Integrating Scene Text and Visual Appearance for Fine-Grained Image Classification

XIANG BAI[1], (Senior Member, IEEE), MINGKUN YANG[1], PENGYUAN LYU[1], YONGCHAO XU[1,2], AND JIEBO LUO[3], (Fellow, IEEE)

[1]School of Electronics Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China
[2]EPITA Research and Development Laboratory, 94270 Le Kremlin-Bicêtre, France
[3]Department of Computer Science, University of Rochester, Rochester, NY 14627, USA

Corresponding author: Yongchao Xu (yongchaoxu@hust.edu.cn)

ABSTRACT Text in natural images contains rich semantics that is often highly relevant to objects or scene. In this paper, we focus on the problem of fully exploiting scene text for visual understanding. The main idea is combining word representations and deep visual features in a globally trainable deep convolutional neural network. First, the recognized words are obtained by a scene text reading system. Next, we combine the word embedding of the recognized words and the deep visual features into a single representation that is optimized by a convolutional neural network for fine-grained image classification. In our framework, the attention mechanism is adopted to compute the relevance between each recognized word and the given image, which further enhances the recognition performance. We have performed experiments on two datasets: con-text dataset and drink bottle dataset, which are proposed for fine-grained classification of business places and drink bottles, respectively. The experimental results consistently demonstrate that the proposed method of combining textual and visual cues significantly outperforms classification with only visual representation. Moreover, we have shown that the learned representation improves the retrieval performance on the drink bottle images by a large margin, making it potentially powerful in product search.

INDEX TERMS Scene text, fine-grained classification, convolutional neural networks, attention mechanism, product search.

I. INTRODUCTION

Combining multi-modal features for the recognition or index of visual data is the inevitable way to automatically understanding image/video content in the recent years [1]–[3]. The multi-modal visual data are often correlated and complementary to each other in visual understanding, especially when they originate from the same source. In this paper, we are concerned on how to efficiently integrate visual and textual cues for fine-grained image classification with convolutional neural networks. The goal of fine-grained classification is to assign the labels to the images having subtle differences in visual appearance such as animal species, product types, and place types. Such a problem is quite challenging, as relevant differences that are not obvious may not be accurately distinguished by a typical classification model. Even for human

perception, domain-specific knowledge is often required for fine-grained classification.

Scene text, which frequently appears in natural scenes including road signs, street views, product packages, and license plates, often possesses rich and precise meanings that are highly related to semantics of the object or scene in the same image. For instance, texts on the buildings or wrapping bags are quite useful for distinguishing among their categories. With the recent developments in text detection and recognition in the wild [4]–[7], it has been shown that textual cues are very beneficial to fine-grained classification, especially in the classification of business places such as bakery, cafe and bookstore [8].

Scene text detection and recognition has been an active research area in both computer vision and document

FIGURE 1. The recognized words in some scene images obtained by a text reading system.

analysis communities. A large number of novel approaches are proposed for the localization and recognition of the text embodied in natural images and videos. Such approaches enable us to automatically extract the text information, which can be considered as an additional information for image classification. In this paper, the recognized words in natural scenes are used as a kind of input features for an image classification model. However, given an image, not every recognized word must have the relation to it. As shown in Figure 1, the recognized words such as 'restaurant', 'bakery', 'pharmacy', 'bread', 'hotel' are very relevant to the scene semantics, but some words like 'grand', 'great', 'edison' don't have the direct connection to them. This requires a model to measure the relevance of each recognized word to the semantics of objects or scenes, in order to make full use of textual information in image classification. It should be mentioned that the existing robust reading system cannot guarantee that all the words embodied in the images can be correctly detected and/or recognized. Yet, even under such a situation, the recognized words are still very helpful for image classification as demonstrated by our experiments.

There are a few existing studies that combine textual and visual cues for cross-domain feature learning [1], image classification [2], or joint inference of objects and scenes [3]. However, it is important to note that the textual cues used in these works are not extracted directly from images. Rather, they are descriptive texts associated with images. Furthermore, to the best of our knowledge, there has been little research on combining textual and visual cues for fine-grained image classification. Therefore, it is imperative to develop such technologies for many fine-grained classification tasks such as the recognition or indexing of films, products, logos, places, athletes, advertisements, and so on, to fulfill the needs of many real-world applications. Since such scenarios often contain texts that are highly relevant and complementary to the visual appearances of objects or scenes, the text semantics are quite helpful for enhancing the performances of fine-grained classification.

Recently, deep neural networks have been the mainstream techniques for visual understanding, thanks to their great successes in various vision tasks. In this paper, we focus on designing a proper deep neural network architecture for fine-grained classification, which is able to efficiently fuse the semantics of text and visual cues. We firstly perform word detection and recognition in the images with a recent state-of-the-art scene text reading system [9]. Then, word embedding is adopted to transfer the recognized words into a vector space. Finally, the word features are used as an extra input to a typical convolutional neural network [10] for fine-grained classification. The proposed unified framework for integrating text and visual features is illustrated in Figure 2, which is end-to-end trainable.

Our contributions in this paper are three folds:

1) We present a simple yet effective pipeline, using the state-of-the-art components for scene text and visual recognition, for fine-grained image classification through combining textual and visual cues. Our results show that the meanings of recognized words are often complementary to the visual information, resulting in the significant improvements on fine-grained classification.

2) We develop an end-to-end trainable neural network with the attention mechanism for implicitly exploiting the semantic relationships between textual and visual cues. Such a unified framework is more elegant and makes the training and test process much simpler, compared with other existing alternatives designed for cross-domain representations. Besides, the end-to-end pipeline does not require to tune each sub-modular, leading to more appropriate learned features for the task at hand.

3) We collect and will release a new image dataset consisting of various types of drink bottle images. Our proposed method works well on drink bottle image classification and retrieval, demonstrating its potential in product search.

The remainder of the paper is organized as follows. Section II briefly reviews some related works. In Section III, we describe in detail the proposed framework on integrating textual and visual cues for fine-grained classification. Section IV provides some experimental results, comparisons, and analyses on two datasets. Finally, conclusion remarks and future works are given in Section V.

II. RELATED WORKS
A. FINE-GRAINED CLASSIFICATION
Recently, fine-grained image classification is an active topic in some domains, such as animal species [12], plant species [13] and man-made objects [14]. Different from the base-class image classification [15], fine-grained image classification needs to distinguish images with more subtle differences among object classes. Discerning the subtle differences among similar classes is more challenging. Many recent works rely on relevant parts of fine-grained classes for domain specific knowledge. Examples are the recent works in [16] and [17] that use salient parts in images.
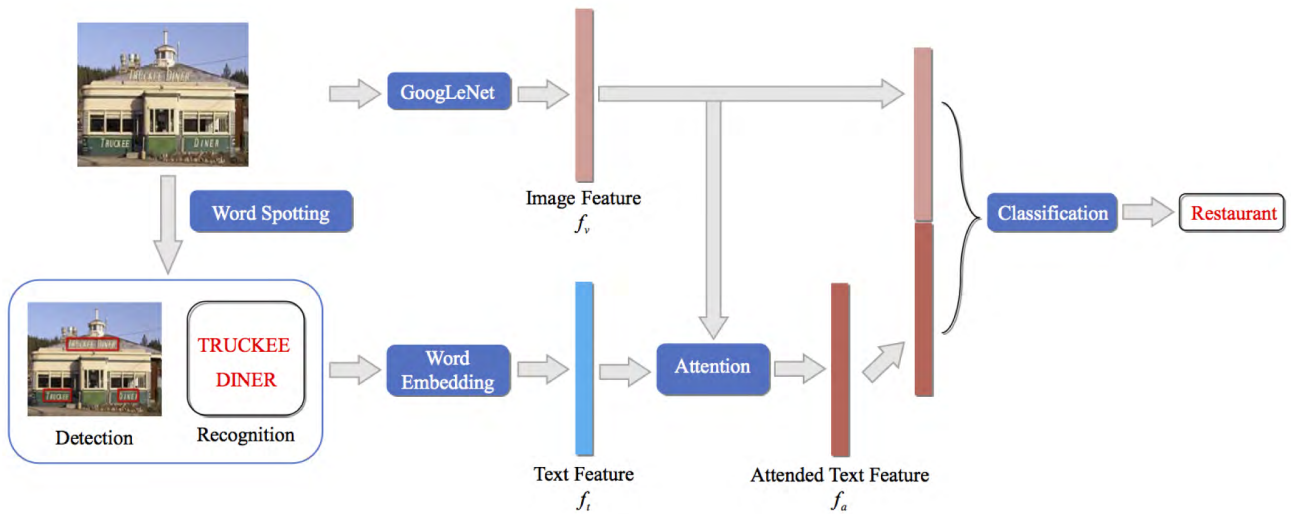
FIGURE 2. Pipeline of the proposed model. Given an input image *I*, the GooLeNet [10] CNN extracts a 1024-dimensional feature vector as the visual representation. A state-of-the-art text detector [9] is used to localize words in the image, followed by a text recognizer [11]. The recognized words are transformed into the textual representation using word embedding. Then the method uses semantic attention mechanism to filter out some noisy words. At last, the multi-modal features are fused for fine-grained classification.

Semantic object parts are used in [12] to isolate the subtle differences. Xiao *et al.* [18] used three attention models to first propose candidate patches, then to localize discriminative parts by selecting relevant patches of a specific object. Zhang *et al.* [19] proposed to select many useful parts from multi-scale part proposals of objects in each image without learning expensive object/part detectors. A weakly supervised strategy that using Multi-Instance Learning (MIL) to learn discriminative patterns for image representation is proposed in [20]. Zhang *et al.* [21] develop a pose-normalized representation for image patches to provide a high degree of pose invariance.

Besides these part-based methods, many other methods have been developed. For example, Wang *et al.* [22] adopted a deep ranking model to learn similarity metric directly from images. A multi-stage distance metric learning framework is proposed in [23] to learn a more discriminative similarity. Xie *et al.* [24] proposed a novel data augmentation approach to address the scarcity of data and a novel regularization technique to improve the generalization performance. Deng *et al.* [17] and Wilber *et al.* [25] proposed to include humans in loop to select discriminative features.

There are also some methods that leverage the multi-modal information for fine-grained classification. Inspired by the fact that concepts in images are usually related to each other, Cui *et al.* [26] proposed an end-to-end framework to embed a general relational knowledge for image representation learning. To fuse the multi-modal information, the general relational knowledge is obtained from extra annotations in [26], while both textual and visual information are directly extracted from the image in our method. Besides, we also propose an attention mechanism to automatically select relevant words for multi-modal fusion. In [27], multimodal information extracted from various wearable devices

are fused together for fine-grained activity recognition. The feature extraction and classification are separated. In addition, this multi-modal information is obtained from different sources, while both the visual cue and textual cue derive from only the image in the proposed method.

The most related work with our proposal in this paper is the one in [8] and [28] where scene texts in the business place images are considered as the domain knowledge. The authors proposed to combine visual and textual cues for fine-grained business place classification. More precisely, they train two one-versus-rest SVM classifiers on visual and respectively text features. Then these two trained SVM kernels are added together to form the final SVM kernel matrix. In this way, the visual and text cues are combined together. In [28], Bag-of-Words are used to represent visual cues. Deep visual cues given by GoogleNet [10] are used in [8]. Sharing the same spirit as part-based methods, Karaoglu *et al.* [8] proposed to extract 100 EdgeBoxes [29]. Then the final visual feature is given by the combination of the GoogleNet visual features for the 100 boxed regions and the GoogleNet visual features for the whole image. In our work, we also propose to integrate both visual and text cues for fine-grained classification. We rely on an attention-based model to select the relevant words and get rid of those irrelevant ones. Hopefully, the combination of the attended text cues and visual feature of the whole image are more discriminative. Besides, as shown in Figure 2, our proposed model is end-to-end trainable, which is more elegant and efficient.

B. SCENE TEXT DETECTION AND RECOGNITION
Scene text detection and recognition have been extensively studied in computer vision and document analysis communities [4], [5], [30]–[55] (see [6], [7] for comprehensive surveys). Since the proposed method mainly leverages the

recognized words to improve fine-grained classification, this section briefly revisits the most relevant works about end-to-end word recognition. In the detection stage, connected components [38], [40], [41] or sliding windows [39], [42] are usually used to extract character candidates. Then, the character candidate regions are further verified and grouped into word candidates based on some geometric rules or learning models. In the recognition stage, character-based methods [39], [40], [56]–[58] and word-based methods [11], [45], [59] are the two mainstream recognition approaches. Recently, following the popular general object detection approaches such as RCNN [60], Faster RCNN [61], YOLO [62], and SSD [63], the scheme of end-to-end word proposal generation followed by proposal classification has been a trend for word recognition in natural images [9], [45], [46], [52], [54]. In this paper, we adopt a recent algorithm for end-to-end word recognition named TextBoxes [9], which is the fastest among the existing text reading systems.

C. MULTI-MODAL FUSION

In many tasks, information about the same phenomenon can be easily obtained from various sources, such as video, audio, images, and text. Generally, information from different modalities is complementary and offers useful knowledge to each other [1]–[3]. Multi-modal approaches have shown superior performances over single-modal ones for many tasks [64]. For example, Karaoglu *et al.* [8], [28] proposed to combine visual and textual cues to improve fine-grained classification of business place images. Similarly, Lu *et al.* [65] discussed some methods that combine visual cues and video captions to improve video classification. Rusiol *et al.* [66] presented a page classification application in a banking workflow by merging visual and textual descriptions.

The main challenge of multi-modal approaches lies on how to optimally combine the multi-modal information. The combination can take place at different levels. The most widely used strategy is to fuse the information at the feature level, which is known as early fusion. This kind of system merges the extracted features using integration mechanisms, such as concatenation [67], neural network architectures [68], etc. Another combination strategy is at decision level or late fusion [69], which merges multiple modalities in the semantic space. A hybrid combination of the early and late fusion is also proposed [70].

Following [8] and [28], we also propose to combine visual and text cues to improve fine-grained classification. We combine the visual cues and textual cues at the feature level.

D. ATTENTION MECHANISM

Attention mechanism is widely used in natural language processing, such as speech recognition [71], machine translation [72], and dialog management [73] to improve the network's ability to find relevant features from corresponding input parts. Recently, visual attention attracts much attention in computer vision society. Ba *et al.* [74] presented an attention-based model for recognizing multiple objects in images. Gregor *et al.* [75] proposed a novel network which combines the attention mechanism and an auto-encoding framework for iterative image construction. The attention mechanism is applied to video description in [76], [77], and [78] studied the spatial attention for image captioning.

In our case, even though a powerful text reading system TextBoxes [9] is used to spot words in natural images, there are still some errors (see Figure 1 due to large variations in both foreground texts and background objects, and uncontrollable lighting conditions). To reduce the influence of false detections and irrelevant words, we introduce attention mechanism to automatically analyze the correlation between texts and images. Hopefully, the attention mechanism helps to select most related words in images for the fine-grained classification.

III. FINE-GRAINED CLASSIFICATION INTEGRATING TEXT AND VISUAL APPEARANCE

A. PROPOSED PIPELINE

In this paper, we present an end-to-end neural-network-based model for improving the fine-grained classification by combining visual and text features in the image. The proposed network structure is depicted in Figure 2. It consists mainly of three parts: visual and text feature extraction, semantic attention mechanism for noisy text filtering, and multi-modal feature fusion. The first part aims at transforming each input source into a feature representation. The second one is dedicated to reducing noisy texts brought in during text spotting. The last part combines different modality representations into a single representation. The whole framework is end-to-end trainable. In the following, we detail the main components of the proposed method.

B. VISUAL REPRESENTATION

In computer vision tasks, convolutional neural networks have become a popular method to extract image features. In [79], it has been shown that CNN models trained on a large image dataset can learn common features, which can be shared in many different tasks. Sharif Razavian *et al.* [8] have also shown that fine-tunning from a trained model can obtain better performance than training from scratch in some tasks. Based on this property, we use GoogleNet [10] trained on millions of ImageNet [80] images as initial CNN model and fine tune its parameters with other parts of our model in an end-to-end way. Given an image I, we use the output of the last average pooling layer of the fine-tuned GoogLeNet as the representation of visual features f_v, a 1024-dimensional feature vector.

C. TEXT REPRESENTATION

Word2Vec [81] and GloVe [82] are two popular word embedding algorithms to construct vector representations for words. These vector-based models represent words in a continuous vector space where semantically similar words are embedded nearby each other. Pennington *et al.* [82] noted
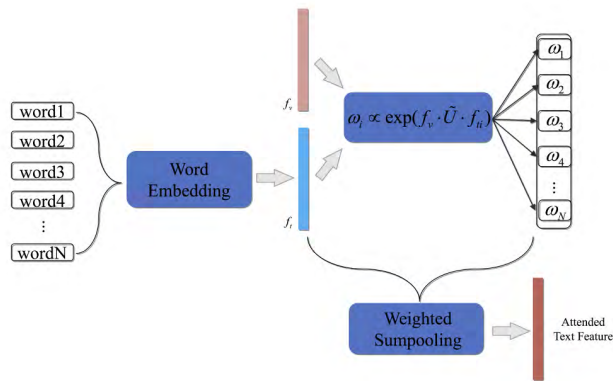
FIGURE 3. The architecture of the proposed semantic attention model. Visual features f_v and text representation f_t are injected into a bilinear function for attention. Different words result in different attention weights, they are fused together by weighted sum-pooling for more discriminative representation.

that GloVe consistently outperforms word2vec for the same corpus, vocabulary, window size and training time. So we adopt GloVe in this paper to transform recognized texts into text features. More specifically, we represent each spotted word T_i by the pre-trained 300-dimensional word vector f_{t_i} of GloVe.

For an image with N spotted words, this results in a text feature f_t whose size is $300 \times N$. To train the network efficiently with batch, we fix the size N with the maximum number of spotted words in each training image denoted as $N_{\max}$. For the training and testing images having less than $N_{\max}$ spotted words, we use zero padding. For the testing images having more than $N_{\max}$ spotted words, we sort the words in decreasing order of their recognition scores and use the first $N_{\max}$ ones to form the text representation. Finally, the text information in each image is represented by a $300 \times N_{\max}$ feature.

D. ATTENTION UNIT
Generally, not all the words in an image must have semantic relations to the scene. Some words may have strong correlations with the image, others may be not relevant at all. Consequently, different words should have different weights representing their correlations with the image. For example, it is reasonable that a word being more related to the image or having a high recognition score should have a large weight, and vice versa. Besides, the text reading system could make some mistakes and result in some noisy texts, which should be discarded. These noisy texts should have a very small weight. Inspired by the work in [78] which proposed a semantic attention mechanism to selectively attend on rich semantic attributes detected from the image, we apply a similar semantic attention mechanism to assign different weights over words in our case. The architecture of the proposed attention model is depicted in Figure 3.

In the attention unit, a bilinear function is used to estimate the weight ω_i for each spotted word T_i in the $N_{\max}$ words

reflecting the image's text feature. This weight ω_i measures the correlation between the spotted word T_i and the image. It is given by:

$$\omega_i \propto exp(f_v^T \cdot \tilde{U} \cdot f_{t_i}), \tag{1}$$

where the exponent is taken to normalize over all words $\mathbb{T} = \{T_i\}$ in a softmax fashion, and $\tilde{U}$ is the 1024×300 bilinear parameter matrix. We then use weighted sum-pooling to fuse all the $N_{\max}$ word features. This fused word feature named attended text representation f_a is given by:

$$f_a = \sum_{i=1}^{N_{\max}} w_i \cdot f_{t_i}. \tag{2}$$

E. MULTI-MODAL FEATURE FUSION
Multi-modal feature fusion has shown superior performance over unimodal approaches in many tasks [83]. A common multi-modal feature fusion method takes two or more modalities to form a preferable feature. In our problem, visual and attended text features provide important and complementary information about the image. The fine-grained classification would benefit from a combination of these two features.

Since the statistical properties are usually not shared across modalities, a transformation of features is required before their fusion. We use a 512-dimensional fully connected layer to map the visual feature into a suitable space $f_{v'}$ having an approximate dimensionality as the attended text feature. Moreover, as shown in many applications, batch-normalization [84] could improve the performance. In this paper, we also adopt this strategy to normalize visual and text features before the multi-modal fusion. Although how to fuse the multi-modal information is not the main focus of this paper, we tried several fusion approaches, like concatenation, linear sum and GMU [85]. The simple concatenation yields the best results. Thus, we adopt the simple concatenation as the final multi-modal feature $f_c = [f_{v'}, f_a]$.

F. SUPERVISION OVER DIFFERENT FEATURES
Multitask Learning (MTL) [86] is an inductive transfer mechanism which leverages the domain-specific information to improve the performance of generalization. It relies on training multi-tasks in parallel using a shared representation, which enables them to help each other to improve their performances. Inspired by MTL, we add two extra classifiers on visual feature f_v and respectively on attended text feature f_a, in parallel with the classifier based on multi-modal feature f_c. Simply put, we train in parallel three classifiers for fine-grained classification: 1) a classifier based on combined visual and attended text feature f_c; 2) a classifier using only visual feature f_v; 3) a classifier relying on attended text feature f_a. Note that for the last classifier trained on f_a only, we add an extra class representing images without recognized text to the set of ground truth categories. All the three classifiers are built in a softmax fashion. The two extra supervisions on f_v and respectively f_a make them more discriminative when used alone, so is their combination. Consequently, this

would improve the performance of the proposed classifier integrating both visual and text appearance.

For an image I having ground truth label Z_i, let W and Φ_i be the parameters of the shared network and respectively the parameters related to the i-th classifier ($i \in \{1, 2, 3\}$). We train the i-th classifier by minimizing the loss function $\mathcal{L}_i(W, \Phi_i)$ given by:

$$\mathcal{L}_i(W, \Phi_i) = -\sum_{k=1}^{K_i} 1(Z_i = k) \log Pr(k \mid I, \mathbb{T}; W, \Phi_i), \quad (3)$$

where K_i is the number of categories for the i-th classifier, and $Pr(k \mid I, \mathbb{T}; W, \Phi_i) \in [0, 1]$ is the predicted score given by the classifier measuring the probability of image I belonging to category k. Note that both K_1 and K_2 denote the number of categories of training set, which is the same for the visual classifier and the fusion classifier. Because the third classifier is trained on attended text feature with an extra class representing images without spotted text, $K_3 = K_1 + 1$.

Let $\emptyset = \{\Phi_i \mid i = 1, 2, 3\}$ denotes the parameters of the three classifiers, then the final loss function is simply given by:

$$\mathcal{L}(W, \emptyset) = \sum_{i=1}^{3} \beta_i \mathcal{L}_i(W, \Phi_i), \quad (4)$$

where β_i is the loss weight for the i-th classifier. We minimize this loss function $\mathcal{L}(W, \emptyset)$ via standard (back-propagation) stochastic gradient descent (SGD):

$$(W, \emptyset)* = \arg\min(\mathcal{L}(W, \emptyset)) \quad (5)$$

IV. EXPERIMENTAL RESULTS
A. EXPERIMENTAL SETUP
1) DATASETS
We evaluate the proposed method on two datasets: *Con-Text dataset* proposed in [28] and our self-collected *Drink Bottle*[1] *dataset*. The Con-Text dataset is dedicated for fine-grained classification of business places, e.g., Cafe, Bookstore, and Pharmacy. It is composed of 24,255 images classified into 28 categories. All the images in this dataset are split into three folds, where two folds are used for training and the other one is used for testing. Each type of experiment on this dataset are repeated three times using respectively the three possible training and testing settings. The average performance is used as the reported results.

To further verify the generality of the proposed method, we have collected a new image dataset named *Drink Bottle dataset* which will be released soon. It consists of various types of drink bottle images contained in *soft drink* and *alcoholic drink* sets in ImageNet dataset [80]. The dataset has 18,488 images divided into 20 categories: *Root Beer, Ginger ale, Coca Cola, Pepsi Cola, Cream Soda, Egg Cream, Birch Beer, Tonic, Sarsaparilla, Orange Soda, Pulque, Kvass,*

Bitter, Guinness, Ouzo, Slivovitz, Drambuie, Vodka, Chablis and Sauterne. The other settings follow Con-Text dataset. As shown in some image samples illustrated in Figure 4, the images in this dataset have low-resolutions and dense words. Moreover, a lot of fonts are in exotic styles. All these aspects make the fine-grained classification on this dataset more challenge than on the Con-Text dataset.

2) EVALUATION MEASURES
The Average Precision (AP) of each category and the mean of AP (mAP) over all categories are used to measure the performance. Following the work in [8], we first sort the classification score in descending order for each category, then the AP is given by:

$$AP = \sum_{k=1}^{n} P(k) * (R(k) - R(k - 1)), \quad (6)$$

where $P(k)$ and $R(k)$ stands for precision and respectively recall at cut-off k in the sorted list, $R(0) = 0$, and n is the number of images in current category.

3) BASELINES
We compare the proposed method with two baselines: a) Visual baseline. We fine tune the pre-trained GoogLeNet with a 28-category softmax classifier; b) Textual baseline. As described in Section III-F, we have to add an extra class representing images without spotted text to train the network using only text information. Although we could evaluate the textual baseline using 29 categories, it is seriously imbalanced between the 28 original categories and the added category. Because most images in this dataset do not have text or the text reading system fails to spot words in images. Consequently, we train and test the textual baseline using 28 categories on a subset of images having spotted words. Since only text information is involved in the textual baseline, instead of using a weighted sum-pooling of word vectors described in Section III-D, we take the classical average of the word vectors given by all spotted words as text feature for this baseline. For a fair comparison, we also train a visual baseline model and a model of the proposed method on this subset.

4) IMPLEMENTATION DETAILS
Our method is built on several pre-trained networks: GoogLeNet[2] [10] to extract image features, TextBoxes[3] [9] to detect texts in images, incremental CNN[4] [87] to recognize the detected words, and GloVe[5] [82] to generate the word embedding. During training, the mini-batch size for SGD is set to 64. The learning rate is set to 0.01 for the first 7K iterations, and then is divided by 10 every

---

[1] https://drive.google.com/open?id=11tPujfvcf2cRYJpaZpSUh8Z_prwci6GY

[2] https://github.com/kmatzen/googlenet-caffe2torch
[3] https://github.com/MhLiao/TextBoxes
[4] http://www.robots.ox.ac.uk/ vgg/research/text/
[5] https://github.com/stanfordnlp/GloVe
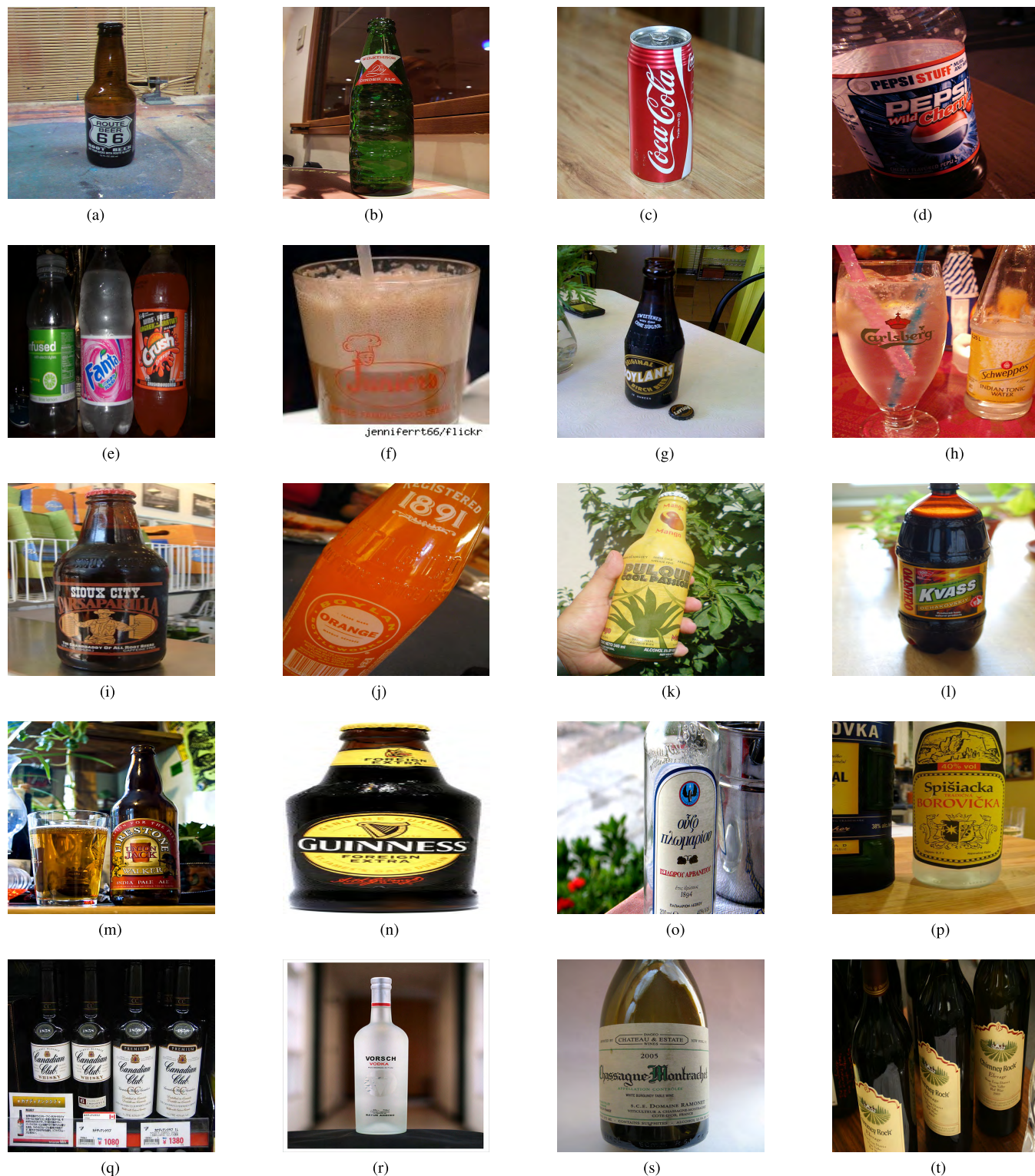
FIGURE 4. Some image samples in the Drink Bottle dataset. They have very similar visual appearances. Distinguishing them is very challenge without textual cues. (a) Root Beer. (b) Ginger ale. (c) Coca Cola. (d) Pepsi Cola. (e) Cream Soda. (f) Egg Cream. (g) Birch Beer. (h) Tonic. (i) Sarsaparilla. (j) Orange Soda. (k) Pulque. (l) Kvass. (m) Bitter. (n) Guinness. (o) Ouzo. (p) Slivovitz. (q) Drambuie. (r) Vodka. (s) Chablis. (t) Sauterne.

10K iterations. The models are trained for up to 30K iterations. We use a weight decay of 0.0001 and a momentum of 0.9.

For the text representation described in Section III-C, we only consider spotted words having more than 3 letters. Because the words having less than 3 letters are more likely to

TABLE 1. mean Average Precision (mAP) (in %) on the two datasets showing the importance of semantic attention mechanism.

	Average-pooling	Weighted sum-pooling
Con-Text	78.6	**79.6**
Drink Bottle	71.5	**72.8**

TABLE 2. mAP (in %) on the two datasets confirming the effect of multi-supervisions. In Single-Supervision, only a classifier trained on the combined feature f_c is used. Multi-Supervisions denotes the use of the three classifiers described in Section III-F respectively on f_c, f_v, and f_a features.

	Single-Supervision	Multi-Supervisions
Con-Text	78.9	**79.6**
Drink Bottle	71.9	**72.8**

be false positives or represent irrelevant information for fine-grained classification. Besides, The value for N_{max} is given by the maximal number of detected words per image in the corresponding dataset. In the following experiments, N_{max} is set to 22 and 21 for Con-Text dataset and Drink Bottle dataset, respectively.

The loss weights in Eq. (4) are set respectively to $\beta_1 = 1.0$, $\beta_2 = 0.5$, and $\beta_3 = 0.5$. In the training stage, we use classical dataset augmentation strategy [10] with different scales and aspect ratios. More precisely, we re-size each training image to 255×255. Then, we randomly crop it with re-size scale in the interval [0.08, 1] and aspect ratio in [3/4, 4/3]. Finally, the cropped patch is re-sized to 224×224. Then randomly horizontal flip is also used with a probability of 0.5. In the testing stage, images are directly re-sized to 224×224 and the standard ten-crop testing [10] is adopted.

The proposed method is implemented using Torch [88] deep learning framework. All the experiments are conducted on an NVIDIA GTX TitanX GPU. The materials to reproduce all the reported results will be released soon.

B. IMPORTANCE OF ATTENTION UNIT AND MULTI-SUPERVISIONS

1) IMPORTANCE OF ATTENTION UNIT

The state-of-the-art text reading system may result in some noisy texts including words that are irrelevant to the scene and incorrectly spotted words. These words may be similar in appearance but semantically far away, such as 'hotpot' and 'hotspot', 'barbera' and 'barber', and etc. We propose to use the attention mechanism to assign appropriate weights to different spotted words. The weighted sum-pooling of word vectors is then used to represent the textual information related to the scene. As depicted in Table 1 on the Con-Text and Drink Bottle dataset, the proposed attention-based method achieves about 1% performance gain over the classical average-pooling scheme.

Some weights given by the attention mechanism over spotted words in several images are given in Figure 5. The proposed attention unit can accurately select the relevant words and eliminate irrelevant ones, even for the failure cases in the third row. The attended weights are reasonable. Their visual features being very close to the incorrectly predicted classes are responsible for their fails. Some failure cases (fourth row) are due to the performance of the text reading system. The relevant words are not detected or incorrectly recognized, such as 'LAUNDROMAT' in (m), 'TEA' in (n), 'ROOT BEER' on the right can in (o), and 'SODA' in (p).

Yet the words having large weights are semantically close to the scene class.

As shown in Table 1 and Figure 5, the attention mechanism can well model the correlation between the visual and textual cues, as well as filter out the irrelevant words and noisy word detections or recognition. Another straightforward method is to take the confidence of text detector and/or text recognizer into account while computing weights in Eq. (1). We also conducted an experiment leveraging the recognition scores and the correlation weights concurrently. Specifically, we update the final weights with these two scores through linear sum. But this variant decreases mAP by 4.2%. This is probably because that there is no relevance between image category and text score (*i.e.*, detection and/or recognition score). Assigning large weights to the irrelevant words having high text scores would harm the classification performance. On the contrary, the words with low text scores may be very relevant to image category, hence deserve large weights. In conclusion, our proposed attention mechanism is more appropriate to discard the noisy text.

2) EFFECT OF MULTI-SUPERVISIONS

Although the main idea in this paper is to use combined text and visual feature f_c to improve fine-grained classification, we rely on multi-supervisions on f_c, f_v, and f_a respectively during the training stage to boost the performance of the proposed classifier based on f_c. Without adjusting carefully the weights over different supervisions in Eq. (4), we simply set the loss weights for combined text and visual classifier, visual classifier, and text classifier as described in Section IV-A.4. As depicted in Table 2 on the Con-Text and Drink Bottle dataset, the multi-supervisions respectively provide about a 0.7% and 0.9% performance gain compared to single-supervision. Note that the two extra supervision classifiers are only used in training stage. During test stage, we simply use the classifier based on combined visual and text feature f_c. Consequently, the 0.7% performance gain thanks to multi-supervisions is achieved without any loss in runtime during the test.

C. COMPARISON WITH BASELINES

We compare the proposed method with several baseline methods described in Section IV-A.3. We have conducted two types of experiments using two different dataset settings: 1) Visual baseline and the proposed method trained on all training images and tested on all testing images; 2) Visual

(a) BARBERSHOP BARBERSHOP
BARBER: 1
SHOP: 7.8e-7
MENUS: 2.8e-8
ROOM: 1.2e-11
BARBS: 3.8e-18

(b) CAFE CAFE
COFFEE: 0.97
ESPRESSO: 0.03
CAPPUCCINO: 2.0e-10
ITALIAN: 2.2e-12

(c) BAKERY BAKERY
CAKES: 0.57
PASTRIES: 0.43
OPEN: 5.5e-9
EGGO: 1.1e-10
DANISH: 3.1e-11

(d) CAFE CAFE
STARBUCKS: 1
SCOFF: 1.1e-8

(e) ROOTBEER ROOTBEER
ROOT: 0.89
BEER: 0.11
BREWED: 1.3e-6
PURE: 2.4e-7
MICRO: 1.1e-9
MADE: 3.8e-10
NATURAL: 2.7e-11
RICH: 1.8e-11

(f) CHABLIS CHABLIS
CHABLIS: 0.99
FRANCE: 8.7e-12
FRANC: 1.1e-12
YIN: 2.4e-16
CON: 2.3e-18
CONTROL: 1.9e-18
BOUTIQUE: 2.5e-19
AFFILIATION: 6.2e-20

(g) BITTER BITTER
BITTER: 0.99
BROWN: 4.05e-5
PREMIUM: 3.5e-9
SPECIAL: 2.8e-9
ENGLISH: 9.4e-11
EXTRA: 6.11e-11

(h) GUINNESS GUINNESS
GUINNESS: 1
SPECIAL: 1.6e-25
EXPORT: 6.4e-27
QUINES: 1.3e-30

(i) PETSHOP COUNTRYSTORE
PET: 0.56
STORE: 0.44
THE: 7.4e-13
VILLAGE: 5.5e-14

(j) FUNERAL MOTEL
CHAPEL: 0.50
MEMORIAL: 0.50
WOODEN: 1.5e-09

(k) CREAMSODA BIRCHBEER
SWEETENED: 0.93
CREME: 0.07
OUNCES: 2.6e-5

(l) GUINNESS BITTER
GUINNESS: 0.50
BITTER: 0.50

(m) DRYCLEANER TOBACCOSHOP
CIGARETTES: 0.91
DRINKS: 0.07
COLD: 0.02
DIAL: 9.7e-05
ACACIA: 1.5e-05
PLAY: 4.3e-08

(n) TEAHOUSE CAFE
CAFE: 1
ROOMS: 2.9e-06
LEA: 5.4e-09

(o) ROOTBEER GINGERALE
GINGER: 0.76
ALE: 0.24
CAFFEINE: 1.03e-14

(p) CREAMSODA ROOTBEER
ROOTBEER: 0.97
CAFFEINE: 0.02
COM: 0.01
FACTS: 3.0e-4
NUTRITION: 3.9e-8

FIGURE 5. Some classification results on the Con-Text dataset and Drink Bottle dataset (better visualization in electronic version). The ground truth and predicted class (in blue and red for correct and respectively incorrect predictions) are given in the first line of each image. The spotted words and their corresponding weights are shown in the rest lines.
(a) BARBERSHOP BARBERSHOP BARBER: 1 SHOP: 7.8e-7 MENUS: 2.8e-8 ROOM: 1.2e-11 BARBS: 3.8e-18. (b) CAFE CAFE COFFEE: 0.97 ESPRESSO: 0.03 CAPPUCCINO: 2.0e-10 ITALIAN: 2.2e-12. (c) BAKERY BAKERY CAKES: 0.57 PASTRIES: 0.43 OPEN: 5.5e-9 EGGO: 1.1e-10 DANISH: 3.1e-11. (d) CAFE CAFE STARBUCKS: 1 SCOFF: 1.1e-8. (e) ROOTBEER ROOTBEER ROOT: 0.89 BEER: 0.11 BREWED: 1.3e-6 PURE: 2.4e-7 MICRO: 1.1e-9 MADE: 3.8e-10 NATURAL: 2.7e-11 RICH: 1.8e-11. (f) CHABLIS CHABLIS CHABLIS: 0.99 FRANCE: 8.7e-12 FRANC: 1.1e-12 YIN: 2.4e-16 CON: 2.3e-18 CONTROL: 1.9e-18 BOUTIQUE: 2.5e-19 AFFILIATION: 6.2e-20. (g) BITTER BITTER BITTER: 0.99 BROWN: 4.05e-5 PREMIUM: 3.5e-9 SPECIAL: 2.8e-9 ENGLISH: 9.4e-11 EXTRA: 6.11e-11. (h) GUINNESS GUINNESS GUINNESS: 1 SPECIAL: 1.6e-25 EXPORT: 6.4e-27 QUINES: 1.3e-30. (i) PETSHOP COUNTRYSTORE PE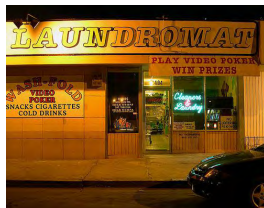T: 0.56 STORE: 0.44 THE: 7.4e-13 VILLAGE: 5.5e-14. (j) FUNERAL MOTEL CHAPEL: 0.50 MEMORIAL: 0.50 WOODEN: 1.5e-09. (k) CREAMSODA BIRCHBEER SWEETENED: 0.93 CREME: 0.07 OUNCES: 2.6e-5. (l) GUINNESS BITTER GUINNESS: 0.50 BITTER: 0.50. (m) DRYCLEANER TOBACCOSHOP CIGARETTES: 0.91 DRINKS: 0.07 COLD: 0.02 DIAL: 9.7e-05 ACACIA: 1.5e-05 PLAY: 4.3e-08. (n) TEAHOUSE CAFE CAFE: 1 ROOMS: 2.9e-06 LEA: 5.4e-09. (o) ROOTBEER GINGERALE GINGER: 0.76 ALE: 0.24 CAFFEINE: 1.03e-14. (p) CREAMSODA ROOTBEER ROOTBEER: 0.97 CAFFEINE: 0.02 COM: 0.01 FACTS: 3.0e-4 NUTRITION: 3.9e-8.

TABLE 3. Classification performance for baselines, our method, and two state-of-the-art methods on the Con-Text dataset. Models with ∗ are trained and tested on the subset of images containing spotted texts. For clearness, the words in the first row are *massage parlor, pet shop, restaurant, computer store, theatre, repair shop, bookstore, bakery, medical center, barbershop, pizzeria, diner, hotspot, bistro, teahouse, school, pharmacy, funeral, country store, tavern, motel, packinghouse, cafe, tobacco shop, dry cleaner, discount house, steakhouse,* and *pawnshop* respectively.

Method	mas.	pet.	res.	com.	theat.	rep.	b.s.	bak.	med.	bar.	piz.	din.	h.p.	bist.	tea	sch.	pha.	fun.	cou.	tav.	motel	pack.	cafe	toba.	d.c.	disc.	steak	pawn	mAP
Visual Baseline∗	60.4	57.8	30.4	63.4	88.2	46.6	76.9	59.9	61.3	66.7	38.6	74.8	90.4	17.9	30.2	71.0	30.8	51.4	60.1	43.4	80.3	53.1	23.8	43.4	50.6	11.1	21.7	42.7	51.7
Textual Baseline∗	28.8	16.8	12.5	28.6	40.6	12.1	42.9	53.8	23.2	56.5	55.2	60.4	81.8	22.8	25.6	44.3	55.9	69.6	46.0	18.6	81.2	47.6	57.1	46.2	87.1	37.5	61.3	75.9	46.1
ours∗	**74.3**	**72.7**	**46.7**	**79.7**	**93.5**	**60.0**	**90.6**	**84.0**	**81.3**	**92.6**	**84.3**	**88.3**	**90.9**	**32.6**	**54.7**	**85.6**	**78.9**	**90.9**	**83.3**	**63.1**	**98.1**	**85.0**	**72.0**	**77.7**	**96.1**	**51.9**	**82.8**	**91.9**	**78.0**
Visual Baseline	81.4	84.3	72.4	74.3	86.1	72.9	85.5	80.4	74.2	85.9	79.9	75.7	69.0	20.9	55.7	66.6	72.7	70.6	60.3	33.8	72.9	60.3	31.5	45.6	49.0	3.1	21.9	30.6	61.3
ours	**81.8**	**89.5**	**78.6**	**80.6**	**92.4**	**80.1**	**94.2**	**89.6**	**83.6**	**95.8**	**90.4**	**86.7**	**80.3**	**32.4**	**68.9**	**81.3**	**88.4**	**88.4**	**78.5**	**52.2**	**93.3**	**85.0**	**57.0**	**72.3**	**93.3**	**51.7**	**74.3**	**87.0**	**79.6**
Karaoglu et al. [28]	34.9	45.2	51.1	33.8	48.5	17.8	60.0	37.8	48.5	55.2	55.2	43.4	65.7	9.1	12.5	44.3	60.8	43.0	35.2	10.6	53.0	38.2	16.2	29.0	50.9	18.7	28.1	44.5	39.0
Karaoglu et al. [8]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	77.3

TABLE 4. Classification performance for baselines and the proposed method on the Drink Bottle dataset. The category of every logogram in this table is depicted in Section IV-A.1. Similar to Table 3, models with ∗ are trained and tested on the subset of images containing spotted texts.

Method	r.b.	g.a.	coke	pepsi	c.s.	e.c.	b.c.	tonic	sarsap.	o.s.	pulque	kvass	bitter	guin.	ouzo	sliv.	dramb.	vodka	chabl.	saut.	mAP
Visual Baseline∗	64.1	63.2	94.8	**98.1**	27.5	40.4	33.1	59.6	29.5	77.8	58.3	26.4	76.4	97.0	64.8	47.3	88.4	82.2	90.9	87.1	65.4
Textual Baseline∗	54.7	61.9	34.1	76.5	27.0	0.2	22.3	33.6	18.6	54.5	37.4	23.2	40.2	90.7	58.7	9.5	75.6	61.5	64.2	55.7	45.0
ours∗	**80.8**	**74.0**	**96.1**	98.0	**54.4**	**41.1**	**47.4**	**69.1**	**48.9**	**85.4**	**65.6**	**40.2**	**84.7**	**98.5**	**79.8**	**59.5**	**91.7**	**90.8**	**93.5**	**90.9**	**74.5**
Visual Baseline	61.9	54.2	94.1	95.6	32.4	60.7	35.7	53.4	41.1	78.5	51.7	31.6	66.5	93.3	45.3	43.6	75.3	75.9	85.1	84.4	63.1
ours	**76.8**	**70.3**	**95.8**	**96.9**	**52.5**	**73.2**	**42.9**	**65.6**	**57.8**	**87.4**	**61.7**	**39.4**	**77.3**	**96.7**	**63.8**	**54.1**	**79.9**	**85.7**	**90.2**	**88.7**	**72.8**
Karaoglu et al. [8]	71.1	64.3	94.1	96.7	43.8	69.4	37.6	56.1	53.2	83.2	52.0	37.5	70.7	95.0	52.4	48.7	77.6	81.9	86.5	84.9	67.8

baseline, textual baseline, and the proposed method trained on the subset of training images and tested on the subset of testing images having spotted texts in each dataset. We add ∗ after corresponding name to distinguish these experiments from previous ones.

Table 3 displays the quantitative comparisons on the Con-Text dataset. In general, the textual baseline achieves results on par with visual baseline on images having spotted texts in this dataset. This shows that the text feature is as useful as visual feature for fine-grained classification on images containing texts. The proposed method using the combination of text and visual features significantly outperforms the baseline based only on text or visual feature. A large improvement is observed particularly for the images having similar visual layout but discriminative textual cues, e.g., *Dry Cleaner*, *Discount House*, *Steak House*, and *Pawn Shop*. For the images having fewer texts or no texts, the improvement is less significant. The observations on the Con-Text dataset also hold for the Drink Bottle dataset. As depicted in Table 4, the proposed method also significantly outperforms the visual and textual baselines. Yet, the improvement over visual baseline (about 10%) is less important as compared to the improvement (about 20%) on the Con-Text dataset. The main reason is likely to be that recognition of the exotic texts in the low-resolution drink bottle images is more challenge. The text reading system would yield more noisy texts. The fact that less accurate text recognition gives less significant improvements also demonstrates the importance and interestingness of integrating text appearances into visual appearances for fine-grained classification.

To further analyze the benefit of integrating text information into visual appearance for fine-grained classification, we compare the performance of the proposed method and visual baseline on the subset of testing images having spotted texts and subset without texts separately in Con-Text and Drink Bottle dataset. In this two datasets, respectively about 42% and 38% images contain texts. The quantitative comparison is given in Table 5. The proposed method improves dramatically the performance on images having texts as observed in the above experiments. Surprisingly, It also boosts slightly the performance for images without texts. One reason for this improvement may lie on the multi-supervisions of the network using shared parameters.

D. COMPARISON WITH OTHER METHODS
We compare the proposed method with the state-of-the-art methods [8], [28] on the Con-Text dataset. As shown in Table 3, the proposed method improves the state-of-the-art

TABLE 5. mAP (in %) on the subset of images having spotted words and the rest images in the Con-Text and Drink Bottle dataset. Both models are trained on all training images.

	Con-Text		Drink Bottle	
	image W texts	image W/O texts	image W texts	image W/O texts
Visual Baseline	52.8	65.8	65.9	57.9
ours	**79.0**	**71.7**	**78.2**	**62.1**

TABLE 6. Evaluation of the proposed method for product retrieval on Drink Bottles dataset. Models with * are trained and tested on the subset of images containing texts (see also Section IV-C).

Method	mAP(%)
Visual Baseline*	60.0
Textual Baseline*	28.9
ours*	**72.6**
Visual Baseline	48.0
ours	**60.8**

results [8] from 77.3% to 79.6% mAP on Con-Text. Besides, the method in [8] leverages extra local object information given by 100 EdgeBoxes [29]. For a fair comparison, the proposed method should be compared with the $Textual(full) + Visual(DEEP-FT)$ variant of [8] which achieves 74.5±0.8% mAP. In this case, the proposed method is significantly superior by 5.1% mAP. Moreover, we also compare with the approach in [8] using their released code[6] on the Drink Bottle dataset. As depicted in Table 4, the proposed method also significantly improves the state-of-the-art results by 5% mAP on this dataset, which further confirms the superiority of the proposed approach. In conclusion, the proposed method only relies on global visual information without the additional local part-based visual features. Moreover, the proposed method is end-to-end trainable, which is more elegant and efficient.

Some fine-grained classification results of the proposed method are given in Figure 5. The proposed attention mechanism accurately selects the most relevant words related the scene. From (a) to (h), the proposed method makes correct predictions regardless the very close visual appearances of different fine-grained classes. Some failure cases are also shown in the third and fourth row. Though relevant spotted words are highlighted by the proposed attention mechanism in the third row, the images are incorrectly classified due to their visual features. In the last row, the most related words are not spotted by the text reading system. The proposed method fails even if the semantically close words are highlighted by the attention mechanism. A better performance could be expected with a more robust text reading system in the future.

E. APPLICATION TO PRODUCT RETRIEVAL

We present an application of the proposed framework for *Product Retrieval* to further demonstrate its potential. Compared to image classification, image retrieval attracts more attention in e-commerce. Given a query image I, the objective of product retrieval is to select all product images from an image set within the same category as I. We have conducted this experiment on the Drink Bottle dataset. We simply extract the input vector of the softmax-based classifier in fine-grained classification model. This input vector is considered as product feature for retrieval. The cosine similarity is used to rank images during matching. Following FlickrLogos-32's

[6]https://github.com/taotaoorange/words-matter-scene-text-for-image-classification

evaluation settings [89], the test set and train set are used as queries and respectively search set. We evaluate this application based on mAP. As shown in Table 6, the proposed method boosts consistently and significantly the retrieval performance.

V. CONCLUSION

In this paper, we have introduced a unified, end-to-end trainable framework which integrates visual and text cues for fine-grained image classification. We have also incorporated the attention mechanism to better exploit the extracted text in images by only selecting the most relevant words. Using several state of the art components in an effective pipeline, our experimental results on two datasets have demonstrated both the added value and improved performance of the proposed method. Furthermore, the product retrieval experiment shows that the learned representation provides a potentially powerful tool for product search. It is worth noting that while it certainly makes sense to employ the state of the art in scene text and visual recognition as the components, our main motivation and contribution are in filling a gap in fine-grained image classification by exploiting multi-modal complementary information when present, and doing so with an end-to-end trainable neural network with the attention mechanism for implicitly exploiting the semantic relationships between textual and visual cues. The main drawback of using textual cues in the proposed framework for fine-grained image classification is clearly the absence of texts in some images. Besides the scene texts in images, other additional textual cues could be user tags, images attributes, and so on. In the future, we would like to incorporate other textual cues to compensate for the problem of no texts in images.

REFERENCES

[1] X. Yang, T. Zhang, and C. Xu, "Cross-domain feature learning in multimedia," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 64–78, Jan. 2015.

[2] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu, "Image classification by cross-media active learning with privileged information," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2494–2502, Dec. 2016.

[3] B. Wang, D. Lin, H. Xiong, and Y. F. Zheng, "Joint inference of objects and scenes with efficient learning of text-object-scene relations," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 507–520, Mar. 2016.

[4] K. L. Bouman, G. Abdollahian, M. Boutin, and E. J. Delp, "A low complexity sign detection and text localization method for mobile applications," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 922–934, Oct. 2011.

[5] L. Wu, P. Shivakumara, T. Lu, and C. L. Tan, "A new technique for multi-oriented scene text line detection and tracking in video," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1137–1152, Aug. 2015.

[6] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1480–1500, Jul. 2015.

[7] Y. Zhu, C. Yao, and X. Bai, "Scene text detection and recognition: Recent advances and future trends," *Frontier Comput. Sci.*, vol. 10, no. 1, pp. 19–36, 2016.

[8] S. Karaoglu, R. Tao, T. Gevers, and A. W. M. Smeulders, "Words matter: Scene text for image classification and retrieval," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1063–1076, May 2017.

[9] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: A fast text detector with a single deep neural network," in *Proc. AAAI*, 2017.

[10] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[11] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, Nov. 2017.

[12] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNS for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 834–849.

[13] A. Angelova and S. Zhu, "Efficient object detection and segmentation for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 811–818.

[14] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 554–561.

[15] S. Bai, X. Bai, and W. Liu, "Multiple stage residual model for image classification and vector compression," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1351–1362, Jul. 2016.

[16] S. Yang, L. Bo, J. Wang, and L. G. Shapiro, "Unsupervised template learning for fine-grained object recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 3122–3130.

[17] J. Deng, J. Krause, and L. Fei-Fei, "Fine-grained crowdsourcing for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 580–587.

[18] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 842–850.

[19] Y. Zhang *et al.*, "Weakly supervised fine-grained categorization with part-based image representation," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1713–1725, Apr. 2016.

[20] P. Tang, X. Wang, B. Feng, and W. Liu, "Learning multi-instance deep discriminative patterns for image classification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3385–3396, Jul. 2017.

[21] N. Zhang, R. Farrell, and T. Darrell, "Pose pooling kernels for sub-category recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 3665–3672.

[22] J. Wang *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1386–1393.

[23] Q. Qian, R. Jin, S. Zhu, and Y. Lin, "Fine-grained visual categorization via multi-stage metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3716–3724.

[24] S. Xie, T. Yang, X. Wang, and Y. Lin, "Hyper-class augmented and regularized deep learning for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2645–2654.

[25] M. J. Wilber, I. S. Kwak, D. Kriegman, and S. Belongie, "Learning concept embeddings with combined human-machine expertise," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 981–989.

[26] P. Cui, S. Liu, and W. Zhu, "General knowledge embedded image representation learning," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 198–207, Jan. 2018, doi: 10.1109/TMM.2017.2724843.

[27] D. De, P. Bharti, S. K. Das, and S. Chellappan, "Multimodal wearable sensing for fine-grained activity recognition in healthcare," *IEEE Internet Comput.*, vol. 19, no. 5, pp. 26–35, Sep./Oct. 2015.

[28] S. Karaoglu, J. C. van Gemert, and T. Gevers, "Con-text: text detection using background connectivity for fine-grained object classification," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 757–760.

[29] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 391–405.

[30] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun./Jul. 2004, pp. 366–373.

[31] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2963–2970.

[32] K. Wang and S. Belongie, "Word spotting in the wild," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2010, pp. 591–604.

[33] Y.-F. Pan, X. Hou, and C.-L. Liu, "A hybrid approach to detect and localize texts in natural scene images," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 800–813, Mar. 2011.

[34] M. Zhao, S. Li, and J. Kwok, "Text detection in images using sparse representation with discriminative dictionaries," *Image Vis. Comput.*, vol. 28, no. 12, pp. 1590–1599, 2010.

[35] C. Yi and Y. Tian, "Text string detection from natural scenes by structure-based partition and grouping," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2594–2605, Sep. 2011.

[36] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 970–983, May 2014.

[37] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1083–1090.

[38] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Proc. Asian Conf. Comput. Vis.* Springer, 2010, pp. 770–783.

[39] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 1457–1464.

[40] C. Yao, X. Bai, and W. Liu, "A unified framework for multioriented text detection and recognition," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4737–4749, Nov. 2014.

[41] L. Neumann and J. Matas, "Scene text localization and recognition with oriented stroke detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 97–104.

[42] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 3304–3308.

[43] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven, "PhotoOCR: Reading text in uncontrolled conditions," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 785–792.

[44] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 512–528.

[45] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *Int. J. Comput. Vis.*, vol. 116, no. 1, pp. 1–20, 2016.

[46] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2315–2324.

[47] W. Huang, Y. Qiao, and X. Tang, "Robust scene text detection with convolution neural network induced MSER trees," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 497–511.

[48] W. Huang, Z. Lin, J. Yang, and J. Wang, "Text localization in natural images using stroke feature transform and text covariance descriptors," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1241–1248.

[49] S. Lu, T. Chen, S. Tian, J.-H. Lim, and C.-L. Tan, "Scene text extraction based on edges and support vector regression," *Int. J. Document Anal. Recognit.*, vol. 18, no. 2, pp. 125–135, 2015.

[50] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 3538–3545.

[51] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. L. Tan, "Text flow: A unified text detection system in natural scene images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4651–4659.

[52] Z. Zhong, L. Jin, S. Zhang, and Z. Feng. (2016). "DeepText: A unified framework for text proposal generation and text detection in natural images." [Online]. Available: https://arxiv.org/abs/1605.07314

[53] D. He, X. Yang, W. Huang, Z. Zhou, D. Kifer, and C. L. Giles, "Aggregating local context for accurate scene text detection," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 280–296.

[54] L. Gomez and D. Karatzas, "TextProposals: A text-specific selective search algorithm for word spotting in the wild," *Pattern Recognit.*, vol. 70, pp. 60–74, Oct. 2017.

[55] S. Calarasanu, J. Fabrizio, and S. Dubuisson, "What is a good evaluation protocol for text localization systems? Concerns, arguments, comparisons and solutions," *Image Vis. Comput.*, vol. 46, pp. 1–17, Feb. 2016.

[56] A. Mishra, K. Alahari, and C. V. Jawahar, "Scene text recognition using higher order language priors," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 127.1–127.11.

[57] A. Mishra, K. Alahari, and C. V. Jawahar, "Top-down and bottom-up cues for scene text recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2687–2694.

[58] X. Bai, C. Yao, and W. Liu, "Strokelets: A learned multi-scale mid-level representation for scene text recognition," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2789–2802, Jun. 2016.

[59] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2552–2566, Dec. 2014.

[60] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[61] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[62] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

[63] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 21–37.

[64] C. A. Bhatt and M. S. Kankanhalli, "Multimedia data mining: State of the art and challenges," in *Multimedia Tools and Applications*. Springer, 2011, pp. 35–76.

[65] T. Lu, S. Palaiahnakote, C. L. Tan, and W. Liu, "Text detection in multimodal video analysis," in *Video Text Detection*. Springer, 2014, pp. 221–246.

[66] M. Rusiñol, V. Frinken, D. Karatzas, A. D. Bagdanov, and J. Lladós, "Multimodal page classification in administrative document image streams," *Int. J. Document Anal. Recognit.*, vol. 17, no. 4, pp. 331–341, 2014.

[67] D. Kiela and L. Bottou, "Learning image embeddings using convolutional neural networks for improved multi-modal semantics," in *Proc. EMNLP*, 2014, pp. 36–45.

[68] G. Andrew, R. Arora, J. A. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. ICML*, vol. 3, 2013, pp. 1247–1255.

[69] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 399–402.

[70] A. D'Ulizia, F. Ferri, and P. Grifoni, "A hybrid grammar-based approach to multimodal languages specification," in *Proc. OTM Confederated Int. Conf. Move Meaningful Internet Syst.* Springer, 2007, pp. 367–376.

[71] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4945–4949.

[72] D. Bahdanau, K. Cho, and Y. Bengio. (2014). "Neural machine translation by jointly learning to align and translate." [Online]. Available: https://arxiv.org/abs/1409.0473

[73] T. Hori *et al.*, "Dialog state tracking with attention-based sequence-to-sequence learning," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2016, pp. 552–558.

[74] J. Ba, V. Mnih, and K. Kavukcuoglu. (2014). "Multiple object recognition with visual attention." [Online]. Available: https://arxiv.org/abs/1412.7755

[75] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. (2015). "DRAW: A recurrent neural network for image generation." [Online]. Available: https://arxiv.org/abs/1502.04623

[76] C. Hori *et al.*, "Attention-based multimodal fusion for video description," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4203–4212.

[77] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, vol. 14, 2015, pp. 77–81.

[78] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4651–4659.

[79] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 512–519.

[80] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[81] T. Mikolov, K. Chen, G. Corrado, and J. Dean. (2013). "Efficient estimation of word representations in vector space." [Online]. Available: https://arxiv.org/abs/1301.3781

[82] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[83] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, nos. 3–4, pp. 197–387, Jun. 2014.

[84] S. Ioffe and C. Szegedy. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167

[85] J. Arevalo, T. Solorio, M. Montes-y-Gómez, and F. A. González. (2017). "Gated multimodal units for information fusion." [Online]. Available: https://arxiv.org/abs/1702.01992

[86] R. Caruana, "Multitask learning," in *Learning to Learn*. Springer, 1998, pp. 95–133.

[87] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. (2014). "Synthetic data and artificial neural networks for natural scene text recognition." [Online]. Available: https://arxiv.org/abs/1406.2227

[88] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *Proc. BigLearn NIPS Workshop*, 2011.

[89] S. Romberg, L. G. Pueyo, R. Lienhart, and R. van Zwol, "Scalable logo recognition in real-world images," in *Proc. 1st ACM Int. Conf. Multimedia Retr.*, 2011, Art. no. 25.

XIANG BAI received the B.S., M.S., and Ph.D. degrees in electronics and information engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2003, 2005, and 2009, respectively. He is currently a Professor with the School of Electronics Information and Communications, HUST, where he is also the Vice Director of the National Center of Anti-Counterfeiting Technology. His research interests include object recognition, shape analysis, scene text recognition, and intelligent systems. He serves as an Associate Editor for *Pattern Recognition*, *Pattern Recognition Letters*, *Neurocomputing*, and *Frontiers of Computer Science*.

MINGKUN YANG received the B.S. degree from the School of Electronics Information and Communications, Huazhong University of Science and Technology, China, in 2016, where he is currently pursuing the master's degree. His main research interests include fine-grained image classification and scene text recognition.

PENGYUAN LYU received the B.S. degree from the School of Electronics Information and Communications, Huazhong University of Science and Technology, China, in 2015, where he is currently pursuing the master's degree. His main research interests include scene text detection and recognition.
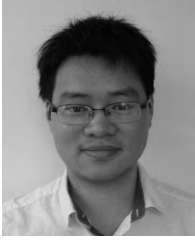
YONGCHAO XU received the B.Eng. degree in electronics and embedded systems from the École polytechnique universitaire de l'université Paris-XI, Orsay, France, the master's degree in signal processing and image processing from Université Paris-Sud, Orsay, in 2010, and the Ph.D. degree in image processing and computer vision from Universite Paris-Est, France, in 2013. He was an Assistant Professor with the EPITA Research and Development Laboratory and a Visiting Professor at LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay. He is currently with the Huazhong University of Science and Technology, Wuhan, China. His research interests include mathematical morphology, image segmentation, medical image analysis, and scene text reading.

JIEBO LUO (S'93–M'96–SM'99–F'09) joined the Department of Computer Science, University of Rochester, in 2011, after a prolific career of over 15 years with Kodak Research. He has authored over 300 technical papers. He holds over 90 U.S. patents. His research interests include computer vision, machine learning, data mining, social media, and biomedical informatics. He has served as the Program Chair for the ACM Multimedia 2010, the IEEE CVPR 2012, ACM ICMR 2016, and the IEEE ICIP 2017. He is on the Editorial Boards of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Pattern Recognition*, the *Journal of Machine Vision and Applications*, and *ACM Transactions on Intelligent Systems and Technology*. He is also a fellow of the SPIE and IAPR.

○ ○ ○