Int J Comput Vis (2018) 126:390–409
https://doi.org/10.1007/s11263-017-1043-5

Transferring Deep Object and Scene Representations for Event Recognition in Still Images

Limin Wang[1] · Zhe Wang[2] · Yu Qiao[3] · Luc Van Gool[1]

Received: 31 March 2016 / Accepted: 1 September 2017 / Published online: 13 September 2017
© Springer Science+Business Media, LLC 2017

Abstract This paper addresses the problem of image-based event recognition by transferring deep representations learned from object and scene datasets. First we empirically investigate the correlation of the concepts of object, scene, and event, thus motivating our representation transfer methods. Based on this empirical study, we propose an iterative selection method to identify a subset of object and scene classes deemed most relevant for representation transfer. Afterwards, we develop three transfer techniques: (1) initialization-based transfer, (2) knowledge-based transfer, and (3) data-based transfer. These newly designed transfer techniques exploit multitask learning frameworks to incorporate extra knowledge from other networks or additional datasets into the fine-tuning procedure of event CNNs. These multitask learning frameworks turn out to be effective in reducing the effect of over-fitting and improving the generalization ability of the learned CNNs. We perform experiments on four event recognition benchmarks: the ChaLearn LAP Cultural Event Recognition dataset, the Web Image Dataset for Event Recognition, the UIUC Sports Event dataset, and the Photo Event Collection dataset. The experimental results show that our proposed algorithm successfully transfers object and scene representations towards the event dataset and achieves the current state-of-the-art performance on all considered datasets.

Keywords Event recognition · Deep learning · Transfer learning · Multitask learning

1 Introduction

Image classification is a fundamental and challenging problem in computer vision and many research efforts have been devoted to this topic during the past few years (Krizhevsky et al. 2012; Ioffe and Szegedy 2015; Everingham et al. 2010; Simonyan and Zisserman 2015; He et al. 2015; Shen et al. 2016; Wang et al. 2017). The majority of these contributions focus on the problem of object recognition or scene recognition, partially due to the simplicity of object and scene concepts and the availability of large-scale datasets (e.g., ImageNet (Deng et al. 2009) and Places (Zhou et al. 2014)). On the other hand, event recognition (Wang et al. 2015b; Salvador et al. 2015; Park and Kwak 2015; Xiong et al. 2015; Li and Li 2007) in static images is also important for semantic image understanding. Being able to selectively retrieve event images helps us to keep nice memories of particular episodes of our lives, to locate where images were taken, to analyze people's culture and so on.

In general, an event captures the complex behavior of a group of people, interacting with multiple objects, and taking place in a specific environment. As illustrated in Fig. 1, the characterization of the concept "event" is relatively compli-

Communicated by Kristen Grauman.

✉ Limin Wang
07wanglimin@gmail.com

Zhe Wang
buptwangzhe2012@gmail.com

Yu Qiao
yu.qiao@siat.ac.cn

Luc Van Gool
vangool@vision.ee.ethz.ch

[1] Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland

[2] Department of Computer Science, University of California, Irvine, CA, USA

[3] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

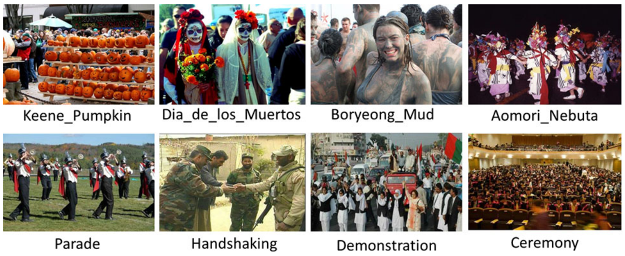

Fig. 1 Examples of event images from the ChaLearn Cultural Event Recognition dataset (top row) and the Web Image Dataset for Event Recognition (WIDER) (bottom row)

cated compared with the concepts of object and scene. Images from the same "event" category may vary even more in visual appearance and structure. Multiple high-level semantic cues, such as interacting objects, scene context, human poses, and garments, can provide useful information for event understanding.

Convolutional Neural Networks (CNNs) (LeCun et al. 1998) have delivered great successes in large-scale image classification, in particular for object recognition (Krizhevsky et al. 2012; He et al. 2015) and scene recognition (Shen et al. 2016; Wang et al. 2017). Large-scale image datasets (Deng et al. 2009; Zhou et al. 2014) (more than 1 million images) with human annotated labels have proven of great importance for this success. However, for event recognition, the current public datasets (Baro et al. 2015; Xiong et al. 2015) are relatively small and could be easily over-fitted by deep models. Moreover, the inherent complexity of the concept of event increases the difficulty of training an event CNN from scratch. Therefore, transferring deep models successfully trained on other datasets to the case of event recognition comes out to be a practical approach. Specifically, in this paper, we aim to study *why the deep representations learned from object and scene datasets are helpful for event recognition* and *how to effectively transfer these deep representations for more accurate event recognition*.

This paper makes three main contributions to transferring deep object and scene representations for event recognition. The first one is an empirical study on the correlation between classes of object, scene, and event. We exploit more accurate and universal concept classifiers to perform this study, including CNNs learned from the ImageNet and Places datasets. Our study empirically proves that there exists correlation among these three concepts. This justifies to pre-train event recognition networks with models learned from object and scene recognition datasets. Furthermore, it guides us to incorporate the tasks of object and scene recognition into the fine-tuning process for event recognition via a multitask learning framework.

The second contribution is to select object and scene categories most relevant for representation transfer. In particular, based on our empirical investigation, we propose an iterative selection method to identify subsets of discriminative and diverse object and scene categories. They increase the relevance between the main task and auxiliary task in our proposed data-based transfer method, resulting in better event recognition performance.

The third contribution is to design new transfer techniques. We propose three transfer techniques to fine tune the CNN models learned from object and scene datasets for event recognition: (1) initialization-based transfer, (2) knowledge-based transfer, and (3) data-based transfer. Our novel multitask based transfer methods are able to reduce the effect of over-fitting and improve the generalization ability of learned models. The experimental results indicate that they are fit to replace the existing fine-tuning for event recognition in still images.

Based on the transferred event models, coined as OS2E-CNNs, we propose an effective event recognition pipeline, that achieves the state-of-the-art performance on four public event recognition datasets. Meanwhile, we extensively study different aspects of our proposed method and try to provide more insights.

The remainder of this paper is organized as follows. In Sect. 2, we review work related to ours. Then, we present an empirical study of the correlation of objects and scenes with events in Sect. 3. Section 4 gives the description of our proposed transfer techniques of adapting object and scene CNNs for event recognition. We propose a simple yet effective event recognition pipeline with our learned OS2E-CNNs in Sect. 5. The experimental evaluation and exploration is described in Sect. 6. Finally, we discuss our method and offer some conclusions in Sect. 7.

2 Related Work

In this section, we briefly review previous work that is related to ours, and highlight its difference from ours.

2.1 Event Recognition in Videos

The analysis of human action and event is an active research area and much of the prior art has focused on video data (Duan et al. 2012; Bhattacharya et al. 2014; Wang et al. 2013, 2016a, b, 2015a, 2016c; Simonyan and Zisserman 2014; Ramanathan et al. 2015; Habibian and Snoek 2014; Mazloom et al. 2014; Izadinia and Shah 2012; Tang et al. 2012). We first review related work on event recognition in videos to well motivate our work on event recognition in still images. Most of the prior work in video based action and event recognition relied on modeling the motion information and temporal dynamics (Simonyan and Zisserman 2014; Ramanathan et al. 2015; Gan et al. 2015; Jain et al. 2015; Tang et al. 2012; Wang et al. 2016c). For examples, Simonyan and Zisserman (2014) developed a two-stream CNN for action recognition, where one captured the static appearance and the other described the temporal motion information. Jain et al. (2015) integrated object responses with motion features for action recognition in videos. Ramanathan et al. (2015) proposed a new temporal embedding method to capture video structure for event retrieval and recognition. Gan et al. (2015) proposed a deep architecture to jointly model spatial and temporal evidence for event recognition in videos. Tang et al. (2012) proposed a latent learning framework to capture the temporal structure of video. Wang et al. (2016c) modeled long-term temporal structure with a sparse sampling strategy and temporal aggregation module, which obtained good performance on the standard action recognition benchmarks.

Meanwhile, there was some work on event recognition, focusing on semantic representations (i.e., the responses of high level concepts, like objects, scenes, and actions) (Ebadollahi et al. 2006; Liu et al. 2013; Mazloom et al. 2014; Habibian and Snoek 2014; Izadinia and Shah 2012). For instance, Ebadollahi et al. (2006) first explored the semantic representation for event recognition. Habibian and Snoek (2014) comprehensively studied the problem of representing videos with the responses of concept detectors, and extensively investigated different aspects of this semantic representations. Liu et al. (2013) modeled the event class with a set of complementary concepts and these concepts were treated as attributes in a semantic space. Mazloom et al. (2014) proposed a new algorithm to learn what concepts are most informative per event, called as *conceptlets*, and solved this problem with an importance sampling method. Izadinia and Shah (2012) proposed a large margin formulation to model the relation between high-level event and low-level event (atomic actions), and used the low-level event detection scores as feature representations. Marszalek et al. (2009) used scene information as contextual cues to improve action recognition performance in videos. Hauptmann et al. (2008) proposed to use semantic concepts of object and scene for video event retrieval.

Our work differs from these methods on event or action recognition in videos from three aspects. First, our method is designed for event recognition in still images, and these methods based on motion dynamics cannot be directly adapted from video domain to image domain. Second, our study is conducted in a deep learning scenario and based on the two standard object and scene benchmarks. Our method relies on more accurate CNN classifiers and the concept scores are more reliable than those traditional detectors. More importantly, although with similar intuition to those semantic representations, our main goal is different from theirs. Instead to directly employing concept responses as visual representations, we aim to use the concept correlation to learn more effective visual representations within our proposed multi-task based transfer framework.

2.2 Event Recognition in Still Images

For still images, action recognition (Yao et al. 2011; Yao and Li 2010; Desai and Ramanan 2012; Delaitre et al. 2011; Gkioxari et al. 2015; Cooper et al. 2003; Bossard et al. 2013; Vu et al. 2014) tended to receive more attention than event recognition. Some work aimed to explore the relationship between multiple cues for holistic image understanding and action recognition. For example, Yao et al. (2012) developed an approach for holistic scene understanding that jointly reasons about regions, location, class and spatial extent of objects, presence of a class in the image, as well as the scene type. Vu et al. (2014) used scene information to predict actions for still images. Among the work on event recognition in still images, Li and Li (2007) proposed a coupled LDA framework to jointly infer the category of event, object, and scene. This method coupled two LDA models and was difficult to scale up to large-scale datasets. Bossard et al. (2013) proposed a latent sub-event approach for event recognition in photo collections and developed a Stopwatch Hidden Markov model. Cooper et al. (2003) presented a similarity-based method to cluster digital photos by time and image content. Xiong et al. (2015) designed a deep CNN architecture by fusing the responses of multiple channels for objects, faces, and people, and performed event recognition in an end-to-end manner.

Recently, at the ChaLearn Looking at People (LAP) challenge (Baro et al. 2015; Escalera et al. 2015), several deep learning based methods for the task of cultural event recognition were presented (Wei et al. 2015; Liu et al. 2015; Rothe et al. 2015; Wang et al. 2015b, c). Liu et al. (2015) proposed to use selective search to generate a set of proposals and to exploit a feature hierarchy to represent these proposals for event recognition. Wei et al. (2015) designed an ensemble method to incorporate the spatial structure into deep representations within a deep spatial pyramid framework (Gao et al. 2015). Rothe et al. (2015) developed a deep linear discriminative retrieval approach for event recognition by extracting features from four layers of CNNs.

This paper is based on our previous challenge solutions (Wang et al. 2015b, c), where we proposed the object-scene convolutional neural networks (OS-CNNs) to transfer the deep representations of object and scene models for event recognition. This network architecture was adopted by other participants (Wei et al. 2015; Liu et al. 2015; Rothe et al. 2015) at the ICCV ChaLearn LAP challenge (Escalera et al. 2015). We substantially extend our previous work by empirically investigating the correlation of object, scene, and event, identifying a small subset of discriminative object and scene classes, and proposing new transfer techniques to improve the generalization capacity of learned event models.

2.3 Transfer Learning

Many approaches have been proposed in recent years to solve the visual domain adaption problem (Fernando et al. 2013; Gong et al. 2014; Kulis et al. 2011), also called as the visual dataset bias problem (Torralba and Efros 2011). These methods recognized that there is a shift in the distribution of the source and target data representations, and they usually tried to learn a feature space transformation to align the source and target representations. Recently, supervised CNN representations have been shown to be effective for a variety of visual recognition tasks (Girshick et al. 2014; Oquab et al. 2014; Chatfield et al. 2014; Sharif Razavian et al. 2014; Azizpour et al. 2015). Sharif Razavian et al. (2014) proposed to treat CNNs as generic feature extractors, yielding an astounding baseline for many visual tasks. Girshick et al. (2014) designed a region-based object detection pipeline and transferred classification models to the detection task. Oquab et al. (2014) proposed a transfer framework to adapt a representation learned from the ImageNet dataset for various tasks on the Pascal VOC dataset (Everingham et al. 2010). Chatfield et al. (2014) comprehensively studied three types of models learned in the ImageNet dataset and transferred these representations to the Pascal VOC dataset. Azizpour et al. (2015) empirically analyzed different factors of transferability for a generic CNN representation on a variety of tasks. Tzeng et al. (2015) recently proposed to jointly learn deep models between source and target domains, where both domains shared the same task.

There are also some works on domain adaption with application in event recognition from videos (Ma et al. 2014; Yang et al. 2012; Yan et al. 2015). Ma et al. (2014) proposed an algorithm to adapt knowledge from another source for event detection even if the features of the source and the target are partially different, but overlapping. Yang et al. (2012) developed a robust cross-media transfer for event detection by taking different types of noisy social multimedia data as input. Yan et al. (2015) developed an event oriented dictionary learning method by leveraging training samples of selected concepts from the Semantic Indexing dataset.

We focus on event recognition in still images by transferring deep representations in object and scene models. In our case, both the distributions of input images and final targets are different from those of the source tasks. Furthermore we propose multitask learning frameworks, which try to utilize extra knowledge or data to guide the fine-tuning procedure on the target dataset. Therefore, our proposed transfer methods can help to reduce the effect of overfitting and improve the generalization ability of the learned models.

3 An Empirical Study

In this section, we quantitatively study the correlation among the concepts of object, scene and event by using more accurate and universal concept classifiers learned from standard benchmarks (i.e., ImageNet and Places). Then, we propose to select a subset of discriminative object and scene categories to better fine tune these deep representations for event recognition.

3.1 Evaluating Object and Scene Responses

In order to empirically study the correlation between the presence of object or scene and event classes, we choose the CNN learned in the ImageNet dataset (Deng et al. 2009) and Places dataset (Zhou et al. 2014) as a accurate and universal object and scene detector, respectively. These two datasets are currently the largest object recognition and scene recognition datasets, where we use 1000 object classes from ImageNet and 205 scene classes from Places. These categories almost cover all common object and scene classes, that possibly correlate with the event classes. We now first introduce the training details of these object and scene CNNs. Then, we describe how to calculate the object and scene responses for each image.

CNN models. We choose the network architecture of inception with Batch Normalization (BN inception) (Ioffe and Szegedy 2015), due to its balance between accuracy and efficiency. This network starts with 2 convolutional and max pooling layers, subsequently has 10 inception layers, and ends with a global average pooling layer and a fully connected layer. For both object and scene CNN models, we use the same training policy. Specifically, we use a multi-GPU parallel version (Wang et al. 2015d) of the Caffe toolbox (Jia et al. 2014), which is publicly available.[1] These networks are learned using the mini-batch stochastic gradient decent algorithm, where the batch size is set to 256 and momentum to 0.9. The learning rate is initialized as 0.1 and reduced by a factor of 10 every 200,000 iterations. The whole training procedure stops at 750,000 iterations. Concerning data augmentation, we use the common techniques such as conner cropping, scale jittering, and horizontal flipping (Wang et al. 2015d). The training images are resized to 256 × 256 and these cropped regions are resized to 224 × 224. Our BN inception models achieve a top-5 error of 7.9% on the ImageNet validation set and 11.8% on the Places validation set with a single crop.

Object and scene responses. After the training of object and scene CNNs, we treat them as universal object and

[1] https://github.com/yjxiong/caffe.

scene detectors, respectively. More concretely, we choose the ChaLearn Culture Event Recognition dataset (Escalera et al. 2015) for our empirical investigation. We use these object and scene CNN models to scan over the training dataset and compute the likelihood of the existence of certain object and scene classes for each image. Specifically, we first resize each image into three different scales of 256 × 256, 384 × 384, and 512 × 512 to handle scale variations. Then, we crop 224 × 224 regions from these resized images in a 3 × 3 grid. Finally, each crop is fed into the CNN models to obtain score vectors S^o and S^s to represent the distribution of object and scene classes. These scores of multiple crops from every single image are averaged to yield the image-level object and scene distribution, denoted by Φ^o and Φ^s.

3.2 Exploring Object and Scene Responses

After having calculated the object and scene responses for the training images, we are ready to quantitatively analyze the correlation among these three concepts: object, scene, and event. Here we try to answer the question of *whether and how different event classes tend to co-exist with distinctive sets of objects and scenes*.

More specifically, given a set of event training images $\{\mathbf{I}_i, y_i, \Phi^o(\mathbf{I}_i), \Phi^s(\mathbf{I}_i)\}_{i=1}^N$, where $\mathbf{I}_i$ denotes the image, y_i is its event label, and $\Phi^o(\mathbf{I}_i)$ and $\Phi^s(\mathbf{I}_i)$ are the scores of object and scene responses calculated as described in the previous subsection. $\Phi^o(\mathbf{I}_i)$ and $\Phi^s(\mathbf{I}_i)$ could be interpreted as distribution $p(o|\mathbf{I}_i)$ and $p(s|\mathbf{I}_i)$, we estimate the conditional distribution $p(o|e)$ and $p(s|e)$ as follows:

$$\begin{aligned} p(o|e) &= \frac{1}{N_e}\sum_{i:y_i=e}\Phi^o(\mathbf{I}_i)[o], \\ p(s|e) &= \frac{1}{N_e}\sum_{i:y_i=e}\Phi^s(\mathbf{I}_i)[s], \end{aligned} \tag{1}$$

where $p(o|e)$ and $p(s|e)$ are the conditional distributions of object class o and scene class s given event class e, $\Phi[j]$ represents the jth element of Φ, and N_e is the number of images belonging to the eth event class (i.e., $N_e = \sum_{i=1}^N \mathbb{I}(y_i = e)$). We take the score average of images from the same event class to estimate the conditional distribution of object and scene given a specific event class. Meanwhile, we estimate the prior probability of event e as follows:

$$p(e) = \frac{N_e}{N}, \tag{2}$$

where N is the total number of training images.

To investigate the correlation between object or scene and event, we first visualize several examples of conditional probabilities $p(o|e)$ and $p(s|e)$ in Fig. 2. We notice that some event classes show strong responses to specific object or scene classes. For example, in the event of `aomori nebuta`, the object class `comic book` has the highest response. In the event of `beltane fire`, there is a strong preference for the object class `torch`. For scene responses, the event of `afrika burn` yields a high response for the `desert` scene class and the event of `bastille day` for the `tower` scene class. In such cases, the corresponding object and scene class may act as a strong visual attribute that can be exploited for event recognition. On the other hand, some event classes, such as `battle of the orange` and `boston marathon`, may have strong response scores for multiple object and scene classes simultaneously. In these situations, the co-occurrence of several object and scene classes may contribute to the prediction of event classes. From these examples in Fig. 2, we conclude that some event classes may have strong responses to a specific object and scene class, or a small subset of object and scene classes. Therefore, the deep representations learned in object and scene models should be also discriminative and helpful for event recognition, and pre-trained models provide good initialization for fine-tuning on the event dataset. In the next subsection, we will define a quantitative measure to evaluate the discriminative capacity of object and scene classes for event recognition.

To further analyze the object and scene responses on the whole event dataset, we estimate the marginal distributions of object and scene as follows:

$$p(o) = \sum_e p(o|e)p(e), \quad p(s) = \sum_e p(s|e)p(e). \tag{3}$$

With the above equations, we can estimate the distributions of common objects and scenes in the ChaLearn Cultural Event Recognition dataset. As shown in Fig. 3, we plot the distribution of object and scene classes, where we only show the top 100 object and scene classes for visual clarity. The distributions of object and scene responses both exhibit the property of long tails, where most of the probability resides in a small subset of these object and scene classes. These universal object and scene classes do not relate to event classes equally and many classes only weakly correlate with the event classes. Hence, it is possible to identify a small subset of object and scene classes for more efficient and effective representation transfer.

3.3 Selecting Object and Scene Classes

In this subsection, we aim to measure the discriminative capacity of object and scene classes, and find a subset that is most discriminative for distinguishing the event classes. These subsets of object and scene classes enhance the relevance of object and scene recognition with event clas-

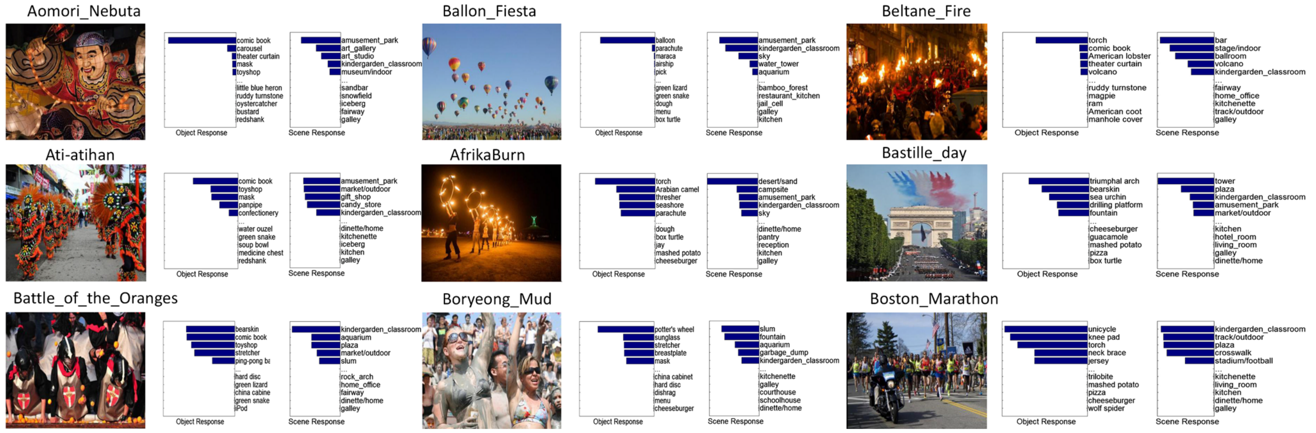

Fig. 2 ChaLearn Cultural Event categories with corresponding histograms of object responses (i.e., $p(o|e)$) and scene responses (i.e., $p(s|e)$). Top row: event categories with relatively low-entropy object responses. Middle row: event categories with relatively low-entropy scene responses. Bottom row: event categories with high-entropy object and scene responses. Best viewed in color (Color figure online)

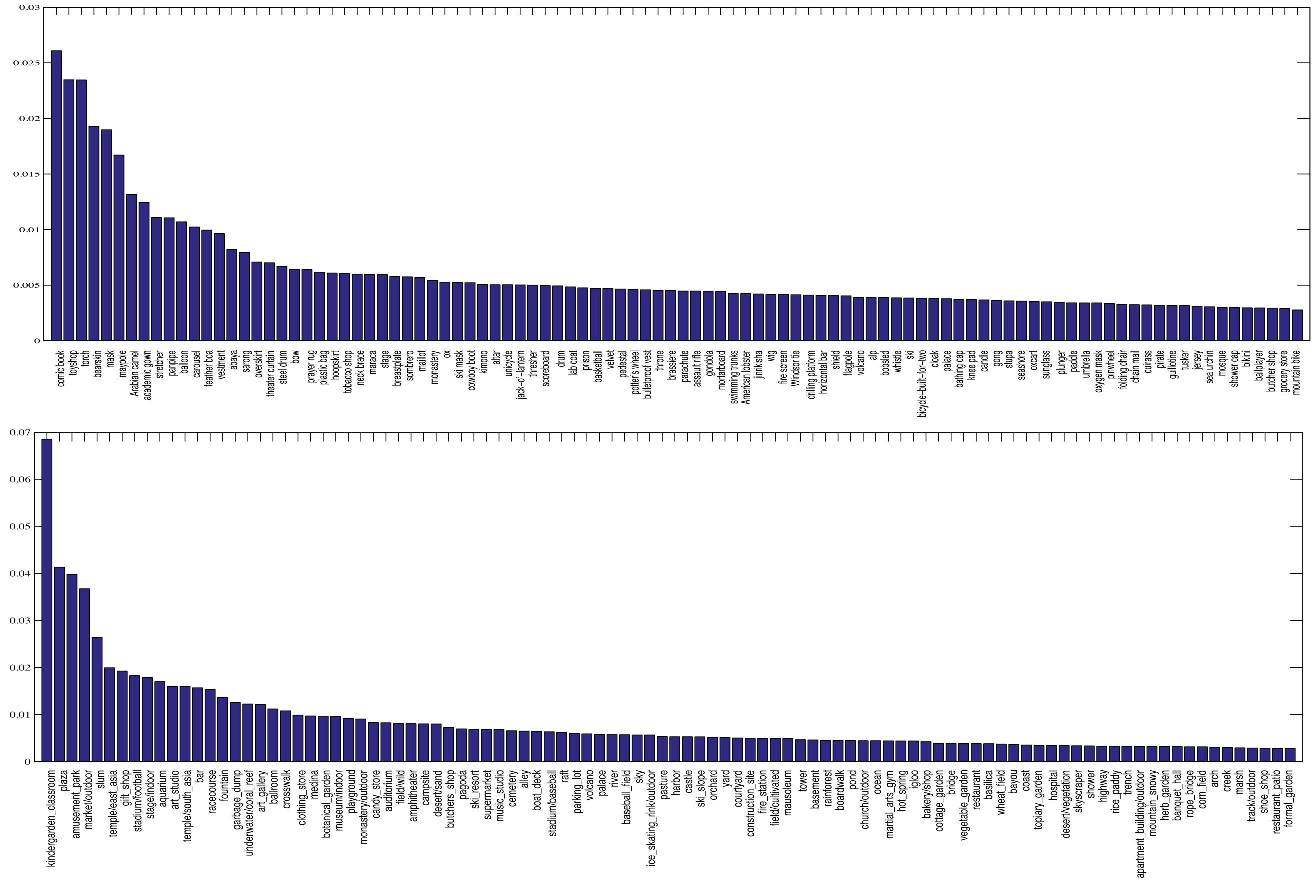

Fig. 3 The overall object responses (i.e., p(o)) and scene responses (i.e., p(s)) on the ChaLearn Cultural Event Recognition dataset. We notice that the distributions of object responses and scene responses yielded a "long tail", which means the most probability goes to a small portion of the object and scene classes. Best viewed in color (Color figure online)

sification, thus contributing to the better performance of our proposed data-based transfer method in Sect. 4.

Specifically, the selection of the subsets of object and scene classes is according to the following two principles:

– Each selected object or scene should only occur in a small number of event classes. We call this property the *discriminative capacity* of a single object or scene.

- Each object or scene in the subset should have a low correlation with the others. We call this property the *diversity capacity* of the subset.

Formulation. Based on these principles, we formulate the subset selection as an inference problem on a fully-connected graph, where each node corresponds to an object or scene class, and each edge encodes the correlation between the connected pair of nodes. Each node is associated with a binary hidden variable $h_i \in \{0, 1\}$, representing whether the object or scene class is selected or not. Therefore, given a set of object or scene classes $\mathbf{h} = \{h_i\}_{i=1}^N$, we want to minimize the following energy function:

$$E(\mathbf{h}) = \sum_{i=1}^{N} \phi(h_i) + \lambda \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \psi(h_i, h_j), \\ \text{s.t.} \sum_{i=1}^{N} h_i = K \tag{4}$$

where $\phi(h_i)$ is a unary term to represent the cost of selecting the ith class, $\psi(h_i, h_j)$ is a pairwise term to denote the cost of having both ith and jth classes, K is the number of categories to be selected, and λ is a weight parameter to balance these two terms (set to 0.5). The unary term is a penalty function to ensure the discriminative capacity of each selected class, and the pairwise term is a penalty function to encourage the diversity of the selected subset. It should be noted that our formulation is similar to plenty of works on graph-based formulation for selection in different problems, such as part selection (Juneja et al. 2013), attribute selection (Zheng et al. 2014), and feature selection (Das et al. 2012).

To meet the requirement of discriminative capacity, the responses of selected objects and scenes should peak for a small subset of events. Entropy is a natural measure to quantify the peaked nature of a probability distribution. Therefore, we adopt the conditional entropy $H(E|o)$ to represent the discriminative capacity of the oth object, which is defined as follows:

$$H(E|o) = -\sum_{e} p(e|o) \log_2 p(e|o), \tag{5}$$

where $p(e|o)$ is the conditional event distribution given a specific object class, which can be computed from Eqs. (1) and (2) using Bayes' formula. So, if $h_i = 1$, then $\phi(h_i) = H(E|i)$.

As to a subset's diversity capacity, we need to consider the correlation between pairs of classes. Instead of using low-level features to calculate their similarity, we utilize the conditional probability $P(e|o)$ to measure their correlation. If two object classes would predict similar events, they should have similar conditional probabilities $p(e|o)$. Specifically, if $h_i = 1$ and $h_j = 1$, then $\psi(h_i, h_j) =< p(e|i), p(e|j) >$.

Algorithm 1: Selecting Objects.

```
Data: conditional distribution $p(e|o)$, number: $K$.
Result: selected object classes: $\mathcal{O} = \{o_i\}_{i=1}^K$.
- Compute the cost of discriminative capacity of each object $\phi(o)$
defined in Eq. (5).
- Initialization: $n \leftarrow 0$, $\mathcal{O} \leftarrow \emptyset$.
while $n < K$ do
    1. For each remaining object $o$, update the correlation
    measure:
       $S(\mathcal{O}, o) = \frac{1}{|\mathcal{O}|} \sum_{o_i \in \mathcal{O}} < p(e|o_i), p(e|o) >$,
    2. Choose the object class : $o^* \leftarrow \arg\min_o \phi(o) + \lambda S(\mathcal{O}, o)$.
    3. Update: $n \leftarrow n + 1$, $\mathcal{O} \leftarrow \mathcal{O} \cup \{o^*\}$
end
- Return object classes: $\mathcal{O}$.
```

Optimization. In general, the pairwise potential $\phi(h_i, h_j)$ is calculated from the event dataset and it is highly dependent on the dataset. So we can not assume any property about the pairwise potential and make sure the it strictly follow the submodular property (Zheng et al. 2014). Therefore, the problem defined in Eq. (4) is a NP-hard problem in general.

Inspired by the part selection method (Juneja et al. 2013), we design a greedy selection method as shown in Algorithm 1. We first pick the object class that has the lowest conditional entropy. Then we update the correlation of the remaining object classes with the selected one with the average similarity. After this, we choose the object class which simultaneously minimizes the cost of discriminative and diversity capacity and go to the next iteration. The whole iteration is repeated until K objects are selected. In practice, we also try other more complex approximate algorithms of sequential tree-reweighted algorithm (TRW-S) (Kolmogorov 2006), which show similar performance to our proposed greedy selection method.

4 Transferring Deep Representations

In this section, we focus on how to transfer these deep object and scene representations for the task of event recognition. The representations of object and scene CNNs are learned to maximize the performance in classifying each image into predefined object and scene classes, respectively. The difference between object or scene recognition and event recognition is double: (1) *domain difference:* the distribution of event images is different from those of object and scene images. (2) *target difference:* the final recognition objective is different. Therefore, transferring these deep representations is necessary to deal with the problems of domain and target difference.

Some challenges remain for this transfer. As said, the sizes of the event recognition datasets (e.g., ChaLearn Cultural Event Recognition dataset (Escalera et al. 2015) are relatively small compared with those of the large-scale object and scene recognition datasets (e.g., ImageNet (Deng et al. 2009) and Places (Zhou et al. 2014)). However, CNNs come with millions of parameters and may easily over-fit with limited training samples. Hence, the question of *how to adapt deep representations to new tasks with limited training samples* needs to be explored. As shown in Fig. 4, we try to design effective transfer techniques, able to reduce over-fitting to the training data and to improve the generalization capacity of the learned model.

4.1 Baseline: Initialization-Based Transfer

Fine-tuning yields a simple yet effective method to transfer deep models learned in a large-scale dataset to a new task, where only a smaller training dataset is available. We choose fine-tuning as a baseline transfer technique. We call this transfer method as *initialization based transfer*, because we simply copy the weights of pre-trained models to the corresponding layers of event models as initialization. Specifically, suppose we have M event classes, we then minimize the cross-entropy loss function in the target dataset D_e defined as follows:

$$\ell_C(D_e) = -\sum_{\mathbf{I}_i \in D_e} \sum_{m=1}^{M} \mathbb{I}(y_i = m) \log p_{i,m}, \tag{6}$$

where $(\mathbf{I}_i, y_i)$ is a pair of an image and its label, D_e is the event dataset, $p_{i,m}$ is the mth output of the softmax layer for image $\mathbf{I}_i$.

In practice, we choose the BN inception architecture (Ioffe and Szegedy 2015), but we make two important modifications when minimizing the above loss function to improve the final recognition performance:

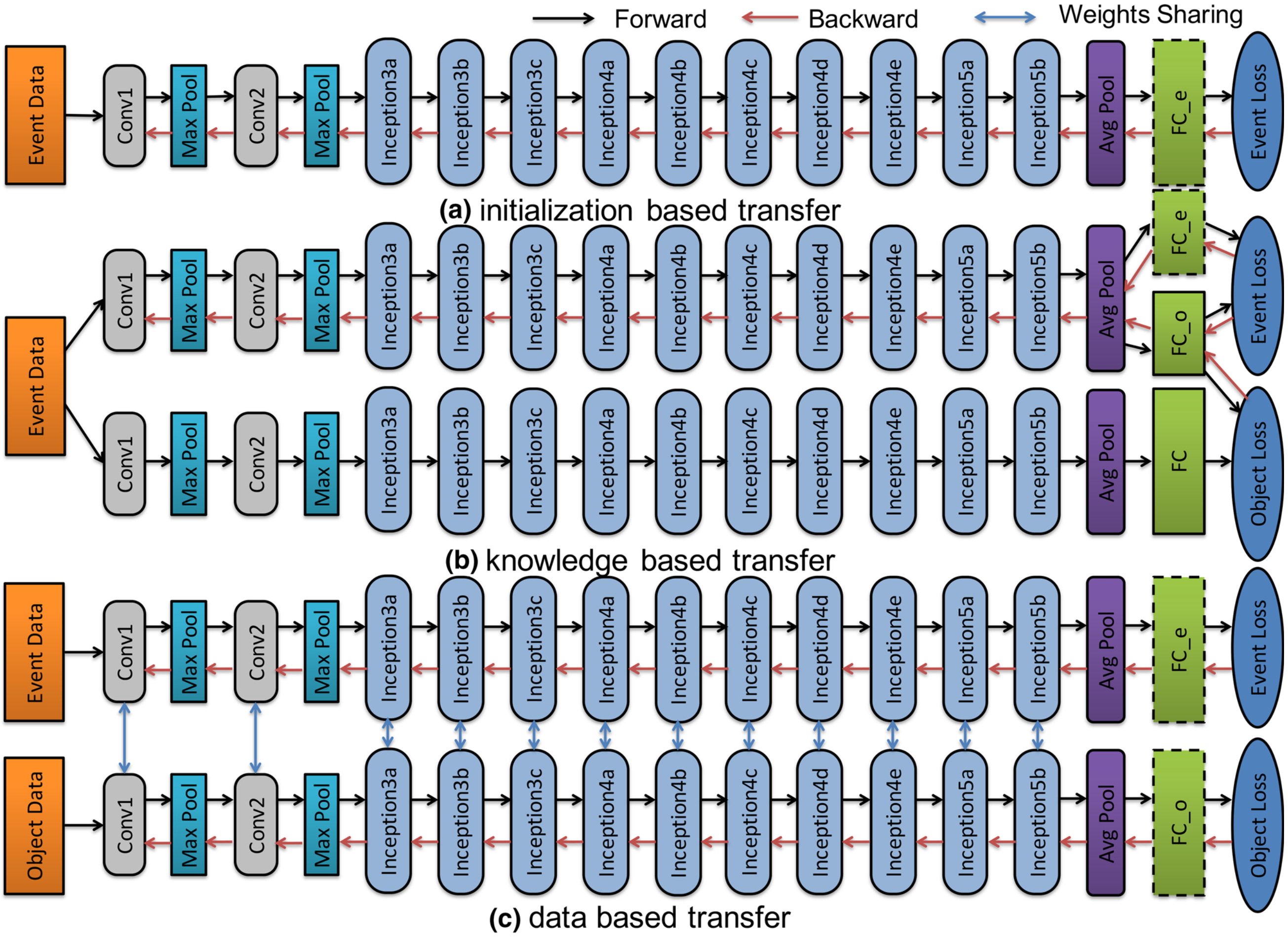

Fig. 4 Illustration of three transfer techniques that we propose: **a** initialization based transfer, **b** knowledge based transfer, and **c** data based transfer. In initialization based transfer, we use the pre-trained models to initialize the event CNNs and fine-tune them on the target dataset. In knowledge based transfer, we utilize the soft codes produced by object or scene networks to guide the training of event CNNs. In data based transfer, we exploit object or scene training data to jointly learn event and object or scene CNNs in a weight-sharing scheme

- We freeze the BN layers in the event CNNs. In the original training of the BN inception network (Ioffe and Szegedy 2015), it estimates the activation mean and variance within each batch, and uses them to transform these activation values into a standard distribution. This data adaptively estimating the activation mean and variance contributes to accelerating the CNN's convergence. Yet, it also increases the risk of over-fitting when having limited training samples. Therefore, we freeze the mean and variance values estimated from the object and scene datasets.
- We add a dropout layer before the fully connected layer in the BN inception network. In the original BN inception network (Ioffe and Szegedy 2015), it turns out BN acts as a kind of regularizer, and there is no need for a dropout layer when training on the large-scale ImageNet dataset. However, the size of the event recognition dataset is much smaller than that of ImageNet, and in our experiments, the effect of over-fitting will be more serious without the dropout layer.

This transfer technique simply employs a pre-trained model as initialization and ignores other information during the fine-tuning procedure. Although the fine-tuning process starts with a semantic and stable initialization, it may still suffer from over-fitting, as we shall see in our experiments. Incorporating other relevant tasks into the procedure of fine-tuning will regularize the learning process of event CNNs and thus relieve the over-fitting problem. We will design two multitask based transfer methods in the following subsections.

4.2 Knowledge-Based Transfer

As shown in Sect. 3, the occurrence of object and scene is highly correlated with specific event class. The fine-tuning technique simply utilizes object and scene models as initialization but ignores the rich information coming from the present object and scene background during the fine-tuning procedure. This subsection aims to propose a multitask based transfer method to incorporate object and scene recognition into the fine-tuning process on the event dataset.

A complicating factor is that we only have event labels on the target dataset, but no object and scene labels. To overcome this, we utilize the pre-trained object and scene CNNs to predict the likelihood of object and scene classes. As shown in Fig. 4, we treat the output scores of object and scene CNNs as soft codes and use these soft codes as supervision to guide the fine-tuning of event CNNs. The advantages of using the soft codes of pre-trained models are two-fold: (1) we do not need to spend much time labeling the images with object and scene annotations. (2) The soft codes also capture the co-occurrence of multiple objects and scenes. Since the knowledge of object and scene CNNs is explicitly exploited to obtain the soft codes of images, we call this transfer technique as *knowledge-based transfer*. Specifically, during the fine-tuning process, we minimize the following loss function:

$$\ell_{know}(D_e, \mathbf{F}) = \ell_C(D_e) + \alpha \ell_{soft}(D_e, \mathbf{F}), \tag{7}$$

where $\ell_C(D_e)$ is the loss function of event recognition as defined in Eq. (6), $\ell_{soft}(D_e, \mathbf{F})$ is the loss of measuring the distance between prediction and the soft codes produced by object or scene CNNs $\mathbf{F}$, and α is a weight to balance these two terms. The loss $\ell_{soft}(D_e, \mathbf{F})$ is formulated as follows:

$$\ell_{soft}(D_e, \mathbf{F}) = -\sum_{\mathbf{I}_i \in D_e} \sum_{k=1}^{K} q_{i,k} \log f_{i,k}, \tag{8}$$

where q_i the softmax output of the event network for object or scene prediction with image $\mathbf{I}_i$, and f_i is the softmax output of the pre-trained object or scene models, namely $f_i = \mathbf{F}(\mathbf{I}_i)$. Essentially, this is the cross-entropy loss and its goal is to make the transferred event models imitate the object or scene models.

In this transfer technique, we propose a multitask learning framework to jointly fine tune the network weights for event recognition and imitate the object and scene networks to recognize the objects and scene present. This additional imitation task exploits these soft codes of pre-trained models to guide the fine-tuning process and acts as a kind of regularizer to improve the generalization capacity of event models.

4.3 Data-Based Transfer

In this subsection, we approach learning the event models together with object and scene recognition from a different perspective. In order to incorporate the object and scene recognition into the fine-tuning process, we jointly train CNNs for relevant tasks on different datasets in a weight sharing scheme. We find that this weight sharing scheme is another effective strategy to regularize the fine-tuning of the event network.

Specifically, we simultaneously fine-tune two different networks on two datasets: one network is fine-tuned on the subset of ImageNet (Deng et al. 2009) or Places (Zhou et al. 2015) for object recognition or scene recognition, and the other one is fine tuned on the event recognition dataset to classify events. As shown in Fig. 4, these two networks have their own data layers to handle different datasets, fully-connected layers to deal with different targets, and loss layers to update the weights for different tasks. The weights of the remaining layers are shared by these two networks, which means that they should be updated synchronously during back propagation. Therefore, during the training process, we minimize the following loss function:

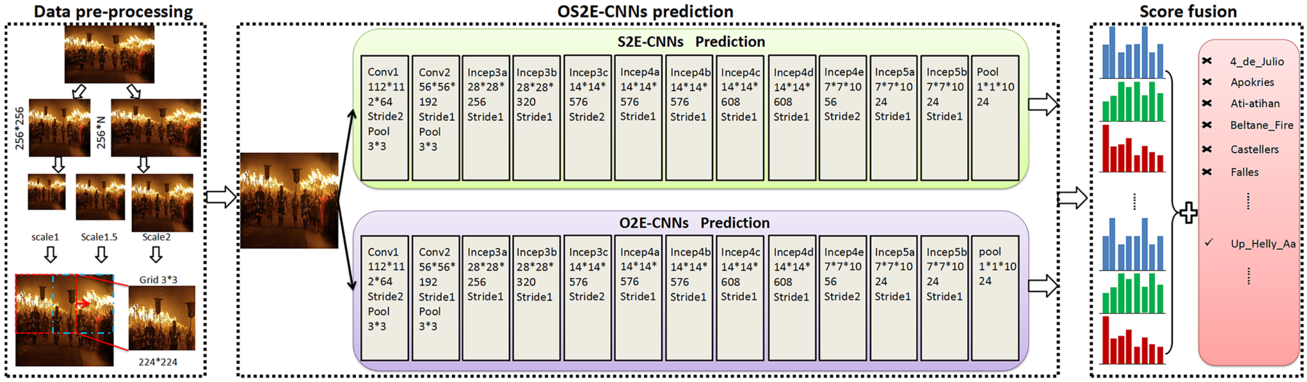
Data pre-processing
256*256
256*N
scale1
Scale1.5
Scale2
Grid 3*3
224*224
OS2E-CNNs prediction
S2E-CNNs Prediction
Conv1 112*112*64 Stride2 Pool 3*3
Conv2 56*56*192 Stride1 Pool 3*3
Incep3a 28*28*256 Stride1
Incep3b 28*28*320 Stride1
Incep3c 14*14*576 Stride2
Incep4a 14*14*576 Stride1
Incep4b 14*14*576 Stride1
Incep4c 14*14*608 Stride1
Incep4d 14*14*608 Stride1
Incep4e 7*7*1056 Stride2
Incep5a 7*7*1024 Stride1
Incep5b 7*7*1024 Stride1
Pool 1*1*1024
O2E-CNNs Prediction
Conv1 112*112*64 Stride2 Pool 3*3
Conv2 56*56*192 Stride1 Pool 3*3
Incep3a 28*28*256 Stride1
Incep3b 28*28*320 Stride1
Incep3c 14*14*576 Stride2
Incep4a 14*14*576 Stride1
Incep4b 14*14*576 Stride1
Incep4c 14*14*608 Stride1
Incep4d 14*14*608 Stride1
Incep4e 7*7*1056 Stride2
Incep5a 7*7*1024 Stride1
Incep5b 7*7*1024 Stride1
pool 1*1*1024
Score fusion
✗ 4_de_Julio
✗ Apokries
✗ Ati-atihan
✗ Beltane_Fire
✗ Castellers
✗ Falles
✓ Up_Helly_Aa

Fig. 5 Pipeline of event recognition with OS2E-CNNs. It is composed of three steps: (1) data pre-processing, (2) OS2E-CNNs prediction, and (3) score fusion. In the data pre-processing step we transform each image into a set of image regions with a multi-ratio and multi-scale cropping strategy. Then, these crop regions are fed into OS2E-CNNs to predict the score of different event classes. Finally, these scores of different image regions are fused to yield the final recognition result

$$\ell_{data}(D_e, D) = \ell_C(D_e) + \beta\ell_C(D), \tag{9}$$

where $\ell_C(D_e)$ is the loss function for event recognition on the target dataset as defined in Eq. (6), and $\ell_C(D)$ is the loss function on the auxiliary dataset for object or scene recognition, β is a parameter to keep a balance between these two terms.

This transfer technique aims to utilize useful information hidden in other datasets to help the fine-tuning of event models and we call this method as *data-based transfer*. The weight sharing scheme couples two networks together and jointly updates the network weights during fine-tuning process. This jointly updating convolutional weights is capable of exploiting the supervision information from two related datasets to guide network training, and can prevent it from over-fitting into any single dataset. The supervision information from the other dataset acts as a regularizer to improve the quality of fine-tuning on the target dataset. The proposed selection algorithm in Sect. 3.3 is able to identify a small set of discriminative and relevant object or scene classes. Simply utilizing the images from these subsets is able to enhance the relevance between main task and auxiliary tasks, thus contributing to improve the performance of event recognition.

5 Event Recognition with OS2E-CNNs

After the introduction of transfer techniques from object and scene models to event CNNs, we are ready to describe how to deploy these fine-tuned CNNs for event recognition.

The pre-trained object and scene CNNs are fine-tuned for event recognition on the event dataset and we call these learned networks **OS2E-CNNs**, which is short for *convolutional neural networks transferred from object and scene recognition to event recognition*. As shown in Fig. 5, the OS2E-CNNs are composed of two streams: (1) O2E-CNNs: event CNNs transferred from object models and object datasets. (2) S2E-CNNs: event CNNs transferred from scene models and scene datasets. In order to efficiently deploy OS2E-CNNs, our current implementation treats OS2E-CNNs as "end-to-end predictors" without training explicit classifiers such as SVMs. As illustrated in Fig. 5, our event recognition pipeline involves three steps: (1) data pre-processing, (2) network prediction, and (3) score fusion.

Data pre-processing We conduct common data augmentation techniques to enrich the image samples. Specifically, we design a *multi-ratio* and *multi-scale* cropping technique, to deal with the aspect and scale variations existed in the event images. More details about this cropping will be explained in Sect. 6.2.

Network prediction We first subtract their mean pixel value from the cropped image regions and then feed the results into the O2E-CNN and S2E-CNN independently. The O2E-CNN and S2E-CNN produce prediction scores at the softmax layer. These score vectors $S(\mathbf{R})$ represent the likelihood of event categories for each image region $\mathbf{R}$.

Score fusion We combine the scores of the different networks for different crops. First, for each cropped region $\mathbf{R}$, the prediction score of the OS2E-CNN is a weighted average of O2E-CNN and S2E-CNN results:

$$S_{os}(\mathbf{R}) = \alpha_o S_o(\mathbf{R}) + \alpha_s S_s(\mathbf{R}), \tag{10}$$

where $S_o(\mathbf{R})$ and $S_s(\mathbf{R})$ are the score vectors of the O2E-CNN and S2E-CNN for region $\mathbf{R}$. α_o and α_s are their fusion weights and in the current implementation are equal. Then, for the whole image $\mathbf{I}$, the prediction score is obtained by fusing across these cropped regions:

$$S_{os}(\mathbf{I}) = \sum_{\mathbf{R}_i \in \mathbf{I}} S_{os}(\mathbf{R}_i). \tag{11}$$

6 Experiments

In this section, we describe the detailed experimental setting and report the performance of our method. First, we introduce the datasets used for evaluation and their corresponding experimental setup. Next, we describe the implementation details of our method. Then, we investigate our method from the aspects of cropping strategy, object and scene selection, and transfer techniques. We also compare our method with the winners of the ICCV15 ChaLearn Looking at People (LAP) challenge. Furthermore, we fix the parameter settings and perform experiments on other event recognition datasets. Finally, we give examples for which our method fails to predict the correct labels.

6.1 Datasets and Evaluation Protocol

The **ICCV15 ChaLearn LAP challenge** (Escalera et al. 2015)[2] provides a large dataset for Cultural Event Recognition in still images. There are 100 event classes in total, including 99 event classes and 1 background class. The whole dataset is divided into three parts: development data (14,332 images), validation data (5,704 images), and evaluation data (8,669 images). The performance is measured by computing the average precision (AP) for each event class and reporting the mean AP over all the classes (mAP). We perform experiments under two different settings on this dataset. The first is the *validation setting*, where we train OS2E-CNN models on the development data and test on the validation data. We study different configurations of our method to determine the optimal setting. The second experiment uses the **challenge setting**, where we merge the development data and validation data into a single training dataset, and re-train OS2E-CNNs on this new training dataset. We send our recognition results to the challenge organizers and obtain the final performance back.

The **Web Image Dataset for Event Recognition** (Xiong et al. 2015)[3] is probably the largest image benchmark for event recognition. In its current version, there are 50,574 images and 61 event categories. The whole dataset is divided into 25,275 training images and 25,299 testing images. The evaluation measure is based on the mean recognition accuracy across all the event classes. WIDER focuses on event categories in our daily life, such as parade, dancing, meeting, press conference, and so on. Therefore, it is complementary to the ChaLearn Cultural Event Recognition dataset from the aspect of event classes.

The **UIUC Sports Event dataset** (Li and Li 2007)[4] is probably the first image benchmark for event recognition. It is composed of 8 sports event categories. The number of images in each event category ranges from 137 to 250. We follow the original evaluation setting, where 70 images of each class are selected as training samples and 60 images are selected as testing samples. The final evaluation is based on the mean recognition accuracy across the 8 event classes. Recognition accuracy has already saturated (around 95%) and it is difficult to achieve improvements over the state-of-the-art. Nevertheless, this dataset can verify the effectiveness of our proposed different transfer techniques if our method is still able to boost the recognition performance.

The **Photo Event Collection dataset** (Bossard et al. 2013)[5] is proposed for classifying the photo collections into pre-defined event classes such as birthday, wedding, graduation, and so on. This dataset contains more than 61,000 images from 807 collections, annotated with 14 social event classes. We use this dataset to verify the effectiveness of our proposed method for image-level event recognition, although it is mainly designed for collection-level event classification. We follow the train and test split released by the original paper, where each class has 30 collections for training and 10 collections for testing. We directly assign the collection-level label to each image contained in this collection and simply use the image itself for event recognition, without any meta information such as temporal information.

6.2 Implementation Details

Training. During the fine-tuning procedure of OS2E-CNNs, we first resize each training image to 256 × 256. At each iteration, we randomly crop a region from the whole image. To deal with scale and aspect ratio variations, we design a multi-scale and multi-ratio cropping strategy, where the cropped width w and height h are randomly picked from $\{256, 224, 192, 160, 128\}$. Then these cropped regions are resized to 224 × 224 for network training. These cropped regions also undergo random horizontal flipping. The network weights are learned using the mini-batch stochastic gradient descent with momentum (set to 0.9). At each iteration, a mini-batch of 256 images is constructed by random sampling. The dropout ratio for the added dropout layer is set as 0.7. As we pre-train the network weights with the ImageNet and Places models, we set a smaller learning rate for fine-tuning. Specifically, we initialize the learning rate as 0.01. After this, we decrease the learning rate by a factor of 10 every K iterations. The whole training proce-

[2] http://gesture.chalearn.org/.

[3] http://personal.ie.cuhk.edu.hk/~xy012/event_recog/WIDER/.

[4] http://vision.stanford.edu/lijiali/event_dataset/.

[5] https://www.vision.ee.ethz.ch/datasets_extra/pec/.

dure stops at $2.5K$ iterations. K is related to the training set size and we set it to 5000 for the validation setting of ChaLearn Cultural Event Recognition, 7000 for the challenge setting of this dataset, 10,000 for the Web Image Dataset for Event Recognition, 300 for the UIUC Sports Event dataset, and 1000 for the Photo Collection Event dataset.

Testing. We first resize the image with two different aspect ratios: (1) keeping the image aspect ratio fixed and setting the smaller side to 256, (2) resizing the image to 256 × 256. Second, to deal with scale variations, we change the image resolutions with the smaller side as 384, 512, or the whole size as 384 × 384, 512 × 512. Finally, we densely crop 224 × 224 regions in a grid of 3 × 3 from these images. Hence, we crop a total number of $2 \times 3 \times 9 = 54$ regions from a single image.

6.3 Exploration of Testing Strategy

We begin our experiment by exploring the effectiveness of the multi-ratio and multi-scale cropping strategy proposed in Sect. 5. Specifically, in this experiment, we use the ChaLearn Cultural Event Recognition dataset under the validation setting and choose initialization-based transfer to learn the OS2E-CNN models. The results are reported in Table 1 and we notice that the strategy of multi-ratio and multi-scale cropping is helpful for improving recognition performance.

First, given a fixed image aspect ratio, we resize the image to three different scales: (1) the original scale, (2) 1.5 times that scale, and (3) double the scale. The performance of using three different scales is improved to 85.3% for the OS2E-CNN compared with the original performance 83.4%. This improvement suggests the multi-scale cropping method can handle scale variations of the test images. Second, we choose two aspect ratios for the testing images (256 × N vs. 256 × 256) and the first aspect ratio obtains better performance (85.3% vs. 85.0%). We fuse the recognition results of these two aspect ratios and boost the recognition performance to 85.6%. This improvement may be ascribed to an aspect ratio difference among test images and testing on different ratios could be helpful to handle this issue. In the remainder of this section, we will use this multi-scale and multi-ratio cropping technique for other experimental explorations.

6.4 Evaluation on Object and Scene Selection

In this subsection, we verify the effectiveness of the object and scene selection method proposed in Sect. 3.3, on the ChaLearn Cultural Event Recognition dataset under the validation setting. Our experiment is performed in two scenarios: (1) using the object and scene responses as features for event recognition, and (2) transferring deep object and scene representations for event recognition via a data based transfer method. We also visualize the conditional probability of top selected object and classes.

First, we treat the object and scene responses $\Phi^{o}(\mathbf{I})$ and $\Phi^{s}(\mathbf{I})$ as features and train a linear SVM to classify event classes. To make the training of SVMs more stable, we normalize these responses with ℓ_2-norm (Vedaldi and Zisserman 2012). The experimental results are reported in Table 2. The object responses achieve a mAP of 70.2% and scene responses obtain a mAP of 61.5%. The feature concatenation of both responses further improves the performance to 72.1%. Following this, we plot the mAP values with different

Table 1 Performance of different cropping strategies on the ChaLearn Cultural Event Recognition dataset under the *validation setting*

Ratio	Scale	O2E-CNNs (%)	S2E-CNNs (%)	OS2E-CNNs (%)
256 × N	Scale 1	82.1	80.5	84.2
	Scale 1.5	81.8	81.1	84.1
	Scale 2	77.2	76.1	79.4
	Combine	83.4	82.8	**85.3**
256 × 256	Scale 1	80.4	78.2	82.8
	Scale 1.5	82.4	80.8	84.3
	Scale 2	81.7	80.5	83.5
	Combine	83.2	82.0	**85.0**
Combine	Scale 1	82.0	80.3	84.1
	Scale 1.5	83.2	82.1	85.0
	Scale 2	81.2	80.3	83.0
	Combine	83.9	83.0	**85.6**

Bold values indicate the best results
We use two image aspect ratios and three different scales.
These cropped regions from different resolutions and scales are complementary to each other

Table 2 Event recognition results using the object and scene responses on the ChaLearn Cultural Event Recognition dataset under the *validation setting*

Responses	Objects (%)	Scenes (%)	Fusion (%)
Performance (mAP)	70.2	61.5	72.1

Objects and scenes can provide useful visual cues for event recognition, which achieves a relatively high recognition performance

The combination of objects and scenes can further boost the recognition performance

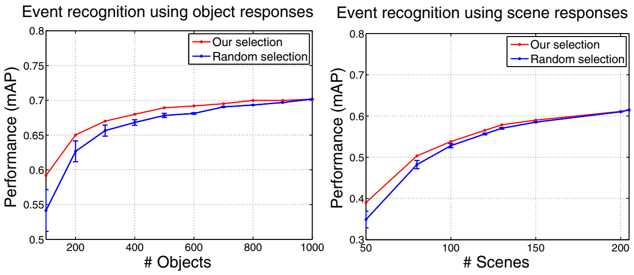

Fig. 6 Exploration of the object and scene selection algorithm on the ChaLearn Cultural Event Recognition dataset under the *validation setting*. The results are compared with the random selection algorithm

numbers of selected object and scene classes in Fig. 6. We compare our method with a baseline of random selection. The performance gap is relatively large when the selected number is small and this gap becomes smaller when more classes are selected. As more classes are selected, those discriminative object and scene classes are more easily picked by a random sampling method. Our method can select a subset of classes, that achieves the 95% performance of using all classes. For instance, it obtains a performance of around 67% with 300 object classes and 59% with 150 scene classes.

Then, we study the effectiveness of object and scene selection for data-based transfer method. We perform an exploration study on the number of selected classes, and the experimental results are shown in the left of Fig. 7. Selecting subsets of discriminative classes is able to enhance the relevance of the main task and auxiliary tasks in the data-based transfer method, thus improving the performance of the main task (i.e., event recognition) (Caruana 1997). From experimental results, the number of selected object and scene classes has an influence on the relevance of the main task and auxiliary tasks, where 300 object classes and 150 scene classes achieve the best performance. Therefore, we fix the number of selected object classes as 300 and scene classes as 150 in the remaining experiments.

Next, we compare with two other methods in the data-based transfer method: (1) using all object and scene classes, (2) selecting 300 object classes and 150 scene classes with random selection. For fair comparison, these three methods all use the same pre-training models, namely CNNs pre-trained on all object and scene classes. The results are summarized in the right of Fig. 7. For the O2E-CNNs, using all the objects (1000 classes) achieves the performance of 85.0%, which is lower than that of employing 300 object classes. Using a smaller number of object classes may enhance the relevance of main and auxiliary tasks. Comparing the performance of random selection and our proposed selection method, this confirms the effectiveness of considering discriminative and diversity capacity during the selection process. For S2E-CNNs, similar results to O2E-CNNs are observed, and our proposed selection algorithm outperforms the other two methods. Meanwhile, we also try the sequential tree-weighted algorithm (TRW-S) (Kolmogorov 2006) to solve the optimization problem and it achieves a similar performance to our greedy selection method (85.4% for O2E-CNN and 84.8% for S2E-CNNs).

Finally, we visualize the conditional probabilities $p(e|o)$ and $p(e|s)$ of selected objects and scenes in Fig. 8. Here we only plot the top 30 selected classes, listed in the same order as they were selected. Our selection algorithm is able to choose discriminative object and scene classes, while keeping their diversity. For example, the first selected object class is `bison`, which is good at discriminating the event `annual bufallo roundup` from other classes, and the first selected scene class is `desert/sand`, a strong indicator for event class `afrika burn` and `sahara festival`.

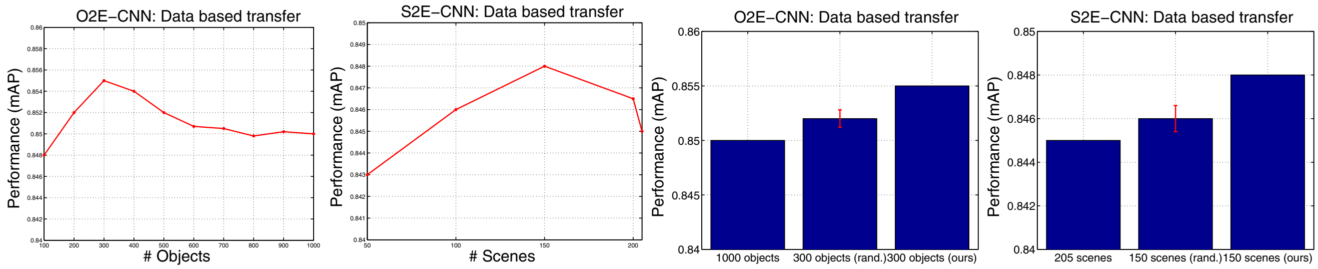

Fig. 7 Exploration of the object and scene selection algorithm on the ChaLearn Cultural Event Recognition dataset under the *validation setting*. We choose the data-based transfer technique to study the effect of selecting a subset of objects and scenes. Left: we study the influence of the number of selected object and scene classes. Right: we compare with the random selection and usage of all classes

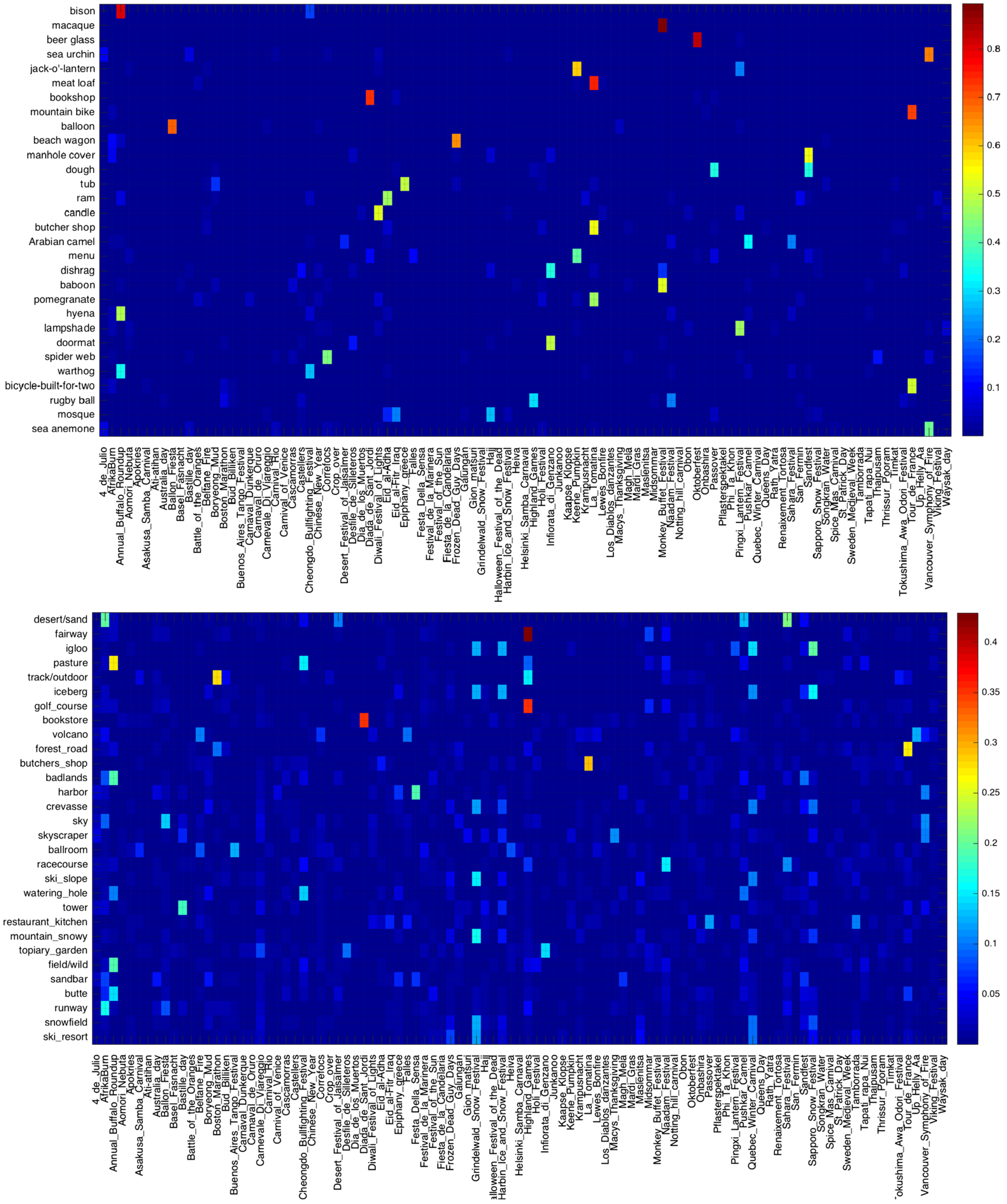

Fig. 8 Visualization of the conditional probabilities $p(e|o)$ and $p(e|s)$ of selected objects and scenes for the ChaLearn Cultural Recognition dataset. For visual clarification, we plot the top 30 objects (top row) and scenes (bottom row) in the order they are selected. From these results, we see that our selection method is able to find a subset of discriminative and diverse objects and scenes

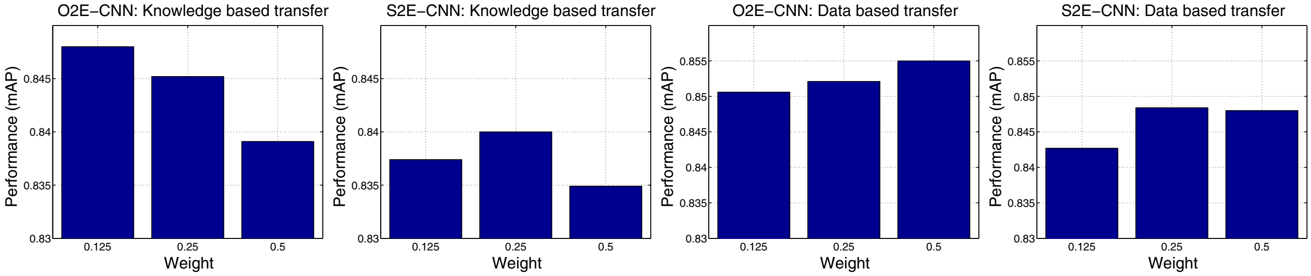

Fig. 9 Performance of different weights of auxiliary tasks on the ChaLearn Cultural Event Recognition dataset under the *validation setting*. We study both knowledge based transfer and data based transfer methods and aim to find the optimal parameter setting

Meanwhile, our selected object and scene classes appear across different event classes and ensure the subset diversity.

6.5 Exploration of Auxiliary Task Weights

We now explore the influence of different weights in the multitask transfer framework: (1) knowledge-based transfer, and (2) data-based transfer, on the ChaLearn Cultural Event Recognition dataset under the validation setting. An important parameter in this multitask framework is the weights of auxiliary tasks, namely parameters α and β in Eqs. (7) and (9).

We first study the effect of weight α in knowledge-based transfer. As our goal is to perform event recognition, we constrain the weight of the auxiliary task to be less than 0.5. Specifically, we choose three different weights 0.125, 0.25, and 0.5, and the experimental results of both the O2E-CNN and S2E-CNN are shown in the left of Fig. 9. For the O2E-CNN, smaller weights achieve better performance and the weight of 0.125 gets the best performance of around 84.8%. However, for the S2E-CNN, the best weight is 0.25, where it obtains a performance of around 84.0%. The effect of over-fitting is more serious for the S2E-CNN than the O2E-CNN, and we need to set a higher weight for the auxiliary task to better regularize the training of event CNNs.

We then compare the performance of O2E-CNNs and S2E-CNNs by using different weight values in data-based transfer method. The results are reported in the right of Fig. 9. The performance of data-based transfer is less sensitive to the weight setting, where weight 0.125 achieves the lowest performance, and the weights 0.25 and 0.5 obtain a similar performance. Hence, in the remaining experimental explorations, we fix the weight of the auxiliary task to 0.5 for both the O2E-CNN and S2E-CNN in the data-based transfer method.

6.6 Comparison of Transfer Techniques

In this subsection we study the performance of different transfer techniques proposed in Sect. 4. We test them on the ChaLearn Cultural Event recognition dataset under the validation setting. For fair comparison, these three transfer methods are all initialized with the models pre-trained on all object and scene classes.

First, we compare the performance of using different pre-trained models: object CNNs pre-trained on the ImageNet dataset and scene CNNs pre-trained on the Places dataset. From these results in Table 3, we observe that deep representations transferred from object CNNs outperform those transferred from scene CNNs. The superior performance of O2E-CNNs may imply that the objects more strongly correlate with events, tallying with the fact that the selected object classes in Fig. 8 have lower conditional entropy and yield stronger discriminative capacity than the selected scene classes. Furthermore, we fuse the prediction results of O2E-CNNs and S2E-CNNs, enabling further improvements in recognition performance.

Then, we compare the recognition results of three transfer methods. We see that initialization-based transfer is already effective for fine-tuning event CNNs, and it obtains a performance of 85.6% for OS2E-CNNs. The newly designed knowledge-based transfer and data-based transfer achieve better performance, which indicates that incorporating relevant tasks into the fine-tuning process contributes to improve the generalization ability of the final event models. Data-based transfer is better than knowledge-based transfer but requires additional training images. Furthermore, we fuse the prediction scores of knowledge-based transfer and data-based transfer, getting a slightly better performance.

Finally, we study the fine-tuning procedure with more details. Specifically, we plot the training and testing loss of three transfer methods in Fig. 10. First, there exists a gap between the training and test loss for all transfer methods. This indicates over-fitting is still a serious problem for fine-tuning CNNs on a small dataset. Second, we see that the effect of over-fitting is more severe in the scenario of initialization-based transfer. As the iteration number increases, the test loss stops decreasing and even increases by around 0.3. On the other hand, for knowledge-based and data-based transfer, the extra tasks are helpful to reduce the degree of over-fitting throughout.

Table 3 Performance of different transfer techniques on the ChaLearn Cultural Event Recognition dataset under the *validation setting*

Method	O2E-CNNs (%)	S2E-CNNs (%)	OS2E-CNNs (%)
Initialization	83.9	83.0	85.6
Knowledge	84.8	84.0	86.3
Data	85.5	84.8	87.0
Know.+Data	85.6	85.4	87.2
ALL	86.0	85.6	87.2

We compare our proposed three transfer techniques and the data-based transfer achieves the best performance for both O2E-CNN and S2E-CNN

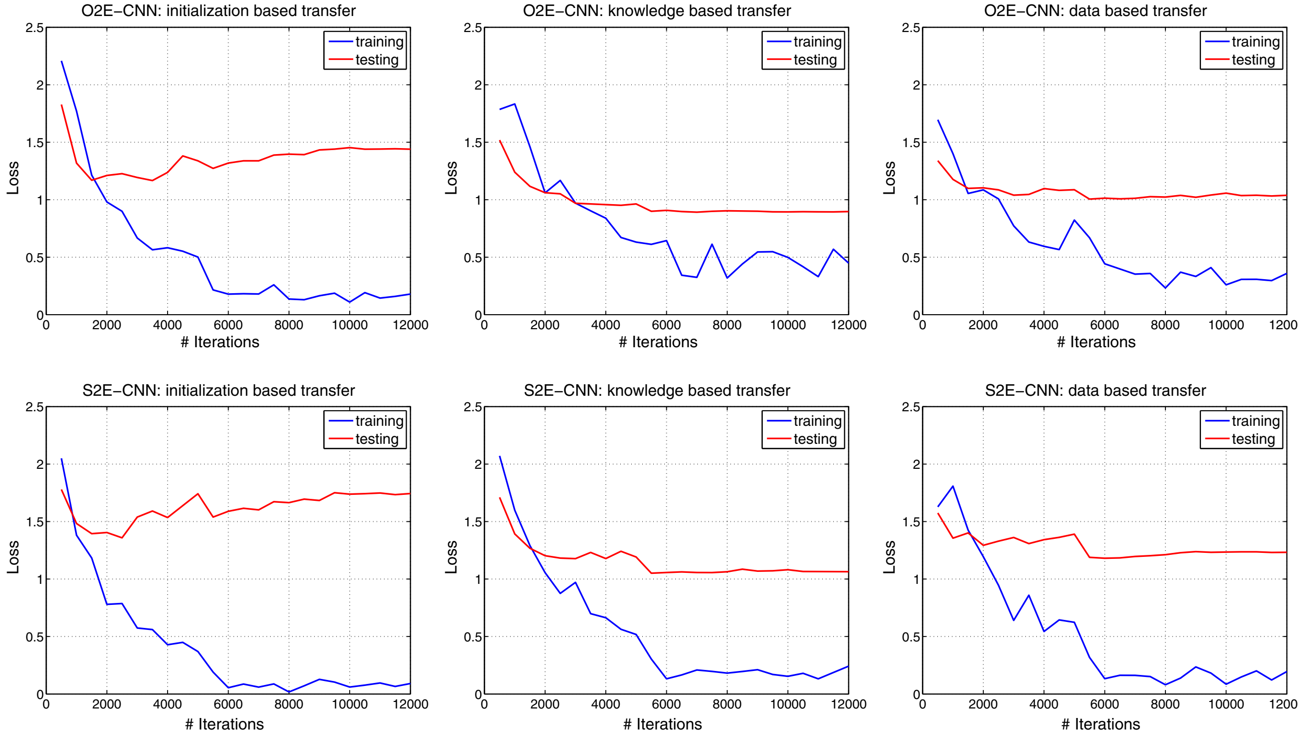

Fig. 10 Training and testing the loss of O2E-CNN (top row) and S2E-CNN (bottom row) on the ChaLearn Cultural Event Recognition dataset under the *validation setting*. We compare our proposed three transfer techniques. The results indicate that knowledge-based transfer and data-based transfer help to reduce the effect of over-fitting

6.7 Challenge Results

In this subsection we report the experimental results on ChaLearn Cultural Event Recognition dataset under the *challenging setting*. We can not access the labels of testing images and the parameter settings are determined according to the study under the validation setting as described in previous subsections.

The numerical results are reported in Table 4. We compare the performance of our proposed method with the winners of the ICCV ChaLearn Looking at People (LAP) challenge (Wei et al. 2015; Liu et al. 2015; Wang et al. 2015c; Rothe et al. 2015). These winner solutions all employ the pre-trained models learning from ImageNet and Places, so it is fair to compare our method with them. The performance of initialization-based transfer achieves a mAP value

Table 4 Performance of different transfer techniques on the ChaLearn Cultural Event Recognition dataset under *challenge setting*

Method	Networks	Explicit classifiers	Performance
CAS	4	LDA + LR	85.4
FV	5	SPM + FV + LR	85.1
MMLAB	4	FV + SVM	84.7
CVL_ETHZ	2	LDA + *k*-NN	79.8
Initialization	2	None	85.9
Knowledge	2	None	86.2
Data	2	None	86.9
Data + Know.	4	None	87.0
All	6	None	**87.1**

Bold value indicates the best results
Our method outperforms these winners of the ICCV ChaLearn Looking at People (LAP) challenge

of 85.9%, which outperforms all the winners. This result may be ascribed to the better network structure and the proposed multi-ratio and multi-scale cropping strategy. We also notice that the newly designed multitask transfer techniques obtain higher mAP values, and the data-based transfer method gets the best performance of 86.9% among the three transfer methods. Finally, we combine the prediction results of different transfer techniques, which yields the best performance of 87.1% on the test data of the ChaLearn Cultural Event Recognition dataset.

6.8 Evaluation on Other Event Datasets

We present the experimental results of our method on the other event recognition datasets in this subsection. Specifically, we perform experiments on the Web Image Dataset for Event Recognition (WIDER) (Xiong et al. 2015), the UIUC Event8 dataset (Li and Li 2007), and the Photo Event Collection dataset (PEC) (Bossard et al. 2013). We use the same parameter settings with the ChaLearn Cultural Event Recognition dataset (i.e., no specific tuning).

First, we report the numerical results on WIDER in Table 5. We see that our newly proposed transfer methods outperform initialization-based transfer, in keeping with our findings on the ChaLearn Cultural Event Recognition dataset. Knowledge-based transfer and data-based transfer improve the performance of initialization-based transfer by 1.2 and 1.6%, respectively. We also compare the performance of our method with two other approaches: (1) baseline CNN models and (2) deep channel fusion (Xiong et al. 2015), which obtained the state-of-the-art performance on this dataset. Our initialization based transfer method is able to improve the performance by 8.4% and our multitask based transfer is better than previous method by around 10.6%.

Second, the results on the UIUC Event8 dataset are summarized in Table 6. This dataset is relatively small and the state-of-the-art performance is very high (around 95%). Our baseline of initialization-based transfer achieves a performance of 96.9%, and our new transfer methods are still able to boost the performance to 98.8 and 98.0%. The performance of data-based transfer is a bit lower than that of knowledge-based transfer. This could be ascribed to the smaller size of the UIUC Event8 dataset, which requires fewer iterations before convergence. So, the fine-tuning is not able to fully exploit the extra ImageNet and Places images in data-based transfer. We compare with the baseline of Couple LDA (Li and Li 2007) and recent deep learning methods (Zhou et al. 2014, 2015). Our proposed transfer methods outperform these previous approaches and obtain the state-of-the-art performance of 98.8% on this dataset.

Finally, we report the performance of event recognition on the Photo Event Collection dataset (PEC). We simply use the original image without temporal information to perform image-level event recognition. As this dataset is designed for collection classification, no previous works report performance on image-level event recognition. For completeness, we also report the performance of several methods using temporal information for collection-level event recognition,

Table 5 Event recognition performance on the Web Image Dataset for Event Recognition (WIDER)

Method	Performance (%)
Baseline CNN (Xiong et al. 2015)	39.7
Deep channel fusion (Xiong et al. 2015)	42.4
Initialization	50.8
Knowledge	52.0
Data	52.6
Data + Know.	**53.0**
All	52.8

Bold value indicates the best results
We test our proposed transfer methods and compare with the state-of-the-art performance

Table 6 Event recognition on the UIUC Sports Event dataset

Method	Accuracy (%)
Couple LDA (Li and Li 2007)	73.4
ImageNet CNN Feature (Zhou et al. 2014)	94.4
Places CNN Feature (Zhou et al. 2014)	94.1
GoogLeNet GAP (Zhou et al. 2015)	95.0
Initialization	96.9
Knowledge	**98.8**
Data	98.0
Data + Know.	98.4
All	98.2

Bold value indicates the best results
We test our proposed transfer methods and compare with the state-of-the-art performance

Table 7 Event recognition on the Photo Collection Event dataset (PEC)

Method	Accuracy (%)
Aggregated SVM (Bossard et al. 2013)	41.4
Bag of Sub-events (Bossard et al. 2013)	51.4
HMM (Bossard et al. 2013)	53.6
SHMM (Bossard et al. 2013)	55.7
Initialization	60.6
Knowledge	**62.0**
Data	61.7
Data + Know.	62.2
All	61.9

Bold value indicates the best results
We test our proposed transfer methods and compare with the state-of-the-art performance

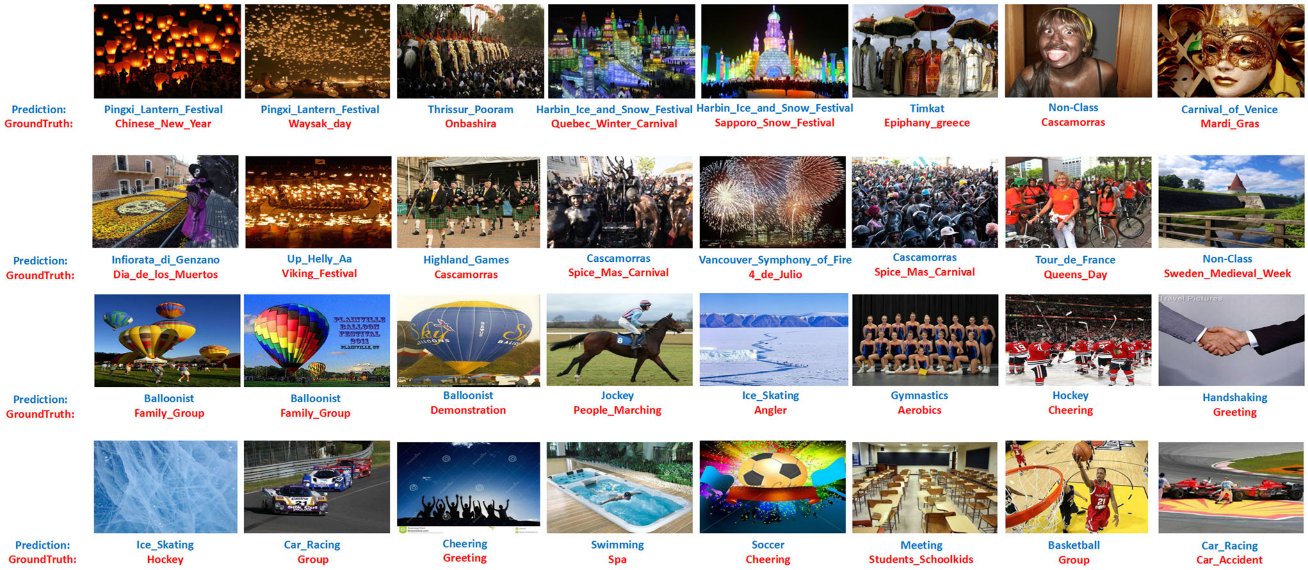

Fig. 11 We show several failure cases from the ChaLearn Cultural Event Recognition dataset at the top 2 rows and from the Web Image Dataset for Event Recognition (WIDER) at the bottom 2 rows. We notice that sometimes the ground truth labels contain noise and our prediction results seem more reasonable. For example, the second image of last row is more likely to be event `car racing` than `group` and the seventh image of last row is more likely to be event of `basketball` than `group`

including Aggregated SVM, HMM, SHMM. It should be noted that these numbers cannot be directly compared with our result due to the different evaluation tasks. We mainly compare the performance of our proposed transfer methods and the results are summarized in Table 7. We see that our newly designed transfer methods outperform the initialization based transfer consistently, which again demonstrates the superiority of our proposed multitask based transfer methods to the fine-tuning baseline.

6.9 Visualization of Recognition Examples

Several failure examples are given in Fig. 11. In these cases, our method produces a wrong label with high confidence. In the top two rows, we show some failure cases from the ChaLearn Cultural Event Recognition dataset. From these samples, we see that the event class `Chinese new year` may be easily confused with the event class `pingxi lattern festival`, that the event `harbin ice and snow festival` comes close in appearance to the events `sapporo snow festival` and `quebec winter carnival`, that the event class `carnival of venice` looks like `mardi gras`, and so on. In the bottom two rows, we give some failure cases from the Web Image Dataset for Event Recognition (WIDER). We see that our method may confuse the class `balloonist` with the class `family group`, the class `jockey` with the class of `people marching`, the class `gymnastics` with the class `aerobics`, and so on. Also, we notice that sometimes the ground truth contains noise and our prediction results seem more reasonable. For example, the second image of last row is more likely to be event `car racing` than `group` and the seventh image of last row is more likely to be event of `basketball` than `group`.

7 Conclusions

We have presented a deep architecture, coined as OS2E-CNN, for event recognition in still images. It transfers deep representations from object and scene models to the event recognition task. Objects, scenes, and events are indeed semantically related. We empirically studied the relation among categories of object, scene, and event. It appears that the likelihood of object and scene classes matters for event understanding in still images. Yet, not all object and scene classes strongly correlate with the event classes, and we designed an effective method to select a subset of discriminative and diverse object and scene classes, which helps us better fine tune deep representations in the data-based transfer method. To adapt these deep learned representations of object and scene models, we developed three transfer methods: (1) initialization-based transfer, (2) knowledge-based transfer, and (3) data-based transfer. The latter two transfer techniques exploit multitask learning frameworks to incorporate the extra knowledge from other networks or extra data from public datasets into the fine-tuning procedure of event models. It turns out that these new transfer methods are effective to reduce over-fitting and to improve the generalization ability. Our method achieves the state-of-the-art performance and outperforms competing approaches on four public benchmarks.

In the future, we may consider incorporating more semantic cues such as human pose and garments into a unified framework for event recognition from still images. The concept of event is a higher-level concept than other semantic ones such as objects and scenes, and we consider investigating into a new recognition framework, that is able to exploit the hierarchical structure among the task of object recognition, scene recognition, and event recognition.

Acknowledgements This work is partially supported by the ERC Advanced Grant VarCity, the Toyota Research Project TRACE-Zurich, the National Key Research and Development Program of China (2016YFC1400704), the National Natural Science Foundation of China (U1613211, 61633021), and the External Cooperation Program of BIC Chinese Academy of Sciences (172644KYSB20150019, 172644KYSB20160033).

References

Azizpour, H., Sharif Razavian, A., Sullivan, J., Maki, A., & Carlsson, S. (2015). From generic to specific deep representations for visual recognition. In *CVPR Workshop on DeepVision*, pp. 36–45.

Baro, X., Gonzalez, J., Fabian, J., Bautista, M. A., Oliu, M., Jair Escalante, H., Guyon, I., & Escalera, S .(2015). Chalearn looking at people 2015 challenges: Action spotting and cultural event recognition. In *CVPR Workshop on ChaLearn Looking at People,* pp. 1–9.

Bhattacharya, S., Kalayeh, M. M., Sukthankar, R., & Shah, M. (2014). Recognition of complex events: Exploiting temporal dynamics between underlying concepts. In *CVPR*, pp. 2243–2250.

Bossard, L., Guillaumin, M., & Gool, L. J. V. (2013). Event recognition in photo collections with a stopwatch HMM. In *ICCV*, pp. 1193–1200.

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*(1), 41–75.

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, pp. 1–12.

Cooper, M. L., Foote, J., Girgensohn, A., & Wilcox, L. (2003). Temporal event clustering for digital photo collections. In *ACM Multimedia*, pp. 364–373.

Das, A., Dasgupta, A., & Kumar, R. (2012). Selecting diverse features via spectral regularization. In *NIPS*, pp. 1592–1600.

Delaitre, V., Sivic, J., & Laptev, I. (2011). Learning person-object interactions for action recognition in still images. In *NIPS*, pp. 1503–1511.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Li, F. (2009). ImageNet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255.

Desai, C., & Ramanan, D. (2012). Detecting actions, poses, and objects with relational phraselets. In *ECCV*, pp. 158–172.

Duan, L., Xu, D., Tsang, I. W., & Luo, J. (2012). Visual event recognition in videos by learning from web data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(9), 1667–1680.

Ebadollahi, S., Xie, L., Chang, S., & Smith, J. R. (2006). Visual event detection using multi-dimensional concept dynamics. In *ICME*, pp. 881–884.

Escalera, S., Fabian, J., Pardo, P., Baro, X., Gonzalez, J., Escalante, H. J., Misevic, D., Steiner, U., & Guyon, I. (2015). Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *ICCV Workshop on ChaLearn Looking at People*, pp. 1–9.

Everingham, M., Gool, L. J. V., Williams, C. K. I., Winn, J. M., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Fernando, B., Habrard, A., Sebban, M., & Tuytelaars, T. (2013). Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, pp. 2960–2967.

Gan, C., Wang, N., Yang, Y., Yeung, D., & Hauptmann, A. G. (2015). Devnet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*, pp. 2568–2577.

Gao, B., Wei, X., Wu, J., & Lin, W. (2015). Deep spatial pyramid: The devil is once again in the details. CoRR abs/1504.05277.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pp. 580–587.

Gkioxari, G., Girshick, R. B., & Malik, J. (2015). Contextual action recognition with r*cnn. In *ICCV*, pp. 1080–1088.

Gong, B., Grauman, K., & Sha, F. (2014). Learning kernels for unsupervised domain adaptation with applications to visual object recognition. *International Journal of Computer Vision*, *109*(1–2), 3–27.

Habibian, A., & Snoek, C. G. M. (2014). Recommendations for recognizing video events by concept vocabularies. *Computer Vision and Image Understanding*, *124*, 110–122.

Hauptmann, A. G., Christel, M. G., & Yan, R. (2008). Video retrieval based on semantic concepts. *Proceedings of the IEEE*, *96*(4), 602–622.

He, K., Zhang, X., Ren, S., & Sun. J. (2015). Deep residual learning for image recognition. CoRR abs/1512.03385.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pp. 448–456.

Izadinia, H., & Shah, M. (2012). Recognizing complex events using large margin joint low-level event model. In *Computer Vision—ECCV 2012–12th European Conference on Computer Vision*, Florence, Italy, October 7–13, 2012, Proceedings, Part IV, pp. 430–444.

Jain, M., van Gemert, J. C., & Snoek, C. G. M. (2015). What do 15,000 object categories tell us about classifying and localizing actions? In *CVPR*, pp. 46–55.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R. B., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. CoRR abs/1408.5093.

Juneja, M., Vedaldi, A., Jawahar, C. V., & Zisserman, A. (2013). Blocks that shout: Distinctive parts for scene classification. In *CVPR*, pp. 923–930.

Kolmogorov, V. (2006). Convergent tree-reweighted message passing for energy minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(10), 1568–1583.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *NIPS*, pp. 1106–1114.

Kulis, B., Saenko, K., & Darrell, T. (2011). What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, pp. 1785–1792.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Li, L., & Li, F. (2007). What, where and who? classifying events by scene and object recognition. In *ICCV*, pp. 1–8.

Liu, J., Yu, Q., Javed, O., Ali, S., Tamrakar, A., Divakaran, A., Cheng, H., & Sawhney, H. S. (2013). Video event recognition using concept attributes. In *WACV*, pp. 339–346.

Liu, M., Liu, X., Li, Y., Chen, X., Hauptmann, A. G., & Shan, S. (2015). Exploiting feature hierarchies with convolutional neural networks for cultural event recognition. In *ICCV Workshop on ChaLearn Looking at People*, pp. 32–37.

Ma, Z., Yang, Y., Sebe, N., & Hauptmann, A. G. (2014). Knowledge adaptation with partiallyshared features for event detectionusing few exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(9), 1789–1802.

Marszalek, M., Laptev, I., & Schmid, C. (2009). Actions in context. In *CVPR*, pp. 2929–2936.

Mazloom, M., Gavves, E., & Snoek, C. G. M. (2014). Conceptlets: Selective semantics for classifying video events. *IEEE Transactions on Multimedia*, *16*(8), 2214–2228.

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, pp. 1717–1724.

Park, S., & Kwak, N. (2015). Cultural event recognition by subregion classification with convolutional neural network. In *CVPR Workshop on ChaLearn Looking at People*, pp. 45–50.

Ramanathan, V., Tang, K. D., Mori, G., & Li, F. (2015). Learning temporal embeddings for complex video analysis. In *ICCV*, pp. 4471–4479.

Rothe, R., Timofte, R., & Van Gool, L. (2015). Dldr: Deep linear discriminative retrieval for cultural event classification from a single image. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, pp. 53–60.

Salvador, A., Zeppelzauer, M., Manchon-Vizuete, D., Calafell, A., & Giro-i Nieto, X. (2015). Cultural event recognition with visual convnets and temporal models. In *CVPR Workshop on ChaLearn Looking at People*, pp. 36–44.

Sharif Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). Cnn features off-the-shelf: An astounding baseline for recognition. In *CVPR Workshop on DeepVision*, pp. 806–813.

Shen, L., Lin, Z., & Huang, Q. (2016). Relay backpropagation for effective learning of deep convolutional neural networks. In *ECCV*, pp. 467–482.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *NIPS*, pp. 568–576.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR*, pp. 1–8.

Tang, K. D., Li, F., & Koller, D. (2012). Learning latent temporal structure for complex event detection. In *CVPR*, pp. 1250–1257.

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *CVPR*, pp. 1521–1528.

Tzeng, E., Hoffman, J., Darrell, T., & Saenko, K. (2015). Simultanously deep transfer across domains and tasks. In *ICCV*, pp. 4068–4076.

Vedaldi, A., & Zisserman, A. (2012). Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(3), 480–492.

Vu, T., Olsson, C., Laptev, I., Oliva, A., & Sivic, J. (2014). Predicting actions from static scenes. In *ECCV*, pp. 421–436.

Wang, L., Guo, S., Huang, W., Xiong, Y., & Qiao, Y. (2017). Knowledge guided disambiguation for large-scale scene classification with multi-resolution cnns. *IEEE Transactions on Image Processing*, *26*(4), 2055–2068.

Wang, H., Kläser, A., Schmid, C., & Liu, C. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, *103*(1), 60–79.

Wang, H., Oneata, D., Verbeek, J. J., & Schmid, C. (2016a). A robust and efficient video representation for action recognition. *International Journal of Computer Vision*, *119*(3), 219–238.

Wang, L., Qiao, Y., & Tang, X. (2015a). Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, pp. 4305–4314.

Wang, L., Qiao, Y., & Tang, X. (2016b). MoFAP: A multi-level representation for action recognition. *International Journal of Computer Vision*, *119*(3), 254–271.

Wang, L., Wang, Z., Du, W., & Qiao, Y. (2015b). Object-scene convolutional neural networks for event recognition in images. In *CVPR Workshop on ChaLearn Looking at People*, pp. 30–35.

Wang, L., Wang, Z., Guo, S., & Qiao, Y. (2015c). Better exploiting oscnns for better event recognition in images. In *ICCV Workshop on ChaLearn Looking at People*, pp. 45–52.

Wang, L., Xiong, Y., Wang, Z., & Qiao, Y. (2015d). Towards good practices for very deep two-stream convnets. CoRR abs/1507.02159.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016c). Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pp. 20–36.

Wei, X. S., Gao, B. B., & Wu, J. (2015). Deep spatial pyramid ensemble for cultural event recognition. In *ICCV Workshop on ChaLearn Looking at People*, pp. 38–44.

Xiong, Y., Zhu, K., Lin, D., & Tang, X. (2015). Recognize complex events from static images by fusing deep channels. In *CVPR*, pp. 1600–1609.

Yan, Y., Yang, Y., Shen, H., Meng, D., Liu, G., Hauptmann, A. G., & Sebe, N. (2015). Complex event detection via event oriented dictionary learning. In *AAAI*, pp. 3841–3847.

Yang, Y., Yang, Y., Huang, Z., Liu, J., & Ma, Z. (2012). Robust cross-media transfer for visual event detection. In *Proceedings of the 20th ACM International Conference on Multimedia*, pp. 1045–1048.

Yao, B., Jiang, X., Khosla, A., Lin, A. L., Guibas, L. J., & Li, F. (2011). Human action recognition by learning bases of action attributes and parts. In *ICCV*, pp. 1331–1338.

Yao, B., & Li, F. (2010). Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, pp. 9–16.

Yao, J., Fidler, S., & Urtasun, R. (2012). Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, pp. 702–709.

Zheng, J., Jiang, Z., Chellappa, R., & Phillips, P. J. (2014). Submodular attribute selection for action recognition in video. In *NIPS*, pp 1341–1349.

Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., & Torralba, A. (2015). Learning deep features for discriminative localization. CoRR abs/1512.04150.

Zhou, B., Lapedriza, À., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In *NIPS*, pp. 487–495.