IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 16, 2020, accepted May 29, 2020, date of publication June 3, 2020, date of current version June 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2999596

Weakly Supervised Object Detection Using Complementary Learning and Instance Clustering

MEHWISH AWAN AND JITAE SHIN, (Member, IEEE)
College of Information and Communication Engineering, Sungkyunkwan University, Suwon 16419, South Korea

Corresponding author: Jitae Shin (jtshin@skku.edu)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2017R1D1A1B03031752, and in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Centre (ITRC) support program, supervised by the Institute for Information & Communications Technology Promotion (IITP) under Grant IITP-2020-2018-0-01798.

ABSTRACT Supervised object detection schemes use fully annotated training data, which is fairly expensive to constitute. Whereas, weakly supervised object detection (WSOD) uses only image-level annotations for training which are much simpler to acquire. WSOD is a challenging task since it aims to learn object localization and detection with image-level labels. In line with this assertion, in this paper, we present an end-to-end framework for WSOD based on discriminative feature learning. We use the objectness technique to get initial proposals from the images. Afterwards, two complementary networks are trained in parallel to obtain discriminative image features, which are channel-wise concatenated with the features of the third network. We name this classification network designed for discriminative feature learning as fused complementary network. This network learns the proposals enclosing whole object instances by complementary features which ultimately learns to predict the high probabilities for whole objects than proposals containing only object parts. Clustering is then hierarchically performed on the region proposals. Our clustering method, named instance clustering, first performs inter-class clustering followed by iterative intra-class clustering using intersection-over-union metric to obtain spatially adjacent cluster members corresponding to each object instance. In each intra-class clustering iteration, the high scoring proposal is set as centroid from each intra-class cluster. Experiments are conducted on PASCAL VOC2007 and PASCAL VOC2012 datasets. Both qualitative and quantitative results have shown improved WSOD performance on these benchmarks.

INDEX TERMS Weakly supervised object detection, complementary learning, discriminative features, instance clustering, and deep learning.

I. INTRODUCTION

In supervised object detection, bounding box annotations are required for training on multi-label images. Gathering ground-truth bounding boxes for natural images is the major limitation in real-world object detection applications since it is a time-consuming and laborious task [1]. Using weakly supervised learning (WSL) to object detection is an appropriate solution to object annotations problem. Weakly supervised object detection (WSOD) refers to learning object detections with only image-level annotations [2], [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea F. F. Abate.

Many WSOD approaches have hitherto been proposed. In literature, the mainstream follows the conventional multiple instance learning (MIL) approach for WSOD problem [4]–[6]. Since MIL is used in many computer vision applications, many variants [7], [8] of MIL have been proposed to date including image classification, object detection, semantic segmentation, etc. In MIL, instances are collected as a set of positive and negative bags. These bags are considered as labels for the classifier. Although MIL networks achieved some promising results, there are problems with the assumptions while optimizing the parameters of the classifier. For instance, it is assumed that positive bags are known in an image. The other assumption, typically made with MIL based methods, is that the most likely positives

are predicted using the existing classifier. Due to predicted false positives, the learning process could be erroneous in such cases as classifier explicitly cannot deduct true positives in a given image [1]. Additionally, due to the remarkable performance delivered by convolutional neural networks (CNNs) [9] in computer vision, several works combined the MIL with CNNs to get better WSOD performance [10], [11]. Many recent works used pre-trained CNN models on large scale datasets as feature extractors [10], [12]. CNN-learned features achieve improved performance in contrast to traditional hand-engineered features as shown by these methods. Some recent methods perform end-to-end training [3], [13] with MIL and object proposals extracted from images to achieve better WSOD performance [3], [14]. However, these pipelines are complex and involve a series of refinements on region proposals, which results in inefficient inference and cannot be used in real-time systems. In this paper, we also have performed fast object detections by training a fully supervised object detector with pseudo ground-truths inferred by the proposed WSOD method to achieve fast WSOD.

Although many methods have been studied for WSOD as previously discussed, it is yet challenging to attain high accuracy in WSOD. In this study, we present a WSOD method to achieve improved detection performance. In particular, our method is based on discriminative feature learning and clustering to proposal extraction for WSOD. The main idea for using complementary classifiers is to mine the proposals containing entire object instances in a given image. Two classifiers (network *A* and *B* in Figure 1) with distinct image features are trained in parallel inspired by [2] and [15]. Network *C* (Figure 1) is the main object detector network which is branched into two data streams computing recognition and detection scores separately after region level spatial pyramid pooling (SPP). The detection stream is proposed to rectify region scores based on entropy values. In our detection branch, we introduce the entropy layer to learn minimized entropy over region proposals in context with object detection.

The final score for all regions is computed by performing the Hadamard product between recognition and detection scores. Hereafter, we apply our clustering method named instance clustering (IC) on these proposals. This method efficiently removes many of the proposal bounding boxes which encloses incomplete objects. Contrasting to standard non-maximum suppression (NMS), the proposed IC is not only efficient in mining significant detections but also intended to overcome the problem of duplicate detections. Duplicate detection occurs when for a single instance there exists another candidate proposal with a high category score and that proposal has intersection-over-union (IoU) value below IoU threshold with the highest scoring proposal for that instance. This problem is mainly observed in WSOD since no bounding box regression is performed. It degrades the overall precision of the object detector. Therefore, the proposed IC is designed to identify such defective detections and remove them.

IC groups the proposal bounding boxes in a hierarchical and iterative way. Inter-class clustering is first employed to generate clusters with respect to the predicted category for each proposal. After that, intra-class clustering is performed iteratively using IoU metric. A cluster contains spatially adjacent proposals that belong to a single object instance. The maximum scoring proposal is chosen from each cluster as a final proposal for the particular object instance. As a result of this mining process, significant proposals are extracted. The proposed approach is simple yet very efficient to deal with a very challenging task of WSOD.

In this paper, our contributions are listed as follows:

1) We propose an end-to-end network for WSOD based on the proposal mining approach. Our WSOD method comprises of two modules, complementary learning and instance clustering. These two modules precisely overcome the two main problems of WSOD, partial object detection and duplicate detections of the same object instance.
2) A complementary network with fused discriminative features named fused complementary network (FuCN) is proposed. This network learns features for the entire object and hence increases the likelihood of selecting correct bounding boxes containing whole object instances instead of selecting incorrect bounding boxes with only parts of objects.
3) We propose entropy loss in the detection branch to learn region proposals with minimum entropy.
4) For mining proposals and removing duplicate detections, we propose the IC method. Inter-class clustering is first performed based on region scores and then intra-class clustering is performed iteratively using the IoU metric to extract spatially adjacent clusters for each instance. The IC method efficiently extracts the multiple object instances for multi-label images.
5) We evaluate our method on PASCAL VOC2007 and PASCAL VOC2012 datasets in terms of average precision (AP) and correct localization (CorLoc) to measure accuracy and inference time in frames per second (FPS).

The rest of the paper is organized as follows. Section II is devoted to related work. In Section III we present our proposed method. Section IV presents the experimental results and discussion. Finally, in Section V we conclude the paper.

II. RELATED WORK

In recent years, WSOD has been broadly investigated. MIL is a weakly supervised learning paradigm and followed by the majority of methods in the literature for WSOD. However, the MIL approach has a potential problem of non-convex optimization. Several studies have intended to standardize optimization by improving the MIL initialization strategy. Cinbis *et al.* [10] proposed a multi-fold MIL approach, this strategy was helpful in avoiding the performance collapse in object localization. Tang *et al.* [14] applied clustering on proposals and then they combined MIL with a series of CNN

classifiers for proposals refinement. The authors integrated the MIL with CNNs into the network training and refined proposals iteratively. Deselaers *et al.* [16] used objectness [17] to initialize boxes and proposed to use the conditional random field (CRF) in order to localize the object instances concurrently and learn an appearance model over the iterations so that the CRF progressively adapts to the new class.

Several methods [3], [5], [13], [14] investigated WSOD in end-to-end training fashion. Wang *et al.* [4] relaxed the MIL optimization constraints into a convex program and optimized by stochastic gradient descent (SGD) to train detectors effectively. Bilen and Vedaldi [3] presented an end-to-end CNN framework initialized by objectness, which divides into two parallel streams of classification and detection. Li *et al.* [18] presented a progressive domain adaptation method and then performed adaptations at both streams, i.e., classification and detection. Kantorov *et al.* [13] proposed deep CNN models by using contextual information for improved localization. Jie *et al.* [19] presented a self-taught learning method to learn object spatial location information for training detector. The detector is learned to localize positive samples progressively. Sangineto *et al.* [1] proposed a self-paced learning approach and trained with Fast RCNN [20]. During network training, the same network at different progression stages is used to predict object localization of positive samples. At each stage, a subset of images is selected whose pseudo ground-truth is the most reliable. Zhang *et al.* [2] used an adversarial complementary learning approach inspired by [15]. The authors trained two parallel networks with the feature maps of the first classifier thresholded and then erased from input features to another classifier. This approach enhances object localization performance.

Shen *et al.* [21] used a generative adversarial learning approach for end-to-end WSOD, they used single shot multibox detector (SSD) [22] as a fast detector. Authors in [23] proposed a collaborative self-paced learning framework with weakly supervised settings using both instance level and image-level prior-knowledge. Recently, Shen *et al.* [24] studied the multi-task learning for WSOD. They treated object detection together with semantic segmentation as a joint learning problem to overcome the failure patterns of segmentation and object detection which are typically encountered in other MIL based self-enforcement methods [5], [10], [12] trained with single-task learning. Li *et al.* [25] studied WSOD as a joint task with weakly supervised segmentation trained in end-to-end fashion, where each individual task supervises the accompanying task in a collaborative loop. Lately, Zhang *et al.* [26] studied WSOD with reinforcement learning approach under region-searching paradigm. They used region correspondence maps as pseudo-target regions to train the agent under weak supervisions. These localization maps are used as the states information.

The aforementioned methods use multiple stages of refinement and these networks [7], [14], [16], [19] employ many steps, making it difficult to perform fast detections efficiently. These methods [14], [23], [25] have achieved suitable performance, however, these approaches still have the problem of partial object detection and missing detection in case of occluded objects. Since the training image is decomposed into thousands of proposals, each approximately correct training instance is flooded with many incorrect training instances. Our method differs from the aforementioned methods in many aspects, our network is trained end-to-end and does not require many stages or steps for refinement of proposals for WSOD. The most important aspect of our network is complementary feature learning which greatly encourages the network to learn whole objects, and thereby improves object detections. We have revised the method in Zhang *et al.* [2] in order to obtain proposals with integral object regions. However, we use an additional parallel network that takes input features combined from complementary networks to learn whole object features. We perform position-aware channel-wise concatenations of feature maps to learn the whole features corresponding to the object category. Hence, our objective of using complementary network is to learn whole object features. These channel-wise concatenated complementary features also work as spatial regularizer unlike the one used in [3] that penalizes the feature map discrepancies between high scoring regions with high overlap. Moreover, we use IC to cluster the proposals for each

TABLE 1. Notations.

Notation	Description
$\boldsymbol{I}$	set of input images
Y_i	label-vector of i^{th} image
F_{VGG}	extracted features from VGG backbone network
F_A	extracted features from network A
F_E	erased F_{VGG} from F_A
F_B	discriminative features extracted from network B
F_C	extracted features from network C
δ_{erase}	feature erase threshold
E_A	extracted heatmap from network A
F_F	channel-wise concatenated features (F_A, F_B, F_C)
δ_{score}	score threshold
δ_{IoU}	IoU threshold
δ_{out}	outlier threshold
x^C	score matrix from classification branch
x^D	score matrix from detection branch
x^F	final score matrix
C^j	inter-class cluster of j^{th} class
r_i^{jmax}	region propsal with maximum class score
k_i^j	intra-class centroid proposal
X_i^{jm}	an instance cluster
$\boldsymbol{Z}$	set to hold number of each cluster members for a particular instance
$\boldsymbol{Y}$	subset of $\boldsymbol{Z}$ to follow a symmetric distribution
μ_Z	mean of $\boldsymbol{Z}$
σ_Z	standard deviation of $\boldsymbol{Z}$
μ_Y	mean of $\boldsymbol{Y}$
σ_Y	standard deviation of $\boldsymbol{Y}$
$\boldsymbol{P}$	final proposals set

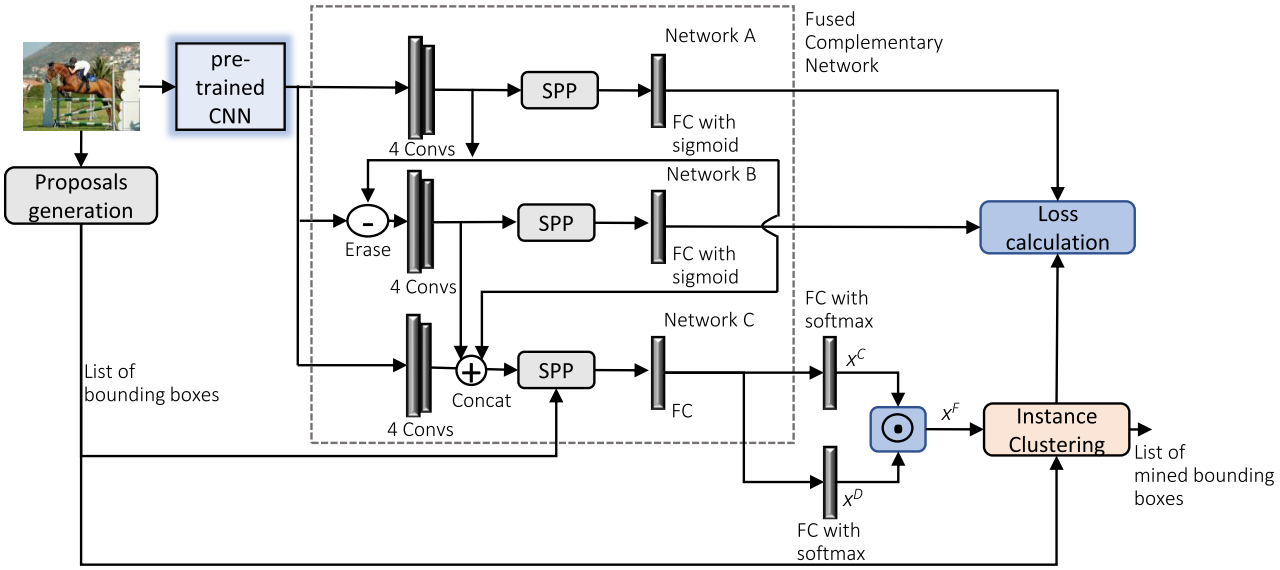


FIGURE 1. An architectural diagram of the proposed WSOD method.

object instance of a particular class separately and iteratively extracting the multiple object instances in multi-label images. Complementary feature learning together with IC is a better, robust and computationally efficient approach than instance classifier refinement used in [5] and [14].

III. PROPOSED METHOD

This section describes the proposed method for WSOD in detail. The architectural diagram of our method is shown in Figure 1. It is an end-to-end object detection network trained with only image-level labeling. FuCN is the major component of the proposed method followed by the IC module. A list of about 2,000 bounding boxes is generated by selective search [27] as the region proposals. The position-aware feature maps from input image are extracted by VGG16 [28] with a reduced fully connected (FC) layer pre-trained on ImageNet [29] classification task. Any pre-trained CNN can be used as a backbone network as the sole purpose is to extract features. These features are fed to FuCN, which is a network of three parallel convolutional networks. Each network consists of 4 convolutional layers followed by a single-level SPP layer. In networks A and B, the output of the SPP layer is fed to the FC layer followed by a sigmoid layer for multi-label classification. We extract feature maps from the entire image which are fed to all three parallel networks as input. In network C, we apply SPP on the feature maps to produce fixed-length representations of proposals to be fed to the subsequent FC layer. Table 1 shows the notations used in this paper.

Here the network is split into two branches, the classification branch, and the detection branch. The classification branch computes class probabilities for each region proposal using softmax nonlinearity (x^C). The proposed detection branch is trained based on the entropy values of the proposals. The region proposals are optimized with respect to the minimum entropy loss function. It uses spatial information to perform regions comparison by computing entropy of probability distribution and is supportive in rectifying proposal scores. In the regions where the randomness in probability distribution is high, there will be high entropy values corresponding to those regions. However, the regions with skewed probability distribution will have less entropy value and correspond to the accurate object localizations in terms of detections. The entropy layer is followed by a softmax layer to rank proposals based on entropy values as region scores (x^D). The final score matrix (x^F) is computed by taking Hadamard product (element-wise product) of two score matrices x^C and x^D from classification and detection branches, respectively. The authors in [3] used x^F for computing image-level classification score, since x^F, which is calculated based on local information of regions, affects the learning process and may lead to convergence to wrong local minima. Unlike [3], we compute image-level classification scores from the classification branch by max-pooling over regions in a class-specific manner. Clustering is then applied to the list of proposals to get the final mined bounding boxes. Each module is discussed in detail in subsequent subsections. We also train a separate fast object detector SSD [22] and feed it with detected bounding boxes by our method as pseudo ground-truths for fast WSOD.

A. THE PROPOSED FuCN

A deep classification network extracts unique patterns for a specific category [2]. However, those features do not essentially cover whole object regions but only emphasize the distinct features corresponding to that category. This yields good recognition performance but cannot be effective for detection where the goal is to detect the entire object. In the proposed FuCN, we use a discriminative feature learning approach

for object detection under weak supervisions of image-level annotations. Two complementary networks (classifiers A and B) are trained with distinct input features, the complementary features from both networks are channel-wise concatenated with the position-aware feature maps of the third network (detection network C). These concatenated features in network C are followed by a 1×1 convolution layer to maintain the number of channels. This complementary network ultimately learns the proposals covering the entire object. A cascade of three complementary classifiers does not further contribute to localize any distinct regions as discussed and analyzed in [2]. Therefore, we employ two complementary networks to obtain discriminative features, since including the third classifier has no significance. The image features learned by the convolutional layers of network A are thresholded and erased (δ_{erase}) from the features extracted by the backbone to input these erased features to network B. An FC layer followed by a sigmoid layer is applied on top of SPP layer for multi-label classification in the networks A and B as shown in Figure 1. By the concatenation of complementary features of networks A and B with the feature maps of network C, the network C attains distinct features of the same object which are imperative to mine the proposals with the whole object. In network C, the second FC layer is followed by the softmax layer. Afterwards, image-level classification score from network C is computed by max-pooling over all regions on the class-specific basis.

In particular, image features are extracted by pre-trained VGG16 [28] on ImageNet [29]. Then, we add four additional convolutional layers (Convs in Figure 1) in all three networks followed by a single-level SPP layer and the generated representations pass through the FC with sigmoid. For feature erasing, the discriminative features of network A are first used as the threshold on its heatmap and then these regions are erased from the input features for classifier B. Erased values are replaced by zeros. Network B is then encouraged to learn from the features of other regions corresponding to the target object. The confidence scores and labels of proposals are obtained from network C for proposal mining. Algorithm 1 illustrates the proposed discriminative feature learning procedure for object detection adapted from [2] and [15].

Algorithm 1 FuCN with Discriminative Feature Learning

Inputs: Image (I_i), label-vector (Y_i), threshold (δ_{erase})
Output: Concatenated feature map (F_F)

Extract image features $F_{VGG}(I_i, Y_i)$ from backbone VGG
Extract features F_A from network A (F_{VGG}, $Convs$, Y_i)
Extract features F_C from network C (F_{VGG}, $Convs$, Y_i)
Extract heatmap E_A from network A (F_A, Y_i)
Obtain discriminative region $R = E_A > \delta_{erase}$
Extract erased features $F_E = F_{VGG} - R$
Feed F_E to $Convs$ in network B
Extract F_B from network B (F_E, $Convs$, Y_i)
Obtain F_F by concatenated (F_A, F_B, F_C)

Algorithm 2 Instance Clustering

Inputs: Image (I_i), Proposals (R_i), IoU threshold (δ_{IoU}), Score threshold (δ_{score})
Output: Final proposals (P_i)

Remove $r_i < \delta_{score}$
Inter-class clustering: make c clusters C^j
Intra-class clustering: *for* $j = 1$ *to* c *do*
  While $C^j \neq$ *empty do*
    Select centroid (k_i^j) with $r_i^{j_{max}}$ for that C^j
    Cluster $X_i^{j_m}$ such that $IoU(k_i^j, r_i^j) > \delta_{IoU}$
    Remove $X_i^{j_m}$ from $C^j$and update C^j
Calculate δ_{out}
for $j = 1$ *to* c *do*
  Repeat m times
    if $length(X_i^{j_m}) <= \delta_{out}$ *then*
      Remove $X_i^{j_m}$ as outlier
    else
      Choose $r_i^{j_{max}}$ as final $p_i^{j_m}$ from $X_i^{j_m}$
      Append ($p_i^{j_m}$) to P_i

B. INSTANCE CLUSTERING

Natural images may contain many instances for the same category. In this section, we explain our IC method for clustering the proposals for object instances in a given image. Our clustering method clusters the proposal bounding boxes hierarchically and iteratively. Algorithm 2 illustrates the procedure for IC. We first filter the proposals based on the region confidence score threshold (δ_{score}). Proposals with less than δ_{score} are straightforwardly dropped. This initial thresholding is performed for two reasons; first, the objective is to extract the final proposals with high scores encompassing integral regions, and secondly to effectively reduce the computational cost for the next clustering step. This thresholding step can be skipped but it results computational overhead due to redundant boxes. Our IC method includes two main phases; inter-class clustering and intra-class clustering.

For inter-class clustering, we make a set of the proposal bounding boxes C^j corresponding to each class separately. If there exist c classes in the given image, then c inter-class clusters are generated. Each inter-class cluster contains proposals with the same category and region score at least above δ_{score}.

Then we perform intra-class clustering iteratively by using bounding boxes spatial relationship information. We perform intra-class clustering up to m iterations depending upon the number of instances presented in the image for that particular category. Hence, for each object instance, intra-class clusters are generated and this process is repeated for m times intended for all the instances in that inter-class cluster. At each iteration, the proposal with maximum region score $r_i^{j_{max}}$ is set as the centroid k_i^j. We compute IoU between

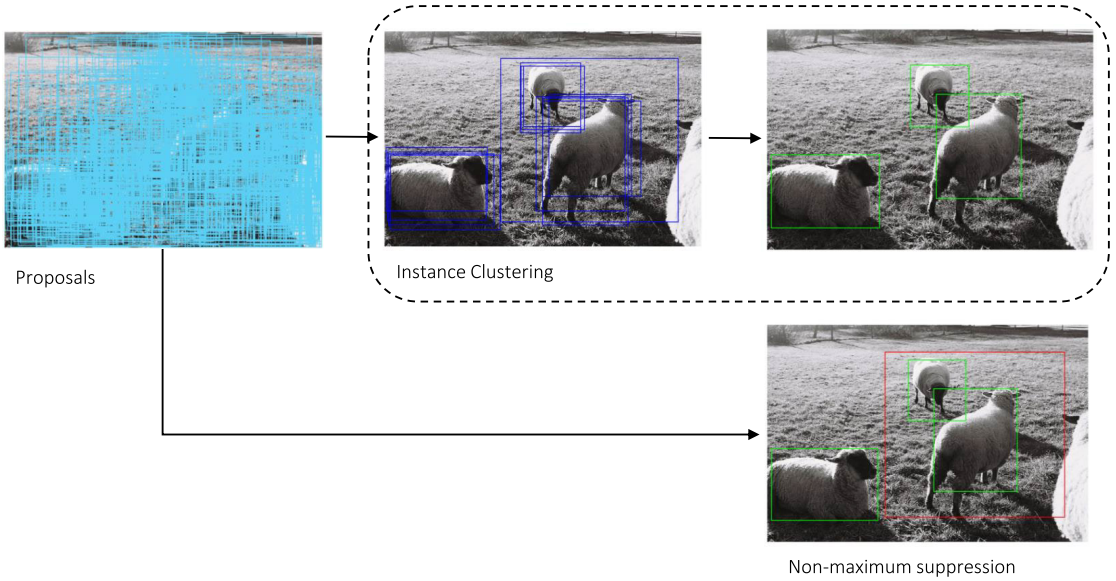


FIGURE 2. Qualitative comparison between proposed instance clustering and non-maximum suppression.

centroid bounding box k_i^j and all other proposals r_i^j in the intra-class cluster. An instance cluster X_i^{jm} is generated on proposals with IoU above δ_{IoU} with centroid. This gives the spatially adjacent bounding boxes for that particular instance. After generating an instance cluster, the intra-class cluster is required to update. Proposals from the generated instance cluster are removed from the intra-class cluster. In the next iteration, the maximum scoring proposal from the updated intra-class is set as centroid, and IoU is computed among centroid and all other proposals in the intra-class cluster. Another spatially adjacent instance cluster is constructed for the other instance of the same category in an image. The same process is implemented for all instances in m iterations. In this way, all intra-class clusters are clustered into instance clusters corresponding to the instances in the image.

It is the common observation that in some cases, the standard NMS has a problem of duplicate detections for a single instance, specifically when used for WSOD systems. This happens when there exist high scoring candidate(s) for that instance but below IoU threshold with the highest scoring proposal. The duplicate detections or the outlier detections degrade the overall precision of the object detector. To overcome this problem, the proposed IC is designed to identify such defective detections and remove them. We believe that the number of candidate proposals surrounding the object instance for a true positive detection will be greater, and therefore, the instance cluster will have a greater number of proposals within the IoU threshold. Whereas, the instance cluster generating the duplicate detection for that instance will have fewer proposals although of high confidence scores. Consequently, the number of members in the instance clusters follows a distribution, while the outlier clusters deviate from that distribution at least to some extent. A set $\boldsymbol{Z}$ is created that holds the number of members of instance clusters in a particular intra-class cluster. Afterwards, we sort $\boldsymbol{Z}$ in descending order. We calculate the mean (μ_Z) and standard deviation (σ_Z) of $\boldsymbol{Z}$. Since the mean and standard deviation of the set are skewed by a widely varying number of cluster members in intra-class clusters and can result in the incorrect threshold for outliers identification. Therefore, a set $\boldsymbol{Y}$ is generated to consider nearly symmetric distribution by following a simple criterion. Each data point in set $\boldsymbol{Z}$ is subtracted from the standard deviation ($o = z - \sigma_Z$), and then compared with mean such that, if o is less than $\mu_Z (o < \mu_Z)$ then z is member of $\boldsymbol{Y}$, $\boldsymbol{Y} = \{z |\ z \in \boldsymbol{Z}, o < \mu_Z, o = z - \sigma_Z\}$.

Afterwards, we calculate the mean (μ_Y) and standard deviation (σ_Y) of the set $\boldsymbol{Y}$. The outlier threshold (δ_{out}) is calculated by computing average of μ_Y and σ_Y. Instead of only considering standard deviation (as discussed in Section IV: I Ablation Study) as the threshold for removing outliers, the average of μ_Y and σ_Y gives a sophisticated threshold even if the outlier data points are close to the mean. Floor function is then applied on average as defined below;

$$\delta_{out} = \left\lfloor \frac{\mu_Y + \sigma_Y}{2} \right\rfloor \tag{1}$$

If an instance cluster has the number of members less than or equal to δ_{out}, it is considered as an outlier cluster and is discarded. This approach enhances the AP significantly. Figure 2 visually illustrates the difference in final detections by the proposed IC method and NMS.

C. TRAINING PROPOSED WSOD METHOD

After feature extraction from pre-trained VGG [28], the network is branched into three further networks, A, B and C. Network A and network B extract discriminative features

which are then concatenated for input to network C (detection network). For N multi-label images in trainval set, label-vector for i^{th} image is $y_i = [y_{i1}, y_{i2}, \ldots, y_{iC}]$, where $y_{ij} = 1(j = 1, \ldots, C)$, if j^{th} class object is present in image and $y_{ij} = 0$ otherwise, and C is the total number of categories. We use binary cross entropy (BCE) loss function for training networks A and B as in (3). Category-specific scores for each image are obtained from sigmoid output. For network C, image-level classification scores are computed by class-specific max-pooling on all regions R, $R = (r_1, r_2, \ldots, r_T)$, here T is the total number of proposals. For i^{th} image, the j^{th} class score is calculated by maximum of all regions j^{th} class softmax probabilities, $p_{ij} = max(p^j_{r_1}, p^j_{r_2}, \ldots, p^j_{r_T})$. Thus, the prediction vector for i^{th} image is computed as $p_i = [p_{i1}, p_{i2}, \ldots, p_{iC}]$. In network C, we employ BCE loss function for training classification branch and minimum entropy loss in detection branch to learn regions with minimum randomness as in (4). SGD with momentum 0.9 and weight decay 5×10^{-4} is used for optimizing object detector. In network C, all steps after the last two parallel FC layers are not included in backpropagation and hence not involved in the network learning process. Our WSOD is trained with learning rate 10^{-3} for the first 30 epochs and then with learning rate 10^{-4} for the remaining 40 epochs. The entire training settings except the loss functions are same in all three networks A, B, and C since it is an end-to-end WSOD network. The loss function for the entire network is defined as follows;

$$L = L_A + L_B + L_C \tag{2}$$

$$L_A = L_B = -\sum_{j=1}^{C} \left(y_{ij}.log\left(p_{ij}\right) + \left(1 - y_{ij}\right).log\left(p_{ij}\right)\right) \tag{3}$$

$$L_C = -\sum_{j=1}^{C} \left(y_{ij}.log\left(p_{ij}\right) + \left(1 - y_{ij}\right).log\left(p_{ij}\right)\right) - \sum_{s=1}^{S} \left(p_s.log\left(p_s\right)\right) \tag{4}$$

where, L is the loss function for the proposed WSOD network, L_A, L_B, and L_C are the loss functions of networks A, B, and C, respectively. S is the number of discrete states (s being the individual state) in the probability distribution.

IV. EXPERIMENTAL RESULTS

A. BENCHMARK DATA

We evaluate the proposed method on PASCAL VOC2007 and PASCAL VOC2012 datasets with 20 object categories which are widely used as benchmarks for object detection. Trainval sets with 5011 images for VOC2007 and 11540 images for VOC2012 are used for training. We use only image-level labels for training. For evaluation, the proposed WSOD method is evaluated on the test set with 4952 images for VOC2007 and 10991 images for VOC2012.

B. PERFORMANCE EVALUATION METRICS

We use two performance measures to evaluate accuracy for object detection. The first metric is AP with 0.5 IoU between detected boxes and ground-truths and mean of AP (mAP). IoU is the metric to evaluate the correctness of the predicted bounding box. It is the ratio between the intersection and the union of the predicted box and the ground-truth box. mAP is an average of the AP computed for all the classes for object detection. Furthermore, we use CorLoc to test the localization accuracy. CorLoc metric is used to evaluate the precision of detections. It is the percentage of images with correctly localized boxes. Both AP and CorLoc metrics measure the quantitative performance of the object detector based on the PASCAL criteria with IoU > 0.5.

C. BACKBONE NETWORK

Backbone is a fully convolutional network to extract position-aware feature maps with multiple channels from the input RGB (red, green, blue) image. We use two separate backbones in our experiments depending on the type of task intended from the deep network. For proposed WSOD, VGG16 [28] pre-trained on ImageNet [29] classification task is used as a backbone. However, for a fast WSOD task, we perform experiments with VGG16 backbone network in first setting and DetNet [30] backbone network in second experimental setting. We trained the detection backbone network i.e., DetNet [30] on PASCAL VOC2012 multi-label dataset for the classification task and then used this network as a backbone for detection task in SSD [22].

D. FAST OBJECT DETECTOR

We train SSD [22] by the proposed WSOD method for fast detections. SSD is fed with image and mined proposals as pseudo ground-truths extracted from our proposal mining method. We made some modifications to the original SSD to improve its performance. The casual convolutions are replaced with dilated convolutions on multi-scale feature layers of SSD. Dilated convolutions have a major benefit of the large receptive field which yields improved spatial resolution and assist in improving detection for small objects with the same cost and memory as casual convolutions. Other hyperparameters and training settings are set as described in [22].

E. IMPLEMENTATION DETAILS

We evaluate our method in two settings. In the first setting, the proposed WSOD with VGG16 backbone is evaluated. In our experiments, FuCN is trained with two complementary networks. For erasing the feature maps extracted by network A from input to network B, 0.6 hard threshold is used as suggested by [2]. In case of too large, network B can be restricted to further extract discriminative object features, and conversely too low threshold can result in similar features as extracted by network A. Therefore, a well-designed threshold

TABLE 2. Comparison (AP in %) on PASCAL VOC2007 test set.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Cinbis *et al.* [10]	39.3	43	28.8	20.4	8	45.5	47.9	22.1	8.4	33.5	23.6	29.2	38.5	47.9	20.3	20	35.8	30.8	41	20.1	30.2
Li *et al.* [18]	49.7	33.6	30.8	19.9	13	40.5	54.3	37.4	14.8	39.8	9.4	28.8	38.1	49.8	14.5	24	27.1	12.1	42.3	39.7	31.0
Bilen *et al.* [3]	42.9	56	32	17.6	10.2	61.8	50.2	29	3.8	36.2	18.5	31.1	45.8	54.5	10.2	15.4	36.3	45.2	50.1	43.8	34.8
Tang *et al.* [5]	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
Tang *et al.* [14]	54.4	69	39.3	19.2	15.7	62.9	64.4	30	25.1	52.5	44.4	19.6	39.3	67.7	17.8	22.9	46.6	57.5	58.6	63	43.5
Zhang *et al.* [23]	63.4	55.0	52.8	36.6	10.7	66.3	57.0	69.5	7.2	52.5	14.4	64.6	69.4	57.7	28.4	15.8	43.7	42.3	69.3	40.5	45.9
Li *et al.* [25]	59.4	71.5	38.9	32.2	21.5	67.7	64.5	68.9	20.4	49.2	47.6	60.9	55.9	67.4	31.2	22.9	45.0	53.2	60.9	64.4	50.2
Our WSOD-NMS	55.3	63.7	46.4	31.2	12.6	65.8	70.3	64.9	12.9	45.7	52.9	68.3	48.8	70.7	38.4	55.4	48.5	52.2	59.5	55.7	50.9
Our WSOD-IC	59.9	64.2	49.3	33.1	13.9	67.9	72.1	64.9	13.5	45.9	53.7	69.3	49.5	71.8	38.5	56.2	50.7	57.1	63.4	58.5	52.6
SSD-NMS*	60.7	65.7	48.9	33.6	14.9	68.2	73.1	65.7	16.2	46.8	54.9	70.1	49.5	73	39.4	59.3	53.4	55.8	62.6	58.7	53.5
SSD-IC*	64.2	66.4	49.9	35.8	15.5	69.5	75.2	67.3	17.7	47.3	55.8	71.6	51.7	73.8	40.3	58.7	54.3	58.7	64.3	62.9	55.0

*Note: SSD-NMS is a fully supervised SSD network [22] trained with pseudo ground-truths extracted by the proposed WSOD using NMS (that is WSOD-NMS). Moreover, it uses NMS for mining the final detections from a pool of candidate proposals. Conversely, the region proposals generated by the proposed WSOD using proposed IC (that is WSOD-IC) are fed as pseudo ground-truths in SSD-IC for training SSD. SSD-IC uses IC for extracting the final detections instead of NMS.

TABLE 3. Comparison (CorLoc in %) on PASCAL VOC2007 trainval set.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	Mean
Cinbis *et al.* [10]	65.3	55	52.4	48.3	18.2	66.4	77.8	35.6	26.5	67	46.9	48.4	70.5	69.1	35.2	35.2	69.1	43.4	64.6	43.7	52.0
Li *et al.* [18]	77.3	62.6	53.3	41.4	28.7	58.6	76.2	61.1	24.5	59.6	18	49.9	56.8	71.4	20.9	44.5	59.4	22.3	60.9	48.8	49.8
Bilen *et al.* [3]	68.5	67.5	56.7	34.3	32.8	69.9	75	45.7	17.1	68.1	30.5	40.6	67.2	82.9	28.8	43.7	71.9	62	62.8	58.2	54.2
Tang *et al.* [5]	81.7	80.4	48.7	49.5	32.8	81.7	85.4	40.1	40.6	79.5	35.7	33.7	60.5	88.8	21.8	57.9	76.3	59.9	75.3	81.4	60.6
Tang *et al.* [14]	79.6	85.5	62.2	47.9	37.0	83.8	83.4	43	38.3	80.1	50.6	30.9	57.8	90.8	27	58.2	75.3	68.5	75.7	78.9	62.7
Zhang *et al.* [23]	83.2	65.0	72.0	64.6	16.8	75.3	79.1	81.3	23.6	80.1	19.0	77.2	84.3	82.9	53.0	28.6	68.8	56.8	87.0	49.6	62.4
Li *et al.* [25]	85.0	83.9	58.9	59.6	43.1	79.7	85.2	77.9	31.3	78.1	50.6	75.6	76.2	88.4	49.7	56.4	73.2	62.6	77.2	79.9	68.6
Our WSOD-NMS	85.3	88.2	69.7	53.8	20.3	77.8	85.1	78.2	27.4	80.8	50.9	75.3	76.3	86.7	72.2	75.8	75.9	65.1	63.8	77.3	69.2
Our WSOD-IC	87.5	89.7	71.3	58.6	22.1	78.9	85.1	78.4	28.5	83.1	50.9	75.9	77.2	88.1	72.5	75.8	76.6	67.8	64.6	78.6	70.5
SSD-NMS	86.6	89.9	71.8	55.9	25.3	79.8	86.3	79.4	29.1	81.5	51.8	76.7	79.8	87.3	73.4	76.4	77.1	68.6	68.1	79.7	71.2
SSD-IC	88.6	90.4	73.5	59.9	25.5	80.9	87.4	79.8	30.3	85.2	51.9	77.7	79.9	89.8	73.8	77.5	78.7	68.8	68.3	79.8	72.3

is significant for learning discriminative feature extraction. For IC in inter-class clustering, first, we mine proposals above 0.6 score from pool of region proposals. We use this threshold to remove many of the proposals which may not contribute to the correct detections. This reduces the computational cost by not focusing on proposals with low scores for further refinement process. Moreover, low score proposals possibly contain object parts. For intra-class clustering, IoU threshold 0.5 is used among top-scoring proposals and other proposals in instance cluster. Under this threshold, the proposals with an appropriate spatial adjacency among them for a single instance can be clustered effectively. Intra-class clustering is repeated m times for remaining instances (if any) present in an image.

In the second setting, for fast detections, SSD300 [22] is trained with pseudo ground-truths extracted by the proposed WSOD method. We use dilated convolutions on multi-scale feature layers with kernel size 3 and padding instead of casual convolutions. We also use DetNet [30] pre-trained on PASCAL VOC2012 multi-label trainval set images with image-level labels as the backbone in the second setting for SSD (reported in Table 6). Some modifications are made in the original DetNet [30] for training it on multi-label images with supervisions as images label-vector. In DetNet, we add a sigmoid layer after the FC layer to get classification scores for multiple categories. ResNet50 [31] pre-trained on ImageNet dataset with reduced FC layer is adopted as backbone network to train DetNet for classification task. DetNet is used as a backbone network followed by dilated multi-scale feature layers of SSD. Since, DetNet is explicitly designed for object detection to preserve the spatial resolution, it improves the detection accuracy of SSD. Other settings for object detector training are same as in [22]. All experiments are conducted on NVIDIA GeForce TITAN XP 4 Parallel GPUs.

F. COMPARISON WITH STATE-OF-THE-ART METHODS

Table 2 and Table 4 show the results in terms of AP and mean AP metrics for proposed WSOD method, SSD (trained by pseudo ground-truths by proposed WSOD) and state-of-the-art methods on PASCAL VOC2007 and

TABLE 4. Comparison (AP in %) on PASCAL VOC2012 test set.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Jie *et al.* [19]	60.8	54.2	34.1	14.9	13.1	54.3	53.4	58.6	3.7	53.1	8.3	43.4	49.8	69.2	4.1	17.5	43.8	25.6	55	50.1	38.3
Tang *et al.* [14]	58.2	66	41.8	24.8	27.2	55.7	55.2	28.5	16.6	51	17.5	28.6	49.7	70.5	7.1	25.7	47.5	36.6	44.1	59.2	40.6
Li *et al.* [25]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	43.5
Our WSOD-NMS	51.9	61.6	41.5	25.1	9.8	58.9	57.5	58.8	9.4	35.2	46.2	55.1	44.3	67.4	31.7	37.2	48.1	42.5	43.2	50.8	43.8
Our WSOD-IC	53.8	62.7	42.2	26.1	15.9	59.9	58.5	59.7	13.8	36.1	46.8	56.3	46.3	69.8	32.9	37.5	51	42.9	44.3	51.9	45.4
SSD-NMS	54.7	62.7	43.8	27	17.5	60.7	60.8	60.8	15.7	37.9	47.5	58.3	46.6	70.3	32.8	40.7	50.3	44.3	45.7	53.7	46.5
SSD-IC	55.7	64.8	43.8	27	18.2	60.7	61.3	60.9	16.8	37.8	47.5	58.3	47.8	70.3	32.8	40.7	51.6	45.1	45.7	53.9	47.0

TABLE 5. Comparison (CorLoc in %) on PASCAL VOC2012 trainval set.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	Mean
Jie *et al.* [19]	82.4	68.1	54.5	38.9	35.9	84.7	73.1	64.8	17.1	78.3	22.5	57	70.8	86.6	18.7	49.7	80.7	45.3	70.1	77.3	58.8
Tang *et al.* [14]	77.2	83	62.1	55	49.3	83	75.8	37.7	43.2	81.6	46.8	42.9	73.3	90.3	21.4	56.7	84.4	55	62.9	82.5	63.2
Li *et al.* [25]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.9
Our WSOD-NMS	78.5	85.7	64.8	55.5	22.1	78.9	77	65.9	26.8	79	48.6	68.4	75.5	88.1	61.8	63.2	84.8	60.3	65.1	78	66.4
Our WSOD-IC	81.2	86.9	65.8	56.7	25.7	79.8	77.8	65.9	30.8	81.2	49.9	69.8	76.8	89.8	63.8	63.8	85.7	60.8	66.8	78.8	68.0
SSD-NMS	79.8	87.4	66.1	57.8	26.8	79.7	78.6	66.3	28.3	81.7	50.7	71.4	76.8	89.8	65.3	64.5	86.2	63.2	67.4	79.4	68.3
SSD-IC	81.2	88.5	66.9	58.9	29.3	80.1	78.8	67.2	30.8	81.9	51.2	71.5	77.2	89.8	65.5	64.5	86.4	64.6	67.4	79.9	69.0

PASCAL VOC 2012 test sets respectively. In all experiments, the results are also compared between IC and standard NMS for both weakly supervised (WSOD-IC and WSOD-NMS) and fully supervised via pseudo ground-truths (SSD-IC and SSD-NMS) settings. Note that SSD-NMS is trained with the pseudo ground-truths generated by WSOD-NMS, and SSD-IC is trained with the pseudo ground-truths of WSOD-IC. Significant improvement in detection has been shown by our method compared to other WSOD methods. From Table 2, it can be observed that, the proposed method WSOD-NMS achieved highest mAP of 50.9% among all compared methods. The proposed IC method further improves mAP by 1.7% with overall 52.6% mAP. On PASCAL VOC2007 and VOC2012 test sets our method outperformed other approaches on 7 and 10 classes respectively. Our method achieves 32.5% boost in AP on ''plant'' class compared to [14] and [25], and 7.2% gain in AP on ''person'' class compared to [25], which is an enormous improvement as shown in Table 2. Moreover, WSOD-IC further improves the accuracy of WSOD-NMS almost for all classes on both datasets. It can be observed that SSD-IC has 1.5% mAP further gain than SSD-NMS. More gain is achieved in weakly supervised setting, this is due to the fact that in WSOD no regression of bounding boxes is performed with respect to ground-truths and candidate proposals are more dispersed in WSOD than in fully supervised object detection. The proposed IC method is very effective in eliminating duplicate detections without harming the true positive detections. Since [25] has the highest mAP for PASCAL VOC2007 among all compared state-of-the-art methods as illustrated in Table 2, however, our method has 0.7% (WSOD-NMS) and 2.4% (WSOD-IC) improvement in mAP as compared to [25]. Table 4 shows 1.9% improved mAP by our WSOD-IC for PASCAL VOC2012 than [25]. WSOD-IC has 1.6% improved mAP than WSOD-NMS for PASCAL VOC2012 test set.

Table 3 and Table 5 illustrate the results in terms of CorLoc on PASCAL VOC2007 and PASCAL VOC2012 trainval sets by the proposed method, SSD and state-of-the-art methods. WSOD-NMS has 0.6% and WSOD-IC shows 1.9% performance improvement in terms of average CorLoc as compared to [25] on PASCAL VOC2007. In comparison with Tang *et al.* [14] mean CorLoc score by proposed WSOD-NMS is quite high, our method outperforms by 6.8% on PASCAL VOC2007 and 3.2% on PASCAL VOC2012. Unexpectedly, on PASCAL VOC2012 trainval set our method WSOD-NMS has 1.5% less mean CorLoc compared to [25]. This is due to the declined performance on classes with relatively smaller size objects particularly ''bottle'' class. However, our method WSOD-IC further achieves 1.6% increase in mean CorLoc than WSOD-NMS.

SSD-IC achieves state-of-the-art results with 72.3% and 69% mean CorLoc for PASCAL VOC2007 and PASCAL VOC2012 respectively as reported in Table 3 and Table 5. Furthermore, PASCAL VOC2007 and VOC2012 trainval sets have 6 and 12 classes correspondingly with maximum CorLoc achieved by the proposed WSOD-NMS method compared to state-of-the-art methods (given that per-category mean scores of [25] for PASCAL VOC2012 dataset are not available). However, we have observed lower performance for certain classes such as ''bottle'' and ''chair'' on both datasets. This is primarily due to the small size of objects and

TABLE 6. Inference time and accuracy in terms of FPS and mAP (%) respectively on PASCAL VOC2007 and PASCAL VOC2012 test sets.

Method		Backbone	VOC2007		VOC2012	
			mAP	FPS	mAP	FPS
Weakly supervised	Our WSOD-NMS	VGG16	50.9	0.85	43.8	0.89
	Our WSOD-IC	VGG16	52.6	0.72	45.4	0.84
Fully supervised	SSD-NMS	VGG16	53.5	46	46.5	46
	SSD-IC	VGG16	55.0	44	47.0	45
	SSD-NMS	DetNet	54.3	26	47.1	26
	SSD-IC	DetNet	55.7	24	48.1	25

not so prominent complementary features of objects of these classes.

The overall results show substantial performance boost by the proposed method due to the use of complementary features, learning regions with minimum entropy, and IC as compared to state-of-the-art methods. Out of 20 categories our method achieved top CorLoc for 12 categories on PASCAL VOC2012 trainval set as shown in Table 5. Nevertheless, the quantitative results have demonstrated relatively low AP on certain classes such as ''bottle'', ''boat'', and ''chair'' than other categories. This is because of the very small size instances, heavy occlusions and overlapping of the instances, and fewer complementary semantic features of these categories. In our second experimental setting, we have used 300×300 resolution of SSD, further improvement is expected by using SSD with 512×512 resolution.

G. INFERENCE TIME

We report the inference time of proposed WSOD-IC and WSOD-NMS with VGG16 backbone in Table 6. Proposals generation time for WSOD is not considered in inference time. Additionally, we compare SSD (trained with pseudo ground-truths from proposed WSOD) with VGG16 and DetNet backbones separately in terms of inference time and accuray on PASCAL VOC2007 and PASCAL VOC2012 datasets. The purpose of using different backbone networks is to further optimize the detection results while residing in the weakly supervised regime. Table 6 shows the trade-off between accuracy and inference time with different settings of object detectors in terms of FPS and mAP, all methods with batch size 1.

On PASCAL VOC2007, SSD-NMS with DetNet backbone achieved 0.8% mAP improvement as compared to SSD-NMS with VGG16 backbone but with a huge drop of speed. Likewise, SSD-IC with DetNet backbone has 1.1% increment in mAP compared to SSD-IC with VGG16 on PASCAL VOC2012 dataset. The compromise between speed and precision is due to extra stages included in DetNet to maintain high resolution feature maps and large receptive field, both of which are important for the object detection task. Inference times by proposed WSOD either with NMS or IC are reasonably close for both datasets. Similarly, the comparable trends are observed between the inference times of SSD-NMS and SSD-IC with VGG16 backbone, and also in the case of DetNet backbone on both datasets.

H. QUALITATIVE RESULTS

Figure 3 shows the detection results on PASCAL VOC2007 and PASCAL VOC2012 test sets by our WSOD method. It is observed that the proposed method effectively detects the whole objects. Moreover, it can also be noticed that false detections are mainly due to the larger size initial proposals generated by selective search [27]. The initialization of proposals is very important for final detections, and our method efficiently extracts the proposals with almost covering the whole object.

Only a few false detections with object parts have been noticed which are primarily due to occlusions. The other observed false detections are in the case of object cluster with the same category (a group of objects in an image located very close to each other which belongs to the same object class). In WSOD, there likely exist some redundant instance clusters for a particular object in an object cluster. This is the state when for an object instance in an object cluster the instance cluster ($X_i^{j_m}$) with tight region proposals and the instance cluster(s) with relatively loose region proposals but high confidence scores have IoU more than the threshold (δ_{out}) between them. In such a circumstance, all the instance clusters can have a significantly high number of members (region proposals). It is worth mentioning that for such cases, IC functions analogous to NMS. This results in true positive detections along with the false duplicate detections. However, such false detections are observed for highly overlapped objects or adjacent objects with high similarity in color intensities. Note that region proposals are generated based on objectness by selective search.

Overall our method has shown an improved performance for WSOD. Discriminative feature learning is imperative in extracting distinct patterns for an object and guides the WSOD network (network C in Figure 1) in enhancing the detection performance. The final detections inferred by the proposed WSOD method as presented in Figure 3 has validated considerably good region proposal mining in most cases. We have observed that most WSOD methods have two common problems, 1) detection of object parts and 2) detections including adjacent object parts. These problems are effectively tackled by our method through the discriminative feature learning approach and instance clustering, respectively. Figure 3 illustrates the examples of detection outputs where the objects are partially occluded by neighboring object instances of same category. Feeding the whole

FIGURE 3. Detection results: green bounding boxes indicate true detections (IoU > 0.5), and red bounding boxes show false detections (IoU < 0.5).

object features encourages the final classifier to learn categories based on whole object features instead of learning only class-specific distinct features. Hence, by learning complementary features, our method can tackle the problem of partially occluded objects by multiple adjacent instances of the same category.

I. ABLATION STUDY

In the proposed IC method, we use the threshold δ_{out} as defined in (1). We use this threshold since it achieves the highest performance gain for IC as compared to other defined thresholds. To analyze IC with different thresholds, we define four additional thresholds ($\delta_{out(a)}$, $\delta_{out(b)}$, $\delta_{out(c)}$, and $\delta_{out(d)}$). We investigate the effect of each threshold in detecting duplicate detections for the same instance and analyze all thresholds with qualitative and quantitative results. Figure 4 demonstrates the intermediate and final results of the IC with different thresholds. Following are the defined thresholds, that are statistically and experimentally observed as effective thresholds for detecting the outlier clusters with IC.

1) Calculate the difference of mean and standard deviation of set $\boldsymbol{Z}$ and then apply floor function on it as defined in (5).

$$\delta_{out(a)} = \lfloor \mu_Z - \sigma_Z \rfloor \tag{5}$$

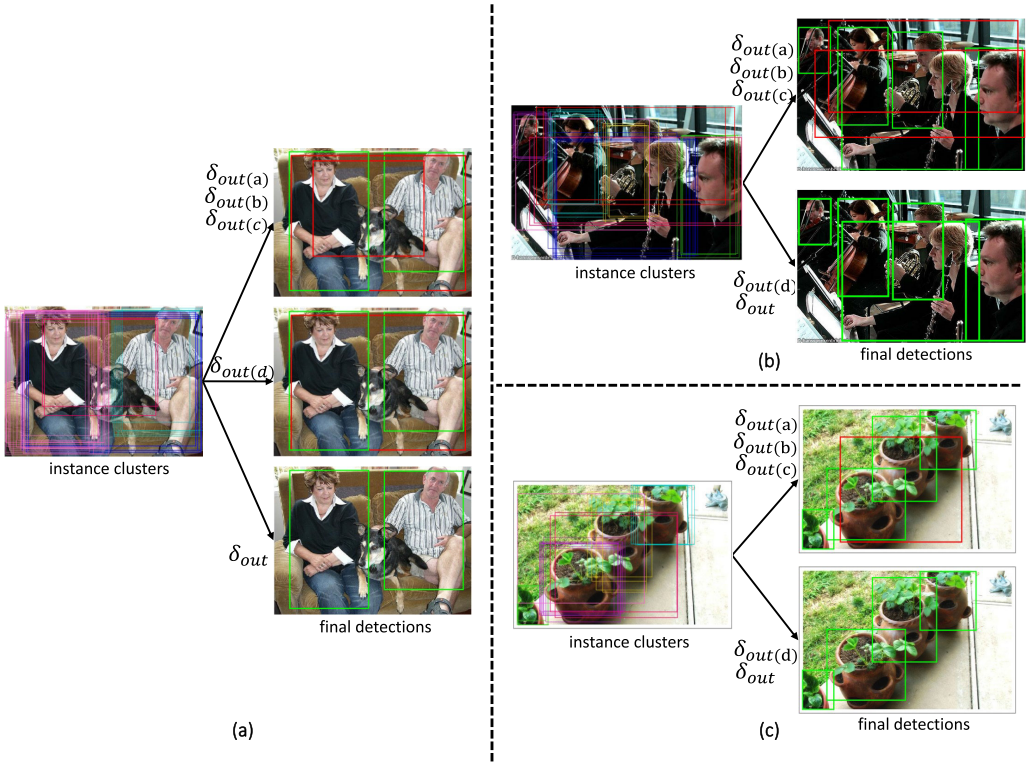


FIGURE 4. Qualitative results (proposals with same color indicate instance clusters): intermediate and final regions by IC inferred with different outlier thresholds. In final detections, green bounding boxes show true detections and red bounding boxes indicate false detections. These detection results are demonstrated for a single object category.

TABLE 7. Comparison of thresholds for the proposed IC method in terms of mAP (%) on PASCAL VOC2007 and PASCAL VOC2012 test sets.

Threshold	mAP	
	VOC2007	VOC2012
δ_{out}	52.6	45.4
$\delta_{out(a)}$	51.2	44.2
$\delta_{out(b)}$	51.4	44.4
$\delta_{out(c)}$	51.3	44.5
$\delta_{out(d)}$	52.1	45.1

2) Compute the floor function on difference of mean and standard deviation of set $\boldsymbol{Y}$ as defined in (6).

$$\delta_{out(b)} = \lfloor \mu_Y - \sigma_Y \rfloor \tag{6}$$

3) Calculate the standard deviation of set $\boldsymbol{Y}$ and then apply floor function as defined in (7).

$$\delta_{out(c)} = \lfloor \sigma_Y \rfloor \tag{7}$$

4) Apply ceil function on the standard deviation of set $\boldsymbol{Y}$ as defined in (8).

$$\delta_{out(d)} = \lceil \sigma_Y \rceil \tag{8}$$

In Table 7, the quantitative results are shown to analyze the proposed IC with different thresholds on PASCAL VOC2007 and PASCAL VOC2012 test sets. The performance of thresholds $\delta_{out(a)}$, $\delta_{out(b)}$, and $\delta_{out(c)}$ is fairly comparable with 51.2%, 51.4%, and 51.3% mAP respectively. However, $\delta_{out(a)}$ has the least mAP caused by two main reasons, 1) smaller threshold value for clusters members with skewed distributions which results in no outlier removal, 2) in case of clusters members with symmetric distribution in set $\boldsymbol{Z}$, it eliminates the true positive clusters as well in some cases due to the result of a large threshold value. A similar performance pattern is observed on PASCAL VOC2012 dataset by these thresholds as shown in Figure 5. Threshold $\delta_{out(a)}$ has the lowest accuracy (44.2% mAP) compared to other thresholds. An increase of 0.2%, 0.3%, and 0.9% is obtained by $\delta_{out(b)}$, $\delta_{out(c)}$, and $\delta_{out(d)}$ respectively compared to $\delta_{out(a)}$. Thresholds $\delta_{out(b)}$ and $\delta_{out(c)}$ are effective in removing outlier clusters in cases where outlier clusters have quite few members compared to the true positive clusters. It is observed that the threshold values computed by $\delta_{out(a)}$ and $\delta_{out(c)}$ are smaller for most cases with skewed distribution and not enough to remove all outlier clusters. The comparison of performance patterns by all defined thresholds with IC on both datasets is illustrated in Figure 5.

Overall experimental results show that the thresholds $\delta_{out(a)}$, $\delta_{out(b)}$, and $\delta_{out(c)}$ are less effective in detecting outlier clusters as compared to δ_{out} and $\delta_{out(d)}$ as also shown in Figure 4. However, $\delta_{out(d)}$ is observed less effective to remove outlier clusters in case of relatively skewed distribution that even persists in set $\boldsymbol{Y}$. In Figure 4 (a), the members of instance clusters have considerable asymmetrical distribution, in such cases, the only successful threshold is δ_{out} in

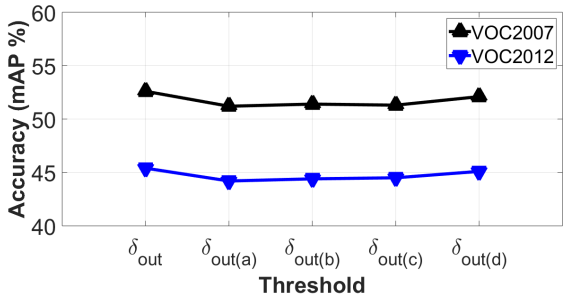


FIGURE 5. Performance trends of IC with different thresholds on PASCAL VOC2007 and PASCAL VOC2012 test sets.

removing all outliers. Figure 4 (b) and (c) are examples of relatively less asymmetrical distribution of members of clusters. The thresholds δ_{out} and $\delta_{out(d)}$ eliminate all outlier clusters, however, $\delta_{out(a)}$, $\delta_{out(b)}$, and $\delta_{out(c)}$ partially remove outliers. We have observed that the performance of the proposed IC method with any of the defined thresholds is better than NMS. The threshold δ_{out} has been observed most effective (with 52.6% mAP for VOC2007 and 45.4% mAP for VOC2012) in detecting outliers while preserving the true positive proposals among other compared thresholds.

V. CONCLUSION

In this paper, we have proposed a WSOD method based on complementary feature learning and instance clustering. We have trained FuCN to get discriminative features from two complementary networks and concatenated with the features of the third classifier to make it learn categories with whole object features. After acquiring the proposal scores by element-wise product of score matrices from classification and detection streams, we refine the proposals via clustering. Our instance clustering method is efficient in mining proposals for multiple instances in the multi-label images. We have identified that WSOD methods suffer from duplicate detection problem for the same instance, the proposed IC method filters such detections as outliers and leverages the detection performance by removing them. The results have shown improved detection performance by the proposed WSOD method compared to the state-of-the-art WSOD methods.

REFERENCES

[1] E. Sangineto, M. Nabi, D. Culibrk, and N. Sebe, "Self paced deep learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 712–725, Mar. 2019.

[2] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1325–1334.

[3] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2846–2854.

[4] X. Wang, Z. Zhu, C. Yao, and X. Bai, "Relaxed multiple-instance SVM with application to object discovery," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1224–1232.

[5] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2843–2851.

[6] Z. Zhang, S. Qiao, C. Xie, W. Shen, B. Wang, and A. L. Yuille, "Single-shot object detection with enriched semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5813–5821.

[7] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognit.*, vol. 74, pp. 15–24, Feb. 2018.

[8] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 577–584.

[9] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[10] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 189–203, Jan. 2017.

[11] M. Shi, H. Caesar, and V. Ferrari, "Weakly supervised object localization using things and stuff transfer," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3381–3390.

[12] W. Ren, K. Huang, D. Tao, and T. Tan, "Weakly supervised large scale object localization with multiple instance learning and bag splitting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 405–416, Feb. 2016.

[13] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, "Contextlocnet: Context-aware deep network models for weakly supervised localization," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 350–365.

[14] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, and A. Yuille, "PCL: Proposal cluster learning for weakly supervised object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 176–191, Jan. 2020.

[15] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1568–1576.

[16] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *Int. J. Comput. Vis.*, vol. 100, no. 3, pp. 275–293, Dec. 2012.

[17] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.

[18] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang, "Weakly supervised object localization with progressive domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3512–3520.

[19] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu, "Deep self-taught learning for weakly supervised object localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1377–1385.

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

[21] Y. Shen, R. Ji, S. Zhang, W. Zuo, and Y. Wang, "Generative adversarial learning towards fast weakly supervised detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5764–5773.

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[23] D. Zhang, J. Han, L. Zhao, and D. Meng, "Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework," *Int. J. Comput. Vis.*, vol. 127, no. 4, pp. 363–380, Apr. 2019.

[24] Y. Shen, R. Ji, Y. Wang, Y. Wu, and L. Cao, "Cyclic guidance for weakly supervised joint detection and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 697–707.

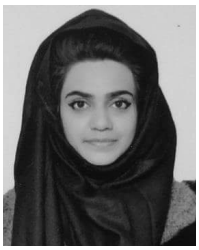
[25] X. Li, M. Kan, S. Shan, and X. Chen, "Weakly supervised object detection with segmentation collaboration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9735–9744.

[26] D. Zhang, J. Han, L. Zhao, and T. Zhao, "From discriminant to complete: Reinforcement searching-agent learning for weakly supervised object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 21, 2020, doi: 10.1109/TNNLS.2020.2969483.

[27] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[29] 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.
[30] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "DetNet: A backbone network for object detection," 2018, *arXiv:1804.06215*. [Online]. Available: http://arxiv.org/abs/1804.06215
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.



MEHWISH AWAN received the M.C.S. degree in computer science (major in machine learning and image processing) from Arid Agriculture University, Rawalpindi, Pakistan, in 2013, and the M.S. degree in computer science (major in machine learning and image processing) from COMSATS University, Islamabad, Pakistan, in 2016. She is currently pursuing the Ph.D. degree in computer engineering with the Media System Laboratory, Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, South Korea. Her major research interests include computer vision with weakly supervised and unsupervised learning approaches particularly in vision tasks for autonomous driving (object detection and semantic segmentation), and other video analysis applications.

JITAE SHIN (Member, IEEE) received the B.S. degree from Seoul National University, in 1986, the M.S. degree from the Korea Advanced Institute of Science and Technology, in 1988, and the M.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, in 1998 and 2001, respectively. He was with Korea Electric Power Corporation and the Korea Atomic Energy Research Institute for a period of eight years. Since 2002, he has been a Professor with the School of Electronic and Electrical Engineering, Sungkyunkwan University, Suwon, South Korea. His current research interests include image/video signal processing, deep learning applications, medical image processing, and video transmission over wireless/mobile communication systems.