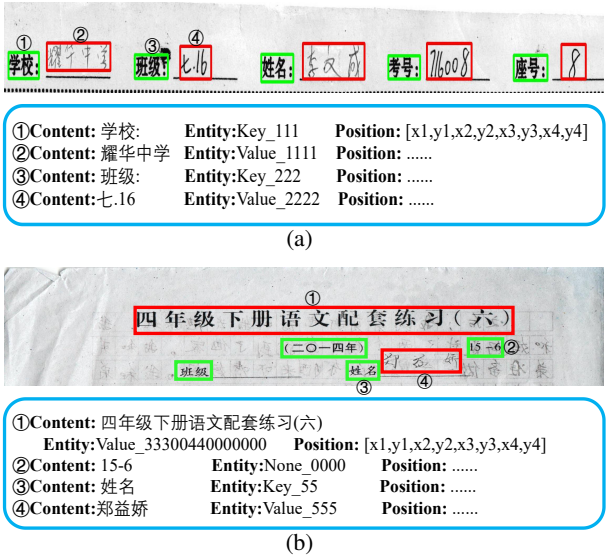


Figure 2. Examples of annotations in EPHOIE. In *Entity* field, 'Key' or 'Value' indicates it's key or value of an entity respectively, whereas 'None' indicates neither of them. The different numbers in *Entity* denotes different categories.

Dataset Description To the best of our knowledge, the EPHOIE benchmark is the first public dataset for both OCR and VIE tasks in Chinese and aims to motivate more advanced works in the fields of both document intelligence and VIE. It contains 1,494 images with 15,771 annotated text instances, including handwritten and printed characters. It is divided into a training set with 1,183 images and a testing set with 311 images respectively. All the images in EPHOIE are collected and scanned from real examination papers of various schools with the diversity of text types and layout distribution. The statistic of our dataset and the comparison with the most widely used public benchmark SROIE are shown in Table 1. For EPHOIE, we only crop the paper head regions that contain all key information.

Annotation Details The detailed annotation forms of EPHOIE are presented in Figure 2. As there exist both horizontal and arbitrary quadrilateral texts, four vertices were required to surround them. In addition to annotating bounding boxes for text detection, text content is also required for both text recognition and information extraction. We annotate all the texts appearing on the image, while additionally label the entity key-value pair for all key information. The number string in *Entity* denotes the category of each token, since there may exists multiple entities in a single segment.

Methodology

The overall framework of our VIES is illustrated in Figure 3. It consists of a shared backbone and three specific branches of text detection, recognition and information extraction. Given an input image, the text spotting branches are responsible for not only localizing and recognizing all the texts in it, but also providing abundant visual and semantic features through vision and semantics coordination mechanisms for subsequent networks. The adaptive feature fusion module in information extraction branch first gathers these abundant representations with additional spatial features extracted from detected boxes to adaptively generate fused features in decoupled levels (segment-level and token-level). In this part, the multi-head self-attention mechanism is introduced to allow each individual to freely attend to all the others. Then, the features in decoupled levels are re-coupled and finally the specific entities are distinguished from recognized strings using sequence labeling module.

Text Detection with Vision Coordination Mechanism (VCM)

Accurate text detection is the prerequisite for text recognition and information extraction. An intuitive idea to boost the detection branch is to make the IE branch provide feedback guidance in an end-to-end manner during training.

Given an input image, our VIES first uses the shared backbone to extract high-level feature representations X. Then, the detection branch takes X as input and outputs boxes B, confidence scores C and even binary masks M for arbitrary quadrilateral texts:

$$B, C, M = TextDetection(X) \tag{1}$$

Here, we introduce an innovative vision coordination mechanism (VCM), which can effectively transfer rich visual features F_{vis} from the detection branch to the IE branch and conversely provide additional supervised information to contribute to the optimization of the detection branch. It can be shown in Figure 4(a) and defined as follows:

$$F_{vis} = Linear(AvgPool(Conv2D(RegionPool(X, B)))) \tag{2}$$

Here, $RegionPool$ denotes region feature pooling methods such as RoIPooling (Girshick 2015) and RoIAlign (He et al. 2017). $AvgPool$ reduces both height and width dimension to unit size. $Linear$ is learned projection to transform F_{vis} into d channels.

For visually rich documents, the key visual clues such as shapes, fonts and colors have been integrated in F_{vis}. The gradients of IE branch can also help the detection branch learn more general representations that are beneficial to the entire framework.

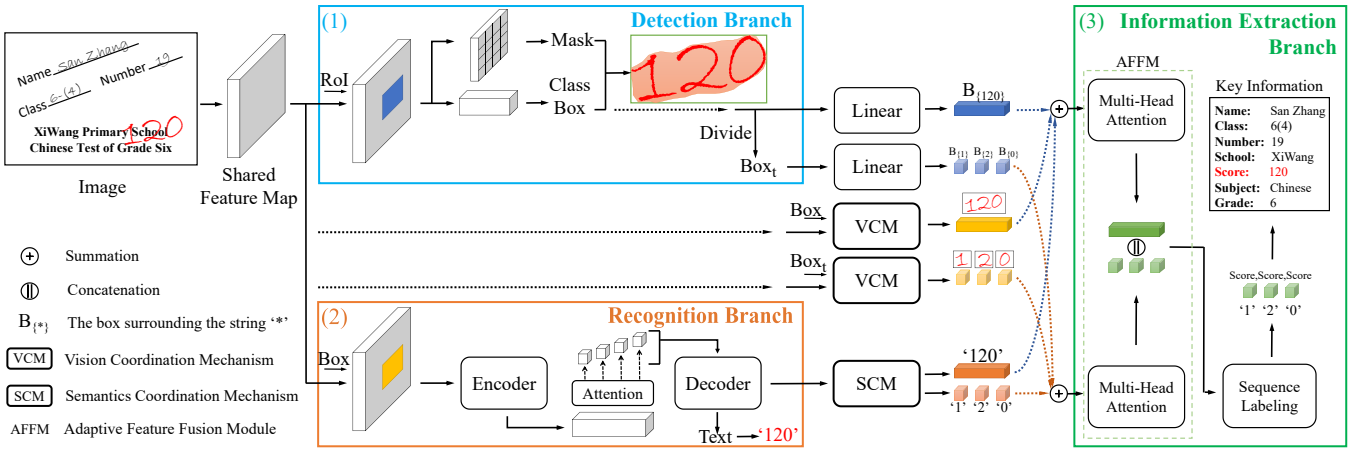


Figure 3. The overall framework of VIES. It consists of a shared backbone network and three specific branches: (1) text detection, (2) text recognition and (3) information extraction. Box_t denotes boxes of single tokens which divided from the box of entire text segment Box.

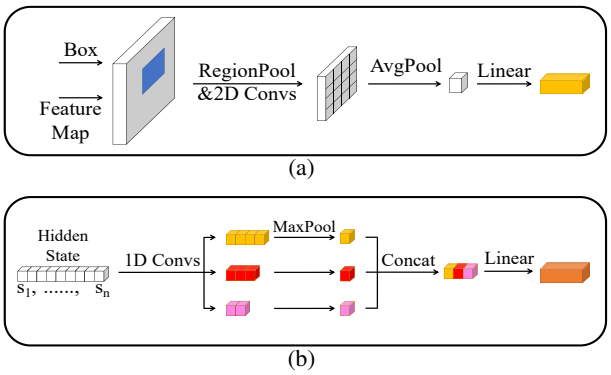


Figure 4. The detailed structures of Vision and Semantics Coordination Mechanisms. (a) Vision Coordination Mechanism (VCM). (b) Semantics Coordination Mechanism (SCM).

Text Recognition with Semantics Coordination Mechanism (SCM)

Text recognition greatly limits the upper bound of the performance of the entire system. If the recognized strings are less accurate, it will always be useless no matter how powerful the IE branch is. Based on this consideration, whether semantic supervision of IE stage can boost the recognition branch is particularly critical.

In our VIES, given the shared features X, high-level representations in specific text regions are collected and fed into an encoder to extract the input feature sequence $H = (h_1, h_2, \cdots, h_N)$, where N is the feature length. Then, an attention-based decoder (Bahdanau, Cho, and Bengio 2015) is adopted to recurrently generate the hidden states $S = (s_1, s_2, \cdots, s_M)$ by referring to the history of recognized characters and H, where M indicates the maximum decoding step. Finally, the output text sequence $O = (o_1, o_2, \cdots, o_M)$ is computed using S.

Here, we introduce our semantics coordination mechanism (SCM) to establish the bidirectional semantics flow between our recognition branch and IE branch. The hidden states S in our recognition branch contain high-level semantic representations of each single token in every decoding step. Therefore, we regard it as token-level semantic features $F_{sem,t}$ and send it to the IE branch:

$$F_{sem,t} = (s_1, s_2, \cdots, s_M) = S, \quad where \quad F_{sem,t_i} = s_i \tag{3}$$

Here, $F_{sem,t_i} \in \mathbb{R}^d$ denodes the d-dimensional vector corresponding to the i-th token in the segment.

Note that, the segment-level semantic features $F_{sem,s}$ also greatly affect the category characteristics. Further, $F_{sem,t}$ captures local clues and $F_{sem,s}$ contains global information, indicating that they are complementary to each other. Inspired by the previous works (Zhang, Zhao, and LeCun 2015; Kim 2014) which adopted CNNs to integrate holistic expression for each sentence from the words' or characters' embeddings, our VIES generates the summarization of each segment $F_{sem,s}$ from $F_{sem,t}$ as follows:

$$F_{sem,t_{1:n}} = F_{sem,t_1} \oplus \cdots \oplus F_{sem,t_n}, \tag{4}$$

$$c_i = Conv1D_i(F_{sem,t_{1:n}}), \tag{5}$$

$$c_i = MaxPool1D(c_i), \quad i = 1, \ldots, n_c \tag{6}$$

$$F_{sem,s} = Linear(c_1 \oplus \cdots \oplus c_{n_c}) \tag{7}$$

Here, $\oplus$ is the concatenation operator, n is the length of the current segment and n_c is the number of 1D convolution kernels. Note that all 1D operations are carried out over the *length* dimension.

The overall structure of SCM is illustrated in Figure 4(b). In this way, the extracted competent semantic representations can be passed forward directly, and the higher-level semantic constraints of IE branch can guide the training process of recognition branch.

Information Extraction with Adaptive Feature Fusion Module (AFFM)

Information extraction requires the most comprehensive and expressive representations to distinguish specific entities from recognized strings. Besides the visual and semantic features provided by text spotting branches above, our IE branch further extracts spatial features from text boxes and decouples token-level representations so that the relatively accurate clues can be obtained regardless of whether the token is attributed to the wrong string or whether the string is over- or under-cut, which often occur in text detection owing to the complex background and the variety of shapes and styles. To encode location information, we generate spatial features F_{spt} from relative bounding box coordinates as:

$$F_{spt} = Linear([\frac{x_{min}}{W_{img}}, \frac{y_{min}}{H_{img}}, \frac{x_{max}}{W_{img}}, \frac{y_{max}}{H_{img}}]) \quad (8)$$

where W_{img} and H_{img} are image width and height respectively. $Linear$ is used to transform F_{spt} into d channels which is the same as that in the visual and semantic features above. We intuitively divide the box of the entire segment evenly along its longest side into several single tokens' boxes B_t according to the length of recognized strings. Then the token-level visual features $F_{vis,t}$ and spatial features $F_{spt,t}$ can be generated according to B_t.

After acquiring the features of multi levels from multi sources as representations in a learned common embedding space, our adaptive feature fusion module (AFFM) introduces two multi-layer multi-head self-attention modules combined with linear transforms to first enrich all projected vectors at different fine-granularities respectively. The multimodal features are summarized and followed by the layer normalization to generate comprehensive representations of each individual. Then, it serves as the K, Q and V in the scaled dot-product attention, which can expressed as:

$$Attention(Q, K, V) = softmax(\frac{Q\,K^T}{\sqrt{d}})V, \quad (9)$$
$$where \quad Q, K, V = LayerNorm(F_{vis} + F_{sem} + F_{spt})$$

$$F_* = Concat(head_1, head_2, \cdots, head_h)\,W^O, \quad (10)$$
$$where \quad head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

where W_i^Q, W_i^K, W_i^V and W^O are projection parameters and F_* denotes the fused features in segment-level or token-level. Then, we re-couple them to combine global and local information:

$$\overline{F}_{j,t_i} = F_{j,t_i} \oplus F_j, \quad (11)$$
$$where \quad i = 1, ..., n_j, \quad j = 1, ..., n_s$$

where n_s is the number of text segments and n_j is the length of j-th segment. $\overline{F}_{j,t_i}$ constitutes input feature sequence for subsequent sequence labeling module.

Sequence Labeling After feature recoupling, we feed the input feature sequence into standard BiLSTM-CRF (Lample et al. 2016) for entity extraction. Intuitively, segment embedding provides extra global representations. The concatenated features are fed into a BiLSTM network to be encoded, and the output is further passed to a fully connected network and then a CRF layer to learn the semantics constraints in an entity sequence.

Optimization Strategy

In the training phase, our proposed framework can be trained in an end-to-end manner with the weighted sum of the losses generated from three branches of text detection, recognition and information extraction:

$$L = L_E + \lambda_D\, L_D + \lambda_R\, L_R \quad (12)$$

where λ_D and λ_R are hyper-parameters that control the tradeoff between losses. L_D and L_R are losses of text detection and recognition branches respectively, and L_E is the loss of information extraction branch.

L_D consists of losses for text classification, box regression and mask identification respectively, as defined in (He et al. 2017). L_R adopts CrossEntropyLoss between output text sequence O and ground truth text sequence. CRFLoss is also adopted as L_E for information extraction.

Experiments

Implement Details

We adopt Mask R-CNN (He et al. 2017) as our text detection branch with ResNet-50 (He et al. 2016) followed by FPN (Lin et al. 2017) as its backbone. We use LSTM (Hochreiter and Schmidhuber 1997) in attention mechanism for text recognition. In SCM, the sizes of three 1D convolutions are 2, 3 and 4. In AFFM, we set the number of heads and sublayers is 4 and 3, the dimension of input features and linear transforms is 256 and 512 respectively.

The hyper-parameters λ_D and λ_R are all set to 1.0 in our experiments. In end-to-end training phase, the initial learning rate is set as 0.1 for text spotting branches and 1.0 for information extraction branch with ADADELTA (Zeiler 2012) optimization after sufficient pre-training of the former. We also decrease it to a tenth every 25 epoches for two times.

Ablation Study

In this section, we evaluate the influences of multiple components of the proposed framework on the EPHOIE dataset.

Effect of End-to-End Optimization To explore the effects of end-to-end optimization manner introduced by VCM and SCM, we perform the following ablation studies and the results are presented in Table 2. **Baseline** denotes the gradients generated by information extraction branch are detached and cannot be back-propagated to the text spotting part. We select two other advanced structures – graph attention network (GAT) (Veličković et al. 2018) similar to (Liu et al. 2019) and the information extraction module in TRIE (Zhang et al. 2020), then combine them with the optimization methods both in TRIE and our VIES for detailed comparison.

From Table 2, it can be seen that **VIES(Ours)** outperforms four counterparts in all of text detection, recognition and information extraction tasks by a large margin,

Task	Mesure	Optimization Strategy				
		Baseline	TRIE	E2E(Ours) + IE stage in TRIE	E2E(Ours) + GAT	**VIES(Ours)**
Detection	IoU=0.5, F1-Score	97.00	97.31	97.36	97.10	**97.48**
	IoU=0.6, F1-Score	96.03	**96.33**	96.25	96.05	96.15
	IoU=0.7, F1-Score	92.72	93.02	92.92	92.48	**93.06**
	IoU=0.8, F1-Score	78.89	78.30	79.04	78.16	**79.60**
Recognition	AR	96.43	96.28	96.40	95.59	**96.79**
	LA	93.98	93.53	93.78	93.55	**94.52**
Information Extraction	F1-Score	80.31	81.24	82.21	80.51	**83.81**

Table 2. Effect of End-to-End Optimization. **LA** indicates the whole line accuracy. **A + B** indicates the combination of optimization method A and IE structure B.

Setting	The design of VCM	F1-Score
1)	Adopting **RoIPooling** as $RegionPool$	83.28
Ours	Adopting **RoIAlign** as $RegionPool$	**83.81**
Setting	**The design of SCM**	**F1-Score**
2)	Taking a further decoding step after predicting token <END>	80.26
3)	Extracting from H with 1D convolutions	82.59
Ours	Extracting from S with 1D convolutions	**83.81**

Table 3. Effects of VCM and SCM. H denotes the input feature sequence and S denotes hidden states in the recognition branch.

Setting	Source			F1-Score
	Semantics	Vision	Location	
(1)	✓			80.51
(2)	✓	✓		83.25
(3)	✓		✓	81.58
(4)	✓	✓	✓	**83.81**

Table 4. Effect of multi-source features.

which reveals the superiority of our framework. **TRIE** performs better in detection task under low IoU and information extraction task than **Baseline**, however, the performance of detection with high IoU and recognition are both evidently reduced. This indicates that the improvement of final achievements does not always mean the overall progress of the entire system. Compared to it, **E2E(Ours) + IE stage in TRIE** achieves comparable detection results under low IoU and significantly better performances on other counts, which fully verifies the advantage of our optimization strategy. Moreover, **VIES(Ours)** shows significant gains in all tasks over **E2E(Ours) + IE stage in TRIE** and **E2E(Ours) + GAT**, revealing both the effectiveness of modeling of our AFFM and the fact that, the co-training method needs to be built under careful considerations to take full advantage of its role.

Effects of VCM and SCM Here we conduct the following experiments to verify the effects of our VCM and SCM. We design several intuitive and effective structures for them and the results are shown in Table 3. It totally indicates that although combining text spotting branches and IE branch is a relatively intuitive idea, how to make the best use of it needs a comprehensive design. Our VCM and SCM can maximize the benefits of end-to-end optimization.

Effect of multi-source features We conduct the following experiments to verify the effectiveness of multi-source features in AFFM, and the results are presented in Table 4. It can be observed that further fusion of the multi-modality representations in Setting **(4)** provides the best performance.

Semantic features are the most distinguishable ones for information extraction. As shown in Setting **(1)**, our method can achieve satisfactory performance by using only semantic features. Moreover, these features are provided from our recognition branch, which may be more effective than the re-extracted ones under traditional routines.

Visual features such as fonts and colors which containing rich semantic clues is also crucial. This brings significant performance gains, which can be observed in Setting **(2)**. When semantics of different texts are highly similar, visual features play the decisive role.

Note that, introducing spatial features in Setting **(3)** outperforms Setting **(1)** slightly, revealing that the shape and location of texts also plays a critical role in representing semantic meanings. Through the adaptive feature fusion process, the expressive features mentioned above belong to different individuals are allowed freely attending to all the others, which enables modeling both inter- and intra- segment relations in a homogeneous manner. The negative effects of errors from different sources can also be mitigated here.

Comparison with the State-of-the-Arts

To comprehensively evaluate our framework, we compared it with several state-of-the-art methods. It is notable that, we re-implement them based on the original papers or source codes if available on open-source platforms.

Results on EPHOIE Dataset As shown in Table 5, our method exhibits superior performance on EPHOIE. (Liu et al. 2019), TRIE (Zhang et al. 2020) and VIES which introduce multimodal representations outperform counterparts by significant margins. Under the **End-to-End** setting where the OCR results are less accurate, the robustness of our

Setting	Method	Entities										
		Subject	Test Time	Name	School	Examination Number	Seat Number	Class	Student Number	Grade	Score	**Mean**
Ground Truth	(Lample et al. 2016)	98.51	**100.0**	98.87	98.80	75.86	72.73	94.04	84.44	98.18	69.57	89.10
	(Liu et al. 2019)	98.18	**100.0**	99.52	**100.0**	88.17	86.00	97.39	80.00	94.44	81.82	92.55
	GraphIE (Qian et al. 2019)	94.00	**100.0**	95.84	97.06	82.19	84.44	93.07	85.33	94.44	76.19	90.26
	TRIE (Zhang et al. 2020)	98.79	**100.0**	99.46	99.64	88.64	85.92	97.94	84.32	97.02	80.39	93.21
	VIES(Ours)	**99.39**	**100.0**	**99.67**	99.28	**91.81**	**88.73**	**99.29**	**89.47**	**98.35**	**86.27**	**95.23**
End-to-End	(Lample et al. 2016)	82.08	89.95	72.61	83.29	62.18	64.56	66.87	63.68	81.17	53.09	71.95
	(Liu et al. 2019)	84.12	90.61	78.35	87.25	68.60	64.45	71.56	68.39	82.19	55.22	75.07
	TRIE (Zhang et al. 2020)	85.92	92.20	85.94	91.92	73.63	69.01	79.91	78.00	**83.82**	62.74	80.31
	VIES(Ours)	**86.14**	**93.50**	**90.35**	**95.47**	**77.72**	**76.05**	**85.65**	**81.05**	83.49	**68.62**	**83.81**

Table 5. Performance (F1-Score) comparison of the state-of-the-art algorithms on the EPHOIE dataset. **Ground Truth** means using ground truth bounding boxes and texts as inputs for information extraction branch, and **End-to-End** denotes using same predictions from text spotting branches instead.

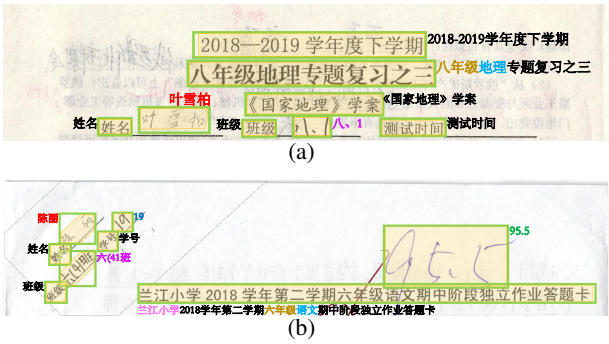


Figure 5. Examples of prediction results of VIES on EPHOIE. Different colors denotes different entities.

Setting	Method	F1-Score
Ground Truth	(Lample et al. 2016)†	90.85
	LayoutLM (Xu et al. 2020)	95.24
	(Liu et al. 2019)	95.10
	PICK (Yu et al. 2020)	96.12
	TRIE (Zhang et al. 2020)	**96.18**
	VIES(Ours)	96.12
End-to-End	NER (Ma and Hovy 2016)†	69.09
	Chargrid (Katti et al. 2018)†	78.24
	(Lample et al. 2016)	78.60
	(Liu et al. 2019)	80.76
	TRIE (Zhang et al. 2020)	82.06
	VIES(Ours)	**91.07**
Competition (Huang et al. 2019)	Rank 1	90.49
	Rank 2	89.70
	Rank 3	89.63

Table 6. Performance comparison of the state-of-the-art algorithms on SROIE dataset. † indicates the result is reported in (Zhang et al. 2020). **Competition** shows the performance of the top three methods during ICDAR 2019 SROIE Competition which inevitably introduced techniques such as model ensemble and complex post-processing.

pipeline is more evident. Its reasonable design effectively reduces the negative effects caused by text spotting errors. Some examples of qualitative results of VIES are shown in Figure 5.

Results on SROIE Dataset The results of experiments on SROIE dataset are shown in Table 6. Our method achieves competitive results under **Ground Truth** setting and outperforms the state-of-the-art results by significant margins (**from 82.06 to 91.07**) under **End-to-End** setting. The methods in **Competition** may inevitably introduce model ensemble techniques for each tasks and complex post-processing. However, our VIES achieves even better results using only a single framework with light-weight network structures. And we only introduce simple regularizations to correct the format of *Total* and *Date* results.

Compared with EPHOIE, the layout of scanned receipts is relatively fixed, the font style is less changeable and there exists less noise in the background. In such a relatively simple scenario, the superiority of our method is further confirmed.

Conclusion

In this paper, we propose a robust visual information extraction system (VIES) towards real-world scenarios, which is an unified end-to-end trainable framework for simultaneous text detection, recognition and information extraction. Additionally, we propose a fully-annotated dataset called EPHOIE, which is the first Chinese benchmark for both OCR and VIE tasks. Extensive experiments demonstrate that our VIES achieves superior performance on the EPHOIE dataset and has 9.01% F-score gains campared with the previous state-of-the-art methods on the widely used SROIE dataset under the end-to-end information extraction scenario.

Visual information extraction is a challenging task in the cross domain of natural language processing and computer vision. Many issues have not been well addressed, including complex layouts and background, over-reliance on complete annotations and continuous accumulation of errors. Therefore, it remains an open research problem and deserves more attention and further investigation.

Acknowledgments

This research is supported in part by NSFC (Grant No.: 61936003, 61771199), GD-NSF (No. 2017A030312006), the National Key Research and Development Program of China (No. 2016YFB1001405), Guangdong Intellectual Property Office Project (2018-10-1), and Guangzhou Science, Technology and Innovation Project (201704020134).

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Carbonell, M.; Fornés, A.; Villegas, M.; and Lladós, J. 2020. A Neural Model for Text Localization, Transcription and Named Entity Recognition in Full Pages. *Pattern Recognition Letters* 219–227.

Esser, D.; Schuster, D.; Muthmann, K.; Berger, M.; and Schill, A. 2012. Automatic indexing of scanned documents: a layout-based approach. In *Document Recognition and Retrieval XIX*, volume 8297, 118–125.

Girshick, R. B. 2015. Fast R-CNN. *In ICCV* 1440–1448.

Guo, H.; Qin, X.; Liu, J.; Han, J.; Liu, J.; and Ding, E. 2019. EATEN: Entity-aware attention for single shot visual text extraction. In *ICDAR*, 254–259.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *ICCV*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Huang, Z.; Chen, K.; He, J.; Bai, X.; Karatzas, D.; Lu, S.; and Jawahar, C. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *ICDAR*, 1516–1520.

Huffman, S. B. 1995. Learning information extraction patterns from examples. In *IJCAI*, 246–260.

Hwang, W.; Kim, S.; Seo, M.; Yim, J.; Park, S.; Park, S.; Lee, J.; Lee, B.; and Lee, H. 2019. Post-OCR parsing: building simple and robust parser via BIO tagging. In *Workshop on Document Intelligence at NeurIPS 2019*.

Kahraman, H. T.; Sagiroglu, S.; and Colak, I. 2010. Development of adaptive and intelligent web-based educational systems. In *AICT*, 1–5.

Katti, A. R.; Reisswig, C.; Guder, C.; Brarda, S.; Bickel, S.; Höhne, J.; and Faddoul, J. B. 2018. Chargrid: Towards Understanding 2D Documents. In *EMNLP*, 4459–4469.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP*, 1746–1751.

Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 2016. Neural Architectures for Named Entity Recognition. In *NAACL-HLT*, 260–270.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *CVPR*, 2117–2125.

Liu, X.; Gao, F.; Zhang, Q.; and Zhao, H. 2019. Graph Convolution for Multimodal Information Extraction from Visually Rich Documents. In *NAACL-HLT*, 32–39.

Ma, X.; and Hovy, E. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *ACL*, 1064–1074.

MUSLEA, I. 1999. Extraction patterns for information extraction tasks: A survey. In *Proc. AAAI-99 Workshop on Machine Learning for Information Extraction, Orlando, Florida, 1999*, 1–6.

Qian, Y.; Santus, E.; Jin, Z.; Guo, J.; and Barzilay, R. 2019. GraphIE: A Graph-Based Framework for Information Extraction. In *NAACL-HLT*, 751–761.

Tremblay, G.; and Labonté, É. 2003. Semi-automatic marking of java programs using junit. In *EISTA*, 42–47.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *ICLR*.

Wong, K. Y.; Casey, R. G.; and Wahl, F. M. 1982. Document analysis system. *IBM journal of research and development* 26(6): 647–656.

Xu, Y.; Li, M.; Cui, L.; Huang, S.; Wei, F.; and Zhou, M. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *ACM-SIGKDD*, 1192–1200.

Yu, W.; Lu, N.; Qi, X.; Gong, P.; and Xiao, R. 2020. PICK: Processing Key Information Extraction from Documents using Improved Graph Learning-Convolutional Networks. In *ICPR*.

Zeiler, M. D. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* .

Zhang, P.; Xu, Y.; Cheng, Z.; Pu, S.; Lu, J.; Qiao, L.; Niu, Y.; and Wu, F. 2020. TRIE: End-to-End Text Reading and Information Extraction for Document Understanding. In *ACM-MM*.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. In *NIPS*, 649–657.