Pattern Recognition xxx (xxxx) 108445

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com

Journal Cover Image

Hierarchical domain adaptation with local feature patterns

Jun Wen [a, b, d, 1], Junsong Yuan [e], Qian Zheng [f], Risheng Liu [g, h], Zhefeng Gong [c], Nenggan Zheng [a, b, *]

[a] Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou, Zhejiang, China
[b] College of Computer Science and Techology, Zhejiang University, Hangzhou, Zhejiang, China
[c] Department of Neurobiology, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China
[d] Department of Biomedical Informatics, Harvard Medical School, USA
[e] Department of Computer Science and Engineering, State University of New York at Buffalo, USA
[f] Nanyang Technological University, Singapore
[g] DUT-RU International School of Information Science and Engineering, Dalian University of Technology, Liaoning, China
[h] Pazhou Lab, Guangzhou, China

ARTICLE INFO

Article history:
Received 26 October 2020
Received in revised form 13 July 2021
Accepted 18 November 2021

Keywords:
Domain adaptation
Local feature patterns
Adversarial learning
Hierarchical alignment

ABSTRACT

Domain adaptation is proposed to generalize learning machines and address performance degradation of models that are trained from one specific source domain but applied to novel target domains. Existing domain adaptation methods focus on transferring holistic features whose discriminability is generally tailored to be source-specific and inferiorly generic to be transferable. As a result, standard domain adaptation on holistic features usually damages feature structures, especially local feature statistics, and deteriorates the learned discriminability. To alleviate this issue, we propose to transfer primitive local feature patterns, whose discriminability are shown to be inherently more sharable, and perform hierarchical feature adaptation. Concretely, we first learn a cluster of domain-shared local feature patterns and partition the feature space into cells. Local features are adaptively aggregated inside each cell to obtain cell features, which are further integrated into holistic features. To achieve fine-grained adaptations, we simultaneously perform alignment on local features, cell features and holistic features, within which process the local and cell features are aligned independently inside each cell to maintain the learned local structures and prevent *negative transfer*. Experimenting on typical *one-to-one* unsupervised domain adaptation for both image classification and action recognition tasks, partial domain adaptation, and domain-agnostic adaptation, we show that the proposed method achieves more reliable feature transfer by consistently outperforming state-of-the-art models and the learned domain-invariant features generalize well to novel domains.

1. Introduction

Domain shift is commonly encountered by machine learning practitioners that models are trained in one particular source distribution but applied to different but related target distributions. Generally *domain shift* causes performance degradations, e.g., an object recognition model trained using images of daylight can hardly generalize to infrared images. Though state-of-the-art deep representations demonstrate certain level of invariance to low-level variations, they are still susceptible to *domain shift* [1], as we cannot manually label sufficient training data that cover diverse application domains. To address such performance drops, one typical solution is to further finetune the learned models on the current task. However, it may be prohibitively difficult and expensive to obtain sufficient labeled data to properly finetune the large-scale parameters employed by deep networks. Instead of recollecting labeled data and retraining learned models for every possible new scenarios, unsupervised domain adaptation (UDA) is proposed to improve model generalization ability by transferring informative knowledge learned from one labeled domain, denoted as source domain, to unlabeled novel domains, denoted as target domains [1,2].

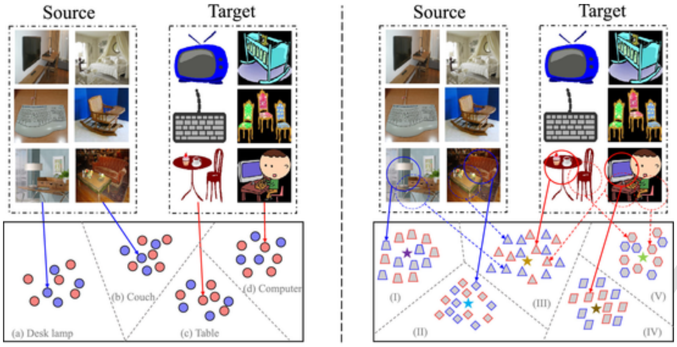
Among the UDA approaches to bridging different domains and alleviating *domain shift*, an important strategy is to learn domain-invariant representations. For example, plenty of traditional methods with shallow features have been proposed to minimize domain discrepancy in the shared subspace [3,4]. Recently, deep neural networks have been exploited to map both domains into a domain-invariant feature space and learn more transferable representations [2,5]. This is generally achieved by optimizing the learned representations to minimize some

[*] Corresponding author.
E-mail address: zng@cs.zju.edu.cn (N. Zheng).
[1] This work is done when Jun Wen was at Zhejinag University.

https://doi.org/10.1016/j.patcog.2021.108445
0031-3203/© 2021

measures of domain discrepancy, e.g., maximum mean discrepancy (MMD) [6,7], reconstruction loss [8], correlation [9,10], moment [11], or adversarial loss [1,2]. Among them, the adversarial learning based domain adaptation methods have become increasingly prevalent because of their excellent performance.

To achieve a successful domain adaptation, two key properties of the learned representations should be satisfied: discriminability and transferability. The feature discriminability is generally induced by a supervised training objective using available source labeled data while the transferability is typically strengthened via domain-invariant feature learning. However, these two properties could be incompatible in that standard feature transferability enhancement tends to be detrimental to the learned feature discriminability as shown in [13,14]. We postulate an important reason for that is that existing domain adaptation methods generally adapt holistic features which are inherently of inferior transferability. The holistic features, which are defined to capture global semantics of samples, e.g., deep features from the final fully-connected layers of deep networks, are usually tailored to capture task-specific semantics and biased toward the source, thus hardly domain-sharable or transferable to novel domains as shown in [15]. Standard domain-invariant feature learning on biased holistic features is risky and generally deteriorates the learned discriminative structures with arbitrary transformations between the source and target, especially when their supports are disjoint [16]. In challenging domain adaptation tasks, e.g., partial domain adaptation [17] or domain agnostic adaptation [18], *negative transfer* is shown to occur frequently. Instead of focusing on such holistic features, we observe that primitive local feature patterns, which are defined to capture local statistics, *e.g.*, regional "objects" of images or "sub-actions" of action sequences, are inherently more domain-sharable. We illustrate their comparisons in the Fig. 1. As shown in Fig. 1 (a), samples of different labels from the source domain "Real-world" and target domain "Clipart" [12] could share none of the holistic features. However, as we can see from Fig. 1 (b), the local feature pattern "Table" is sharable not only across domains but also among the four different categories, namely "Desk lamp", "Couch", "Table" and "Computer". But to be noted, the semantics of local feature patterns are generic and do not necessarily correspond to specific category labels. Upon the sharable local patterns, we further build holistic features for final prediction, which are tailored to be more task-specific and thus discriminative. These generic local feature patterns inherently bridge source and target domains with joint source-target supports and could make more reliable feature transfer.



(a) Holistic feature transfer

(b) Local feature transfer

Fig. 1. Comparisons of (a) existing holistic feature adaptation and (b) the proposed adaptation approach based on local feature patterns. The circles denote source or target data samples and their holistic features (source: blue; target: red). The trapezoids, diamonds, triangles, parallelograms and hexagons denote different local feature patterns, upon which discriminative holistic features are built. In (b), the local semantic "Table", represented by triangles, is shared not only between the "Real-world" and "Clipart" domains [12] but also among the four different categories, namely "Desk lamp", "Couch", "Table" and "Computer" (the solid lines indicate the most important local patterns for classifying the samples). In contrast, none of holistic features are sharable among different categories as shown in (a) (best viewed in color).. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Motivated by the above observations, we propose a novel domain adaptation approach that bridges source and target domains by learning transferable local feature patterns and with hierarchical feature alignment, as shown in Fig. 2. Concretely, we first partition feature space into cells by learning several typical local feature patterns using samples from both domains, which are shared both across domains and among different categories. Regarding to the residuals and similarity to the typical local feature patterns, the local features are then adaptively aggregated within each cell to obtain cell features, which are concatenated and normalized to build the final sample-level holistic feature. To prevent the damage to the learned feature structures or discriminability, hierarchical feature alignment is performed for fine-grained feature adaptation. First, we adapt local features inside each local feature pattern cell independently, which maintains the diversified local statistics of each cell and reduces the influences of irrelevant features. We further perform adaptation of cell features inside each cell, which can be regarded as re-weighted composites of local features with weights determined by the similarity of each local feature to the local patterns. Since the above feature adaptation is performed over the domain-shared local feature patterns with joint source-target support, the feature alignment is easier in optimization and more reliable. Finally, we perform standard adaptation of holistic features which are aggregated from the local and cell features. Built on the primitive local feature patterns, we ex-

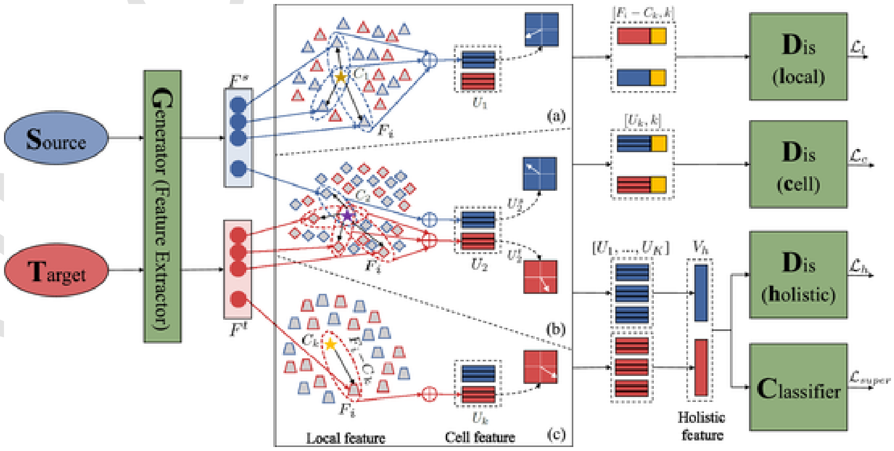


Fig. 2. Architecture of the proposed method. K typical local feature patterns are learned using data from both the source and target domains. Multi-level feature alignments are performed to achieve fine-grained feature adaptation while maintaining learned discriminative local statistics (best viewed in color).

pect such hierarchical feature adaptation to achieve more reliable feature transfer in varied domain adaptation tasks.

We summarize the main contributions as follows:

- We propose to learn primitive local feature patterns for unsupervised domain adaptation, whose discriminability is inherently more transferable than the typically-adopted holistic features.
- We propose a hierarchical feature adaptation strategy to achieve fine-grained feature alignment, through which discriminative local feature structures are maintained and *negative transfer* of irrelevant features is attenuated.
- Experimenting on the typical *one-to-one* domain adaptation for image classification and action recognition tasks, challenging partial domain adaptation [17] and domain-agnostic learning [18], the consistently distinct improvements demonstrate the superiority of the proposed method over state-of-the-art approaches. Further, our learned domain-invariant features are shown to generalize well to novel domains.

This paper makes a substantial extension to its conference version [19] in all the three aspects of contributions above. 1) We show that besides the local and holistic alignments, additional cell feature alignment is necessary, and thus propose a novel hierarchical feature adaptation strategy to achieve fine-grained feature alignment while maintaining the learned discriminative structures in Section 3.2. 2) We further theoretically justify the proposed method in Section 3.4, and experimentally show that it is able to reduce the upper bound of the expected target error in Section 4.4.2. 3) We show that the local feature patterns are not limited to be spatial in images, but can also apply to temporal or spatiotemporal patterns in action recognition tasks in Section 4.1.2 and Section 4.1.3, respectively. 4) Beside the typical unsupervised domain adaptation, we demonstrate that the learned local feature patterns also enjoy more superior transferability thanholistic features on other domain adaptation tasks, i.*e.*, partial domain adaptation, domain-agnostic learning, and domain generalization in Section 4.2, Section 4.3 and Section 4.4.1, respectively.

2. Related works

In practical applications, machine learning models usually have to work on data of varied distributions, namely from different domains. When the *domain shift* is large enough, a model trained on one domain typically performs poorly on another. Supervised fine-tuning on novel domains could be infeasible when labeled data in novel domains are prohibitively difficult or expensive to obtain, and meanwhile minimal amount of novel data usually causes model over-fitting. Unsupervised domain adaptation is proposed to adapt source-learned model and generalize it to target domains only with additional unlabeled target data.

Domain shift is prevalently reduced by domain adaptation methods through domain-invariant feature learning [20]. Previous methods usually seek to align source and target feature through subspace learning [3,4,21]. Recently deep domain adaptation approaches become prevalent as deep networks can learn more transferable representations [2,6,22]. Different measures of domain discrepancy have been minimized to bring close source and target domains. Maximum Mean Discrepancy (MMD) over (multi-layer) deep features between source and target is previously used to reduce domain distribution divergency [6,7,22]. Sun et al. propose to align the second-order statistics of deep features across domains [9]. Moments matching is used to transfer features learned from multiple source domains to one target domain [11]. The most prevalent approach is based on adversarial learning by training an additional discriminator network to distinguish source features from the target ones and reducing domain shift by encouraging feature extractors to produce features that are able to confuse the discriminator [1,2,5]. There are also methods that combine different measures, *e.g.*, Rahman et al. propose to incorporate correlation alignment along with adversarial training [10].

Besides feature transferability, increasing attentions are attached to feature discriminability in order to enhance the two simultaneously. Shu et al. show that feature distribution matching is a weak constraint for successful domain adaptation and propose to enforce a cluster assumption in feature space to prevent the classifier's decision boundary from crossing high-density feature regions [16]. Saito et al. propose to pay attention to classifier decision boundary and maximize the classifiers' prediction discrepancy to align source and target features [23]. To prevent the influence of domain-invariant feature learning on feature discriminability, a two-stream architecture is proposed with each stream operating in one domain [13]. Chen et al. [14] show that adversarial feature adaptation could damage the original feature structures and deteriorate feature discriminability, and propose to maintain the source-learned feature discriminability during domain-invariant feature learning.

All these deep domain adaptation methods are built on holistic features which are inherently of inferior transferability than local feature patterns, which is the motivation of our work. In this paper, we propose to build domain adaptation on local feature patterns, which are more primitive and sharable. A closely related work is [24], which adapts image patch features with MMD measuring domain divergency. Another close work is [25], which proposes to focus on the most transferable local regions via a attention module. The main difference between [25] and our work is that we adaptively perform adaptation over all local regions hierarchically, instead of just the regions selected by the attention part. By paying attention to local or low-level features, DCAN [26] attempts to learn domain-specific discriminative features by exciting different feature channels also via an attention-based strategy. We perform local and cell feature adaptation locally within each separated cell for fine-grained feature alignment, which also shares the same spirit of [5] in the sense of local statistic adaptation. However, it aligns the holistic features locally within each category to respect the statistics of each category. For the hierarchical feature adaptation strategy, DAN [6], RTN [22], and [27] propose to simultaneously adapt deep holistic features of multiple layers. However, they focus on matching the overall distributions of holistic features in each deep layer, without considering their diversified local statistics. This is the first paper to investigate fine-grained adaptation of hierarchical features over primitive local feature patterns.

Local Feature Aggregation There are various methods to aggregate local features into holistic ones, such as vectors of locally aggregated descriptors (VLAD) [28], bag of visual words (BoW) [29], etc. Previously, these methods have usually been applied to aggregate hand-crafted keypoint descriptors, such as SIFT, as a post-processing step. Only recently have them been extended to encode deep convolutional features in a end-to-end training manner [30,31]. In this paper, we employ the end-to-end trainable NetVLAD [30] for local feature aggregation, over which hierarchical feature adaptation is performed. To be noted, the focus of this paper is not on the local feature aggregation strategy but to explore approach to achieving reliable feature transfer over the local features.

3. The proposed method

In this section, we present the proposed domain adaptation method. Given labeled source domain data $D_s = \left(X_s, Y_s\right)$ and unlabeled target domain data $D_t = (X_t)$, the goal of unsupervised domain adaptation is to learn adapted models that minimize the expected target prediction errors. The source and target domain are sampled from joint distribution $P_s\left(X_s, Y_s\right)$ and $P_t\left(X_t, Y_t\right)$, respectively, with $P_s \neq P_t$. Existing domain adaptation methods typically bridge source and target domains by learning domain-invariant holistic feature V_h, namely achieving

$P_s(V_h) = P_t(V_h)$. However, holistic discriminability is usually tailored to be biased towards the source domain and thus with inferior transferability, especially when the labeled source data are limited. Adapting biased holistic feature tends to mixup up local feature patterns arbitrarily and damages feature structures with deteriorated feature discriminability.

To alleviate this issue, we build domain adaptation on more sharable local feature patterns with hierarchical alignment for reliable transfer. Concretely, we decompose holistic features with K more primitive and transferable local feature patterns $C_1, C_2, \ldots, C_K$, which partition feature space into K cells. By adaptively aggregating local features, we obtain cell features, which are further integrated into holistic features. To achieve fine-grained feature alignment, we simultaneously adapt local features, cell features and holistic features whilewith the local and cell feature alignments performed inside each cell to maintain the learned feature structures. Architecture of the proposed method is illustrated in Fig. 2. In the following, we first describe the learning of local feature patterns, and then present the proposed hierarchical feature adaptation strategy. The main notations used in this paper are summarized in Table 1.

3.1. Learning local feature patterns

In this section, we learn transferable local feature patterns for reliable domain adaptation. We assume there are several typical and primitive local feature patterns that are sharable both across domains and among different categories, e.g., the common “objects” in different images or “sub-actions” in action sequences. We employ the end-to-end trainable NetVLAD [30] to learn discriminative local feature patterns and aggregate local features to obtain holistic features. We first learn an initial cluster of typical local feature patterns and then adapt them for cross-domain feature transfer. Given a collection of local features, we perform *k-means* clustering to obtain K clustering centers to initialize the K typical local feature patterns, $C_1, C_2, \ldots, C_K$. Each local feature F_i is then assigned a k-dimensional similarity vector S_i according to its distances to the K local feature patterns, defined as:

$$S_i[k] = \frac{e^{-\alpha\|F_i - C_k\|^2}}{\sum_{k'=1}^{K} e^{-\alpha\|F_i - C_{k'}\|^2}}, \tag{1}$$

which soft-assigns each local feature F_i to the local pattern C_k with the similarity being proportional to its distances to the C_k in the feature space. $S_i[k]$ ranges between 0 and 1, with the highest *similarity* value assigned to the closest local feature pattern. α is a tunable hyper-parameter (positive constant) and controls the decay of the similarity responses to the magnitude of the distances. Note that for $\alpha \to +\infty$, F_i is hard-assigned to the nearest local feature pattern.

The NetVLAD encoding converts the multiple local features into a single $d \times K$ dimensional vector V_h, describing the distribution of local features regarding to the K typical local feature patterns, which is formulated as:

$$V_h = [U_1, U_2, \ldots, U_K], \tag{2}$$

where U_k is d-dimensional cell feature which is aggregated from local features and defined as:

$$U_k = \sum_{i=1}^{N} S_i[k]\left(F_i[d] - C_k[d]\right), \tag{3}$$

where $F_i[d]$ and $C_k[d]$ are the d-th dimension of feature F_i and local feature pattern C_k, respectively. $F_i[d] - C_k[d]$ is the residual of feature F_i to local feature pattern C_k. N denotes the number of input local features. The intuition is that residuals record the differences between local features and the typical local feature patterns. The residuals are aggregated inside each of the local feature pattern cell, and the similarity vector defined above determines the contribution of the residual of each feature to the total cell residuals. The representation of each cell is stacked and normalized into a $d \times K$ dimensional holistic descriptor V_h [32]. As we can see, the aggregation of holistic feature V_h is guided by the similarity vector and thus independent of, either temporally or spatially, the positional variations of the local features.

To encourage the local feature patterns to be more discriminative, we enforce a clustering assumption over them with a sparse loss $\mathscr{L}_s$ which is defined as:

$$\mathscr{L}_s = \frac{1}{N}\sum_{i=1}^{N} max\left(-\sum_{k=1}^{K} S_i[k]\log S_i[k], m\right), \tag{4}$$

where m is the information entropy threshold. S is the similarity vector described in Eq. 1, but here we use a much smaller decay weight α_s. By minimizing the sparse loss, we expect sparse soft-assignments of local features to the learned local feature patterns and thus less boundary local features lying between different local feature pattern cells.

Table 1
Summary of used notations.

Notation	Description	Notation	Description
X	Input feature	Y	Input label
Subscript S	Denote source	Subscript T	Denote Target
F	Local feature	N	Number of local features
V_h	Holistic feature	U	Cell feature
C	Local feature pattern	S	Similarity vector
$\mathscr{L}$	Training loss	α	Similarity decay weight
G	Feature extractor	**D**	Feature discriminator
β	Weight of holistic adaptation	γ	Weight of cell adaptation
λ	Weight of local adaptation	η	Weight of sparse loss

3.2. Feature adaptation

In this section, we perform hierarchical feature adaptation to achieve fine-grained feature alignment and reliable domain adaptation. We perform feature adaptation based on adversarial training. But to be noted, our method could also apply to various domain discrepancy measures [6,9,11]. We first describe the typical adversarial domain adaptation and then present the proposed local feature adaptation, cell feature adaptation and holistic feature adaptation.

3.2.1. Adversarial domain adapation

We employ adversarial training to match the source and target feature distributions and learn domain-invariant features [2,33]. The adversarial domain adaptation procedure can be regarded as a two-player game, where the first player is the domain discriminator $\mathbf{D}$ that is trained to distinguish the source features from the target features, while the second player, the feature extractor $\mathbf{G}$, is trained to confuse the domain discriminator. By optimizing the discriminator to best discriminate target from source features, the feature extractor is guided to learn features that are domain-invariant. Formally, the $\mathbf{D}$ and $\mathbf{G}$ are trained with the following *minmax* procedure:

$$\min_{\mathbf{G}}\max_{\mathbf{D}} \mathbb{E}_{x_i\sim X_s} \log\left(\mathbf{D}\left(\mathbf{G}\left(x_i\right)\right)\right) + \mathbb{E}_{x_j\sim X_t} log\left(1 - \mathbf{D}\left(\mathbf{G}\left(x_j\right)\right)\right). \tag{5}$$

3.2.2. Local feature adaptation

Existing adversarial domain adaptation methods focus on aligning holistic features, without considering the statistics of the more transferable local feature patterns. As a result, local features tend to be poorly mixed up by the typical brute-force holistic feature alignment. We propose to maintain the multi-mode local statistics of local features learned from the source by matching their distributions locally and independently within each separated cells, which enables fine-grained local feature alignments and reduces the influences of irrelevant local features to combat *negative transfer*. Besides, since the local features are soft-assigned to the K cells according to the semantic similarity, defined in Eq. 1, the local feature alignment is invariant to the positional variation of the local features.

We achieve local feature adaptation with a conditional adversarial training [34]. We assign each local feature F_i to its nearest local feature pattern $a_i = \arg\max_k (S_i[k])$, where $S_i[\cdot]$ is a similarity vector, defined in Eq. 1. Instead of focusing on the original local features, we propose to match the distribution of their residuals to the assigned local feature patterns, which could enable easier optimization with improved feature alignments for the following two reasons. Firstly, both the holistic and cell features are built on the residuals of the local features, thus alignments of the residuals directly promote the alignments of holistic and cell features. Further, we observe that though the local features vary greatly in value during training, their residuals to the cell centers, which are also trained end-to-end, are relatively stable, namely with more stable statistic gap across domains, which enable easier domain-invariance learning.

In addition, we progressively align the local features within each cell by re-weighting the adversarial adaptation loss, according to their similarity to the typical local feature patterns. This is because there could be many boundary local features lying between the cell patterns, and strong adaptation loss would mislead the local features to be initially assigned to the incorrect cells, thus producing false feature alignment. The training objective is formulated as:

$$\begin{aligned}\min_{\mathbf{G}}\max_{\mathbf{D}_l}\mathcal{L}_l = {} & \mathbb{E}_{x_i\sim X_s} S_i\left[a_i\right] * log\left(\mathbf{D}_l\left(F_i^s - C_{a_i}, a_i\right)\right) \\ & + \mathbb{E}_{x_j\sim X_t} S_j\left[a_j\right] * log\left(1 - \mathbf{D}_l\left(F_j^t - C_{a_j}, a_j\right)\right),\end{aligned} \tag{6}$$

where F_i^s and F_j^t denote the local features of source sample x_i and target x_j, respectively, and C_{a_i} and C_{a_j} are the assigned local feature patterns, with index of a_i and a_j, respectively. $\mathbf{D}_l$ is the discriminator trained to distinguish source and target local features.

3.2.3. Cell adaptation

Well-aligned local feature cannot necessarily guarantee alignment of the cell features, which are re-weighted composite of local features. As shown in Fig. 2(b), there could be multiple local features of a sample assigned to one cell. Though the local features are well aligned within the C_2 across domains as a whole, the cell feature distribution U_2^s and U_2^t could still be distinct. To achieve fine-grained feature alignment, we further perform cell feature adaptations. In the same spirit to respect the local statistics of each cell, we employ the conditional adversarial training to adapt the cell features, which is formulated as:

$$\begin{aligned}\min_{\mathbf{G}}\max_{\mathbf{D}_c}\mathcal{L}_c = {} & \mathbb{E}_{x_i\sim X_s} \log\left(\mathbf{D}_c\left(U_k^s, k\right)\right) \\ & + \mathbb{E}_{x_j\sim X_t} \log\left(1 - \mathbf{D}_c\left(U_k^t, k\right)\right),\end{aligned} \tag{7}$$

where U_k^s and U_k^t denote the aggregated cell feature from cell k of source sample x_i and target sample x_j, respectively. $\mathbf{D}_c$ is the discriminator trained to distinguish the source from the target cell features.

3.2.4. Holistic adaptation

While the above domain-invariance learning over local and cell features brings closer the source and target, domain shift could still linger in the holistic features. In this section, we perform adaptation on holistic features. We share the source-learned classifier and learn domain-invariant holistic feature V_h in a typical adversarial training procedure:

$$\begin{aligned}\min_{\mathbf{G}}\max_{\mathbf{D}_h}\mathcal{L}_h = {} & \mathbb{E}_{x_i\sim X_s} \log\left(\mathbf{D}_h\left(V_h^s\right)\right) \\ & + \mathbb{E}_{x_j\sim X_t} log\left(1 - \mathbf{D}_h\left(V_h^t\right)\right),\end{aligned} \tag{8}$$

where V_h^s and V_h^t denote the holistic feature of source sample x_i and target sample x_j, respectively. $\mathbf{D}_h$ is the discriminator trained to distinguish the source from the target holistic features.

To be noted, the main goal to this paper is to explore domain adaptation over local feature patterns. Therefore, we do not additionally consider conditional shift in the classifier, as in [16,35], or perform structural regularizations over the holistic feature statistics, *e.g.*, clustering assumptions [16,36]. Our method is reasonably expected to make a good complement to these methods in practice for stronger models by performing the proposed local feature structure preserved domain adaptation.

3.3. Network training

To learn discriminative features that are transferable across domains, we train the network with the following objective:

$$\min_{\mathbf{G}}\max_{\mathbf{D}_l,\mathbf{D}_c,\mathbf{D}_h} \mathcal{L}_{super} + \beta * \mathcal{L}_l + \gamma * \mathcal{L}_c + \lambda * \mathcal{L}_h + \eta * \mathcal{L}_s, \tag{9}$$

where $\mathcal{L}_{super} = J\left(\mathbf{G}\left(X_s\right), Y_s\right)$, a typical cross-entropy loss for the classification task, is the source-domain supervised training objective for feature discriminability and optimized using the available source labeled data. $\mathcal{L}_l$, $\mathcal{L}_c$ and $\mathcal{L}_h$ are the domain-invariant feature learning objective to reduce domain shift. $\mathcal{L}_s$ is the sparse loss to encourage local feature patterns to be more discriminative. β, γ, λ and η are hyper-parameters that trade-off the objectives in the unified optimization problem.

We optimize the network in three steps. Firstly, we perform standard adversarial training to adapt the holistic features as done in DANN [1]. By firstly bringing closer the source and target in holistic feature space, the domain shift of local features can also be alleviated, as shown in Fig. 5(d), based on which we expect to learn more domain-sharable typical local feature patterns. Then, we perform *K-means* clustering over the local features that are from both the source and target domains, initialize the typical local feature patterns with the K clustering centers, and train the classifier to minimize source supervised loss while keeping the layers of the backbone before the local feature aggregation layer frozen. Finally, we jointly train the classifier, feature extractor, and the typical local feature patterns with the final objective in Eq. 9.

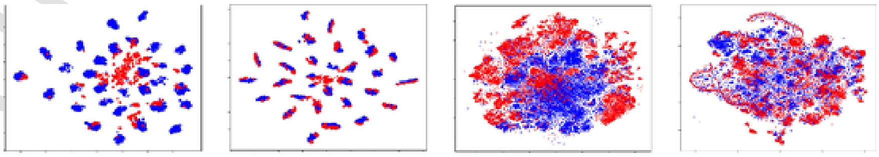

(a) AlexNet (holistic) (b) DANN (holistic) (c) AlexNet (local) (d) DANN (local)

Fig. 5. The t-SNE visualizations of holistic and local features learned by the fine-tuned AlexNet in (a) and (c) and by DANN in (b) and (d), respectively, on the $A \rightarrow W$ task (blue: **A**, red: **W**, best viewed in color). In (a) and (b), the category labels are indicated using different markers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.4. Theoretical justification

In this section, we justify that our method reduces the upper bound of the expected error on the target samples based on the theoretical analysis of domain adaptation [20]. Given the source domain S and the target domain T, let V be the domain-invariant feature and $\mathscr{H}$ the hypothesis class, the probabilistic error of hypothesis $\mathscr{H}$ on the target domain is upper-bounded by the following three terms:

$$\forall h \in \mathscr{H}, \epsilon_T\left(h\left(V_T\right)\right) \leq \epsilon_S\left(h\left(V_S\right)\right) + \frac{1}{2} d_{\mathscr{H}\triangle\mathscr{H}}\left(V_S, V_T\right) + R, \tag{10}$$

where 1) $\epsilon_S(h)$ denotes the expected source-domain error which can be easily minimized using available source labels; 2) $d_{\mathscr{H}\triangle\mathscr{H}}\left(V_S, V_T\right)$ measures domain discrepancy between the two domains w.r.t. a hypothesis set $\mathscr{H}$; 3) R, the error of the ideal hypothesis on both domains. The R is defined as $R = \min_{h\in\mathscr{H}} \epsilon_S\left(h\left(V_S\right), f_S\left(V_S\right)\right) + \epsilon_T\left(h\left(V_T\right), f_T\left(V_T\right)\right)$, where f_S and f_T are labeling functions for the source and target domains, respectively. Most existing domain adaption methods treat the third term to be negligibly small. However, as shown in [14] and Fig. 4, while standard domain adaptation methods are able to reduce the second term (domain discrepancy), the third term, especially the target-domain error, tends to be unboundedly enlarged. This is because the learned feature discriminability would inevitably be damaged during the adaptation.

We postulate that an important reason of why standard feature transferability enhancement is generally harmful to the learned discriminability is that the holistic feature discriminability is tailored by the source labeled data to be biased, thus with inferior transferability. Standard domain adaptation of biased holistic features tends to mix up local feature structures arbitrarily as shown in Fig. 5(d), which weakens feature discriminability, especially when the source-target supports are disjoint [16].

To maintain the source-learned feature discriminability, we perform hierarchical feature adaptations upon local feature patterns. Each holistic feature is decomposed with K local feature patterns $C_1, C_2, \dots, C_K$, which are more primitive and transferable. We preserve the learned local statistics by additionally performing local adaptations independently inside each cell C_k, which also promotes the transfer of relevant local feature patterns and prevents the *negative transfer* of irrelevant features. As shown in Fig. 6 and Fig. 4, our method achieves fine-grained feature alignment and significantly reduces the second term in Eq. 10 (domain discrepancy), respectively, while keeping the third term (feature discriminability) negligibly deteriorated, thus reducing the upper bound of the target error.

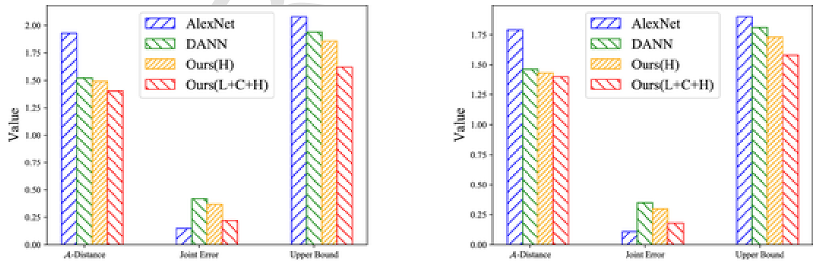
4. Experiments

To evaluate the reliability of the proposed domain adaptation method, we experiment with three settings: 1) typical *one-to-one* unsupervised domain adaptation where domain adaptation is performed from one source domain to one target domain, 2) partial domain adaptation where the target label space is a subset of the source domain, and 3) domain-agnostic adaptation which is a more difficult but practical problem of learning from one labeled source domain and adapting to severalunlabeled target domains.

To investigate the effectiveness of the proposed hierarchical feature adaptation, we evaluate our method with different variants, they are, 1) *Ours(H)*, which only aligns holistic features, 2) *Ours(L+H)*, which aligns both local and holistic features, 3) *Ours(C+H)*, which aligns both cell and holistic features, 4) *Ours(L+C+H)*, which simultaneously aligns local, cell and holistic features as proposed. To show the importance of the re-weighting of local adaptation loss, we also evaluate *Ours (L+H, w/o rw)*, which performs adaptation without re-weighting the local adversarial loss. Besides, we evaluate *Ours(L+H, ori)*, which aligns the original local features, to verify that alignment of local residuals promotes stable training and alignments of cell and holistic features.

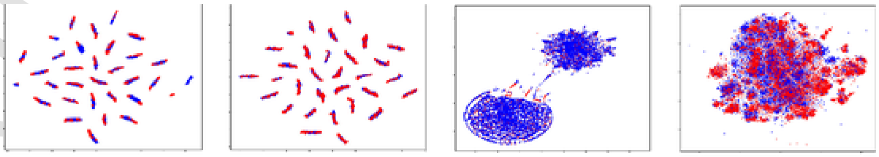
4.1. One-to-one unsupervised domain adaptation

In this section, we evaluate the proposed method on the typical *one-to-one* unsupervised domain adaptation . We experiment on both image classification and action recognition tasks. The local feature pattern learning and local feature aggregation are performed on features with



(a) A→W (b) W→ADC

Fig. 4. Analysis of the upper bound of the target error. Local and cell feature adaptation enable *Ours(L+C+H)* to reduce domain discrepancy $\mathscr{A}$-distance significantly while keeping the feature discriminability less attenuated, thus reducing the upper bound of target error.



(a) Ours (H,holistic) (b) Ours (full, holistic) (c) Ours (full, cell) (d) Ours (full, local)

Fig. 6. The t-SNE visualizations of (a) adapted holistic features of Ours (H), (b) adapted holistic features of the proposed full model Ours (L+C+H), (c) adapted cell features of the full model Ours (L+C+H), (d) adapted local features of the full model Ours (L+C+H) on the $A \to W$ task. Source **A**: blue, and target **W**: red, best viewed in color. In (a) and (b), the category labels are indicated using different markers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

varied semantics, namely on “spatially” local 2D convolutional features of the image recognition task, on “spatiotemporally” local 3D convolutional features of the action recognition task with RGB videos, and on “temporally” local sequential features of the skeleton-based action recognition.

4.1.1. Image classification

Setup We compare to varied domain adaptation methods to comprehensively verify the effectiveness of the proposed methods, including both traditional domain adaptation methods using subspace alignment [3,4] or component analysis [21], and end-to-end deep domain adaptation methods. For deep domain adaptation, we report methods with a variety of domain discrepancy measures, including DAN with maximum mean discrepancies (MMD) on multiple feature layers [6], D-CORAL with second-order statistics [9], and RTN with MMD on hierarchically fused features [22]), and the adversarial learning based DANN [1], ADDA [2], JAN [7], CDAN-M [34], GCAN [37], DCAN [26], CAADA [10]. For deep domain adaptation, both one-layer (D-CORAL, DANN, ADDA, and CDAN-M) and multi-layer (DAN, RTN, DCAN) holistic feature adaptation methods are compared. To further study the transferability of holistic features, we also report the performance of the typical adversarial domain adaptation method DANN with the fully-connected layers randomly initialized, denoted as DANN(random). We experiment on the most popular *Office-31* dataset [38] and the challenging *Office-home* dataset [12].

Office-31 [38]. This dataset is the most popular for visual domain adaptation. It consists of 4652 images from 31 categories with three different domains: Amazon (A), with 2817 images from amazon.com, Webcam (W) and DSLR (D), with 795 images and 498 images taken by web camera and digital SLR camera in different environmental settings, respectively.

Office-home [12]. This is a challenging domain adaptation dataset, which comprises 15,588 images with 65 categories of objects in office and home settings. There are 4 significantly different domains: Art (Ar) with 2427 painting, sketches or artistic depiction images, Clipart (Cl) with 4365 images, Product (Pr) containing 4439 images and Real-world (Rw) with 4357 regularly captured images. Some example images are shown in Fig. 3.

Office-Caltech10 [3]. This dataset contains the 10 common categories shared by *Office-31* and *Caltech-256* datasets, as shown in Fig. 3. Besides the *Amazon, DSLR,* and *Webcam,* it includes a novel domain: Caltech (C), which are sampled from the *Caltech-256* dataset.

Implementation Details

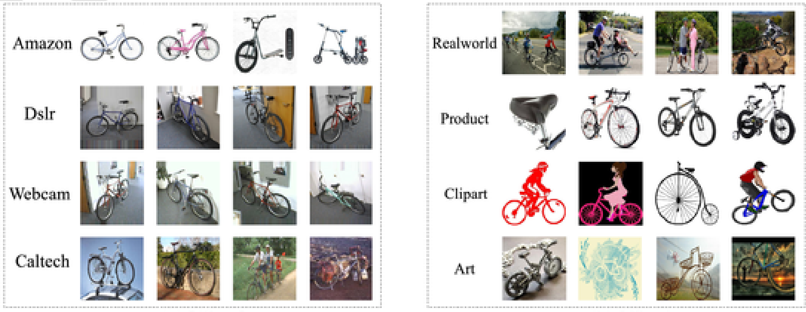
We use the AlexNet network, pre-trained from the ImageNet, as the feature extractor. The local feature pattern learning and local feature aggregation are performed on the last convolutional layer where spatially local semantics are learned. It is because this feature layer achieves the best performance, which is consistent with the conference version [19]. We share the parameters of the source and target feature extractors. We keep the number of local feature patterns fixed, that is $K = 32$. Model performance is non-sensitive to the cluster number K, as shown in the conference version [19]. But larger K calls for stronger local/cell feature discriminators accordingly to distinguish the diversified local statistics. For local feature aggregation, we use a large $\alpha = 5000.0$ to encourage independent residual accumulation within each local feature pattern cell. We use a small similarity decay $\alpha_s = 0.01$ and a small sparsity threshold $m = 0.02$. For adversarial feature adaptation, all feature discriminators $\mathbf{D}_l$, $\mathbf{D}_c$, and $\mathbf{D}_h$ consist of 3 fully connected layers: two hidden layers with 2048 and 2048 units, respectively, followed by a final discriminator output layer. The model is implemented in Tensorflow framework and optimized using Adam with initial learning rate of 0.001 for the classifier layers and 0.0001 for the pre-trained layers. If not explicitly pointed out, the hyper-parameters settings also apply to the following parts. The reported results are averaged from 5 random runs.

Experimental Results We report the results on the *Office-31* and *Office-home* datasets in Table 2 and Table 3, respectively. The proposed model *Ours(L + C + H)* achieves consistent improvements over the compared methods. End-to-end deep domain adaptation methods learn more transferable features and show significant advantages over the traditional approaches, *i.e.*, TCA, SA and GFK. Adversarial domain adaptation methods outperform D-CORAL, DAN, RTN, which are with traditional domain discrepancy measures. Comparing to the adaptation of multi-layer holistic deep features in DAN, RTN and JAN, the proposed *Ours(L + C + H)* shows significant advantages, which verify the transferability of local feature patterns.

When the holistic fully-connected feature layers are randomly initialized, DANN suffers averaged performance drops of 5.7% on the *Office-31* and 4.3% on the *Office-home*. The performance drops are more distinct when plain or less diversified domains, e.g., *DSLR* or *Webcam* in the *Office-31*, act as the source. The results show that when the holistic fully-connected layers are tailored to be source-specific or biased by the limited source labeled data, *e.g.*, no available large-scale pre-training data, it is generally difficult for the learned holistic discriminability to be reliably adapted. The distinct performance gaps between DANN (random) between Ours (H), in which there are hidden no fully-connected layers, also show the superiority and wide applicability of employing local feature patterns for domain adaptation.

The introduced local and cell feature adaptation together bring *Ours (L + C + H)* averaged performance improvements of 3.1% on the *Office-31* and 3.8% on the *Office-home* over the holistic-only *Ours(H)*. The advantages of the proposed method are more obvious when more diversified domains are included in the adaptation, e.g., the *Amazon* in the *Office-31* and *Realworld* in the *Office-home*, in which case more diversified local feature patterns can be learned and transferred.

Adaptive re-weighting of local adaptation loss according to the semantic similarity brings *Ours(L + H)* performance improvements over *Ours(L + H, w/o rw)* of 1.1% on the *Office-31* and 1.3% on the *Office-home*. Further, the average variance of *Ours(L + H, w/o rw)* on the *Office-31* is 0.43, which is much larger than the 0.23 of *Ours(L + H)* on the



(a) Office-Caltech (b) Office-home

Fig. 3. Example images (“bikes”) of Office-Caltech [3] and Office-home [12] datasets.

Table 2
Accuracy (%) on the *Office31* dataset for the *one-to-one* unsupervised domain adaptation task.

Method	A→W	W→D	D→W	A→D	W→A	D→A	Avg
TCA [21]	45.4 ± 0.0	92.2 ± 0.0	81.1 ± 0.0	46.8 ± 0.0	39.5 ± 0.0	36.4 ± 0.0	56.9
SA [4]	47.4 ± 0.0	93.8 ± 0.0	89.1 ± 0.0	50.6 ± 0.0	37.6 ± 0.0	39.5 ± 0.0	59.7
GFK [3]	54.7 ± 0.0	96.2 ± 0.0	92.1 ± 0.0	52.4 ± 0.0	41.8 ± 0.0	43.2 ± 0.0	63.4
AlexNet	61.6 ± 0.5	99.0 ± 0.2	95.4 ± 0.3	63.8 ± 0.5	49.8 ± 0.4	51.1 ± 0.6	70.1
D-CORAL [9]	66.8 ± 0.6	99.2 ± 0.1	95.7 ± 0.3	66.4 ± 0.4	51.5 ± 0.3	52.8 ± 0.2	72.1
DAN [6]	68.5 ± 0.5	99.0 ± 0.3	96.0 ± 0.3	67.0 ± 0.4	53.1 ± 0.5	54.8 ± 0.5	72.9
RTN [22]	73.3 ± 0.3	99.6 ± 0.1	96.8 ± 0.2	72.3 ± 0.3	51.0 ± 0.1	50.5 ± 0.3	73.7
DANN [1]	73.0 ± 0.5	99.2 ± 0.3	96.4 ± 0.3	72.3 ± 0.3	51.2 ± 0.5	53.4 ± 0.4	74.3
ADDA [2]	73.5 ± 0.6	98.8 ± 0.4	96.2 ± 0.4	71.6 ± 0.4	53.5 ± 0.6	54.6 ± 0.5	74.7
JAN [7]	74.9 ± 0.3	99.5 ± 0.2	96.6 ± 0.2	71.8 ± 0.2	55.0 ± 0.4	**58.3** ± 0.3	76.0
CDAN-M [34]	78.3 ± 0.2	**100.0** ± 0.0	97.2 ± 0.1	76.3 ± 0.1	57.3 ± 0.3	57.3 ± 0.2	77.7
DCAN [26]	79.3 ± 0.4	99.1 ± 0.2	**97.4** ± 0.3	77.8 ± 0.4	58.1 ± 0.4	57.5 ± 0.4	78.2
CAADA [10]	**80.2**	99.2	97.1	77.7	57.4	58.1	78.3
DANN(random) [1]	65.7 ± 0.7	98.3 ± 0.8	95.6 ± 0.7	64.1 ± 0.6	43.2 ± 0.8	44.7 ± 0.7	68.6
Ours(H)	74.2 ± 0.2	99.6 ± 0.3	96.6 ± 0.2	73.3 ± 0.4	54.5 ± 0.2	54.1 ± 0.4	75.4
Ours(L + H, w/o rw)	75.3 ± 0.4	99.5 ± 0.3	96.4 ± 0.4	75.0 ± 0.4	56.1 ± 0.5	55.8 ± 0.6	76.4
Ours(L + H, ori)	76.3 ± 0.3	99.6 ± 0.3	96.8 ± 0.2	76.0 ± 0.3	56.9 ± 0.4	57.0 ± 0.4	77.1
Ours(L + H)	76.8 ± 0.2	99.7 ± 0.2	96.7 ± 0.3	76.5 ± 0.1	57.8 ± 0.3	57.4 ± 0.3	77.5
Ours(C + H)	77.2 ± 0.2	99.6 ± 0.4	96.9 ± 0.2	77.6 ± 0.1	58.7 ± 0.3	57.7 ± 0.2	78.0
Ours(L + C + H)	78.8 ± 0.4	99.8 ± 0.2	97.0 ± 0.2	**78.2** ± 0.3	**59.1** ± 0.5	**58.3** ± 0.4	**78.5**

Table 3
Accuracy (%) on the *Office-home* dataset for the *one-to-one* unsupervised domain adaptation task.

Method	Ar:Cl	Ar:Pr	Ar:Rw	Cl:Ar	Cl:Pr	Cl:Rw	Pr:Ar	Pr:Cl	Pr:Rw	Rw:Ar	Rw:Cl	Rw:Pr	Avg
AlexNet	26.4	32.6	41.3	22.1	41.7	42.1	20.5	20.3	51.1	31.0	27.9	54.9	34.3
DAN [6]	31.7	43.2	55.1	33.8	48.6	50.8	30.1	35.1	57.7	44.6	39.3	63.7	44.5
DANN [1]	36.4	45.2	54.7	35.2	51.8	55.1	31.6	39.7	59.3	45.7	46.4	65.9	47.3
JAN [7]	35.5	46.1	57.7	36.4	53.4	54.5	33.4	40.3	60.1	45.9	47.4	67.9	48.2
CAADA [10]	35.3	46.2	56.6	34.9	51.8	**60.0**	34.9	40.0	60.2	47.8	44.5	67.9	48.2
MSTN [36]	34.9	46.2	56.8	36.6	55.0	55.4	33.3	41.7	60.7	47.0	45.9	68.3	48.5
CDAN-M [34]	38.1	50.3	60.3	39.7	56.4	57.8	35.5	43.1	63.2	48.4	48.5	71.1	51.0
GCAN [37]	36.4	47.3	61.1	37.9	**58.3**	57.0	35.8	42.7	64.5	**50.1**	49.1	72.5	51.1
DCAN [26]	39.1	51.6	62.2	**42.0**	57.6	58.4	35.7	43.2	63.7	49.3	48.8	**72.7**	52.0
DANN (random) [1]	32.2	38.7	50.5	31.4	46.7	50.3	27.3	35.8	54.8	42.6	43.2	62.8	43.0
Ours(H)	37.5	45.7	56.1	36.6	52.7	55.5	32.8	39.8	60.9	47.4	46.9	68.8	48.4
Ours(L + H, w/o rw)	37.9	46.5	57.4	37.3	53.0	55.2	33.2	39.2	60.3	48.0	47.2	69.4	48.7
Ours(L + H, ori)	38.0	47.1	57.8	38.1	53.3	55.4	35.2	40.1	62.5	49.2	48.1	70.2	49.6
Ours(L + H)	38.6	47.9	58.5	38.6	53.7	55.9	35.3	40.2	63.1	49.3	48.2	70.7	50.0
Ours(C + H)	39.1	51.7	62.6	39.5	55.4	56.7	35.5	41.8	63.9	49.8	49.5	71.8	51.4
Ours(L + C + H)	**39.5**	**52.1**	**63.0**	40.1	56.9	57.9	**36.0**	**43.5**	**64.6**	**50.1**	**49.9**	72.3	**52.2**

Office-31. Adaptive re-weighting enables progressive adaptation, which promotes stable optimization and prevents false feature alignment.

Instead of aligning local features in the original feature space, as done in *Ours(L + H, ori), Ours(L + H)* achieves improvements over it of 0.4% on both the *Office-31*and *Office-home*. The improvements of *Ours (L + H)* over *Ours(L + H, ori)* are more distinct on tasks where cell alignment plays an important role, *e.g.*, 0.8% on the Ar→Pr task of *Office-home*. This is because residual alignment is able to promote cell and holistic feature alignments.

4.1.2. *Action recognition on RGB videos*

Setup We evaluate the proposed method on action recognition task using RGB video and learn transferable spatiotemporal local feature patterns for domain adaptation. We experiment on the UCF-Olympic and UCF-HMDB [39] datasets.

UCF-Olympic [39]. This dataset consists of the shared 6 categories of Olympic dataset and UCF50 dataset, namely “Basketball”, “Pole vault”, “Tennis serve”, “Diving”, “Clean jerk”, and “Throw discus”. The videos in Olympic mainly come from the Olympic Games with the actions performed by professional athletes under specific platforms. However, the UCF50 are performed in unconstrained scenarios, by unprofessional individuals.

UCF-HMDB [39]. This dataset comprises the 12 categories of the overlapping categories between the UCF101 and HMDB51 dataset. The actions of HMDB are performed in more diversified settings and with more varied camera angles than the UCF101, thus more challenging. We follow the official protocol to split the training and validation sets to evaluate the domain adaptation performance.

Implementation Details An inflated 3D (I3D) ConvNet pre-trained from the Kinetics dataset [40] is used to extract spatiotemporal features from RGB videos. We perform local feature patterns learning and local feature aggregation on the last convolutional layer, where spatiotemporally local semantics are learned. For the I3D net, we follow the same implementations as in [40]. The same feature discriminators and training hyper-parameters as the above Image Classification task are used.

Experimental Results The results of action recognition on RGB videos are reported in Table 4. We compare to the typical unsupervised domain adaptation methods DANN [1], JAN [7], and MCD [23], all of which adapt the holistic features before the classifier. The TA3N [39] performs multi-level feature adaptations to reduce domain shift in frame-level features, temporal dynamics, and video-level holistic features. As shown in Table 4, transferring from UCF to Olympic (or HMDB to UCF) is easier than transferring from Olympic to UCF(or UCF to HMDB) because UCF(or HMDB) is more diversified, and the learned features are less biased and thus more transferable. Local and cell feature adaptations with local spatiotemporal feature patterns enable *Ours (L + C + H)* to achieve an improvement of 2.8% over the holistic only

Table 4
Results of *one-to-one* unsupervised domain adaptation on action recognition with RGB videos.

Method	UCF→ Olympic	Olympic→ UCF	UCF→ HMDB	HMDB→ UCF	Avg
DANN [1]	92.5 ± 0.4	88.5 ± 0.5	73.3 ± 0.4	81.9 ± 0.5	84.1
JAN [7]	94.3 ± 0.3	87.9 ± 0.4	73.1 ± 0.3	82.2 ± 0.3	84.4
MCD [23]	96.5 ± 0.3	90.9 ± 0.3	75.3 ± 0.4	83.3 ± 0.4	86.5
TA3N [39]	98.2	93.0	78.3	81.8	87.8
Ours(H)	97.9 ± 0.5	90.1 ± 0.5	75.5 ± 0.4	81.9 ± 0.3	86.1
Ours(L + H)	98.3 ± 0.3	93.5 ± 0.3	76.8 ± 0.3	82.8 ± 0.4	87.9
Ours(C + H)	98.5 ± 0.3	93.9 ± 0.3	77.0 ± 0.2	83.2 ± 0.3	88.2
Ours (L + C + H)	**98.7** ± 0.2	**94.5** ± 0.2	**78.7** ± 0.3	**83.7** ± 0.2	**88.9**

model *Ours(H)* and outperform DANN, JAN, and MCD. Comparing to the multi-level feature adaptation of TA3N that adapts spatial (image) features and temporal dynamic features separately, our method directly adapts the local spatiotemporal patterns ("sub-actions") and achieves an improvement of 1.1%.

4.1.3. Skeleton-based action recognition

Setup We experiment on the CMU-HDM05 dataset for the skeleton-based action recognition task. We select the shared 9 action categories of CMU dataset and HDM05 [42], namely "cartwheelg, "grabg, "jumpg, "pick", "punch", "run", "sit", "throw" and "walk" ("walk back"). 3D coordinates of 31 joints are recorded in both the CMU and HDM05 while with some differences in the joint position between them. The sequences of HDM05 are performed by 5 actors while the CMU by 144 non-professional actors with larger intra-class diversities and viewpoint variations. We follow the standard split of training and test sets for HDM05 and CMU dataset as in [42].

Implementation Details A two-layer bidirectional GRUs network with each layer of 800 hidden unites pre-trained from the CMU dataset is used to model the temporal dependency of the skeleton sequences. We perform local feature patterns learning and aggregation on the outputs of each time step of the last GRUs layer, where temporally local semantics are learned. We follow the data pre-processing strategy and network implementations as in [42]. The same feature discriminators and training hyper-parameters as the above Image Classification task are used.

Experimental Results The results of skeleton-based action recognition are reported in Table 5. The proposed *Ours(L + C + H)* consistently outperforms the compared state-of-the-art domain adaptation methods DANN [1], DSN [41], ADDA [2] and MCD [23]. The same as the action recognition on RGB videos above, transferring from challenging source domain (CMU dataset) to plain domain (HMDB05) is more reliable because the learned features would be less biased and more transferable. Local and cell feature adaptation promote positive transfer of discriminative local spatiotemporal patterns across domains, bringing *Ours (L + C + H)* an averaged improvement of 3.3% over the holistic-only model *Ours(H)*.

Table 5
Results of *one-to-one* unsupervised domain adaptation on skeleton-based action recognition.

Method	CMU→HMD05	HMD05→CMU	Avg
Two-layer GRUs	81.1 ± 0.3	74.8 ± 0.4	78.0
ADDA [2]	87.6 ± 0.3	78.7 ± 0.3	83.2
DANN [1]	88.5 ± 0.5	82.1 ± 0.4	85.3
DSN [41]	90.2 ± 0.3	83.4 ± 0.4	86.8
MCD [23]	89.8 ± 0.3	85.7 ± 0.2	87.8
Ours(H)	88.9 ± 0.4	82.5 ± 0.3	85.7
Ours(L + H)	90.8 ± 0.2	85.8 ± 0.4	88.3
Ours(C + H)	90.5 ± 0.5	85.2 ± 0.4	87.9
Ours(L + C + H)	**91.3** ± 0.3	**86.7** ± 0.2	**89.0**

4.2. Partial domain adaptation

In this section, we evaluate the effectiveness of the proposed method on the more practical and challenging partial domain adaption task where the source domain is assumed to be diverse enough and the target label space is a subspace of the source label space [17]. In this case, many source points in the feature space (semantic features) are irrelevant to the target domain, which tends to cause *negative transfer* of holistic features with standard domain adaptation approaches.

4.2.1. Setup

We experiment on the *Office-31* dataset [38], with the source domain containing the whole 31 categories while the target domain only includes the 10 categories shared by the *Office-31* and *Caltech-256*, as done in [17]. We evaluate with the following six transfer tasks: A31→W10, D31→W10, W31→D10, A31→D10, D31→A10 and W31→A10.

4.2.2. Experimental results

The results of partial domain adaptation are shown in Table 6. *Negative transfer* occurs on the typical adversarial domain adaptation method DANN which is outperformed by the finetuned AlexNet. The SAN is specially designed for partial domain adaptation and attempts to get rid of the negative transfer of irrelevant features by selecting out the outlier source features. The proposed method adapts local features and cell features within each separated local cell to respect the local statistic of each cell, which promotes the positive transfer of domain-sharable local feature patterns and reduces the influences of irrelevant features. As shown in Table 6, with the introduced local feature and cell adaptation, *Ours(L + H)* and *Ours(C + H)*, reduce the *negative transfer* significantly over *Ours(H)*, with improvements of 5.4% and 4.6%, respectively. The combination of the two brings *Ours(L + C + H)* a considerable averaged improvement of 6.9% over the holistic-only model *Ours(H)*. When combined with the SAN, we are able to achieve the state-of-the-art performance.

4.3. Domain-agnostic adaptation

In this section, we evaluate the transferability of the learned local feature patterns on the domain-agnostic learning task [18], a more difficult and practical problem of adaptation from one labeled source domain to multiple target domains in which both the category and domain labels are unavailable. For examples, the source domain might be Amazon while the target be a combination of Dlsr, Webcam, and Caltech. We argue that safely transferable features should be safely applicable across multiple target domains. We experiment on the *Office-Caltech10* dataset [3]. The results of domain-agnostic adaptation are shown in Table 7. Our method achieves performance comparable to the state-of-the-art method DADA which is specifically designed for this

Table 6
Accuracy (%) of the partial domain adaptation task on the *Office31* dataset.

Method	A31→ W10	D31→ W10	W31→ D10	A31→ D10	D31→ A10	W31→ A10	Avg
AlexNet	58.5	95.1	98.1	71.2	70.6	67.7	76.9
DAN [6]	56.5	71.9	86.8	51.9	50.4	52.3	61.6
DANN [1]	49.5	93.6	90.4	49.7	46.7	48.8	63.1
RTN [22]	66.8	86.8	99.4	70.1	73.5	76.4	78.8
Ours(H)	65.9	94.9	93.6	66.3	65.8	64.3	75.1
Ours(L + H)	68.5	95.2	96.7	73.2	74.6	74.8	80.5
Ours(C + H)	67.7	94.7	96.2	73.3	73.8	72.3	79.7
Ours(L + C + H)	72.3	95.9	98.3	74.6	75.4	75.3	82.0
SAN [17]	80.0	98.6	100.0	81.3	80.6	83.1	87.3
Ours (L + C + H) + SAN [17]	**82.1**	**98.8**	**100.0**	**83.8**	**80.9**	**83.6**	**88.2**

Table 7
Accuracy (%) of domain-agnostic adaptation task on the *Office31* dataset. The model is trained using data from one labeled source domain and multiple unlabeled target domains.

Method	A→C,D,W	C→A,D,W	D→A,C,W	W→A,C,D	Avg
AlexNet	83.1 ± 0.2	88.9 ± 0.4	86.7 ± 0.4	82.2 ± 0.3	85.2
RTN [22]	85.2 ± 0.4	89.8 ± 0.3	81.7 ± 0.3	83.7 ± 0.4	85.1
JAN [7]	83.5 ± 0.3	88.5 ± 0.2	80.1 ± 0.3	85.9 ± 0.4	84.5
DANN [1]	84.8 ± 0.3	89.7 ± 0.2	87.8 ± 0.4	89.1 ± 0.4	87.9
Ours(H)	85.0 ± 0.4	90.3 ± 0.2	88.3 ± 0.4	90.2 ± 0.4	88.5
Ours(L + H)	85.3 ± 0.3	90.9 ± 0.3	88.6 ± 0.3	90.4 ± 0.4	88.8
Ours(C + H)	85.5 ± 0.3	91.7 ± 0.4	88.5 ± 0.4	90.3 ± 0.3	89.0
Ours(L + C + H)	85.9 ± 0.2	**92.3** ± 0.3	89.0 ± 0.4	90.7 ± 0.3	89.5
DADA [18]	**86.3** ± 0.3	91.7 ± 0.4	**89.9** ± 0.3	**91.3** ± 0.3	**89.8**

task through explicit feature disentanglement. *Negative transfer* occurs in the standard domain adaptation methods RTN and JAN, both of which adapt multi-layer holistic deep features. With the transferable local feature patterns, *Ours(H)* outperforms DANN, and the introduced local and cell feature adaptation bring *Ours(L + C + H)* an averaged improvement of 1.0% over the holistic-only *Ours(H)*.

4.4. Discussions

4.4.1. Feature generalization

We study the cross-domain generalization ability to evaluate the transferability of the learned domain-invariant features on the *Office-Caltech10* dataset. We postulate that a safely transferable feature is able to generalize to multiple novel domains. We perform adaptation from one source domain to one target domain, and evaluate the performance of the learned features in novel unseen domains. For example, we may use the *Dslr* as the source domain and A*mazon* as the target domain, and after the *one-to-one* adaptation, the feature transferability is evaluated in the *Webcam* and *Caltech* combined. We experiment with four representative scenarios which are built based on the complexity of source and target domains. The results are reported in Table 8. To be noted, *Dslr* and *Webcam* are small and plain while *Amazon* and *Caltech* are more complex and diversified. As we can see, DANN improves the performance over the AlexNet in the target domain while deteriorates the feature generalization ability to novel domains. The proposed method enhances feature transferability both in the target domains and unseen domains. Specifically, both *Ours(L + H)* and *Ours(C + H)* show distinct advantages over *Ours(H)*, which verifies the importance of local and cell feature adaptation in strengthening feature transferability. *Ours (L + C + H)* achieves a significant improvement of 4.1% over the DANN.

4.4.2. Target error bound

In this section, we study the target domain error bound on two representative tasks: *A→W* on the *Office-31* and *W→A,D,C* on the *Office-Caltech10*. As formulated in Eq. 10, the expected target error is upper-bounded by the following three terms: the expected source error, domain discrepancy, error of ideal joint hypothesis on both domains. Considering that source error is usually optimized to be ignorably small with the available source labels, only the latter two terms are considered.

Domain Discrepancy. The $\mathcal{A}$-distance is suggested as a measure of domain discrepancy in [20], defined as $d_A = 2(1 - 2\epsilon)$, where ϵ is the generalization error of a domain classifier trained to distinguish the source domain and target domain features. As shown in Fig. 4, the proposed method *Ours(L + C + H)* reduces domain discrepancy over the fine-tuned AlexNet more significantly than DANN and *Ours(H)* on both tasks. The results show that local and cell feature adaptation make for fine-grained domain shift reduction.

Error of Ideal Joint Hypothesis. We study the third term, joint error of an ideal hypothesis, by training a two-layer MLP classifier on the adapted features from both source and target, using their category labels, as done in [14]. Better classification accuracy indicates the learned features are more discriminative in classifying different categories. As we can see in Fig. 4, comparing to the *Ours(H)* or DANN, the feature discriminability of *Ours(L + C + H)* are less deteriorated while reducing the second term to strenghen transferability and thus with smaller joint errors, which verify that local feature patterns are more transferable than holistic features.

Combining the $\mathcal{A}$-distance and joint error of the ideal hypothesis, we show that the proposed *Ours(L + C + H)* reduces domain discrepancy (the second term in Eq. 4) more significantly while keeping feature discriminability (the third term) less deteriorated, thus successfully reducing the upper bound of target error.

4.4.3. Alignment visualization

We visualize the learned holistic, cell, and local features of fine-tuned AlexNet, DANN, and the proposed *Ours(L + C + H)* in Fig. 5 and Fig. 6 using the t-SNE embedding on the $A \rightarrow W$ transfer task. For the fine-tuned AlexNet, there exists obvious domain shift in both the holistic features, shown in Fig. 5 (a), and local features, shown in Fig. 5 (c). DANN successfully brings closer the source and target in both holistic and local feature space, as shown in Fig. 5 (b) and (d). However, there are still many boundary confusing holistic features lying between different category clusters which tend to be misclassified.Meanwhile, the local features are mixed up arbitrarily. As shown in Fig. 6, while still with a few mismatched samples, Ours (H) achieves better alignments over the DANN by only aligning the holistic features, which verify the effectiveness of sharing local feature patterns on bridging source and target domains. Through hierarchical adaptation, the holistic features, cell features, and local features are all well aligned by the proposed model Ours (L + C + H).

5. Conclusions

In this paper, we proposed a novel hierarchical domain adaptation approach which is built on local feature patterns. We showed that the typically-adopted holistic features are more biased and less transferablethan local feature patterns, which are more primitive and sharable not only across domains but also among different categories. Domain adaptation on holistic features is likely to cause unreliable feature transfer, especially when the holistic features are not well pre-trained.

Table 8
Accuracy (%) of domain generalization on the *Office-Caltech10* dataset. The model is trained using data from one labeled source domain and one unlabeled target domain while the performance is evaluated both on the target domain and on other unseen domains.

Method	A→W		W→D		D→A		C→A		Avg	
	W	C,D	D	A,C	A	C,W	A	D,W	Target	Unseen
AlexNet	75.0 ± 0.3	86.2 ± 0.4	98.6 ± 0.2	69.8 ± 0.6	83.6 ± 0.4	86.1 ± 0.6	91.7 ± 0.3	85.0 ± 0.4	87.2	81.8
DANN	88.7 ± 0.4	85.5 ± 0.5	99.2 ± 0.2	66.2 ± 0.5	87.3 ± 0.3	81.1 ± 0.5	93.8 ± 0.2	87.1 ± 0.3	92.3	80.0
Ours(H)	89.0 ± 0.3	86.9 ± 0.3	99.5 ± 0.3	68.3 ± 0.3	87.6 ± 0.3	83.7 ± 0.3	94.0 ± 0.3	87.4 ± 0.3	92.5	81.6
Ours(L + H)	89.3 ± 0.3	88.7 ± 0.4	99.6 ± 0.4	69.9 ± 0.5	88.6 ± 0.3	86.5 ± 0.5	94.1 ± 0.3	87.6 ± 0.4	92.9	83.2
Ours(C + H)	89.8 ± 0.4	89.3 ± 0.3	99.5 ± 0.3	70.6 ± 0.4	89.4 ± 0.2	87.2 ± 0.4	94.1 ± 0.2	87.5 ± 0.3	92.3	83.7
Ours(L + C + H)	**90.1** ± 0.3	**89.8** ± 0.3	**99.8** ± 0.3	**71.5** ± 0.3	**89.9** ± 0.3	**87.5** ± 0.3	**94.3** ± 0.3	**87.7** ± 0.3	**93.5**	**84.1**

To preserve the learned discriminative feature structures and prevent false feature alignments, we proposed to partition local feature space into cells by learning a cluster of local feature patterns and then perform feature adaptation in multiple levels. We experimentally showed that the proposed method boosts performance by minimizing the domain discrepancy without severely deteriorating the feature discriminability or structures. We showed that the local feature patterns are not limited to be spatial in images, but also applicable to temporal patterns in sequential data, or spatiotemporal patterns in videos. We showed that the proposed method is not only beneficial to the typical *one-to-one* unsupervised domain adaptation but also to the other domain adaptation tasks, *e.g.*, partial domain adaptation, domain-agnostic learning, domain generalization, which are commonly encountered in machine learning practice.

Despite the promise of the proposed method, the performance improvements on the benchmarks are still not very significant. This is mainly because we only perform the typical adversarial adaptation on local, cell and holistic features, without any further structural regularization on their statistics, *e.g.*, clustering assumptions. When combined with these techniques in practice, further boosted performances are highly expected. Another limitation of the proposed method is that we borrowed the local feature aggregation strategy from the NetVLAD, while which is not specifically designed for feature transfer. In the future, we intend to explore more effective feature aggregation strategies tfor domain adaptation. Further, besides the single modality data, we aim to explore feature aggregation strategies for domain adaptation on multi-modality data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Zhejiang Lab (2020KB0AC02), National Key R&D Program of China (SQ2020YFB130047, 2020YFB1313501, 2020YFB1313503), Zhejiang Provincial Natural Science Foundation (LR19F020005), the National Natural Science Foundation of China (61972347, 31070944, 31271147, 61922019, 31471063, 31671074, and 61572433) and the Fundamental Research Funds for the Central Universities, China (2017FZA7003). Qian Zheng is supported by the Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University, Singapore.

References

[1] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, The Journal of Machine Learning Research 17 (1) (2016). 2096–2030

[2] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7167–7176.

[3] B. Gong, Y. Shi, F. Sha, K. Grauman, Geodesic flow kernel for unsupervised domain adaptation. Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2066–2073.

[4] B. Fernando, A. Habrard, M. Sebban, T. Tuytelaars, Unsupervised visual domain adaptation using subspace alignment. Proceedings of the IEEE international conference on computer vision, 2013, pp. 2960–2967.

[5] Z. Pei, Z. Cao, M. Long, J. Wang, Multi-adversarial domain adaptation. AAAI Conference on Artificial Intelligence, 2018.

[6] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks. International conference on machine learning, PMLR, 2015, pp. 97–105.

[7] M. Long, H. Zhu, J. Wang, M.I. Jordan, Deep transfer learning with joint adaptation networks. Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 2208–2217.

[8] M. Ghifary, W.B. Kleijn, M. Zhang, D. Balduzzi, W. Li, Deep reconstruction-classification networks for unsupervised domain adaptation. European Conference on Computer Vision, Springer, 2016, pp. 597–613.

[9] B. Sun, K. Saenko, Deep coral: Correlation alignment for deep domain adaptation. European Conference on Computer Vision, Springer, 2016, pp. 443–450.

[10] M.M. Rahman, C. Fookes, M. Baktashmotlagh, S. Sridharan, Correlation-aware adversarial domain adaptation and generalization, Pattern Recognit 100 (2020) 107124.

[11] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, B. Wang, Moment matching for multi-source domain adaptation. Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1406–1415.

[12] H. Venkateswara, J. Eusebio, S. Chakraborty, S. Panchanathan, Deep hashing network for unsupervised domain adaptation. Proc. CVPR, 2017, pp. 5018–5027.

[13] A. Rozantsev, M. Salzmann, P. Fua, Beyond sharing weights for deep domain adaptation, IEEE Trans Pattern Anal Mach Intell 41 (4) (2018) 801–814.

[14] X. Chen, S. Wang, M. Long, J. Wang, Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. International Conference on Machine Learning, 2019, pp. 1081–1090.

[15] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks? Advances in Neural Information Processing systems, 2014, pp. 3320–3328.

[16] R. Shu, H.H. Bui, H. Narui, S. Ermon, A dirt-t approach to unsupervised domain adaptation. Proc. 6th International Conference on Learning Representations, 2018.

[17] Z. Cao, M. Long, J. Wang, M.I. Jordan, Partial transfer learning with selective adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2724–2732.

[18] X. Peng, Z. Huang, X. Sun, K. Saenko, Domain agnostic learning with disentangled representations. International Conference on Machine Learning, 2019, pp. 5102–5112.

[19] J. Wen, R. Liu, N. Zheng, Q. Zheng, Z. Gong, J. Yuan, Exploiting local feature patterns for unsupervised domain adaptation. Thirty-Third AAAI Conference on Artificial Intelligence, 2019.

[20] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J.W. Vaughan, A theory of learning from different domains, Mach Learn 79 (1–2) (2010) 151–175.

[21] S.J. Pan, I.W. Tsang, J.T. Kwok, Q. Yang, Domain adaptation via transfer component analysis, IEEE Trans. Neural Networks 22 (2) (2011) 199–210.

[22] M. Long, H. Zhu, J. Wang, M.I. Jordan, Unsupervised domain adaptation with residual transfer networks. Advances in Neural Information Processing Systems, 2016, pp. 136–144.

[23] K. Saito, K. Watanabe, Y. Ushiku, T. Harada, Maximum classifier discrepancy for unsupervised domain adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3723–3732.

[24] J. Gao, A local domain adaptation feature extraction method. 2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), IEEE, 2013, pp. 526–530.

[25] X. Wang, L. Li, W. Ye, M. Long, J. Wang, Transferable attention for domain adaptation. Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 5345–5352.

[26] S. Li, C. Liu, Q. Lin, B. Xie, Z. Ding, G. Huang, J. Tang, Domain conditioned adaptation network. Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 11386–11393.

[27] Z. Luo, Y. Zou, J. Hoffman, L.F. Fei-Fei, Label efficient learning of transferable representations acrosss domains and tasks. Advances in Neural Information Processing Systems, 2017, pp. 164–176.

[28] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation. Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 3304–3311.

[29] J. Sivic, A. Zisserman, Video google: A text retrieval approach to object matching in videos. null, IEEE, 2003, p. 1470.

[30] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, J. Sivic, Netvlad: Cnn architecture for weakly supervised place recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5297–5307.

[31] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, B. Russell, Actionvlad: Learning spatio-temporal aggregation for action classification. CVPR, volume 2, 2017, p. 3.

[32] R. Arandjelovic, A. Zisserman, All about vlad. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2013, pp. 1578–1585.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets. Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[34] M. Long, Z. CAO, J. Wang, M.I. Jordan, Conditional adversarial domain adaptation. Advances in Neural Information Processing Systems 31, Curran Associates, Inc., 2018, pp. 1647–1657.

[35] J. Wen, N. Zheng, J. Yuan, Z. Gong, C. Chen, Bayesian uncertainty matching for unsupervised domain adaptation. Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, 2019, pp. 3849–3855.

[36] S. Xie, Z. Zheng, L. Chen, C. Chen, Learning semantic representations for unsupervised domain adaptation. International Conference on Machine Learning, 2018, pp. 5423–5432.

[37] X. Ma, T. Zhang, C. Xu, Gcan: Graph convolutional adversarial network for unsupervised domain adaptation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8266–8276.

[38] K. Saenko, B. Kulis, M. Fritz, T. Darrell, Adapting visual category models to new domains. European conference on computer vision, Springer, 2010, pp. 213–226.

[39] M.-H. Chen, Z. Kira, G. AlRegib, J. Yoo, R. Chen, J. Zheng, Temporal attentive alignment for large-scale video domain adaptation. Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6321–6330.

[40] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset. proceedings of the IEEE Conference on Computer Vision and

Pattern Recognition, 2017, pp. 6299–6308.

[41] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, D. Erhan, Domain separation networks. Advances in Neural Information Processing Systems, 2016, pp. 343–351.

[42] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, Z. Gong, Unsupervised representation learning with long-term dynamics for skeleton based action recognition. Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 2644–2651.

Jun Wen is currently a postdoctoral Research Fellow with Harvard Medical School. He received the Ph.D. degrees in computer science from Zhejiang University in 2020, China. His research interests include transfer learning and healthcare data mining. He served as a reviewer for Pattern Recognition, T-IP, NeurIPS, CVPR, ICML, AAAI, ICCV, WACV, etc.

Junsong Yuan is Professor and Director of Visual Computing Lab at Department of Computer Science and Engineering (CSE), State University of New York at Buffalo (UB), USA. Before joining UB, he was Associate Professor (2015-2018) and Nanyang Assistant Professor (2009-2015) at Nanyang Technological University (NTU), Singapore. He obtained his Ph.D. from Northwestern University in 2009, M.Eng. from National University of Singapore in 2005, and B.Eng. from Huazhong University of Science Technology (HUST) in 2002. His research interests include computer vision, pattern recognition, video analytics, human action and gesture analysis, large-scale visual search and mining. He received Best Paper Award from IEEE Trans. on Multimedia, Nanyang Assistant Professorship from NTU, and Outstanding EECS Ph.D. Thesis award from Northwestern University. He served as Associate Editor of IEEE Trans. on Image Processing (T-IP), IEEE Trans. on Circuits and Systems for Video Technology (T-CSVT), Machine Vision and Applications (MVA), and Senior Area Editor of Journal of Visual Communications and Image Representation (JVCI). He serves as General Co-Chair, Program Co-Chair, or Area Chair for many conferences including CVPR/ICCV/ACM MM/ICME/ICIP/ICPR/WACV/ACCV. He is a Fellow of IEEE and IAPR.

Qian Zheng is currently a Research Fellow with the ROSE lab, Nanyang Technological University. His current research interests include computational photography and computer vision. He has published several papers in international journals and conferences, such as T-PAMI, T-IP, T-IFS, CVPR, ICCV, AAAI, and IJCAI. He is a guest editor of Frontiers in Neuroscience and a reviewer of T-IP, CVPR, ICCV, AAAI, and IJCAI.

Risheng Liu is currently a professor with DUT-RU International School of Information Science and Engineering, Dalian University of Technology. He was awarded the "Outstanding Youth Science Foundation" of the National Natural Science Foundation of China. His research interests include machine learning, optimization, computer vision and multimedia. He was a co-recipient of the IEEE ICME Best Student Paper Award in both 2014 and 2015. His two papers were selected as Finalist of the Best Paper Award in ICME 2017. His one paper was also awarded as the Best Open Source Project Award in ICME 2021.

Zhefeng Gong is currently a full Professor with the Department of Neurobiology, Zhejiang University School of Medicine. His current interests include neural decoding of Drosophila and neurosciences-inspired neural networks.

Nenggan Zheng is currently a full Professor in computer science with the Qiushi Academy for Advanced Studies, Zhejiang University. His current research interests include artificial intelligence, braincomputer interface and embedded systems.