Few Shot Transfer Active Learning for Logo Detection in Sports Video

Hang Su[a,*], Qiu Guoping[a,b,c,d,e]

[a] *College of Electronic and Information Engineering, Shenzhen University, Shenzhen 518052, China*
[b] *Guangdong Key Lab for Intelligent Information Processing, Shenzhen University, Shenzhen 518052, China*
[c] *Shenzhen Institute of AI and Robotics for Society, Shenzhen 518172, China*
[d] *Pengcheng Laboratory, Shenzhen 518055, China*
[e] *School of Computer Science, University of Nottingham, Nottingham NG8 1BB, UK*

Abstract

We exploit the power of deep convolutional neural network (DCNN) and take advantage of established datasets from existing applications to develop a deep transfer active learning (DTAL) algorithm to select the most valuable samples such that we can label the smallest number of samples to achieve the maximum performance improvements in training video object detection models. By exploiting possible shared common deep feature space between static dataset and video dataset through transfer learning based on highly adaptable DCNN features, DTAL implements a diversity based active learning to select the most informative samples from a sequence of similar image frames for video object detection. We have successfully applied the new DTAL algorithm to implement active learning for logo detection from live streaming sports videos as well as pedestrian and face detection from video data. We show that DTAL is a better active learning method outperforming state of the art deep learning based active learning techniques. In addition, we contribute one of the largest video based logo datasets, sports match video logo (SMVL), to facilitate research in logo detection in general and in the applications of transfer learning and active learning to exploit static object detection datasets for video object detection in

*Corresponding author
Email address: `suhang@szu.edu.cn` (Hang Su)

particular.

Keywords: Logo Detection, Deep Learning, Active Learning, Transfer Learning, Video Dataset, Object Detection

1. Introduction

Product logos are ubiquitous in sports videos and there is great interest in determining when and where a logo has appeared on the screen. As new products are emerging all the time, a logo detector needs to constantly learn to recognize new logo classes. Furthermore, different sports venue may have very different physical layouts, lighting conditions and camera angle, the same logo may appear very differently in different sports matches both in terms of shape deformation and surface reluctance, causing any pre-trained logo detection models to fail or to perform poorly. One way to learn new logo classes and to improve detection accuracy on existing logo classes is to re-train the models with samples from the current event. To do so, it is necessary to label samples online. This means that it is crucial to use the minimum number of labelled samples to achieve the maximum improvements.

Active learning seeks the most informative or most valuable samples such that we need only to label the smallest number of samples to achieve the maximum performance improvements in training a machine learning model [1]. It is therefore most suited for tackling aforementioned online sports match logo detection problem. In this paper, we make use of deep learning architecture and adapt transfer learning principle to develop new techniques for implementing active learning in logo detection from video data.

Transfer learning aims to use existing data in other domains or applications to learn a model and adapt the model for the current application [2]. There are many such datasets [3, 4, 5, 6, 7, 8, 9, 10, 11] that may be used to construct a detector and then adapt it to develop active learning solutions to our sports match logo detection problem. These datasets may not contain exactly the logo classes we aim to detect, but a shared common feature space can be obtained.

Therefore we can initially treat the logo classes in the current video event as a meta-class, thus a meta detector can be designed to detect and classify un-seen logo objects on unlabelled data.

In this paper, we exploit the power of deep convolutional neural network (DCNN) and take advantage of existing datasets to develop a deep transfer active learning (DTAL) approach to sports video logo detection. Using existing static image logo datasets, we train a DCNN to equip it with powerful discriminative feature extraction and mata-class object detection capabilities. We then adapt this trained DCNN to implement an active learning solution to sports match logo detection task. By exploiting the discriminative features extracted by the DCNN and its meta object detection ability, the new DTAL algorithm uses a feature diversity criterion to select the most diverse set of valuable samples such that we can use the smallest number of such samples to train the best sports match logo detectors. It turns out the problem and challenges that exist in logo detection are much more generic. We will show that our new DTAL algorithm can be equally applied to other video object detection tasks such as pedestrian detection [12, 13] and face detection [14, 15].

We make the following contributions:

(1) To research and tackle many unique issues in logo detection from live streaming sport events including multiple instances, small objects and data redundancy, we first contribute a large scale sports video logo detection database called Sports Match Video Logo Dataset (SMVL) to facilitate research in logo detection in general and in the applications of transfer learning and active leaning to exploit static object detection datasets for video object detection in particular. To our knowledge, this is one of the largest video based logo detection datasets. The SMVL dataset is publicly available at `https://dataset2021.github.io/SMVL/`.

(2) We introduce a deep transfer active learning (DTAL) method to implement transfer learning based active learning for object detection in video datasets. By exploiting the possible shared common deep feature space between static dataset and video dataset through transfer learning based on highly adaptable DCNN features, the method implements a diversity based active learning to select the most informative samples from a sequence of similar image frames for video object detection.

(3) Specifically, DCNN trained with auxiliary static datasets are utilized to extract common features of video object detection datasets, thus enabling the construction of deep feature based object level active learning selection algorithm. The selection algorithm aggregates two criteria of different levels: (1) An instance level max-min metric measuring the deep feature similarities between labelled and unlabelled instances. (2) An image level spatial information metric measuring the overlapping ratios between the bounding boxes in labelled and unlabelled images.

We show that our new DTAL algorithm is a general method which is not only suitable for sports match logo detection but is equally effective for other object detection in video and have successfully applied DTAL to the tasks of pedestrian detection and face detection in videos.

2. Related Work

Logo Detection Dataset As a long-standing problem, logo detection task have been explored with different datasets and detection models. Recent state-of-the-art logo detection methods leverage generic object detection networks [16, 17, 18] with deep convolutional neural network features. However, fully-supervised learning methods requires object-level bounding box annotation, which brings the requirement for more logo datasets. Currently, there are various static logo detection datasets with bbox level annotations [3, 4, 5, 6, 7, 19, 8, 9, 10, 20, 11] . It is always insufficient for real world applications considering the inherently infinite growing logo designs and variations, and the data collection are labour costing process for logo datasets [11]. To address the problem of scalable logo detection, we propose a *SMVL* dataset using video data as data source. However, video based object detection datasets also have their own problems, including heavy annotation labour and redundant similar frames.

Video Object Detection Dataset There exist many video datasets provided for specific applications [21, 22, 23, 24, 25, 26, 27],. Bounding box level annotated object detection video datasets are also explored for various applications, including pedestrian detection [13], face identification [15], kitchen object detection [28], general object detection [29] and YouTube video face detection [30].

Compared with static image datasets [31, 32, 14], public domain video object detection datasets with analogous tasks [29, 30, 15] are limited in data scale, especially number of classes. To enable the advancement of video object detection, more effort is needed to better exploit information from video based datasets.

Active Learning There exists many research on object detection by active learning utilizing various information as acquisition scores [33, 34, 35] to select informative samples to be queried. In deep learning area, predicted losses [36], deep network classification scores [37, 38], localization-aware informations [39, 40] are explored. However, above criteria tends to generate similar acquisition scores on similar images, thus suffering from redundant labelling set on video data. Diversity based active learning [37, 41, 42] aims to select most diverse labelling samples to maximize the value of training batch, by querying in areas of the sample space where labels are sparse. It is achieved by formulating a feature space and select/distinguish the geometrically distant points as diverse samples for labelling.

Although diversity based methods could avoid the redundant selection on similar frames, current diversity based methods only applies to image classification where the whole images can be extracted as a single point in feature space to calculate the geometrical distances. In this work, to grant the diversity based active learning with the ability to localize the target instances, and to correctly establish a feature space for geometrical evaluation, a meta-class transfer learning method is exploited.

Transfer Learning Transfer learning is used to improve a learner from one

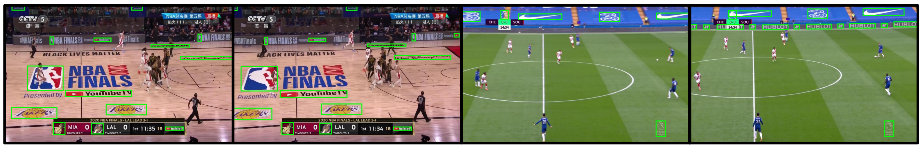


Figure 1: Examples of our SMVL dataset. Sequential video image frames are selected from basketball matches and soccer matches. The green boxes are the annotated logo bounding boxes that occurring multiple times in each of the images.

data domain by transferring information from a related data domain [43]. The need for transfer learning occurs when there is a limited supply of target training data, due to the data being rare, being expensive to collect and label, or being inaccessible [44]. Transfer learning has been successfully applied to many machine learning applications [45, 46, 47, 48, 49, 50]. Specifically, homogeneous transfer learning approaches are developed and proposed for handling the situations where the domains are of the same feature space [2]. An alternative view of homogeneous transfer learning environment is that two datasets exist in different sub-domains are linked by a high-level common domain [44].

In this work, a meta-class category is introduced to exploit the high-level common feature space of static and video dataset object level deep feature.

3. Sports Match Video Logo Dataset

We first contribute asports video logo detection database Sports Match Video Logo (SMVL) to facilitate research in logo detection in general and in the applications of transfer learning and active leaning to exploit static object detection datasets for video object detection in particular. It is large scale in image number, logo classes and object instances compared with existing publically available logo detection datasets. Example images are shown in Figure.1.

3.1. Dataset Construction

Data Selection Popular basketball and soccer match videos are selected as the data source.The dataset contains videos from 40 live matches, 20 basketball and 20 soccer games. From each match, 5 video clips of around 5 to 10 seconds are extracted from various camera angles. Frame images are extracted from the clips at a rate of 24 frames per second. The dataset contains 199 video clips with around 120 images per clip, resulting in 24,348 frame images and 223 logo classes in total. The resolution of the frames is 3840 × 2160 pixels.

Annotation At data annotation stage, all collected images are manually annotated with logo classes and bounding boxes. At an initial annotation stage, 30 workers are employed who spent a total of 800 hours to manually label the images. Then at the data verification and refinement stage, the dataset was verified manually by multiple workers and refined multiple times, costing 20 days in the correction/verification loop.

Training and Test Data Split For performance evaluation, we randomly choose 12,369 images as testing data. In the active learning context, we set the remaining 11,979 images as active learning image pool, where annotated data will not be utilized unless selected as training data. This setting is to simulate the active learning task when given a data pool across all clips, and aims to choose training batches to achieve the best training performances with least labelling effort.

3.2. Properties of SMVL

Small and multiple Logos: In real world live streaming of sports matches, brand logo advertisements often appear in multiple areas in the video simultaneously, including seating areas, the playing field, athletes' uniform and video post-processing captions. Also in real world sports broadcasting, the cameras will focus on the actual playing actions, often resulting in the advertisement logos appear on the edge of the image and far away from the cameras. This brings two notable features to SMVL, that the average size of logo instances relative to the whole image is very small and there are more instances occurring per image (0.0036% and 15.06, respectively) compared with popular existing logo datasets, e.g., FlickrLogos-32 [5] (0.0916%, 1.60), WebLogo-2M [51] (0.0769%, 1.20) and Openlogo [11] (0.0417, 1.89).

Video Based Image Frames: Due to the nature of video data, the dataset is composed of continuous similar image frames. This property presents one of the most important challenge to the dataset for active learning because there exist high redundancy in the data which could harm the performances of active learning as will be shown in Sec.5.

Scalability: Since the data is collected from popular sports matches and large number of images could be extracted from the sports match videos, the data collecting process is relatively easy compared to manually collecting individual images. There are potentially unlimited data as sports matches are happening almost continuously. This permits the scalability of the data for further expansion, which is desired for deep learning model development. A similar attempt of collecting and annotating sports match video logos is presented in [8], however our dataset is 10 times larger compared to existing ones, both in terms of the logo classes and the number of images.

4. Methodology

Problem definition In active learning object detection, we have access to an unlabelled data pool X_u. Using a certain selection strategy, we select a subset X_l from X_u as training set, and manually label X_l with bounding box level annotation. Then we train an object detection model based on a deep convolutional neural network. The goal is to choose an automatic selection strategy to select the elements for X_l to improve the performance of the detection model.

More specifically, for each sample x of $x \subset X_u$, a selection metric value $f(x)$ is assigned to determine it's *informativeness* or effectiveness for training the detection model. A higher $f(x)$ indicates that x is more valuable. Thus X_l is selected by ranking all samples in X_u based on the values of $f(x)$.

Specifically for video data based active learning in this work, we propose a much smaller X_l size compared with active learning tasks for static image datasets, e.g 0.2% to 1% training samples of the training data pool. This is due to the data redundancy by the continuity and similarity of the image frames in video datasets, thus less labelling training data should be required for video based data compared with static image data.

4.1. General Meta-Class Detection

Model Training To generate selection metric $f(x)$, an initial meta-class detection model is needed, which is able to extract useful deep features and output logo location information from unlabelled image pool X_u. However, in this paradigm there are no labelled training samples to initialize the model.

To address this issue, we utilize a sequential transfer learning strategy by first deploying an existing static object detection dataset as auxiliary data. This is supported by the ability of the DCNN models to adapt to various datasets, as shown in the cross-dataset learning paradigm [52, 53]. A model is first trained with static datasets to acquire meta-class feature extraction and object localization ability for both static dataset and video dataset. This strategy echoes the principle of common domain transfer learning, where two datasets in different sub-domains, in this case static dataset and video dataset, share a high-level common feature domain [44].

Model Architecture The DCNN model takes the auxiliary data as input and splits into two branches at the fully-connected classification layer. Specifically, the ordinary FC layer classifies specific logo classes with the classification loss $\mathcal{L}_{class}$, while a meta-class layer outputs the probability of the object belonging to a target meta-class with the meta-class loss $\mathcal{L}_{logo}$. The loss function of the DCNN model is:

$$\mathcal{L} = \mathcal{L}_{loc} + \lambda_1 \mathcal{L}_{class} + \lambda_2 \mathcal{L}_{logo} \tag{1}$$

where $\mathcal{L}_{\text{loc}}$ represents the bounding box regression loss of the detection model, the hyper-parameters λ_1 and λ_2 control the relative importance of the two loss terms.

By training this two branch architecture with auxiliary general object datasets, the convolutional layers of the model could learn discriminative features, and learn to detect target meta-class instances despite the auxiliary dataset does not contain the same classes as those in the target video dataset. This will not only enable the model to generate discriminate features at the object level but also equip it with the ability of localizing unseen meta-class instances in the

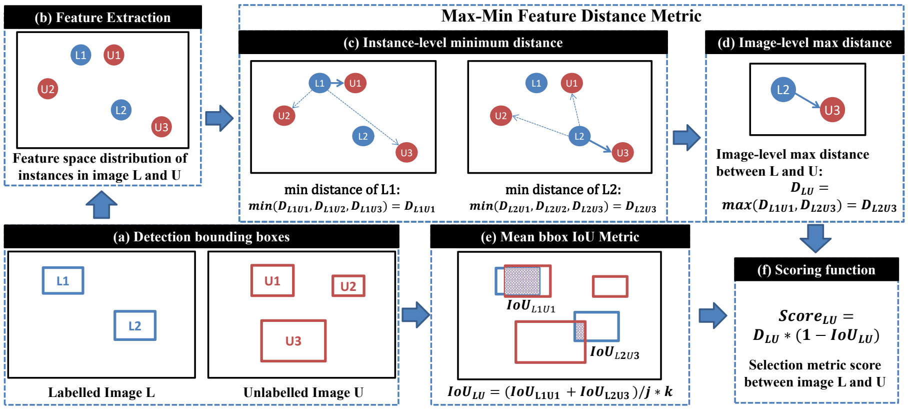

Max-Min Feature Distance Metric
(b) Feature Extraction
L1
U1
U2
L2
U3
Feature space distribution of instances in image L and U
(c) Instance-level minimum distance
min distance of L1:
$min(D_{L1U1}, D_{L1U2}, D_{L1U3}) = D_{L1U1}$
min distance of L2:
$min(D_{L2U1}, D_{L2U2}, D_{L2U3}) = D_{L2U3}$
(d) Image-level max distance
Image-level max distance between L and U:
$D_{LU} = max(D_{L1U1}, D_{L2U3}) = D_{L2U3}$
(a) Detection bounding boxes
Labelled Image L
Unlabelled Image U
(e) Mean bbox IoU Metric
IoU_{L1U1}
IoU_{L2U3}
$IoU_{LU} = (IoU_{L1U1} + IoU_{L2U3})/j * k$
(f) Scoring function
$Score_{LU} = D_{LU} * (1 - IoU_{LU})$
Selection metric score between image L and U

Figure 2: Calculation process of the scoring function between labelled image L and unlabelled image U.

image pool X_u. After feeding the image pool X_u into the meta-class detection model, the instance-level latent vectors, *meta-class scores* of each instance and detection bounding boxes are obtained, which can then be used to construct sample selection metric $f(x)$.

4.2. Selection Metric

Our method aims to select the most diverse instances and images between the labelled training batch and the unlabelled image pool to maximize the informativeness of labelled batch. The process of constructing the selection metric is shown in Fig.2. Meta-class detector is first applied to obtain detection bounding boxes (Fig.2 (a)) and instance-level feature distribution in shared meta-class feature space (Fig.2 (b)).

Max-Min Feature Distance Metric We utilize the instance-level latent vectors to calculate the distance between the labelled image object features and un-labelled image object features. A further distance represents the un-labelled instance is more diverse from the labelled instance, thus it is more valuable. As each image contains multiple object instances, the selection metric is determined in a max-min fashion. Given a selected image L and a un-selected image U, each containing j and k object instances respectively, first we calculate:

$$D_{min_{L_1}} = \min(D_{L_1U_1}, D_{L_1U_2}...D_{L_1U_k}) \tag{2}$$

where $D_{L_1U_x}$ represents the Euclidean distance between one of the instance latent vectors v_{L_1} in image L with each instance vector v_{U_x} in image U, as shown in Fig.2 (c). $D_{min_{L_1}}$ is the minimum distance between one of the instances in L and every instance in U, which means that this minimum distance is determined by the object instance in U that is the most similar to one of the instance v_{L_1} in L. Then we calculate:

$$D_{LU} = \max(D_{min_{L_1}}, D_{min_{L_2}} ... D_{min_{L_j}}) \tag{3}$$

where D_{LU} is the max distance of every minimum instance vector distance $D_{min_{L_x}}$, as shown in Fig.2 (d). It represents the furthest distance between every instances in image L and image U in the instance-level feature space. We use this max-min latent vector distance as the distance score between the selected images and un-selected images.

Max-Min feature distance metric alone would still select neighbouring video frames when some of the logo instances generate higher distance values. We assume this is caused by relatively higher feature values due to specific noise instances or drastic image change. This noise is further magnified by the nature of max-min instance based selection method, where the value of a single max-distance instance will determine the selection criteria despite there are other instances in same image. To mitigate this effect and prevent redundant selection on continuous frames, we also propose a image level spatial based BBox IoU metric as a weight function for feature distance metric.

Mean Bounding Box IoU Metric With *meta-class* scores of each instances and detection bounding boxes, we can estimate the context and spatial similarity between different images. We first perform a non-maximum suppression with *meta-class scores* and bounding box data to obtain the effective bounding boxes of each image. The Intersection over Union (IoU) is calculated for every bounding box between two images to obtain the image pair's mean IoU. Specifically, for a selected image L and a un-selected image U, each containing j and k object instances respectively, the mean IoU is:

$$IoU_{LU} = mean(IoU_{b_{L_1}b_{U_1}}, IoU_{b_{L_1}b_{U_2}} ... IoU_{b_{L_j}b_{U_k}}) \tag{4}$$

where b_{L_x}, b_{U_x} represent each of the bounding boxes in image L and U, as shown in Fig.2 (e). A lower IoU_{LU} indicates more diverse image level spatial information between the image pairs, thus more likely to be selected.

Scoring Function and Algorithm After obtaining the max-min feature distance and detection bbox IoU of each image, the overall weighted distance metric score between image L and U is calculated as:

$$Score_{LU} = D_{LU} * (1 - IoU_{LU}) \tag{5}$$

where the selection metric score $Score_{LU}$ is higher when the max-min feature distance D_{LU} increases and mean bbox IoU IoU_{LU} decreases, as shown in Fig.2 (f).

After the selection metric between selected images and un-selected images are determined as $f = Score_{LU}$, the active learning image selection process is performed following Algorithm 1. For algorithm initial bootstrap, we sum the meta-class classification scores of all bounding boxes of each image and select the image with the highest sum as I_{boot}. A larger sum indicates more target instances and higher classification confidence. Note that a similar max-min selection method is also applied in the process.

5. Logo Detection Experiments

Performance metrics. For performance evaluation, we used the common Average Precision (AP) for individual classes, and the mean Average Precision (mAP) over all classes [54]. We considered a detection as being correct if the Intersection over Union (IoU) between the detected and ground-truth boxes exceeds 50%.

Implementation details. We build the model with ImageNet dataset pre-trained resnet101 [55] as deep convolutional layer architecture and a detection structure similar to Faster R-CNN [17]. [56]. We set the learning rate to 0.001 and the batch size to 4. For active learning labelling sizes, we tested different labelling size settings of 20, 30, ..., 100.

Algorithm 1: Training batch selection approach

```
Data: meta-class object detection model $D$, training batch size $k$,
  unlabelled image pool $\mathcal{U}$, labelled training batch $\mathcal{L}$, initial image $I_{boot}$.
Initialisation: Select the initial image $I_{boot}$ and move it from $\mathcal{U}$ to $\mathcal{L}$.
repeat
    forall image $L$ in $\mathcal{L}$ do
        forall image $U$ in $\mathcal{U}$ do
            Apply $D$ to calculate $f = Score_{LU}$ for every image pairs
              between $U$ and $L$, then the score matrix $M_{score}$ of
              $size(\mathcal{U}) \times size(\mathcal{L})$ is obtained;
        Get minimum score array $A_{min-score} = min(M_{score})$, which
          indicate the scores of most similar images in $\mathcal{U}$ for each images
          in $\mathcal{L}$;
    The maximum score image is selected by $max(A_{min-score})$,
      indicating the most diverse image in $\mathcal{U}$ across all image pairs, then
      move it to $\mathcal{L}$;
until $size(\mathcal{L}) = k$;
Return Updated labelled training data $\mathcal{L}$ of size k, Updated unlabelled
  data pool $\mathcal{U}$;
```

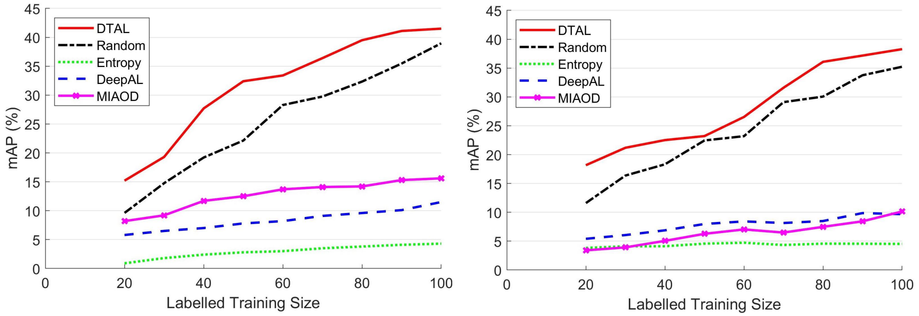
5.1. Baselines

We compared our DTAL method with four different active learning methods: (1) Random: Active learning training batches are sampled randomly from the unlabelled data pool. Specifically, we generated 3 different random training batches, the final result shows the average mAP of 3 random batch trained models. (2) Entropy: Entropy ranking uses the images entropy score as the informativeness metric to select active learning training batches. (3) DeepAL [34]: A state-of-the-art object detection active learning method based on deep learning pixel-level informativeness score. (4) MIAOD [57]: A state-of-the-art active learning object detection model with multiple instance learning (MIL) module. For MIAOD training setting, first an initial training set ${X_L}^0$ is randomly sampled, then a number of training sets ${X_L}^n$ are selected in each cycle. In this work, we applied the initial labelled set size of ${X_L}^0 = 10$, and selected sets size in each cycle to ${X_L}^n = 10$.

Note that the objective of this paper is on the development of active learning solutions which can label the smallest number of samples to achieve the maximum improvements in video object (logo) detection, we therefore focus on comparing active learning algorithms and only use basic static image based object detection models rather than advanced detection models that take full advantages of temporal relations of video frames in the experiments. To compare specifically the effectiveness of active learning selection methods and avoid performance change due to different detector models, we used the same detector models based on Resnet101 [55] and Faster R-CNN [17] on every comparison method. All settings were the same (if possible by design) for fair comparison. As a reference, the same detector model trained on all samples from the active learning image pool (11,979) will reach a performance of mAP = 92.10%.

5.2. Evaluations

Dataset and Setting. We utilised the public QMUL-OpenLogo logo detection dataset [11] as the auxiliary logo dataset, and our newly proposed SMVL as the active learning target and evaluation dataset.



(a) (b)

Figure 3: (a): Case 1 mAP results on SMVL dataset; (b): Case 3 mAP results on SMVL dataset.

Model Performances, case 1: Training from scratch. We first evaluate DTAL in situations where we have to construct the model from scratch. mAP performances of DTAL and 4 other methods are shown in Fig.3 (a), it is seen that DTAL can select better training batches for labelling and model training, with more advantages at smaller numbers of labelled samples, achieving 5.58% improvement at $k = 20$ and 10.26% improvement at $k = 50$. The performance improvement then drops as k increases, to only 2.54% at $k = 100$. This result indicates that our method could select more diverse and informative training images, while the advantage is reduced when labelled sample size increases. This is totally expected as more samples are labelled the chance of labelling enough diverse training data increases. Furthermore, the design objective of active learning is to train a high performance model when it is difficult or too costly to obtain large number of labelled samples. These results demonstrate that our DTAL algorithm has achieved the objective of active learning and is a better algorithm than the other 4 methods, including two state of the art deep learning based techniques.

Model Performances, case 2: Different video clip split for training and testing. We also evaluate DTAL in situations where the training frames and testing frames are from different video clips. In this setting, the training pool and testing data are from different video clips. For each of the sports matches containing 5 video clips, 3 video clips are used as training data pool and 2 clips are used as testing data. The mAP performances of DTAL and 4 other methods are shown in Fig.3 (b). DTAL can still select better training batches for labelling and model training, with less improvement compared with case 1, achieving 6.56% improvement at $k = 20$ and 3.05% improvement at $k = 100$. Note that the shared logo classes are reduced from 223 to 144 because of the absence of minor logo classes only occurring on training or testing sets.

Model Performances, case 3: Fine-tuning an existing model. We now evaluate DTAL for the situations where a model has already been trained with some pre-labelled data, and we would like to use DTAL to fine-tune the model with the objective of using the smallest number of labelled samples to achieve the best improvements. The original pre-trained model have a mAP of 39.41%. Results of fine-tune the model with $k = 100$ labelled samples are shown in Table.1. It is seen that DTAL can also be applied to fine-tuning an existing model, improving model performance by 16.43% with 100 labelled samples, again better than competitors.

Table 1: Case 3 mAP results on SMVL dataset.

DTAL	Random	DeepAL	MIAOD	Entropy
55.84%	52.58%	40.18%	41.48%	30.33%

Feature Distribution Analysis.

It is notable that DeepAL and MIAOD method output lower performances compared to even random selection. This is mainly due to the biased selection by similar acquisition scores on similar images. The latent feature distribution of the selected images are visualized by t-SNE, and shown in Fig.4. It is seen in the Figure that there are two main regions, bottom left and upper-right, representing the ’Soccer’ and the ’Basketball’ image clusters respectively.

It is evident that DeepAL method formed a cluster on the ’Basketball’ images, with only a few images in ’Soccer’. MIAOD method also formed a cluster on the ’Soccer’ images with more diverse examples exist in the ’Basketball’ area, however many of it’s images belong to a single video clip at the bottom.

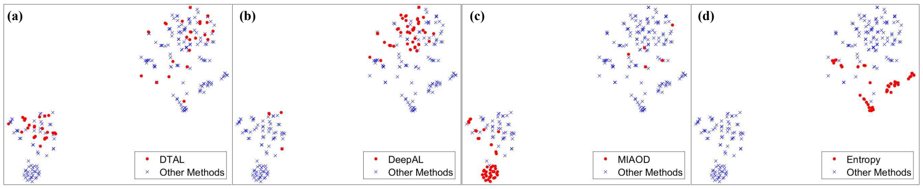


Figure 4: The feature distribution of selected training images by different methods, visualized by t-SNE method. The labelling size is set to $k = 50$. Training models from scratch.

the entropy method only selected the basketball images as training batches, and forming several clusters which represents several video clips, resulting in the lowest performances. Compared with the distributions above, our DTAL method can select training batches with a more diverse distribution across both 'Soccer' and 'Basketball' cluster areas, resulting in the best overall performances. Random method should have the most diverse distributions in theory, however it doesn't contain any informativeness criteria to assist active learning, thus it is not shown.

5.3. Analysis of Selection Metric

In this section, we evaluate 3 variations of our selection metric to assess the role of each component of the metric. For all selection metrics, we set $k = 50$ as base training batch for comparison. The performances of the models trained with different selection metric variations are shown in Table.2.

Table 2: More selection metrics. Main represents our main selection metric, random represents random selected samples. Feature distance only represents the selection metric without bbox IoU weighting. Bbox weight only represents the selection metric without feature distance. Bbox weight squared represents making the spatial information more important.

Main	32.4%
Random	22.5%
Feature distance only	29.6%
Bbox weight only	19.0%
Bbox weight squared	20.8%

Where *feature distance only* removes the bounding box IoU score, changing the final score metric function Eq.5 to:

$$Score_{LU} = D_{LU} \tag{6}$$

bbox weight only removes the latent vector score, changing the final score metric function Eq.5 to:

$$Score_{LU} = 1 - IoU_{LU} \tag{7}$$

bbox weight squared represents the case when the score metric function eq. 5 is changed to:

$$Score_{LU} = D_{LU} * (1 - IoU_{LU})^2 \tag{8}$$

The results show that the original Eq.5 score metric have the best performance. The *feature distance only* metric shows a lower performance after removing bounding box information as scoring criteria. *bbox weight only* shows significantly lower performance than the original metric by 13.4%, indicating the latent vector distance is more important in selecting informative images in deep active learning. *bbox weight squared* also shows significant decrease in model performance by 11.6%, it is also notable that both *bbox weight only* and *bbox weight squared* metric showing lower performance than simple random selection. We assume this is because the bounding box IoU metric by itself is a selection metric that limits the variation of the selected training batch, as it only considers bounding box spatial relation between different images instead of overall image content.

6. Extension to other Applications

As an active learning method designed for video object detection, DTAL can also be applied to various video object detection applications. Here we explore two applications, namely *pedestrian detection in video data* and *face identification in video data.* To better evaluate the performance of our method for different applications, additional datasets are introduced. Other training and evaluation settings are the same as in Sec.5.

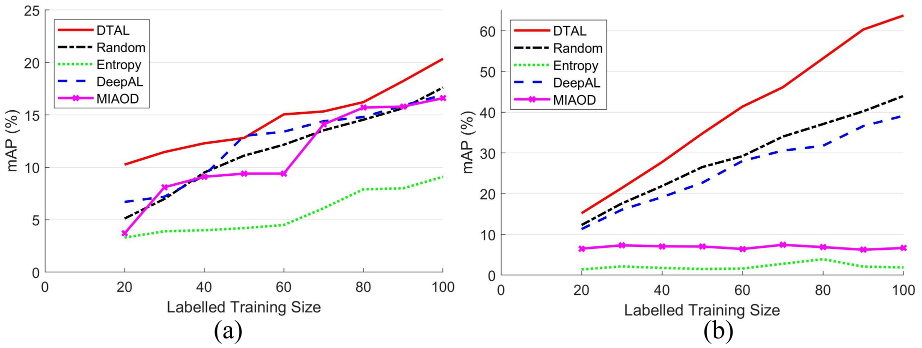


Figure 5: (a): mAP result of Caltech Pedestrian; (b): mAP result of YouTubeFaces.

6.1. Pedestrian Dataset

Dataset and setting For pedestrian video detection dataset, we selected BDD100k [12] as auxiliary dataset and Caltech Pedestrian Dataset [13] as active learning target and evaluation dataset. The BDD100k is a static street view object detection dataset with 10 object classes, where we select its most frequent ”pedestrian” class as the target meta-class. The Caltech Pedestrian Dataset [13] is a street view video dataset for pedestrian detection, with a single class ”pedestrian” and multiple-instance in each images. From BDD100k, we selected 22,065 images with labelled pedestrians. From Caltech, we selected 12,288 images as unlabelled data pool and 12,150 images as testing data, both containing ”pedestrian” and ”person” labelled instances.

Model Performances From Fig.5 (a), it is seen that our active learning selection metric can also select better training samples for the pedestrian datasets. Specifically, although the overall performance improvement is lower than that of SMVL dataset, our method still achieves 3.56% improvement compared with DeepAL (the best performer in the compared methods) for labelled size of $k = 20$ and 2.74% improvement compared with random selection (the best performer in the compared methods) for labelled size of $k = 100$.

6.2. Face Identification Dataset

Dataset and setting For video frame based face identification dataset, we selected CelebA [14] as auxiliary dataset, and YouTubeFaces Dataset [15] as the active learning target and evaluation dataset. CelebA is a static face identification dataset collected from various celebrity photos, where we implement a meta-class of ”face” for meta-class detection. YouTubeFaces is a video face identification dataset collected from youtube video data. Specially, YouTubeFaces only contains single person face in each frame. From CelebA, we selected 60,838 images with labelled faces of 2,360 identification classes. From YouTube Faces, we selected 29,326 images as unlabelled data pool and 9,728 images as testing data, containing 100 identification classes.

Selection Metric for Single Instance Dataset Due to the attributes of YouTubeFaces Dataset, where all images only contain a single target face and the target face occurs at the center of the image across all images, the max-min feature distance metric designed for multi-instance calculation and mean bbox IoU metric designed for bounding box spatial relation does not apply to this dataset. Therefore we implemented an instance-level feature clustering algorithm on this application, similar to image-level feature clustering method for key frame extraction [58]. Specifically, we first apply our DTAL main transfer active learning strategy, where an auxiliary-dataset-trained model is used to find meta-class feature space and extract instance-level features from unlabelled data pool. Then, instead of max-min metric, k-means clustering is applied to the extracted instance-level features. The cluster number is set equal to the number of identification classes. The nearest samples to the cluster center are selected iteratively as active learning training batches for each of the classes.

Model Performances From Fig.5 (b), it is seen that our active learning selection metric could also select better training batch for single instance face identification datasets. Specifically, our method achieves 2.90% improvement for labelled size of $k = 20$ and 19.74% improvement for labelled size of $k = 100$ compared with random selection. This indicates the meta-class transfer learning strategy can also apply to single instance sequential frame datasets for active learning with simple clustering selection metrics and still achieve better performance compared with existing active learning methods.

7. Conclusion

In this work, we presented a *deep transfer active learning (DTAL)* method for implementing active learning and have successfully applied it to the detection of logos, pedestrians and faces from video data. The method utilises existing static object detection datasets as auxiliary data to train an meta-class detection model that could both output instance-level deep features of shared meta-class and generate bounding boxes as image-level spatial information. The final metric aggregates the instance-level latent vector distance and image-level bounding box IoU to generate a selection score for informative and diverse active learning training batch. Empirical evaluations show the performance advantages of our method over state of the art deep learning based active learning methods.

A possible limitation of this method is that it's applications require auxiliary datasets to exploit the target meta-class feature space. It is also notable that the method is specifically designed for sequential video data but not for static data.

8. Acknowledgement

This work is partially supported by the Education Department of Guangdong Province, PR China, under project No 2019KZDZX1028, Guangdong Basic and Applied Basic Research Foundation (Grant 2019B151502001) and Shenzhen R&D Program (Grant JCYJ20200109105008228)

References

[1] B. Settles, Active learning literature survey, Tech. Rep. 1648, University of Wisconsin-Madison Department of Computer Sciences (2009).

[2] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proceedings of the IEEE 109 (1) (2020) 43–76.

[3] A. Joly, O. Buisson, Logo retrieval with a contrario visual query expansion, in: ACM International Conference on Multimedia, 2009, pp. 581–584.

[4] Y. Kalantidis, L. G. Pueyo, M. Trevisiol, R. van Zwol, Y. Avrithis, Scalable triangulation-based logo recognition, in: ACM International Conference on Multimedia Retrieval, 2011, p. 20.

[5] S. Romberg, L. G. Pueyo, R. Lienhart, R. Van Zwol, Scalable logo recognition in real-world images, in: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ACM, 2011, p. 25.

[6] S. Bianco, M. Buzzelli, D. Mazzini, R. Schettini, Deep learning for logo recognition, Neurocomputing 245 (2017) 23–30.

[7] A. Tüzkö., C. Herrmann., D. Manger., J. Beyerer., Open set logo detection and retrieval, in: Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Vol. 5, 2018, pp. 284–292.

[8] Y. Liao, X. Lu, C. Zhang, Y. Wang, Z. Tang, Mutual enhancement for detection of multiple logos in sports videos, in: IEEE International Conference on Computer Vision, 2017, pp. 4856–4865.

[9] H. Sahbi, L. Ballan, G. Serra, A. Del Bimbo, Context-dependent logo matching and recognition, IEEE Transactions on Image Processing 22 (3) (2012) 1018–1031.

[10] S. C. Hoi, X. Wu, H. Liu, Y. Wu, H. Wang, H. Xue, Q. Wu, Logo-net: Large-scale deep logo detection and brand recognition with deep region-based convolutional networks, arXiv preprint arXiv:1511.02462.

[11] H. Su, X. Zhu, S. Gong, Open logo detection challenge, in: British Machine Vision Conference, 2018, p. 16.

[12] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, T. Darrell, Bdd100k: A diverse driving video database with scalable annotation tooling, arXiv preprint arXiv:1805.04687 2 (5) (2018) 6.

[13] P. Dollár, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: A benchmark, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 304–311.

[14] Z. Liu, P. Luo, X. Wang, X. Tang, Large-scale celebfaces attributes (celeba) dataset, Retrieved August 15 (2018) (2018) 11.

[15] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, in: CVPR 2011, IEEE, 2011, pp. 529–534.

[16] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[17] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[18] J. Redmon, A. Farhadi, Yolo9000: better, faster, stronger, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7263–7271.

[19] F. N. Iandola, A. Shen, P. Gao, K. Keutzer, Deeplogo: Hitting logo recognition with the deep neural network hammer, arXiv preprint arXiv:1510.02131.

[20] H. Su, X. Zhu, S. Gong, Deep learning logo detection with data expansion by synthesising context, in: IEEE Winter Conference on Applications of Computer Vision, IEEE, 2017, pp. 530–539.

[21] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 724–732.

[22] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, P. Ishwar, Cdnet 2014: An expanded change detection benchmark dataset, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2014, pp. 387–394.

[23] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, G. Nebehay, R. Pflugfelder, The visual object tracking vot2015 challenge results, in: Proceedings of the IEEE international conference on computer vision workshops, 2015, pp. 1–23.

[24] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, K. Schindler, Motchallenge 2015: Towards a benchmark for multi-target tracking, arXiv preprint arXiv:1504.01942.

[25] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.

[26] H. Jhuang, H. Garrote, E. Poggio, T. Serre, T. Hmdb, A large video database for human motion recognition, in: Proc. of IEEE International Conference on Computer Vision, Vol. 4, 2011, p. 6.

[27] G. Awad, A. Butt, J. Fiscus, D. Joy, A. Delgado, W. Mcclinton, M. Michel, A. Smeaton, Y. Graham, W. Kraaij, et al., Trecvid 2017: evaluating ad-hoc and instance video search, events detection, video captioning, and hyperlinking, in: TREC Video Retrieval Evaluation (TRECVID), 2017.

[28] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al., Scaling egocentric vision: The epic-kitchens dataset, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 720–736.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.

[30] E. Real, J. Shlens, S. Mazzocchi, X. Pan, V. Vanhoucke, Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5296–5305.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2009.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[33] E. Haussmann, M. Fenzi, K. Chitta, J. Ivanecky, H. Xu, D. Roy, A. Mittel, N. Koumchatzky, C. Farabet, J. M. Alvarez, Scalable active learning for object detection, in: 2020 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2020, pp. 1430–1435.

[34] H. H. Aghdam, A. Gonzalez-Garcia, J. v. d. Weijer, A. M. López, Active learning for deep detection neural networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3672–3680.

[35] S. Roy, A. Unmesh, V. P. Namboodiri, Deep active learning for object detection., in: BMVC, Vol. 362, 2018, p. 91.

[36] D. Yoo, I. S. Kweon, Learning loss for active learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 93–102.

[37] O. Sener, S. Savarese, Active learning for convolutional neural networks: A core-set approach, arXiv preprint arXiv:1708.00489.

[38] C.-A. Brust, C. Käding, J. Denzler, Active learning for deep object detection, arXiv preprint arXiv:1809.09875.

[39] C.-C. Kao, T.-Y. Lee, P. Sen, M.-Y. Liu, Localization-aware active learning for object detection, in: Asian Conference on Computer Vision, Springer, 2018, pp. 506–522.

[40] J. Choi, I. Elezi, H.-J. Lee, C. Farabet, J. M. Alvarez, Active learning for deep object detection via probabilistic modeling, arXiv preprint arXiv:2103.16130.

[41] D. Gissin, S. Shalev-Shwartz, Discriminative active learning, arXiv preprint arXiv:1907.06347.

[42] S. Sinha, S. Ebrahimi, T. Darrell, Variational adversarial active learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5972–5981.

[43] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on knowledge and data engineering 22 (10) (2009) 1345–1359.

[44] K. Weiss, T. M. Khoshgoftaar, D. Wang, A survey of transfer learning, Journal of Big data 3 (1) (2016) 9.

[45] C. Wang, S. Mahadevan, Heterogeneous domain adaptation using manifold alignment, in: Twenty-Second International Joint Conference on Artificial Intelligence, 2011, pp. 1541–1546.

[46] L. Duan, D. Xu, I. Tsang, Learning with augmented features for heterogeneous domain adaptation, arXiv preprint arXiv:1206.4660.

[47] M. Harel, S. Mannor, Learning from multiple outlooks, arXiv preprint arXiv:1005.0027.

[48] J. Nam, W. Fu, S. Kim, T. Menzies, L. Tan, Heterogeneous defect prediction, IEEE Transactions on Software Engineering 44 (9) (2017) 874–896.

[49] J. T. Zhou, I. W. Tsang, S. J. Pan, M. Tan, Heterogeneous domain adaptation for multiple classes, in: Artificial Intelligence and Statistics, 2014, pp. 1095–1103.

[50] J. T. Zhou, S. J. Pan, I. W. Tsang, Y. Yan, Hybrid heterogeneous transfer learning through deep learning, in: Twenty-eighth AAAI conference on artificial intelligence, 2014.

[51] H. Su, S. Gong, X. Zhu, Weblogo-2m: Scalable logo detection by deep learning from the web, in: IEEE International Conference on Computer Vision Workshops, 2017, pp. 270–279.

[52] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, Y. Tian, Unsupervised cross-dataset transfer learning for person re-identification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1306–1315.

[53] T. Baltrušaitis, M. Mahmoud, P. Robinson, Cross-dataset learning and person-specific normalisation for automatic action unit detection, in: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Vol. 6, IEEE, 2015, pp. 1–6.

[54] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International journal of computer vision 88 (2) (2010) 303–338.

[55] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[56] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, International Conference on Learning Representations (2014) 13.

[57] T. Yuan, F. Wan, M. Fu, J. Liu, S. Xu, X. Ji, Q. Ye, Multiple instance active learning for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5330–5339.

[58] X. Li, B. Zhao, X. Lu, Key frame extraction in the summary space, IEEE transactions on cybernetics 48 (6) (2017) 1923–1934.