Coarse-to-Fine Attention: Self-Luminous Object Detection for Autonomous Driving

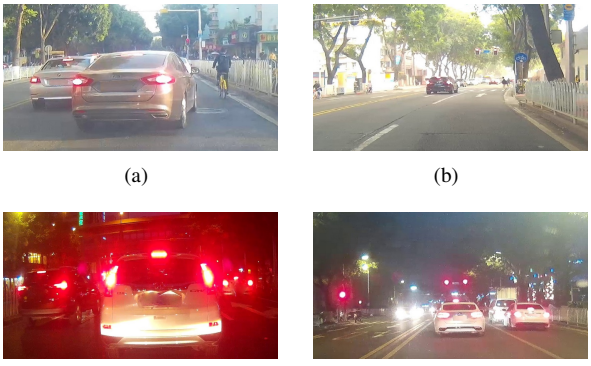
Shenao Zhang, Bo Wu, Xin He, Lei Kuang, and Delu Zeng

***Abstract*—Recent progress in deep learning and large-scale datasets enormously benefit image understanding. Several works addressing semantic segmentation or general object detection focus on the visual understanding of street scenes. However, few datasets and algorithms are concerned with the understandings of car tail light signals, which are important for autonomous driving systems. This task is challenging due to the abstract semantic meanings and vague boundaries of the light signals. In this paper, we introduce the Vehicle Light Signal (VLS) dataset, a large-scale image dataset for the light signal detection task, which is challenging due to the vague object boundaries in different environments as well as the interference of the noise, *e.g.*, lights in the surroundings. We propose a Coarse-to-Fine Attention mechanism to combat these challenges to dynamically localize the informative patterns that correspond to these regions. Specifically, in a Mixture of Experts manner, features that are responsible for bounding box localization and category classification are first clustered to expert channels with our coarse-attention module. By mixing the information from the learned expert channels with the fine-attention module, the noisy features in the coarse region proposals can be effectively eliminated and thus benefiting the training process. We evaluate the proposed attention mechanism on the VLS dataset to show the effectiveness of extracting informative patterns from rough regions, and demonstrate how it works on the CIFAR dataset.**

***Index Terms*—Object detection, attention mechanism, computer vision, deep learning.**

I. Introduction

DETECTION is a fundamental task in computer vision and has achieved successes in many real-world scenarios, including general object detection [1], [2], road object detection [3] and face detection [4]. Despite these progresses, self-luminous target detection remains an open problem with important research significance [5], [6], *e.g.*, the perception of tail light signals of surrounding cars in autonomous driving systems. Understanding the behaviors of cars by light signal perception will facilitate the decision making of the autonomous cars. However, recognizing such light signals is non-trivial due to the vague object boundaries and the interference of noise. Previous works using complex image processing or adopting prior knowledge in the specific datasets often suffer from low accuracy and the resulting algorithms are narrowly specified, leaving the common properties of the self-luminous objects underexplored.



(a) (b) (c) (d)

Fig. 1: Illustration of self-luminous object detection task. (a) & (b): The contrast of light signals and the background has large variance, *i.e.*, it is much easier to recognize signals in (a) than in (b); (c): Vague boundaries of bounding boxes that characterize the semantic categories; (d): Interference noise of environment light.

Self-luminous target detection is challenging and different from general object detection for the following reasons. Firstly, much more noise exists in such scenarios, *e.g.*, bounding box regressor will be fooled to include the light noise from other cars and the surrounding road lamps, and signal category classifier will give wrong predictions. Secondly, semantic information within the bounding boxes are important since the general contours and appearances look the same for all ground truths and the semantics are key to make correct predictions. Thirdly, the brightness and contrast of images have large variance, causing boundaries of objects sometimes vague (the lights from many sources are mixed) and separated light signals hard to distinguish. We provide an illustration in Fig. 1.

Such differences also result in different detection algorithms. In two-stage general object detectors, region proposals are generated in the first stage while the accurate locations of bounding boxes and object categories are predicted in the second stage [7], [8], [9]. While this pipeline is shown suitable for general object detection tasks, in this work, which aims to detect self-luminous targets with much noise impacting the prediction of the vague boundaries and extracting semantic

Shenao Zhang is with the Department of Electrical and Information Engineering, South China University of Technology, Guangzhou, China (e-mail: shenaozhang@mail.scut.edu.cn).

Bo Wu is with the MIT-IBM Watson AI Lab, IBM Research, Cambridge, MA, 02142 USA (e-mail: bo.wu@ibm.com).

Xin He is with the School of Mathematics, South China University of Technology, Guangzhou, China (e-mail: msxinhe@mail.scut.edu.cn).

Lei Kuang is with the Department of Computer Science in Columbia University in the City of New York, NY, 10027 USA (e-mail: lk2807@columbia.edu).

Delu Zeng is with the School of Mathematics, South China University of Technology, Guangzhou, China (e-mail: dlzeng@scut.edu.cn).

Manuscript received April 19, 2005; revised August 26, 2015.

information, will lead to bad performance. Compared with general object detection tasks, self-luminous objects have specific structures, and focusing on the discriminative parts is needed when such structural prior exists. For example, boundaries of objects under low light environments [10] are hard to decide and features of traffic light signals [5] under different scenarios,*e.g.*, in the rain or at night, have large variance. Thus, performance of general two-stage detectors will be constrained. The region proposal generator, *e.g.*, RPN in Faster R-CNN [9], will generate low-quality proposals due to the interference in the surroundings. Thus, accuracy of bounding box positions are limited and predicted categories which are highly dependent on the quality of region proposals will also be significantly influenced by confusing information.

To solve the limitations, we can either learn a better region proposal network that generates both proposal locations and foreground prediction more accurately, or dynamically extract informative patterns from region proposals, allowing more noise in proposal regions. In this work, we follow the second efficient way to insert an attention module for noise filtering and information delivery. We propose a novel attention mechanism to combat the challenges lying in self-luminous target detection tasks, called Coarse-to-Fine Attention (CFA) mechanism. CFA first extracts features from low-quality coarse region proposals by encouraging a set of expert extractors in network, and then discovers accurate regions in spatial domain to guide detection. With this idea, both the above issues can be alleviated. On the one hand, the accurate bounding boxes can be predicted by considering into account the spatial locations of certain patterns generated by the expert extractors in coarse attention, especially with structured objects where both the structures and the correlations between them needed to be considered to make good predictions. On the other hand, informative patterns which are crucial for classifying such self-luminous targets, *e.g.*, signal semantics of traffic lights, can be effectively integrated to obtain performance boosting using our attention mechanism, which we name Coarse-to-Fine Attention mechanism that can be integrated to two-stage detectors for performance boosting.

Specifically, coarse attention can be interpreted as an attention mechanism in channel dimension which encourages expert channels to extract information in coarse proposals. Modal feature vectors, which are responsible for distinguishing the sub-parts without concrete appearance semantics in object, are generated to represent certain patterns underlying data in an unsupervised manner. Since our coarse attention builds a bridge between channel dimension and spatial dimension, expert feature extractors are expected to focus on specific informative spatial patterns. Benefited from information delivery to modal feature vectors at expert channels, fine attention are followed to predict accurate attention maps in spatial domain. In this way, region proposal networks are implicitly optimized to contain desired information with backpropagation, *i.e.*, the areas that characterize the semantic meanings and boundaries.

In this paper, we publish our Vehicle Light Signal (VLS) dataset, which contains traffic images of real-world roads as the complex real situations cars might encounter, with bounding boxes of surrounding cars and labels of their behavior, according to their tail light signals. We focus on this challenging task: perception of tail light signals of other vehicles in real-world scenarios. The challenges lie in not only the high variance of light signals under different conditions and thus the need for extracting and interpreting semantic patterns, but also interference noise contained in environments. That is, structures in light signal bounding boxes needed to be discovered and correlations are needed for understanding the semantic meanings to avoid interfering by surrounding road lights. Details and statics can be found at Section IV.

The contributions of our work can be divided into four folds:

- We propose a Coarse-to-Fine Attention (CFA) module that can be inserted directly into Faster RCNN or other conventional two-stage detectors to tackle the challenges in self-luminous object detection tasks, which boosts the detection performance.
- We give mathematics proof and intuitive inspiration of the proposed CFA mechanism to validate its effectiveness.
- We publish our challenging Vehicle Light Signal (VLS) dataset for perception of the behaviours of surrounding vehicles with their light signals.
- We conduct experiments on CIFAR-10 dataset for proof-of-correctness of our coarse attention module, and report the performance of the proposed attention mechanism on VLS dataset, which beats the baseline by a large margin.

II. Related work

A. Object Detection

Popular object detection tasks mainly include detecting general objects, like COCO [1] and PASCAL VOC [2] challenges. Traditional detection approaches can be classified into two categories, i.e., two-stage detector and one-stage detector, based on whether they have a separate region-of-interest proposal network. While one-stage detectors [11], [12] are relatively simple by combining processes of generating proposals and regression using predetermined anchors, two-stage detectors adopt an intuitive idea to first roughly localize where objects we are interested in are, i.e., proposals. Then a simple classifier is followed to regress the final object category and location. Among two-stage detectors, R-CNN [7] adopts selective search to generate proposals, and is the first work combining region proposals with CNNs. However, R-CNN is slow due to that it is a multi-stage pipeline and needs to learn within every object proposal.Fast R-CNN [8] accelerates it by sharing computation. Faster R-CNN [9] proposes RPN for efficiently generating high-quality proposals, and achieved excellent performance on many datasets. It has been discovered in two-stage detectors that the fine-grained and interference property between classes is a drawback which can be improved via our attention methods, while Faster R-CNN is considered as a mainstream in two-stage detectors in recent years. In this paper, we mainly focus on the design of two-stage detectors such as Faster R-CNN, Our CFA method tends to reduce the drawback and the performance is verified on our VLS dataset.

B. Self-Luminous Object Detection

Despite of general object detection tasks, different underlying properties in data can lead to significantly varied detection

mechanisms. When object of interests come to human faces [4], structured dependencies are essential for good performance. For small object detection task [13], general detectors will also lead to low recall rate due to the lack of high-resolution information. In this paper, we mainly focus on a challenging area of interest, called self-luminous object detection. For example, the rear light perception of a car is difficult when road lamps and surrounding rear light of other cars exist, causing prediction of bounding boxes and categories of labels a challenge.

Light signal detection [14], [15], [16] shares similarities with the focused self-luminous object detection. To the best of our knowledge, we are the first work to combat the challenges brought by self-luminous objects in light signal detection tasks, which we will show dominate the performance at Section IV. Previous works that directly adopt general object detectors [17], [18], [16], [19], [20], [21], [22] like RCNN [7], Faster RCNN [9] or YOLOv3 [23] do not take the properties of self-luminous objects into account, and are thus performance constrained.

Other lines for solving light signal perception tasks include dealing with the small objects in images [16] and conducting image processing to make the boundaries clear [24], [25]. Previous works that adopt image processing and feature engineering [14], [26], [27] analyze the structures of data, and are constrained to certain properties of the dataset. On the contrary, we adopt attention mechanism to efficiently localize and recognise the light signals and can be directly trained on related datasets. [15] adopts prior knowledge of data and trains a neural network, which is different than ours that our proposed Coarse-to-Fine Attention mechanism can be inserted to any two-stage detector and solve tasks without assuming the particular structures or appearances of data.

C. Attention Mechanism

Attention mechanism has proven useful to boost the discriminative power to obtain expressive models. Deep neural network models are highly benefited from attention mechanism, which is an effective way for capacitating the ability of dynamically extracting informative features. SENet [28] assigns each channel in a neural network a weight and optimizes the weights in an end to end manner. Although our coarse attention also weights each channel of feature map, CA iteratively clusters features and generates channel weights (like EM), while SE leans channel weights implicitly as black-box. For our method, no additional parameters are ideally needed if we use channel scores directly, instead of adding layers to fit the superficial statistics. As for spatial attention methods, spatial transformer networks (STN) [29] is designed to learn transformation parameters in the spatial domain. We also provide a Fine Attention (FA) module to localize the precise discriminative region in a much simpler way with the help of CA.Though the goal of our method is the same of using both channel attention and spatial attention. Our Coarse-to-Fine Attention is different from simple combining a channel attention and a spatial attention as we used a pseudo label to train the coarse attention block. Other kinds of attention

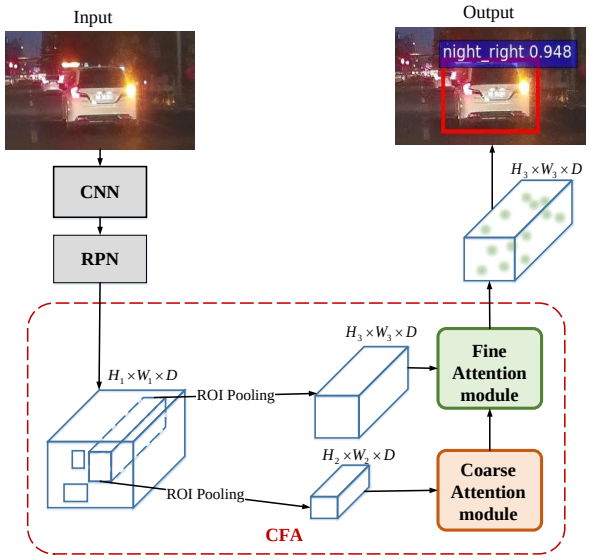


Fig. 2: The overview of CFA. Our attention has two modules: coarse attention module and fine attention module. Coarse attention builds a bridge between channel dimension and spatial dimension by clustering information in feature maps to expert channels which have the potential to extract certain patterns, fine attention locates informative spatial parts benefiting from expert channels.

mechanisms are also widely adopted for better extracting features in both computer vision and natural language processing fields. [30] shed light on attention mechanism to solve the problem of lacking long-range relationships in machine translation problems. [31], [32], [33] also provide evidence that attention can facilitate different tasks.

We notice that the same term coarse-to-fine is adopted by previous works [34], [35]. However, the coarse and fine in their context are both spatial attention mechanisms, in terms of a large global region towards a smaller local region in it, which greatly differ from ours. Specifically, [35] proposed a heuristic attention mechanism for image classification by first focusing on the object region and then determine its category. And recurrent neural network is adopted to make predictions. [34] developed a hierarchical inception module to aggregate multi-scales spatial features for text detection. Different from these works, the coarse attention in our method is an EM-style algorithm that iteratively estimate central features and optimize model parameters. Thus, by clustering the discriminitive patterns to high-weight channels, the high-weight expert channels have the potential to focus on one of the vital components of the target. By mixing the information from the learned expert channels, the noisy features in the coarse region proposals can be effectively eliminated and benefit the training of two-stage detectors. Once again, our algorithm ideally needs no additional parameters, which is different from previous attention methods.

III. Approach

In this section, we will introduce the proposed Coarse-to-Fine Attention (CFA) mechanism for detection of luminous objects. As stated in section I, CFA is designed to deal with the challenges of what characterize the objects, including the bounding boxes and semantic meanings, alleviating the problems caused by noise in coarse proposal regions, *e.g.*, surrounding road lights will cause confusion if they are also included in proposal regions.

A. Framework Overview

Coarse-to-Fine Attention (CFA) consists of two coupled attention modules: coarse attention and fine attention. Coarse attention module clusters informative features to expert channels as central feature vectors and assign attention weight for each channel iteratively. And fine attention module leverages the information from coarse attention module and extract spatial attention regions in the rough proposals. By generating the central feature vectors from coarse attention, fine attention is followed. In this work, we use central feature vectors and modal feature vectors interchangeably, which both refer to the features that contain discriminative features and are responsible for the bounding box localization and category classification.

Coarse attention bridges the connections between channel dimension and spatial dimension by considering informative patterns underlying the data each channel possesses. Fine attention fuses the central feature vectors to give a spatial attention map for bounding box regression and class prediction. CFA can be inserted in any two-stage detectors and the overall network can be optimized in an end-to-end manner. One example of our CFA module integrated with Faster RCNN [9] is shown in Fig. 2. Since fine attention is responsible for generating spatial attention map, larger size of spatial dimensions are desired, *e.g.*, $h_3 = 2h_2$, $w_3 = 2w_2$ as in our experiment. By assigning each channel an attention weight and encourage central features to channels with high attention weight, *i.e.*, expert channels, patterns that correspond to self-luminous objects in rough proposals are extracted from the noise.

B. Coarse Attention Mechanism

We begin with our coarse attention module, which extracts expressive modal vectors that are responsible for localizing disentangled spatial patterns with a set of expert feature extractors, *i.e.*, a subset of channels with highest attention scores that together characterize the objects. For each class of object, the corresponding expert channels are excited and optimized to extract useful features from the inaccurate region proposals. Thus, the useless noise which varies in different images will not obstruct the training process of the expert feature extractors. By doing so, important features in the coarse regions is delivered to certain channels and form the central feature vectors.

The main idea of coarse attention is showed as in Fig. 3. Expert feature extractors are encouraged to focus on specific

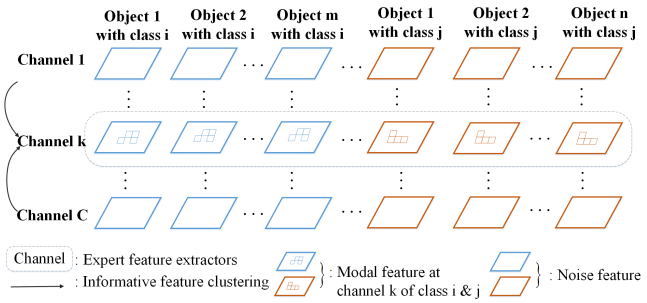


Fig. 3: The key insight of the coarse attention is shown. To extract informative features from coarse proposal regions, expert channels with high attention weights are excited and the informative features are clustered to the certain expert channels, by encouraging modal features of each class to be extracted by experts. Such expert channels are expected to focus on one of the vital components of the target. The fine attention module then follows to mix the information from expert channels, which we will discuss later. The noisy features in the coarse region proposals can thus be effectively eliminated and benefit training.

patterns by forcing similar modal vectors to be generated with the same categories which contain similar semantic information. The spatial positions are expected to benefit bounding box localization, and symmetrically, *i.e.*. in a iterative way. With backpropagation, correlated distinguishable features can be delivered to the clustering of specific patterns in a way similar to mixture models [36] to update modal vectors and send useful information that are responsible for such patterns in a dynamic way.

Each channel of features are position-aware [9], [37] and we expect all instances of the same categories to share central features if with several perfect pattern extractors. The properties above do not hold for features lie in each channel, but we can encourage important information to cluster to several channels and implicitly disentangle features with sparse representation by backpropagation. That is, we add regularization to model the desired modal vectors for each class by dynamically weighting the importance of channels. Channels with high attention score embed certain patterns while avoiding integrating noisy features.

We implement the above idea of generating candidate modal features for each category at every channel by averaging previously trained samples at the particular channel. Then importance weight of each channel is dynamically assigned by calculating similarities with assumed modal features of the same category as well as dissimilarities with other categories to regularize discriminative features clustering to a set of experts, which shares similarity with mixture models [36] to update modal vectors and weight assignments iteratively by EM algorithms [38]. As training time increases, central features of important patterns can be obtained by channels with high attention scores optimized in an end-to-end manner.

The illustration of our coarse attention module is at Fig. 4. To learn expert feature extractors, central features for each

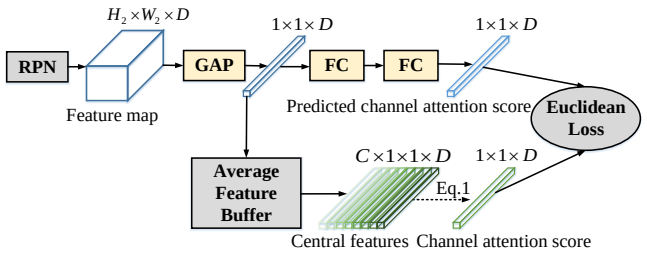


Fig. 4: Coarse attention module learns a set of expert feature extractors by encouraging features to cluster to channels with high attention scores. Central features of each category are represented by the average of all previous features. D is the channel dimension, C is total number of categories. Channel score is calculated by Equation 1.

class are approximated by averaging all generated features at the same channel of the same class. Scores at each channel of current samples are needed to ensure important features extracted by channels with high scores. By encouraging features at expert channels share similarity with central features of the same class, while distinguishing from central features of different classes, channel scores S of sample i with category c can be calculated as:

$$S = -(A^c - F_i)^2 + \frac{1}{C-1}\sum_{c' \neq c}(A^{c'} - F_i)^2, \tag{1}$$

where F_i is feature of current sample i, A^c is the averaged central feature of class c calculated by averaging central feature from buffer and F_i, and C is the total number of classes, $\frac{1}{C-1}$ controls the weight of contribution for samples with different labels.The above process shares similarities with EM algorithms [38] to first generate channel attention scores with weights before optimization as Expectation step, followed by delivering informative features to channels with high scores and optimizing parameters of RPN as Maximization step. Although it is possible to directly use the generated channel score, we optimize a MLP to fit the channel attention scores so that the training process is stable when noise dominates the proposals and the average feature buffer is avoided when testing. Also, using the predicted channel weights when deployment avoids additional memory caused by checkpointing and keeping central features.

Next, we show how our attention mechanism works mathematically.

Beginning with global average pooling [39], discriminative image regions are depicted by generating class activation map (CAM) for classifying particular categories. More specifically, the feature of sample i with class c after GAP is:

$$F_d = \frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y} f_d(x, y), \tag{2}$$

where d denotes the index of channel dimension, $f_d(x, y)$ represents the feature map at channel d at position (x, y). Then FC layer is followed to compute the original classification output score O_{org}:

$$O_{\text{org}} = \sum_d (W_d F_d), \tag{3}$$

where W_d is the weight of FC layer of class c at channel dimension d. After optimizing, scores for the same class of current sample,*i.e.*, O_{org} should be higher. With the proposed coarse attention mechanism, each channel is assigned an attention weight S as in Eq. 1, and the corresponding output score is:

$$O = \sum_d (S_d W_d F_d) \tag{4}$$

By eliminating F_d^2 in S_d, the output score O can be calculated as:

$$\begin{aligned} O &= \sum_d (-(A_d^c - F_d)^2 + \frac{1}{C-1}\sum_{c' \neq c}(A_d^{c'} - F_d)^2 W_d F_d) \\ &= \sum_d ((\frac{1}{C-1}\sum_{c' \neq c} A_d^{c'2} - A_d^{c2}) \\ &+ 2F_d(A_d^c - \frac{1}{C-1}\sum_{c' \neq c} A_d^{c'} - A_d^c) W_d F_d) \\ &= \sum_d ((\alpha F_d + \beta) W_d F_d), \end{aligned} \tag{5}$$

where F_d is features of current sample at channel d, A_d^c is the central features at channel d of category c. In our implementation, A_d^c is approximated by averaging all previously trained samples of instances with category c at channel c and is saved in the buffer at each iteration. So $\alpha = 2(A_d^c - \frac{1}{C-1}\sum_{c' \neq c} A_d^{c'} - A_d^c) W_d F_d)$ and $\beta = \frac{1}{C-1}\sum_{c' \neq c} A_d^{c'2} - A_d^{c2}$ are treated as constant vectors.

The optimization can thus be viewed as two iterative steps. Firstly, the central features and attention weights are estimated. Then the model parameter is optimized for loss minimization. For each channel d, we first calculate F_d^{avg} and $F_d F_d^{\text{avg}}$ as modal feature vector and channel attention weight respectively. Since weights W_d are parameters corresponding to the same class c, maximizing O is desired. When attention score is high at channel d, $W_d F_d$ is also expected to be high, and W_d is optimized accordingly. Thus, in the training process, F_d is encouraged to contain modal features for the same category to make output score O large with expert channel weight W_d. In this way, central features as well as W_d and parameters that are responsible for generating F_d are optimized iteratively in a similar manner with EM algorithms. We also conduct a proof-of-correctness experiment on CIFAR-10 at Section IV-B and verify that informative features are clustered to expert channels.

C. Fine Attention Mechanism

Coarse attention builds the connection between the channel dimension and spatial dimension by clustering distinguishable spatial features to a set of channels. However, to get a better understanding of semantic meanings and thus predicting bounding boxes and categories more precisely, we need a

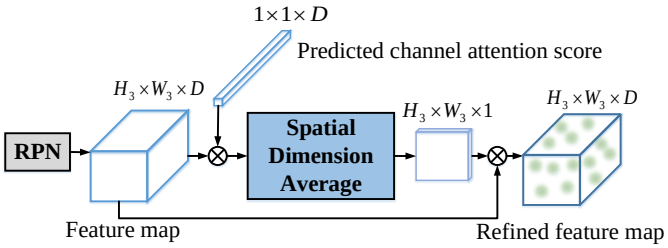


Fig. 5: Fine attention module generates spatial attention map to locate the discriminative parts that are responsible for predicting bounding boxes and semantic categories.

fine attention module to localize the exact spatial regions that are responsible for the semantic categories and ground truth bounding boxes. From the coarse attention module, we obtained multiple channel experts, each of which focuses on specific central features and is expected to represent a certain part of the object of interest. We propose to aggregate features from the optimized expert channels by generating a spatial attention mask.

The fine attention is implemented as the weighted-sum of the channel features maps. High weights of the channels indicate that the channel feature map has high confidence to represent a subset of the vital parts, *e.g.*, the right taillight of the vehicle in the tail signal detection task. In this way, the spatial areas that the experts focus together play the role to characterize the semantic categories and bounding box locations.

The architecture of our fine attention module is illustrated in Fig. 5. The input feature maps are generated from RPN through ROI pooling. Then the predicted channel attention scores obtained from the previous CA module are adopted to generate the spatial attention mask, and features are refined accordingly. That is,

$$f(x,y) = \sigma(\sum_{d}(S_d \cdot f_d(x,y)))f(x,y), \tag{6}$$

where $f(x,y)$ is input features, d is channel index, S_d is channel attention score at channel d generated by coarse attention module, σ is sigmoid activation function.

IV. Experiments

In this section, we conduct experiments to evaluate the proposed Coarse-to-Fine Attention (CFA) mechanism. We first provide details and instances of our Vehicle Light Signal (VLS) dataset, and then evaluate the performance of CFA on the VLS dataset, including both comparisons with other state-of-the-art approaches, visualizations and analysis. Then experiments are designed on the CIFAR-10 dataset to validate the effectiveness of the proposed attention mechanism.

A. Experiments on Vehicle Light Signal Dataset

In this part, we first provide details of our dataset and some image instances. We then report the performance of the our models, comparing it to other object detection algorithms. Visualization and analysis are also given.

TABLE I: Instance numbers of each category in our VLS dataset. The total number of classes are 8, containing moving forward, brake, turning left and turning right, in daytime mode and nighttime mode respectively.

Day				Night			
forward	brake	left	right	forward	brake	left	right
2213	1750	889	931	2123	1075	783	807

Dataset. We adopt our Vehicle Light Signal (VLS) dataset for results analysis due to the challenges of localizing light signals and extracting the semantic information. VLS dataset contains 4 common behaviours of vehicles: driving forward, braking, turning left, and turning right. We classify each behaviour signal at 2 scenarios: day and night, since the lighting signals are not the same when during the day and night. We do that since the signals have differences (determined by headlights). Day and night characterize the light signals, rather than the environment (each signal belongs to one category). We collect our data from the driving recorder by selecting dissimilar images from each video sequence (15 frames from 15−minutes video). The VLS dataset includes 7720 images, 8 categories, and 10571 instances totally. We classify the bounding boxes by driving forward, braking, turning left and turning right, in daytime and night respectively. We randomly choose 60% of samples as training data, 20% as validation and 20% as testing in experiments, and this ratio can be fixed for different models in other experiments. Table I shows the instances distribution of each vehicle light signal in our VLS dataset.

Implementation Details. We implement our models using Caffe [40]. The models are trained on one NVIDIA GTX 1080Ti. For all our models, the initial learning rate is set to 10^{-3}, the momentum is set to 0.9 and the weight decay is 5×10^{-4}. In our task, the input of the system is original image, and the output contains the position of vehicle and the classification result of vehicle light signal in this image. Thus we use the average precision (AP) of each category and the mean average precision (mAP) in object detection as our performance evaluation criteria.

Performance Comparison. In this section, we compare the performance between our method and other algorithms. A large part of state-of-the-art methods which are used in light signal detection tasks adopt general object detectors. We also follow previous works to first evaluate the performance of the popular two-stage general object detector Faster RCNN [9] with different backbones, and evaluate the performance when integrating our Coarse-to-Fine Attention mechanisms. The mAP of Faster RCNN with backbone VGG16 is 61.05%, compared with 67.24% when also integrating our Coarse-to-Fine Attention mechanisms, increased by 6.19%. With resnet50 [41] as backbone, the average precision of some categories are increased with our attention modules. The results are shown at Table II. We can see that with the same backbone network, when trained with CFA, the detector outperforms the original models, especially when the backbone

Fig. 6: Samples of the dataset and the detection results of our algorithm. Detection categories ended with 'no' denote that the vehicles are normally moving forward.

Methods	Backbone	Day				Night				mAP
		forward	brake	left	right	forward	brake	left	right	
Faster RCNN	VGG16	67.17	84.31	39.65	53.46	84.33	80.79	39.53	39.16	61.05
Ours	VGG16	**79.88**	**86.51**	45.69	55.21	**86.14**	**83.50**	49.28	51.67	**67.24**
Faster RCNN	Resnet50	74.35	85.59	41.15	49.46	83.72	80.85	47.69	48.00	63.85
Ours	Resnet50	70.14	84.96	**46.25**	48.97	83.75	83.49	**58.67**	**58.15**	66.80
RFnet [16]		69.44	83.20	35.67	39.71	72.23	74.91	27.67	29.86	54.09

TABLE II: Performance comparison on VLS dataset. The proposed CFA mechanism consistently boosts the performance of Faster RCNN baselines with different backbone network architectures, and outperforms RFnet by a large margin.

Backbone	C-A	F-A	Day				Night				mAP
			forward	brake	left	right	forward	brake	left	right	
VGG16	√		78.42	86.32	42.02	45.75	85.97	81.91	43.16	47.22	63.85
	√	√	**79.88**	**86.51**	45.69	55.21	**86.14**	**83.50**	49.28	51.67	**67.24**
Resnet50	√		77.08	85.18	44.75	**56.56**	82.59	81.63	46.44	52.67	65.86
	√	√	70.14	84.96	**46.25**	48.97	83.75	83.49	**58.67**	**58.15**	66.80

TABLE III: Ablation studies of the two attention components. C-A refers to Fatser RCNN integrated with our coarse-attention module and F-A refers to the proposed fine-attention module.

network performs poorly, *e.g.*, VGG16 [42]. That is, our CFA module have the ability to extract informative features from low-quality proposals with more noise or interference.

Other lines of state-of-the-art models that detect light signals analyze the properties of data and adopt either feature engineering or image processing to combat certain challenges. We refer the readers to Section II-B. We choose RFnet [16] among such works, which mainly alleviate the problems that the overall appearances in some scenarios are hard to recognise. Our results outperforms RFNet by a large margin, which we argue is because the properties of self-luminous objects dominant the performance. We provide some samples of the dataset images and the detection results of our algorithm in Fig. 6.

Ablation Study. We then conduct ablation experiments to evaluate the influence of our coarse attention and fine attention components. Results are provided at Table III. We can notice that the overall performance gets better when adopting coarse attention, gets best when adopting fine attention after coarse attention, validating the effectiveness of both modules.

Analysis of Region Proposals. To give an intuitive understanding of the influence of our attention modules, we analyze the statistics of proposal regions in Table IV of Faster RCNN baseline and our CFA module, all with Resnet50 as backbone network. We provide the statistics of proposals generated from the first stage of two-stage detectors, *i.e.*, the proposals generated by RPN in Faster RCNN from the test set. We select foreground proposals whose confidence score are in top 300 and fall into the image after NMS from each sample,

TABLE IV: The number of foreground proposals, average score and average size of the foreground proposals generated by RPN, with Faster RCNN baseline and our CFA module. For our methods, to obtain better results, spatial regions with low confidence scores are also included to provide enough information that is needed. Our CFA module extracts informative patterns from the coarser regions to predict more accurate categories and boundaries, *i.e.*, higher mAP.

Methods	# proposals	avg score	avg size	mAP
Faster RCNN	463110	0.057	46559.6	63.85
Ours	463110	0.053	49256.9	66.80

Fig. 7: Coarse attention visualization. The left two columns and the right two columns correspond to the highest and the second highest coarse attention scores of two sample proposals respectively,*i.e*., two expert feature extractors in proposal regions. **Top:** original image; **Mid:** the feature map of channels whose coarse attention scores are top2; **Bottom:** visualized patterns with deconvolution network [43] which correspond to the appearance or contours in original images. Patterns that contain important features are focused by the expert channels, *e.g*., the right part of the rear lamps at the last column and contours of the symmetry light signals at the first column.

obtaining 463110 foreground proposals in total from 1544 test images. We then analyze the average score and average size of them. We can notice that with our CFA module, the average score and average size of foreground proposals will get smaller compared with the scores and sizes in original Faster RCNN. We can see our attention modules implicitly encourage the proposal regions to contain regions of interests even if more noise exists We analyze the reason is that our Coarse-to-Fine Attention mechanisms facilitate extracting informative features and spatial regions in proposals. To obtain better results, spatial regions with low confidence scores are included to provide enough information that is needed to predict the accurate categories and boundaries. As can be seen at Table IV, higher mAP is obtained with the coarser proposal regions, benefited from the proposed attention mechanisms.

Visualization. To validate that our methods can boost the performance by training expert feature extractors, focusing on discriminative patterns with coarse attention and generating accurate spatial attention maps with fine attention, we analyze the extracted features of our coarse attention module. We use the toolbox provided by [43] to visualize where coarse attention pays attention. Some examples are provided in Fig. 7. We visualize what coarse attention has learned using Deconv-net proposed in [44] and get two pairs by choosing the same channel whose attention scores are within top2, predicted by the coarse attention module, as a pair. We can see in Fig. 7 that coarse attention focus on patterns that contain spatial information like the center part of the whole car tile or the right part of the car, which is desired with a set of expert feature extractors optimized by modeling modal feature. In this case, the right tile light is focused by both the two experts. Then fine attention is followed to consider all experts in a voting manner to give accurate spatial attention maps. Discriminative parts, *e.g*.the right tile light in Fig. 7, are thus paid more attention.

Example Instances. In order to compare the detection effects more intuitively, we selected some hard samples to explain. Here, we have carried out occlusion processing on the vehicle's license plate in order to protect the owner's information. The detection results in Fig. 8 show that our method outperform baseline in the following points: (I) The first column (brake) and the second column (normal forward) contain different light signals of the same car with very similar patterns under strong sunlight reflection. Our method can detect this subtle differences with high confidence, and can also detect small objects. (II) Our method can correctly detect target that is with large variance. For example, the uncommon angle of the leftmost vehicle in the third column. (III) The vehicle under low light with only silent part of rear lights in fourth column, adding the light interference of the surrounding buildings. Our method can effectively remove the environmental interference information and correctly detect the brake signal. (IV) In the fifth column, our method preferentially detects the vehicle at positive angle as left turn correctly which means more sensitive to the vehicle light signal of abnormal driving semantics. (V) The last one shows our method can detect the turn-right light signal of the vehicle coming from other lanes under heavy rain.

B. Experimental Analysis of Coarse Attention

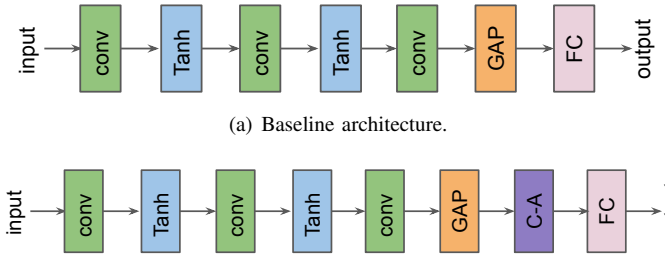
Although our coarse attention is proposed for detection models to extract central features in rough regions and allow low-quality proposals, it can also be inserted to recognition models to extract the informative patterns. In this part, we conduct experiment on CIFAR-10 [45] dataset to validate the effectiveness of our coarse attention mechanism.

Implementation Details. We choose two classes of images with labels automobile and bird in CIFAR-10 to form a binary classification task. All the training images and test images that belong to these two categories are used for training and evaluation. That is, the total number of training samples are 4000 with each of the two classes 2000. We train two models for comparison with the only difference lies in the integration of our coarse-attention module. For better visualization and avoid dimensionality reduction, we use a simple network as baseline network, and compare with the network that also integrated our coarse-attention module, as in Fig. 9.

Analysis. We train both the above two networks for 120 epochs and compare their accuracy performance and feature differences. The accuracy of the baseline network is 93.67% and achieves a 93.79% accuracy with our coarse attention, outperforms the baseline by 0.12%. Although the improvements are marginal and simply enlarging model capacity will probably give higher results, our purpose is to analyze how the informative features are clustered.

Thus, we visualize features of all instances after GAP as in Fig. 10. It gives a better understanding of how informative

Fig. 8: Comparison of detection results. Detection categories ended with 'no' denote that the vehicles are normally moving forward. Here we show some samples from our dataset and compare the performance of baseline method and our CFA method. The baseline method is Faster RCNN with VGG16 as backbone network. The shown challenging scenarios including noisy environment like strong sunlight reflection, mixed light of neighboring vehicles, environment light; as well as variance of target like angle variance. Our method shows superior performance compared with the baseline.



(a) Baseline architecture.

(b) Baseline architecture with our coarse-attention module.

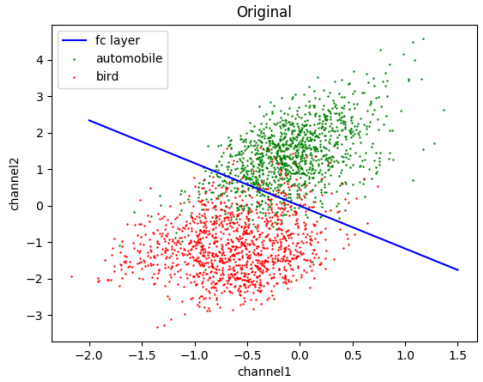
Fig. 9: Baseline network architecture and network with the proposed coarse-attention for binary classification of automobile and bird instances in CIFAR-10.

features are clustered in expert channels, *e.g.*, the two expert channels in Fig. 10. Compared with the original network, the splitting bound is obviously better for classification in our method. As analyzed in Section III-B, central features of these two classes are clustering to the two channels, *i.e.*, the features for these two classes are mainly encoded by channel 0 and 1 respectively, which is desired with proposed coarse attention module. The features with coarse attention are more distinguishable, easing the classification.

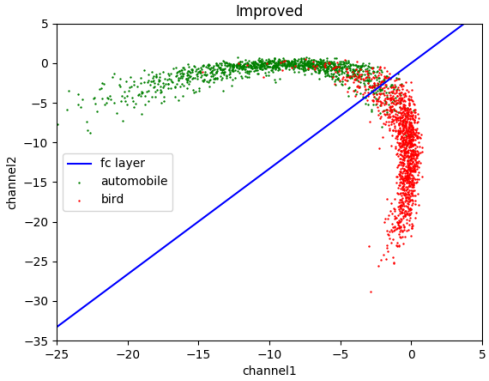
Since we only consider two channels in this binary classification task, both the channels are acting its own role as expert channels to extract the informative central features in the two classes. When the number of channels gets larger for more complicated tasks, *e.g.*, 512 in our detection models, central features of each category will be delivered to a set of expert channels by assigning the attention weights as in Eq. 1. And more informative features that expert channels get, the larger attention weights they are, and iteratively aggregate more accurate central features.

V. Conclusion

In this paper, we found and summarized the difference between luminous object detection and general object detection tasks, and aimed at these detection difficulties like vague boundaries and abstract semantic information, we have introduced the Coarse-to-Fine Attention mechanism for vehicle



(a) Baseline feature visulization with accuracy 93.67%.



(b) Improved network with our coarse-attention module feature visulization with accuracy 93.79%.

Fig. 10: Feature visulization before the FC layer. Informative features of the two classes are clustered to expert channels,*i.e.*, channel 1 and channel 2 extract certain patterns in automobile and bird samples respectively and play the roles as expert channels.

light signal perception task, which can be used as an effective add-on to two-stage detectors while keeping end-to-end training paradigm. We provide reliable theoretical derivation and experimental verification of the proposed attention modules. Besides, we publish a challenging vehicle light signal (VLS) dataset for researches in this field.

For future works, a promising direction is to develop more general algorithms that are not only suitable for self-luminous object detection tasks, but also for other tasks where extracting distinctive features and filtering interfering noise are important, *e.g.*, fine-grained image classification. For engineering deployment in driving systems, the proposed algorithm may work better by taking other detection techniques into consideration. For example, techniques for small object detection are useful when encountering light signals that are far from the source location.

ACKNOWLEDGMENT

This work was supported in part by grants from National Science Foundation of China (61571005), the fundamental research program of Guangdong, China (2020B1515310023), the Science and Technology Research Program of Guangzhou, China (201804010429).

REFERENCES

[1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[2] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[3] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2174–2182.

[4] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 696–706, 2002.

[5] M. Omachi and S. Omachi, "Traffic light detection with color and edge information," in *2009 2nd IEEE International Conference on Computer Science and Information Technology*. IEEE, 2009, pp. 284–287.

[6] J. Müller and K. Dietmayer, "Detecting traffic lights by single shot detection," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 266–273.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[8] R. Girshick, "Fast r-cnn," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[10] P. Liebman and G. Entine, "Sensitive low-light-level microspectrophotometer: detection of photosensitive pigments of retinal cones," *JOSA*, vol. 54, no. 12, pp. 1451–1459, 1964.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[13] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1222–1230.

[14] D.-Y. Chen, Y.-H. Lin, and Y.-J. Peng, "Nighttime brake-light detection by nakagami imaging," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 13, pp. 1627–1637, 12 2012.

[15] J.-G. Wang, L. Zhou, Y. Pan, S. Lee, Z. Song, B. Han, and V. Saputra, "Appearance-based brake-lights recognition using deep learning and vehicle detection," 06 2016, pp. 815–820.

[16] Z. Zhang, X. Zhou, S. Chan, S. Chen, and H. Liu, "Faster r-cnn for small traffic sign detection," in *CCF Chinese Conference on Computer Vision*. Springer, 2017, pp. 155–165.

[17] T.-W. Yeh, S.-Y. Lin, H.-Y. Lin, S.-W. Chan, A. Lin, and Y.-Y. Lin, "Traffic light detection using convolutional neural networks and lidar data," 12 2019, pp. 1–2.

[18] A. N. Aneesh, L. Shine, R. Pradeep, and V. Sajith, "Real-time traffic light detection and recognition based on deep retinanet for self driving cars," in *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, vol. 1, 2019, pp. 1554–1557.

[19] P. Cheng, W. Liu, Y. Zhang, and H. Ma, *LOCO: Local Context Based Faster R-CNN for Small Traffic Sign Detection*, 01 2018, pp. 329–341.

[20] J. Li and Z. Wang, "Real-time traffic sign recognition based on efficient cnns in the wild," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, pp. 1–10, 06 2018.

[21] R. Kulkarni, S. Dhavalikar, and S. Bangar, "Traffic light detection and recognition for self driving cars using deep learning," 08 2018, pp. 1–4.

[22] W. Pan, Y. Chen, and B. Liu, "Traffic light detection for self-driving vehicles based on deep learning," 12 2019, pp. 63–67.

[23] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[24] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Computer Vision and Image Understanding*, vol. 178, 11 2018.

[25] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," 06 2018, pp. 3291–3300.

[26] J. Guo, R. You, and L. Huang, "Mixed vertical-and-horizontal-text traffic sign detection and recognition for street-level scene," *IEEE Access*, vol. PP, pp. 1–1, 04 2020.

[27] X. Wu, R. Hu, and Y. Bao, "Parallelism optimized architecture on fpga for real-time traffic light detection," *IEEE Access*, vol. 7, pp. 178 167–178 176, 2019.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[29] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[31] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[32] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[33] S. Khandelwal and L. Sigal, "Attentionrnn: A structured spatial attention mechanism," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[34] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3047–3055.

[35] F. Lyu, F. Hu, V. S. Sheng, Z. Wu, Q. Fu, and B. Fu, "Coarse to fine: Multi-label image classification with global/local attention," in *2018 IEEE International Smart Cities Conference (ISC2)*. IEEE, 2018, pp. 1–7.

[36] G. J. McLachlan and K. E. Basford, *Mixture models: Inference and applications to clustering*, vol. 38.

[37] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[38] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[39] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[40] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[43] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015.
[44] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.
[45] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.

Shenao Zhang is an undergraduate student majoring in Electrical and Information Engineering at South China University of Technology, Guangzhou China. This work is done when Shenao was a summer intern at Columbia University. His research interests include computer vision, deep learning, and machine learning.

Bo Wu Bo Wu is currently a Researcher with MIT-IBM Watson AI Lab in MA, USA. Before that, he was a Research Scientist at Columbia University. He was the Visiting Scholar in Microsoft Research Asia (MSRA) and Academia Sinica. He received the Ph.D. degree in computer science from the Chinese Academy of Sciences (ICT, CAS), Beijing, China. His current research interests are deep learning, social computing, computer vision, and natural language learning, which focus on visual, language, or user behavior understanding, forecasting, and reasoning. He served as Area Chair for ACM Multimedia, Senior Program Committee Member for IJCAI, and organized global SMP Challenges 2017-2020. He received the following research and competition awards: ACL 2020 Best Demo Paper Award, ACM Turing 50th Student Scholarship, Schlumberger Ph.D. Award, DARPA Event Engine Competition (Top1), Alibaba Global Vision Challenge (Top3/4K), ICIP2020 Prediction Challenge (Top1), ST-VQA Challenge (Top1).

Xin He graduated from the School of Mathematics, South China University of Technology with a master's degree in computational mathematics.

Lei Kuang received the bachelor's degree in internet of things engineering from Beijing University of Posts and Telecommunications, Beijing China, in 2019. He is currently pursuing the master's degree in computer science in Columbia University in the City of New York, NY, USA. He will join NVIDIA as a solution architect upon graduation. His research focuses on computer vision, high performance computing and robotics.

Delu Zeng received his Ph.D. degree in electronic and information engineering from South China University of Technology, China, in 2009. He is now a full professor in the School of Mathematics in South China University of Technology, China. He has been the visiting scholar of Columbia University, University of Oulu, University of Waterloo. He has been focusing his research in applied mathematics and its interdisciplinary applications. His research interests include numerical calculations, applications of partial differential equations, optimizations, machine learning and their applications in image processing, and data analysis.