Received November 14, 2017, accepted December 12, 2017, date of publication December 27, 2017,
date of current version February 14, 2018.

Digital Object Identifier 10.1109/ACCESS.2017.2787155

Cascaded Regional Spatio-Temporal Feature-Routing Networks for Video Object Detection

HUI SHUAI, QINGSHAN LIU, (Senior Member, IEEE), KAIHUA ZHANG, JING YANG, AND JIANKANG DENG
School of Information and Control, Nanjing University of Information Science and Technology, Nanjing 210044, China

Corresponding author: Hui Shuai (huishuai13@gmail.com)

The work of Q. Liu was supported by the Natural Science Foundation of China under Grant 61532009. The work of K. Zhang was supported in part by the Natural Science Foundation of Jiangsu Province of China under Grant BK20170040 and in part by the Six Talent Peaks Project in Jiangsu Province under Grant R2017L07.

ABSTRACT This paper presents a cascaded regional spatiotemporal feature-routing networks for video object detection. Region proposal networks in faster region-based convolutional neural network (CNN) generate spatial proposals, whereas neglecting the temporal property of the videos. We incorporate the correlation filter tracking on the convolutional feature maps to explore an efficient and effective spatiotemporal region proposal generation method. To gradually refine the bounding boxes of proposals, three region classification and regression networks are cascaded. Feature maps from different layers in CNNs extract hierarchical information of the input, so we propose a router function which selects feature maps according to the scale of proposals. In addition, object co-occurrence inference is exploited to suppress conflicting false positives, which leads to a semantically coherent interpretation on the video. Extensive experiments on the Pascal VOC 2007 dataset and the ImageNet VID dataset show that the proposed method achieves the state-of-the-art performance for detecting unconstrained objects in cluttered scenes.

INDEX TERMS Video object detection, correlation filter tracking, router-function, regression networks, co-occurrence inference.

I. INTRODUCTION

Object detection [1], [2] is a fundamental computer vision task that aims at automatically localizing objects from images, which has great potential in multimedia applications [3], [4]. Early methods can effectively and efficiently detect certain limited object categories (e.g., face [5] and person [6]) by sliding windows or cascaded classifiers, but they cannot work well on multiple categories. Recently, object detection on multiple categories has been significantly improved due to the advances of deep convolutional neural networks (CNNs) [7], of which one particularly successful paradigm is Region based CNN (R-CNN) [8] that composed of sequential region proposal [9] and region classification [7] module. R-CNN transforms object detection into an object classification problem, and fine-tunes a pre-trained ImageNet [10] classification network with end-to-end training for region classification.

Fast R-CNN [11] incorporates a simplified spatial pyramid pooling layer into the R-CNN which can handle input images of random size. In addition, a multi-task loss for classification and regression is employed, which makes training a single stage. Faster R-CNN [12] extends R-CNN by introducing a region proposal network that shares full-image convolutional features with region classification network. Both networks share the convolutional feature maps and they are trained by a simple alternating optimization. Moreover, RPN and classification network are essentially the same. They are just two different architectures of the classifier based on convolutional feature maps. Therefore, with continual appearance of new structure such as GoogLeNet [13], ResNet [14], performance of Faster R-CNN is gradually improved. The revised version of Faster R-CNN maintain state-of-the-art results in object detection on images [15].

Although numerous works [8], [11], [12], [16] have been proposed for object detection on images, there are few works for video object detection. Video object detection is much more challenging due to the large appearance changes caused by occlusion, deformation, abrupt motion,

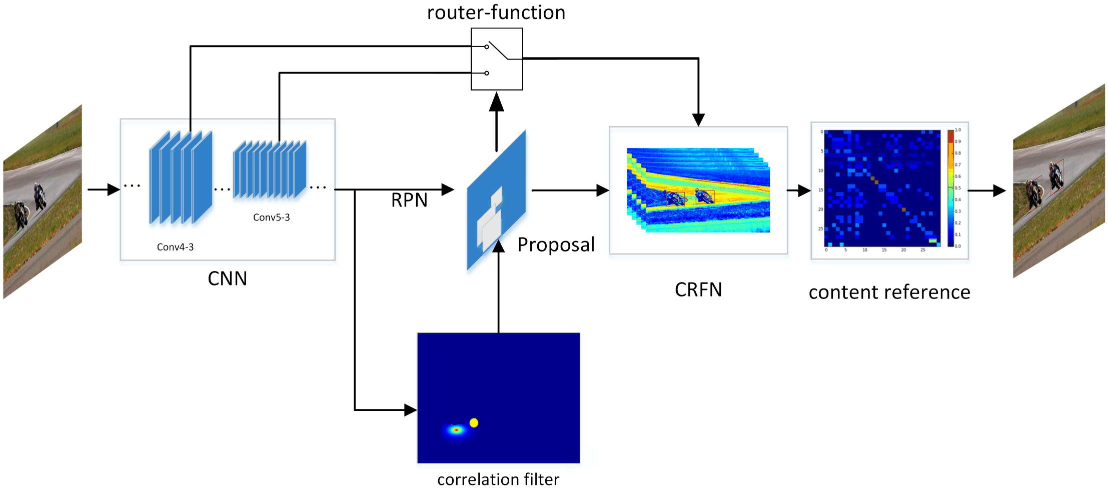


FIGURE 1. Overview of the proposed cascaded regional spatio-temporal feature-routing network(CRSFN). RPN and correlation filter jointly generate spatio-temporal regional proposals. Router-function selection feature maps for subsequent cascaded regional classification and regression networks(CRCR) according to the scale of proposals. CRCR gradually refine the bounding boxes of proposals. Ultimately, content inference is utilized to suppress conflicting false positives.

illumination variation, and background clutter, etc. Since Faster R-CNN [12], [17] has achieved good results in object detection on images, it is thus of great interest to understand how to exploit this framework for robust video object detection. Faster R-CNN benefits from the advances in image classification [17], but it confronts two problems when applied to object detection in videos. The first one is that region proposals on each frame are not always effective due to extremely large appearance variations, which may cause much loss on the recall rate. The second one is that the independent region classification cannot incorporate spatio-temporal video context, which causes the low scores of some obscure regions.

To alleviate these problems, a possible solution is to decrease the target variations by the ‘‘divide and conquer’’ scheme, which trains multiple classifiers to handle the variations of each sub-problem [18]. Meanwhile, the context information on videos can be utilized to propagate the true positives while suppressing the false ones. We introduce a cascaded regional feature-routing networks(CRFN) for robust object detection. CRFN stacks multiple regional feature-routing classification and regression networks on the top of convolutional feature maps, which can gradually refine the bounding box of the target object, and hence improve the detection performance. Taking account of temporal property of video object, we exploit the correlation filter tracking to generate temporal region proposals from the high-confident detection results and propose a cascaded regional spatio-temporal feature-routing network(CRSFN). In addition, co-occurrence inference between the dominant class and others is used for effectively suppressing the false positives. We utilize the context information via an efficient Look-Up-Table method, which can effectively suppress the conflicting false positives and guide the detector to produce a semantically coherent interpretation on the video. Figure 1 illustrates the overview of the proposed method.

Extensive experiments on the Pascal VOC 2007 dataset and the ImageNet VID dataset show that the proposed method achieves the state-of-the-art performance for detecting unconstrained objects in cluttered scenes.

II. RELATED WORK

The proposed method is inspired by the following related works:

A. OBJECT DETECTION

Most recently proposed popular object detection approaches are based on deep CNNs [8], [11], [12], [16], [19], [20]. Detection systems such as R-CNN [8], SPP-net [16] and Fast R-CNN [11] can be divided into two steps: salient object proposal generation and region proposal classification. Networks on the convolutional feature maps(NoCs) [19] extracts features with a fixed pre-trained deep CNN and explores different networks on the convolutional feature maps as object classifiers. Extensive experiments show that a well-designed NoCs on the top of convolutional network with maxout [21] performs extremely well when trained from a random initialization, which indicates that there are still significant gains to design new classification networks on top of a fixed convolutional network. MultiBox [22] generates region proposals for region classification and simultaneously predicts multiple boxes. YOLO [20] predicts bounding boxes and class probabilities directly from the whole image domain in one evaluation. Faster R-CNN [12] unifies region proposal generation and classifier training on the shared convolutional feature maps with a multi-task loss function. In this paper, we first train the Faster R-CNN and then fix the convolutional layers. After that, three steps of feature-routing NoC architectures with a multi-task loss function are fine-tuned on the convolutional feature maps.

TABLE 1. Detection results of Fast R-CNN and Faster R-CNN on the PASCAL VOC 07 test set using ZF [35] and VGG [36] models. The training set is PASCAL VOC 07+12 trainval. If we decrease the threshold of IoU, the mAPs increase obviously.

IoU	0.25	0.30	0.40	**0.50**	0.60	0.70	0.75	0.80	0.90
Fast[11]_ZF	67.5	66.7	63.9	**58.7**	49.6	36.2	28.1	17.8	4.9
Fater[12]_ZF	69.5	68.3	64.6	**59.9**	50.1	34.5	25.4	15.2	2.6
Fast[11]_VGG	77.0	76.3	74.2	**70.1**	62.7	50.5	41.8	31.6	8.4
Faster[12]_VGG	79.2	78.5	75.6	**73.2**	65.3	51.8	41.3	29.7	6.7

B. SHAPE OR POSE-INDEXED FEATURE AND CASCADED REGRESSION

Shape or pose-indexed features [23] can gradually refine shapes or poses via cascaded regression. This method has been successfully applied for face alignment [24]. Similarly, features from different layers extract hierarchical information of the input. These features can be used to refine the location of regions in different scales. In this paper, we iteratively apply the location-indexed convolutional features routed by router-function to cascaded classification and regression of region proposals via multi-task learning [11].

C. CORRELATION FILTER TRACKING

In visual object tracking, correlation filter tracking have attracted considerable attention owning to its remarkable computational efficiency with Fast Fourier Transforms (FFT) [25]–[28]. The correlation filter tracking regress all the shifted versions of input features to a Gaussian function and update their weights in an online manner. Ma *et al.* [27] develop a hierarchical correlation filter based visual tracking method over a set of multi-dimensional convolutional feature maps, which achieves the state-of-the-art results on the Object tracking benchmark [29]. In this work, we propose a correlation filter tracking based region proposal generation method for video object detection, which integrates the class-specified priors into the correlation filter framework with the convolutional layers being fine-tuned by the object detection task.

D. CONTEXT MODEL

Recently proposed object detection methods [8], [11], [12], [16] have moved focus to varying categories with large appearance variations, such as two hundred categories in ImageNet [30] and eighty categories in COCO [31]. There has been a growing interest in exploiting contextual information in addition to local features to detect multiple object categories in an image. A context model [32], [33] is able to rule out some conflicting combinations and guide detectors to produce a semantically coherent interpretation of an image. Choi *et al.* [34] propose a tree-based context model which incorporates global image features, the dependencies between object categories, and the outputs of the local detectors into a probabilistic framework. In this work, we transform the global video context information into the constraints between the dominant class and other classes in the video, in which the confliction between different classes is built on an efficient Look-Up-Table, which helps to suppress incoherent false positives effectively.

III. METHODOLOGY

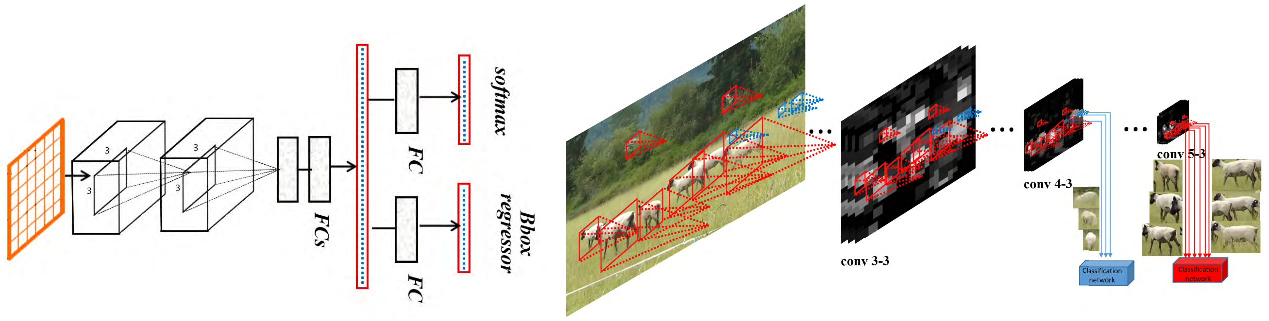
A. CASCADED REGIONAL FEATURE-ROUTING NETWORKS

The region classification step is usually consisted of two important components: a feature extractor and a classifier. R-CNN can be treated as a convolutional feature extractor followed by a multi-layer perceptron (MLP) classifier. Motivated by this, we first train the Faster R-CNN to yield a set of fixed convolutional layers, and then we fine-tune three steps of cascaded regional feature-routing classification and regression networks via multi-task learning for region classification, which are combined to improve the performance of region classification.

1) CASCADED REGION CLASSIFICATION AND REGRESSION

In [8], with bounding box regression for post-processing, R-CNN obtains a performance gain of 4.2%, which can be attributed to the fact that the region regression with the location-indexed features can provide a more accurate location. Specifically, as shown by Table 1, if we decrease the threshold of the intersection-over-union (IoU), the mAPs of the Fast R-CNN and Faster R-CNN both increase obviously. Here, the standard threshold of IoU is 0.5, and there are some false positives generated from inexact localizations

Since the location-indexed features are able to provide a more accurate location [23], we train several cascaded region regressors on the location-indexed convolutional feature maps to gradually refine the detection results. As shown in Figure 2(a), we follow [11], [19] to transform the convolutional features to a constant length (7 × 7 × 512) by adaptive pooling. Motivated by the part models [37], [38], We also add two additional 3 × 3 convolutional layers [19] before the three-layer perceptron with a multi-task loss function [11]. For training this model, we generate a set of training data with different IoU lower bounds for each network: The negative examples of each step are the region proposals whose IoU ratios with the ground truth are less than 0.3 while the positive ones of the first step is the region proposals whose IoU ratios with the ground truth are more than 0.4. Then, the first region classification and regression network is trained to update all of the region proposals. In the second step, the IoU threshold of the positive example is set to 0.5, and 0.6 in the last step. Since the IoU ratio of a successful detection is 0.5, there is no need to add more steps.



(a) (b)

FIGURE 2. Cascaded regional feature-routing classification and regression. (a) The convolutional features are adaptively pooled into 7 × 7, followed by two additional 3 × 3 convolutional layers and a three-layer perceptron with multi-task loss. (b) The sizes of sheep exhibit large variations, so we use the convolutional features from conv5-3 for the large proposal regions (> 56 pixels), and conv4-3 for the small proposal regions (< 56 pixels). (a) Region classification and regression network. (b) Multi-scale feature selection.

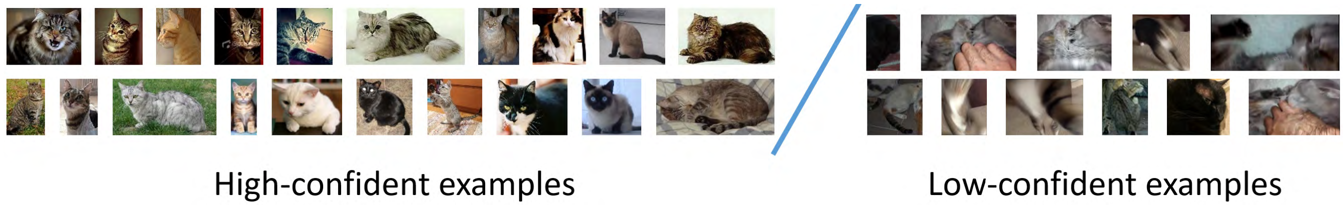


FIGURE 3. The class of "domestic cat" exhibits large appearance variations caused by occlusion, deformation, abrupt motion, illumination variation, and background clutter,etc. We separate the ground truth into two groups: the high-confident (score > 0.6) positive examples and the low-confident (score < 0.6) positive examples.

2) FEATURE ROUTING

As shown in Figure 2(b), we adaptively use the convolutional features from different layers to classify and regress the region proposals. These features are selected by the router-function according to the scale of regions. In other words, local and contextual visual patterns are selectively transmitted to cascaded classify and regression module. The route of message is controlled by the router-function. Specifically, we fix the convolutional layers of VGG-16 fine-tuned by Faster R-CNN and add two independently cascaded region classification and regression networks to conv 4-3 and conv 5-3, respectively. In inference, the route is determined by the router-function according to the scale of regions. When the proposal region in the original image is larger than 56 pixels, the corresponding conv5-3 features are selected. Otherwise, the conv4-3 features are used. The router-function is formulated as:

$$f^* = \begin{cases} f_{conv4-3} & p \leq threshold \\ f_{conv5-3} & p > threshold \end{cases} \tag{1}$$

f^* is the feature map selected to refine the bounding box, p is the number of pixels in the proposal, the threshold we use is 56 pixels. This process is handcrafted based on prior knowledge and it could potentially be replaced by learning method in future work.

3) TRAINING DATA RE-WEIGHTING

The training data from the videos exhibit large appearance variations caused by occlusion, deformation, abrupt motion, illumination variation, and background clutter, etc. As shown in Figure 3, there are large appearance variations within the training data of "domestic cat". If we directly use all of the training data, the detection accuracy will deteriorate. With the pre-trained Faster R-CNN, we can get the score of each ground-truth region. Then, we divide the training data into two group: the high-confident (score $>$ 0.6) positive examples and the low-confident (score $<$ 0.6) positive examples. After that, two CRFN networks are independently trained on the high-confident positive examples and the whole training data. During testing, the scores predicted by these two models are averaged as the final score for each region proposal. The model trained on the high-confident positive examples only gives high scores to the salient examples, which is helpful for keeping a high precision, whereas the model trained on the whole training data tends to give high scores to the hard positive examples and false positives, which is helpful to improve the recall.

B. SPATIO-TEMPORAL REGION PROPOSAL

Region proposal networks in Faster R-CNN generate spatial proposals while neglecting the temporal property of the videos. In order to incorporate temporal property into Faster R-CNN architecture, we propose a correlation filter based temporal region proposal generation method to generate spatio-temporal region proposals. The correlation filter learn a generative model and estimate the translation of the target objects by searching for the maximum response on the correlation map. The initial locations are the detection results with high-confidences (score $>$ 0.6), and then the temporal region proposals are generated near the initial regions forward and backward through the video frames.

Multi-channel feature x of size $W \times H \times D$ is cropped from the convolutional feature maps, where W and H indicate the width and height of the region (2 times of the target object size), and D indicates the number of channels. We only utilize the correlation filter to estimate the translation of the target object, and the scale of the target object is not changed. The scope of the search space covers the whole x, and each shifted sample $x_{i,j}$, $(i,j) \in \{0, 1, \ldots, W-1\} \times \{0, 1, \ldots, H-1\}$ has a corresponding Gaussian distribution function label $y(i,j) = e^{-\frac{(i-W/2)^2+(j-H/2)^2}{2\sigma^2}}$, where the kernel width σ is set to be 0.1. The correlation filter r is learned by solving the following minimization problem:

$$r^* = \arg\min_{r} \sum_{i,j}^{W,H} \left\| r \cdot x_{i,j} - y(i,j) \right\|_2^2 + \lambda \left\| r \right\|_2^2, \qquad (2)$$

where $r \cdot x_{i,j} = \sum_{k=1}^{D} r_{i,j,k}^{T} x_{i,j,k}$, λ is a regularization parameter. This minimization problem can be solved in each individual feature channel using Fast Fourier Transform (FFT), and the learned filter in the frequency domain on the k-th ($k \in \{1, \ldots, D\}$) channel can be calculated as

$$R^k = \frac{Y \odot \overline{X}^k}{\sum_{k=1}^{D} X^k \odot \overline{X}^k + \lambda}, \qquad (3)$$

where Y is the Fourier transform of y, $\overline{X}$ is the complex conjugation of the Fourier transform of x, and the operator $\odot$ is the element-wise product. Given a convolutional feature crop z of size $W \times H \times D$ from the next frame, the response map is computed by an inverse FFT transform $\mathcal{F}^{-1}(\sum\limits_{k=1}^{D} R^k \odot \overline{Z}^k)$, where $\mathcal{F}^{-1}$ denotes the inverse FFT. The translation of the target object can be estimated by searching for the position of maximum value on the correlation response map. Since the temporal region proposal is initialized with a class-specific region, we can incorporate the class-specific prior to update the proposal generation model for the current video. The detection results with high-confidences (score $>$ 0.6) are selected from the whole video, and we only keep the most similar instance for each frame to avoid the confusions between different instances of the same class. These instances are allocated normalization weights according to their temporal distances to the current initial region. Moreover, the numerator and denominator of the correlation filter R^k are updated separately by each instance. Afterwards, the temporal region proposal is conducted forward and backward to the video, where the correlation filter R^k is updated via a moving average:

$$\begin{aligned} A_t^k &= 0.3A_0 + (0.7-\mu)\, A_{t-1}^k + \mu Y \odot \overline{X}_t^k \\ B_t^k &= 0.3B_0 + (0.7-\mu)\, B_{t-1}^k + \mu \sum\nolimits_{k=1}^{D} X_t^k \odot \overline{X}_t^k \\ R_t^k &= \frac{A_t^k}{B_t^k + \lambda}, \end{aligned} \qquad (4)$$

where A_0 and B_0 are the class-specific priors learned from the high-confidence detection results, and μ is the temporal update rate. The class-specific priors keep the generative properties of our method stable while the online update strategy enables the correlation filter well adapt to the appearance changes.

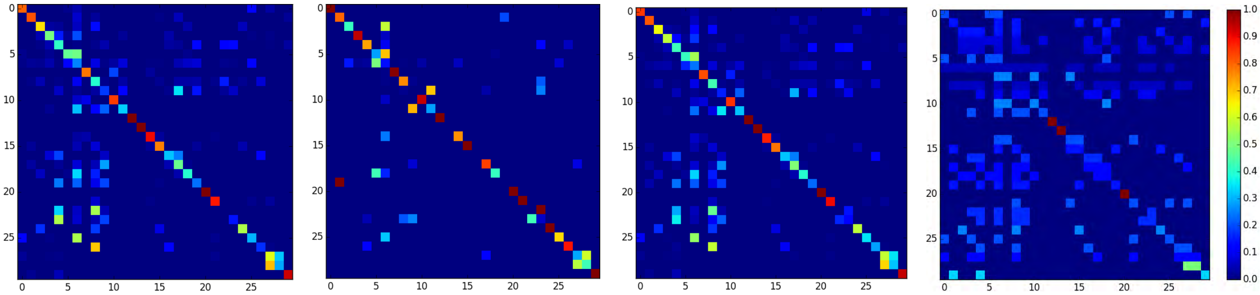
The size of the target object changes on different frames, however, the correlation filter should be the same size with the search window x. Therefore, we resize the convolutional feature maps of each frame to make the target objects with the same scale (the longer side is 28). Moreover, to remove the boundary discontinuities, the cropped convolutional features are weighted by a cosine window. In addition, since the convolutional layers are location-sensitive, they are beneficial to enhance objection localization. We use the conv5-3 features to generate the temporal region proposals.

C. CONTEXT INFERENCE

The region based object detection methods often focus on locally identifying a particular object region. Since each proposal region is processed independently from others, the outcome of detection may be semantically incorrect. To deal with this problem, we exploit the contextual information such as the global features of a video (it is a grassland scene) and dependencies among the object categories (e.g., sheep and cattle often co-occur, and lions and whales rarely co-occur) besides the local convolutional features. With a semantically coherent inference, some false positives on the videos can be effectively removed.

The scene context is an effective clue for object detection. Some particular classes have strong correlations with their environments, such as the correlation between sheep and grassland. Since there is no scene annotation on the ImageNet VID dataset, we transform the correlation between objects and their environments in an indirect way. We first run the Faster R-CNN on the whole video, and select the class with the most high-confident (score $>$ 0.6) detection results and a wide distribution on the whole video as the dominant class of this video. Then, the correlations between the objects and their scene context are transformed into the co-occurrence of object pairs.

We select all of the multi-instance frames from the ImageNet VID training set and validation set, among which there are 336,219 multi-instance frames in the training set and 53,192 multi-instance frames in the validation set. Figure 4 shows all of the co-occurrence relationships of the instance pairs. We can observe that most of the co-occurrence happens within the same class except for the class of "snake (Class ID: 23)", which rarely co-occurs with each other within one frame. We think this is caused by the dataset bias. Furthermore, as shown in Figure 4(c), the co-occurrence matrix is relatively sparse, and some classes rarely co-occur in one frame. Thus, we can use this co-occurrence information to effectively suppress the incoherent false positives. Figure 4(d) is generated from Figure 4(c) by binarizing the co-occurrence matrix, among which three classes, i.e., "giant



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

FIGURE 4. Object co-occurrence distributions on the ImageNet VID dataset. The intra-class co-occurrence is very common, however, the inter-class co-occurrence does not always exist. The class ID ranges from 0 – 29 with the same sequence of definition provided by the ImageNet toolkit. (d) is the binary co-occurrence matrix on the ImageNet VID dataset. Three classes, i.e., "giant panda", "hamster" and "red panda," show high independence. The sparsity of the co-occurrence matrix is 699/900. (a) Train. (b) Validation. (c) Train and Validation. (d) Co-occurrence.

panda", "hamster" and "red panda", show high independence, and the co-occurrence only exists within class. There are only 201 co-occurrence relationships within the ImageNet VID dataset, and we use this binary co-occurrence table to suppress false positives. After the dominant class of the video is determined, all of the conflicting object classes are selected from the binary co-occurrence table. There may be some bias between the training set and the test set, so we reduce the scores of the conflicting detection results by half instead of just removing them from the final results.

IV. EXPERIMENTS

A. EXPERIMENTAL SETTING

1) IMAGE DATASET

We validate the proposed CRFN method on the widely used PASCAL VOC 2007 object detection benchmark [39]. It covers 20 object categories, in which the test set contains 5,000 images. Following [12], we use an augmented set of 1,6000 images as our training data, which consists the VOC 2007 training and validation images and VOC 2012 training and validation images.

2) VIDEO DATASET

We further investigate our CRSFN on the ImageNet VID dataset, which contains 3,862 videos for training, 555 videos for validation, and 937 videos for testing. Besides, it contains 30 basic-level categories which are selected from the 200 categories of ImageNet DET dataset, which are carefully chosen considering different factors such as movement type, level of video clutterness, and average number of object instance. The data distributions of these two datasets are quite different. Specifically, the instances of each class from the video dataset exhibit large appearance variation, but the instance number is limited, and hence the model cannot benefit from the repeated instance examples. However, the training data from the image dataset exhibit large instance diversity of each class. Therefore, these two datasets are complementary with each other, which can be used together to boost the performance of our method.

3) DEEP MODEL

As a common practice [8], [11], [12], [16], we use the deep CNN model pre-trained on the 1000-class ImageNet dataset [10]. Specifically, we investigate the VGG 16 model [36], which has 13 convolutional layers and three fc layers.

4) TRAINING DETAILS

We first train the Faster R-CNN on the ImageNet DET subset, and then fix the convolutional layers and train the CRFN model using data from the ImageNet VID dataset. Moreover, to setup the Faster R-CNN baseline, we change three parameters. First, four anchor scales are used in this paper with areas of 64^2, 128^2, 256^2, and 512^2 pixels on the original image. Second, the threshold of intersection-over-union (IoU) overlap for the positive examples is set to $min(0.5, w*h/(w+10)*(h+10))$. Third, the lower threshold of the IoU overlap for the negative examples in the region classification step is set to be 0 instead of 0.1. The former two are designed for objects with low resolution, and the last one is designed for negative example mining during the "image-centric" training. All the training data are resized to a short side of $s = 600$ pixels and a long side less than $s = 1000$ pixels. We use a learning rate of 0.001 for 240,000 mini-batches, and 0.0001 for the next 120,000 mini-batches on the 122,000 images from the ImageNet DET subset. We also use a momentum of 0.9 and a weight decay of 0.0005 [7].

After finishing training the Faster R-CNN on the ImageNet DET subset, we fix the convolutional layers and train the CRFN model in a "region-centric" way [16]. Specifically, we generate a fixed-resolution feature map region via a region pooling operation that has a fixed output resolution. Formally, we define a desired fixed output spatial resolution 7×7, which is the output spatial size of the last pooling layer in the pre-trained model. For an arbitrary feature map region of size $w \times h$, we produce the 7×7 output by max pooling in spatial bins of the size $\frac{w}{7} \times \frac{h}{7}$. Moreover, the additional two convolutional filter have a spatial size of 3×3 and a padding of 1, so the output spatial resolution is unchanged (7×7).

TABLE 2. Detection results for PASCAL VOC 2007 test set using the VGG-16 model [36]. Here "bb" denotes bounding box regression [8].

method	train set	mAP	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
R-CNN[8]	07	62.2	71.6	73.5	58.1	42.2	39.4	70.7	76.0	74.5	38.7	71.0	56.9	74.5	67.9	69.6	59.3	35.7	62.1	64.0	66.5	71.2
R-CNN[8],bb	07	66.0	73.4	77.0	63.4	45.4	44.6	75.1	78.1	79.8	40.5	73.7	62.2	79.4	78.1	73.1	64.2	35.6	66.8	67.2	70.4	71.1
SPP[16]	07	60.4	69.4	70.4	58.8	47.3	39.2	72.2	70.4	71.5	38.1	70.3	52.8	69.3	69.8	71.1	51.4	33.5	58.5	52.6	67.1	73.8
SPP[16]	07+12	64.6	70.8	78.1	65.6	51.0	43.4	74.4	71.2	76.6	43.6	73.8	55.0	76.9	73.8	73.1	55.2	33.7	65.3	65.0	69.4	75.6
NoC[19]	07+12	68.8	74.6	77.7	68.5	53.3	45.5	78.0	75.5	82.1	47.9	77.2	63.0	81.1	75.9	75.1	61.6	41.7	72.9	73.3	73.8	77.7
NoC[19],bb	07+12	71.6	75.4	79.4	71.9	57.4	50.9	83.0	77.4	85.9	51.3	77.5	65.9	82.8	82.7	77.7	65.2	45.6	70.2	75.7	76.8	78.6
Fast[11]	07	66.9	74.5	78.3	69.2	53.2	36.6	77.3	78.2	82.0	40.7	72.7	67.9	79.6	79.2	73.0	69.0	30.1	65.4	70.2	75.8	65.8
Fast[11]	07+12	70.0	77.0	78.1	69.3	59.4	38.3	81.6	78.6	86.7	42.8	78.8	68.9	84.7	82.0	76.6	69.9	31.8	70.1	74.8	80.4	70.4
Faster[12]	07	69.9	70.0	80.6	70.1	57.3	49.9	78.2	80.4	82.0	52.2	75.3	67.2	80.3	79.8	75.0	76.3	39.1	68.3	67.3	81.1	67.6
Faster[12]	07+12	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
Multi-scale	07+12	73.8	77.0	80.7	73.5	66.1	55.9	81.3	82.0	86.2	53.0	81.2	65.9	86.2	84.0	78.8	77.1	46.2	75.8	68.0	83.3	74.2
CRCR	07+12	74.6	78.1	80.6	72.7	66.8	54.1	83.9	84.6	87.0	54.6	83.0	66.6	85.5	86.7	79.1	77.3	42.3	76.2	77.1	84.0	72.9
CRFN	07+12	75.6	79.0	81.4	73.9	66.5	56.4	84.8	85.3	86.6	55.1	84.3	66.7	86.2	87.1	79.9	78.9	47.0	76.5	77.9	83.9	75.2

These two convolutional layers are initialized by the identity filter. Finally, the following fully-connected (fc) layers are initialized by the corresponding fc layers in the region classification model. Thus, the whole fine-tuning procedure becomes equivalent to the forth step of Faster R-CNN. Except the training example distribution, the cascaded networks are trained similarly. We train 2M mini-batches using the learning rate 1e-4, and then M mini-batches using 1e-5, where M is the number of images in the training dataset. During training, the multi-scale feature selection is determined by the region proposal size. We resize the image such that $\min(w, h) = s \in \mathcal{S} = \{200, 300, 400, 600\}$, and compute the feature maps for each scale. Especially if the size of object is less than 10 pixels after down-sampling, it is neglected during training.

We use the training data re-weighting technique in this paper, which improves the diversity of the model ensemble. The CRFN networks are trained independently by these two group of data from III-A.3.

B. EXPERIMENTS ON IMAGES

As shown in Table 2, we compare the proposed CRFN network with the previously leading methods with the VGG-16 model [36]. R-CNN fine-tunes all the layers on the VOC 07 trainval dataset, and outperforms the SPPnet by 1.8%, which suggests that fine-tuning all the layers improves the performance compared with using a fixed pre-trained image classification model. Moreover, compared with the SPPnet, NoC improves performance by 4.2%, which indicates that we can boost the performance by designing a complex classification network on the fixed convolutional feature maps. In addition, the bounding box regression improves the mAP of 3.8% for R-CNN and 2.8% for NoC, which demonstrates that the region regression step is able to improve the localization accuracy. Fast R-CNN equips with all the experiences mentioned above, in which the convolutional layers are fine-tuned and the region regression step is incorporated with a multi-task learning strategy. Faster R-CNN further improves the performance by alternating optimization between the tasks of region proposal and region classification. The multi-scale region classification and regression network improves the mAP by 0.6%, and the cascaded region classification and regression network improves the mAP by 1.4%. The proposed CRFN boosts the performance to 75.6% by six additional networks (two scale × three cascade).

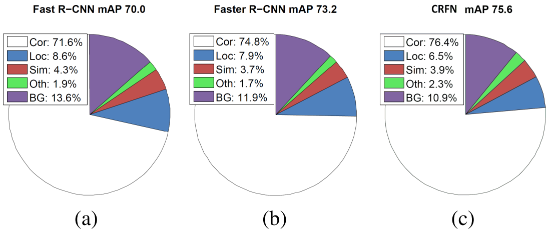


FIGURE 5. Error Analysis: Fast R-CNN, Faster R-CNN and CRFN. These charts show the percentage of localization and background errors in the top *N* detections for various categories (*N* is the object number in that category). (a) Fast R-CNN. (b) Faster R-CNN. (c) CRFN.

To better understand the effect of the CRFN regression network, we use the diagnosing tool of [40] to analyze the top-ranked false-positive predictions. The false positives due to poor localization are denoted as "Loc" while the false positives due to object recognition error consist of "Sim" (confusion with a similar category), "Oth" (confusion with a dissimilar category), "BG" (confusion with on background). As shown by Figure 5, the positive effect of the proposed CRFN is to considerably reduce the localization errors, *e.g.*, reducing from 7.9% to 6.5% compared to Faster R-CNN. Better region localization helps to improve the region classification accuracy, which further enhances the overall detection performance.

C. EXPERIMENTS ON VIDEOS

1) EFFECT OF THE PROPOSED SRP

We first validate the proposed spatio-temporal region proposal (SRP) method on the test set of the ImageNet VID dataset. The baseline is 300 proposals predicted by the Region Proposal Network (RPN), which is trained on the ImageNet DET subset. In Figure 6, we compute the recall of proposals at different IoU ratios with annotation boxes. Although the Recall to the IoU metric is just loosely related to the ultimate detection accuracy, we can use this metric to diagnose the proposal method due to the fact that if an object is missed in the object proposal step, the detection system would definitely miss the object. To avoid generating too many redundancy

TABLE 3. Detection results on the ImageNet VID validation set. "+" denotes that these methods are incrementally incorporated on the faster R-CNN baseline.

method	Training data	mAP on the VID validation set
Faster R-CNN	DET subset	46.8
Faster R-CNN	VID	52.4
Faster R-CNN	DET subset+VID	54.6
+CRCR	DET subset+VID	59.1
+CRFN	DET subset+VID	61.3
+SRP	DET subset+VID	65.9
+Context Inference	DET subset+VID	69.2
+Traning data re-weighting	DET subset+VID	71.8

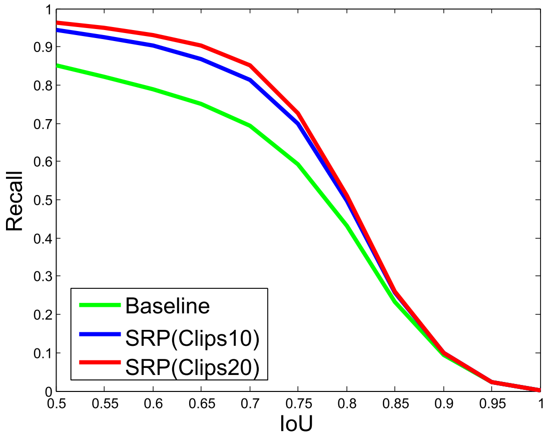


FIGURE 6. Temporal region proposal (TRP) improves the region proposal on the ImageNet VID dataset. The baseline method is Faster R-CNN trained on the ImageNet DET subset. "Clip 10" and "Clip 20" mean that the video is divided into ten or twenty clips. We only take one frame as the initial for correlation filter.

temporal region proposals, we divide the whole video into N clips. In each clip, we only choose one frame with the highest detection confidence results as the start point for the correlation filter. The plot shows that the proposed SRP method greatly improves the recall of the proposals compared to the baseline RPN method, which indicates that the correlation filter on the convolutional feature maps is able to effectively propagate the region proposals from the high-confidence detection results. When the clip number increases from 10 to 20, the recall of the proposals slightly rises, and we finally set the clip number as 10 for our experiments, and the proposal number on each frame is usually less than 500.

2) ABLATIVE STUDY

To comprehensively investigate the behavior of the proposed method, we conduct several ablative studies on the ImageNet VID validation set. First, we train Faster R-CNN on the DET subset and the VID dataset. For the DET subset, the training data of "dog" and "bird" are condensed. For the VID training set, 4% of the frames are sampled from the whole videos to avoid redundant training examples. Detailed information about the training data is given in Table 4. These two datasets are quite different from each other, and their combination improves the performance. We fix the convolutional layers trained by Faster R-CNN, and train the cascaded region classification and regression network, leading to an improvement of 4.5%. Using multi-scale training slightly improves the result by 2.2%. Next, we investigate the role of temporal information. When using the correlation filter to generate region proposals, the mAP increases to 65.9%, suggesting that the temporal information from the video is very effective to prorogate region proposals from the high-confidence detection results. On the other hand, incorporating context information by an efficient Look-Up-Table method increases the result by 3.3%. This suggests that the context information is able to suppress some conflicting false positives and guide the proposed detector to produce a semantically coherent interpretation of a video. Finally, we re-weight the training data and ensemble two separate model. The mAP improves to 71.8%. We have trained 12 region classification and regression networks (2 scales × three cascade × two group), which aim to improve the localization accuracy and give a more reasonable score for each region proposal by decreasing the variance of the original problem.

We also evaluate the proposed method on the ImageNet DET validation set. The training data are the combination of the DET subset and VID sampling set. As is shown in Table 4, CRFN obtains a performance gain of 3.6% compared to Faster R-CNN. This suggests that the cascaded regional feature-routing regression method helps to improve the localization accuracy, and thus increases the region classification performance. Using the re-weighted training data, the CRFN ensemble further increases the mAP by 1.8%. We further compare our result with the state-of-the-art CUvideo on the ImageNet VID test set, in which our result is 72.1%, which is 2.4% higher than the result of CUvideo.

3) EXPERIMENTS ON IMAGENET2016

We slightly improved the proposed method and participated in the VID 2016 task. In contrast to the region-based Fast R-CNN that apply a costly per-region subnetwork hundreds of times, we employed a more efficient region-based fully convolutional networks with almost all computation shared on the entire image [41]. In addition, we utilized a more efficient network GoogleNet v2 [42] instead of the

TABLE 4. Detection results on the ImageNet DET validation set and VID test set. The statistics of the training data are also given in the left.

	Traing	data	DET	Validation	set	VID	Test	set
	DET Subset all:107910	VID Sampling all:69276	Faster R-CNN	CRFN	CRFN Ensemble	CUVideo provided data	CUVideo outside data	Our method
airplane	1745	3443	62.3	69.5	72.8	85.5	87.6	89.4
antelope	2569	2376	70.9	74.4	76.3	45.5	48.7	54.9
bear	3025	2076	81.4	85.4	87.8	39.6	40.4	51.8
bicycle	1849	1396	54.7	62.5	65.1	60.3	61.4	57.6
bird	7858	5158	87.6	88.3	89.5	57.7	60.2	58.8
bus	2957	1207	67.8	69.5	72.1	88.6	88.1	90.7
car	10518	4583	55.2	58	61	49.3	47.8	54.1
cattle	1328	2122	55.8	63.8	66.3	54.2	56.8	72.5
dog	13383	5186	91.5	91.4	91.8	75.8	77.2	84.6
domestic_cat	3504	2356	72.3	76.1	77.9	54.3	56.8	61.6
elephant	2084	3363	78.5	83.5	85.9	98.2	98.6	94.9
fox	2678	1487	81.2	84.3	86.6	87.9	89.8	94.7
giant_panda	961	2119	68.7	71.5	72.3	87.4	87.5	77.2
hamster	874	1543	64.4	68.2	68.3	81.5	81.8	82.6
horse	2215	2173	66.1	68.1	69.2	84.5	84.9	79.5
lion	999	1288	77.8	81.6	82.4	65.4	67.5	65.8
lizard	5911	1273	79.1	87.8	90.4	67.1	70.5	72.6
monkey	8763	2832	75.4	81.5	84.3	52.6	54.1	60.5
motorcycle	2616	1378	66.3	67.5	67.3	56.2	59.9	63.3
rabbit	2188	1567	86.7	88.7	89.1	82.4	84.2	84.9
red_panda	1025	1917	78.9	78.5	78.9	89.4	89.1	79.8
sheep	1875	1436	69.2	71.1	73.4	71.5	75.4	68.5
snake	8694	1285	71.8	78.2	81	28.4	36	57.7
squirrel	925	1876	74.9	76.2	77.4	47	52.4	70.8
tiger	1191	842	84.6	85.8	85.5	74.9	75.3	82.7
train	1352	4216	71.1	73.3	75.9	74.4	76.4	69.8
turtle	2985	1783	76.3	80.1	82.8	71.8	72.9	79.5
watercraft	9170	2351	53.3	59.2	63.4	69.4	73	72.9
whale	1396	1566	70.2	77.2	82.5	46.6	48	44.3
zebra	1272	3078	79.6	81	81.4	87.1	87.5	84.2
mAP			72.5	76.1	77.9	67.8	69.7	72.1

TABLE 5. Detection results on the ImageNet VID validation set. "+" denotes these methods are incrementally incorporated on the RPN and R-FCN(GNv2) baseline.

method	mAP on the VID validation set
Baseline:RPN+R-FCN(GNv2)	63.85
+CRFN	67.37
+SRP (increase recall from 86.2% to 95.7%)	73.02
+Context inference(supress FP)	76.19
+training data re-weighting	77.30
+Multi-scale test	78.44
+Ensemble with a self-trained deeper inception + shortcut network	81.15

VGG-16 network. As is shown in Table 5, the baseline method obtained the mAP of 63.85% by a frame by frame detection. The performance gradually increased to 77.30% by incorporating the proposed cascaded regional feature-routing classification and regression, temporal region proposal, context inference, and training data re-weighting. This is a real-time object detector on GPU with high accuracy. When multi-scale testing and model ensemble with deeper inception and shortcut network are used, the final mAP is 81.15% on the validation set and 80.8% on the test set, and our algorithm has achieved the first place in the ImageNet ILSVRC2016 Object Detection from Video task.

V. CONCLUSION

Video object detection is an important vision task, yet has received little consideration in the context of general object detection. In this work, we suggest to train several cascaded regional feature-routing classification and regression networks on top of the convolutional feature maps, which are able to improve the localization accuracy, leading to outstanding results on the Pascal VOC 2007 object detection benchmark. Moreover, for video object detection, we explore the correlation filter tracking on the convolutional feature maps to efficiently generate region proposals from the high-confident detection results. In addition, we also perform object co-occurrence inference via an efficient Look-Up-Table method, which can suppress the conflicting false positives. Extensive evaluations on the ImageNet VID dataset demonstrate that the proposed CRSFN outperforms the ImageNet ILSVRC2015 winner CUvideo on the task of video object detection.

ACKNOWLEDGE

The authors would like to thank the two anonymous referees for their constructive comments which are extremely helpful for improving this work.

REFERENCES

[1] J. Lei *et al.*, "A universal framework for salient object detection," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1783–1795, Sep. 2016.

[2] J. Li *et al.*, "Attentive contexts for object detection," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 944–954, May 2017.

[3] M. Wang, R. Hong, X.-T. Yuan, S. Yan, and T.-S. Chua, "Movie2Comics: Towards a lively video content presentation," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 858–870, Jun. 2012.

[4] A. Tawari and M. M. Trivedi, "Face expression recognition by cross modal data association," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1543–1552, Jul. 2013.

[5] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[7] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[8] 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.

[9] K. E. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders, "Segmentation as selective search for object recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1879–1886.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[11] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[12] 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.

[13] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, 2015, pp. 1–9.

[14] K. He, X. Zhang, S. Ren, and J. Sun. (2015). "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification." [Online]. Available: https://arxiv.org/abs/1502.01852

[15] J. Huang *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proc. CVPR*, 2017, pp. 3296–3297.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 346–361.

[17] K. He, X. Zhang, S. Ren, and J. Sun. (2016). "Deep residual learning for image recognition." [Online]. Available: https://arxiv.org/abs/1502.01852

[18] X. Xiong and F. De la Torre, "Global supervised descent method," in *Proc. CVPR*, 2015, pp. 2664–2673.

[19] S. Ren, K. He, R. Girshick, X. Zhang, and J. Sun. (2015). "Object detection networks on convolutional feature maps." [Online]. Available: https://arxiv.org/abs/1504.06066

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. (2015). "You only look once: Unified, real-time object detection." [Online]. Available: https://arxiv.org/abs/1506.02640

[21] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. (2013). "Maxout networks." [Online]. Available: https://arxiv.org/abs/1302.4389

[22] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proc. CVPR*, 2014, pp. 2147–2154.

[23] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1078–1085.

[24] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. CVPR*, 2013, pp. 532–539.

[25] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[26] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang, "Fast visual tracking via dense spatio-temporal context learning," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 127–141.

[27] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3074–3082.

[28] R. Yao, S. Xia, Z. Zhang, and Y. Zhang, "Real-time correlation filter tracking by efficient dense belief propagation with structure preserving," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 772–784, Apr. 2017.

[29] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

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

[31] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[32] L. Wang, X. Zhao, Y. Si, L. Cao, and Y. Liu, "Context-associative hierarchical memory model for human activity recognition and prediction," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 646–659, Mar. 2017.

[33] P. Pourashraf and F. Safaei, "Perceptual pruning: A context-aware transcoder for immersive video conferencing systems," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1327–1338, Jun. 2017.

[34] M. J. Choi, A. Torralba, and A. S. Willsky, "A tree-based context model for object recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 240–252, Feb. 2012.

[35] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[36] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[37] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2241–2248.

[38] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks," in *Proc. CVPR*, 2015, pp. 437–446.

[39] M. Everingham *et al.*, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[40] D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing error in object detectors," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 340–353.

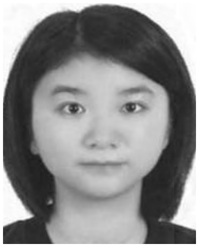
[41] Y. Li *et al.*, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[42] 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

HUI SHUAI received the bachelor's degree in automation from the Nanjing University of Information Science and Technology in 2015. He has been pursuing the master's degree at the School of Information and Control, Nanjing University of Information Science and Technology since 2015. His research interests include machine learning and computer vision.

QINGSHAN LIU (SM'08) received the M.S. degree from Southeast University, Nanjing, China, in 2000, and the Ph.D. degree from the Chinese Academy of Sciences, Beijing, China, in 2003. During 2004–2005, he was an Associate Researcher with the Multimedia Laboratory, Chinese University of Hong Kong, Hong Kong. He was an Associate Professor with the National Laboratory of Pattern Recognition, Chinese Academy of Sciences. From 2010 to 2011, he was an Assistant Research Professor with the Computational Biomedicine Imaging and Modeling Center, Department of Computer Science, Rutgers, The State University of New Jersey, Piscataway, NJ, USA. He is currently a Professor with the School of Information and Control Engineering, Nanjing University of Information Science and Technology, Nanjing. His research interests include image and vision analysis and machine learning.

KAIHUA ZHANG received the B.S. degree in technology and science of electronic information from the Ocean University of China in 2006, the M.S. degree in signal and information processing from the University of Science and Technology of China in 2009, and the Ph.D. degree from the Department of Computing, The Hong Kong Polytechnic University, in 2013. From 2009 to 2010, he was a Research Assistant with the Department of Computing, The Hong Kong Polytechnic University. He is currently a Professor with the School of Information and Control, Nanjing University of Information Science and Technology, Nanjing, China. His research interests include image segmentation, level sets, and visual tracking.



JING YANG received the M.S. degree from the School of Information and Control, Nanjing University of Information Science and Technology, in 2017. Her research interests include machine learning and computer vision.

JIANKANG DENG received the M.S. degree from the School of Information and Control, Nanjing University of Information Science and Technology, in 2015. His research interests include machine learning and computer vision.