IQ-STAN: IMAGE QUALITY GUIDED SPATIO-TEMPORAL ATTENTION NETWORK FOR LICENSE PLATE RECOGNITION

Cong Zhang, Qi Wang, Xuelong Li*

School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P.R. China.

ABSTRACT

License plate recognition (LPR) is one of the essential components in intelligent transportation systems. Although the image processing algorithms for LPR have been extensively studied in the past several years, the recognition performance is still not satisfactory especially in unconstrained complex scenes. In order to tackle this issue, a novel deep multi-task learning-based method is proposed in this paper by introducing contextual information in multiple license plate frames. Specifically, an end-to-end trainable multi-task architecture, namely IQ-STAN, is developed by joint license plate recognition and image quality scoring. Moreover, we propose an image quality-guided spatio-temporal attention mechanism, which is utilized in the frame-level feature representation during the phase of plate recognition. Extensive experiments are conducted and the competitive results demonstrate the effectiveness of our proposed framework.

Index Terms— Spatio-Temporal Attention, License Plate Recognition,Image Quality, Multi-Task, Multi-Frame

1. INTRODUCTION

License plate recognition (LPR) plays an important role in intelligent transportation systems. The technology LPR has attracted considerable attention from the related academic community due to its wide range applications such as parking lot access control and road traffic law enforcement [1, 2, 3, 4, 5].

The main task of the LPR system is to read each character automatically in license plate images captured by surveillance cameras [6]. Over the past decades, the technology LPR has been studied widely. As mentioned in [1], a traditional LPR system mainly contains three stages: firstly localize license plates in the full image based on hand-crafted features [7]; secondly segment the detected plate into individual character blocks; finally recognize the segmented characters one by one with a pre-designed classifier such as support vector machine (SVM) and artificial neural networks (ANN) [8]. Thanks to the rapid development of deep learning, many object detectors like Faster R-CNN [9] and YOLO [10] have been demonstrated to meet requirements for the LPR systems [11, 12]. However, LPR especially in complex environments is still a difficult problem waiting to be solved. Moreover, plate segmentation is also a challenging task [13]. Recently some segmentation-free methods have been proposed and achieved better performance [2, 13], which attracts considerable interests. In [2], Li *et al.* regard the task LPR as a sequence labeling problem and adopt recurrent neural networks (RNNs) followed by connectionist temporal classification (CTC) to recognize the plate sequences without segmentation.

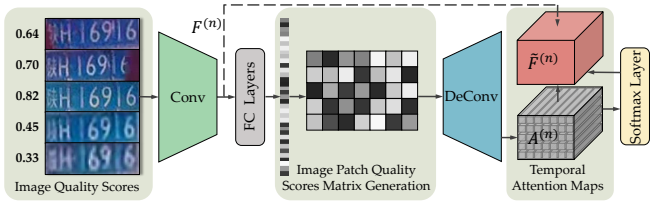

Fig. 1. There are five frames involved with the same license plate but captured at different times. Two transpose convolutional layers (often named "Deconv") are exploited in the process of generating temporal attention feature maps $\widetilde{F}^{(n)}$.

These previous LPR works have made improvements in various scenarios. However, there are some important issues to be addressed. (1) Most algorithms are proposed for the LPR task in strictly limited conditions. With the auxiliary equipment, character recognition from well-aligned plate images has achieved high accuracy. Nevertheless, in real-life applications, the captured images usually suffer from distortion such as motion blur, causing some algorithms not working. (2) Many existing methods exploit the super-resolution technology to recover the high-resolution plate images [6, 14, 15, 16], which may improve the performance of LPR systems. However, such a super-resolution operation will increase the computational complexity and then impact the real-time capabilities in LPR systems. Furthermore, the acquisition of training sets for super-resolution is also laborious and extremely time-consuming. Intuitively, the quality of plates directly affects recognition results which can be considered as prior information. It is not difficult to obtain the image quality in ad-

*Qi Wang is the corresponding author. This work was supported by the National Natural Science Foundation of China under Grant U1864204, 61773316, U1801262, 61871470 and 61761130079.

vance yet such prior information may be usually ignored. (3) Most existing LPR algorithms focus on single-frame recognition without contextual information involved in consecutive frames. However, plate frames captured at different times may contain valuable and complementary information. As shown in Fig. 1, quality scores vary with different frames. More importantly, their features are complementary but correlative, which affects the prediction of the final results in a different way. Thus multi-frame LPR algorithms can improve recognition accuracy, especially in complex environments.

To remedy these problems above, in this paper, we propose the image quality (IQ) guided spatio-temporal attention network for LPR, namely IQ-STAN. In order to tackle the multi-frame information more efficiently, we insert a novel IQ-guided temporal attention component based on the sharing features. Moreover, image quality scoring and plate recognition are integrated into a unified framework innovatively. The main contributions of this work are summarized as follows:

- A unified and compact multi-task architecture named IQ-STAN is proposed by joint image quality scoring and license plate recognition. Both tasks are based on supervised learning and IQ-STAN can be trained in an end-to-end manner via deep multi-task learning.
- The IQ-guided spatio-temporal attention mechanism is introduced to explore valuable contextual information from multi-frame license plate images. The proposed spatio-temporal attention mechanism improves the system performance and robustness especially in unconstrained and complicated scenarios.

2. IQ-GUIDED SPATIO-TEMPORAL ATTENTION NETWORK

In general, a complete and serviceable LPR system consists of two or three components: license plate detection (required), tracking (optional) and recognition (required) [1]. As mentioned above, some pre-trained deep learning-based detectors like Faster R-CNN [9] can favorably localize the plates from images or videos captured by traffic surveillance cameras. For video sequences, template matching and MSER (maximally stable extremal region) have been demonstrated to work well on license plate tracking [17, 18]. Therefore, this paper mainly focuses on the final step, *i.e.*, license plate recognition.

Inspired by [2], we consider the detected plates as text sequences, and then an attention-based segmentation-free method is adopted to read them. Different from other approaches [2, 13], the proposed IQ-STAN handles the contextual information in multiple frames instead of the individual images. As presented in Fig. 2, it is composed of two components, the shared convolutional neural network (CNN) and IQ-guided spatio-temporal attention modules, which will be introduced respectively in the following sections.

2.1. Shared CNN for Deep Multi-task Learning

As depicted in Fig. 2, the proposed multi-task architecture IQ-STAN involves two tasks, image quality scoring and license plate recognition. The former task aims to estimate the objective quality of input images and further generate temporal attention maps. The latter focuses on reading characters from the cropped license plates with the prior information of quality scores, which are assigned by scoring networks. For better performance, license plate recognition usually operates on discriminative features of the input images, as is image quality scoring. Consequently, we develop the CNN module for both tasks simultaneously, in which the parameters are shared via deep multi-task learning. Time consumption and computational resources are greatly reduced in this way.

The input images are fixed to the size of 840×50 and presented to the CNN in RGB color space, which can be denoted as $I \in R^{840\times50\times3}$. As summarized in Fig. 2, the CNN architecture used in this paper is relatively shallow for the sake of fast training and inference. It consists of six fully convolutional layers interleaved by three max pooling layers which are utilized to reduce the size of feature maps to 210×7. The extracted features are denoted as $F^{(n)} = CNN(I_n)$, where I_n represents the n-th input image. Batch normalization and dropout are employed as a means of regularization.

2.2. IQ-guided Spatio-temporal Attention

In order to recognize the license plate and evaluate its image quality simultaneously, we design two parallel branches following the shared CNN for these two subtasks, as illustrated in Fig. 2. More importantly, the IQ scores attained from multi-frame discriminative features are further collected to generate temporal attention maps. As a guide, the attention map works on the encoder in the recognition module using element-wise operations. Inspired by recent text recognition methods [19], our LPR algorithm relies on the attention-based encoder-decoder model where the attention mechanism is related to the visual spatial position of the characters. Therefore, the novel spatio-temporal attention mechanism is developed by joint IQ-guided temporal attention and spatial attention-based decoder in the deep multi-task framework.

As shown in Fig. 2, the image quality scoring module consists of two different pooling layers followed by two fully connected layers. In practice, with the deep feature maps $F^{(n)}$, we divide it into 35 patches (an input image contains 5 frames and 7 characters per-frame) in the vertical direction and evaluate their quality scores separately. Thus $F^{(n)} = (F_1^{(n)}, F_2^{(n)}, \cdots, F_M^{(n)})$ where $F_m^{(n)}$ represents the m-th feature patch and $M = 35$ in our dataset. Formally denote the predicted quality scores as $\hat{s}^{(n)} = \{\hat{s}_1^{(n)}, \hat{s}_2^{(n)}, \cdots, \hat{s}_M^{(n)}\}$, and the temporal attention maps can be further computed by:

$$\widetilde{F}^{(n)} = A^{(n)} \odot F^{(n)} = Attention(F^{(n)}, \hat{s}^{(n)}), \quad (1)$$

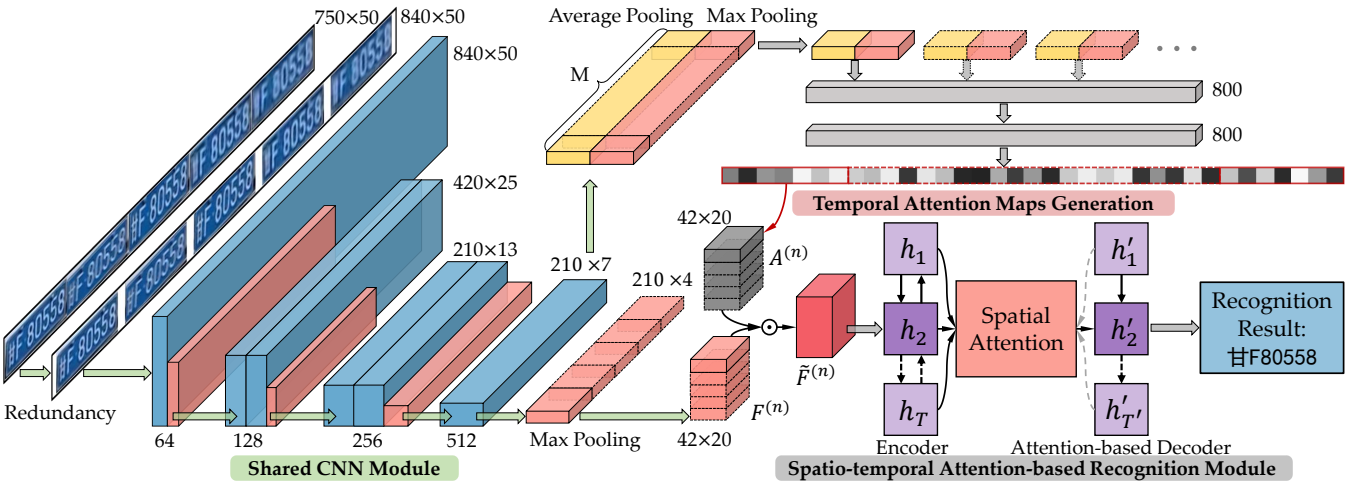

Fig. 2. Overview of the proposed multi-task architecture IQ-STAN which acts on 5 different frames of the same license plate. Each input image is extended to the resolution of 840×50 from 750×50 by providing redundancy between multiple frames.

where $A^{(n)}$ represents the generated attention map and $\odot$ means element-wise multiplication. The process of temporal attention maps generation is illustrated in Fig. 1. It attempts to exploit temporal contextual information in multi-frame images on the feature level by principled learning.

For plate recognition, bidirectional long short-term memories (LSTMs) are exploited in the encoder-decoder model to deal with both future and past contextual information. With the initial state h_0, the encoder hidden state at time t is updated via:

$$\begin{aligned} h_t^{(1)} &= LSTM(h_{t+1}^{(1)}, \widetilde{F}_t^{(n)}; \theta), \\ h_t^{(2)} &= LSTM(h_{t-1}^{(2)}, \widetilde{F}_t^{(n)}; \theta), \end{aligned} \tag{2}$$

where the two equations are related to the backward cell and the forward one, respectively. In the decoder, another LSTM is employed to predict the target sequence $\hat{y} = (\hat{y}_1, \cdots, \hat{y}_T)$ based on visual attention mechanism. Formally denote $(\tilde{f}_1^{(n)}, \cdots, \tilde{f}_T^{(n)})$ as sequential feature vectors generated from $\widetilde{F}^{(n)}$, then the decoding phase can be formulated as:

$$\hat{y} = Decoder(h'_t, V_t), \quad V_t = \sum_{j=1}^{T} \alpha_{t,j} \tilde{f}_j^{(n)}, \tag{3}$$

where h'_t is the hidden state of the LSTM decoder at step t, while $\alpha_t \in \mathbb{R}^T$ is a vector of spatial attention weights as defined and computed in [19]. Finally, the spatial attention-based decoder predicts the recognition results of the multi-frame license plate sequences.

2.3. Loss Functions for Multi-task Training

Most of the previous attention models are inserted as intermediate modules into the whole frameworks, where the attention modules may not be directly supervised and trained. However, unlike them, the proposed attention model is explicitly trained in a straightforward manner. Assume the plate images in the training set as $\mathcal{I} = \bigcup_{n=1}^{N} I_n$. Given an input image $I \in \mathcal{I}$, its labels can be denoted as $\{y, \tilde{s}\}$, where y represents the recognition label while $\tilde{s} = \{s_1, s_2, \cdots, s_L\}$ ($L = 5$ in our dataset). s_i is the ground truth quality score of the i-th license plate frame in I. Considering both plate recognition and image quality scoring, the total loss for deep multi-task training can be formulated as:

$$\mathcal{L} = \mu L_R + (1 - \mu) L_S + \lambda \mathcal{R}, \tag{4}$$

where μ is a tunable parameter to weigh the importance between the recognition loss L_R and the quality scoring loss L_S which is set to 0.5. In addition, $\mathcal{R}$ represents ℓ_2 regularization to alleviate overfitting and its weight $\lambda = 0.5$ in our experiments. Given the feature map $\widetilde{F}$, the recognition loss can be computed as mentioned in [19]:

$$L_R = -\sum_{t} \ln P(y \mid \widetilde{F}). \tag{5}$$

We exploit ℓ_1 norm defined in [20] as the loss function of image quality scoring, given by:

$$L_S = \sum_{i=1}^{L} \|s_i - \hat{s}_i\|_1 = \sum_{i=1}^{L} \Big| s_i - \frac{1}{M/L} \sum_{j=1}^{M/L} \hat{s}_{i,j} \Big|. \tag{6}$$

In addition, the Adam optimizer is utilized to minimize the total loss $\mathcal{L}$ in practice.

3. EXPERIMENTS

In this section, we perform extensive experiments to demonstrate the capabilities of the proposed framework IQ-STAN and utilize different protocols to measure its performance. Moreover, the effectiveness of the presented IQ-guided spatio-temporal attention is confirmed by the ablation study.

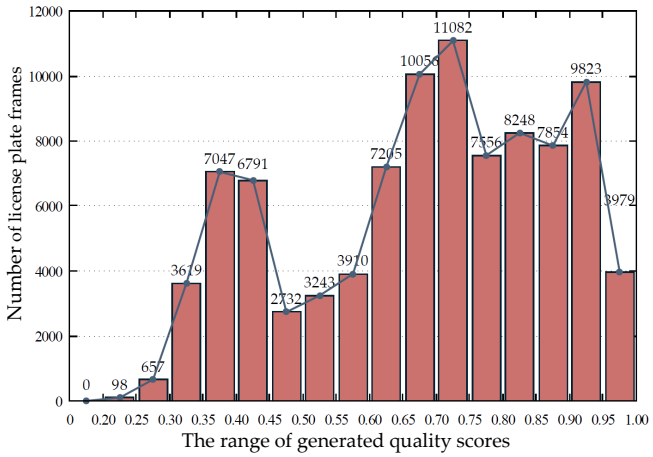

Fig. 3. The distribution of image quality scores, which are generated using the trained NR-IQA model in [20] and then normalized to the range of $[0, 1]$.

3.1. Datasets and Experimental Settings

Since there exists no available dataset for multi-frame license plate recognition with image quality scores, we build a new dataset and perform experiments on the self-built real-world LPR dataset named OPT-MFLPR. It is constituted of 18,780 Chinese vehicle license plate images with the resolution of 750×50 pixels, which are all acquired by the same camera equipment. Unlike other LPR datasets, each image in OPT-MFLPR contains 5 frames of the same license plate captured and detected at different times, as exhibited in Fig. 1.

However, training the multi-task architecture IQ-STAN requires not only recognition labels but also the ground truth of image quality scores, yet the latter's acquisition is extremely costly and labor-intensive. In fact, the quality scoring branch in IQ-STAN is similar to no-reference image quality assessment (NR-IQA) task [20], which aims to evaluate the objective image quality without pristine reference images. Inspired by this, we yield the ground truth quality scores based on the excellent trained NR-IQA model [20] and the distribution of generated quality scores is illustrated in Fig. 3. For each image I_n in OPT-MFLPR dataset, its labels can be denoted as $\{y^{(n)}, \tilde{s}^{(n)}\}$ and $\tilde{s}^{(n)} = \{s_1^{(n)}, s_2^{(n)}, \cdots, s_5^{(n)}\}$, as mentioned in section 2.3. In the evaluation, 12,000 images are used as the training set and others as the test set.

3.2. Performance Evaluation and Ablation Study

In order to comprehensively evaluate the proposed multi-task framework IQ-STAN, extensive experiments with various experimental protocols are conducted. For the fair competition, all experiments are performed on the OPT-MFLPR dataset using the same training strategy, where the batch size is set to 10 and the learning rate is initialized to 0.001.

The performance of recognition is evaluated by the recognition accuracy, which is widely used in [2, 5, 6]. The recognition accuracy is defined as the total number of plates dividing into the number of correctly recognized plates. Note that recognizing all characters accurately means the correct recognition. We use EasyPR[1], BaiduLPR[2] and AlibabaLPR[3] as the comparative LPR methods since they are famous open source projects for Chinese plate recognition. In order to clarify the effectiveness of IQ-guided spatio-temporal attention, the remainder after removing IQ scoring branch is viewed as baseline method in the experiments. Moreover, we employ two different measures for recognition accuracy, including or excluding Chinese characters. The proposed IQ-STAN achieves the highest accuracy in OPT-MFLPR dataset, as shown in Table 1. In addition, two metrics in [20], namely linear correlation coefficient (LCC) and Spearman rank order correlation coefficient (SROCC), are exploited to evaluate the IQ scoring performance which directly affects the temporal attention maps generation. In comparison with training only IQ scoring branch, the multi-task framework IQ-STAN performs better for the metric LCC in Table 2.

In terms of the computational speed, the proposed IQ-STAN takes about 157ms per five frames (approximately at 31.85 fps), which is efficient for LPR systems.

Table 1. Performance comparison of different methods on OPT-MFLPR dataset.

Methods	Recognition Accuracy (%)	
	Including Chinese characters	Excluding Chinese characters
EasyPR[1]	71.78	78.04
BaiduLPR[2]	86.23	91.80
AlibabaLPR[3]	88.28	**94.27**
Baseline (Ours)	**90.11**	93.92
IQ-STAN (Ours)	**93.07**	**95.35**

Table 2. Performance evaluation of the IQ scoring module.

Methods	LCC	SROCC
Only IQ Scoring	0.830	**0.819**
IQ-STAN	**0.847**	0.805

4. CONCLUSIONS

In this work, we propose a deep learning-based multi-task architecture named IQ-STAN for LPR systems, which jointly evaluates perceptual image quality and recognizes the characters in license plates. IQ-STAN involves a novel IQ-guided spatio-temporal attention network that can be trained in an end-to-end manner. The experiments on the self-built real-world LPR dataset indicates its effectiveness and efficiency.

[1]https://github.com/liuruoze/EasyPR
[2]https://cloud.baidu.com/doc/OCR/OCR-API.html
[3]https://help.aliyun.com/document_detail/56879.html

5. REFERENCES

[1] Shan Du, Mahmoud Ibrahim, Mohamed Shehata, and Wael Badawy, "Automatic license plate recognition (alpr): A state-of-the-art review," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 2, pp. 311–325, 2013.

[2] Hui Li, Peng Wang, Mingyu You, and Chunhua Shen, "Reading car license plates using deep neural networks," *Image and Vision Computing*, vol. 72, pp. 14–23, 2018.

[3] Yuan Yuan, Zhitong Xiong, and Qi Wang, "An incremental framework for video-based traffic sign detection, tracking, and recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 7, pp. 1918–1929, 2016.

[4] Qi Wang, Junyu Gao, and Yuan Yuan, "Embedding structured contour and location prior in siamesed fully convolutional networks for road detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 230–241, 2018.

[5] Geesern Hsu, Jiunchang Chen, and Yuzu Chung, "Application-oriented license plate recognition," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 2, pp. 552–561, 2013.

[6] Minghui Zhang, Wu Liu, and Huadong Ma, "Joint license plate super-resolution and recognition in one multi-task gan framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 1443–1447.

[7] Amir Hossein Ashtari, Md Jan Nordin, and Mahmood Fathy, "An iranian license plate recognition system based on color features," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 4, pp. 1690–1705, 2014.

[8] Jianbin Jiao, Qixiang Ye, and Qingming Huang, "A configurable method for multi-style license plate recognition," *Pattern Recognition*, vol. 42, no. 3, pp. 358–369, 2009.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, , no. 6, pp. 1137–1149, 2017.

[10] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[11] Rayson Laroca, Evair Severo, Luiz A Zanlorensi, Luiz S Oliveira, Gabriel Resende Gonçalves, William Robson Schwartz, and David Menotti, "A robust real-time automatic license plate recognition based on the yolo detector," in *International Joint Conference on Neural Networks*, 2018, pp. 1–10.

[12] Zhi Yang, Feng-Lin Du, Yi Xia, Chun-Hou Zheng, and Jun Zhang, "Automatic license plate recognition based on faster r-cnn algorithm," in *International Conference on Intelligent Computing Methodologies*, 2018, pp. 319–326.

[13] Hui Li, Peng Wang, and Chunhua Shen, "Toward end-to-end car license plate detection and recognition with deep neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 1126–1136, 2019.

[14] Vojtech Vasek, Vojtech Franc, and Martin Urban, "License plate recognition and super-resolution from low-resolution videos by convolutional neural networks," in *British Machine Vision Conference*, 2018, p. 132.

[15] Yuexian Zou, Yi Wang, Wenjie Guan, and Wenwu Wang, "Semantic super-resolution for extremely low-resolution vehicle license plate," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 3772–3776.

[16] Hilario Seibel, Siome Goldenstein, and Anderson Rocha, "Eyes on the target: Super-resolution and license-plate recognition in low-quality surveillance videos," *IEEE Access*, vol. 5, pp. 20020–20035, 2017.

[17] Thidarat Pinthong, Worawut Yimyam, Narumol Chumuang, and Mahasak Ketcham, "License plate tracking based on template matching technique," in *IEEE International Symposium on Communications and Information Technologies*, 2018, pp. 299–303.

[18] Michael Donoser and Horst Bischof, "Efficient maximally stable extremal region (mser) tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 553–560.

[19] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou, "Focusing attention: Towards accurate text recognition in natural images," in *IEEE International Conference on Computer Vision*, 2017, pp. 5076–5084.

[20] Le Kang, Peng Ye, Yi Li, and David Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.