Towards Universal Representation for Unseen Action Recognition

Yi Zhu[1], Yang Long[*2], Yu Guan[2], Shawn Newsam[1], and Ling Shao[3]

[1]University of California, Merced
[2]Open Lab, School of Computing, Newcastle University, UK.
[3]Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE.

Abstract

Unseen Action Recognition (UAR) aims to recognise novel action categories without training examples. While previous methods focus on inner-dataset seen/unseen splits, this paper proposes a pipeline using a large-scale training source to achieve a Universal Representation (UR) that can generalise to a more realistic Cross-Dataset UAR (CD-UAR) scenario. We first address UAR as a Generalised Multiple-Instance Learning (GMIL) problem and discover 'building-blocks' from the large-scale ActivityNet dataset using distribution kernels. Essential visual and semantic components are preserved in a shared space to achieve the UR that can efficiently generalise to new datasets. Predicted UR exemplars can be improved by a simple semantic adaptation, and then an unseen action can be directly recognised using UR during the test. Without further training, extensive experiments manifest significant improvements over the UCF101 and HMDB51 benchmarks.

1. Introduction

The field of human action recognition has advanced rapidly over the past few years. We have moved from manually designed features [41, 8] to learned convolutional neural network (CNN) features [39, 15]; from encoding appearance information to encoding motion information [37, 40, 36]; and from learning local features to learning global video features [42, 6, 19]. The performance has continued to soar higher as we incorporate more of the steps into an end-to-end learning framework [56, 55]. However, such robust and accurate action classifiers often rely on large-scale training video datasets using deep neural networks, which require large numbers of expensive annotated samples per action class. Although several large-scale video datasets have been proposed like Sports-1M [15], Activi-

tyNet [13], YouTube-8M [1] and Kinetics [16], it is practically infeasible and extremely costly to annotate action videos with the ever-growing need of new categories.

Zero-shot action recognition has recently drawn considerable attention because of its ability to recognize unseen action categories without any labelled examples. The key idea is to make a trained model that can generalise to unseen categories with a shared semantic representation. The most popular side information being used are attributes, word vectors and visual-semantic embeddings. Such zero-shot learning frameworks effectively bypass the data collection limitations of traditional supervised learning approaches, which makes them more promising paradigms for UAR.

Extensive work on zero-shot action recognition has been done in the past five years. [22, 10, 12, 26, 24] considered using attributes for classifications. These attribute-based methods are easy to understand and implement, but hard to define and scale up to a large-scale scenario. Semantic representations like word vectors [9, 3, 25] are thus preferred since only category names are required for constructing the label embeddings. There also has been much recent work on using visual-semantic embeddings extracted from pre-trained deep networks [14, 44, 28] due to their superior performance over single view word vectors or attributes.

However, whichever side information we adopt, the generalisation capability of these approaches is not promising, which is referred to as the domain shift problem. Most previous work thus still focuses on inner-dataset seen/unseen splits. This is not very practical since each new dataset or each category will require re-training. Motivated by such a fact we propose to utilise a large-scale training source to achieve a *Universal Representation* (UR) that can automatically generalise to a more realistic Cross-Dataset UAR (CD-UAR) scenario. Unseen actions from new datasets can be directly recognised via the UR without further training or fine-tuning on the target dataset.

The proposed pipeline is illustrated in Fig. 1. We first leverage the power of deep neural networks to extract vi-
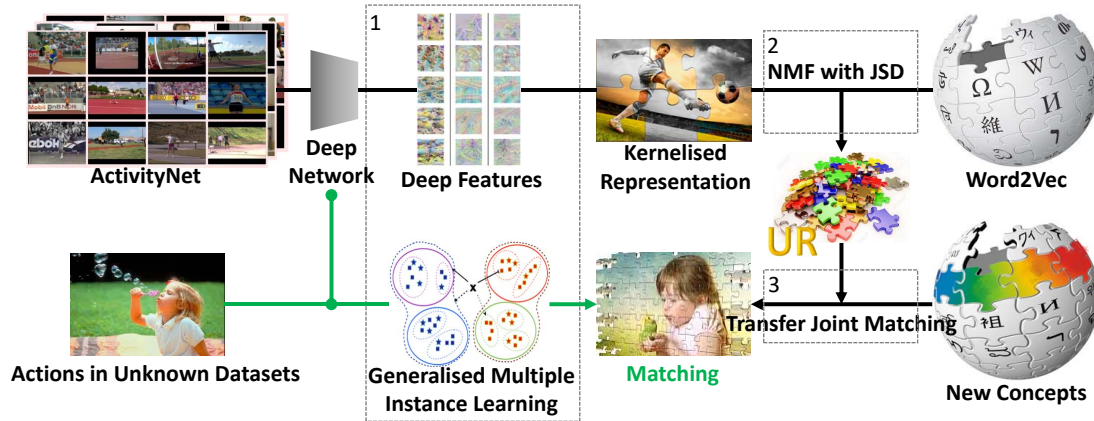

Figure 1. The proposed CD-UAR pipeline: 1) Extract deep features for each frame and summarise the video by essential components that are kernelised by GMIL; 2) Preserve shared components with the label embedding to achieve UR using NMF with JSD; 3) New concepts can be represented by UR and adjusted by domain adaptation. Test (green line): unseen actions are encoded by GMIL using the same essential components in ActivityNet to achieve a matching using UR.

sual features, which results in a Generative Multiple Instance Learning (GMIL) problem. Namely, all the visual features (instances) in a video share the label while only a small portion is determinative. Compared to conventional global summaries of visual features using Bag-of-Visual-Word or Fisher Vector encoding, GMIL aims to discover those essential "building-blocks" to represent actions in the source and target domains and suppress the ambiguous instances. We then introduce our novel Universal Representation Learning (URL) algorithm composed of Non-negative Matrix Factorisation (NMF) with a Jensen-Shannon Divergence (JSD) constraint. The non-negativity property of NMF allows us to learn a part-based representation, which serves as the key bases between the visual and semantic modalities. JSD is a symmetrised and bounded version of the Kullback-Leibler divergence, which can make balanced generalisation to new distributions of both visual and semantic features. A representation that can generalise to both visual and semantic views, and both source and target domains, is referred to as the UR. More insighs of NMF, JSD, and UR will be discussed in the experiments. Our main contributions can be summarised as follows:

- This paper extends conventional UAR tasks to more realistic CD-UAR scenarios. Unseen actions in new datasets can be directly recognised via the UR without further training or fine-tuning on the target dataset.

- We propose a CD-UAR pipeline that incorporates deep feature extraction, Generative Multiple Instance Learning, Universal Representation Learning, and semantic domain adaptation.

- Our novel URL algorithm unifies NMF with a JSD constraint. The resultant UR can substantially preserve both the shared and generative bases of visual semantic

features so as to withstand the challenging CD-UAR scenario.

- Extensive experiments manifest that the UR can effectively generalise across different datasets and outperform state-of-the-art approaches in inductive UAR scenarios using either low-level or deep features.

2. Related Work

Zero-shot human action recognition has advanced rapidly due to its importance and necessity as aforementioned. The common practice of zero-shot learning is to transfer action knowledge through a semantic embedding space, such as attributes, word vectors or visual features.

Initial work [22] has considered a set of manually defined attributes to describe the spatial-temporal evolution of the action in a video. Gan et al. [12] investigated the problem of how to accurately and robustly detect attributes from images or videos, and the learned high-quality attribute detectors are shown to generalize well across different categories. However, attribute-based methods suffer from several drawbacks: (1) Actions are complex compositions including various human motions and human-object interaction. It is extremely hard (e.g., subjective, labor-intensive, lack of domain knowledge) to determine a set of attributes for describing all actions; (2) Attribute-based approaches are not applicable for large-scale settings since they always require re-training of the model when adding new attributes; (3) Despite the fact that the attributes can be data-driven learned or semi-automatically defined [10], their semantic meanings may be unknown or inappropriate.

Hence, word vectors have been preferred for zero-shot action recognition, since only category names are required for constructing the label embeddings. [9, 45] are among the first works to adopt semantic word vector spaces as the

intermediate-level embedding for zero-shot action recognition. Following [45], Alexiou et al. [3] proposed to explore broader semantic contextual information (e.g., synonyms) in the text domain to enrich the word vector representation of action classes. However, word vectors alone are deficient for discriminating various classes because of the semantic gap between visual and textual information.

Thus, a large number of recent works [14, 21, 44, 43] exploit large object/scene recognition datasets to map object/scene scores in videos to actions. This makes sense since objects and scenes could serve as the basis to construct arbitrary action videos and the semantic representation can alleviate such visual gaps. The motivation can also be ascribed to the success of CNNs [49, 51, 48]. With the help of off-the-shelf object detectors, such methods [28] could even perform zero-shot spatio-temporal action localization.

There are also other alternatives to solve zero-shot action recognition. Gan et al. [11] leveraged the semantic inter-class relationships between the known and unknown actions followed by label transfer learning. Such similarity mapping doesn't require attributes. Qin et al. [32] formulated zero-shot learning as designing error-correcting output codes, which bypass the drawbacks of using attributes or word vectors. Due to the domain shift problem, several works have extended the methods above using either transductive learning [9, 46] or domain adaptation [17, 47].

However, all previous methods focus on inner-dataset seen/unseen splits while we extend the problem to CD-UAR. This scenario is more realistic and practical; for example, we can directly recognise unseen categories from new datasets without further training or fine-tuning. Though promising, CD-UAR is much more challenging compared to conventional UAR. We contend that when both CD and UAR are considered, the severe domain shift exceeds the generalization capability of existing approaches. Hence, we propose the URL algorithm to obtain a more robust universal representation. Our novel CD-UAR pipeline dramatically outperforms both conventional benchmarks and state-of-the-art approaches, which are in inductive UAR scenarios using low-level features and CD-UAR using deep features, respectively. One related work also applies NMF to zero-shot image classification [50]. Despite the fact that promising generalisation is reported, which supports our insights, it still focuses on inner-class splits without considering CD-UAR. Also, their sparsity constrained NMF has completely different goals to our methods with JSD.

3. Approach

In this section, we first formalise the problem and clarify each step as below. We then introduce our CD-UAR pipeline in detail, which includes Genearalised Multiple-Instance Learning, Universal Representation Learning and semantic adaptation.
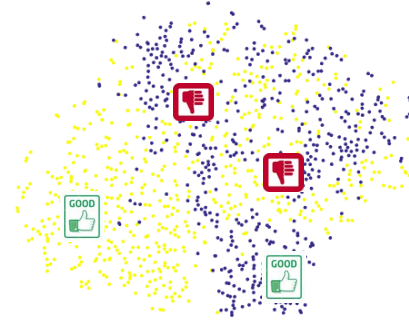

Figure 2. Visualisation of feature distributions of action 'long-jump' and 'triple-jump' in the ActivityNet dataset using tSNE.

Training Let $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_{N_s}, y_{N_s}) \subseteq \boldsymbol{X}_s \times \boldsymbol{Y}_s$ denote the training actions and their class labels in pairs in the source domain $\mathcal{D}_s$, where N_s is the training sample size; each action $\boldsymbol{x}_i$ has L_i frames in a D-dimensional visual feature space $[\boldsymbol{x}_i] = (\boldsymbol{x}_i^1, ..., \boldsymbol{x}_i^{L_i}) \in \mathbb{R}^{D \times L_i}$; $y_i \in \{1, \cdots, C\}$ consists of C discrete labels of training classes.

Inference Given a new dataset in the target domain $\mathcal{D}_t$ with C_u unseen action classes that are novel and distinct, *i.e.* $\boldsymbol{Y}_u = \{C+1, ..., C+C_u\}$ and $\boldsymbol{Y}_u \cap \boldsymbol{Y}_s = \emptyset$, the key solution to UAR needs to associate these novel concepts to $\mathcal{D}_s$ by human teaching. To avoid expensive annotations, we adopt Word2vec semantic ($\boldsymbol{S}$) label embedding $(\hat{\boldsymbol{s}}_1, \hat{y}_1), \cdots, (\hat{\boldsymbol{s}}_{C_u}, \hat{y}_{C_u}) \subseteq \boldsymbol{S}_u \times \boldsymbol{Y}_u$. Hat and subscript u denote information about unseen classes. Inference then can be achieved by learning a visual-semantic compatibility function $\min \mathcal{L}(\Phi(\boldsymbol{X}_s), \Psi(\boldsymbol{S}_s))$ that can generalise to $\boldsymbol{S}_u$.

Test Using the learned $\mathcal{L}$, an unseen action $\hat{x}$ can be recognised by $f : \Phi(\hat{\boldsymbol{x}}) \to \Psi(\boldsymbol{S}_u) \times \boldsymbol{Y}_u$.

3.1. Genearalised Multiple-Instance Learning

Conventional summary of $\boldsymbol{x}_i$ can be achieved by Bag-of-Visual-Words or Fisher Vectors [31]. In GMIL, it is assumed that instances in the same class can be drawn from different distributions. Let $P(\cdot)$ denote the space of Borel probability measures over its argument, which is known as a bag. Conventionally, it is assumed that some instances are *attractive* $P_+(\boldsymbol{x})$ while others are *repulsive* $P_-(\boldsymbol{x})$. This paper argues that many instances may exist in *neutral* bags. In Fig. 2, we show an example of visual feature distributions of 'long-jump' and 'triple-jump'. Each point denotes a frame. While most frames fall in the neutral bags (red thumb), only a few frames (green thumb) are attractive to one class and repulsive to others. The neutral bags may contain many basic action bases shared by classes or just background noise. Conventional *Maximum Mean Discrepancy* [7] may not well represent such distributions. Instead, this paper adopts the odds ratio embedding, which aims to discover the most *attractive* bases to each class c and suppress the neutral ones. This can be simply implemented by

the pooled Naive Bayes Nearest Neighbor (NBNN) kernel [33] at the 'bag-level'. We conduct k-means on each class to cluster them into H bags. The associated kernel function is:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^T \phi(\boldsymbol{x}'), \tag{1}$$

where $\Phi(\boldsymbol{X}_s) = [\phi(\boldsymbol{x}_i)] = [\phi^1(\boldsymbol{x}_i), ..., \phi^{CH}(\boldsymbol{x}_i)]^T$ is the kernelised representation with odds ratio [27] applies to each kernel embedding: $\phi^{c_i}(\boldsymbol{x}) = \sum_{l=1}^{L_i} \log \frac{p(c|\boldsymbol{x}^l)}{p(\bar{c}|\boldsymbol{x}^l)}$. Specific implementation details can be found in the supplementary material. In this way, we discover $C \times H$ bases as 'building-blocks' to represent any actions in both the source and target domains.

3.2. Universal Representation Learning

For clarity, we use $A = \Phi(\boldsymbol{X}_s)$ and $B = \boldsymbol{S}_s$ to define the visual and semantic embeddings in the source domain $\mathcal{D}_s : A \times B$. Towards universal representation, we aim to find a shared space that can: 1) well preserve the key bases between visual and semantic modalities; 2) generalise to new distributions of unseen datasets. For the former, let $A = [\mathbf{a}_1, \cdots, \mathbf{a}_{N_s}] \in \mathbb{R}_{\geq 0}^{M_1 \times N_s}$ and $B = [\mathbf{b}_1, \cdots, \mathbf{b}_{N_s}] \in \mathbb{R}_{\geq 0}^{M_2 \times N_s}$; $M_1 = C \times H$ and $M_2 = L$. NMF is employed to find two nonnegative matrices from A: $U \in \mathbb{R}_{\geq 0}^{M_1 \times D_1}$ and $V_1 \in \mathbb{R}_{\geq 0}^{D_1 \times N_s}$ and two nonnegative matrices from B: $W \in \mathbb{R}_{\geq 0}^{M_2 \times D_2}$ and $V_2 \in \mathbb{R}_{\geq 0}^{D_2 \times N_s}$ with full rank whose product can approximately represent the original matrix A and B, i.e., $A \approx UV_1$ and $B \approx WV_2$. In practice, we set $D_1 < \min(M_1, N_s)$ and $D_2 < \min(M_2, N_s)$. We constrain the shared coefficient matrix: $V_1 = V_2 = V \in \mathbb{R}_{\geq 0}^{D \times N_s}$. For the latter aim, we introduce JSD to preserve the generative components from the GMIL and use these essential 'building-blocks' to generalise to unseen datasets. Hence, the overall objective function is given as:

$$\mathcal{L} = \min_{U,W,V} \|A - UV\|_F^2 + \|B - WV\|_F^2 \\ + \eta \, \text{JSD}, s.t. \, U, W, V \geq 0, \tag{2}$$

where $\|\cdot\|_F$ is the Frobenius norm; η is a smoothness parameter; JSD is short for the following equation:

$$\text{JSD}(P_A \| P_B) = \frac{1}{2} KL(P_A \| Q) + \frac{1}{2} KL(P_B \| Q) \\ = \frac{1}{2} \sum_i \sum_j p_A^{ij} \log p_A^{ij} - p_A^{ij} \log q_{ij} \tag{3} \\ + \frac{1}{2} \sum_i \sum_j p_B^{ij} \log p_B^{ij} - p_B^{ij} \log q_{ij}$$

where P_A and P_B are probability distributions in space A and B. We aim to find the joint probability distribution Q in the shared space V that is generalised to by P_A and P_B and their shifted distributions in the target domain. Specifically,

JSD can be estimated pairwise as:

$$\begin{cases} p_A^{ij} = \dfrac{g(\mathbf{a}^i, \mathbf{a}^j)}{\sum_{k \neq l} g(\mathbf{a}^k, \mathbf{a}^l)} \\[3mm] p_B^{ij} = \dfrac{g(\mathbf{b}^i, \mathbf{b}^j)}{\sum_{k \neq l} g(\mathbf{b}^k, \mathbf{b}^l)} \\[3mm] q_{ij} = \dfrac{(1 + \|\mathbf{v}_i - \mathbf{v}_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|\mathbf{v}_k - \mathbf{v}_l\|^2)^{-1}}. \end{cases} \tag{4}$$

Without loss of generality, this paper use the cross-entropy distance to implement $g(\cdot)$.

3.2.1 Optimization

Let the Lagrangian of Eq. 2 be:

$$\mathcal{L} = \|A - UV\|^2 + \|B - WV\|^2 + \eta \, \text{JSD} \\ + tr(\Phi U^T) + tr(\Theta W^T) + tr(\Psi V^T), \tag{5}$$

where Φ, Θ and Ψ are three Lagrangian multiplier matrices. $tr(\cdot)$ denotes the trace of a matrix. For clarity, JSD in Eq. 3 is simply denoted as G. We define two auxiliary variables d_{ij} and Z as follows:

$$d_{ij} = \|\mathbf{v}_i - \mathbf{v}_j\| \text{ and } Z = \sum_{k \neq l}(1 + d_{kl}^2)^{-1}. \tag{6}$$

Note that if $\mathbf{v}_i$ changes, the only pairwise distances that change are d_{ij} and d_{ji}. Therefore, the gradient of function G with respect to $\mathbf{v}_i$ is given by:

$$\frac{\partial G}{\partial \mathbf{v}_i} = 2 \sum_{j=1}^{N} \frac{\partial G}{\partial d_{ij}}(\mathbf{v}_i - \mathbf{v}_j). \tag{7}$$

Then $\frac{\partial G}{\partial d_{ij}}$ can be calculated by JS divergence in Eq. (3):

$$\frac{\partial G}{\partial d_{ij}} = -\frac{\eta}{2} \sum_{k \neq l}(p_A^{kl} + p_B^{kl}) \left(\frac{1}{q_{kl}Z} \frac{\partial((1 + d_{kl}^2)^{-1})}{\partial d_{ij}} - \frac{1}{Z} \frac{\partial Z}{\partial d_{ij}} \right). \tag{8}$$

Since $\frac{\partial((1 + d_{kl}^2)^{-1})}{\partial d_{ij}}$ is nonzero if and only if $k = i$ and $l = j$, and $\sum_{k \neq l} p_{kl} = 1$, it can be simplified as:

$$\frac{\partial G}{\partial d_{ij}} = \eta(p_A^{ij} + p_B^{ij} - 2q_{ij})(1 + d_{ij}^2)^{-1}. \tag{9}$$

Substituting Eq. (9) into Eq. (7), we have the gradient of the JS divergence as:

$$\frac{\partial G}{\partial \mathbf{v}_i} = 2\eta \sum_{j=1}^{N}(p_A^{ij} + p_B^{ij} - 2q_{ij})(\mathbf{v}_i - \mathbf{v}_j)(1 + \|\mathbf{v}_i - \mathbf{v}_j\|^2)^{-1}. \tag{10}$$

Let the gradients of $\mathcal{L}$ be zeros to minimize O_f:

$$\frac{\partial \mathcal{L}}{\partial V} = 2(-U^T A + U^T U V - W^T B + W^T W V) + \frac{\partial G}{\partial V} + \Psi = \mathbf{0}, \quad (11)$$

$$\frac{\partial \mathcal{L}}{\partial U} = 2(-AV^T + UVV^T) + \Phi = \mathbf{0}, \quad (12)$$

$$\frac{\partial \mathcal{L}}{\partial W} = 2(-BW^T + WVV^T) + \Theta = \mathbf{0}. \quad (13)$$

In addition, we also have KKT conditions: $\Phi_{ij} U_{ij} = 0$, $\Theta_{ij} W_{ij} = 0$ and $\Psi_{ij} V_{ij} = 0$, $\forall i, j$. Then multiplying V_{ij}, U_{ij} and W_{ij} in the corresponding positions on both sides of Eqs. (11), (12) and (13) respectively, we obtain:

$$\left(2(-U^T A + U^T U V - W^T B + W^T W V) + \frac{\partial G}{\partial \mathbf{v}_i} \right)_{ij} V_{ij} = 0, \quad (14)$$

$$2(-AV^T + UVV^T)_{ij} U_{ij} = 0, \quad (15)$$

$$2(-BV^T + WVV^T)_{ij} W_{ij} = 0. \quad (16)$$

Note that

$$\left(\frac{\partial G}{\partial \mathbf{v}_j} \right)_i = \left(2\eta \sum_{k=1}^{N} \frac{(p_A^{jk} + p_B^{jk} - 2q_{jk})(\mathbf{v}_j - \mathbf{v}_k)}{1 + \|\mathbf{v}_j - \mathbf{v}_k\|^2} \right)_i$$
$$= 2\eta \sum_{k=1}^{N} \frac{(p_A^{jk} + p_B^{jk} - 2q_{jk})(V_{ij} - V_{ik})}{1 + \|\mathbf{v}_j - \mathbf{v}_k\|^2}.$$

The multiplicative update rules of the bases of both W and U for any i and j are obtained as:

$$U_{ij} \leftarrow \frac{(AV^T)_{ij}}{(UVV^T)_{ij}} U_{ij}, \quad (17)$$

$$W_{ij} \leftarrow \frac{(BV^T)_{ij}}{(WVV^T)_{ij}} W_{ij}. \quad (18)$$

The update rule of the shared space preserving the coefficient matrix V between the visual and semantic data spaces is:

$$V_{ij} \leftarrow \frac{(U^T A)_{ij} + (W^T B)_{ij} + \Upsilon}{(U^T U V)_{ij} + (W^T W V)_{ij} + \Gamma} V_{ij}, \quad (19)$$

where for simplicity, we let $\Upsilon = \eta \sum_{k=1}^{N} \frac{(p_A^{jk} + p_B^{jk})V_{ik} + 2q_{jk}V_{ij}}{1 + \|\mathbf{v}_j - \mathbf{v}_k\|^2}$, $\Gamma = \eta \sum_{k=1}^{N} \frac{(p_A^{jk} + p_B^{jk})V_{ij} + 2q_{jk}V_{ik}}{1 + \|\mathbf{v}_j - \mathbf{v}_k\|^2}$.

All the elements in U, W and V can be guaranteed to be nonnegative from the allocation. [20] proves that the objective function is monotonically non-increasing after each update of U, W or V. The proof of convergence about U, W and V is similar to that in [53, 5].

3.2.2 Orthogonal Projection

After U, W and V have converged, we need two projection matrices $\mathcal{P}_A$ and $\mathcal{P}_B$ to project A and B into V. However, since our algorithm is NMF-based, a direct projection to the shared space does not exist. Inspired by [4], we learn two rotations to protect the data originality while projecting it into the universal space, which is known as the Orthogonal Procrustes problem [35]:

$$\begin{cases} \min_{\mathcal{P}_A} \|\mathcal{P}_A A - V\|, s.t. \mathcal{P}_A^T \mathcal{P}_A = I, \\ \min_{\mathcal{P}_B} \|\mathcal{P}_B B - V\|, s.t. \mathcal{P}_B^T \mathcal{P}_B = I, \end{cases} \quad (20)$$

where I is an identity matrix. According to [52], orthogonal projection has the following advantages: 1) It can preserve the data structure; 2) It can redistribute the variance more evenly, which maximally decorrelates dimensions. The optimisation is simple. We first use the singular value decomposition (SVD) algorithm to decompose the matrix: $A^T V = Q \Sigma S^T$. Then $\mathcal{P}_A = S \Lambda Q^T$, where Λ is a connection matrix as $\Lambda = [I, \mathbf{0}] \in \mathbb{R}^{D \times M}$ and $\mathbf{0}$ indicates all zeros in the matrix. $\mathcal{P}_B$ is achieved in the same way. Given a new dataset $\mathcal{D}_t$, semantic embeddings $B_u = \boldsymbol{S}_u$ can be projected into V as class-level UR prototypes in an unseen action gallery $\hat{V}_B = \mathcal{P}_B B_u$. A test example $\hat{\mathbf{a}}$ can be simply predicted by nearest neighbour search:

$$\hat{y} = \underset{C+1 \leqslant u \leqslant C+C_u}{\arg\max} \|\mathcal{P}_A \hat{\mathbf{a}} - \hat{\boldsymbol{v}}_{B_u}\|_2^2, \quad (21)$$

where $\hat{\boldsymbol{v}}_{B_u} \in \hat{V}_B$. The overall Universal Representation Learning (URL) is summarised in Algorithm 1.

Algorithm 1 Universal Representation Learning (URL)

Require:
    Source domain $\mathcal{D}_s$: $A \in \mathbb{R}^{M_1 \times N}$ and $B \in \mathbb{R}^{M_2 \times N}$; number of bases D; hyper-parameter η;

Ensure: The basis matrices U, W, orthogonal projections $\mathcal{P}_A$ and $\mathcal{P}_B$.

1: Initialize U, W and V with uniformly distributed random values between 0 and 1.
2: **repeat**
3: Compute the basis matrices U and W and UR matrix V via Eqs. (17), (18) and (19), respectively;
4: **until** convergence
5: SVD decomposes the matrices $A^T V$ and $B^T V$ to obtain $Q_A \Sigma S_A^T$ and $Q_B \Sigma S_B^T$
6: $\mathcal{P}_A = S_A \Omega Q_A^T$; $\mathcal{P}_B = S_B \Omega Q_B^T$

3.3. Computational Complexity Analysis

The UAR test can be achieved by efficient NN search among a small number of prototypes. The training con-

sists of three parts. For NMF optimisation, each iteration takes $O(max\{M_1ND, M_2ND\})$. In comparison, the basic NMF algorithm in [20] applied to A and B separately will have complexity of $O(M_1ND)$ and $O(M_2ND)$ respectively. In other words, our algorithm is no more complex than the basic NMF. The second regression requires SVD decomposition which has complexity $O(2N^2D)$. Therefore, the total computational complexity is: $O(max\{M_1ND, M_2ND\}t + 2N^2D)$, w.r.t. the number of iterations t.

3.4. Semantic Adaptation

Since we aim to make the UR generalise to new datasets, the domain shift between $\mathcal{D}_s$ and $\mathcal{D}_u$ is unknown. For improved performance, we can use the semantic information of the target domain to approximate the shift. The key insight is to measure new unseen class labels using our discovered 'building blocks'. Because the learnt UR can reliably associate visual and semantic modalities, *i.e.* $\hat{V}_A \sim \hat{V}_B$ we could approximate the seen-unseen discrepancy $V_A \rightarrow \hat{V}_A$ by $V_A \rightarrow \hat{V}_B$.

To this end, we employ Transfer Joint Matching (TJM) [23], which achieves feature matching and instance reweighing in a unified framework. We first mix the projected semantic embeddings of unseen classes with our training samples in the UR space by $[V_A, \hat{V}_B] \in \mathbb{R}^{D \times (N_s + C_u)}$, where $V_A = \mathcal{P}_A A$. TJM can provide an adaptive matrix A and a kernel matrix K:

$$\mathcal{L}_{TJM}(V_A, \hat{V}_B) \rightarrow (A, K), \tag{22}$$

through which we can achieve the adapted unseen class prototypes $\hat{V}'_B$ in the UR space via $Z = A^T K = [V'_A, \hat{V}'_B]$.

Unseen Action Recognition Given a test action $\hat{x}$, we first convert it into a kernelised representation using the trained GMIL kernel embedding in Eq. 1: $\hat{a} = [\phi^1(\hat{x}), ..., \phi^{CH}(\hat{x})]^T$. Similar to Eq. 21, we can now make a prediction using the adapted unseen prototypes:

$$\hat{y} = \underset{C+1 \leqslant u \leqslant C+C_u}{\arg\max} \|\mathcal{P}_A \hat{a} - \hat{v}'_{B_u}\|_2^2. \tag{23}$$

4. Experiments

We perform the URL on the large-scale ActivityNet [13] dataset. Cross-dataset UAR experiments are conducted on two widely-used benchmarks, UCF101 [38] and HMDB51 [18]. UCF101 and HMDB51 contain trimmed videos while ActivityNet contains untrimmed ones. We first compare our approach to state-of-the-art methods using either low-level or deep features. To understand the contribution of each component of our method, we also provide detailed analysis of possible alternative baselines.

Method	Feature	Setting	HMDB51	UCF101
ST [45]	BoW	T	15.0±3.0	15.8±2.3
ESZSL [34]	FV	I	18.5±2.0	15.0±1.3
SJE [2]	FV	I	13.3±2.4	9.9±1.4
MTE [47]	FV	I	19.7±1.6	15.8±1.3
ZSECOC [32]	FV	I	22.6±1.2	15.1±1.7
Ours	FV	I	**24.4±1.6**	**17.5±1.6**
Ours	FV	T	28.9±1.2	20.1±1.4
Ours	GMIL-D	CD	51.8±0.7	42.5±0.9

Table 1. Comparison with state-of-the-art methods using standard low-level features. Last two sets of results are just for reference. T: transductive; I: inductive; Results are in %.

4.1. Settings

Datasets ActivityNet[1] consists of 10024 training, 4926 validation, and 5044 test videos from 200 activity classes. Each class has at least 100 videos. Since the videos are untrimmed, a large proportion of videos have a duration between 5 and 10 minutes. UCF101 is composed of realistic action videos from YouTube. It contains 13320 video clips distributed among 101 action classes. Each class has at least 100 video clips and each clip lasts an average duration of 7.2s. HMDB51 includes 6766 videos of 51 action classes extracted from a wide range of sources, such as web videos and movies. Each class has at least 101 video clips and each clip lasts an average duration of 4.3s.

Visual and Semantic Representation For all three datasets, we use a single CNN model to obtain the video features. The model is a ResNet-200 initially trained on ImageNet and fine-tuned on ActivityNet dataset. Overlapping classes between ActivityNet and UCF101 are not used during fine-tuning. We adopt the good practices from temporal segment networks (TSN) [42], which is one of the state-of-the-art action classification frameworks. We extract feature from the last average pooling layer (2048-d) as our frame-level representation. Note that we only use features extracted from a single RGB frame. We believe better performance could be achieved by considering motion information, e.g. features extracted from multiple RGB frames [39] or consecutive optical flow [37, 57, 54]. However, our primary aim is to demonstrate the ability of universal representations. Without loss of generality, we use the widely-used skip-gram neural network model [29] that is trained on Google News dataset and represent each category name by an L2-normalized 300-d word vector. For multi-word names, we use accumulated word vectors [30].

Implementation Details For GMIL, we estimate the pooled local NBNN kernel [33] using $k_{nn} = 200$ to estimate the odds-ratio in [27]. The best hyper-parameter η for URL and that in TJM are achieved through cross-validation. In order to enhance the robustness, we propose a leave-one-hop-away cross validation. Specifically, the training set of

---

[1] We use the latest release 1.3 of ActivityNet for our experiments

Method	Train	Test	Splits	Accuracy (%)
Jain *et al.* [14]	-	101	3	30.3
Mettes and Snoek [28]	-	101	3	32.8
Ours	-	101	3	**34.2**
Kodirov *et al.* [17]	51	50	10	14.0
Liu *et al.* [22]	51	50	5	14.9
Xu *et al.* [47]	51	50	50	22.9
Li *et al.* [21]	51	50	30	26.8
Mettes and Snoek [28]	-	50	10	40.4
Ours	-	50	10	**42.5**
Kodirov *et al.* [17]	81	20	10	22.5
Gan *et al.* [12]	81	20	10	31.1
Mettes and Snoek [28]	-	20	10	51.2
Ours	-	20	10	**53.8**

Table 2. Comparison with state-of-art methods on different splits using deep features.

ActivityNet is evenly divided into 5 hops according to the ontological structure. In each iteration, we use 1 hop for validation while the other furthest 3 hops are used for training. Except for feature extraction, the whole experiment is conducted on a PC with an Intel quad-core 3.4GHz CPU and 32GB memory.

4.2. Comparison with State-of-the-art Methods

Comparison Using Low-level Features Since most existing methods are based on low-level features, we observe a significant performance gap. For fair comparison, we first follow [32] and conduct experiments in a conventional *inductive* scenario. The seen/unseen splits for HMDB51 and UCF101 are 27/26 and 51/50, respectively. Visual features are 50688-d Fisher Vectors of improved dense trajectory [41], which are provided by [47]. Semantic features use the same Word2vec model. Without local features for each frame, our training starts from the URL. Note some methods [45] are also based on a *transductive* assumption. Our method can simply address such a scenario by incorporating $\hat{V}_A$ into the TJM domain adaptation. We report our results in Table 1. The accuracy is averaged over 10 random splits.

Our method outperforms all of the compared state-of-the-art methods in the same inductive scenario. Although the transductive setting to some extent violates the 'unseen' action recognition constraint, the TJM domain adaptation method shows significant improvements. However, none of the compared methods are competitive to the proposed pipeline even though it is completely inductive plus cross-dataset challenge.

Comparison Using Deep Features In Table 2, we follow recent work [28] which provides the most comparisons to related zero-shot approaches. Due to many different data splits and evaluation metrics, the comparison is divided into the three most common settings, *i.e.* using the standard supervised test splits; using 50 randomly selected actions for testing; and using 20 actions randomly for testing.

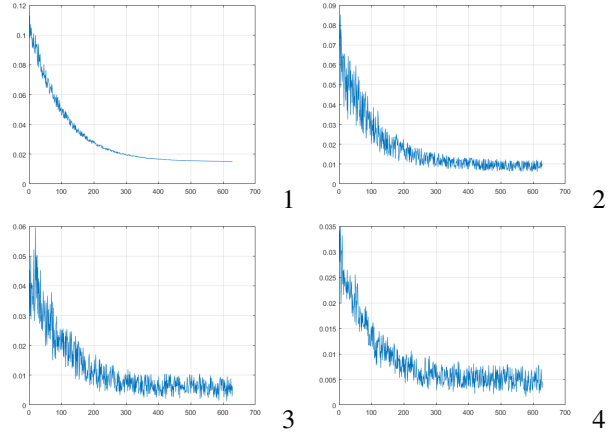The highlights of the comparison are summarised as fol-

Figure 3. Convergence analysis w.r.t. # iterations. (1) is the overall loss in Eq. 2. (2) is the JSD loss. (3) and (4) show decomposition losses of A and B, respectively.

lows. First, [28] is also a deep-feature based approach, which employs a GoogLeNet network, pre-trained on a 12,988-category shuffle of ImageNet. In addition, it adopts the Faster R-CNN pre-trained on the MS-COCO dataset. Secondly, it also does not need training or fine-tuning on the test datasets. In other words, [28] shares the same spirit to our cross-dataset scenario, but from an object detection perspective. By contrast, our CD-UAR is achieved by pure representation learning. Overall, this is a fair comparison and worthy of a thorough discussion.

Our method consistently outperforms all of the compared approaches, with minimum margins of 1.4%, 2.1%, and 2.6% over [28], respectively. Note that, other than [14] which is also deep-model-based, there are no other competitive results. Such a finding suggests future UAR research should focus on deep features instead. Besides visual features, we use the similar skip-gram model of Word2vec for label embeddings.Therefore, the credit of performance improvements should be given to the method itself.

4.3. In-depth Analysis

Since our method outperforms all of the compared benchmarks, to further understand the success of the method, we conduct 5 baselines as alternatives to our main approach. The results are summarised in Table 3.

Convergence Analysis Before analysing baselines, we first show examples of convergence curves in Fig. 3 during our URL optimisation. It can be seen the overall loss reliably converges after approximately 400 iterations. The JSD constraint in (2) gradually resolves while the decomposition losses (3) and (4) tend to be competing to each other. This can be ascribed to the difference of ranks between A and B. While A is instance-level kernelised features, B is class-level Word2vec that has much lower rank than that of A. The alternation in each iteration reweighs A and B once in turn, despite the overall converged loss.

Dataset	HMDB51				UCF101			
Setting	Cross-Dataset		Transductive		Cross-Dataset		Transductive	
GMIL+ESZSL[34]	25.7		30.2		19.8		24.9	
UR Dimensionality	Low	High	Low	High	Low	High	Low	High
Fisher Vector	47.7	48.6	53.9	54.6	35.8	39.7	42.2	43.0
NMF (no JSD)	17.2	18.0	19.2	20.4	15.5	17.4	18.2	19.8
CCA	13.8	12.2	18.2	17.1	8.2	9.6	12.9	13.6
No TJM	48.9	50.5	51.8	53.9	32.5	36.6	38.1	38.6
Ours	49.6	51.8	57.8	58.2	36.1	42.5	47.4	49.9

Table 3. In-depth analysis with baseline approaches. 'Ours' refers to the complete pipeline with deep features, GMIL kernel embedding, URL with NMF and JSD, and TJM. (Results are in %).

Pipeline Validation Due to the power of deep features demonstrated by the above comparison, an intuitive assumption is that the CD-UAR can be easily resolved by deep features. We thus use the same GMIL features followed by a state-of-the-art ESZSL [34] using RBF kernels. The performance in Table 1 (15.0%) is improved to (19.8%), which is marginal to our surprise. Such a results shows the difficulty of CD-UAR while confirms the contribution of the proposed pipeline.

GMIL vs FV As we stated earlier, the frame-based action features can be viewed as the GMIL problem. Therefore, we change the encoding to conventional FV and keep the rest of the pipeline. It can be seen that the average performance drop is 2% with as high as 6.9% in transductive scenario on UCF101.

Separated Contribution Our URL algorithm is arguably the main contribution in this paper. To see our progress over conventional NMF, we set $\eta = 0$ to remove the JSD constraint. As shown in Table 3, the performance is severely degraded. This is because NMF can only find the shared bases regardless of the data structural change. GNMF [5] may not address this problem as well (not proved) because we need to preserve the distributions of those generative bases rather than data structures. While generative bases are 'building blocks' for new actions, the data structure may completely change in new datasets. However, NMF is better at preserving bases than canonical correlation analysis (CCA) which is purely based on mutual-information maximisation. Therefore, a significant performance gap can be observed between the results of CCA and NMF.

Without Domain Adaptation In our pipeline, TJM is used to adjust the inferred unseen prototypes from Word2vec. The key insight is to align the inferred bases to that of GMIL in the source domain that is also used to represent unseen actions. In this way, visual and semantic UR is connected by $\hat{V}_B \sim V_A \sim \hat{V}_A$. Without such a scheme, however, we observe marginal performance degradation in the CD-UAR scenario (roughly 3%). This is probably because ActivityNet is rich and the concepts of HMDB51 and UCF101 are not very distinctive. We further investigate the CD transductive scenario, which assumes $\hat{V}_A$ can be observed for

TJM. As a result, the benefit from domain adaptation is large (roughly 5% on HMDB51 and 1% on UCF101 between 'Ours' and 'No TJM').

Basis Space Size We propose two sets of size according to the original sizes of A and B (recall section 3.2), namely the high one $D_{high} = \frac{1}{2}(M_1 + M_2)$ and the low one $D_{low} = \frac{1}{4}(M_1 + M_2)$. As shown in Table 3, the higher dimension gives better results in most cases. Note that the performance difference is not significant. We can thus conclude that our method is not sensitive to the basis space size.

5. Conclusion

This paper studied a challenging Cross-Dataset Unseen Action Recognition problem. We proposed a pipeline consisting of deep feature extraction, Generative Multiple-Instance Learning, Universal Representation Learning, and Domain Adaptation. A novel URL algorithm was proposed to incorporate Non-negative Matrix Factorisation with a Jensen-Shannon Divergence constraint. NMF was shown to be advantageous for finding shared bases between visual and semantic spaces, while the remarkable improvement of JSD was empirically demonstrated in distributive basis preserving for unseen dataset generalisation. The resulting Universal Representation effectively generalises to unseen actions without further training or fine-tuning on the new dataset. Our experimental results exceeded that of state-of-the-art methods using both conventional and deep features. Detailed evaluation manifests that most of contribution should be credited to the URL approach.

We leave several interesting open questions. For methodology, we have not examined other variations of NMF or divergences. The GMIL problem is proposed without in-depth discussion, although a simple trial using pooled local-NBNN kernel showed promising progress. In addition, the improvement of TJM was not significant in inductive CD-UAR. A unified framework for GMIL, URL and domain adaptation could be a better solution in the future.

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. YouTube-8M: A Large-Scale Video Classification Benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 1

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of Output Embeddings for Fine-Grained Image Classification. In *CVPR*, 2015. 6

[3] I. Alexiou, T. Xiang, and S. Gong. Exploring Synonyms as Context in Zero-Shot Action Recognition. In *ICIP*, 2016. 1, 3

[4] D. Cai, X. He, and J. Han. Spectral Regression for Efficient Regularized Subspace Learning. In *ICCV*, 2007. 5

[5] D. Cai, X. He, J. Han, and T. S. Huang. Graph Regularized Nonnegative Matrix Factorization for Data Representation. *PAMI*, 2011. 5, 8

[6] A. Diba, V. Sharma, and L. V. Gool. Deep Temporal Linear Encoding Networks. In *CVPR*, 2017. 1

[7] G. Doran, A. Latham, and S. Ray. A Unifying Framework for Learning Bag Labels from Generalized Multiple-Instance Data. In *IJCAI*, 2016. 3

[8] B. Fernando, E. Gavves, J. O. M., A. Ghodrati, and T. Tuytelaars. Modeling Video Evolution for Action Recognition. In *CVPR*, 2015. 1

[9] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong. Transductive Multi-view Embedding for Zero-Shot Recognition and Annotation. In *ECCV*, 2014. 1, 2, 3

[10] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Learning Multi-modal Latent Attributes. *PAMI*, 2013. 1, 2

[11] C. Gan, M. Lin, Y. Yang, Y. Zhuang, and A. G.Hauptmann. Exploring Semantic Inter-Class Relationships (SIR) for Zero-Shot Action Recognition. In *AAAI*, 2015. 3

[12] C. Gan, T. Yang, and B. Gong. Learning Attributes Equals Multi-Source Domain Generalization. In *CVPR*, 2016. 1, 2, 7

[13] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *CVPR*, 2015. 1, 6

[14] M. Jain, J. C. van Gemert, T. Mensink, and C. G. Snoek. Objects2action: Classifying and Localizing Actions without Any Video Example. In *ICCV*, 2015. 1, 3, 7

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale Video Classification with Convolutional Neural Networks. In *CVPR*, 2014. 1

[16] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[17] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised Domain Adaptation for Zero-Shot Learning. In *ICCV*, 2015. 3, 7

[18] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A Large Video Database for Human Motion Recognition. In *ICCV*, 2011. 6

[19] Z. Lan, Y. Zhu, A. G. Hauptmann, and S. Newsam. Deep Local Video Feature for Action Recognition. In *CVPR Workshops*, 2017. 1

[20] D. D. Lee and H. S. Seung. Algorithms for Non-Negative Matrix Factorization. In *NIPS*, 2001. 5, 6

[21] Y. Li, S. H. Hu, and B. Li. Recognizing Unseen Actions in a Domain-Adapted Embedding Space. In *ICIP*, 2016. 3, 7

[22] J. Liu, B. Kuipers, and S. Savarese. Recognizing Human Actions by Attributes. In *CVPR*, 2011. 1, 2, 7

[23] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer Joint Matching for Unsupervised Domain Adaptation. In *CVPR*, 2014. 6

[24] Y. Long, L. Liu, F. Shen, L. Shao, and X. Li. Zero-shot learning using synthesised unseen visual data with diffusion regularisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1

[25] Y. Long, L. Liu, Y. Shen, and L. Shao. Towards affordable semantic searching: Zero-shot retrieval via dominant attributes. In *AAAI*, 2018. 1

[26] Y. Long and L. Shao. Learning to recognise unseen classes by a few similes. In *ACMMM*, 2017. 1

[27] S. McCann and D. G. Lowe. Local Naive Bayes Nearest Neighbor for Image Classification. In *CVPR*, 2012. 4, 6

[28] P. Mettes and C. G. Snoek. Spatial-Aware Object Embeddings for Zero-Shot Localization and Classification of Actions. *ICCV*, 2017. 1, 3, 7

[29] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and Their Compositionality. In *NIPS*, 2013. 6

[30] D. Milajevs, D. Kartsaklis, M. Sadrzadeh, and M. Purver. Evaluating Neural Word Representations in Tensor-Based Compositional Settings. *EMNLP*, 2014. 6

[31] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher Kernel for Large-Scale Image Classification. *ECCV*, 2010. 3

[32] J. Qin, L. Liu, L. Shao, F. Shen, B. Ni, J. Chen, and Y. Wang. Zero-Shot Action Recognition with Error-Correcting Output Codes. In *CVPR*, 2017. 3, 6, 7

[33] K. Rematas, M. Fritz, and T. Tuytelaars. The Pooled NBNN Kernel: Beyond Image-to-Class and Image-to-Image. In *ACCV*, 2012. 4, 6

[34] B. Romera-Paredes and P. Torr. An Embarrassingly Simple Approach to Zero-Shot Learning. In *ICML*, 2015. 6, 8

[35] P. H. Schönemann. A Generalized Solution of the Orthogonal Procrustes Problem. *Psychometrika*, 1966. 5

[36] G. A. Sigurdsson, S. Divvala, A. Farhadi, and A. Gupta. Asynchronous Temporal Fields for Action Recognition. In *CVPR*, 2017. 1

[37] K. Simonyan and A. Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In *NIPS*, 2014. 1, 6

[38] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild. In *CRCV-TR-12-01*, 2012. 6

[39] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *ICCV*, 2015. 1, 6

[40] G. Varol, I. Laptev, and C. Schmid. Long-term Temporal Convolutions for Action Recognition. 2017. 1

[41] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In *ICCV*, 2013. 1, 7

[42] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Val Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *ECCV*, 2016. 1, 6

[43] Q. Wang and K. Chen. Alternative Semantic Representations for Zero-Shot Human Action Recognition. In *ECML-PKDD*, 2017. 3

[44] Z. Wu, Y. Fu, Y.-G. Jiang, and L. Sigal. Harnessing Object and Scene Semantics for Large-Scale Video Understanding. In *CVPR*, 2016. 1, 3

[45] X. Xu, T. Hospedales, and S. Gong. Semantic Embedding Space for Zero-Shot Action Recognition. In *ICIP*, 2015. 2, 3, 6, 7

[46] X. Xu, T. Hospedales, and S. Gong. Transductive Zero-Shot Action Recognition by Word-Vector Embedding. *IJCV*, 2017. 3

[47] X. Xu, T. M. Hospedales, and S. Gong. Multi-Task Zero-Shot Action Recognition with Prioritised Data Augmentation. In *ECCV*, 2016. 3, 6, 7

[48] J. Xue, H. Zhang, and K. Dana. Deep Texture Manifold for Ground Terrain Recognition. In *CVPR*, 2018. 3

[49] J. Xue, H. Zhang, K. Dana, and K. Nishino. Differential Angular Imaging for Material Recognition. In *CVPR*, 2017. 3

[50] M. Ye and Y. Guo. Zero-shot classification with discriminative semantic representation learning. In *CVPR*, 2017. 3

[51] H. Zhang, V. Sindagi, and V. M. Patel. Image De-raining Using a Conditional Generative Adversarial Network. *arXiv preprint arXiv:1701.05957*, 2017. 3

[52] X. Zhang, F. X. Yu, R. Guo, S. Kumar, S. Wang, and S.-F. Chang. Fast Orthogonal Projection based on Kronecker Product. In *ICCV*, 2015. 5

[53] W. Zheng, Y. Qian, and H. Tang. Dimensionality Reduction with Category Information Fusion and Non-Negative Matrix Factorization for Text Categorization. In *AICI*, 2011. 5

[54] Y. Zhu, Z. Lan, S. Newsam, and A. G. Hauptmann. Guided Optical Flow Learning. *arXiv preprint arXiv:1702.02295*, 2017. 6

[55] Y. Zhu, Z. Lan, S. Newsam, and A. G. Hauptmann. Hidden Two-Stream Convolutional Networks for Action Recognition. *arXiv preprint arXiv:1704.00389*, 2017. 1

[56] Y. Zhu and S. Newsam. Depth2Action: Exploring Embedded Depth for Large-Scale Action Recognition. In *ECCV Workshops*, 2016. 1

[57] Y. Zhu and S. Newsam. DenseNet for Dense Flow. In *ICIP*, 2017. 6