IET Computer Vision

Research Article

IET Journals
The Institution of Engineering and Technology

Heterogeneous domain adaptation method for video annotation

ISSN 1751-9632
Received on 15th May 2016
Revised 8th October 2016
Accepted on 25th October 2016
E-First on 29th November 2016
doi: 10.1049/iet-cvi.2016.0148
www.ietdl.org

Han Wang[1] ✉, *Xinxiao Wu*[2], *Yunde Jia*[2]

[1]*School of Information Science and Technology, Institute of Visual Media, Beijing Forestry University, Beijing, People's Republic of China*
[2]*School of Computer Science, Beijing Laboratory of Intelligent Information Technology, Beijing Institute of Technology, Beijing, People's Republic of China*

✉ *E-mail: wanghan@bjfu.edu.cn*

Abstract: In this study, the authors study the video annotation problem over heterogeneous domains, in which data from the image source domain and the video target domain is represented by heterogeneous features with different dimensions and physical meanings. A novel feature learning method, called heterogeneous discriminative analysis of canonical correlation (HDCC), is proposed to discover a common feature subspace in which heterogeneous features can be compared. The HDCC utilises discriminative information from the source domain as well as topology information from the target domain to learn two different projection matrices. By using these two matrices, heterogeneous data can be projected onto a common subspace and different features can be compared. They additionally design a group weighting learning framework for multi-domain adaptation to effectively leverage knowledge learned from the source domain. Under this framework, source domain images are organised in groups according to their semantic meanings, and different weights are assigned to these groups according to their relevancies to the target domain videos. Extensive experiments on the Columbia Consumer Video and Kodak datasets demonstrate the effectiveness of their HDCC and group weighting methods.

1 Introduction

In real-world application of consumer video annotation, many factors – such as amateur capturing, randomly annotation, or unconstraint environment – make it a challenging task to find desired videos from the Internet. It is often expensive and time-consuming to collect enough labelled videos to train well generalised classifiers for video annotation. To alleviate human from burdensome labelling work, transfer learning has attracted growing attention because it can learn robust classifiers with very few or even no labelled data from the target domain by leveraging a large amount of labelled data from existing domains. Recent advances in transfer learning [1] have shown encouraging progress in using existing source domain to build models for the target domain. Many researchers have tried to seek other sources of labelled data and transfer the related knowledge from these data to videos [2–5]. With the development of the Internet, large quantities of labelled image data can be easily obtained and can provide rich sources of information. By leveraging this information knowledge from web images can be transferred to consumer videos. However, most existing transfer learning methods assume that the source and the target domains share the same type of features with the same dimension. As such, they cannot well handle the problem of transferring knowledge between domains with different data dimensions, especially when the image source domain is constructed by static image features and the video target domain is constructed by spatial–temporal video features. The two domains are represented by heterogeneous features with different dimensions and physical meanings.

Realising the limitations of existing works, in this paper, we first propose a novel unsupervised heterogeneous domain adaptation (DA) method: namely, heterogeneous discriminative analysis of canonical correlation (HDCC), for heterogeneous feature adaptation. We borrow the idea of CC analysis (CCA) and propose a novel heterogeneous feature mapping method by preserving both topological structures of the target videos and discriminative information of the source images. Unlike traditional supervised CCA method which requires labelled information in both source and target domains. In this paper, we consider the case where labelled data only exists in the image source domain and no labelled training data is obtained from the target domain. Under this circumstance, HDCC seeks to leverage both unlabelled static and motion features of the target videos as well as labelled static information from the source images. HDCC learns two projection matrices by transferring discriminative knowledge from the labelled source domain to the unlabelled target domain, under which the intra-class variations are maximised and the inter-class variations are minimised. The topology of the target videos is also considered in HDCC to provide the symbiotic relationship of the heterogeneous features, simultaneously. By introducing these two projection matrices, static features from the source domain and motion features from the target domain can be projected onto a common feature subspace via their corresponding projection matrices. With HDCC, the label information in the source domain and unlabelled information in the target domain are incorporated into a unified framework for knowledge transformation.

We further extend our HDCC into a joint group-weight learning framework for multiple-source knowledge transfer. Under this framework, web images are leveraged according to their semantic relevances with the target domain videos. The motivation of group-weight learning is that brute force transfer learning may degrade the performance of classifiers for videos. To decrease the risk of this negative transfer, our strategy is to organise images according to their semantic meanings instead of origins. To collect knowledge from the Internet, we propose to use event-related keywords to query from image search engines. Since events in real world are complex and vary widely, single query word is not sufficient to describe the complicated situations. We therefore apply multiple associational keywords to represent an event. For example, we may associate the event 'basketball' with keywords of 'basketball match', 'NBA' (National Basketball Association), 'Kobe', and 'basketball dancer' etc. For each keyword, we collect a group of related web images. For each event, multiple groups of images are collected by querying web image search engines. By computing weights for all the groups of each event in a joint manner and assigning different weights to different source image groups based on their relevances to target events, our method can increase the chance of borrowing positive knowledge to the target domain.

The main contributions of this paper are: (i) we introduce an HDCC method for image–video DA by discriminatively learning a new common feature subspace. (ii) We propose to annotate videos by leveraging knowledge transferred from multiple web image groups which are queried from image search engines with different associational keywords. (iii) We develop a joint group-weight learning framework to effectively adapt different related source image groups to target videos according to the corresponding events. Figure 2 shows the framework of our method.

The remainder of this paper is organised as follows. We briefly review the related work in Section 2. The problem setting and definition are shown in Section 3. We then introduce HDCC and joint group-weight learning framework in Section 4. Extensive experimental results are presented in Section 5, followed by conclusions and future work in Section 6.

2 Relayed work

Transfer learning [1, 6] has been deployed in a wide variety of applications such as sign language recognition [7] and text classification [8]. The motivation behind transfer learning is that people can intelligently apply knowledge learned previously to solve new problems faster or with better solutions. In [9], Ma *et al.* transfer knowledge from laboratory dataset such as Kungliga Tekniska högskolan (KTH) to improve the action recognition accuracy in the real-world videos. Recently, knowledge transfer in multimedia content analysis has attracted much more attention of researchers [4, 10, 11]. Yang *et al.* [12] proposed adaptive SVM (A-SVM) to enhance the prediction performance of video concept detection, in which a new SVM classifier is adapted from an existing classifier. Duan *et al.* [13] proposed to simultaneously learn the optimal linear combination of base kernels and the target classifier by minimising a regularised structural risk function in the SVM framework. Moreover then, they proposed adaptive multiple kernel learning [14] to add the pre-learned classifiers as the prior. These methods mainly focus on the single source domain setting. To utilise numerous labelled image data in the Internet, multiple-source DA methods [5, 15] are proposed, which leverage different pre-computed classifiers learned from multiple-source domains. However, these methods take no account on intrinsic connections among data within and between domains, and assign different weights to different source domains only according to the difference between the target data and particular source data. In this paper, we propose to leverage different groups of images queried by different associational keywords to the web. By this means, we insure that the data in each group are of the same concept, and also insure that different groups within the same event are correlated to each other.

A few works have been done on investigation of knowledge transform from images to videos. In [16], transfer models are learned from loosely labelled web images. This work cannot distinguish actions such as 'standing-up' and 'sitting-down' because it does not utilise temporal information of actions in the image-based model. In [17], Chen *et al.* propose to discover semantic concept automatically from weakly labelled images to annotate consumer videos. Web images are also used in video summarisation [18] for providing some canonical point of view. Recently, Duan *et al.* [5] developed a new event recognition approach for consumer videos by using web images. In their work, video features and image features are integrated in a target decision function to jointly determine events in videos. Wang *et al.* [19] proposed organising web images in groups and assigning different weights to different groups to measure the relevance between the source and the target. In [20], they further extend the method to transfer incremental learning. Among these methods, features in different domains are used separately; potential connections among different feature spaces are ignored.

To adapt classifiers in different feature spaces, one simple way is to translate all the training data into a common feature space. The idea has already been demonstrated successful in several applications such as cross-lingual text classification [21]. To accomplish knowledge transfer across different feature spaces, researchers have proposed multi-view learning methods [22], in which each instance has multiple views in different feature spaces. Different from multi-view learning, in this paper, we focus on the situation where the training data are in a source feature space, and the test data are in a different target feature space, and that there is no correspondence between the instances in these spaces. The source and target feature spaces can be very different, as in the case of image and video. Li *et al.* [23] study the heterogeneous feature adaptation problem by augmenting the heterogeneous features to a common feature subspace. Recently, to solve more general translated learning problem such as the translation between documents and images, CCA [24] is introduced to capture the relationship between heterogeneous features. In [25], authors proposed to use kernel CCA to learn a description that exploits the relations between the ordered tag words and the images visual descriptors, and compute similarities across the two views. However, CCA is a supervised feature extraction method, and in our transfer learning setting the labels of the target domain data are not exploited, resulting in the limitation of the annotation performance. In this paper, we apply HDCC to translate image features and video features to a common feature subspace by using both labelled information in source domain and unlabelled information in target domain. Thus, classifiers learned on this common subspace can be adapted in both domains.

3 Problem statement

3.1 Motivation

Our aim is to improve the learning of the target predictive function $f_{\mathrm{t}}(\cdot)$ using knowledge in both source and target domains. In our learning scheme, we assume that some unlabelled data in the target domain can be seen at the training stage, which is known as *transductive learning* [1]. Under such setting, the predictive function learned in the source domain can be adapted in the target domain through some unlabelled target domain data.

Our method leverages different groups of source data and gradually adapts them to the target classifier by building a common feature subspace for two heterogeneous domains. As a result, our work not only provides a framework for automatic discovering which part of knowledge is helpful to the target domain or tasks, but also provides an effective way to transfer this knowledge from the source to the target.

3.2 Problem description

We set two domains in our work: a source domain $\mathcal{D}^{\mathrm{s}} = (\mathcal{X}^{\mathrm{s}}, P(X^{\mathrm{s}}))$, in which we have abundant labelled web images, and a target domain $\mathcal{D}^{\mathrm{t}} = (\mathcal{X}^{\mathrm{t}}, P(X^{\mathrm{t}}))$, in which we have a large number of unlabelled real-world consumer videos. Here, $P(X^{\mathrm{s}})$ and $P(X^{\mathrm{t}})$ are marginal distributions of the source domain feature space $\mathcal{X}^{\mathrm{s}}$ and the target domain feature space $\mathcal{X}^{\mathrm{t}}$, respectively.

We apply different associational keywords for each event to collect multiple image sets. Here, we refer to an image set returned by one keyword as a *group*, and each group represents one concept of the corresponding event. Consequently, the instances $\boldsymbol{X}^{\mathrm{s}}$ retrieved from image search engines are divided into different groups, according to their corresponding associational keywords. Resort to G multiple groups, we have knowledge of the events covering different concepts. We define the gth group of an event as $\boldsymbol{X}^g = \{\boldsymbol{x}_i^g\}_{i=1}^{N_g}$, where $g \in \{1, \ldots, G\}$ and $\boldsymbol{x}_i^g \in \mathbb{R}^{d_{\mathrm{s}}}$ is the ith image in the gth group. d_{s} represents the dimensionality of the source domain features and N_g represents the number of images in the gth group. In addition, we define N_{t} unlabelled videos in the target domain as $\boldsymbol{X}^{\mathrm{t}} = \{\boldsymbol{x}_i^{\mathrm{t}}\}_{i=1}^{N_{\mathrm{t}}}$, where $\boldsymbol{x}_i^{\mathrm{t}} \in \mathbb{R}^{d_{\mathrm{t}}}$ is the ith video in the target domain and d_{t} represents the dimensionality of the target domain features.

4 Transferring knowledge from web images to videos

4.1 Heterogeneous discriminative analysis of CC

As image feature and video feature are from two heterogeneous feature spaces, classifiers learned on images cannot be directly applied to videos. Any source sample (web image) and any target sample (video) can be projected onto this common space by using two projection matrices $\boldsymbol{w}_{\mathrm{s}} \in \mathbb{R}^{d_{\mathrm{s}} \times d_{\mathrm{c}}}$ and $\boldsymbol{w}_{\mathrm{t}} \in \mathbb{R}^{d_{\mathrm{t}} \times d_{\mathrm{c}}}$, respectively. Here, d_{c} is the dimension of common feature subspace. Specifically, we define a general function

$$\psi(\boldsymbol{x}) = \begin{cases} \phi(\boldsymbol{x}) = \boldsymbol{w}_{\mathrm{s}}' \boldsymbol{x}, & \text{if } \boldsymbol{x} \in \mathbb{R}^{d_{\mathrm{s}}} \\ \varphi(\boldsymbol{x}) = \boldsymbol{w}_{\mathrm{t}}' \boldsymbol{x}, & \text{if } \boldsymbol{x} \in \mathbb{R}^{d_{\mathrm{t}}} \end{cases}, \tag{1}$$

where x is either the source sample or the target sample. In this paper, we propose HDCC to learn the two projection matrices (i.e. $\boldsymbol{w}_{\mathrm{s}}$ and $\boldsymbol{w}_{\mathrm{t}}$). The assumption behind these methods is that the representations in these two spaces contain some common information that is reflected in correlations between them. Since the videos in the target domain are unlabelled, traditional supervised CCA which needs corresponding label information to connect features from these two domains is not suitable in such case. To solve the unsupervised problem in heterogeneous feature adaptation, we incorporate both discriminative information in the source domain and topology information in the target domain in a joint manner. Specifically, we find a correspondence between a video and its own keyframe, which provides a natural correlation between image/keyframe feature space and video feature space. Motivated by this correspondence, we further explore the topology of a video by dividing a video into several clips. A video is divided into several clips and a frame is randomly selected from each clip.

To make the projected samples more discriminative using CCs, we maximise the similarities of any pair of within-class samples and minimise the similarities of any pair of between-class samples. Formally, given N samples of paired data $\{(S_1, T_1), \ldots, (S_N, T_N)\}$, where $S_i \in \mathbb{R}^{d_{\mathrm{s}}}$ and $T_i \in \mathbb{R}^{d_{\mathrm{t}}}$ denote the image feature space and video feature space, respectively. The goal is to learn two projection matrices $\boldsymbol{w}_{\mathrm{s}} \in \mathbb{R}^{d_{\mathrm{s}} \times d_{\mathrm{c}}}$ and $\boldsymbol{w}_{\mathrm{t}} \in \mathbb{R}^{d_{\mathrm{t}} \times d_{\mathrm{c}}}$ by maximising the objective function

$$\begin{aligned} \max_{\boldsymbol{w}_{\mathrm{s}}, \boldsymbol{w}_{\mathrm{t}}} \quad & (\boldsymbol{w}_{\mathrm{s}}' \boldsymbol{C}_{ST} \boldsymbol{w}_{\mathrm{t}} + \boldsymbol{w}_{\mathrm{s}}' \boldsymbol{C}_{w} \boldsymbol{w}_{\mathrm{s}} - \eta \boldsymbol{w}_{\mathrm{s}}' \boldsymbol{C}_{b} \boldsymbol{w}_{\mathrm{s}}) \\ \text{s.t.} \quad & \boldsymbol{w}_{\mathrm{s}}' \boldsymbol{C}_{SS} \boldsymbol{w}_{\mathrm{s}} = 1 \\ & \boldsymbol{w}_{\mathrm{t}}' \boldsymbol{C}_{TT} \boldsymbol{w}_{\mathrm{t}} = 1 \end{aligned} \tag{2}$$

where $\boldsymbol{C}_{ST} = \sum_{i=1}^{N} \sum_{j=1}^{N} S_{ij}^{x} (\boldsymbol{x}_i^{\mathrm{t}} - \boldsymbol{x}_j^{\mathrm{t}}) S_{ij}^{y} (\boldsymbol{y}_i^{\mathrm{t}} - \boldsymbol{y}_j^{\mathrm{t}})^{\mathrm{T}}$ denotes the cross-feature covariance matrix. Here, $S_{ij}^{x} = \boldsymbol{x}_i^{\mathrm{t}'} \boldsymbol{x}_j^{\mathrm{t}}$ and $S_{ij}^{y} = \boldsymbol{y}_i^{\mathrm{t}'} \boldsymbol{y}_j^{\mathrm{t}}$. $\boldsymbol{x}_i^{\mathrm{t}}$ represents the image feature extracted from the keyframe of video clip $\boldsymbol{y}_i^{\mathrm{t}}$. $\boldsymbol{C}_{\mathrm{SS}}$ and $\boldsymbol{C}_{\mathrm{TT}}$ denote the auto-covariance matrices for image feature domain S and video feature domain T, respectively. $\boldsymbol{C}_b = \sum_{i=1}^{N} \sum_{i \in B_i}^{N} \boldsymbol{x}_i^{\mathrm{s}} \boldsymbol{x}_l^{\mathrm{s}'}$ and $\boldsymbol{C}_w = \sum_{i=1}^{N} \sum_{k \in W_i}^{N} \boldsymbol{x}_i^{\mathrm{s}} \boldsymbol{x}_k^{\mathrm{s}'}$ denote the inter-class covariance matrix and intra-class covariance matrix of the source domain data, respectively. B_i denotes the set of source domain images with the same label of $\boldsymbol{x}_i^{\mathrm{s}}$ and W_i denotes the set of source domain images with the different label of $\boldsymbol{x}_i^{\mathrm{s}}$. By introducing $\boldsymbol{C}_b$ and $\boldsymbol{C}_w$, the discriminative information of the source domain can be transferred to the common feature representation by maximising the inter-class variation and minimising the intra-class variation.

To solve the objective function in (2), we first obtain the corresponding Lagrangian

$$\begin{aligned} L(\lambda, \boldsymbol{w}_{\mathrm{s}}, \boldsymbol{w}_{\mathrm{t}}) = {} & \boldsymbol{w}_{\mathrm{s}}' \boldsymbol{C}_{ST} \boldsymbol{w}_{\mathrm{t}} + \boldsymbol{w}_{\mathrm{s}}' \boldsymbol{C}_{b} \boldsymbol{w}_{\mathrm{s}} - \eta \boldsymbol{w}_{\mathrm{s}}' \boldsymbol{C}_{w} \boldsymbol{w}_{\mathrm{s}} \\ & - \frac{\lambda_S}{2} (\boldsymbol{w}_{\mathrm{s}}' \boldsymbol{C}_{SS} \boldsymbol{w}_{\mathrm{s}} - 1) - \frac{\lambda_T}{2} (\boldsymbol{w}_{\mathrm{t}}' \boldsymbol{C}_{TT} \boldsymbol{w}_{\mathrm{t}} - 1). \end{aligned} \tag{3}$$

Then, the derivative of f with respect to $\boldsymbol{w}_{\mathrm{s}}$ and $\boldsymbol{w}_{\mathrm{t}}$ becomes

$$\frac{\partial f}{\partial w_{\mathrm{s}}} = \boldsymbol{C}_{ST} \boldsymbol{w}_{\mathrm{t}} + 2\boldsymbol{C}_b \boldsymbol{w}_{\mathrm{s}} - 2\eta \boldsymbol{C}_w \boldsymbol{w}_{\mathrm{s}} - \lambda_S \boldsymbol{C}_{SS} \boldsymbol{w}_{\mathrm{s}} = 0 \tag{4}$$

$$\frac{\partial f}{\partial w_{\mathrm{t}}} = \boldsymbol{C}_{TS} \boldsymbol{w}_{\mathrm{s}} - \lambda_T \boldsymbol{C}_{TT} \boldsymbol{w}_{\mathrm{t}} = 0 \tag{5}$$

From (5), we have

$$\boldsymbol{w}_{\mathrm{t}} = \frac{\boldsymbol{C}_{TT}^{-1} \boldsymbol{C}_{TS} \boldsymbol{w}_{\mathrm{s}}}{\lambda T} \tag{6}$$

By substituting (6) into (4), we arrive at the following problem:

$$(\boldsymbol{C}_{ST} \boldsymbol{C}_{TT}^{-1} \boldsymbol{C}_{TS} + 2\lambda_T \boldsymbol{C}_b - 2\lambda_T \eta \boldsymbol{C}_w) \boldsymbol{w}_{\mathrm{s}} = \lambda_S \lambda_T \boldsymbol{C}_{SS} \boldsymbol{w}_{\mathrm{s}} \tag{7}$$

As the covariance matrices $\boldsymbol{C}_{SS}$ is symmetric positive definite, we are able to decompose them using a complete Cholesky decomposition

$$\boldsymbol{C}_{SS} = \boldsymbol{R}_{SS} \cdot \boldsymbol{R}_{ss}' \tag{8}$$

where $\boldsymbol{R}_{SS}$ is a lower triangular matrix. If we let $u_{\mathrm{s}} = \boldsymbol{R}_{ss}' \cdot \boldsymbol{w}_{\mathrm{s}}$ and $\lambda = \lambda_S \lambda_T$, we are able to rewrite (7) as

$$\boldsymbol{R}_{SS}^{-1} (\boldsymbol{C}_{ST} \boldsymbol{R}_{TT}^{-1} \boldsymbol{R}_{TS} + \lambda_T \boldsymbol{C}_b + \lambda_T \eta \boldsymbol{C}_w) \boldsymbol{R}_{SS}^{-1'} u_{\mathrm{s}} = \lambda u_{\mathrm{s}} \tag{9}$$

We are therefore left with a symmetric eigenproblem of the form $Ax = \lambda x$. The optimisation of $\boldsymbol{w}_{\mathrm{s}}$ and $\boldsymbol{w}_{\mathrm{t}}$ involves the variables of λ_S, λ_T, $\boldsymbol{w}_{\mathrm{s}}$, and $\boldsymbol{w}_{\mathrm{t}}$. As the other variables are not explicitly represented by $\boldsymbol{w}_{\mathrm{s}}$, it is difficult to find a closed-form solution for $\boldsymbol{w}_{\mathrm{s}}$. In this paper, an iterative optimisation algorithm is proposed. We compute an optimal solution for two of the three variables at a time by fixing the rest one and repeating this for a certain number of iterations. After initialising λ_T, the problem in (9) can be solved by an eigenvalue decomposition problem.

With this common feature subspace, the classifiers learned on the source domain images can be easily adapted to the target domain videos. Our method explores how to take advantage of labelled information of source domain data to maintain discriminations and temporal information of target domain to obtain topological knowledge. It is worth mentioning that our HDCC method can be readily applied to other transfer learning tasks when heterogeneous features exist in the target domain data.

4.2 Training pre-learned classifiers

In this section, we discuss the pre-learned classifiers on G groups of web images. Formally, we define the gth classifier $f_{\mathrm{s}}^{g}(\boldsymbol{x}^{\mathrm{s},g})$ for the gth group in source domain as

$$f_{\mathrm{s}}^{g}(\boldsymbol{x}^{\mathrm{s},g}) = w_1 \psi(\boldsymbol{x}^{\mathrm{s},g}) + w_2 v(\boldsymbol{x}^{\mathrm{s},g}), \tag{10}$$

where $\boldsymbol{w} = [w_1; w_2]$ is the weighting parameter. $\boldsymbol{x}^{\mathrm{s},g}$ is the image sample of the gth group from the source domain. $\psi(\boldsymbol{x}^{\mathrm{s},g})$ and $v(\boldsymbol{x}^{\mathrm{s},g})$ are the common feature and the scale-invariant feature transform (SIFT) feature for $\boldsymbol{x}^{\mathrm{s},g}$, respectively. It is worth mentioning that our newly proposed intermediate subspace feature for the source and target samples can be readily incorporated into different methods, making these methods applicable for the transductive learning problem.

Algorithm 1

Require:

$\{X^g\}_{g=1}^G$: the set of image groups;

X^t : unlabeled target videos;

Ensure:

$\{f^g\}_{g=1}^G$: pre-learned classifiers;

$\{\alpha_g\}_{g=1}^G$: group weights.

Phase-I

Initialize: $\alpha_g = 1/G$;

Get initial source classifier $f_s^{g(0)}(x)$ using standard SVMs

Set $i = 1$

repeat

  Calculate $f_s^{g(i)}(x)$ for all target and source samples

  Choose largest τ target samples falling in margin band $\mathcal{M} = \{x| -1 \le f_s^{g(i)}(x^t) \le 1\}$

  Remove largest τ source data according to $f_s^{g(i)}(x^s)$

  Update $f_s^{g(i)}$

  $i = i + 1$

until Convergence

Phase-II

Initialize target classifier $f_t = \sum_{g=1}^G \alpha_g f_s^g(x)$

Use Quadratic Programming to minimize Eq. (16) to obtain α_g

return f_s^g and α_g

Fig. 1 *Algorithm 1: Joint group weighting learning*

Note that the distribution of image/keyframe feature in the target domain videos is different from that in the source domain images. Therefore, we employ DASVM [26] to train $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ of pre-learned group-classifiers to make them more adaptive to the target domain. In DASVM, source domain samples are only used for initialising the pre-learned classifiers for the target function [Algorithm 1 – phase I (see Fig. 1)]. After initialisation, the source domain samples are gradually replaced by the target domain samples which are used to learn the final separation hyperplane.

4.3 Joint group-weight learning

Given the initial classifiers from the previous step in Section 4.1, we now extend the situation of single group to multiple groups, in which each group is collected by an event-related associational keyword query from the Internet. We propose a novel joint group-weight learning scheme to integrate different groups of source data according to their corresponding weights. The weight for each source group represents its contribution to the target video.

The target video classifier in our group-weight learning method is defined by

$$f_{\mathrm{t}}(\boldsymbol{x}) = \sum_{g=1}^{G} \alpha_g f_{\mathrm{s}}^g(\boldsymbol{x}), \tag{11}$$

where $\alpha_g > 0$ is the weight for the gth group. We assume that the weights are normalised, that is, $\sum_{g=1}^G \alpha_g = 1$.

On the basis of the smoothness assumption for different groups, we minimise both the loss of the labelled source data and the difference between different group-classifiers on the unlabelled target data.

The proposed framework is given by

$$\min_{f_{\mathrm{t}}} \Omega(f_{\mathrm{t}}) + \lambda_L \Omega_L(f_t) + \lambda_T \Omega_T(f_{\mathrm{t}}) + \lambda_G \Omega_G(f_{\mathrm{t}}), \tag{12}$$

where λ_L, λ_G, $\lambda_T > 0$ are tradeoff parameters. The details of each term in (12) are described as follows.

Here, $\Omega(f_{\mathrm{t}}) = (1/2)\sum_{g=1}^G ||\alpha_g||^2$ controls the complexity of the target classifier f_{t}, where α_g, $g \in \{1, \ldots, G\}$ is the weight of the gth group.

$\Omega_L(f_{\mathrm{t}})$ is a loss function of the target classifier f_{t} on the labelled instances of the source domain, defined by

$$\Omega_L(f_{\mathrm{t}}) = \sum_{i=1}^{N_{\mathrm{s}}} ||f_{\mathrm{t}}(\boldsymbol{x}_i^{\mathrm{s}}) - y_i^{\mathrm{s}}||^2, \tag{13}$$

where $\boldsymbol{x}_i^{\mathrm{s}}$ is the ith web image, $\boldsymbol{y}_i^{\mathrm{s}}$ is the event label of $\boldsymbol{x}_i^{\mathrm{s}}$, and N_{s} is the number of training samples in the source domain. This regulariser enforces the decision value of the target classifier f_{t} on the source domain similar to the ground-truth of event label.

If we only use labelled source domain data to obtain the target function, the target function may overfit on these data, and the generalisation ability maybe degraded. As shown in the traditional transductive learning methods [1, 3], unlabelled data can be employed to improve the classification performance. Therefore, we use a group loss function $\Omega_G(f_{\mathrm{t}})$ to ensure the smoothness on the target domain data, which is parameterised as

$$\Omega_G(f_{\mathrm{t}}) = \sum_{i=1}^{N_{\mathrm{t}}} \sum_{g=1}^{G} \alpha_g \sum_{k=1,\, k\neq g}^{G} ||f_{\mathrm{s}}^k(\boldsymbol{x}_i^{\mathrm{t}}) - f_{\mathrm{s}}^g(\boldsymbol{x}_i^{\mathrm{t}})||^2. \tag{14}$$

This loss function enforces the target function to be smooth on the data: namely, different groups belonging to the same event should have similar decision values. For DA, we similarly assume that the pre-learned classifiers in the source domain should have similar decision values on the unlabelled samples in the target domain. For example, if the gth group and the kth group are from the same event, we ensure that $f_{\mathrm{s}}^k(\boldsymbol{x})$ is close to $f_{\mathrm{s}}^g(\boldsymbol{x})$. Actually, we introduce $\Omega_G(f_{\mathrm{t}})$ to penalise those groups far from major event-related groups.

In our framework, we also use the unlabelled instances in the target domain to enhance the generalisation ability of the classification model. So, we have the target data-driven regulariser formulated by

$$\Omega_T(f_{\mathrm{t}}) = \sum_{i=1}^{N_{\mathrm{t}}} ||f_{\mathrm{t}}(\boldsymbol{x}_i^{\mathrm{t}})||^2, \tag{15}$$

where N_t is the number of unlabelled target samples. This regulariser reduces distribution mismatching by purely using labelled instances from the source domains.

Putting everything together, we have the following optimisation problem:

$$\min_{\alpha} \frac{1}{2}||\alpha||^2 + \lambda_L \sum_{i=1}^{N_s} ||f_t(\boldsymbol{x}_i^s) - y_i^s||^2 + \lambda_T \sum_{i=1}^{N_t} ||f_t(\boldsymbol{x}_i^t)||^2 + \lambda_G \sum_{i=1}^{N_t} \sum_{g=1}^{G} \alpha_g \sum_{k=1, k\neq g}^{G} ||f_s^k(\boldsymbol{x}_i^t) - f_s^f(\boldsymbol{x}_i^t)||^2 \quad (16)$$

$$\text{s.t.} \sum_{g=1}^{G} \alpha_g = 1.$$

The optimisation problem of (16) can be solved by a standard quadratic programming.

The algorithm framework is summarised in Algorithm 1 (Fig. 1). In Phase-I of the training stage, we train the pre-learned classifiers for each individual group in the source domain. In Phase-II of the training stage, weights are assigned to different groups in a joint manner based on (16) to generate the target classifier f^t. In the test stage, for each test video, the image feature is extracted from the keyframe, and the video feature is represented by the common feature. We input these features to the target classifier f^t to obtain the final event label.

5 Experiments

5.1 Datasets

Columbia Consumer Video (CCV) dataset: This is a consumer video dataset collected by Columbia University [27]. It contains a training set of 4659 videos and a test set of 4658 videos which are annotated to 20 semantic categories. Since our work focuses on event analysis, we do not consider non-event categories (i.e. 'playground', 'bird', 'beach', 'cat', and 'dog'). To facilitate the keyword-based image collection using the web search engine, we merge 'wedding ceremony', 'wedding reception', and 'wedding dance' into one event as 'wedding'. We also merge 'non-music performance' and 'music performance' into 'performance'. Finally, there are 12 event categories: 'basketball', 'baseball', 'soccer', 'ice-skating', 'biking', 'swimming', 'skinning', 'graduation', 'birthday', 'wedding', 'show', and 'parade'.

Kodak dataset: This dataset is collected by Kodak [28] from about 100 real users over 1 year. We only consider six event categories (i.e. 'wedding', 'birthday', 'picnic', 'parade', 'show', and 'sports') in our experiments.

For each video, we apply two types of features: video feature and image feature. For video feature, we extract 144-dimensional (144D) 3D space–time interest point [27] on the CCV dataset. The 96D histograms of oriented gradients and the 108D histograms of optical flow [14] are used on the Kodak dataset. For image feature, we first randomly sample one frame from each video as its keyframe and then extract 128D SIFT features from salient regions detected by the difference of Gaussians detectors [29].

We use the Google image search engine to collect a large number of knowledge for all events. Each event is represented by five groups of knowledge. Specifically, each group is obtained by querying an associational keyword of the event from the image search engine. All these images constitute the source domain. Your paper must be in single column format with double spacing to ensure that reviewers can easily read and mark up this paper if required. All paragraphs must be justified, i.e. both left-justified and right-justified.

According to the above video datasets, we collect web images for 13 events: basketball, baseball, soccer, ice-skating, biking, swimming, graduation, birthday, wedding, show, parade, and picnic. Table 1 lists the associational keywords for each event in our experiment. Each row shows the five groups of an event and each column corresponds to one group of associational keywords. For each image, we extract 128D SIFT features in the same way used in keyframes.

5.2 Experimental setup

In the experiments, we use the bag-of-words for both image and video features. Specifically, we cluster the SIFT features, which are extracted from all the training web images and keyframes of the videos, into 2000 words by using *k*-means clustering. Each image–video keyframe is then represented as a 2000D token frequency feature by quantising its SIFT features with respect to the visual codebook. For the videos in both datasets, we directly use the 5000D and 2000D features provided by Duan *et al.* [14] and Jiang *et al.* [27] for the CCV dataset and Kodak dataset, respectively.

For a given event, we use five associational keywords as query to search relevant images, and collect the original top 300 images for each query keyword. To train an initial pre-learned classifier for each group, we use the queried 300 images in the corresponding group as positive samples and randomly select 300 images from other groups as negative samples. In the training stage, for the CCV dataset we use the training set defined by Jiang *et al.* [27] as the unlabelled target domain. For the Kodak dataset, the target domain contains the total 195 videos. Then, we have labelled web image groups in the source domain and unlabelled videos in the target domain to form our training set.

We compare our method with standard SVM, DASVM [26], two-stage weighing multiple DA (2SW-MDA) [30], and domain selection machine (DSM) [5], as these methods can work when there is no labelled training data in the target domain. The standard SVM and DASVM can transfer knowledge only when the source data and the target data are with the same type of features. Therefore, they can only use image features (i.e. SIFT) to learn classifiers for the target domain. Specifically, since we do not have any labelled data in the target domain, we learn one classifier from each group for the standard SVM. The standard SVM and DASVM could not handle multiple groups, and the final results are equal

Table 1 Examples of the associational query keywords for each event

Event	Group I	Group 2	Group 3	Group 4	Group 5
baseball	baseball	baseball games	foul line	softball	baseball American
basketball	basketball	basket	NBA	rebound	basketball court
biking	biking	bicycle	bike	cycle	biking clipart
birthday	birthday	cake	candle	celebrate	birthday dinner
graduation	graduation	finish school	scholar	graduation cap	graduate
ice-skating	ice-skating	patinar	skate	skater	skating cartoon
show	non-music performance	music performance	concert	show	acrobat performance
parade	parade	demonstration	procession	protest	halloween parade
skiing	skiing	skier	sled	slcgc	alpine skiing
soccer	soccer	fifa	football	real Madrid	AC Tinian
swimming	natation	swim	swimming	synchronised	aquatics
wedding	wedding	wedding ceremony	wedding dance	wedding reception	wedding cake
picnic	picnic	barbecue	cookout	food	dine together

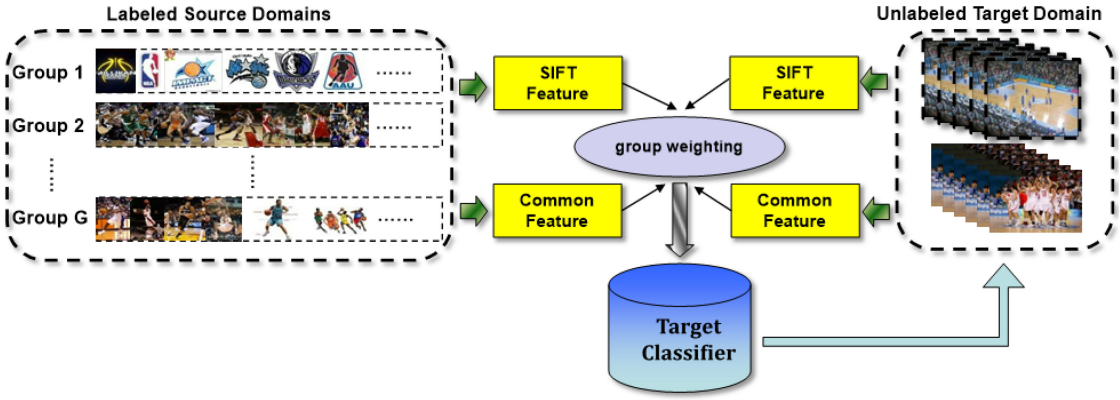

Fig. 2 *Illustration of the framework*

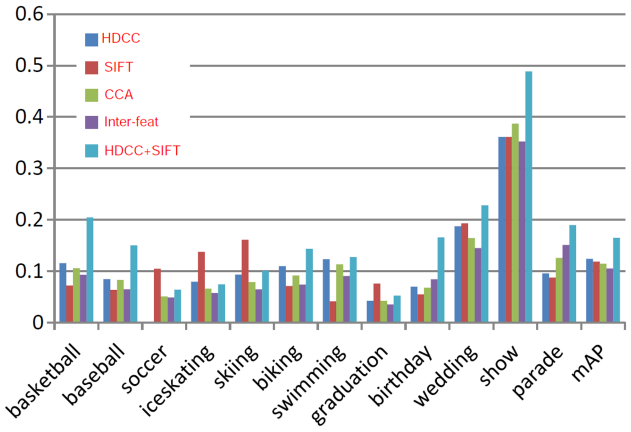

Fig. 3 *Evaluation of different features on the CCV dataset*

combination of all group-classifiers. For the DASVM, we obtain the initial classifier by labelled data in one source group and gradually update it using the unlabelled videos from the target domain. Similar to the standard SVM, we equally fuse the DASVM classifiers from all five source groups to obtain the final results. For the DSM and 2SW-MDA, we use the non-linear χ^2 kernel and average the decision values of pre-learned SVM classifiers using video keyframes to generate the prediction for each video. We do experiments on the newly proposed convolutional neural networks (CNN) features on the CCV dataset. The images and frames are extracted as 4096D feature vectors by using the Caffe [31] implementation of the CNNs described by Krizhevsky [32].

It is worth mentioning that we do not have any labelled data videos or keyframes in the target domain. Our initial single grouped data classifier is learned with the labelled training data only from the source domain. Moreover, we gradually remove the source data and use the video data instead of the source data, the detailed solution referred to [26]. For all the methods, we use the average precision (AP) for performance evaluation and define mean AP (mAP) as the mean of APs over all events.

5.3 Results

We first compare our method with existing approaches on the CCV and Kodak datasets. The mAPs of all methods on these datasets are shown in Table 2.

From the results, we observe that:

i. Our method achieves the best performance on both datasets, which show that jointly learning different groups of knowledge is beneficial to positive transform.
ii. Our method outperforms 2SW-MDA, which demonstrates the effectiveness of simultaneously minimising the loss of labelled source domain data on the target classifier as well as different regularisers defined on the unlabelled target domain data.
iii. Our method is better than DSM on mAPs, which illustrate the benefit of using grouped event-related images returned from web search engine and associating different weights to different groups. It is interesting to note that the data from all groups queried by associational keywords can benefit understanding video events.
iv. In terms of per-event APs, there is no consistent winner among the four methods. This indicates the existence of irrelevant data which hinds these transfer learning methods for acquiring good target classifiers. Our method achieves more stable performance, which demonstrates that jointly weighting different groups can cope with noisy web images.

We then evaluate the proposed HDCC on both the CCV and Kodak datasets. Figs. 3–4 show per-event AP results of different features: the proposed common feature (HDCC), SIFT feature (SIFT), intermediate feature [3], and the combination of SIFT and HDCC (HDCC+SIFT). For most events, we observe that the proposed common feature achieves comparable performance. Especially for those events which are closely related to motions such as ‘birthday’, ‘parade’, and ‘picnic’, the HDCC plays a more important role in improving the performance. The best results come from the integration of SIFT feature and common feature, which obviously demonstrate the beneficial of combining both image and video features for video annotation.

To verify the influence of the number of groups in our method, we report the performances using different group numbers. As shown in Fig. 5, the results on multiple groups consistently perform better than that on a single group. This indicates that our group weighting scheme can gain useful information to guide video annotation. We also note that the mAPs do not monotonically increase with the increase of group numbers. A possible explanation is that the ability to transfer knowledge from the source to the target depends on how they are related. The transferred knowledge becomes more useful when the relationships are closer.

We also investigate the effects of each term in our optimisation function in (16) for learning weights of groups. Table 3 shows the results when $\lambda_G = 0$ and $\lambda_T = 0$. From the results, we can observe that the mAP dramatically decreases when either $\Omega_G(f^t)$ or $\Omega_T(f^t)$ is removed from the optimisation function. In Table 3, we also

Table 2 Comparison of mAPs (%) between our method and other methods on the CCV and Kodak datasets

Method	SVM	SVM with CNN	DASVM	2SW-MDA	DSM	Ours	Ours with CNN
CCV	8.52	18.95	10.90	10.45	12.63	**17.38**	**25.41**
Kodak	23.92	—	28.63	25.34	31.54	**34.69**	—

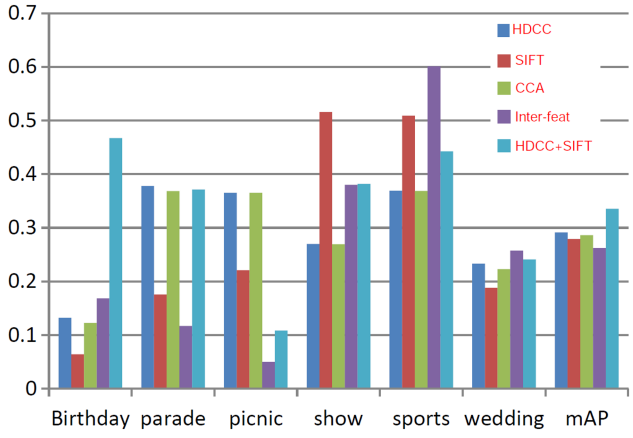

Fig. 4 *Evaluation of different features on the Kodak dataset*

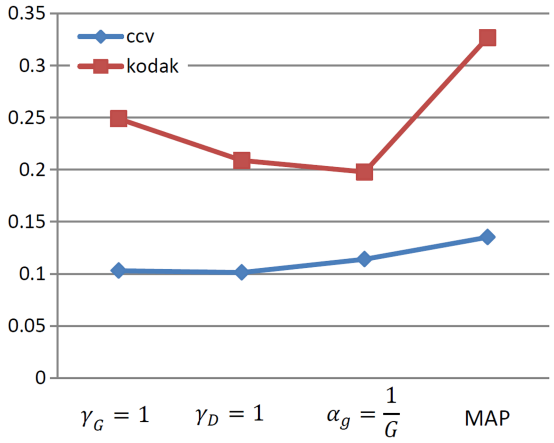

Fig. 5 *Evaluation of different group numbers*

Table 3 Evaluation of different components of (18) using mAPs (%)

	$\lambda_G = 0$	$\lambda_T = 0$	$\alpha_g = 1/G$	**Our method**
CCV	10.30	10.13	11.40	**17.38**
Kodak	24.90	20.90	19.79	**34.69**

show the results when the weights of all the groups are equal, i.e. $\alpha_g = (1/G)$.

6 Conclusion

In this paper, we have proposed a novel HDCC method to learn a common feature subspace for heterogeneous DA. Given labelled images and unlabelled videos, HDCC learns two projection matrices by integrating both discriminative information of the source domain and topology information of the target domain. By introducing these two matrices the different domain features can be compared in a common feature space. We further introduce a jointly group-weighted learning framework to utilise different groups of web images to annotate unlabelled real-world videos. In our framework, we divide the image data into different groups by querying the web image search engine with different associational keywords. Under this framework, different weights are assigned to different groups in a joint manner which takes account of correlations among the source groups, as well as the correlations between the source and the target.

In future, we will investigate how to incorporate convolutional neural network into our heterogeneous DA framework. Another important direction is to analyse the generalisation bound for heterogeneous DA [5].

7 Acknowledgment

This work was supported by the Fundamental Research Funds for the Central Universities with grant no. BLX2014-28, and the Fundamental Research Funds for the Central Universities (NO. 2015ZCQ-XX).

8 References

[1] Pan, S., Yang, Q.: 'A survey on transfer learning', *Knowl. Data Eng.*, 2010, **22**, (10), pp. 1345–1359
[2] Bergamo, A., Torresani, L.: 'Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach, 2010, pp. 181–189
[3] Gopalan, R., Li, R., Chellappa, R.: 'Domain adaptation for object recognition: an unsupervised approach'. ICCV, 2011, pp. 999–1006
[4] Kulis, B., Saenko, K., Darrell, T.: 'What you saw is not what you get: domain adaptation using asymmetric kernel transforms'. CVPR, 2011, pp. 1785–1792
[5] Duan, L., Xu, D., Tsang, S.-F.C.: 'Exploiting web images for event recognition in consumer videos: a multiple source domain adaptation approach'. CVPR, 2012, pp. 1959–1966
[6] Hu, J., Lu, J., Tan, Y.P.: 'Deep transfer metric learning, in: computer vision and pattern recognition', *IEEE*, 2015, pp. 325–333
[7] Farhadi, A., Forsyth, D., White, R.: 'Transfer learning in sign language'. CVPR, 2007, pp. 1–8
[8] Wang, H., Huang, H., Nie, F., *et al.*: 'Cross-language web page classification via dual knowledge transfer using nonnegative matrix tri-factorization'. Proc. of the 34th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2011, pp. 933–942
[9] Ma, Z., Yang, Y., Nie, F., *et al.*: 'Harnessing lab knowledge for real-world action recognition', *Int. J. Comput. Vis.*, 2014, **109**, (1-2), pp. 60–73
[10] Rahmani, H., Mian, A.: 'Learning a non-linear knowledge transfer model for cross-view action recognition'. Int. Conf. on Computer Vision and Pattern Recognition, 2015
[11] Gong, B., Shi, Y., Sha, F., *et al.*: 'Geodesic flow kernel for unsupervised domain adaptation'. CVPR, 2012, pp. 2066–2073
[12] Yang, J., Yan, R., Hauptmann, A.: 'Cross-domain video concept detection using adaptive SVMs'. Int. Conf. on Multimedia, 2007, pp. 188–197
[13] Duan, L., Tsang, I., Xu, D., *et al.*: 'Domain transfer SVM for video concept detection'. CVPR, 2009, pp. 1375–1381
[14] Duan, L., Xu, D., Tsang, I., *et al.*: 'Visual event recognition in videos by learning from web data'. CVPR, 2010, pp. 1959–1966
[15] Doretto, G., Yao, Y.: 'Boosting for transfer learning with multiple auxiliary domains'. CVPR, 2010
[16] Ikizler-Cinbis, N., Cinbis, R., Sclaroff, S.: 'Learning actions from the web'. CVPR, 2009, pp. 995–1002
[17] Chen, J., Cui, Y., Ye, G., *et al.*: 'Event-driven semantic concept discovery by exploiting weakly tagged Internet images'. Int. Conf. on Multimedia Retrieval, 2014, pp. 1–8
[18] Xiong, B., Grauman, K.: 'Detecting snap points in egocentric video with a web photo prior'. ECCV 2014, 2014, pp. 282–298
[19] Wang, H., Wu, X., Jia, Y.: 'Video annotation via image groups from the web', *IEEE Trans. Multimed.*, 2014, **16**, (5), pp. 1282–1291
[20] Wang, H., Song, H., Wu, X., *et al.*: 'Video annotation by incremental learning from grouped heterogeneous sources'. ACCV, 2014
[21] Bel, N., Koster, C., Villegas, M.: 'Cross-lingual text categorization', *Res. Adv. Technol. Digit. Libr.*, Lecture Notes in Computer Science 2003, **18**, (2769), pp. 126–139
[22] Muslea, I., Minton, S., Knoblock, C.: 'Active+ semi-supervised learning= robust multi-view learning'. Machine Learning-Int. Workshop then Conf., 2002, pp. 435–442
[23] Li, W., Duan, L., Xu, D., *et al.*: 'Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation, pattern analysis and machine intelligence', *IEEE Trans.*, 2014, **36**, (6), pp. 134–1148
[24] Hardoon, D., Szedmak, S., Shawe-Taylor, J.: 'Canonical correlation analysis: an overview with application to learning methods', *Neural Comput.*, 2004, **16**, (12), pp. 2639–2664
[25] Hwang, S., Grauman, K.: 'Accounting for the relative importance of objects in image retrieval'. Proc. of the British Machine Vision Conf., 2010, pp. 1–12
[26] Bruzzone, L., Marconcini, M.: 'Domain adaptation problems: a DASVM classification technique and a circular validation strategy', *PAMI*, 2010, **32**, (5), pp. 770–787
[27] Jiang, Y., Ye, G., Chang, S., *et al.*: 'Consumer video understanding: a benchmark database and an evaluation of human and machine performance'. ICMR, 2011, p. 29
[28] Loui, A., Luo, J., Chang, S., *et al.*: 'Kodak's consumer video benchmark data set: concept definition and annotation'. Workshop on Multimedia Information Retrieval, 2007, pp. 245–254
[29] Lowe, D.: 'Distinctive image features from scale-invariant keypoints', *IJCV*, 2004, **60**, (2), pp. 91–110
[30] Sun, Q., Chattopadhyay, R., Panchanathan, S., *et al.*: 'A two-stage weighting framework for multi-source domain adaptation', *Adv. Neural Inf. Process. Syst.*, (2011), pp. 505–513
[31] Jia, Y., Shelhamer, E., Donahue, J., *et al.*: 'Caffe: convolutional architecture for fast feature embedding', *Eprint Arxiv*, 2014, pp. 675–678
[32] Krizhevsky, A., Sutskever, I., Hinton, G.E.: 'Image net classification with deep convolutional neural networks', *Adv. Neural Inf. Process. Syst.*, 2012, **25**, pp. 25–33