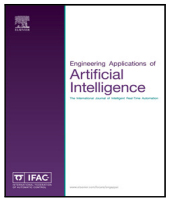
Engineering Applications of Artificial Intelligence 119 (2023) 105698

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Survey paper

Video surveillance using deep transfer learning and deep domain adaptation: Towards better generalization

Yassine Himeur [a,b,*], Somaya Al-Maadeed [a], Hamza Kheddar [c], Noor Al-Maadeed [a], Khalid Abualsaud [a], Amr Mohamed [a], Tamer Khattab [d]

[a] Computer Science and Engineering, Qatar University, Qatar
[b] College of Engineering and Information Technology, University of Dubai, Dubai, United Arab Emirates
[c] LSEA Laboratory, Electrical Engineering Department, University of Medea, Algeria
[d] Electrical Engineering Department, Qatar University, Qatar

ARTICLE INFO

Keywords:
Video surveillance
Deep learning
Deep transfer learning
Deep domain adaptation
Fine-tuning

ABSTRACT

Recently, developing automated video surveillance systems (VSSs) has become crucial to ensure the security and safety of the population, especially during events involving large crowds, such as sporting events. While artificial intelligence (AI) smooths the path of computers to think like humans, machine learning (ML) and deep learning (DL) pave the way more, even by adding training and learning components. DL algorithms require data labeling and high-performance computers to effectively analyze and understand surveillance data recorded from fixed or mobile cameras installed in indoor or outdoor environments. However, they might not perform as expected, take much time in training, or not have enough input data to generalize well. To that end, deep transfer learning (DTL) and deep domain adaptation (DDA) have recently been proposed as promising solutions to alleviate these issues. Typically, they can (i) ease the training process, (ii) improve the generalizability of ML and DL models, and (iii) overcome data scarcity problems by transferring knowledge from one domain to another or from one task to another. Although the increasing number of articles proposed to develop DTL- and DDA-based VSSs, a thorough review that summarizes and criticizes the state-of-the-art is still missing. To that end, this paper introduces, to the best of the authors' knowledge, the first overview of existing DTL- and DDA-based video surveillance to (i) shed light on their benefits, (ii) discuss their challenges, and (iii) highlight their future perspectives.

1. Introduction

1.1. Elementary

Deep learning (DL)-based video surveillance systems (VSSs) have attained various inspiring results in recent years when applied to different tasks, including crowd counting (CC) (Sánchez et al., 2020), abnormal event detection (AED) (Belhadi et al., 2021), object detection (OD) (Zaidi et al., 2022), human action recognition (HAR) (Sun et al., 2019), etc. By representing high-level abstractions using multiple layers of non-linear transformations, deep networks simulate the perception of human brains. In doing so, DL models require copious amounts of data to be trained. However, these algorithms work well with some applications where it is easy to get the data, but they put many other applications in disadvantageous positions. This is because (i) there are not sufficient resources or scales of data needed for training DL models from scratch; (ii) collecting large-scale datasets is an expensive and time-consuming process; (iii) most DL models rely on supervised learning where datasets should be labeled prior to the training process, and (iv) involving human experts in manually labeling training datasets represents another substantial cost and effort (Jiao et al., 2021b).

Existing DL architectures, including convolutional neural networks (CNNs) (Gu et al., 2018), stacked autoencoders (SAEs) (Zhang et al., 2021e), and deep belief networks (DBNs) (Gochoo et al., 2021), among others, have the ability of learning deep and transferable representations. However, domain shifts between the training and testing datasets can significantly affect their performance. This is mainly due to the transition of deep features from general to specific representations and the sharp decrease in representation transferability in higher layers. Specifically, in many studies, it is assumed that testing and training samples have been taken from the same domain, meaning that the data distribution and the input feature space are the same. Nevertheless, in numerous real-world applications, this assumption

* Corresponding author at: College of Engineering and Information Technology, University of Dubai, Dubai, United Arab Emirates.
E-mail addresses: yassine.himeur@qu.edu.qa (Y. Himeur), s_alali@qu.edu.qa (S. Al-Maadeed), kheddar.hamza@univ-medea.dz (H. Kheddar), n.alali@qu.edu.qa (N. Al-Maadeed), k.abualsaud@qu.edu.qa (K. Abualsaud), amrm@qu.edu.qa (A. Mohamed), tkhattab@ieee.org (T. Khattab).

https://doi.org/10.1016/j.engappai.2022.105698
Received 16 July 2022; Received in revised form 5 November 2022; Accepted 28 November 2022
Available online xxxx
0952-1976/

Abbreviations

AI	artificial intelligence
AdaIN	adaptative instance normalization
AED	abnormal event detection
ASNet	adversarial scoring network
ATLnet	adaptive TL network
CC	crowd counting
CDAD	cross-domain anomaly detection
CDAR	cross-domain human action recognition
CDCC	cross-domain crowd counting
CDDF	cross-domain data fusion
CDOD	cross-domain object detection
CMAR	cross-media action recognition
CNN	convolutional neural networks
CSCC	cross-scene crowd counting
CSPD	cross-spectral pedestrian detection
CSAR	cross-spectral action recognition
CVAR	cross-view action recognition
CVCS	cross-view cross-scene
DAN	density adaption network
DBN	deep belief networks
DDA	dynamic distribution alignment
DDAN	domain-adversarial neural network
DML	distance metric learning
DL	deep learning
DKPNet	domain-specific knowledge propagating network
DSPNet	deep scale purification network
DTL	deep transfer learning
EDIREC-Net	error-aware density isomorphism reconstruction network
ELM	extreme learning machine
FCL	fully-connected layer
GAN	generative adversarial networks
GRU	gated recurrent unit
HAR	human action recognition
HCN	high-density counter network
HOG	histogram of gradient
ILAN	instance-level adaptation network
ILRT	inter-layer relation transfer
ILSVRC	ImageNet large scale visual recognition challenge
IPT	intra-layer pattern transfer
IRN	Inception-residual network
KAIST	Korea advanced institute of science & technology
KNN	K-nearest neighbors
LDCN	low-density counter network
LRCN	long-term recurrent convolutional network
OD	object detection
M2AR	multi-modal action recognition
MAML	model agnostic meta-learning
MCNN	multi-column CNN
MCDCD	maximum cross-domain classifier discrepancy
MDDA	multi-source DDA
MDNet	multiple descriptor network
MIST	multiple instance self-training technique
MMD	maximum mean discrepancy
ML	machine learning
MSDN	multi-scale detection network
NLT	neuron linear transformation
PTM	pretrained model
RAE	recursive auto-encoders
ReLu	rectified linear unit
ROI	region of interest
RRN	region reconstruction network
S2V	sensor-to-vision
SAE	stacked autoencoder
SbE	serial-based extended
SD	source domain
SDDA	single-source deep domain adaptation
SHA	Shanghaitech Part_A
SHB	Shanghaitech Part_B
SKT	structured knowledge transfer
STL	subtask-dominated transfer learning
SVAR	sensor-to-vision action recognition
TD	target domain
TL	transfer learning
TSM	temporal shift module
UDDA	unsupervised deep domain adaptation
VGG	visual geometry group network
VSS	video surveillance systems
YOLO	you only look once

does not hold (Hazarika et al., 2021). Different research groups have proposed many datasets for VSS applications, such as CC, AED, OD, HAR, etc. Data collected by a research group might only include certain types of variations. For example, Shanghaitech Part_B (SHB) is taken from the streets of metropolitan Shanghai. In the WorldExpo'10 dataset, video sequences are captured by surveillance cameras from Shanghai 2010 WorldExpo. Shanghaitech Part_A (SHA), UCF, and AHU datasets are scrawled from the Internet. For instance, as illustrated in Fig. 1, it can be seen that considerable variations in data distributions exist among these datasets. While video frames in ShanghaiTech Part_A (SHA) (Zhang et al., 2016) illustrate congested crowds, those in UCF-QNRF (Idrees et al., 2018) present highly-congested crowds with more background scenarios. Additionally, those in NWPU-Crowd (Wang et al., 2020b) have much more diversities in scales, density, background, etc. However, images of ShanghaiTech B (SHB) (Zhang et al., 2016) refer to low-density crowds and usual street-based scenes. Because of these variations, learning a general and robust estimating is challenging to correctly predict density crowds or perform other related tasks. Moreover, many supervised learning-based VSSs can perform well using large-scale annotated data. However, most of them fail in real-world scenarios, where there are limited labeled data, complex backgrounds, different exposures, and points of view. Additionally, overfitting and the difficulty of generalizing to other benchmark repositories still need to be solved.

On the other hand, successful DL models are data-hungry and depend on the availability of comprehensive training labeled datasets and computational resources. However, many VSS tasks cannot secure enough annotated data to train DL models (Rezaee et al., 2022). Moreover, an increasing demand is recently witnessed to implement DL models on edge devices with limited computation capacities, especially after the progress made in federated learning. Additionally, most DL models can only perform one single task, while generalizing the acquired knowledge to other tasks requires a new set of data points and equal or more quantities of training data, which is not practical in real-world scenarios (Sayed et al., 2022). This means that the generalization of DL models needs to be improved, and their complexity

Fig. 1. Data distribution comparison between: (a) SHA (Zhang et al., 2016), (b) UCF-QNRF (Idrees et al., 2018), (c) NWPU (Wang et al., 2020b) and (d) SHB (Zhang et al., 2016).

must be reduced to support real-time VSS applications and improve their performance (Rajasekhar et al., 2021).

To that end, deep transfer learning (DTL) and deep domain adaptation (DDA) have recently been adopted to overcome the issues mentioned above. Typically, DTL, which consists of training a DL model for a task (or on a specific domain) and then utilizing the knowledge learned for a distinct but related task (pr on another different but related domain), has been investigated. Concretely, it has been shown that "reusing features – the specific quantifiable characteristics of a phenomenon being observed – plays an essential role in successful transfers" (Maschler and Weyrich, 2021). Humans have inspired this as they do not learn everything from scratch but can transfer their knowledge from previously learned domains to newer domains and tasks. Thus, considering the importance of this problem, the AI research community has been highly collaborative by developing several large-scale datasets, research works, and numerous models and making them open-source to build on them and facilitate the reproduction of research results (Loey et al., 2021)s. DTL uses considerably less data, which helps eliminate the necessity of data annotation. For instance, different CNN models have been trained on the ImageNet dataset for computer vision applications, representing a progressing work since 2005 that includes more than 14 million images classified and annotated across 80,000 groups. These pre-trained models (PTMs) can be used to perform other related tasks on small target datasets with or without labels (Han et al., 2021a). Additionally, training most existing DL models with AI-optimized, cutting-edge hardware can necessitate days, if not weeks. To that end, using DTL and building on top of the current DL architectures available as pre-trained and open-source weights is advisable (Han et al., 2021c). When screening the VSS literature, two main categories of DTL techniques are found. Fine-tuning is one approach that refers to taking a network model that has already been trained for a given task and making it perform a similar task. DDA is the other methodology, a specific case of DTL, where the feature spaces between the SD and TD are the same. At the same time, the marginal probability distributions of the input data are different. It assumes that the labeled and unlabeled data come from other, but related domains (Triess et al., 2021).

1.2. Our contributions

Although there are some generic reviews summarizing the concept and progress made in transfer learning (TL), e.g., Agarwal et al. (2021), Zhuang et al. (2020), Durrani and Arshad (2021), Yu et al. (2022), this paper presents the first specific review discussing the contributions of DTL for VSSs and creating a complete unified framework of categories and concepts. This enables the reader to scrutinize and understand the field of DTL and its applications for VSS tasks, such as OD, CC, AED, and HAR. Typically, the focus was on discussing both the fine-tuning- and DA-based VSS studies, where most of the latest contributions correspond to cross-domain object detection (CDOD), cross-domain anomaly detection (CDAD), cross-domain human action recognition (CDHAR), and cross-domain crowd counting (CDCC) are presented and analyzed. Moreover, interesting insights on the use of DTL for data fusion contexts are posed with reference to cross-domain data fusion (CDDF). In addition to the critical challenges determined in this review, future research directions serve as the pull factor toward generalized DL models in VSS tasks. All in all, the novel contributions of this overview can be summarized as follows:

- Presenting the DTL and DDA background that defines and examines the different aspects contributing to the development of DTL and DDA models.
- Introducing a comprehensive taxonomy of DTL and DDA models and thoroughly analyzing the existing literature on DTL-b and DDA-based VSSs and related concepts, covering more than 200 studies sorted based on various criteria.
- Shedding light on the interesting crossroads of data fusion and DTL/DDA, and highlighting how DTL/DDA can help develop efficient cross-domain data fusion (CDDF)-based VSSs.
- Enumerating a series of critical challenges of DTL and DDA models used in VSSs, which still need to be resolved or insufficiently investigated. Typically, different research needs are identified, including (i) the accuracy saturation issue, (ii) the concepts and metrics to measure the knowledge gain of DTL/DDA networks, (iii) the problem of negative transfer, (iv) the overfitting problem, etc.
- Outlining future research directions towards more generalized DTL/DDA models with less computational complexity.

Moving on, Table 1 depicts the results of the contribution comparison between the proposed review and other existing TL-based surveys. In addition to being the first review on TL-based VSSs, it is clearly seen that the proposed study presents many new contributions in terms of (i) including the DA part, (ii) discussing the different applications of TL in VSSs, (iii) discussing the pre-trained models, (iv) identifying currents challenges (e.g., negative transfer, knowledge gain measurement, unification of TL), and (iv) deriving future directions.

Table 1
Contribution comparison of the proposed study against other TL surveys. The tick mark (✓) indicates that the specific field has been addressed, whereas the cross mark (✗) means addressing the specific fields has been missed.

Survey	Description	TL Background	Domain adaptation	Applications of TL in VSSs	Pretrained models	Current challenges			Future directions
						Negative transfer	Knowledge gain	Unification of TL	
Lu et al. (2015)	TL for computational intelligence	✗	✗	✗	✗	✗	✗	✗	✗
Weiss et al. (2016)	General information on TL	✓	✗	✗	✗	✗	✗	✗	✗
Niu et al. (2020)	Generic TL contributions	✗	✗	✗	✗	✗	✗	✗	✗
Agarwal et al. (2021)	Categorization and general applications of TL	✗	✗	✗	✗	✗	✗	✗	✗
Zhuang et al. (2020)	Focus on homogeneous TL	✓	✗	✗	✗	✗	✗	✗	✗
Durrani and Arshad (2021)	TL for NLP	✓	✗	✗	✗	✗	✗	✗	✗
Wan et al. (2021)	TL in EEG	✓	✗	✗	✗	✗	✗	✗	✗
Bashath et al. (2022)	TL for text data	✓	✗	✗	✗	✗	✗	✗	✗
Li et al. (2021b)	TL for EEG	✗	✗	✗	✗	✗	✗	✗	✗
Yu et al. (2022)	TL for medical image analysis	✗	✗	✗	✗	✗	✗	✗	✗
Ours	TL for VSSs	✓	✓	✓	✓	✓	✓	✓	✓

Table 2
Distribution overviewed articles with reference to the name of journal/conference.

Journal/conference name	# articles	Journal/conference name	# articles
arXiv preprint	33	IEEE Trans. Pattern Analysis and Machine Intelligence	5
IEEE Conference on Computer Vision and Patter recognition	28	International Conference on Machine Learning	4
Neurocomputing	11	IEEE Acces	4
AAAI Conference on Artificial Intelligence	9	International Journal of Computer Vision	4
Neurocomputing	10	Sensors	4
IEEE Trans. Image Processing	8	IEEE Trans. Industrial Informatics	3
Information Fusion	8	International Conference on Image Processing	3
ACM International Conference on Multimedia	8	Engineering Applications of Artificial Intelligence	3
European Conference on Computer Vision (ECCV)	6	Int. Conf. on Pervasive Computing and Communications Workshops	3

1.3. Bibliometric analysis

To explore and analyze the scientific studies considered in this review, a bibliometric analysis is conducted. Fig. 2 gives a brief overview of the studies' long-term interest in DTL-based VSSs, where it can be seen that since 2015 an increasing interest has been witnessed in developing DTL-based VSS solutions. Typically, this interest is shown in Fig. 2(A), where the number of published articles has exponentially increased. The number of papers reached 83 in 2021 and 20 in the first quarter of 2022. Besides, Fig. 2(B) illustrates the most active researchers in the field of VSS-based DTL since 2015, in which only the authors that have produced more than two papers in the last half-decade are considered. In this review, 298 articles are discussed, which have been classified based on the application, as explained in Fig. 2(C). Accordingly, it is clearly shown that most of the frameworks have investigated the CDCC, CDHAR, CDOD, and CDAD tasks. At the same time, the CDDF area has received the most attention. On the other hand, the discussed articles include 146 research papers, 21 surveys, 111 conference papers, and 20 chapter books. Fig. 2(D) illustrates the percentage of each type of papers in our review. We can easily distinguish that the research journal papers take the lion's share of the whole published studies(49%), followed by conference papers (37%).

Table 2 presents statistics about the number of articles published in different journals, conference proceedings, preprints, etc. Typically, it is clearly seen that preprints come in the first position, followed by the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) and Neurocomputing. Besides, Fig. 3 illustrates the most significant themes covered in this review, which have been extracted using VOSviewer.[1] Accordingly, it can be shown that "transfer learning", "object detection", "domain adaptation", and "action recognition" were the most investigated themes.

Lastly, Table 3 provides some statistics concerning the authors' number of publications per affiliation and country. Explicitly, the top 10 institutions are illustrated in this table, where it can be seen that the University of Oxford is ranked first, followed by Zhejiang University and Nanjing University.

1.4. Paper organization

The rest of this paper is organized as follows. Section 2 presents the background and taxonomy of TL. Moving on, Section 3 overviews existing DTL studies and critically analyzes their pros and cons. Next, DDA contributions are discussed in Section 4 while Section 5 highlights the applications of DTL and DDA. Following, Section 6 discusses the key challenges of using DTL and DDA in VSSs before deriving their future directions in Section 7. Lastly, concluding remarks are drawn in Section 8. Fig. 4 illustrates the detailed road-map of our review article and outlines the main titles investigated in this study.

2. DTL background and taxonomy

2.1. Background

This section presents the background of DTL and explains its different categories. Moreover, it highlights its differences compared to DL. As depicted in Fig. 5, in DTL, the knowledge learned in one task is shared with other related but different tasks.

Def. 1 - Domain: Let us consider a specific dataset $X = \{x_1, \ldots, x_n\} \in \chi$, in which χ represents the feature space, and $P(X)$ refers to the marginal probability distribution of X. A domain is defined as $\mathbb{D} = \{X,\ P(X)\}$. In DTL, the domain that contains the initial knowledge is defined as the source domain (SD), and it is represented by $\mathbb{D}_S$. By contrast, the domain including the unknown knowledge to be learned is named the target domain (TD), it refers to $\mathbb{D}_T$ (Lu et al., 2021a).

Definition 2 - Task: Considering the previously defined dataset $X = \{x_1, \ldots, x_n\} \in \chi$, which corresponds to a set of labels $Y = \{y_1, \ldots, y_n\} \in \gamma$, where γ represents the label space. A task can be defined as $\mathbb{T} = \{Y, \mathbb{F}(X)\}$, where $\mathbb{F}$ denotes the learning objective predictive function

[1] https://www.vosviewer.com/

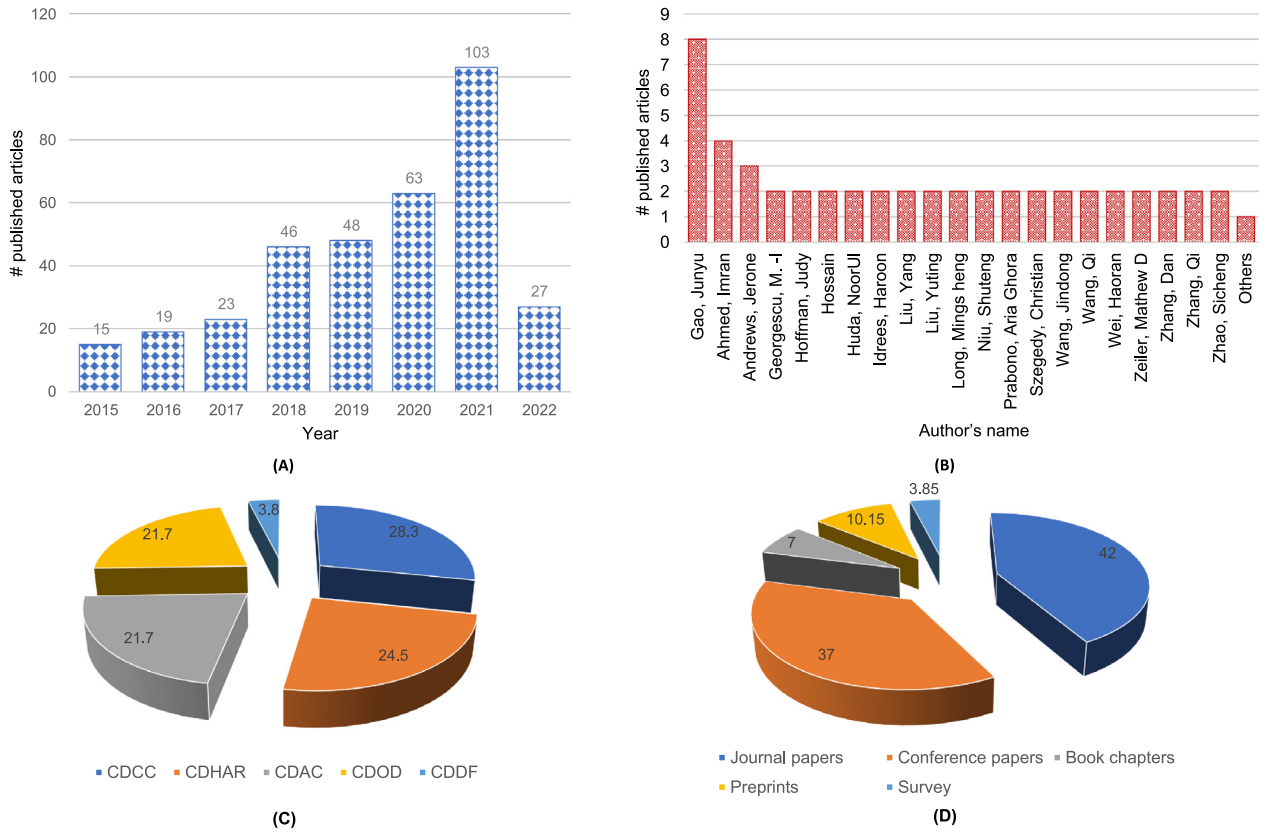


Fig. 2. Bibliometric analysis in terms of (A) the number of articles involved in this review in each year, (B) the most active and influenced authors in the DTL-based VSSs, (C) the percentage of papers involved in describing the applications of VSS-based DTL, and (D) the type of papers and percentage of the research articles involved in our survey.

Table 3
Publications per affiliation and country of the authors.

Affiliation	Country	# of articles
university of oxford	UK	25
Zhejiang University	China	13
Nanjing University	China	12
kings college london	UK	11
Comsats University Islamabad	Pakistan	10
Sandia National Laboratories	USA	10
University of Augsburg	Germany	10
Hitec University Taxila	Pakistan	9
Carnegie Mellon University	USA	8
Delhi Technological University	India	8
Huazhong University of Science and Technology	China	8

that could be represented as well as a conditional distribution $P(Y|X)$. Following the definition of task, the label spaces of the SD and TD are represented as γ_S and γ_T, respectively (Ramirez et al., 2019).

Although there is no comprehensive and standardized categorization of DTL methods, we can initially categorize them into various types based on *what*, *when*, and *how* knowledge is transferred. Thus, a well-defined taxonomy of DTL algorithms is presented in this section.

(a) What knowledge is transferred: inquires which characteristics of knowledge are transferable across domains or tasks. Some information is particular to certain domains or tasks, while other knowledge is shared across domains and can aid increase the performance in the target task or domain. Based on this definition, DTL is either feature-based, instance-based, relation-based, or model-based (Morid et al., 2021).

(b) How knowledge is transferred: inquires about which learning algorithms must be implemented to transfer the knowledge.

(c) When knowledge is transferred: inquires as to when and under what circumstances knowledge should or should not be transferred.

2.2. Taxonomy

A well-defined taxonomy of DTL methodologies used for VSSs is performed in this section. Fig. 6 depicts the proposed taxonomy, which has been sorted by (i) learning style, (ii) methodology, (iii) surveillance type, (iv) data annotation, and (v) popular DTL models. Typically, DTL techniques can be divided into several groups according to whether the source and target's domains and tasks are similar. Table 4 summarizes these possibilities.

2.2.1. Inductive DTL

In comparison to classical ML, which may be used as a reference for the DTL comparison, and given that the target tasks $\mathbb{T}_T$ are distinct from the source tasks $\mathbb{T}_S$, the goal of inductive DTL is to enhance the target prediction function $\mathbb{F}_T$ in the TD, mentioned above. However, the SD $\mathbb{D}_S$ and TD $\mathbb{D}_T$ may not always be the same (Table 4). The inductive

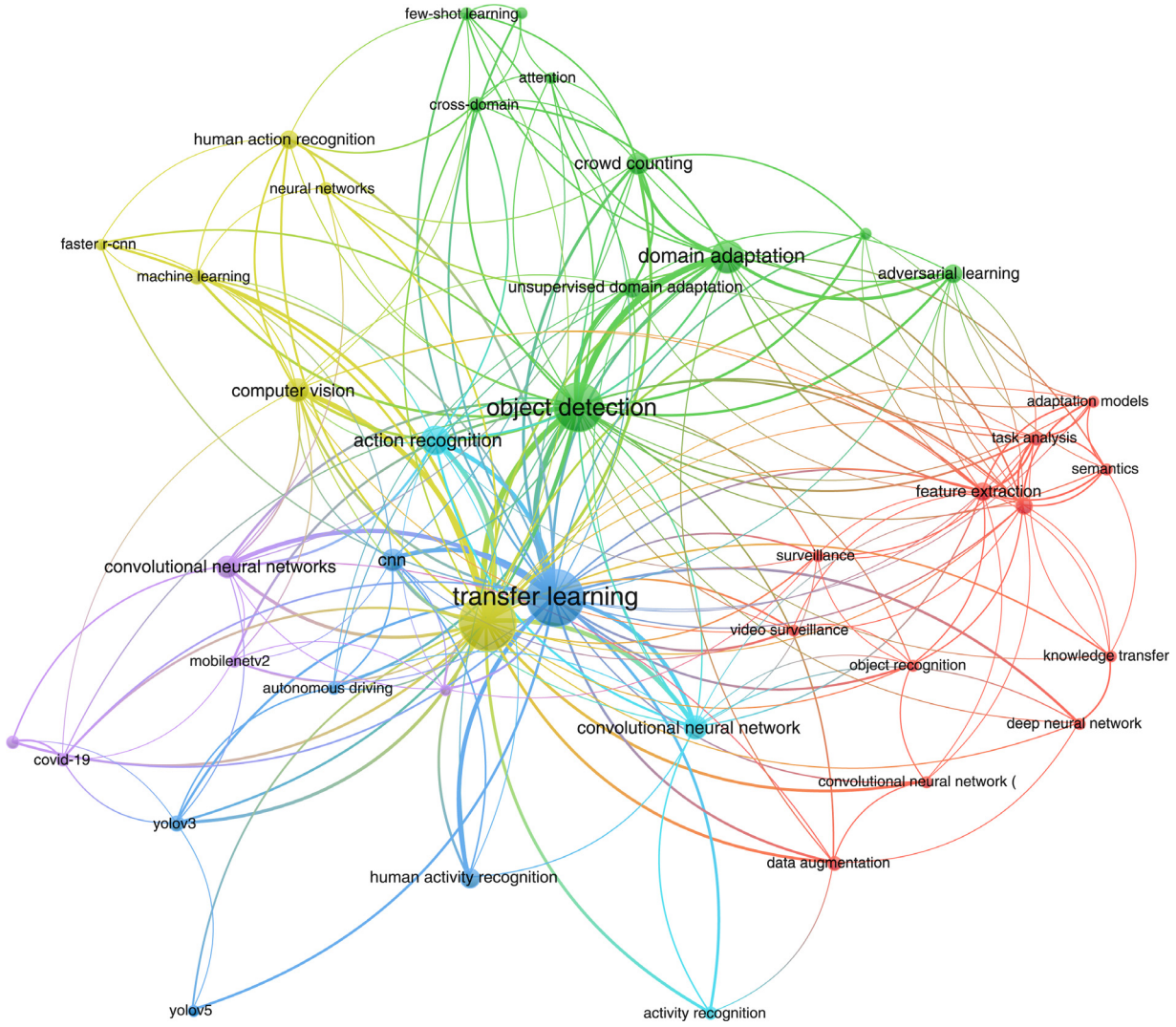


Fig. 3. Summary of the most significant themes covered in this review identified using knowledge graphs.

Table 4
DTL categorization is based on the similarity between the domains and tasks of the source and target, where the mark ($\subsetneq$) indicates that the domains/tasks are different but related, ($\exists!$) indicates that there exists one and only one domain/task, and ($\cong$) means that the domains, tasks, or spaces do not always equal.

	Domains	Tasks	Math. propriety	Sub-categories/Usage
Traditional ML/DL	$\mathbb{D}_S = \mathbb{D}_T$	$\mathbb{T}_S = \mathbb{T}_T$	$X_S \neq X_T$, $Y_S = Y_T$	The DL model is trained on the X_S dataset and used to recognize the X_T dataset.
Inductive DTL	$\mathbb{D}_S \cong \mathbb{D}_T$	$\mathbb{T}_S \neq \mathbb{T}_T$	$X_S \neq X_T$, $Y_S \exists, Y_T \exists$	If $Y_S \exists$, DTL is a multitask learning. If $Y_S \nexists$, DTL is a self-taught learning, thus $\chi_S \cong \chi_T$.
Transductive DTL	$\mathbb{D}_S \neq \mathbb{D}_T$	$\mathbb{T}_S = \mathbb{T}_T$	$P(X_S) \neq P(X_T)$, $Y_S \exists, Y_T \nexists$, $\chi_S = \chi_T$	When $\chi_S = \chi_T$, DTL is related to DDA. If $\mathbb{D}_T \exists!$ and $\mathbb{T}_T \exists!$, DTL is used for sample selection bias or covariate shift.
Cross-modality DTL	$\mathbb{D}_S \neq \mathbb{D}_T$	$\mathbb{T}_S \neq \mathbb{T}_T$	$P(Y_S/X_S) \neq P(Y_T/X_T)$, $Y_S \neq Y_T$, $\chi_S \neq \chi_T$	The SD and TD represent different data modalities, e.g., the dataset X_S of $\mathbb{D}_S$ is collected from wearable-sensors while the dataset X_T of $\mathbb{D}_T$ is from vision-sensors.
Unsupervised DTL	$\mathbb{D}_S \subsetneq \mathbb{D}_T$	$\mathbb{T}_S \subsetneq \mathbb{T}_T$	$Y_S \nexists, Y_T \nexists$	The DTL is used for clustering, dimensionality reduction, and density estimation,..., etc.

DTL can be stated similarly to the following two cases, depending on whether labeled or unlabeled data is available:

(a) Multi-task DTL: the SD has a huge labeled database (X_S labeled with Y_S), which is a distinctive form of multi-task learning. However, with the multi-task approaches, many tasks ($T_1, T_2, \ldots, T_n$) are learned at the same time (in parallel), including both the source and target tasks (Li et al., 2021b).

(b) Sequential learning: or commonly known as the *self-taught learning*, refers to the case where the dataset is not labeled in the SD. Sequential learning is based on (i) the feature representation transfer, learned from an extensive collection of unlabeled datasets, and (ii) the

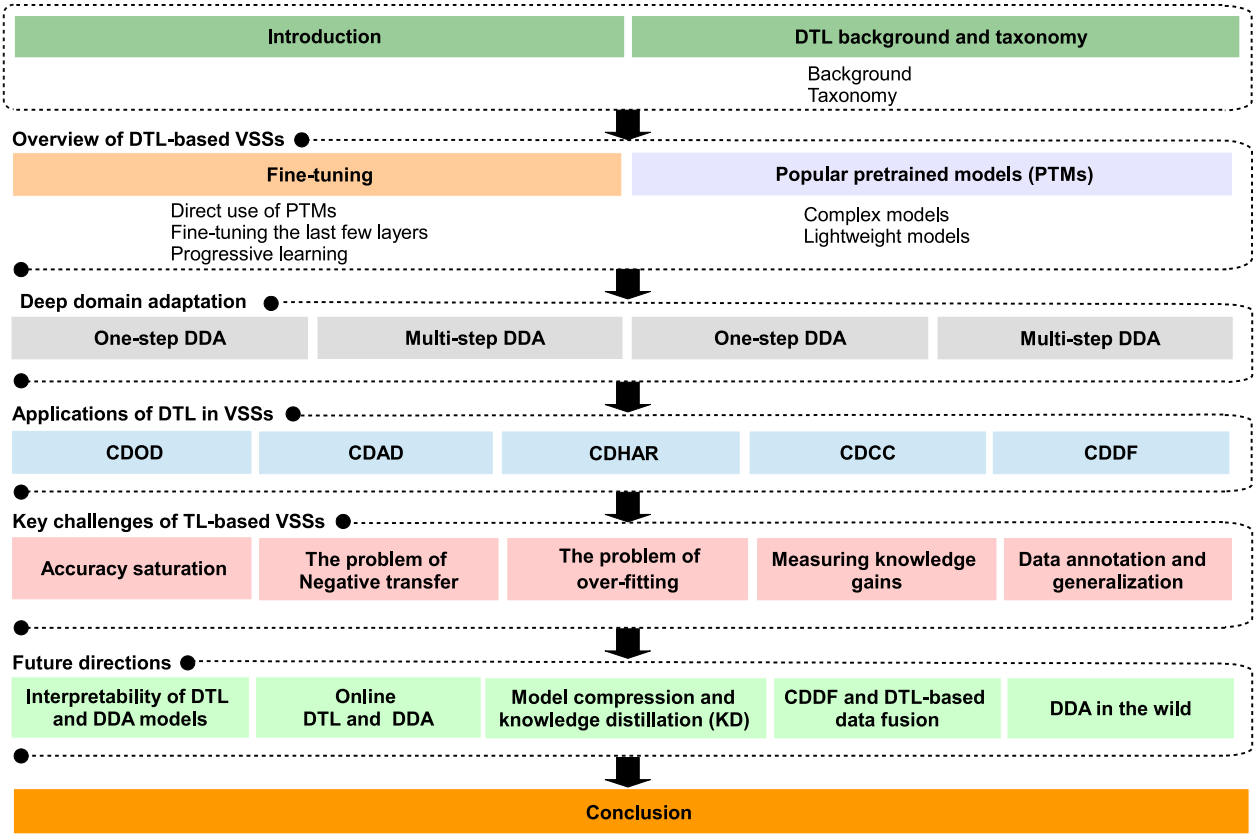


Fig. 4. Road-map of the review paper.

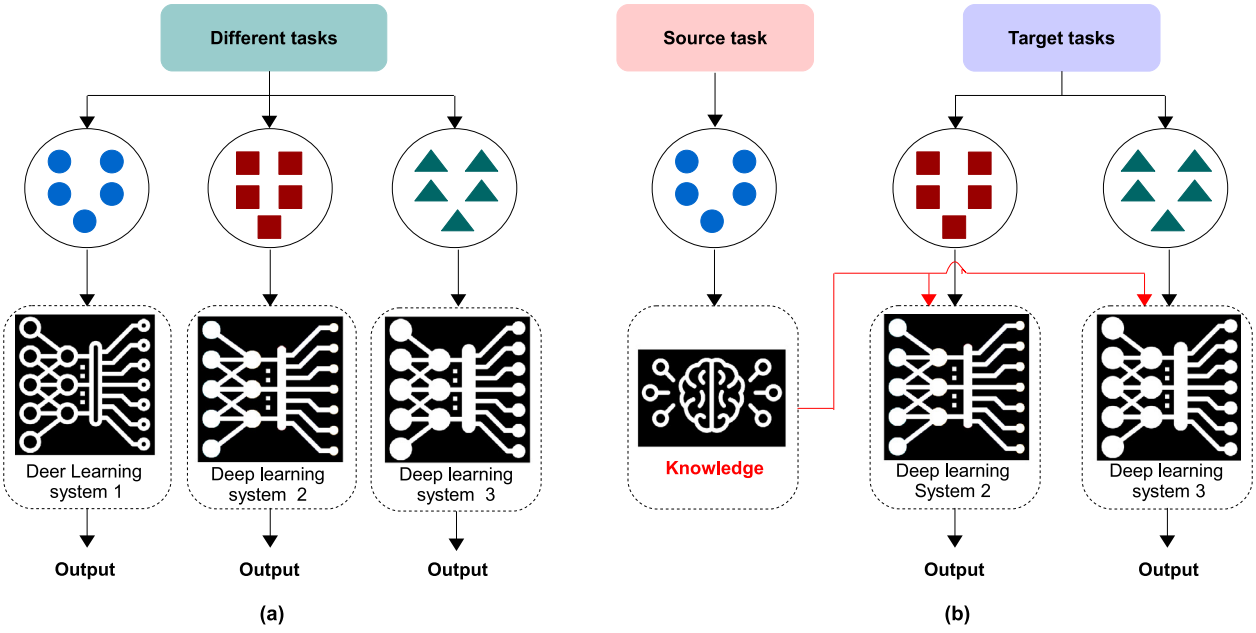


Fig. 5. Difference between conventional DL and DTL techniques for multiple tasks: (a) conventional ML and (b) TL.

learned representation is applied to labeled data to accomplish classification tasks. Hence, this DTL scheme refers to sequentially learning several activities (tasks) where the gaps between the SD and TD may differ. For example, let us suppose that we have a PTM M and consider applying DTL to multiple tasks $(T_1, T_2, \ldots, T_n)$, a specific task $\mathbb{T}_T$ at each time step t is learned, which is slower than the multi-task learning. However, when not all the tasks are present at the time of the training, it might be beneficial. Sequential learning can additionally be classified into several types (Alyafeai et al., 2020):

1- **Fine-tuning:** it is based on learning a new function $\mathbb{F}_T$ that translates the parameters $\mathbb{F}_T(W_S) = W_T$ by using a PTM M. W_S and W_T are the weights of source and target tasks $\mathbb{T}_S$ and $\mathbb{T}_T$, respectively. The settings can be adjusted across all the layers or just partially (Fig. 7) and the learning rate for each layer can be

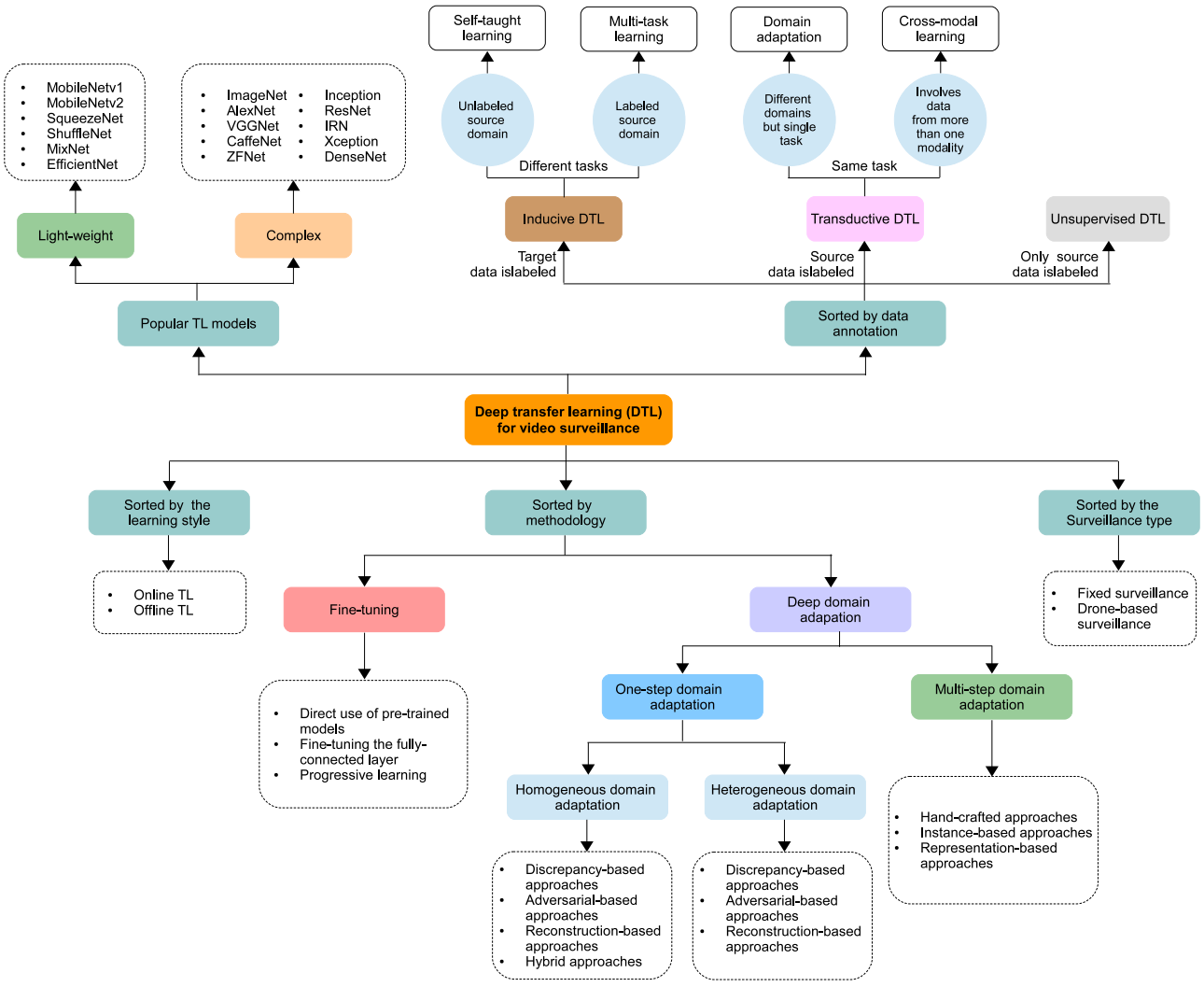


Fig. 6. Proposed taxonomy of existing DTL algorithms for VSSs.

distinct (discriminative fine-tuning). A new set of parameters K can be added to most of the tasks so that (Ribani and Marengoni, 2019):

$$\mathbb{F}_T(W_T, K) = W_S \times K \tag{1}$$

2- **Adapter modules:** given an M_S model that has been pretrained, and output the weights W_S for a target task $\mathbb{T}_T$. The adapter module aims to lunch a different set of parameters K lower than W_S, i.e., $K \ll W_S$. K and W_S must have the ability to be decomposed into more compact modules such that $W_S = \{w\}_n$ and $K = \{k\}_n$. The adapter module enables learning the following new function $\mathbb{F}_T$:

$$\mathbb{F}_T(K, W_S) = k'_1 \times w_1 \times \cdots k'_n \times w_n \tag{2}$$

According to Eq. (2), during the adaptation procedure, the set of original weights $W_S = \{w\}_n$ is left unaltered, but the set of weights K is changed to $K' = \{k'\}_n$. The principle of DDA is illustrated in Fig. 7.

3- **Feature based:** interested only in learning concepts and representations at various image levels, such as corners/interest points, blobs/regions of interest points, ridges, or edges E. The collection of E based on a PTM M remains unaltered, i.e., $\mathbb{F}_T(W_S, E) = E \times W'$, in the way that W' is fine-tuned.

4- **Zero-shot:** is the easiest method among all of the others. Making the assumption that the parameters W_S cannot be modified or add K as a new parameter to a PTM M_S using W_S. To put this into context, there is no training technique to optimize or learn new parameters in zero-shot.

2.2.2. Transductive DTL

Compared to the traditional ML, which can be considered as a reference for DTL comparison, and given that in practical scenarios, the TD $\mathbb{D}_T$ is distinct from the SD $\mathbb{D}_S$. The SD has a labeled dataset (X_S labeled with Y_S), whereas the TD has no labeled dataset. The source and target tasks are similar (Table 4). The goal of transductive DTL is to build a target prediction function $\mathbb{F}_T$ in the $\mathbb{D}_T$ using the knowledge of the $\mathbb{D}_S$ and $\mathbb{T}_T$. Furthermore, the transductive DTL environment may be further classified into two categories depending on different conditions between the source and destination domains (Wan et al., 2021):

(a) Deep domain adaptation (DDA): the feature spaces across domains, χ_S and χ_T, are identical. Still, the marginal probability distributions of the input dataset are not, $P(Y_S/X_S) \neq P(Y_T/X_T)$ (Liu et al., 2021a). DDA is most effective when the $\mathbb{T}_T$ has a distinct distribution or labeled data is scarce (Tan et al., 2018).

(b) Cross-modality DTL: in most DTL methods, more or less, a relation between feature spaces (or label spaces) should exist, i.e., $\mathbb{D}_S$ and $\mathbb{D}_T$. Put differently, DTL can only occur when the source and destination data are both in the same modality, like video, speech, or text. Cross-modality DTL, in contrast to all other DTL approaches, is one of the

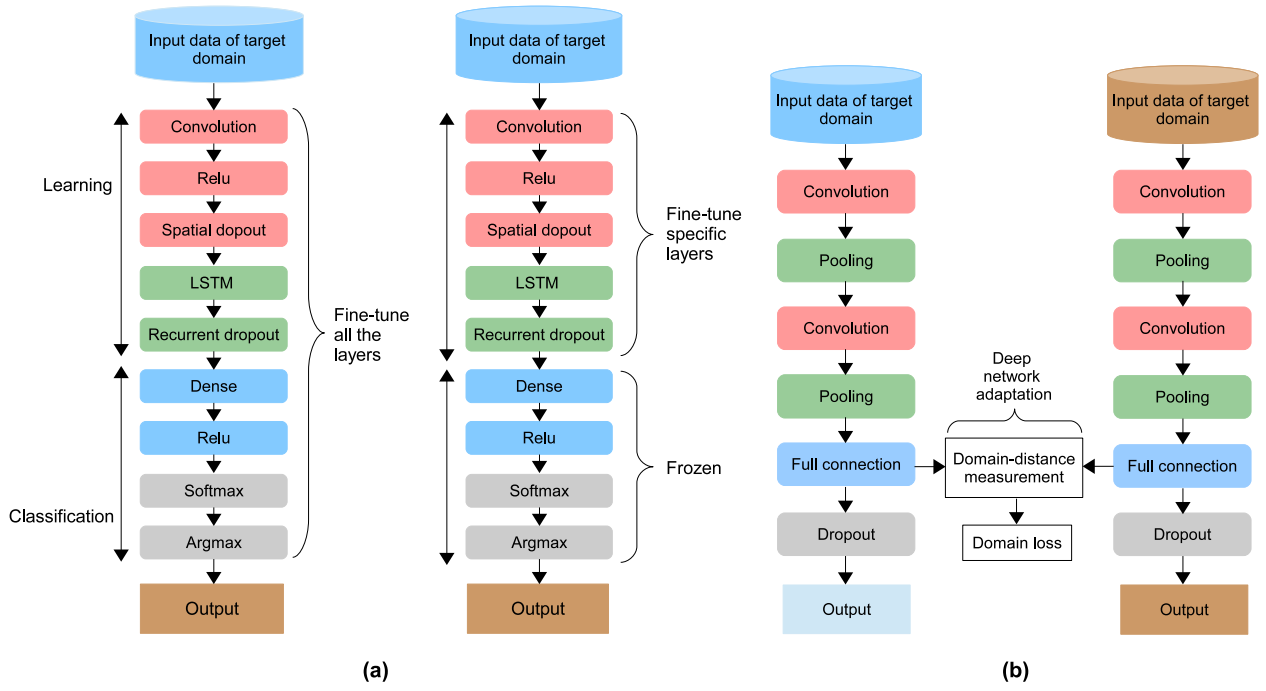


Fig. 7. Example of DTL models used in VSSs: (a) fine-tuning, and (b) DDA.

most complicated issues in DTL. It is assumed that the feature spaces of the source and destination domains are completely distinct ($\chi_S \neq \chi_T$), as in speech-to-image, image-to-text, and text-to-speech. Furthermore, the label spaces of source Y_S and destination Y_S domains might differ ($Y_S \neq Y_T$) (Niu et al., 2021).

(c) Unsupervised DTL: it aims to enhance the learning of the target predictive function $\mathbb{F}_T$ in $\mathbb{D}_T$ using the knowledge in $\mathbb{D}_S$ and $\mathbb{T}_S$, where $\mathbb{T}_S$ is different from $\mathbb{T}_T$ but related, and the labels Y_S and Y_T are not available (Si et al., 2021).

2.2.3. Adversarial DTL

In contrast to the methods described above, adversarial learning (Zhou et al., 2020) aids in learning more transferable and discriminative representations. The study in Ganin et al. (2016) was the first that introduced the domain-adversarial neural network (DANN). Instead of using a predefined distance function, e.g., the maximum mean discrepancy (MMD), the core idea is to use a domain-adversarial loss in the network. This has greatly aided the network's ability to learn more discriminative data. Many VSS studies have used domain-adversarial training as a result of DANN's idea (Shen et al., 2018; Georgescu et al., 2020; Choi et al., 2021; Georgescu et al., 2020; Soleimani and Nazerfard, 2021). All the previous works ignore the different effects of marginal and conditional distributions in adversarial DTL. In contrast, in Wang et al. (2020a), an approach based on dynamic distribution alignment is introduced, which can dynamically evaluate the importance of each distribution. To gradually bridge the gap among domains from coarse to fine granularity, a unique adversarial scoring network (ASNet) was developed in Zou et al. (2021). In particular, during the coarse-grained stage, adversarial learning is used to build a dual-discriminator technique to adjust the SD to be near the targets from both the global and local feature space viewpoints. Thus, the distributions of the two domains may be generally aligned. The transferability of source attributes is investigated at the fine-grained stage by scoring how similar source samples are to target samples at many levels using generative probability generated from the coarse stage. After that, the transferable source elements are carefully chosen to aid DTL during the adaptation process. The generalization bottleneck caused by the domain disparity may be successfully reduced using the coarse-to-fine architecture, as portrayed in Fig. 8. Specifically, the input photographs are encoded into density maps by the generator before classifying the density maps as SD or TD using the dual-discriminator. Next, domain distributions are pulled close through adversarial training between the dual discriminator and generator. In the meantime, the dual-discriminator generates four different kinds of scores as a signal to help optimize the density of the SD during adaptation, resulting in fine-grained transfer (Zou et al., 2021).

3. Overview of DTL-based VSSs

3.1. Fine-tuning

Model fine-tuning is a commonly adopted DTL scheme in VSS tasks, which helps fine-tune a pretrained DL network in a TD instead of training the whole architecture from scratch. This is a delicate task in VSSs because of the variations inside the TD, including distinct camera viewpoints, occlusion, and illumination changes. PTMs, e.g., ResNet-50, visual geometry group network (VGG), EfficientNet, and Inceptionv3 have reached significant success and become a milestone in computer vision. To that end, knowledge is encoded into huge parameters and fine-tuned on particular tasks. The rich knowledge implicitly stored in huge parameters can benefit a variety of downstream tasks. Three main techniques are typically adopted to fine-tune pretrained CNN models:

3.1.1. Direct use of PTMs

Directly applying an overall pretrained DL model (end-to-end) from a source task is the simplest approach for solving a VSS target task. Pretrained DL models include millions of parameters, trained for days or even weeks on state-of-the-art machines. In Sahoo et al. (2019), the weights of pretrained VGG-16 and InceptionV3 are fine-tuned to extract features from a target dataset. Similarly, in Atghaei et al. (2020), a pretrained VGG-16 is employed to identify the originality of input patches that are produced by the generator of a GAN model in an attempt to learn the normal data distribution in videos. In Liu et al. (2020c), different 2D-CNN models are fine-tuned to recognize actions in video sequences, including ResNet-50, ResNet-101, ResNeSt-50 and ResNeSt-101. Besides, in Doshi and Yilmaz (2020), to extract location and appearance features, a pretrained OD system based on YOLOv3 is utilized.

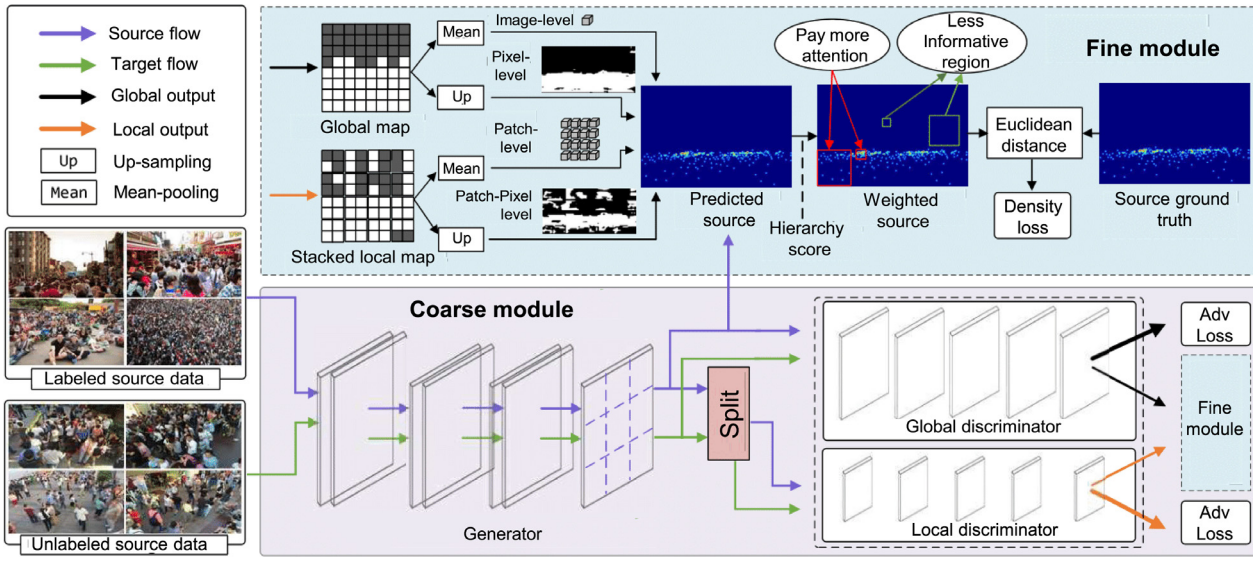


Fig. 8. The block diagram of the ASNet model based on adversarial DTL proposed in Zou et al. (2021).

3.1.2. Fine-tuning the last few layers

In this case, the first several layers usually used as a feature extractor are fixed (or frozen), and only the last few layers are fine-tuned. Typically, the bottom and mid-level layers are considered general features, while the top layers represent the problem-specific features. In this regard, by fine-tuning the last few layers, two main benefits are achieved: (i) the knowledge learned in the SD can be preserved when freezing the first several layers, (ii) the model is adapted to the TD by fine-tuning the last few layers. Therefore, the knowledge from the SD and TD can be combined to improve the performance (Zhang et al., 2016). This fine-tuning strategy has been widely adopted in VSS problems. For instance, in Zhang et al. (2016), a DTL-based CC approach is developed using a multi-column CNN (MCNN) model. The latter has been trained on a large-scale dataset that contains heads of very different sizes before transferring it to other datasets whose crowd heads have different sizes. If the TD dataset has a few training samples, the first layers in every column of MCNN are frozen, and only the last few layers are fine-tuned. There are two advantages to fine-tuning the last few layers in this case.

In Ahmadi et al. (2020), Ahmadi et al. fine-tune a PTM by replacing the softmax layer and running backpropagation for video OD. Yet, another work in Yu et al. (2021) study the problem of few-shot learning for video OD produced from the frequently used ImageNet VID dataset. To efficiently train the video object detector on many base-class items and a few video clips of novel-class objects, a method called *Thaw* is proposed. Typically, the authors first freeze the feature extractor and fine-tune the detection head when developing a few-shot learning scheme for video OD. Zhang et al. (2016) propose an MCNN for CC based on density map estimation TL. In doing so, two scenarios are implemented via (i) fine-tuning the overall network and (ii) fine-tuning the last two layers. Typically, it has been seen that the second strategy presents a better performance in terms of the MAE and MSE. Similarly, in Liu et al. (2020b), unsupervised CC is explored in a DTL setting, where people counting is learned in an unlabeled TD by transferring bi-knowledge learned from regression- and detection-based models in a labeled SD. Moving on, Delussu et al. (2022) introduce a TL-based scheme to learn scene-specific CC algorithms when either representative unlabeled or labeled images are missing. The validation has been conducted using four global regression models (GRM)-based and nine CNN-based CC techniques. In Liu et al. (2018c), a CC baseline network is derived from the VGG-16 network (Simonyan and Zisserman, 2014), which consists of 13 convolutional layers followed by three fully-connected layers (FCLs). The network is adapted to regress the person density maps by removing its two FCLs, and the max-pooling layer (pool5) to prevent further spatial resolution reduction. In Wang et al. (2018b), a lightweight DL combining density adaption network (DAN), low-density counter network (LCN), and a high-density counter network (HCN) is proposed. DTL is performed by fine-tuning the base model pretrained on the SD n the TD. The paper (Ahmed et al., 2021a) presents a DTL-based multiple-person surveillance system using top-view perspectives, providing extensive coverage of the scene or field of view. This approach is conducted in two stages: (i) detecting persons using YOLOv3 and (ii) tracking them using the DeepSORT algorithm. Moreover, a DTL approach is introduced by considering a PTM augmented with an extra trained layer using a top-view dataset.

3.1.3. Progressive learning

Progressive learning relies on using PTMs to perform a continuous learning process, where tasks can be sequentially learned with the possibility to utilize prior knowledge from previously learned tasks to ease the learning and execution of new ones. In Bilal et al. (2021), a spatiotemporal DTL-based HAR framework to recognize overlapping human behaviors in lengthy films is proposed. Various CNN models, i.e., Xception, VGG16, VGG19, ResNet152v2, ResNet-101v2, Inceptionv3, and DenseNet201 have been fine-tuned to learn the spatial connection at the frame level.

Table 5 portrays some of the relevant VSS studies based on fine-tuning.

3.2. Popular pretrained models (PTMs)

Using PTMs as the backbone for target tasks instead of learning models from scratch has become the consensus of the VSS research community. In section, we briefly present the popular PTMs and recent VSS frameworks built upon them, especially by adopting DTL, to highlight some of the critical contributions of PTMs in VSS applications.

3.2.1. Complex models

- ***ImageNet:*** a CNN model called Imagenet has been proposed in the ILSVRC, a large-scale object recognition challenge. Typically, it has been trained on the ImageNet repository (Russakovsky et al., 2015), including more than 15 million labeled images. By using DTL, the pretrained ImageNet network is leveraged and fine-tuned on curated datasets to perform different VSS tasks, such as fire detection (Bari et al., 2021), classification of crowd movements (Bendali-Braham et al., 2019), high-level semantic concept recognition (Su et al., 2014), etc.

Table 5
Summary of existing TL-based VSS frameworks using fine-tuning.

Work	Backbone	Description	Application	Best FMD performance	Limitation/advantage
Zhang et al. (2016)	MCNN	• Single-image CC via MCNN	UCF_CC_50, SHA	MAE=295.1, MSE=490.23	• Although the good generalizability achieved, the performance needs further improvement.
Sahoo et al. (2019)	VGG-16, InceptionV3	• Spatial and temporal feature extraction using pretrained CNN models	UCF crime (Sultani et al., 2018)	AUC=50.16%	• Low detection performance under the UCF crime dataset.
Atghaei et al. (2020)	VGG-16	• Freezing the first five layers and use the learned parameters	UCSD anomaly detection (Mahadevan et al., 2010)	AUC=93%	• Still cannot support real-time applications.
Liu et al. (2020c)	ResNet-50, ResNet-101, ResNeSt-50, ResNeSt-101	• Fine-tuning 2D-CNN models for better action recognition	CitySCENE (Anon, 2022)	AUC=89.2%	• Stronger temporal modules and deeper networks do not bring performance improvement.
Doshi and Yilmaz (2020)	YOLOv3	• AED based on Fine-tuning of MCNN architecture (whole MCNN)	CUHK Avenue UCSD SHA/SHB	AUC=97.8% (UCSD Ped 2)	• Moderate performance on CUHK Avenue and SHA/B datasets.
Ahmadi et al. (2020)	CNN	• Fine-tune a PTM by replacing the softmax and running backpropagation layers	Penn-Fudan (Wang et al., 2007) Daimler (Flohr et al., 2013) Inria person detection (Dalal and Triggs, 2005).	F1=91.0% (DPSB)	• Fail in extremely complicated situations, i.e., when a person is hidden behind another.
Yu et al. (2021)	ResNeXt-101	• Freeze the feature extractor and fine-tune on the detection head	Private dataset	mAP50=51.38%	• Improve novel-class performance on weak base datasets and competitive novel-class performance on strong base datasets.
Liu et al. (2020b)	DSSINet	• CC using regression-detection bi-knowledge transfer from a labeled SD to an unlabeled TD	SHA, UCF_CC_50, UCF_QNRF	MAE=112.24, MSE=218.18	• No limitation is reported.
Delussu et al. (2022)	CNN, GRM	• Scene-specific CC by transferring the knowledge learned on synthetic datasets to real data.	Mall UCSD PETS2009	MAE=4.6, RMSE=6.54	• The performance can significantly dropped due to low image illumination and color degradation.
Ahmed et al. (2021a)	Deep SORT and YOLOv3	• Top view multiple people tracking using fine-tuning of PTMs	COCO private data	Acc=96%	• No information about the generalization of the presented model to other existing datasets, as it is only validated on one dataset that is not available online.
Bilal et al. (2021)	Xception, VGG16, VGG19, ResNet152v2, ResNet-101v2, Inceptionv3 and DenseNet201	Fine-tuning based spatiotemporal HAR framework for long and overlapping action classes	UCF-101	Acc=96.03%	• Do not perform well to learn complex actions containing multiple sub-actions and or multi-view-points.
Liu et al. (2021b)	ResNet-18	• CC by exploiting sample correlation with multi-expert network and fine-tuning	SHA/SHB UCF-QNRF (Idrees et al., 2018) NWPU-Crowd	MAE=63.1, MSE=94.7	• The performance needs further improvement compared to the state-of-the-art.

- ***AlexNet:*** it is the CNN architecture that won the ILSVRC2012 (Krizhevsky et al., 2012) consisting of five convolutional layers combined with max-pooling (to reduce the dimensions of data) followed by three FCLs. The activation function is a rectified linear unit (ReLu) that presents a fast training advantage over other activation functions (Ramachandran et al., 2017; Silver et al., 2016). In Serpush and Rezaei (2020), an AlexNet-based DTL is used to develop a HAR framework. In doing so, a hybrid approach based on background subtraction and histogram of gradient (HOG) is first implemented, followed by applying AlexNet-based DTL and long short-term memory (LSTM) to select the best features. Moving on, human actions are labeled using a Softmax K-nearest neighbors (KNN) classifier. Similarly, the representational power allowed by the AlexNet-based DTL for a HAR task is demonstrated in Giel and Diaz (2015).
- ***VGGNet:*** VGG-16 was first proposed in ILSVRC2014 before introducing VGG-19, which both represent two models for improving AlexNet. In doing so, large kernel-sized filters of AlexNet were replaced with multiple small kernel-sized filters, which resulted in 13 and 16 convolution layers for VGG-16 and VGG-19 (Simonyan and Zisserman, 2014). In Sen and Deb (2021), an action classification scheme for soccer videos is introduced by combining VGG-19 and gated recurrent unit (GRU).
- ***CaffeNet:*** it is an improved version of AlexNet without using data augmentation and placing the pooling layer before normalization operation. Concretely, CaffeNet helped slightly reduce the computational cost of AlexNet, as a result of making data dimensionality reduction before the normalization process (Jia et al., 2014).
- ***ZFNet:*** this model won first place in ILSVRC2013 and has been built upon the architecture of AlexNet with a similar number of architecture and other improvements (Zeiler and Fergus, 2014). Typically, ZFNet introduces the idea of deconvolutional network (Zeiler et al., 2010) to address the black-box nature of CNN algorithms by illustrating their use to learn the feature representation. Accordingly, a deconvolutional network helps map characteristics learned into input pixel spaces, thus, enhancing CNN interpretability.
- ***Inception:*** Inception-V1, also named GoogleNet, is proposed to improve the performance of VGGNet regarding memory usage and runtime while maintaining a good accuracy (Szegedy et al., 2015). In doing so, the redundant or zero activation functions of VGGNet due to the correlations between them are eliminated. Thus, Inception-v1

has been augmented with Inception's module to approximate sparse connections between the activation functions. Following Inception-V1, three variants have been introduced to refine the architecture by (i) using batch normalization for training in Inception-V2 (Ioffe and Szegedy, 2015), (ii) using a factorization approach to reduce the computational cost of convolution layers in Inception-V3 (Szegedy et al., 2016), and (iv) introducing a simplified uniform variant of Inception-v3 with more inception modules in Inception V-4 (Szegedy et al., 2017). For instance, in Mathew et al. (2017), the knowledge learned by Inception v3 on 1,28 million images (categorized into 1000 classes) from the ImageNet LSVRC 2014 is utilized for intrusion detection on small ATM surveillance video dataset of 4719 images.

- ***ResNet:*** augmenting CNN models with more layers can result in vanishing gradients and accuracy saturation. Residual learning is the backbone of the ResNet architecture and is used to solve these issues (Kensert et al., 2019). Before ResNet, CNN architectures learned the characteristics at distinct abstraction levels (at the end of every convolution layer). By contrast, ResNet learns residuals instead of the features, representing the subtraction of characteristics learned from the input of every layer. This has been made using the identity shortcut connections concept (i.e., connecting the input of a layer to x layers after that) (He et al., 2016). Following, different improved versions of ResNet have been developed using distinct numbers of layers, e.g., ResNet-34, ResNet-50, and ResNet-101.
- ***Inception-residual network (IRN):*** it relies on combining the strengths of ResNet and Inception. Typically, ResNet helps the model in having deeper CNN for learning more complex characteristics while maintaining good performance, whereas Inception helps in efficiently learning the characteristics at distinct resolutions within the same convolution layer. Thus, the IRN combines these advantages in two versions (i) Inception-ResNet-V1, which is based on Inception-V3, and (ii) Inception-ResNet-V2 which is based on Inception-V4 (Szegedy et al., 2017). In Khan et al. (2022), abnormal events related to smoking in public areas are detected by transferring the knowledge of the pretrained InceptionResNet-V2 model. In Suresh and Visumathi (2020), a HAR scheme is performed on a small dataset using Inception-ResNet-v2-based DTL and LSTM. Concretely, the model is first trained to be deriving features from Inception-ResNet-v2 prior to applying the output features on the LSTM to learn action sequences.
- ***Xception:*** this model refers to extreme Inception and represents an improved variant of Inception-V3 (Chollet, 2017). It utilizes depthwise separable convolution to separately entail images' spatial dimensions and channel dimensions in the training stage. Moreover, it has better performance than Inception-v3 on ImageNet, although they have almost the same number of parameters. In Wilie et al. (2018), the knowledge of the pretrained Xception model is adopted to develop a CC system named CountNet. Typically, the pretrained Xception is fine-tuned, where only the FCLs have been trained again. This results in better performance, especially when an augmented dataset robust to slice and scale variations is used.
- ***DenseNet:*** in this model, every convolution layer obtains the feature maps (i.e., output) of all preceding layers as input and transfers its feature maps (i.e., output) to all subsequent layers (Huang et al., 2017). Accordingly, every layer receives the combined knowledge of all preceding layers, making the resulting CNN more compact and thinner because of the decreasing number of feature maps. Different versions of Densenet have been proposed, including DenseNet-121, DeneNet-169, and DenseNet-201.
- ***Hybrid models:*** it is worth mentioning that some TL-based VSSs have been developed by combining different CNN-based DTL architectures. For example, Huang et al. (2020) propose a hybrid CNN-based DTL system to detect the distracted behavior of drivers using a cooperative PTM that combines ResNet-50, Inception-v3, and Xception. In Leong et al. (2020), a fusion of the pretrained 2D CNN models, namely VGG-16, ResNets, and DenseNets, is proposed to develop an effective video action recognition system. Typically, the DTL methodology based on these models is used to extract spatial features. Then, a temporal encoding is performed before connecting the output to 3D convolution layers at the top of the architecture. In Abdulazeem et al. (2021), TL-based HAR is adopted by pretraining a standard CNN model on a generic dataset to adjust weights prior to applying it to a TD dataset. Five different CNN architectures and LSTM are considered in the recognition phase. The first three architectures are single-stream and stand-alone, while the last two models combine the first three networks. In Khan et al. (2021), a TL-based HAR is introduced, in which the pretrained DensNet 201 and Inception-v3 are used to map relevant features. Moving on, the serial-based extended (SbE) scheme is used to fuse extracted features. Next, the kurtosis-controlled weighted KNN is employed to select the pertinent features. Lastly, different supervised ML models are used to classify the selected features and perform HAR on different datasets, including KTH (Chen et al., 2021b), IXMAS (Melhart et al., 2021), WVU (Hassan et al., 2018), and Hollywood (Joshi et al., 2020).

3.2.2. Lightweight models

To overcome the high computation cost problem encountered with complex CNN models and enable their implementation on mobile terminals, numerous lightweight but efficient CNN architectures have been proposed.

- ***Mobilenetv1:*** it relies on using a depthwise convolution for performing lightweight filtering by applying a single convolutional filter per input channel. Moreover, it does not use pooling layers, while a depthwise separable convolution with a step size of 2 is utilized for downsampling operations.
- ***Mobilenetv2:*** in addition to inheriting the depthwise separable convolution of MobileNetVl, the width factor and resolution factor are used for compressing the model scale. Moreover, the residual unit used in ResNet is adopted in this architecture, where two improvements are performed by introducing an inverse residual structure and proposing a linear bottleneck structure. In Khaire and Kumar (2022), the power of DTL is leveraged to extract relevant video features using MobileNetv2 and develop an efficient real-time AED framework by reducing computational complexity. This helps implement this approach on the edge and mobile devices at the ATM surveillance sites.
- ***SqueezeNet:*** it has almost the same accuracy as AlexNet but with 50 times fewer parameters and relies on the idea of stacking to construct the network. It reduces the feature map with a 1×1 convolution kernel to decrease network parameters. The concept of stacking is used when creating the network (Iandola et al., 2016).
- ***ShuffleNet:*** it relies on solving the limitations of information flow between channels caused by group convolution through shuffle operations (Zhang et al., 2018).
- ***MixNet:*** although depthwise separable convolutions have vastly been utilized in many lightweight CNNs, limited attention was devoted to investigating the size of the convolution kernels. In contrast, MixNet (Tan and Le, 2019) is proposed to consistently study the influence of the convolution kernel size on the final result based on MobileNet.
- ***EfficientNet:*** because the generalizability of a CNN model can be improved by increasing the depth, width, and resolution of the network, EfficientNet introduces a compound coefficient for proactively improving these parameters and optimizing them via a composite model scaling.

4. Deep domain adaptation

Despite the simplicity of implementing fine-tuning, its efficiency significantly drops when the distributions of SD and TD are different. To that end, domain-distance measurement (DDM) has been considered in

the original networks, which is called DDA. Typically, the cost function of the initial model is adjusted by including a domain loss to quantify the distribution of the SD and the TD. Fig. 5(b) displays a CNN example of DDA that adjusts the distribution in FCLs through DDM. Most VSS frameworks have separately been validated on single-domain datasets with similar characteristics (Wang et al., 2021d; Saponara et al., 2021; Che Aminudin and Suandi, 2022; Lamas et al., 2022) while a small effort has been dedicated to exploring cross-domain VSS.

4.1. One-step DDA

4.1.1. Homogeneous DDA

(a) Discrepancy-based: DDA and DTL aim at performing a VSS task in a TD based on the knowledge learned from an SD. To transfer knowledge, the SD and TD are aligned by reducing the MMD, including DDC (Tzeng et al., 2014), JAN (Long et al., 2017), DAN (Long et al., 2015) and RTN (Hinton et al., 2015). To enhance the model generalization of TD dataset with unlabeled crowd scenes, a set of diverse and decorrelated regressors are learned to prevent overfitting in the SD. While in Xu et al. (2019), a learn-to-scale module is introduced to address the density pattern shift, which maintains good transferability across datasets. Moving on, in Shi et al. (2018), Xu et al. (2019), the crowd information in the TD dataset, which is unlabeled, is not exploited, which has limited the performance of cross-domain CC. Unlike the abovementioned studies, the mutual transformations between the output of individual detection models and density regression in the SD are modeled based on a deep structured scale integration network (DSSINet) in Liu et al. (2020b), which uses VGG-16 as a backbone. Next, the regression-detection bi-knowledge is propagated over modeled transformers to the TD using a self-supervised learning approach. This process is repeated until the performance is converged in the TD. In Prabono et al. (2021a), the discrepancy between the SD and TD datasets is minimized by reducing statistical distance, which helps implement a DDA scheme based on autoencoder for HAR. Rather than simultaneously learning the representation of SD and TD, this method attempts to learn the representation for the domain of interest separately to guarantee its optimality. In Guo et al. (2021), a maximum cross-domain classifier discrepancy (MCDCD) technique is introduced to perform a multi-source unsupervised deep domain adaptation (UDDA) for abnormal human gait detection. In this respect, the information from multiple training SDs is leveraged to enhance the classification performance on the TD.

In some applications, such as cross-domain facial recognition, it is challenging to simultaneously manage the domain shift and the semantic gap during the DDA. Most existing techniques can only reduce domain discrepancy for transferable characteristics but fail to decrease the semantic one. To close this gap, a Joint Discriminative and Mutual Adaptation Networks (JDMAN) is introduced in Li et al. (2021a), which helps in collaboratively bridging the domain shift and semantic gap by domain- and category-level co-adaptation based on mutual information and discriminative metric learning techniques. In Qi et al. (2019), a camera-aware DDA for person re-identification is developed to diminish the discrepancy between the SD and TD as well as across the camera-level sub-domains. Moreover, the temporal continuity in every camera of the TD is exploited for creating discriminative information. It has been executed by proactively producing online triplets within every batch to fully capitalize on the steadily enhanced feature representation in the training stage.

(b) Adversarial adaptation: the second class of methods, such as the adversarial discriminative domain adaptation (ADDA) (Tzeng et al., 2017), multi-adversarial domain adaptation (MADA) (Pei et al., 2018) and conditional adversarial domain adaptation (CDAN) (Long et al., 2018), unsupervised image-to-image translation (UNIT) (Liu et al., 2017), Cycle-Gans (Zhu et al., 2017), CoGAN (Taigman et al., 2016) and Disco-GAN (Kim et al., 2017). In fact, GANs play a major role in the generative modeling of images, although their operation is restricted by different assumptions, which question the efficiency of DA. They attempt to apply adversarial learning to the training of the network. The differences between domains are gradually eliminated since the feature extractor keeps trying to confuse the domain classifier. For instance, In Wang et al. (2019c), a data collection and labeling method is developed for generating synthetic crowd scenes and concurrently annotating without human intervention. Following, the produced synthetic data is then used to improve the CC accuracy in the wild by (i) using an SSIM embedding (SE) Cycle GAN (SE-Cycle-GAN) to transform the synthetic video frames into the photo-realistic frames, (ii) training a spatial fully convolutional network (SFCN) on the translated data. Moving forward, Gao et al. (2019) introduce a domain-adaptation CC (DACC) scheme, which relies on inter-domain features segregation (IFS). Specifically, it consists of transforming synthetic images into realistic data and density map reconstruction. Typically, it prompts the translation quality by segregating domain-shared/independent features and designing content-aware consistency loss. Thus, pseudo labels are generated on real scenes and retrain a final counter, improving the prediction quality.

(c) Reconstruction-based: this scheme utilizes an auxiliary reconstruction process for creating a shared representation between the SD and TD. For example, the Deep Reconstruction Classification Network (DRCN) aims to solve the following two tasks concurrently: (i) classifying SD samples and (ii) reconstructing unlabeled TD samples. This results in helping the model learn to correctly discriminate between the SD and TD samples in addition to preserving information about the TD. For the case of autoencoders, encoder/decoder reconstruction-based DDA techniques aim at concretely learning the domain-invariant representation by a shared encoder and maintaining the domain-special representation by a reconstruction loss in the SD and TD. In this regard, aiming at learning cross-domain shared contents by suppressing domain-specific variations, Deng et al. (2021b) propose the Deep Ladder Reconstruction-Classification Network (DLaReC) approach. The latter utilizes an encoder with cross-domain sharing and a TD reconstruction decoder. Residual shortcuts connect the encoder and decoder at every intermediate layer. In this regard, the domain-specific components are directly fed to the decoder for reconstruction, which helps alleviate the pressure of learning domain-specific variations at later layers of the shared encoder.

To utilize data from both the SD and TD, a large-scale synthetic dataset is established in Wang et al. (2019c) as the source, and a UDDA is considered for reducing the discrepancy between the synthetic SD and the real-world TD using the Cycle-Gan (Zhu et al., 2017). By contrast, domain features in the semantic space have been selected to be aligned using adversarial learning in Han et al. (2020). Whereas in Zhu et al. (2017), Han et al. (2020) density regression networks for DDA are utilized where the TD data is taken as a whole with a domain label.

(d) Hybrid approaches there are also application scenarios where the methods above are simultaneously utilized to obtain better performance. For instance, the authors in Hoffman et al. (2017) combine a soft label loss and domain confusion loss. In contrast, Long et al. (2016) uses both architecture criteria (adapt classifier by residual function) and statistic (MMD) for UDDA. Moving on, class-specific auxiliary weights are assigned by the pseudo-labels into the initial MMD in Yan et al. (2017) In Wei et al. (2018a), different domains are linked up using common characteristics, and domain divergences are simultaneously reduced by learning the translations between common characteristics and domain-specific characteristics. Following, learned translations are cross-used for transferring domain-specific characteristics of one domain to another before composing a homogeneous space, where domain divergences can be reduced. Most existing adversarial learning techniques have focused on resorting to learning domain-transferable feature representations by bounding the feature distribution discrepancy cross-domain. However, this can result in poor generalization and

misalignment performance without capitalizing on task-special adaptation and class information. To overcome these issues, Zhang et al. (2020b) propose joint adversarial learning with a domain alignment DNN architecture and class information, namely the hybrid adversarial network (HAN). The letter relies on (i) incorporating a classification loss for learning a discriminative classifier and (ii) adopting a domain adversarial network for learning a domain-transferable representation that diminishes domain discrepancy. Moving forward, a DNN-based HAN for end-to-end CDHAR, namely HydraNet, is proposed in Prabono et al. (2021b). It is built upon learning reliable domain-invariant latent representations of common characteristics by decreasing statistical shifts between domains.

4.1.2. Heterogeneous DDA

By using homogeneous feature spaces, existing techniques built upon homogeneous DDA are less practical in real-world scenarios as there is more chance of having heterogeneous feature spaces. Although efforts are paid to develop heterogeneous DDA schemes, additional information, such as instance correspondence, is still required. This is challenging to satisfy when sensor data is processed (Prabono et al., 2021b).

Discrepancy-based approaches: in discrepancy-based homogeneous DDA, a model usually reuses or shares the first layers between the SD and TD, where their features spaces have the same dimension. By contrast, the dimensions of the feature spaces of the SD can vary from those of the TD. When addressing the first group of heterogeneous DDA, the video frames in different domains could be promptly rescaled into the same dimensions. Thus, the statistic and class criteria are still efficient and primarily utilized. For instance, by considering an RGB video frame and its paired depth frame in Gupta et al. (2016), mid-level representations learned by a CNN model are used as supervisory signals for re-training a CNN on depth images.

Adversarial-based approaches: by adopting generative models, heterogeneous target data can be generated while transferring some information from the SD to it. Recently, there have been several works on video DDA. For instance, Jamal et al. (2018) utilize an adversarial learning framework with 3D CNN to align the SD and TD. TA3N (Chen et al., 2019) leverages a multi-level adversarial framework with temporal relation and attention mechanism to align the temporal dynamics of feature space for videos. TCoN (Pan et al., 2020) matches the feature distributions between source and TDs, for temporal alignment using the cross-domain co-attention mechanism.

Reconstruction-based approach: heterogeneous DDA can also employ adversarial reconstruction as it is presented in Zhu et al. (2017), Kim et al. (2017) and Yi et al. (2017), where the Cycle-Gan, dual GAN, and disco GAN are implemented, respectively. They use two generators, GA and GB, to generate sketches from images and images from sketches, respectively.

Hybrid approaches: hybrid DDA is a particular scenario of heterogeneous DDA in which common features between domains exist. Moreover, hybrid DDA can be more realistic when it is easier to satisfy the feature commonality. Whereas existing techniques operate using common features in the original feature space, which indeed can still have distribution differences. Additionally, the existing ones require extracting hand-crafted features to perform more informative descriptions while classifying video frames.

A specific scenario of heterogeneous DDA is the Hybrid DDA (Wei et al., 2018a), which is considered an emerging framework on the DA. It leverages the intersections of the feature space between the SD and TD. This case can be comparatively more applicable, notably for HAR, as various wearable devices can have similar sensing modalities and their device-specific sensing modalities. The earliest attempts to solve hybrid DDA problems have focused on taking similar features from both domains and discarding the domain-specific features. In Prabono et al. (2021b), a deep model for hybrid DDA is introduced, which systematically extracts high-level pertinent characteristics from raw data and learns the domain-invariant latent representations of the common characteristics between domains. Typically, latent representations of the common characteristics are eventually utilized as the bridge to transfer relevant characteristics between domains.

4.2. Multi-step DDA

4.2.1. Hand-crafted approaches

Since the intermediate domain is occasionally chosen based on experience, it is decided in advance. For instance, if the SD consists of synthetic video frames and the TD contains real video sequences, some annotated synthetic frames will evidently be explored as intermediate domain data. In this respect, Kim et al. (2021) use visual DDA and image segmentation to measure water elevation from side-view and top-view video data. Concretely, the SD consists of multi-view synthetic data, and the DDA is then applied for estimating water levels from top-view videos to enable the generalization of this strategy and its application to other data repositories. Besides, a simple to complex action TL approach, namely SCA-TLM, is proposed in Liu et al. (2015) for complex HAR using dense trajectories to extract features. By handling the abundant labeled simple actions, SCA-TLM helps in (i) improving the performance of complex HAR and (ii) optimizing the weight parameters. This enables the learning of complex actions to be reconstructed by simple actions. Yan et al. (2018) develop an online TL for 3D LiDAR-based person classification based on multisensor-based tracking. In doing so, a Bayesian tracker and an SVM classifier are used to develop a human classifier, which could be learned from the deployment environment without relying on the training annotated data.

4.2.2. Instance-based approaches

There are other problems where various intermediate domains are candidates. Thus, an automatic selection process might be adopted. As presented in Pan and Yang (2009) and similar to instance-transfer techniques, since the observations of the SD cannot directly be utilized, a combination of samples from both the SD and TD can be effective in constructing the intermediate domain.

4.2.3. Multi-source DDA (MDDA)

MDDA aims to shift the knowledge learned from different SDs to an unlabeled TD. This makes it a challenge, keeping in mind the acute domain discrepancy that exists not only between the SD and TD but also between the also exists among diverse sources. Prior studies on MMDA either estimate a mixed distribution of SDs or combine multiple single-source models, but only some delve into the relevant information among diverse SDs. Recently, some MDDA approaches have been introduced (Hoffman et al., 2018; Zhao et al., 2020; Sun et al., 2015). It has been proved in many UDDA studies with multiple annotated SDs that MDDA is better than single-source deep domain adaptation (SDDA) (Yang et al., 2020; Zhao et al., 2019; Lin et al., 2020; Guo et al., 2020). Existing MDDA techniques are categorized into two groups, where the first one utilizes shared feature spaces (Peng et al., 2019; Xu et al., 2018; Ren et al., 2022; Ahmed et al., 2021d; Nguyen et al., 2021) to bridge the distributions of multiple SDs and TD through the alignment of all domains using a shared network. However, domain-specific knowledge is not completely investigated. Another group relies on multi-model combinations (Zhu et al., 2019; Zhao et al., 2020), which pair the TD with each SD for separately training various classifiers. Moving forward, multiple predictions generated by different classifiers for the same target sample in the test stage are combined to obtain the final prediction, Techniques of this group cannot be directly utilized in most VSS applications because of the distinct sizes and numbers of regions produced by different detection algorithms on every target sample. This makes obtaining the final prediction by weighting these regions in the test stage challenging, even impossible.

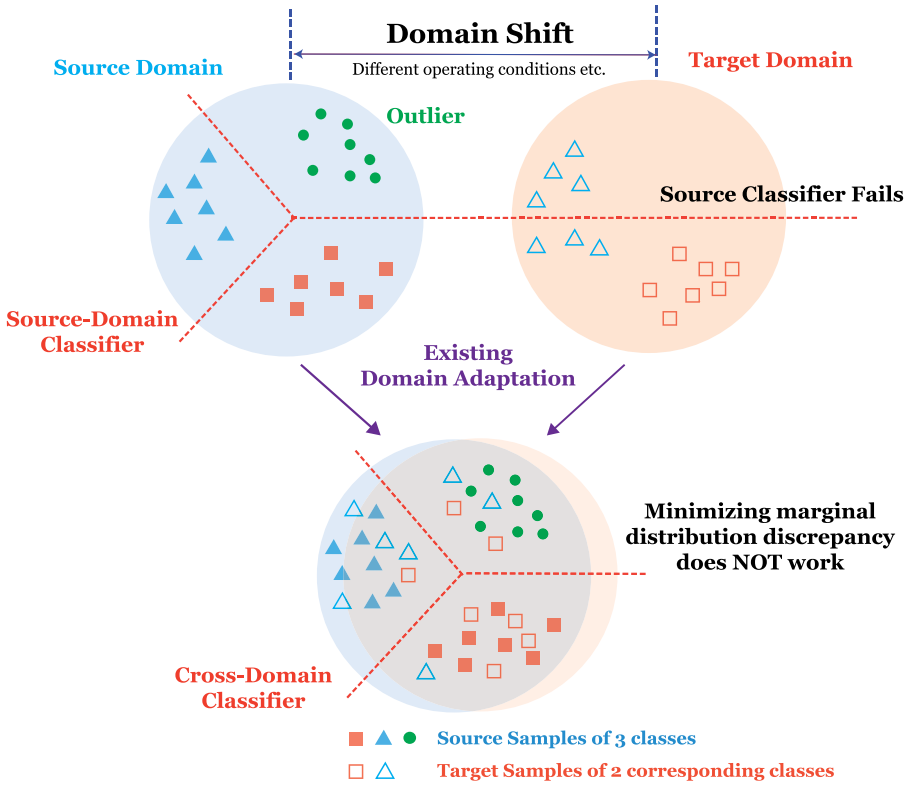


Fig. 9. The partial transfer learning problem (Li et al., 2020).

4.2.4. Representation-based approaches

These techniques freeze a previously trained model and utilize its intermediate representations as input to a new model. A progressive neural network (NN) that can accumulate and transfer knowledge to new domains over a sequence of experiences is proposed in Rusu et al. (2016). Typically, a new NN is constructed for each domain to avoid the target model losing its capability of solving the SD. Besides, the transfer is performed by later connections to previously learned network features. Moreover, to remember the knowledge of intermediate domains, the parameters in the latest network are frozen.

4.3. Partial transfer learning

While most DTL algorithms minimizing the marginal distributions discrepancy have generally relied on assuming identical label spaces across different domains, a partial transfer learning process can be more common for some scenarios. Typically, a discrepancy between the SD and TD label spaces can exist, or the TD label space can be a subspace of the SD label space. In this case, partial transfer learning has emerged as a complex problem since it is challenging to know where to transfer due to the non-availability of shared label spaces. Fig. 9 explain the partial transfer learning problem.

Recently, various studies have been proposed to target partial transfer learning. For instance, Cao et al. (2018) introduced partial transfer learning based on selective adversarial networks (SAN) to circumvent negative transfer concurrently. Specifically, this has helped relax the shared label space assumption to make the TD label space a subspace of the SD label space. In doing so, the outlier source classes were ignored, and the positive transfer was promoted by maximally matching the data distributions in the shared label space. Similarly, in Li et al. (2020), Li et al. address the partial transfer learning issue by adopting a DL-based DA approach. Typically, a class-weighted adversarial neural network has been developed to discard SD outliers, encouraging the positive transfer of the shared classes. Moving forward, a double-layer attention-based GAN is designed in Deng et al. (2021a) to tackle the partial transfer problem. TIn this regard, a transfer that constructs two attention matrices for domains and samples is proposed, where the matrices can guide the model to understand which data parts to ignore or concentrate on before performing domain adaptation. In the same way, a GAN-based architecture, namely a deep partial transfer learning network, is proposed in Yang et al. (2021). Notably, a domain discriminator has been adopted to learn domain-asymmetry factors automatically. The SD data is weighted to block irrelevant knowledge in the maximum mean discrepancy-based distribution adaptation.

4.4. Open-set transfer learning

While conventional DL and ML models aim to train classifiers in the closed-set world, in which the same label space is shared between SD and TD samples, open-set learning refers to the case of having test samples from the unseen classes during training. The authors in Fang et al. (2021) presented the first bold attempt to investigate open-set transfer learning by exploring its generalization error-given training samples with size n. This work has provided a generalization bound for open-set transfer learning, which has theoretically been made by investigating the risk of the TD classifier on unknown classes.

Moreover, some progress has been made in open-set transfer learning, where the major issue of recognizing unknown classes has been addressed. Typically, various techniques have been introduced, e.g., the extreme value theory (Rudd et al., 2017; Perera and Patel, 2019) and open-space risk and Gunther et al. (2017), Geng et al. (2020). Additionally, some studies have focused on adapting DL models to support open-set transfer learning. For example, Han et al. (2021b) propose open-set crowdsourcing using multiple-source transfer learning, while open-set face recognition with DTL and extreme value statistics is proposed in Xie et al. (2018).

On another side, although the studies proposed to target open-set and partial domain adaptations, no prior data on the TD can be found in some real-world applications, such as anomaly and fault detection. This represents another challenging problem in TL. To that end, Zhang et al.

(2021b) introduce a universal DA method for fault diagnosis, which does not use any assumption on the TD label set. In doing so, a hybrid scheme with source and target instance-wise weighting mechanism is developed for selective adaptation. In addition to what has been done in Zhang et al. (2021b), the authors in Zhang et al. (2021c) propose a selective adaptation by using an additional outlier identifier. Typically, unknown fault models can automatically be identified while enabling class-level alignments for the shared health states without prior information about the TD label set.

5. Applications of DTL in VSSs

5.1. Cross-domain object detection (CDOD)

In contrast to other video surveillance tasks, CDOD is significantly challenging because object location and category need to be predicted. State-of-the-art OD frameworks leverage either cross-spectrum or one single spectrum (thermal or visible). For the thermal spectrum, Kieu et al. (2020) introduce a task-conditioned DDA between daytime and nighttime. Concretely, the principal detection task has been augmented with an auxiliary classification stage distinguishing between nighttime and daytime thermal images. Moving on, the classification stage has been utilized for conditioning a YOLOv3 to enhance its adaptation to the thermal domain.

Chen et al. (2018) propose one of the first attempts to apply a UDDA for CDOD, namely DDA Faster-RCNN (DDA-Faster). In this regard, to decrease the domain discrepancy, instance-level and image-level adaptation components have been introduced. Moreover, to learn a domain invariant RPN of the Faster R-CNN model, a consistency regularization has been deployed as well. The suggested scheme in Ahmadi et al. (2020) employed a PTM named YOLOv2, which is an object detector based on CNN. The TL-based fine-tuning is then used to overcome the problems faced in traditional CNN deep networks, such as dealing with different sizes, high definition, or colored images, turning any suggested AI model slower and less precise in real-time applications. In Arruda et al. (2022), a UDDA problem is addressed by detecting objects across different domains. Typically, a two-stage technique is introduced, which (i) trains an unsupervised image-to-image translation algorithm for generating a synthetic dataset that is similar to the TD (fake-data), and (ii) trains object detectors based on Cycle-GAN and adaptative instance normalization (AdaIN) with the new artificial data. Unlike most CDOD techniques requiring labeled datasets for both thermal and visible domains, a UDDA is performed in Marnissi et al. (2022) without requiring thermal data annotation, and it was validated on the KAIST dataset (Hwang et al., 2015). It is worth mentioning that only a few studies addressing the UDDA for CDOD have been investigated in the literature (Chen et al., 2020b, 2018; Saito et al., 2019). For instance, Fig. 10 portrays the flowchart of the CDOD scheme proposed in Saito et al. (2019). It carries out weak-global and strong-local alignments using global domain classifier and local domain classifier networks, respectively. Following, the domain classifiers extract a context vector, which is concatenated in the layer before the final FCL.

Other frameworks focus on conducting adaptation from thermal to visible domains. More specifically, some aim to use an input thermal image to generate a perceptually realistic RGB image. This is called colorization, which is generally approached using generative networks (Devaguptapu et al., 2019; Berg et al., 2018; Kuang et al., 2020). For better detection, other works perform this transformation as explained in Devaguptapu et al. (2019). Typically, an improved multimodal Faster-RCNN is introduced along with a Cycle-GAN for unpaired image-to-image translation of thermal to pseudo-RGB data. In Munir et al. (2020), a DDA approach built upon the style consistency has been utilized for transferring low-level features from the visible to the infrared domains. The OD in the infrared spectrum is conducted using the cross-domain network with style consistency. Compared to the study in Kim et al. (2019), a unified detection model that defines a common feature space is proposed, which enables making intermediate features from the two domains for cross-spectral pedestrian detection (CSPD). The miss rate (MR) proposed in Hwang et al. (2015) has been used to evaluate the performance of this technique.

For adapting visible domains to thermal domains, the work in Guo et al. (2019) generates synthetic thermal images from visible images. This transformation operates as a data augmentation task to train a pedestrian detector and run them on thermal images. Additionally, Xu et al. (2017) propose a cross-modality learning technique that relies on a multi-scale detection network (MSDN) and region reconstruction network (RRN). The latter aims at transferring the non-linear mapping knowledge from the RGB channels to the thermal channel to improve detection performance from visible data. The work in Fuhl et al. (2018) presents a multiple annotation maturation (MAM), a revolutionary self-training approach for fully automated labeling of vast volumes of picture data fed from a previously trained detector. MAM creates detectors that may then be utilized online. Because of their close association with the objects, shadows are sometimes misclassified as foreground or part of it. The authors in Walambe et al. (2021) use an ensemble DTL for multiscale OD from drone images. Specifically, three pre-trained object detectors, including RetinaNet, SSD and YOLOv3, are used to detect object in UAV images from the VisDrone dataset (Zhu et al., 2020).

The cross-domain pedestrian detection (CDPD) is a part of CDOD that attracts significant attention. It is built upon the assumption that the training and test video frames are drawn from different data distributions. Existing frameworks aim to align the descriptions of whole candidate instances between the SD and TD. Due to a substantial visual difference between the candidate instances, the inter-instance difference cannot be overcome through the alignment of the whole candidate instances between two domains. In this respect, Jiao et al. (2021a) assume the separate alignment of every type of instance can be more efficient. Thus, a selective alignment network for CDPD is introduced, consisting of developing (i) a base detector, (ii) an image-level adaptation network, and (iii) an instance-level adaptation network (ILAN). Fig. 11 presents the flowchart of this method. Specifically, $\mathcal{F}$ represents the feature module employed for extracting the feature map of a given image, and the RPN is utilized for generating various candidate proposals. Additionally, the detection module is considered for predicting the location of pedestrians and their corresponding labels. The extracted feature representation is then fed into the image-level adaptation network (ImLAN), which includes a domain classifier D for domain alignment. First, the candidate proposal are grouped by the ILAN with a "Group" before applying the instance-level DA on the corresponding groups in the SD and TD using two classifiers $\mathcal{F}_1$ and $\mathcal{F}_2$ of alignment module. Specifically, the parameters are shared between the modules with the same color.

Table 6 provides a summary of relevant CDOD frameworks discussed in this paper. Their characteristics have also been described in terms of the adopted backbones, methodology, dataset, best performance, and limitation/advantage. Obviously, based on the summarized CDOD studies, the Faster RCNN is the most used backbone then comes the Cycle-GAN. This is mainly due to their computational efficiency compared t other backbones.

5.2. Cross-domain anomaly detection (CDAD)

Numerous works have explored DTL and DDA for abnormal event detection. For instance, anomaly detectors for TDs are inferred without re-training using the latent domain vectors concept. By contrast, learned image representations across different image datasets had been reused in Andrews et al. (2016b). Moving on, a robust one-class DTL approach is designed in Xiao et al. (2015). The anomaly detectors developed in these methods necessitate labeled target instances. By contrast, in Fan et al. (2021), this issue is overcome by transferring

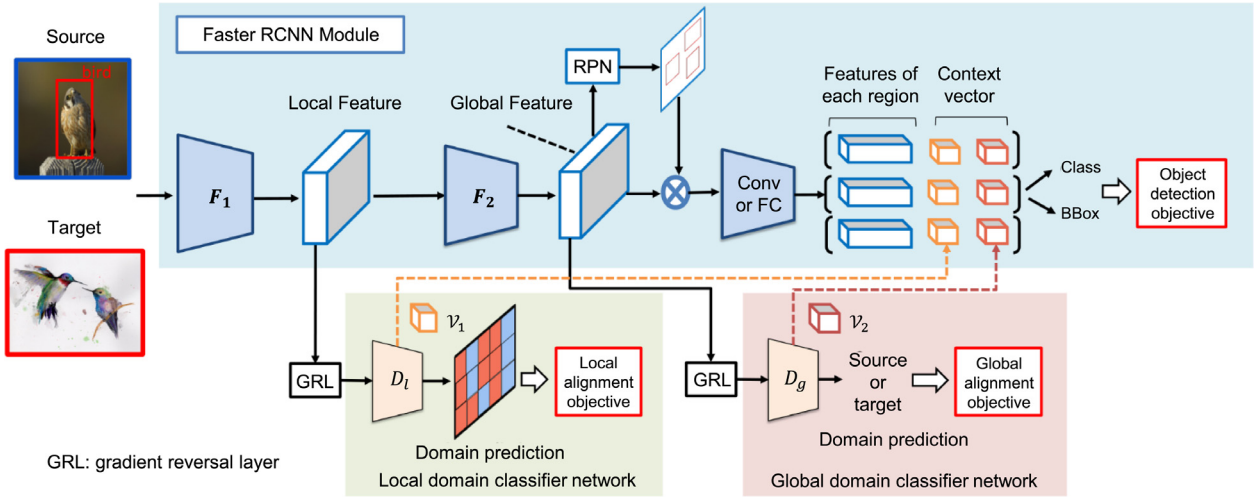


Fig. 10. Flowchart of the CDOD framework proposed in Saito et al. (2019).

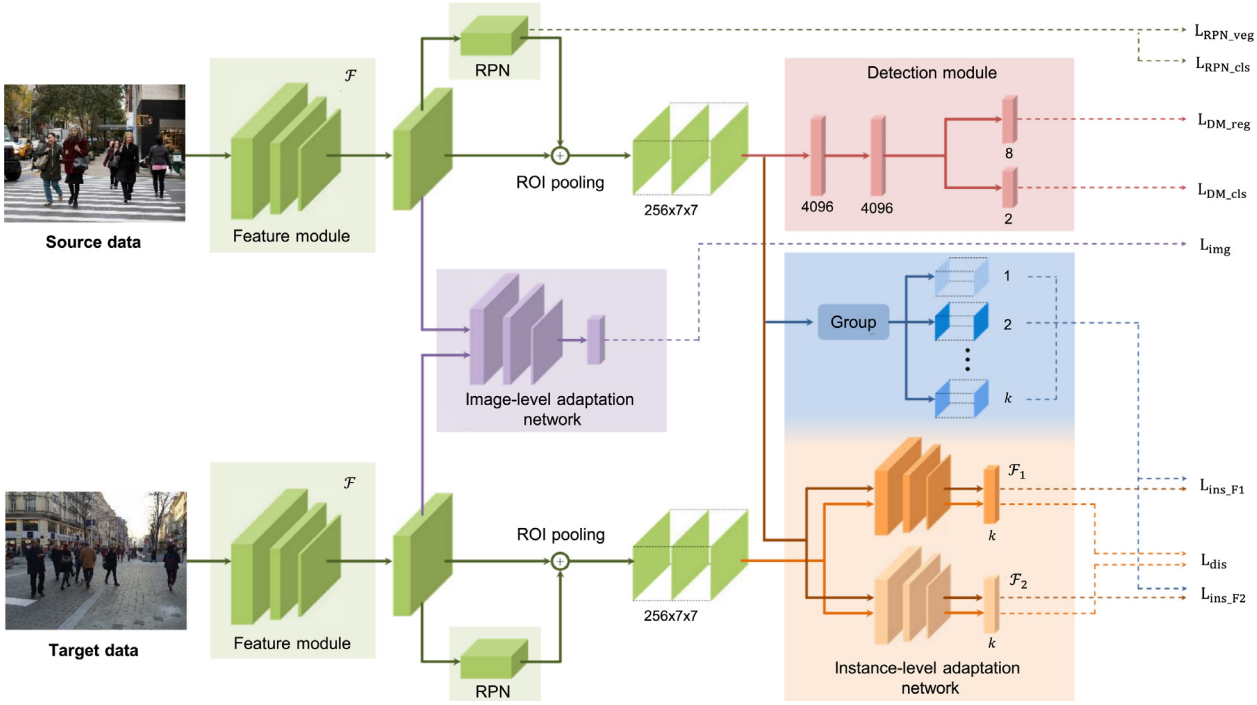


Fig. 11. Flowchart of the CDPD framework proposed in Jiao et al. (2021a).

the anomaly detection knowledge in a supervised manner using importance weighted adversarial autoencoder-based approach. Moreover, a localized instance-transfer algorithm (LocIT) (Vincent et al., 2020) selects labeled source instances to transfer by a local distribution-based approach and constructs a KNN classifier based on these chosen source instances and unlabeled target instances. Although LocIT can handle the situation where SD only contains normal instances, this method degenerates into a KNN-based unsupervised anomaly detection method without knowledge transfer.

Regarding the feature extraction in Sahoo et al. (2019), a two-stream two-dimensional CNN is used. The PTM parameters of VGG-16 and InceptionV3 are fine-tuned via TL. The spatial properties are learned in one stream, while the temporal features are learned in the other. Both learned spatial and temporal characteristics are enhanced to build a robust feature representation. The research in Atghaei et al. (2020) is based on DL approaches and shows how to use spatio-temporal data to recognize and pinpoint abnormal moments in videos. The proposed method employs generative adversarial networks (GANs) and DTL on a pretrained CNN to produce a precise and efficient model. Processing the video's optical-flow information improves the model's efficiency even further. The research in ZhanLi and JiaWei (2019) proposes a DTL recognition model based on the Inception-V3 NN for detecting anomalous behavior in picture samples. There are two significant parts to the procedure: The first step is extracting characteristics of picture samples using the Inception-V3 NN. The second step aims to categorize the acquired characteristics, and an unusual behavior recognition model is created.

The authors in Liu et al. (2020c) fine-tune two types of action recognition models using 2D-CNN- and 3D-CNN, where the latter has better anomaly detection accuracy due to learning fewer parameters. In Bansod and Nandedkar (2019), the spatial level appearance characteristics for abnormal and normal patterns are learned using a pre-trained VGG-16. To detect abnormalities, two methods are investigated. A homogeneous DTL scheme, where a PTM is utilized to fine-tune

Table 6
Summary of existing CDOD frameworks.

Work	Backbone	Description	Dataset	Best performance	Advantage/limitation
Chen et al. (2018)	Faster RCNN	• DDA Faster R-CNN for OD in the Wild	Cityscapes (Cordts et al., 2016), KITTI (Geiger et al., 2012), SIM10K	AP=38.97%	• The computational complexity can be reduced and the performance needs more improvement.
Arruda et al. (2022)	Cycle-Gan, AdaIN	• CDOD using unsupervised image translation	Cityscapes KITTI Foggy Cityscapes (Sakaridis et al., 2018)	mAP=36.8%	• Benefit from generating fake-data even when the qualitative results seem inaccurate, however, the semantic consistency of the translation needs improvement.
Marnissi et al. (2022)	Faster RCNN	• Unsupervised thermal-to-visible DDA for pedestrian detection	KAIST	MR=40.01%	• Feature distribution alignments of Faster R-CNN can be replaced by other deep detectors.
Saito et al. (2019)	Faster RCNN	• Unsupervised CDOB by transferring knowledge from label-rich to label-poor domains	PASCAL VOC Clipart (Inoue et al., 2018) Watercolor (Inoue et al., 2018) Cityscapes FoggyCityscapes (Sakaridis et al., 2018) GTA (Johnson-Roberson et al., 2016)	AP=53.1%	• The performance needs further improvement.
Munir et al. (2020)	MSDN	• CDPD to learn cross-modal deep representations	KAIST	MR=49.55%	• Enable knowledge transfer from multispectral data and accurate detection although under challenging illumination conditions.
Kim et al. (2019)	RetinaNet-C, RetinaNet-T, CMT-CNN , CMT-CNN-SA	• Unpaired CSPD using adversarial feature learning	KAIST	MR=41.51%	• Address the challenging illumination conditions for pedestrian detection (especially at nighttime).
Guo et al. (2019)	Cycle-Gan	• CDPD in thermal images	KAIST	Log-average MR=42.65	• Reduce the log-average MR by up to 12%
Doersch and Zisserman (2019)	DANN	• 3D human pose estimation using knowledge transfer from simulation to reality	SURREAL (Varol et al., 2017)	PA-MPJPE=88.9%	• High computational complexity.
Soviany et al. (2021)	Cycle-GAN	• CDOD using urCriculum self-paced learning	Sim10k, Cityscapes, KITTI, PASCAL VOC 2007, Clipart1k	mAP=27.64%	• Simple and effective without overheads during inference but still there is significant performance gaps compared to Faster RCNN algorithms.
Zhang et al. (2021a)	Faster RCNN	• CDOD using local–global attentive adaptation	PASCAL VOC 2010 Clipart1K Watercolor2K	mAP=43.8%	• Achieve state-of-the-art performance with small and large distribution shifts between the SD and TD, however, further improvement can be made.
Zhang et al. (2022b)	Faster R-CNN, VGG-16, ResNet-101	• Multi-source unsupervised CDOD with information fusion	Pascal VOC2007, Clipart, Watercolor, Comic, Cityscapes, Foggy Cityscapes, SIM10K, and KITTI.	mAP=64.5%	• Help suppress negative transfer caused by abnormal samples.

CNN for each dataset, but only one dataset is evaluated during testing, Second, a hybrid approach, in which VGG-16 has fine-tuned one dataset before being used on another. Similarly, the work in Doshi and Yilmaz (2020) presents a hybrid strategy for detecting video abnormalities with small training data samples using NNs and statistical online KNN decision-making. Object tracking and AED are two components of the VSS proposed in Kale and Shriram (2020). The whole framework identifies and tracks anomalous items. Regarding object tracking, a DTL-based ResNet tracking approach is employed, while the distance metric learning (DML) scheme has been used to detect abnormal activities.

In Shin and Cho (2018), the data scarcity issue is overcome using a GAN model to develop an underlying DTL architecture. The GAN model consists of a generator that creates video sequences and a discriminator that follows the long-term recurrent convolutional network (LRCN) structure. In Keçeli and Kaya (2017), Mumtaz et al. (2018), pre-trained CNN and GoogleNet are used for violent activity detection. Typically, in Keçeli and Kaya (2017), the Lucas–Kanade technique is first utilized to calculate the optical flows of the input videos. Then, by adopting overlapping optical flow magnitudes and orientations, numerous 2D templates are created. The latter is fed into a pretrained CNN, which extracts deep features from several layers. Empirical validation is lastly conducted on violent-flows (Hassner et al., 2012), Hockey (Bermejo Nievas et al., 2011), and movies (Bermejo Nievas et al., 2011) datasets.

In Zhang et al. (2020a), a DDA is adopted to minimize the distribution gap between the test and training data and develop a CDAD approach. Guo et al. (2021) develop an MCDCD that is a unique multi-source UDDA strategy for enhancing anomaly detection using information from several labeled training subjects (i.e., SDs). This approach includes (i) a feature extractor to develop discriminative gait characteristics and (ii) a domain-specific category classifier that maximizes cross-domain discrepancy loss between any two category classifiers. Thus, this helps reduce the domain gap between multiple SDs and the TD while minimizing the cross-entropy loss and reliably identifying source samples. In Arifoglu and Bouchachia (2019), the authors propose to examine recursive auto-encoders (RAE)-based DTL in the context of unusual behavior detection to deal with the problem of data scarcity. They provide a strategy for creating synthetic data to reflect on specific dementia-related behavior.

To increase detection accuracy, DTL extracts human motion characteristics from RGB video frames in Al-Dhamari et al. (2020). Then,

Table 7
Summary of existing CDAD studies with a brief description of their characteristics.

Work	Backbone	Description	Dataset	Best FMD performance	Limitation/advantage
Sahoo et al. (2019)	InceptionV3 and VGG-16	• Study the performance of different classifiers using a two-stream CNN architecture for unusual event detection.	UCF crime	Acc= 50.16%	• Less efficient than the SVM in terms of accuracy.
Atghaei et al. (2020)	VGG-16	• A GAN-based method is proposed to address abnormal event detection problem	UCSD Peds1 UCSD Peds2	AUC =93% (Frame-level)	• Can be applicable for real-time with higher video frame rates.
Liu et al. (2020c)	ResNeSt50	• Anomaly detection in surveillance of video based action recognition.	ImageNet	Acc= 91.14% (Frame level)	• Achieve promising performance and obtain the first prize for general anomaly detection.
Doshi and Yilmaz (2020)	YOLOv3	• An online anomaly detection method for surveillance videos using DTL and any-shot learning	CUHK Avenue UCSD Ped2	AUC= 86.4% AUC= 97.8%	• Significantly outperform the state-of-the-art in terms of any-shot learning.
Bendali-Braham et al. (2019)	TwoStream-I3D	• Ensure the safety of people on the public area by automatic recognition of a crowd movement	Crowd-11	Acc= 70.6%	Outperform the state-of-the-art on Crowd-11 by a consequent margin of ≈ 10% accuracy.
Andrews et al. (2016b)	VGG-F	• Assess transfer representation-learning for anomaly detection by: (i) DTL from PTMs, (ii) DTL from an auxiliary task.	XRAY (Andrews et al., 2016a) Augmented CASIA-WebFace (Yi et al., 2014)	AUROC= 0.99%	• Not all the PTMs perform well.
Fan et al. (2021)	CNN	• Transfer anomaly detection knowledge in an unsupervised manner.	Mnist (M) USPS (U)	AUC=82% (M→U) AUC=96% (U→M)	• The performance is comparable to that of semi-supervised methods.
ZhanLi and JiaWei (2019)	Inception-V3	• Identify abnormal behavior in image samples	NBA	Acc=0.94%	• Less efficient than the three-layer feedforward NN.
Bansod and Nandedkar (2019)	VGG-16	• Anomaly detection from crowd in the field of computer vision using: i) homogeneous approach and ii) hybrid approach (fine-tune CNN for one dataset and then for other dataset.)	UCSD Ped2 UMN (Gray)	Acc=99.56% Acc=99.78%	• Both hybrid and homogeneous approaches detect anomalies with or without sharing the parameters of previous network.

VGGNet-19 is utilized as a source architecture to build a new structure and extract descriptive features. Next, extracted features are then fed into a binary SVM classifier. Similarly, in Lin et al. (2021a), abnormal events in worksites are detected using a DTL-based Faster R-CNN. Feng et al. (2021) build a multiple instance self-training technique (MIST) to refine task-specific discriminative representations effectively for CDAD of video sequences. To boost the performance and reduce the domain shift between the SD and TD, MIST fine-tuned features generated using a self-guided attention-boosted feature encoder have then been used.

Because abnormal events are generally rare, most existing datasets are imbalanced. Therefore, most existing methods have learned the discriminative characteristics from normal data using either semi-supervised or unsupervised procedures. However, they are significantly limited in capturing the abnormal discriminative features, which results in low anomaly detection performance. To close this gap, the authors in Sun et al. (2021) introduce a CD few-shot anomaly detection that can utilize the knowledge learned from various SD videos for solving few-shot anomaly detection in the TD. Specifically, self-supervised training is leveraged on the normal TD data to reduce the domain gap. In Lin et al. (2021b), a CDAD scheme that relies on transferring the knowledge learned on a new synthetic anomaly event dataset is developed. While using 3D-CNN for detecting abnormal events has shown significant performance degradation due to the gap between synthetic and real data, a cyclic 3D-GAN has been then adopted for DA. Table 7 presents a summary of relevant CDAD frameworks discussed in this review.

5.3. Cross-domain human action recognition (CDHAR)

The lack of annotated samples in the TD makes it challenging to develop efficient HAR solutions. To overcome this problem, unsupervised DTL, particularly DDA, has been explored when PTMs, already trained on SDs with a substantial amount of labeled data, are fine-tuned on unlabeled TDs. One of the challenging tasks of HAR is multimodal action recognition, which uses data from cameras and wearable devices to recognize human actions effectively. Existing techniques are categorized into three groups: (i) cross-media action recognition (CMAR) that designs typical multimodal characteristic learning strategies to solve image-to-video HAR problems (Yu et al., 2019; Liu et al., 2019a); (ii) cross-spectral action recognition (CSAR) that uses DDA to address visible-to-infrared HAR problems (Shahroudy et al., 2017; Liu et al., 2018b). For instance, a spatial–optical sequential learning and data organization technique using spatial–optical action data is introduced in Yuan et al. (2018); and (iii) cross-view action recognition (CVAR) that uses DTL techniques to diminish domain discrepancy of action data from different points of view (Liu et al., 2018a; Wang et al., 2019b).

Overall, various frameworks have deployed homogeneous DDA (Khan et al., 2018; Wang et al., 2018a) to adjust class-wise domain discrepancy and promote excellent representation learning. For example, Khan et al. (2018) propose a CNN-based technique for automatically extracting the high-level characteristics in an end-to-end fashion. To quantify the gap of the hidden layers between domains, the Kullback–Leibler divergence has been used. Then, the quantified discrepancy is considered one of the terms in the objective function. The main drawback of homogeneous DDA techniques is their ability to only handle domain inputs with the same feature space. It is suggested in Khan et al. (2015) that action recognition is sub-optimal and that action class labels should be taken into account at the detection stage. Person re-identification (Re-ID) in real-world scenarios presents a key DL challenge, which aims at developing a DL with millions of parameters on a tiny training set with few or no labels. To enable that and overcome the data scarcity problem, a DTL model is suggested in Geng et al. (2016)., which: (a) it is better suited to transfer representations learned from large-scale image/video classification repositories, and (b) combines classification and verification losses, each with its dropout strategy. Besides, the work in Liu et al. (2021e) presents a subtask-dominated TL (STL) approach, which adopts the ResNet50-IBN-a (Pan et al., 2018) as a backbone, to answer the problem of how to handle the heavy unbalanced identity distributions for the one-step person search. The STL technique tackles the long-tail problem in the pretraining stage

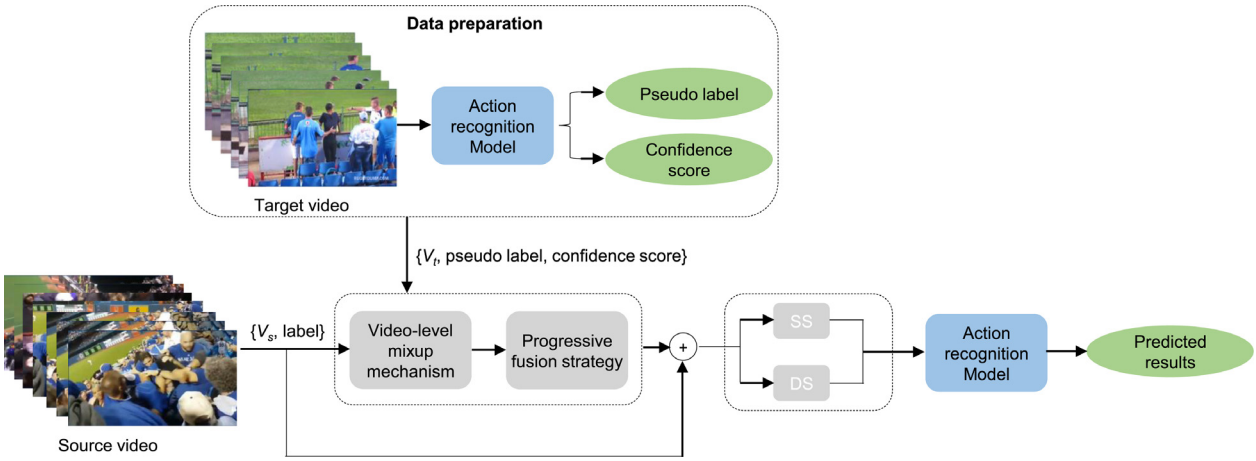


Fig. 12. Flowchart of the CDHAR approach proposed in Wu et al. (2022b).

of the dominating Re-ID subtask and enhances the one-step person search.

In Wu et al. (2022b), a CDHAR scheme is proposed using a temporal shift module (TSM), which uses ResNet-50 as a backbone. This approach relies on (i) pre-training the model on the SD, (ii) processing the TD data (V_t), and (iii) using the PTM for classifying the TD videos to extract its predicted labels (pseudo labels) along with the confidence score of pseudo labels. Fig. 12 portrays the flowchart of this scheme, where $V_s, label$ refers to the annotated SD video and its label. After adding $V_s, label$ and TD videos having pseudo-labels into a mix-up data and fusion modules to generate mixed data, the output is embedded into a model training with sampling strategies, static sampling, and dynamic sampling. This helps in generating fusion data with an increasing proportion of TD data. Lastly, the model trained with the SD and fusion data is utilized for cross-domain prediction.

Hu et al. (2021) propose a deep-frozen TL approach, namely FT-MDnet, to extract Re-ID features from a pretrained detection network to speed up the design and ease implementation. First, an adaptive TL network (ATLnet) is utilized to transform the sharing data of the underlying detection network into a Re-ID feature map using a channel-wise attention process. Then, to extract Re-ID features from the Re-ID feature map, a multi-branch feature representation network called the multiple descriptor network (MDnet) is developed. Various mainstream PTMs have been used to evaluate this approach, including CenterNet, Mask RCNN, YOLOv3, and YOLOv4. The work in Wei et al. (2018b) describes a DTL-based strategy for detecting and classifying people in video data acquired by a high-power lens video camera from distances of many kilometers. A series of computationally efficient image processing procedures are evaluated to recognize moving regions that include humans. These regions are then fed into a pretrained CNN to train only the FCL. Similarly, a real-time top view- and DTL-based person detection system is presented in Ahmed et al. (2021b) using a PTM, namely, CenterNet (Zhou et al., 2019).

Thermal cameras are famous in HAR because of their accuracy in monitoring in the dark and their ability to protect privacy, as explained in Huda et al. (2020, 2021). Manually discovering and annotating large-scale datasets is a costly and time-consuming task. To overcome that, Huda et al. (2020) propose a YOLOv3-based DTL approach to annotate person detection datasets and use the weights for model adaptation of new domains. The authors in Wei and Kehtarnavaz (2019) introduce a semi-supervised faster-RCNN strategy to detect persons and classify the load they carry in video sequences. The moving regions in video frames that include people are first extracted and fed into a quicker RCNN classifier with ResNet-50-based DTL convolutional layers. In Sambolek and Ivašić-Kos (2021), the fine-tuning-based DTL, instances-based DTL, and mapping-based DTL models are built based on YOLOv4 for person detection in drone images. Typically, this study aims at improving the performance of detecting people in search and rescue scenes, where the VisDrone dataset (Zhu et al., 2020) has been deployed as a benchmark.

A real-time person detection based on DTL is proposed in Ahmad et al. (2021), a pretrained cascade RCNN. Using the same process, a social distancing tracking system is proposed in Ahmed et al. (2021c) by detecting people in video sequences using a YOLOv3 OD system. Also, a DTL scheme is considered to reduce computational costs and improve detection accuracy. Fig. 13 illustrates the flowchart of the DTL-based pedestrian detection system using overhead video frames, which has been employed to measure the physical distances between pedestrians. Typically, fine-tuning is adopted by freezing all the layers of the pre-trained YOLOv3 architecture, and only one new layer is trained on the real-world dataset.

On the other hand, the multi-modal action recognition (M2AR) based on sensor-to-vision (S2V) adaptation (from wearable sensors to vision sensors) is mainly a heterogeneous DDA problem that slightly varies from CDHAR problems. This is because of the modality differences between vision sensors and wearable sensors (in terms of data dimensionality), inherent information content, and data distribution. However, only a few studies have explored sensor-to-vision action recognition (SVAR). A multi-modal transfer module to fuse knowledge from different unimodal CNNs is proposed in Joze et al. (2020) and validated on three multi-modal fusion tasks: HAR, audio-visual speech enhancement, and gesture recognition. However, HAR is performed in these methods using raw 1D time-series sensor data, which means they lack texture information, color, and local temporal relationship. This has affected the representativity of wearable-sensor data and made adapting pretrained DL models (e.g., ResNet, VGGNet, AlexNet, etc.) challenging. Moreover, the semantic relationship between wearable sensors and vision-sensors action data that can guide the knowledge transfer is missing.

5.4. Cross-domain crowd counting (CDCC)

CC is an application-oriented activity, and the effectiveness of its inference is critical in real-world applications. Most earlier efforts, on the other hand, depended on large backbone networks and needed expensive run-time consumption, thus limiting their deployment scopes and causing poor scalability. To overcome these issues and make the crowd-counting models more accessible, CDCC methods have been widely investigated. Before discussing some of the relevant studies, we briefly highlight existing datasets used to validate CDCC frameworks. Table 8 summarizes the relevant CC datasets and their basic information, in which Max, Ave, and Min represent the maximum, average, and minimum number of people per image, respectively.

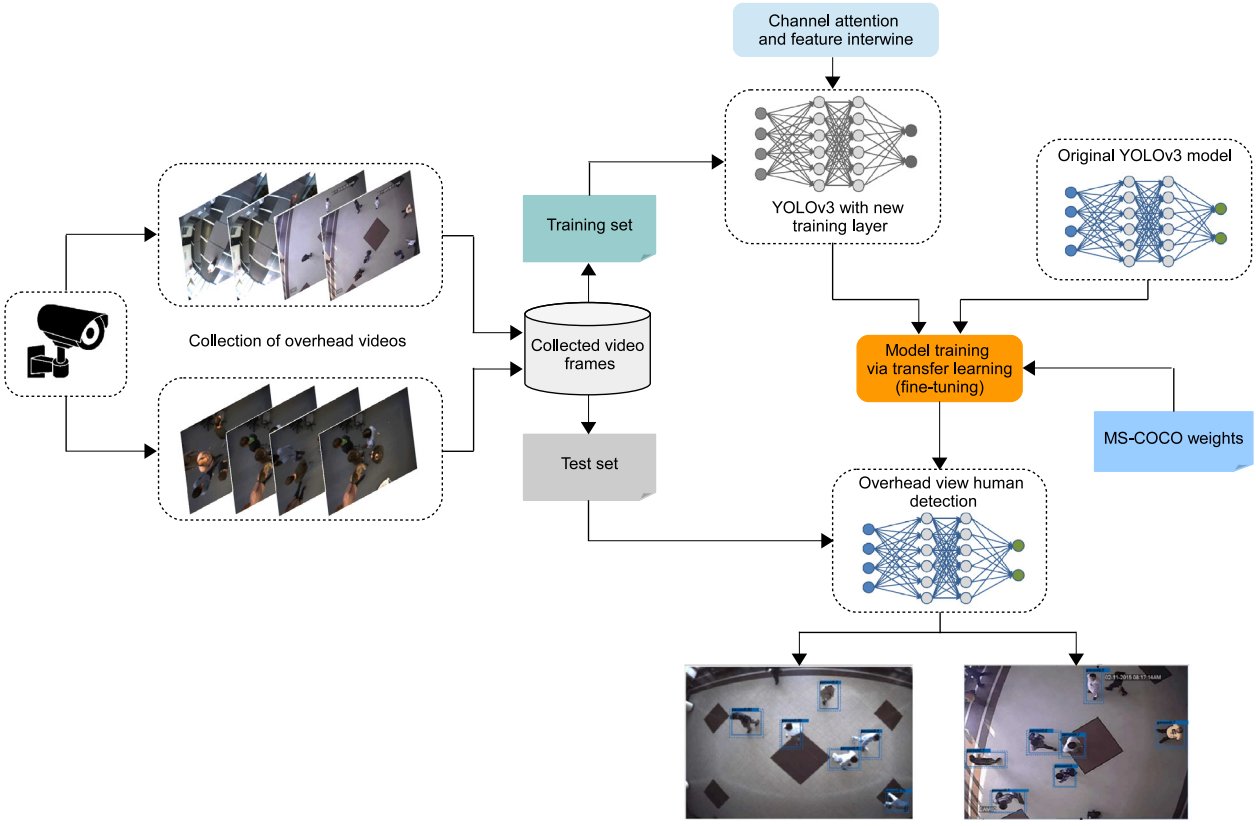


Fig. 13. TL-based pedestrian detection using overhead video frames (Ahmed et al., 2021c).

Table 8
Summary of existing datasets used to validate CDCC techniques.

Dataset	Total number of persons	Max	Ave	Min	Resolution	No. of images
SHA (Zhang et al., 2016)	241,677	3139	501	33	–	482
SHB (Zhang et al., 2016)	88,488	578	123	9	768 × 1024	716
WorldExpo'10 (Zhang et al., 2015)	199,923	253	50	1	576 × 720	3980
UCF_CC_50 (Idrees et al., 2013)	63974	4543	1280	94	–	50
Mall (Chen et al., 2012)	63,325	53	–	13	320 × 240	2000
UCSD (Chan et al., 2008)	49,885	46	25	11	158 × 238	2000
UCF-QNRF (Idrees et al., 2018)	1,251,642	12,865	815	49	–	1535
NWPU-Crowd (Wang et al., 2020b)	2,133,375	20,033	418	0	2191 × 3209	5190
GCC (Wang et al., 2019c)	7,625,843	3995	501	0	1080 × 1920	15,212
JHU-Crowd (Sindagi et al., 2020)	1,515,005	25,791	346	0	910 × 1430	4372
City street (Zhang and Chan, 2019)	63,974	150	–	70	1520 × 2704	500
Venice (Liu et al., 2019b)					1280 × 720	167
BRAINWASH (Stewart et al., 2016)	90,330	–	–	–	480*640	11,769
FDST (Fang et al., 2019)	394,081	57	26.7	9	1920 × 1080 1280 × 720	150,000

The variation of the CC datasets and their challenges have helped boost the number of CDCC studies. For instance, in Xiong et al. (2017), two DTL tasks are considered to transfer the knowledge acquired by bidirectional convolutional long short-term memory (Bi-ConvLSTM) on UCSD and Mall datasets. Specifically, 800 frames from the SD dataset are utilized for training Bi-ConvLSTM, and 50 frames of the target dataset are considered the adaptation set. While in Yao et al. (2017), a deep spatial regression model is proposed based on CNN and LSTM. Concretely, images are fed into a pretrained CNN to extract an ensemble of high-level characteristics. Following, the characteristics in neighboring areas are utilized for regressing the local counts with an LSTM model, considering the spatial information. Lastly, the total count is achieved by summing local patches. In Tong et al. (2018), the CNN-based DTL approach is introduced to count crowds, which uses only five convolutional layers and removes FCLs in the traditional CNN model for achieving an end-to-end system. Geometry-adaptive kernels are used to get the proper density map. The Shanghaitech Part_A dataset was employed to pre-train the model by fixing the first three convolutional layers of the network and training the last two layers on the ZJU_CLASS dataset. In Wilie et al. (2018), the Xception network pretrained parameter, proposed in Boominathan et al. (2016), is utilized as DTL to be trained again with the FCLs. CountNet then achieved a better CC performance by training it with an augmented dataset robust to scale and slice variations.

While the dominant focus within the CC literature has been on the single-frame case or applying CNNs to videos in a frame-by-frame fashion without leveraging motion information, Hossain et al. (2020a) propose a multiscale optical flow pyramid network. The latter considers the spatiotemporal information captured in a video stream by combining an optical flow pyramid with an appearance-based CNN. Then, after putting the SD baseline in place, fine-tuning is performed by simply updating the decoder of the baseline model. The authors in Liu et al. (2020b) investigate a DTL setting in which they learn to recognize and count people in an unlabeled target set by transferring

bi-knowledge from regression and detection-based models to a labeled source set. Ilyas et al. (2021) employ a CNN-based dense feature extraction network for accurate CDCC, consisting of three main modules called (i) backbone network, which has a strong DTL ability, (ii) dense feature extraction modules, and (ii) channel attention module.

In Chen et al. (2020a), a modified VGG-16 network is adopted as a backbone due to its strong DTL ability. Typically, a DDA is applied to adapt the structure to arbitrary resolution by removing three FCLs. Then, the tradeoff between the resource cost and accuracy is considered by eliminating the last two pooling layers. A region relation-aware module is applied at the end of VGG, followed by a bilinear upsample. An end-to-end deep-scale purification network (DSPNet) for dense CC that can encode multiscale characteristics and decrease contextual information loss is proposed in Zeng et al. (2020). DTL is considered for transferring the knowledge learned by DSPNet on the SD datasets, i.e., a transfer from UCF-QNRF or SHB to the TD dataset SHA. Moving forward, different camera angles, exposures, location heights, complicated backdrops, and insufficient annotation data cause supervised learning approaches to fail in real-world applications, and many have overfitting issues. The work in Hou et al. (2021) focuses on training synthetic crowd data and investigating methods to transfer the knowledge to real-world datasets while lowering the need for manual annotation to meet the above challenges. An adaptive domain-invariant feature-extracting module is proposed to increase DDA in feature extraction. In Hossain et al. (2020b), A CDCC is developed using DDA, where the SD and a TD refer to images recorded from cameras in distinct places (e.g., with differing viewpoints, illumination conditions, environment objects, crowd densities, etc.). Sufficient annotated training data from the SD is available, while only a small number of annotated data or completely unlabeled data is available in the TD. Similarly, an error-aware density isomorphism reconstruction network (EDIREC-Net) is proposed in He et al. (2021) for CDCC. This model aims to jointly transfer a counting PTM to TDs and error reasoning to describe reconstruction erroneousness. The density maps in neighboring frames are isomorphic because video crowd dynamics are sequential. Moreover, the estimation–reconstruction consistency is used to track density reconstruction errors and reduce unreliable density reconstructions during training.

In Wang et al. (2021b), the domain discrepancy is described at the parameter level using a neuron linear transformation (NLT) approach. Domain shift is learned by utilizing bias weights and domain factors. Typically, NLT utilizes a few labeled samples from the TD to learn domain shift parameters for a specific neuron of a source model. A linear transformation is adopted to generate the target neuron. Fig. 14 portrays the flowchart of the NLT consisting of three steps: (i) training the source model with synthetic data; (ii) Second, using learnable parameters θ^f and θ^b, defined according to the source model, for modeling the domain shift. Precisely, for a source neuron $\theta_i^S \in \theta^S$ (i is the index of neurons), there is a θ_i^f and a θ_i^b, which have been considered for generating the target neuron θ_i^ρ using a linear operation; (iii), feeding the few shot data into the target model for updating the domain shift parameters after loading the transferred parameters θ^ρ to the target model.

In Zhang et al. (2021d), a multi-view CC paradigm called cross-view cross-scene (CVCS) is presented using unsupervised domain transfer, in which the training and test are processed on multiple scenes with variable camera layouts. CVCS attentively selects and fuses numerous views using camera layout geometry and noise view regularization methods to train the model and handle non-correspondence errors. When only a few training samples are available in a new scene, cross-scene counting becomes challenging. In Yang et al. (2018), information from other scenes is used to learn a cross-scene counting model. Regression is used to count crowds by mapping the properties of crowds to their numbers. Hand-crafted features are generated using block robust principal component analysis segregated crowd foregrounds. Through DDA, existing scene samples (i.e., the SD) are adaptively transferred into the new scene (i.e., the TD). Then, using iterative optimization and training data from both domains, a counting model based on a DDA-extreme learning machine (DDA-ELM) is efficiently learned. In Wu et al. (2021), the C^2MoT method is introduced, a novel CC architecture based on an external momentum template that allows domain-specific information to be encoded via an external template representation. The momentum template is specifically learned in a momentum-updating manner during offline training and then dynamically changed for each test image in an online cross-dataset assessment. The framework of multiple-instance learning (Liu et al., 2021d) presents a strategy for learning crowd segmentation using point-level CC annotations. The generated segments present a crowd-aware fine-grained DDA framework for CC, including two new adaptation modules. The crowd region transfer module restricts the target-domain crowd density distributions.

Besides, while supervised approaches require time-consuming labeling, unsupervised CDCC research utilizing synthetic datasets has become a viable option. In Cai et al. (2021), a two-step DDA approach with multi-level feature response branches is proposed, which takes advantage of intra-domain knowledge to improve TD's adaptability. The scheme in Reddy et al. (2020) is developed to tackle the few-shot scene adaptive CC problem. The model parameters are trained via meta-learning to make successful fine-tuning to a new scene with a few annotated images possible. This technique does not have the fine-tuning limitation of updating particular layers in the decoder closer to the output. Still, it may be used to adjust any decoder parameters. Wang et al. (2019c) introduce a CDCC by proposing a two-stage DDA scheme and establishing a large-scale synthetic repository. Specifically, a SE-CycleGan is used to (i) move synthetic data closer to real-world observations and (ii) apply the developed network in the wild. By contrast, the authors in Li et al. (2019) develop a density adaptation network that aims at discriminating between the density maps produced by the SD and TD. Following, a feature-aware adaptation scheme is proposed in Gao et al. (2020), which extracts domain-invariant characteristics for reducing domain discrepancy in feature layers. Next, a semantic extractor for efficiently distinguishing the crowd and the background in the high-level semantic information is presented in Han et al. (2020). Similarly, a cross-scene CC (CSCC) approach is proposed in Li et al. (2022), which relies on supervised adaptive network parameters. In Liu et al. (2022b), a UDDA across domains using available unlabeled target data is introduced for developing a CDCC approach. In doing so, discovering bi-knowledge transfer between detection- and regression-based networks from an annotated SD has been learned.

Table 9 illustrates the test performance of some popular CDCC methods, where the GCC dataset is considered as the unique SD (Wang et al., 2019c) while for the TD, six datasets are adopted, i.e., UCF-QNRF, SHA, SHB, Mall, UCSD, and WorldExpo'10. In this regard, it is obvious that it is challenging to reach a satisfying performance with no adaptation (NoAdpt). This helps in validating the significant distance between synthetic and real datasets. Also, the performance of supervised training (FA) and fine-tuning (IFS) on a pretrained GCC model with few-shot samples can improve the performance compared to without adaptation. Moreover, it has clearly been seen that NLT has reached better performance than the aforementioned techniques. Additionally, combining NLT and IFS has further improved the performance. For example, considering MAE under UCF-QNRF, this combination reduced the counting error by 3.7% compared to NLT.

Table 10 summarizes some of the relevant CDCC frameworks discussed above. It is clear that most of these works have used VGG-16or VGG-19 as backbones for their models.

5.5. Cross-domain data fusion (CDDF)

Conventional ML techniques generally process data from a unique domain; however, dealing with diverse data modalities in modern big data applications has become the trend. Typically, data collected from different sources can have other representations, distributions,

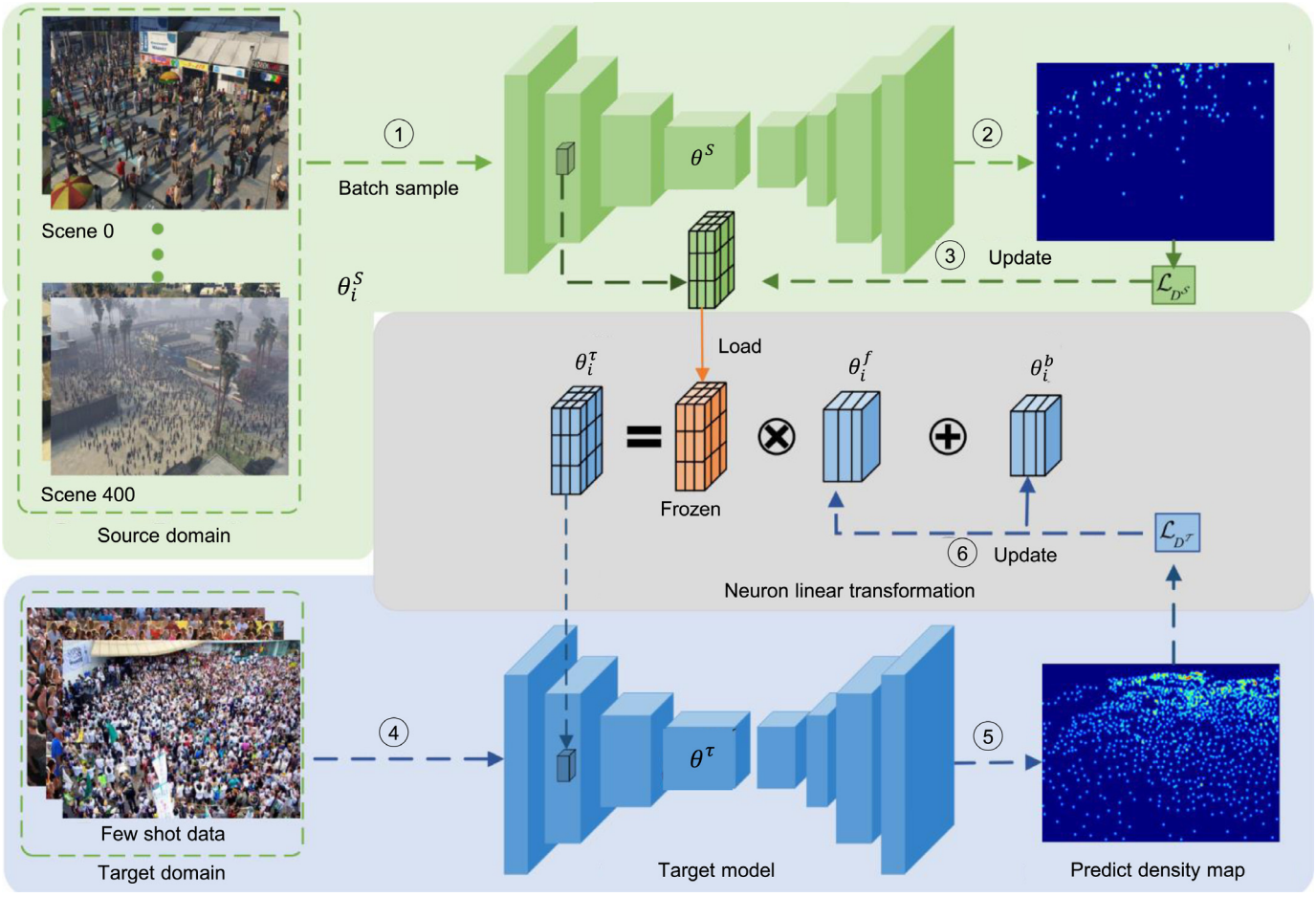


Fig. 14. Flowchart of the CDCC system proposed in Wang et al. (2021b).

Table 9
Performance of some popular CDCC techniques on six different datasets. FS represents 10% shot data from the TD.

Method	Backbone	DA	FS	UCF-QNRF				Shanghai Tech Part B				Shanghai Tech Part A39.7			
				MSE	MAE	SSIM	PSNR	MSE	MAE	SSIM	PSNR	MSE	MAE	SSIM	PSNR
Cycle-Gan (Zhu et al., 2017)	VGG-16	✓	✗	400.6	257.3	0.480	20.80	39.7	25.4	0.763	24.60	204.3	143.3	0.379	19.27
SE Cycle-Gan (Wang et al., 2019c)	VGG-16	✓	✗	384.5	230.4	0.660	21.03	28.3	19.9	0.765	24.78	193.4	123.4	0.407	18.61
SE-Cycle-Gan (JT) (Wang et al., 2021a)	VGG-16	✓	✗	385.7	225.9	0.642	21.10	25.8	16.4	0.786	26.17	189.1	119.6	0.429	18.69
FSC (Han et al., 2020)	VGG-16	✓	✗	390.2	221.2	0.708	23.10	24.7	16.9	0.818	26.20	187.6	129.3	0.513	21.58
IFS (Gao et al., 2019)	VGG-16	✓	✗	357.9	211.7	0.687	21.94	19.4	13.1	0.888	28.03	176.6	112.4	0.502	21.94
FA (Gao et al., 2020)	VGG-16	✓	✗	407.9	269.5	-	-	24.7	16.0	-	-	200.6	144.6	-	-
LIDK (Cai et al., 2021)	VGG-16	✓	✗	375.8	224.3	-	-	22.8	14.3	-	-	-	-	-	-
NoAdpt (Han et al., 2020)	VGG-16	✗	✗	453.7	276.8	0.692	22.22	29.2	20.1	0.895	26.62	279.6	188.0	0.670	20.91
DACC (Han et al., 2020)	VGG-16	✓	✗	343.0	203.5	0.717	21.99	19.4	13.1	0.888	28.03	176.9	112.4	0.502	21.94
NLT (Wang et al., 2021b)	VGG-16	✓	✓	307.1	172.3	0.729	22.8	19.2	11.8	0.937	27.58	157.2	93.8	0.729	21.89
IFS (Gao et al., 2019)+NLT Wang et al. (2021b)	VGG-16	✓	✓	263.1	157.2	0.744	23.01	18.3	10.4	0.942	27.79	151.6	90.1	0.741	22.01
NLT (Wang et al., 2021b)	ResNet-50	✓	✓	279.7	165.8	0.734	22.89	18.8	10.4	0.942	27.79	153.4	91.4	0.749	21.45

Method	Backbone	DA	FS	MALL				UCSD				WorldExpo'10 (only MAE)					
				MSE	MAE	SSIM	PSNR	MSE	MAE	SSIM	PSNR	S1	S2	S3	S4	S5	Avg
Cycle-Gan (Zhu et al., 2017)	VGG-16	✓	✗	-	-	-	-	-	-	-	-	4.4	69.6	49.9	29.2	9.0	32.4
SE Cycle-Gan (Wang et al., 2019c)	VGG-16	✓	✗	-	-	-	-	-	-	-	-	4.3	59.1	43.7	17.0	7.6	26.3
SE-Cycle-Gan (JT) (Wang et al., 2021a)	VGG-16	✓	✗	-	-	-	-	-	-	-	-	4.2	49.6	41.3	19.8	7.2	24.4
FA (Gao et al., 2020)	VGG-16	✓	✗	3.25	2.47	-	-	2.43	2.0	-	-	5.7	59.9	19.7	14.5	8.1	21.6
IFS (Gao et al., 2019)	VGG-16	✓	✗	2.96	2.31	0.933	25.54	2.09	1.76	0.950	24.42	4.5	33.6	14.1	30.4	4.4	17.4
NoAdpt (Han et al., 2020)	VGG-16	✗	✗	6.96	6.20	0.879	24.65	13.22	12.79	0.899	23.94	5.0	89.9	63.1	20.8	17.1	39.2
DACC (Han et al., 2020)	VGG-16	✓	✗	2.96	2.31	0.933	25.54	2.09	1.76	0.950	24.42	4.5	33.6	14.1	30.4	4.4	17.4
NLT (Wang et al., 2021b)	VGG-16	✓	✓	2.55	1.96	0.967	26.92	1.97	1.58	0.942	25.29	2.3	22.8	16.7	19.7	3.9	13.1
IFS (Gao et al., 2019)+NLT Wang et al. (2021b)	VGG-16	✓	✓	2.39	1.86	0.944	27.03	1.81	1.48	0.965	25.58	2.0	15.3	14.7	18.8	3.4	10.8
NLT (Wang et al., 2021b)	ResNet-50	✓	✓	2.42	1.80	0.940	26.84	1.76	1.42	0.964	25.56	3.1	17.8	17.9	20.6	3.2	12.5

Table 10
Summary of the relevant CDCC frameworks discussed in this paper.

Work	Backbone	Description	Dataset	Best performance	Advantage/limitation
Zou et al. (2021)	ASNet	• CDCC using adversarial scoring network	ShanghaiTech UCSD Mall Trancos (Guerrero-Gómez-Olmedo et al., 2015)	MAE=2.76, MSE=3.55	• Mitigate the domain gap at multiple perspectives and boost the adaptation accuracy.
Zeng et al. (2020)	DeconvNet	• CDCC using deep scale purifier network	UCF-QNRF, SHA/SHB, UCF_CC_50	MAE=8.4, RMSE=13.6	• Cannot count multiple types of objects and further improvement can be achieved.
Hou et al. (2021)	CSRNet	• CDCC with MAML	SHA/SHB MALL UCSD GCC NWPU-Crowd	MAE=2.03, MSE=2.41	• High computation cost for training domain-invariant features and more synthetic data is needed for covering different scenarios.
He et al. (2021)	VGG-19	• Unsupervised CDCD using error-aware density isomorphism reconstruction	UCF-QNRF UCSD MALL VENICE FDST (Fang et al., 2019)	MAE=1.79, MSE=2.47	• The erroneousness of density reconstruction is monitored using a reconstruction erroneousness modeling procedure.
Zhang et al. (2021d)	VGG19	• Multi-view CDCC paradigm using CVCS	PETS2009, DukeMTMC CityStreet	MAE=2.83, NAE=0.525	• Can only be applied to real scenes via UDDA.
Wu et al. (2021)	VGG-19	• Zero-shot CDCC using dynamic momentum adaptation	SHA/SHB UCF-QNRF	MAE=12.4, RMSE=21.1	• Achieve leading zero-shot CDCC performance without model fine-tuning.
Liu et al. (2021d)	VGG-16	• Point-derived segmentation for fine-grained DACC	GCC (Wang et al., 2019c), SHA/SHB JHU-CROWD	MAE=13.5, RMSE=21.8	• Learn CC using pixel-level labels without extra annotation effort in the context of multiple instance learning.
Cai et al. (2021)	Masksembles	• Leveraging self-supervision for CDCD	GCC SHA/SHB UCF-QNRF UCF_CC_50 WorldExpo'10	MAE=11.4, RMSE=17.3	• Improve CDCC performance only when labels of synthetic image are available.
Reddy et al. (2020)	CSRNet	• Using meta learning for few-shot scene CDCC	WorldExpo'10, Mall and UCSD	MAE=3.08, RMSE=4.16	• Learn model parameters and facilitate fast adaptation to new target scenes.
Gao et al. (2020)	VGG-16	• CDCC using density alignment and feature-aware adaptation	SHB, WorldExpo'10, Mall, UCSD	MAE=16.4, MSE=25.4	• The counting performance in the real-world needs to be boosted by introducing high-level semantic data in CDCC.
Han et al. (2020)	VGG-16	• Concentrate on semantic consistency for CDCC	SHA/SHB UCF-QNRF	MAE=16.9, MSE=24.7	• Deep transfer should be combined with semantic information to reach higher precision.
Wang et al. (2021b)	ResNet-50	• NLT: Modeling domain shifts for CDCC	SHA/SHB UCF-QNRF WorldExpo'10 UCSD MALL	MAE=1.42, MSE=1.76	• Perform well with increasing few-shot learning data and enhance density maps in counting values and details.
Li et al. (2022)	VGG-16	• Supervised adaptive network parameters for CSCC	WorldExpo'10, Mall, PETS and FDST	MAE=3.47, MSE=4.32	• The performance significantly varies from a dataset to another.

scales, and densities. In this respect, unlocking the knowledge power from distinct but related datasets is challenging, mainly when DL models are used. To that end, developing advanced fusion schemes that can aggregate knowledge of different tasks or domains is paramount. These techniques concentrate on knowledge fusion instead of data merging or schema mapping and rely on implementing CDDF rather than conventional data fusion. Fig. 15 explains the difference between conventional data fusion and CDDF: (a) paradigm of conventional DF, and (b) paradigm of CDDF (Lin et al., 2022).

When developing VSSs, various datasets from the SD and TD can be aggregated at different levels, although every domain has its specific distribution and representation. Typically, advanced fusion techniques should be deployed to combine the knowledge and features from different datasets. One option toward this goal is adopting TL-based fusion techniques, which rely on aggregating knowledge from different tasks or domains and not only merging data (Zheng, 2015). After recognizing objects in the SD and TD concurrently, pertinent feature sets are extracted at both domains to build knowledge bases. The latter can then be fused to collect a more comprehensive knowledge base and thus accurately classify objects (Ghaith et al., 2021).

In Hernandez et al. (2018), a proof-of-concept to collaborate features from multiple modalities for HAR is presented using the teaching–learning approach of TL. Besides, Lin et al. (2022) develop a data fusion-based DTL-empowered granular trust evaluation mechanism. Overall, considering the relevance of CDDF for VSSs, little work has been identified at this study's stage. Therefore, more research and development efforts should be paid to this topic in the near future to improve the performance of VSSs and widen their deployment.

6. Discussion of key challenges

6.1. Accuracy saturation

Performing TL by fine-tuning pre-trained DL models with large-scale datasets, e.g., ImageNet helps considerably in improving and accelerating training. At the same time, the accuracy is often bottlenecked by the limited dataset sizes of the target tasks. To close this gap,

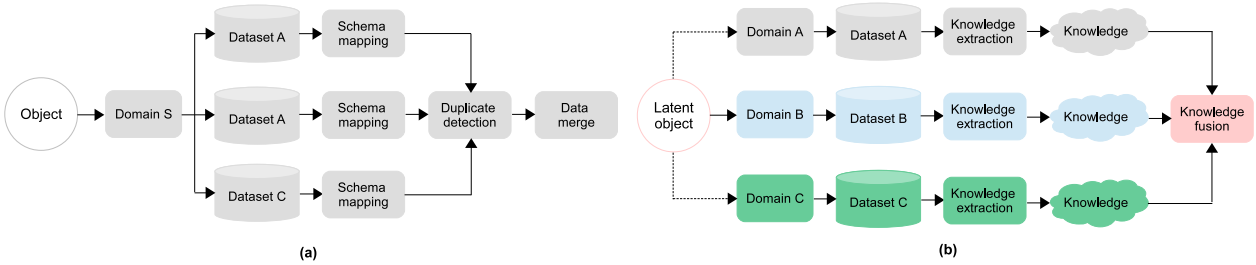


Fig. 15. Difference between conventional data fusion and CDDF: (a) paradigm of conventional DF, and (b) paradigm of CDDF.

some regularization techniques that constrain the outer layer weights of the target network using the starting point as references (SPAR), were investigated. For instance, in Li et al. (2021d), a regularized DTL approach, namely DL transfer using a feature map with attention (DELTA), is proposed. Rather than constraining the weights of a DL model, DELTA attempts to preserve the outer layer outputs of the source model. In this regard, DELTA minimizes the empirical loss before aligning the outer layer outputs of the two models. Accordingly, it constrains a subset of feature maps, which have accurately been chosen by the attention learned in a supervised learning process.

6.2. The problem of negative transfer

When the DTL degrades the classification of prediction performance (or accuracy) of the newly developed model, it is due to negative transfer. Indeed, DTL works perfectly if the SD and TD are sufficiently similar. Put differently, when the data used to pre-train the DTL model is different enough than the data used to re-train this model (or some of its parts), the performance might be worse than expected. Moreover, regardless that the SD and TD can appear similar to humans, algorithms may only sometimes agree with them. In this context, most existing DTL- and DDA-based VSS solutions cannot handle the level variations across domains due to image/video level features, e.g., the category and amount of foreground objects or object co-occurrence. This can result in critical feature misalignment and even negative transfer.

Despite the pervasiveness of negative transfer, it is generally defined informally, lacking precise definition, meticulous analysis, and systematic treatment. In Wang et al. (2019a), propose to filter out unrelated source samples to reduce the impact of negative transfer in adversarial networks. This generic approach can be used in a wide range of TL-based VSS solutions. Recently, some studies have been proposed to investigate this severe issue. For example, Zhang et al. (2022b) demonstrate that the attention mechanism could promote the transfer of similar patterns between multiple SDs and RDs and alleviate negative transfer.

It is also worth mentioning that some studies, such as Sousa et al. (2014), have demonstrated that the quality of the DTL performance is directly related to the Kullback–Leibler divergence estimated between the SDs and TD datasets. Put simply, it may not be practical for some application scenarios to use a DTL, or the successful DTL's architecture should be reliable for heterogeneous problems. Additionally, although these intuitive ideas have experimentally shown a relation between domain divergence and DTL algorithms' performance, theoretical descriptions for these behaviors still need to be discovered.

Moreover, most existing DDA techniques generally address the different OD problems using direct feature alignment between the SD and TD at the video frame level, the instance level (e.g., region proposals), or both. However, it has been demonstrated that a direct feature alignment of all object instances from the two domains can result in a negative transfer. This is because of (i) the existing outlier target instances containing confusing objects, which do not belong to any category of the SD, and therefore it is challenging to detect them, and (ii) the low-relevance source instances, which are significantly statistically distinct from target instances despite that their contained objects are from the same class/group (Chen et al., 2021a).

6.3. The problem of overfitting

One of the challenges in developing DTL-based techniques for VSSs is overcoming overfitting, mainly due to training complex DL models with insufficient data. Although this issue is familiar with all ML models, overfitting in DTL also occurs when the developed model learns details and noises from SD data that negatively impact its outputs (Mutasa et al., 2020). In DTL, the network layers cannot be removed to identify with confidence the best classification/prediction parameters of the DL models. Typically, removing the first layers may negatively impact the dense layers since the number of trainable parameters will change. On the flip side, the number of dense layers can be reduced; however, analyzing the number of layers to be removed while avoiding the overfitting of the model is time-consuming and challenging.

6.4. Measuring knowledge gains

Measuring the knowledge gained when a DTL model is adopted to conduct specific tasks is of utmost importance. However, this challenge has yet to receive its merit, and a few research works have targeted it. Bengio et al. in Glorot et al. (2011) have attempted to analyze how to quantify the DTL gain. In this regard, four measures have been introduced to quantify the gain knowledge, i.e., transfer error, transfer loss, transfer ratio, and in-domain ratio. Despite the fact that these measures can overcome some interpretation issues related to the performance results occurring when dealing with various SDs, it is undetermined how they will behave in other TL-based methods, especially for VSS applications where class sets are different between problems. Further, they can result in non-definite performance if a perfect baseline model is obtained.

6.5. Data annotation and generalization

Most VSS applications still face severe labeling problems where the intractability to collect new large annotated amounts of data (including plentiful images with large diversity) is perceived. In this regard, learning generalized models is an actual problem that can be resolved by training them with data from multiple benchmarks. This is the case of Chen et al. (2021c), a multi-domain joint learning approach is proposed, and a domain-specific knowledge propagating network (DKPNet) is introduced. This helps unbiasedly and simultaneously learn the knowledge from different data domains. Typically, a variational attention scheme models the attention distributions for different domains. In contrast, an intrinsic variational attention technique is developed to leverage the dilemma of overlapped domains and sub-domains.

7. Future directions

Despite the great achievements of numerous DTL- and DDA-based VSS methods, there is still rooms for improving the performance gap between them and the upper bound. Following the aforementioned the open challenges of existing DTL- and DDA techniques, a set of potential future directions are identified in this section.

7.1. Interpretability and explainability of DTL and DDA models

The popularity of DL in VSS applications is booming due to the availability of powerful computing boards and graphic processing units (GPUs). Typically, more complex problems have been introduced, such as CDOD, CDAD, CDHAR, CDCC, and CDDF, which require explaining the outputs of DL models when applied to these problems. However, most explanations for DL algorithms are not appropriate to the video analysis field (Wang and Breckon, 2022). This is because most of them are overshadowed by image methods, and video analysis is usually based on complex scene Understanding and recognition, which makes adding explanations a challenging task (Wu et al., 2022a). To that end, more effort has recently been devoted by the scientific community to investigating the development of explainability techniques for DL-based video analysis. On the other side, it has been shown in recent studies that adding interpretability and explainability to ML models in general, and particularly DL algorithms, can help practitioners (with or without an ML background) in quickly inspecting their inner mechanics and gaining some intuitions on why they are likely to succeed or fail (Dasari et al., 2021). This is also a fundamental property in DTL and DDA models, as interpretations can result in valuable insights, which not only strengthen our understanding of their principles but also help us get ideas on to fix their drawbacks (Angelov and Soares, 2020; Danso et al., 2021).

Currently, most existing TL-based VSS algorithms introduced to perform different tasks, such as CC, HAR, and OD, are unexplainable. These techniques may operate well under particular scenes; however, their performance can be dropped when the scenes in the TD differ from those in the SD. For instance, the characteristics automatically generated by DL models lack interpretability; thus, they do not provide any insights to explain gaps between the SD and TD (Zhang et al., 2022a; Krishnan et al., 2020). To that end, adding interpretability and explainability (e.g., rule reasoning) to DTL- and DDA-based VSSs algorithms is one of the promising research axes in the near future. Hence, only achieving the highest performance is targeted, but also reasonable levels of interpretability of used DTL/DDA models (Lu et al., 2021b).

In this context, some studies have recently started investigating this research direction. For example, to overcome the arbitrary selection of layers in TL models, which fail to reach a high classification accuracy for target tasks, the authors in Arefeen et al. (2021) develop an explainable TL scheme that intelligently selects layers from DNNs. Similarly, in Liu et al. (2022a), an interpretable DTL technique is built by selecting features using a group-wise feature importance scoring scheme. Moving on, federated TL and explainable AI are combined in Raza et al. (2022) to alleviate the data scarcity and privacy preservation issues. Besides, an explainable by design supervised DA scheme, namely XSDA-Net, is proposed in Kamakshi and Krishnan (2022) by integrating a case-based reasoning strategy into the XSDA-Net. This enables explaining the predictions of test instances related to similar-looking regions in the SD and TD. Moving forward, in Zunino et al. (2021), explainability is used as an approach to bridge the visual-semantic discrepancy between distinct domains. Put differently; model explanations have been deployed to disentangle domain-specific data from other pertinent features.

In Nourani et al. (2020), the authors investigate the role of explainable AI and first impressions for interactive video analysis. Moving on, Zhuo et al. (2019) deploy prior knowledge and state transitions to develop an explainable human action analysis and understanding system. Similarly, in Han et al. (2022), Han et al. propose an explainable action reasoning approach using one-shot video graph generation. In Roy et al. (2021), the accuracy-explanability gap is addressed by adopting an interpretable, tractable probabilistic DL algorithm. Besides, explain AI has been also used for content based retrieval of video frames, as explained in Chittajallu et al. (2019).

Very recently, the efforts made for developing DTL and explainable AI have been joined to design explainable DTL. Specifically, fine-tuning-based DTL relies on selecting some pre-trained layers of a DL model to be re-trained and freezing the rest. However, arbitrarily selecting these layers can fail to reach the desired results on the TD task. To that end, explainability can be introduced to accurately choose the appropriate that maximizes the performance of the DTL model on a target task while maintaining a low overhead of successive training (Arefeen et al., 2021; Islam et al., 2022). Moving on, explainable DTL is introduced in Liu et al. (2022a) to analyze high-dimensional genomic data. Typically, a dimensionality reduction approach has been developed using explainable group-wise feature importance scores.

7.2. Trust and transparency in DTL-based VSSs

AI and DL's usefulness comes from applying computer intelligence skills, including object detection, crowd counting, abnormal event detection, etc. It is no surprise that AI- and DL-based VSSs have become a hotspot for many applications across different sectors (Meske and Bunde, 2020). Although the increasing performance achieved by AI/DL-based VSSs, using these models results in the "black box" problem, which decreases trust towards AI/DL. Typically, most deployed techniques cannot provide explainability as they use uninterpretable feature representations that hide the decision-making process when the different VSS tasks are conducted. On the other hand, while increasing effort has been put into assessing the VSS trustworthiness by measuring the accuracy, reliability, and generalizability, among other characteristics, deciding that a VSS is trustworthy since it meets its system requirements would not guarantee the widespread use of AI/DL (Buhrmester et al., 2021). However, making AI- and DL-based VSSs more transparent by adopting transparency and explainability principles can help them eventually choose the right, unbiased algorithms. Specifically, transparent AI and DL systems can answer why a particular decision or prediction has been made or avoided given the same inputs. In this context, using explainable AI has been identified as an emerging scheme to increase trust in VSSs.

To that end, some studies have recently focused on building transparency in AI and DL algorithms to help users trust computer decisions and know how AI/DL systems arrive at their recommendations and conclusions. For example, the authors in Meske and Bunde (2020) present a case study of medical image analysis to demonstrate how explainable artificial intelligence can help open the "black box" and enhance the level of AI transparency and trust. Besides, a three-fold explanation approach is proposed in Druce et al. (2021) to improve users' trust in deep reinforcement learning-driven autonomous systems. However, it is worth mentioning that more effort should be dedicated to fostering trust in AI-based VSSs, especially those relying on DTL and DDA models.

7.3. Online DTL and DDA

Most DTL- and DDA-based VSS frameworks have focused on offline learning, which is not the best option for various real-world applications. By contrast, adopting online learning is inevitable, especially when small amounts of real-time data are available (Wang et al., 2013). Specifically, online DTL relies on sequentially receiving data in the new domains and using the knowledge of models/classifiers learned in the SDs. However, the application of online DTL faces some challenges, including (i) negative transfer that can be generated when using online DTL on homogeneous domains, especially with the existence of a considerable difference between two conditional probabilities; (ii) difficulty in continuously training old classifiers (learned on the SDs) with the new features due to the high discrepancy between the two feature spaces (SD and TD), and (iii) difficulty to measure the distribution difference between the SD and TD since only a predictive model of the SD is given while the data samples on the TD are received on-the-fly sequentially and therefore should be predicted immediately (Wu et al., 2017,?).

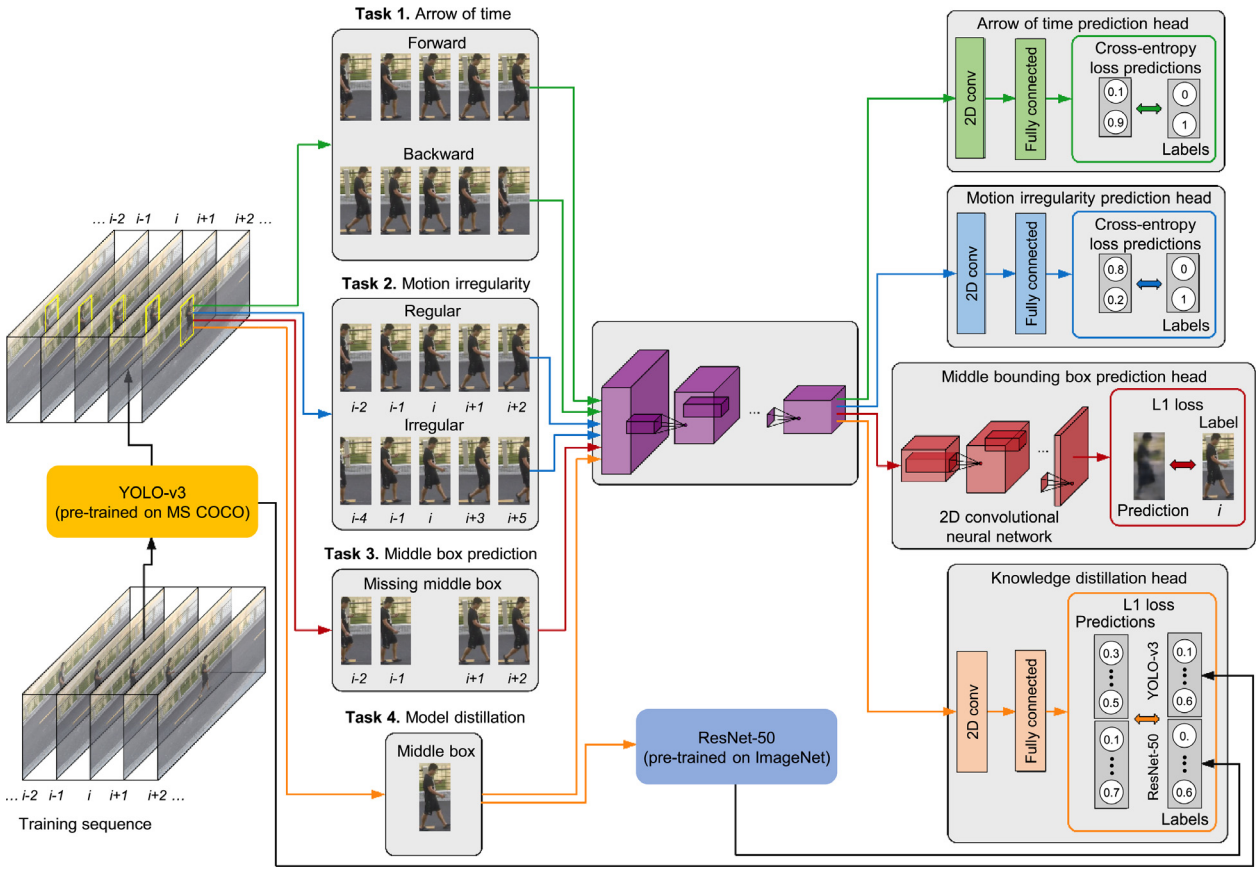


Fig. 16. Example of an AED approach based on knowledge distillation (Georgescu et al., 2021).

7.4. Model compression and knowledge distillation (KD)

One promising way to reduce the complexity of VSSs is by developing model compression and acceleration methods. KD, a representation type of the latter, helps efficiently learn a small student model from a large teacher model. In this respect, a smaller student model with few parameters can perform similarly to a larger teacher model. Most existing CC frameworks require prohibitive computation as they rely on heavy backbone networks, restricting their real-time implementation and causing poor scalability. To overcome these issues, a structured knowledge transfer (SKT) technique, exploiting the structured knowledge of a well-trained teacher network, is proposed in Liu et al. (2020a) for generating a compact but efficient student network. Similarly, in Gu (2020), a CC scheme based on perspective-aware distillation is introduced. Moving on, in Jiang et al. (2021), a task-specific KD scheme for CC, namely ShuffleCount, is proposed. It uses hierarchic feature regulation to learn from the teacher network and avoid negative knowledge transfer from the teacher. Besides, in most HAR systems, significant modality differences exist between data recorded by vision sensors and wearable sensors. To that end, a semantic-aware adaptive knowledge distillation network is proposed in Liu et al. (2021c) for enhancing the accuracy of action recognition in vision-sensor modality. The knowledge from different wearable sensors has effectively been transferred and distilled. This approach utilizes RGB videos and multiple wearable sensors as student and teacher modalities, respectively. In Georgescu et al. (2021), an AED using knowledge distillation is introduced, which relies on multi-task and self-supervised learning. Fig. 16 portrays the flowchart of this approach. Typically, objects in videos are first detected using YOLOv3. Then, three self-supervised tasks are devised for every detected object (which aim to learn the arrow of time, predict motion irregularities, and expect the object appearances in the middleboxes) and a knowledge distillation task (based on ResNet-50 and YOLOv3 as teachers). Moving on, a 3D-CNN has jointly been trained on the four tasks.

Besides, in Wang et al. (2020c), a TL-based CC approach is developed, namely MobileCount, appropriate for embedded and mobile devices with limited computation resources. Moreover, a multi-layer knowledge distillation scheme is adopted to transfer a complex model's information to numerous layers of the developed architecture. This helps in improving performance without increasing the number of floating-point operations. In Kong et al. (2019), a multi-modality distillation framework with the attention process is proposed by realizing a heterogeneous S2V DA. Besides, the CDCC framework developed in Liu et al. (2020a) relies on transferring the structured knowledge of a well-trained teacher network to build a lightweight but powerful target network for CC. It includes two complementary transfer modules: an intra-layer pattern transfer (IPT) that sequentially distills the knowledge embedded in layer-wise features of the source network to guide feature learning of the target network and an inter-layer relation transfer (ILRT) that densely distills the source's cross-layer correlation knowledge to better regulate the target's feature progression.

7.5. CDDF and DTL-based data fusion

CDDF is a promising research direction where there is still room for improvement. Typically, some CDDF strategies and DTL-based fusion techniques have not been wholly used in VSSs. For instance, few DTL-based fusion studies have discussed the aggregation of heterogeneous tasks or features from different domains in VSS tasks. Existing ML-based and DL-based fusion techniques proved to be efficient in other applications (e.g., the multimodal fusion of remote sensing data) and are potential candidates to be combined with DTL and DDA schemes (De et al., 2021). For instance, a multimodal CDDF technique is introduced in De et al. (2021) to analyze the environmental impacts of large-scale

events in public areas. Moving on, a GAN-based deformation invariant CDDF scheme is proposed in Wang et al. (2021c) to synthesize medical images.

7.6. DDA in the wild

Up to date, most DDA frameworks principally concentrate on neat settings. However, in the real world, DDA problems can be a quite complicated aggregation of various "clean" settings (Tran et al., 2019). For instance, in some practical DDA settings, several SDs could be available for which some could have no labeled observations, some could have abundant labeled samples, and some could have few labeled patterns. Simultaneously, the label spaces of the SDs and TDs may have a significant discrepancy. Additionally, there may be numerous TDs, where some TDs can have classes not existing in any SDs. In this direction, resolving real-world DDA issues is an underexplored field, where more research effort should be put in the near future (Li et al., 2021c; Abnar et al., 2021).

8. Conclusion

This comprehensive survey has revolved around VSSs using DTL, which has recently been found as a promising research topic to reduce the computation complexity of DL models and overcome various issues related to the non-availability of sufficient real data, the lack of annotated datasets, the discrepancy between training and testing data, etc. Typically, we have elaborated on this topic by (i) describing the various purposes that motivate this research, (ii) presenting the background of DTL and clarifying its concepts, and (ii) presenting a well-defined taxonomy. This study has resulted in an overall taxonomy of existing DTL-based VSS proposals discussed in the literature by classifying them based on different aspects. Given the prevalence of contributions to developing DTL systems, the literature dealing with fine-tuning and DDA methodologies has thoroughly been inspected.

Moving on, the implications of adopting CDDF in multi-modal VSSs have been explored, unveiling the potential knowledge fusion instead of data aggregation or schema mapping. Next, our discussions have been moved beyond the contributions in DTL-based VSSs to developing more generalized and real-time solutions for real-world scenarios. Current challenges related to the DTL and those specific to VSS tasks have also been discussed. The future of DTL-based VSS methodologies has been explored via the reflections carried out in the discussions held throughout this review, agreeing on the imperative requirement for a proper understanding of the potentiality and caveats opened up by DTL approaches.

Although DTL and DDA have recently achieved success, numerous challenges and issues still need further investigation. Typically, homogeneous DDA has been the focus of most existing DTL-based VSS techniques, where it is assumed that the feature spaces between the SD and TD are the same. However, that is only true in some real-world scenarios. To that end, most recent studies have focused on heterogeneous DDA attempting to transfer knowledge without this critical limitation and benefit from existing datasets to help with more tasks. Moreover, heterogeneous DDA will receive further attention in the near future. Despite the progress summarized in this review, there is still room for improvement. On the other hand, alternative techniques for video frames' translation and style transfer might be investigated, particularly those that help enhance the semantic consistencies of the translation. This will likely boost the performance of object detectors and VSSs in general.

CRediT authorship contribution statement

Yassine Himeur: Conceptualization, Formal analysis, Methodology, Writing – review & editing. **Somaya Al-Maadeed:** Funding acquisition, Writing – review & editing, Project administration, Supervision. **Hamza Kheddar:** Conceptualization, Formal analysis, Methodology, Writing – review & editing. **Noor Al-Maadeed:** Funding acquisition, Writing – review & editing, Project administration, Supervision. **Khalid Abualsaud:** Funding acquisition, Writing – review & editing, Project administration, Supervision. **Amr Mohamed:** Funding acquisition, Writing – review & editing, Project administration, Supervision. **Tamer Khattab:** Funding acquisition, Writing – review & editing, Project administration, Supervision.

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.

Data availability

Data will be made available on request.

Acknowledgments

This research work was made possible by research grant support (QUEX-CENG-SCDL-19/20-1) from Supreme Committee for Delivery and Legacy (SC) in Qatar. The statements made herein are solely the responsibility of the authors. Open Access funding provided by the Qatar National Library.

References

Abdulazeem, Y., Balaha, H.M., Bahgat, W.M., Badawy, M., 2021. Human action recognition based on transfer learning approach. IEEE Access 9, 82058–82069.

Abnar, S., Berg, R.v.d., Ghiasi, G., Dehghani, M., Kalchbrenner, N., Sedghi, H., 2021. Gradual domain adaptation in the wild: When intermediate distributions are absent. arXiv preprint arXiv:2106.06080.

Agarwal, N., Sondhi, A., Chopra, K., Singh, G., 2021. Transfer learning: Survey and classification. In: Smart Innovations in Communication and Computational Sciences. Springer, pp. 145–155.

Ahmad, M., Ahmed, I., Jeon, G., 2021. An IoT-enabled real-time overhead view person detection system based on cascade-RCNN and transfer learning. J. Real-Time Image Process. 1–11.

Ahmadi, M., Ouarda, W., Alimi, A.M., 2020. Efficient and fast objects detection technique for intelligent video surveillance using transfer learning and fine-tuning. Arab. J. Sci. Eng. 45 (3), 1421–1433.

Ahmed, I., Ahmad, M., Ahmad, A., Jeon, G., 2021a. Top view multiple people tracking by detection using deep SORT and YOLOV3 with transfer learning: Within 5G infrastructure. Int. J. Mach. Learn. Cybern. 12 (11), 3053–3067.

Ahmed, I., Ahmad, M., Rodrigues, J.J., Jeon, G., 2021b. Edge computing-based person detection system for top view surveillance: Using CenterNet with transfer learning. Appl. Soft Comput. 107, 107489.

Ahmed, I., Ahmad, M., Rodrigues, J.J., Jeon, G., Din, S., 2021c. A deep learning-based social distance monitoring framework for COVID-19. Sustainable Cities Soc. 65, 102571.

Ahmed, S.M., Raychaudhuri, D.S., Paul, S., Oymak, S., Roy-Chowdhury, A.K., 2021d. Unsupervised multi-source domain adaptation without access to source data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10103–10112.

Al-Dhamari, A., Sudirman, R., Mahmood, N.H., 2020. Transfer deep learning along with binary support vector machine for abnormal behavior detection. IEEE Access 8, 61085–61095.

Alyafeai, Z., AlShaibani, M.S., Ahmad, I., 2020. A survey on transfer learning in natural language processing. arXiv preprint arXiv:2007.04239.

Andrews, J., Morton, E., Griffin, L., 2016a. Detecting anomalous data using auto-encoders. Int. J. Mach. Learn. Comput. 6, 21.

Andrews, J., Tanay, T., Morton, E.J., Griffin, L.D., 2016b. Transfer Representation-Learning for Anomaly Detection. JMLR.

Angelov, P., Soares, E., 2020. Towards explainable deep neural networks (xDNN). Neural Netw. 130, 185–194.

Anon, 2022. CitySCENE. Available online: https://cityscene.github.io/#/. (Accessed 19 April 2022).

Arefeen, M.A., Nimi, S.T., Uddin, M.Y.S., Lee, Y., 2021. TransJury: Towards explainable transfer learning through selection of layers from deep neural networks. In: 2021 IEEE International Conference on Big Data. Big Data, IEEE, pp. 978–984.

Arifoglu, D., Bouchachia, A., 2019. Abnormal behaviour detection for dementia sufferers via transfer learning and recursive auto-encoders. In: 2019 IEEE International Conference on Pervasive Computing and Communications Workshops. PerCom Workshops, IEEE, pp. 529–534.

Arruda, V.F., Berriel, R.F., Paixão, T.M., Badue, C., De Souza, A.F., Sebe, N., Oliveira-Santos, T., 2022. Cross-domain object detection using unsupervised image translation. Expert Syst. Appl. 192, 116334.

Atghaei, A., Ziaeinejad, S., Rahmati, M., 2020. Abnormal event detection in urban surveillance videos using GAN and transfer learning. arXiv preprint arXiv:2011.09619.

Bansod, S., Nandedkar, A., 2019. Transfer learning for video anomaly detection. J. Intell. Fuzzy Systems 36 (3), 1967–1975.

Bari, A., Saini, T., Kumar, A., 2021. Fire detection using deep transfer learning on surveillance videos. In: 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks. ICICV, IEEE, pp. 1061–1067.

Bashath, S., Perera, N., Tripathi, S., Manjang, K., Dehmer, M., Streib, F.E., 2022. A data-centric review of deep transfer learning with applications to text data. Inform. Sci. 585, 498–528.

Belhadi, A., Djenouri, Y., Srivastava, G., Djenouri, D., Lin, J.C.-W., Fortino, G., 2021. Deep learning for pedestrian collective behavior analysis in smart cities: A model of group trajectory outlier detection. Inf. Fusion 65, 13–20.

Bendali-Braham, M., Weber, J., Forestier, G., Idoumghar, L., Muller, P.-A., 2019. Transfer learning for the classification of video-recorded crowd movements. In: 2019 11th International Symposium on Image and Signal Processing and Analysis. ISPA, IEEE, pp. 271–276.

Berg, A., Ahlberg, J., Felsberg, M., 2018. Generating visible spectrum images from thermal infrared. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1143–1152.

Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R., 2011. Violence detection in video using computer vision techniques. In: International Conference on Computer Analysis of Images and Patterns. Springer, pp. 332–339.

Bilal, M., Maqsood, M., Yasmin, S., Hasan, N.U., Rho, S., 2021. A transfer learning-based efficient spatiotemporal human action recognition framework for long and overlapping action classes. J. Supercomput. 1–36.

Boominathan, L., Kruthiventi, S.S., Babu, R.V., 2016. Crowdnet: A deep convolutional network for dense crowd counting. In: Proceedings of the 24th ACM International Conference on Multimedia. pp. 640–644.

Buhrmester, V., Münch, D., Arens, M., 2021. Analysis of explainers of black box deep neural networks for computer vision: A survey. Mach. Learn. Knowl. Extr. 3 (4), 966–989.

Cai, Y., Chen, L., Ma, Z., Lu, C., Wang, C., He, G., 2021. Leveraging intra-domain knowledge to strengthen cross-domain crowd counting. In: 2021 IEEE International Conference on Multimedia and Expo. ICME, IEEE, pp. 1–6.

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

Chan, A.B., Liang, Z.-S.J., Vasconcelos, N., 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–7.

Che Aminudin, M.F., Suandi, S.A., 2022. Video surveillance image enhancement via a convolutional neural network and stacked denoising autoencoder. Neural Comput. Appl. 34 (4), 3079–3095.

Chen, X., Bin, Y., Gao, C., Sang, N., Tang, H., 2020a. Relevant region prediction for crowd counting. Neurocomputing 407, 399–408.

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

Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L., 2018. Domain adaptive faster R-CNN for object detection in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3339–3348.

Chen, K., Loy, C.C., Gong, S., Xiang, T., 2012. Feature mining for localised crowd counting. In: Bmvc, Vol. 1, no. 2. p. 3.

Chen, J., Wu, X., Duan, L., Chen, L., 2021a. Sequential instance refinement for cross-domain object detection in images. IEEE Trans. Image Process. 30, 3970–3984.

Chen, X., Xu, L., Cao, M., Zhang, T., Shang, Z., Zhang, L., 2021b. Design and implementation of human-computer interaction systems based on transfer support vector machine and EEG signal for depression patients' emotion recognition. J. Med. Imag. Health Inform. 11 (3), 948–954.

Chen, B., Yan, Z., Li, K., Li, P., Wang, B., Zuo, W., Zhang, L., 2021c. Variational attention: Propagating domain-specific knowledge for multi-domain learning in crowd counting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16065–16075.

Chen, C., Zheng, Z., Ding, X., Huang, Y., Dou, Q., 2020b. Harmonizing transferability and discriminability for adapting object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8869–8878.

Chittajallu, D.R., Dong, B., Tunison, P., Collins, R., Wells, K., Fleshman, J., Sankaranarayanan, G., Schwaitzberg, S., Cavuoto, L., Enquobahrie, A., 2019. XAI-CBIR: Explainable AI system for content based retrieval of video frames from minimally invasive surgery videos. In: 2019 IEEE 16th International Symposium on Biomedical Imaging. ISBI 2019, IEEE, pp. 66–69.

Choi, W., Yang, W., Na, J., Park, J., Lee, G., Nam, W., 2021. Unsupervised gait phase estimation with domain-adversarial neural network and adaptive window. IEEE J. Biomed. Health Inf..

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1251–1258.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3213–3223.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1. CVPR'05, Ieee, pp. 886–893.

Danso, S.O., Zeng, Z., Muniz-Terrera, G., Ritchie, C.W., 2021. Developing an explainable machine learning-based personalised dementia risk prediction model: A transfer learning approach with ensemble learning algorithms. Front. Big Data 4, 21.

Dasari, C.M., Amilpur, S., Bhukya, R., 2021. Exploring variable-length features (motifs) for predicting binding sites through interpretable deep neural networks. Eng. Appl. Artif. Intell. 106, 104485.

De, S., Wang, W., Zhou, Y., Perera, C., Moessner, K., Alraja, M.N., 2021. Analysing environmental impact of large-scale events in public spaces with cross-domain multimodal data fusion. Computing 103 (9), 1959–1981.

Delussu, R., Putzu, L., Fumera, G., 2022. Scene-specific crowd counting using synthetic training images. Pattern Recognit. 124, 108484.

Deng, Y., Huang, D., Du, S., Li, G., Zhao, C., Lv, J., 2021a. A double-layer attention based adversarial network for partial transfer learning in machinery fault diagnosis. Comput. Ind. 127, 103399.

Deng, W., Su, Z., Qiu, Q., Zhao, L., Kuang, G., Pietikäinen, M., Xiao, H., Liu, L., 2021b. Deep ladder reconstruction-classification network for unsupervised domain adaptation. Pattern Recognit. Lett. 152, 398–405.

Devaguptapu, C., Akolekar, N., M Sharma, M., N Balasubramanian, V., 2019. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.

Doersch, C., Zisserman, A., 2019. Sim2real transfer learning for 3d human pose estimation: Motion to the rescue. Adv. Neural Inf. Process. Syst. 32, 12949–12961.

Doshi, K., Yilmaz, Y., 2020. Any-shot sequential anomaly detection in surveillance videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 934–935.

Druce, J., Harradon, M., Tittle, J., 2021. Explainable artificial intelligence (XAI) for increasing user trust in deep reinforcement learning driven autonomous systems. arXiv preprint arXiv:2106.03775.

Durrani, S., Arshad, U., 2021. Transfer learning from high-resource to low-resource language improves speech affect recognition classification accuracy. arXiv preprint arXiv:2103.11764.

Fan, C., Zhang, F., Liu, P., Sun, X., Li, H., Xiao, T., Zhao, W., Tang, X., 2021. Importance weighted adversarial discriminative transfer for anomaly detection. arXiv preprint arXiv:2105.06649.

Fang, Z., Lu, J., Liu, A., Liu, F., Zhang, G., 2021. Learning bounds for open-set learning. In: International Conference on Machine Learning. PMLR, pp. 3122–3132.

Fang, Y., Zhan, B., Cai, W., Gao, S., Hu, B., 2019. Locality-constrained spatial transformer network for video crowd counting. In: 2019 IEEE International Conference on Multimedia and Expo. ICME, IEEE, pp. 814–819.

Feng, J.-C., Hong, F.-T., Zheng, W.-S., 2021. Mist: Multiple instance self-training framework for video anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14009–14018.

Flohr, F., Gavrila, D., et al., 2013. PedCut: An iterative framework for pedestrian segmentation combining shape models and multiple data cues. In: BMVC.

Fuhl, W., Castner, N., Zhuang, L., Holzer, M., Rosenstiel, W., Kasneci, E., 2018. Mam: Transfer learning for fully automatic video annotation and specialized detector creation. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17 (1), 2096–2130.

Gao, J., Han, T., Wang, Q., Yuan, Y., 2019. Domain-adaptive crowd counting via inter-domain features segregation and gaussian-prior reconstruction. arXiv preprint arXiv:1912.03677.

Gao, J., Yuan, Y., Wang, Q., 2020. Feature-aware adaptation and density alignment for crowd counting in video surveillance. IEEE Trans. Cybern. 51 (10), 4822–4833.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3354–3361.

Geng, C., Huang, S.-j., Chen, S., 2020. Recent advances in open set recognition: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 43 (10), 3614–3631.

Geng, M., Wang, Y., Xiang, T., Tian, Y., 2016. Deep transfer learning for person re-identification. arXiv preprint arXiv:1611.05244.

Georgescu, M.-I., Barbalau, A., Ionescu, R.T., Khan, F.S., Popescu, M., Shah, M., 2021. Anomaly detection in video via self-supervised and multi-task learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12742–12752.

Georgescu, M.-I., Ionescu, R.T., Khan, F.S., Popescu, M., Shah, M., 2020. A background-agnostic framework with adversarial training for abnormal event detection in video. arXiv preprint arXiv:2008.12328.

Ghaith, I.H., Rawashdeh, A., Al Zubi, S., 2021. Transfer learning in data fusion at autonomous driving. In: 2021 International Conference on Information Technology. ICIT, IEEE, pp. 714–718.

Giel, A., Diaz, R., 2015. Recurrent Neural Networks and Transfer Learning for Action Recognition. Stanford University.

Glorot, X., Bordes, A., Bengio, Y., 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In: ICML.

Gochoo, M., Akhter, I., Jalal, A., Kim, K., 2021. Stochastic remote sensing event classification over adaptive posture estimation via multifused data and deep belief network. Remote Sens. 13 (5), 912.

Gu, Y., 2020. Perspective-aware distillation-based crowd counting. In: Proceedings of the 2020 4th International Conference on Deep Learning Technologies. ICDLT, pp. 123–128.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al., 2018. Recent advances in convolutional neural networks. Pattern Recognit. 77, 354–377.

Guerrero-Gómez-Olmedo, R., Torre-Jiménez, B., López-Sastre, R., Maldonado-Bascón, S., Onoro-Rubio, D., 2015. Extremely overlapping vehicle counting. In: Iberian Conference on Pattern Recognition and Image Analysis. Springer, pp. 423–431.

Gunther, M., Cruz, S., Rudd, E.M., Boult, T.E., 2017. Toward open-set face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 71–80.

Guo, Y., Gu, X., Yang, G.-Z., 2021. MCDCD: Multi-source unsupervised domain adaptation for abnormal human gait detection. IEEE J. Biomed. Health Inf..

Guo, T., Huynh, C.P., Solh, M., 2019. Domain-adaptive pedestrian detection in thermal images. In: 2019 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 1660–1664.

Guo, H., Pasunuru, R., Bansal, M., 2020. Multi-source domain adaptation for text classification via distancenet-bandits. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, (no. 05), pp. 7830–7838.

Gupta, S., Hoffman, J., Malik, J., 2016. Cross modal distillation for supervision transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2827–2836.

Han, T., Gao, J., Yuan, Y., Wang, Q., 2020. Focus on semantic consistency for cross-domain crowd understanding. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1848–1852.

Han, T., Liu, C., Wu, R., Jiang, D., 2021a. Deep transfer learning with limited data for machinery fault diagnosis. Appl. Soft Comput. 103, 107150.

Han, G., Yu, G., Liu, L., Cui, L., Domeniconi, C., Zhang, X., 2021b. Open-set crowdsourcing using multiple-source transfer learning. arXiv preprint arXiv:2111.04073.

Han, L., Zhao, Y., Chen, H., Chandrasekar, V., 2021c. Advancing radar nowcasting through deep transfer learning. IEEE Trans. Geosci. Remote Sens. 60, 1–9.

Han, Y., Zhuo, T., Zhang, P., Huang, W., Zha, Y., Zhang, Y., Kankanhalli, M., 2022. One-shot video graph generation for explainable action reasoning. Neurocomputing 488, 212–225.

Hassan, M.M., Uddin, M.Z., Mohamed, A., Almogren, A., 2018. A robust human activity recognition system using smartphone sensors and deep learning. Future Gener. Comput. Syst. 81, 307–313.

Hassner, T., Itcher, Y., Kliper-Gross, O., 2012. Violent flows: Real-time detection of violent crowd behavior. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, pp. 1–6.

Hazarika, D., Poria, S., Zimmermann, R., Mihalcea, R., 2021. Conversational transfer learning for emotion recognition. Inf. Fusion 65, 1–12.

He, Y., Ma, Z., Wei, X., Hong, X., Ke, W., Gong, Y., 2021. Error-aware density isomorphism reconstruction for unsupervised cross-domain crowd counting. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 1540–1548.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hernandez, N., Razzaq, M.A., Nugent, C., McChesney, I., Zhang, S., 2018. Transfer learning and data fusion approach to recognize activities of daily life. In: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare. pp. 227–231.

Hinton, G., Vinyals, O., Dean, J., et al., 2015. Distilling the knowledge in a neural network 2 (7). arXiv preprint arXiv:1503.02531.

Hoffman, J., Mohri, M., Zhang, N., 2018. Algorithms and theory for multiple-source adaptation. Adv. Neural Inf. Process. Syst. 31.

Hoffman, J., Tzeng, E., Darrell, T., Saenko, K., 2017. Simultaneous deep transfer across domains and tasks. In: Domain Adaptation in Computer Vision Applications. Springer, pp. 173–187.

Hossain, M.A., Cannons, K., Jang, D., Cuzzolin, F., Xu, Z., 2020a. Video-based crowd counting using a multi-scale optical flow pyramid network. In: Proceedings of the Asian Conference on Computer Vision.

Hossain, M.A., Reddy, M.K.K., Cannons, K., Xu, Z., Wang, Y., 2020b. Domain adaptation in crowd counting. In: 2020 17th Conference on Computer and Robot Vision. CRV, IEEE, pp. 150–157.

Hou, X., Xu, J., Wu, J., Xu, H., 2021. Cross domain adaptation of crowd counting with model-agnostic meta-learning. Appl. Sci. 11 (24), 12037.

Hu, R., Wang, T., Zhou, Y., Snoussi, H., Cherouat, A., 2021. FT-MDnet: A deep-frozen transfer learning framework for person search. IEEE Trans. Inf. Forensics Secur. 16, 4721–4732.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.

Huang, C., Wang, X., Cao, J., Wang, S., Zhang, Y., 2020. HCF: A hybrid CNN framework for behavior detection of distracted drivers. IEEE Access 8, 109335–109349.

Huda, N.U., Gade, R., Moeslund, T.B., 2021. Effects of pre-processing on the performance of transfer learning based person detection in thermal images. In: 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning. PRML, IEEE, pp. 86–91.

Huda, N.U., Hansen, B.D., Gade, R., Moeslund, T.B., 2020. The effect of a diverse dataset for transfer learning in thermal person detection. Sensors 20 (7), 1982.

Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I., 2015. Multispectral pedestrian detection: Benchmark dataset and baseline. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1037–1045.

Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K., 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360.

Idrees, H., Saleemi, I., Seibert, C., Shah, M., 2013. Multi-source multi-scale counting in extremely dense crowd images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2547–2554.

Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., Shah, M., 2018. Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 532–546.

Ilyas, N., Lee, B., Kim, K., 2021. HADF-crowd: A hierarchical attention-based dense feature extraction network for single-image crowd counting. Sensors 21 (10), 3483.

Inoue, N., Furuta, R., Yamasaki, T., Aizawa, K., 2018. Cross-domain weakly-supervised object detection through progressive domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5001–5009.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. PMLR, pp. 448–456.

Islam, M.N., Hasan, M., Hossain, M., Alam, M., Rabiul, G., Uddin, M.Z., Soylu, A., et al., 2022. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. Sci. Rep. 12 (1), 1–14.

Jamal, A., Namboodiri, V.P., Deodhare, D., Venkatesh, K., 2018. Deep domain adaptation in action space. In: BMVC, Vol. 2, no. 3. p. 5.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia. pp. 675–678.

Jiang, M., Lin, J., Wang, Z.J., 2021. ShuffleCount: Task-specific knowledge distillation for crowd counting. In: 2021 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 999–1003.

Jiao, Y., Yao, H., Xu, C., 2021a. SAN: Selective alignment network for cross-domain pedestrian detection. IEEE Trans. Image Process. 30, 2155–2167.

Jiao, L., Zhang, R., Liu, F., Yang, S., Hou, B., Li, L., Tang, X., 2021b. New generation deep learning for video object detection: A survey. IEEE Trans. Neural Netw. Learn. Syst..

Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S.N., Rosaen, K., Vasudevan, R., 2016. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? arXiv preprint arXiv:1610.01983.

Joshi, A.B., Kumar, D., Gaffar, A., Mishra, D., 2020. Triple color image encryption based on 2D multiple parameter fractional discrete Fourier transform and 3D arnold transform. Opt. Lasers Eng. 133, 106139.

Joze, H.R.V., Shaban, A., Iuzzolino, M.L., Koishida, K., 2020. MMTM: Multimodal transfer module for CNN fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13289–13299.

Kale, S., Shriram, R., 2020. Suspicious activity detection using transfer learning based ResNet tracking from surveillance videos. In: SoCPaR. pp. 208–220.

Kamakshi, V., Krishnan, N.C., 2022. Explainable supervised domain adaptation. arXiv preprint arXiv:2205.09943.

Keçeli, A., Kaya, A., 2017. Violent activity detection with transfer learning method. Electron. Lett. 53 (15), 1047–1048.

Kensert, A., Harrison, P.J., Spjuth, O., 2019. Transfer learning with deep convolutional neural networks for classifying cellular morphological changes. SLAS Discov.: Adv. Life Sci. R D 24 (4), 466–475.

Khaire, P., Kumar, P., 2022. A semi-supervised deep learning based video anomaly detection framework using RGB-d for surveillance of real-world critical environments. Forensic Sci. Int.: Digit. Invest. 40, 301346.
Khan, S., Khan, M.A., Alhaisoni, M., Tariq, U., Yong, H.-S., Armghan, A., Alenezi, F., 2021. Human action recognition: A paradigm of best deep learning features selection and serial based extended fusion. Sensors 21 (23), 7941.
Khan, A., Khan, S., Hassan, B., Zheng, Z., 2022. CNN-based smoker classification and detection in smart city application. Sensors 22 (3), 892.
Khan, M.A.A.H., Roy, N., Misra, A., 2018. Scaling human activity recognition via deep learning-based domain adaptation. In: 2018 IEEE International Conference on Pervasive Computing and Communications. PerCom, IEEE, pp. 1–9.
Khan, F.S., Xu, J., Van De Weijer, J., Bagdanov, A.D., Anwer, R.M., Lopez, A.M., 2015. Recognizing actions through action-specific person detection. IEEE Trans. Image Process. 24 (11), 4422–4432.
Kieu, M., Bagdanov, A.D., Bertini, M., Bimbo, A.d., 2020. Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In: European Conference on Computer Vision. Springer, pp. 546–562.
Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J., 2017. Learning to discover cross-domain relations with generative adversarial networks. In: International Conference on Machine Learning. PMLR, pp. 1857–1865.
Kim, M., Joung, S., Park, K., Kim, S., Sohn, K., 2019. Unpaired cross-spectral pedestrian detection via adversarial feature learning. In: 2019 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 1650–1654.
Kim, J., Kim, T., Oh, S.-H., Do, K., Ryu, J.-G., Kim, J., 2021. Deep visual domain adaptation and semi-supervised segmentation for understanding wave elevation using wave flume video images. Sci. Rep. 11 (1), 1–12.
Kong, Q., Wu, Z., Deng, Z., Klinkigt, M., Tong, B., Murakami, T., 2019. Mmact: A large-scale dataset for cross modal human action understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8658–8667.
Krishnan, J., Purohit, H., Rangwala, H., 2020. Unsupervised and interpretable domain adaptation to rapidly filter tweets for emergency services. In: Proceedings of the 12th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE Press, pp. 409–416.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25.
Kuang, X., Zhu, J., Sui, X., Liu, Y., Liu, C., Chen, Q., Gu, G., 2020. Thermal infrared colorization via conditional generative adversarial network. Infrared Phys. Technol. 107, 103338.
Lamas, A., Tabik, S., Montes, A.C., Pérez-Hernández, F., García, J., Olmos, R., Herrera, F., 2022. Human pose estimation for mitigating false negatives in weapon detection in video-surveillance. Neurocomputing.
Leong, M.C., Prasad, D.K., Lee, Y.T., Lin, F., 2020. Semi-CNN architecture for effective spatio-temporal learning in action recognition. Appl. Sci. 10 (2), 557.
Li, Y., Gao, Y., Chen, B., Zhang, Z., Zhu, L., Lu, G., 2021a. JDMAN: Joint discriminative and mutual adaptation networks for cross-domain facial expression recognition. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3312–3320.
Li, S., Hu, Z., Zhao, M., Bi, S., Sun, Z., 2022. Cross-scene crowd counting based on supervised adaptive network parameters. Signal, Image Video Process. 1–8.
Li, W., Huan, W., Hou, B., Tian, Y., Zhang, Z., Song, A., 2021b. Can emotion be transferred?–A review on transfer learning for EEG-based emotion recognition. IEEE Trans. Cogn. Dev. Syst..
Li, H., Wan, R., Wang, S., Kot, A.C., 2021c. Unsupervised domain adaptation in the wild via disentangling representation learning. Int. J. Comput. Vis. 129 (2), 267–283.
Li, X., Xiong, H., Chen, Z., Huan, J., Liu, J., Xu, C.-Z., Dou, D., 2021d. Knowledge distillation with attention for deep transfer learning of convolutional networks. ACM Trans. Knowl. Discov. Data (TKDD) 16 (3), 1–20.
Li, W., Yongbo, L., Xiangyang, X., 2019. CODA: Counting objects via scale-aware adversarial density adaption. In: 2019 IEEE International Conference on Multimedia and Expo. ICME, IEEE, pp. 193–198.
Li, X., Zhang, W., Ma, H., Luo, Z., Li, X., 2020. Partial transfer learning in machinery cross-domain fault diagnostics using class-weighted adversarial networks. Neural Netw. 129, 313–322.
Lin, Z.-H., Chen, A.Y., Hsieh, S.-H., 2021a. Temporal image analytics for abnormal construction activity identification. Autom. Constr. 124, 103572.
Lin, W., Gao, J., Wang, Q., Li, X., 2021b. Learning to detect anomaly events in crowd scenes from synthetic data. Neurocomputing 436, 248–259.
Lin, H., Garg, S., Hu, J., Wang, X., Piran, M.J., Hossain, M.S., 2022. Data fusion and transfer learning empowered granular trust evaluation for internet of things. Inf. Fusion 78, 149–157.
Lin, C., Zhao, S., Meng, L., Chua, T.-S., 2020. Multi-source domain adaptation for visual sentiment classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, (no. 03), pp. 2661–2668.
Liu, M.-Y., Breuel, T., Kautz, J., 2017. Unsupervised image-to-image translation networks. Adv. Neural Inf. Process. Syst. 30.
Liu, L., Chen, J., Wu, H., Chen, T., Li, G., Lin, L., 2020a. Efficient crowd counting via structured knowledge transfer. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2645–2654.
Liu, Z.-H., Jiang, L.-B., Wei, H.-L., Chen, L., Li, X.-H., 2021a. Optimal transport-based deep domain adaptation approach for fault diagnosis of rotating machine. IEEE Trans. Instrum. Meas. 70, 1–12.
Liu, X., Li, G., Han, Z., Zhang, W., Yang, Y., Huang, Q., Sebe, N., 2021b. Exploiting sample correlation for crowd counting with multi-expert network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3215–3224.
Liu, Y., Lu, Z., Li, J., Yang, T., 2018a. Hierarchically learned view-invariant representations for cross-view action recognition. IEEE Trans. Circuits Syst. Video Technol. 29 (8), 2416–2430.
Liu, Y., Lu, Z., Li, J., Yang, T., Yao, C., 2019a. Deep image-to-video adaptation and fusion networks for action recognition. IEEE Trans. Image Process. 29, 3168–3182.
Liu, Y., Lu, Z., Li, J., Yao, C., Deng, Y., 2018b. Transferable feature representation for visible-to-infrared cross-dataset human action recognition. Complexity 2018.
Liu, L., Meng, Q., Weng, C., Lu, Q., Wang, T., Wen, Y., 2022a. Explainable deep transfer learning model for disease risk prediction using high-dimensional genomic data. PLoS Comput. Biol. 18 (7), e1010328.
Liu, W., Salzmann, M., Fua, P., 2019b. Context-aware crowd counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5099–5108.
Liu, X., Van De Weijer, J., Bagdanov, A.D., 2018c. Leveraging unlabeled data for crowd counting by learning to rank. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7661–7669.
Liu, Y., Wang, K., Li, G., Lin, L., 2021c. Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition. IEEE Trans. Image Process. 30, 5573–5588.
Liu, Y., Wang, Z., Shi, M., Satoh, S., Zhao, Q., Yang, H., 2020b. Towards unsupervised crowd counting via regression-detection bi-knowledge transfer. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 129–137.
Liu, Y., Wang, Z., Shi, M., Satoh, S., Zhao, Q., Yang, H., 2022b. Discovering regression-detection bi-knowledge transfer for unsupervised cross-domain crowd counting. Neurocomputing.
Liu, F., Xu, X., Qiu, S., Qing, C., Tao, D., 2015. Simple to complex transfer learning for action recognition. IEEE Trans. Image Process. 25 (2), 949–960.
Liu, Y., Xu, D., Ren, S., Wu, H., Cai, H., He, S., 2021d. Fine-grained domain adaptive crowd counting via point-derived segmentation. arXiv preprint arXiv:2108.02980.
Liu, C., Yang, H., Zhou, Q., Zheng, S., 2021e. Subtask-dominated transfer learning for long-tail person search. arXiv preprint arXiv:2112.00527.
Liu, K., Zhu, M., Fu, H., Ma, H., Chua, T.-S., 2020c. Enhancing anomaly detection in surveillance videos with transfer learning from action recognition. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 4664–4668.
Loey, M., Manogaran, G., Taha, M.H.N., Khalifa, N.E.M., 2021. A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic. Measurement 167, 108288.
Long, M., Cao, Y., Wang, J., Jordan, M.I., 2015. Learning transferable features with deep adaptation networks. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. ICML '15, JMLR.org, pp. 97–105.
Long, M., Cao, Z., Wang, J., Jordan, M.I., 2018. Conditional adversarial domain adaptation. Adv. Neural Inf. Process. Syst. 31.
Long, M., Zhu, H., Wang, J., Jordan, M.I., 2016. Unsupervised domain adaptation with residual transfer networks. Adv. Neural Inf. Process. Syst. 29.
Long, M., Zhu, H., Wang, J., Jordan, M.I., 2017. Deep transfer learning with joint adaptation networks. In: International Conference on Machine Learning. PMLR, pp. 2208–2217.
Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., Zhang, G., 2015. Transfer learning using computational intelligence: A survey. Knowl.-Based Syst. 80, 14–23.
Lu, Y., Tian, Z., Zhou, R., Liu, W., 2021a. A general transfer learning-based framework for thermal load prediction in regional energy system. Energy 217, 119322.
Lu, S., Wu, D., Zhang, Z., Wang, S.-H., 2021b. An explainable framework for diagnosis of COVID-19 pneumonia via transfer learning and discriminant correlation analysis. ACM Trans. Multimedia Comput., Commun., Appl. (TOMM) 17 (3s), 1–16.
Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N., 2010. Anomaly detection in crowded scenes. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1975–1981.
Marnissi, M.A., Fradi, H., Sahbani, A., Amara, N.E.B., 2022. Unsupervised thermal-to-visible domain adaptation method for pedestrian detection. Pattern Recognit. Lett. 153, 222–231.
Maschler, B., Weyrich, M., 2021. Deep transfer learning for industrial automation: A review and discussion of new techniques for data-driven machine learning. IEEE Ind. Electron. Mag. 15 (2), 65–75.
Mathew, A., Mathew, J., Govind, M., Mooppan, A., 2017. An improved transfer learning approach for intrusion detection. Procedia Comput. Sci. 115, 251–257.
Melhart, D., Liapis, A., Yannakakis, G.N., 2021. The affect game Annotation (AGAIN) dataset. arXiv preprint arXiv:2104.02643.
Meske, C., Bunde, E., 2020. Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support. In: International Conference on Human-Computer Interaction. Springer, pp. 54–69.
Morid, M.A., Borjali, A., Del Fiol, G., 2021. A scoping review of transfer learning research on medical image analysis using ImageNet. Comput. Biol. Med. 128, 104115.
Mumtaz, A., Sargano, A.B., Habib, Z., 2018. Violence detection in surveillance videos with deep network using transfer learning. In: 2018 2nd European Conference on Electrical Engineering and Computer Science. EECS, IEEE, pp. 558–563.
Munir, F., Azam, S., Rafique, M.A., Sheri, A.M., Jeon, M., 2020. Thermal object detection using domain adaptation through style consistency.

Mutasa, S., Sun, S., Ha, R., 2020. Understanding artificial intelligence based radiology studies: What is overfitting? Clinical Imaging 65, 96–99.

Nguyen, V.-A., Nguyen, T., Le, T., Tran, Q.H., Phung, D., 2021. Stem: An approach to multi-source domain adaptation with guarantees. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9352–9363.

Niu, S., Jiang, Y., Chen, B., Wang, J., Liu, Y., Song, H., 2021. Cross-modality transfer learning for image-text information management. ACM Trans. Manag. Inf. Syst. (TMIS) 13 (1), 1–14.

Niu, S., Liu, Y., Wang, J., Song, H., 2020. A decade survey of transfer learning (2010–2020). IEEE Trans. Artif. Intell. 1 (2), 151–166.

Nourani, M., Honeycutt, D.R., Block, J.E., Roy, C., Rahman, T., Ragan, E.D., Gogate, V., 2020. Investigating the importance of first impressions and explainable ai with interactive video analysis. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–8.

Pan, B., Cao, Z., Adeli, E., Niebles, J.C., 2020. Adversarial cross-domain action recognition with co-attention. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, (no. 07), pp. 11815–11822.

Pan, X., Luo, P., Shi, J., Tang, X., 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 464–479.

Pan, S.J., Yang, Q., 2009. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.

Pei, Z., Cao, Z., Long, M., Wang, J., 2018. Multi-adversarial domain adaptation. In: Thirty-Second AAAI Conference on Artificial Intelligence.

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

Perera, P., Patel, V.M., 2019. Deep transfer learning for multiple class novelty detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11544–11552.

Prabono, A.G., Yahya, B.N., Lee, S.-L., 2021a. A typical sample regularizer autoencoder for cross-domain human activity recognition. Inf. Syst. Front. 23 (1), 71–80.

Prabono, A.G., Yahya, B.N., Lee, S.-L., 2021b. Hybrid domain adaptation with deep network architecture for end-to-end cross-domain human activity recognition. Comput. Ind. Eng. 151, 106953.

Qi, L., Wang, L., Huo, J., Zhou, L., Shi, Y., Gao, Y., 2019. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8080–8089.

Rajasekhar, G.P., Granger, E., Cardinal, P., 2021. Deep domain adaptation with ordinal regression for pain assessment using weakly-labeled videos. Image Vis. Comput. 110, 104167.

Ramachandran, P., Zoph, B., Le, Q.V., 2017. Searching for activation functions. arXiv preprint arXiv:1710.05941.

Ramirez, P.Z., Tonioni, A., Salti, S., Stefano, L.D., 2019. Learning across tasks and domains. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8110–8119.

Raza, A., Tran, K.P., Koehl, L., Li, S., 2022. Designing ecg monitoring healthcare system with federated transfer learning and explainable AI. Knowl.-Based Syst. 236, 107763.

Reddy, M.K.K., Hossain, M., Rochan, M., Wang, Y., 2020. Few-shot scene adaptive crowd counting using meta-learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2814–2823.

Ren, C.-X., Liu, Y.-H., Zhang, X.-W., Huang, K.-K., 2022. Multi-source unsupervised domain adaptation via pseudo target domain. IEEE Trans. Image Process. 31, 2122–2135.

Rezaee, K., Zadeh, H.G., Chakraborty, C., Khosravi, M.R., Jeon, G., 2022. Smart visual sensing for overcrowding in COVID-19 infected cities using modified deep transfer learning. IEEE Trans. Ind. Inform..

Ribani, R., Marengoni, M., 2019. A survey of transfer learning for convolutional neural networks. In: 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images Tutorials. SIBGRAPI-T, IEEE, pp. 47–57.

Roy, C., Nourani, M., Honeycutt, D.R., Block, J.E., Rahman, T., Ragan, E.D., Ruozzi, N., Gogate, V., 2021. Explainable activity recognition in videos: Lessons learned. Appl. AI Lett. 2 (4), e59.

Rudd, E.M., Jain, L.P., Scheirer, W.J., Boult, T.E., 2017. The extreme value machine. IEEE Trans. Pattern Anal. Mach. Intell. 40 (3), 762–768.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.

Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R., 2016. Progressive neural networks. arXiv preprint arXiv:1606.04671.

Sahoo, S.R., Dash, R., Mahapatra, R.K., Sahu, B., 2019. Unusual event detection in surveillance video using transfer learning. In: 2019 International Conference on Information Technology. ICIT, IEEE, pp. 319–324.

Saito, K., Ushiku, Y., Harada, T., Saenko, K., 2019. Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6956–6965.

Sakaridis, C., Dai, D., Van Gool, L., 2018. Semantic foggy scene understanding with synthetic data. Int. J. Comput. Vis. 126 (9), 973–992.

Sambolek, S., Ivašić-Kos, M., 2021. Transfer learning methods for training person detector in drone imagery. In: Proceedings of SAI Intelligent Systems Conference. Springer, pp. 688–701.

Sánchez, F.L., Hupont, I., Tabik, S., Herrera, F., 2020. Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. Inf. Fusion 64, 318–335.

Saponara, S., Elhanashi, A., Gagliardi, A., 2021. Real-time video fire/smoke detection based on CNN in antifire surveillance systems. J. Real-Time Image Process. 18 (3), 889–900.

Sayed, A.N., Himeur, Y., Bensaali, F., 2022. Deep and transfer learning for building occupancy detection: A review and comparative analysis. Eng. Appl. Artif. Intell. 115, 105254.

Sen, A., Deb, K., 2021. Categorization of actions in soccer videos using a combination of transfer learning and gated recurrent unit. ICT Express.

Serpush, F., Rezaei, M., 2020. Complex human action recognition in live videos using hybrid fr-dl method. arXiv preprint arXiv:2007.02811.

Shahroudy, A., Ng, T.-T., Gong, Y., Wang, G., 2017. Deep multimodal feature analysis for action recognition in rgb+ d videos. IEEE Trans. Pattern Anal. Mach. Intell. 40 (5), 1045–1058.

Shen, Z., Xu, Y., Ni, B., Wang, M., Hu, J., Yang, X., 2018. Crowd counting via adversarial cross-scale consistency pursuit. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5245–5254.

Shi, Z., Zhang, L., Liu, Y., Cao, X., Ye, Y., Cheng, M.-M., Zheng, G., 2018. Crowd counting with deep negative correlation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5382–5390.

Shin, W., Cho, S.-B., 2018. CCTV image sequence generation and modeling method for video anomaly detection using generative adversarial network. In: International Conference on Intelligent Data Engineering and Automated Learning. Springer, pp. 457–467.

Si, J., Shi, H., Chen, J., Zheng, C., 2021. Unsupervised deep transfer learning with moment matching: A new intelligent fault diagnosis approach for bearings. Measurement 172, 108827.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al., 2016. Mastering the game of go with deep neural networks and tree search. Nature 529 (7587), 484–489.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sindagi, V., Yasarla, R., Patel, V.M., 2020. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. IEEE Trans. Pattern Anal. Mach. Intell..

Soleimani, E., Nazerfard, E., 2021. Cross-subject transfer learning in human activity recognition systems using generative adversarial networks. Neurocomputing 426, 26–34.

Sousa, R., Silva, L.M., Alexandre, L.A., Santos, J., De Sá, J.M., 2014. Transfer learning: Current status, trends and challenges. In: 20th Portuguese Conference on Pattern Recognition, RecPad. pp. 57–58.

Soviany, P., Ionescu, R.T., Rota, P., Sebe, N., 2021. Curriculum self-paced learning for cross-domain object detection. Comput. Vis. Image Underst. 204, 103166.

Stewart, R., Andriluka, M., Ng, A.Y., 2016. End-to-end people detection in crowded scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2325–2333.

Su, Y.-C., Chiu, T.-H., Yeh, C.-Y., Huang, H.-F., Hsu, W.H., 2014. Transfer learning for video recognition with scarce training data for deep convolutional neural network. arXiv preprint arXiv:1409.4127.

Sultani, W., Chen, C., Shah, M., 2018. Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6479–6488.

Sun, S., Liu, Y., Mao, L., 2019. Multi-view learning for visual violence recognition with maximum entropy discrimination and deep features. Inf. Fusion 50, 43–53.

Sun, G., Liu, Z., Wen, L., Shi, J., Xu, C., 2021. Anomaly crossing: A new method for video anomaly detection as cross-domain few-shot learning. arXiv preprint arXiv:2112.06320.

Sun, S., Shi, H., Wu, Y., 2015. A survey of multi-source domain adaptation. Inf. Fusion 24, 84–92.

Suresh, A.J., Visumathi, J., 2020. Inception ResNet deep transfer learning model for human action recognition using LSTM. Mater. Today: Proc..

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826.

Taigman, Y., Polyak, A., Wolf, L., 2016. Unsupervised cross-domain image generation. arXiv preprint arXiv:1611.02200.

Tan, M., Le, Q.V., 2019. Mixconv: Mixed depthwise convolutional kernels. arXiv preprint arXiv:1907.09595.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C., 2018. A survey on deep transfer learning. In: International Conference on Artificial Neural Networks. Springer, pp. 270–279.
Tong, F., Zhang, Z., Wang, H., Wang, Y., 2018. Concise convolutional neural network for crowd counting. In: 2018 10th International Conference on Advanced Infocomm Technology. ICAIT, IEEE, pp. 174–178.
Tran, L., Sohn, K., Yu, X., Liu, X., Chandraker, M., 2019. Gotta adapt'em all: Joint pixel and feature-level domain adaptation for recognition in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2672–2681.
Triess, L.T., Dreissig, M., Rist, C.B., Zöllner, J.M., 2021. A survey on deep domain adaptation for lidar perception. In: 2021 IEEE Intelligent Vehicles Symposium Workshops. IV Workshops, IEEE, pp. 350–357.
Tzeng, E., Hoffman, J., Saenko, K., Darrell, T., 2017. Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7167–7176.
Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T., 2014. Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474.
Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C., 2017. Learning from synthetic humans. In: CVPR.
Vincent, V., Wannes, M., Jesse, D., 2020. Transfer learning for anomaly detection through localized and unsupervised instance selection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, (no. 04), pp. 6054–6061.
Walambe, R., Marathe, A., Kotecha, K., 2021. Multiscale object detection from drone imagery using ensemble transfer learning. Drones 5 (3), 66.
Wan, Z., Yang, R., Huang, M., Zeng, N., Liu, X., 2021. A review on transfer learning in EEG signal analysis. Neurocomputing 421, 1–14.
Wang, Q., Breckon, T.P., 2022. Crowd counting via segmentation guided attention networks and curriculum loss. IEEE Trans. Intell. Transp. Syst..
Wang, J., Chen, Y., Feng, W., Yu, H., Huang, M., Yang, Q., 2020a. Transfer learning with dynamic distribution adaptation. ACM Trans. Intell. Syst. Technol. 11 (1), 1–25.
Wang, J., Chen, Y., Hu, L., Peng, X., Philip, S.Y., 2018a. Stratified transfer learning for cross-domain activity recognition. In: 2018 IEEE International Conference on Pervasive Computing and Communications. PerCom, IEEE, pp. 1–10.
Wang, T., Chen, J., Snoussi, H., 2013. Online detection of abnormal events in video streams. J. Electr. Comput. Eng. 2013.
Wang, Z., Dai, Z., Póczos, B., Carbonell, J., 2019a. Characterizing and avoiding negative transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11293–11302.
Wang, L., Ding, Z., Tao, Z., Liu, Y., Fu, Y., 2019b. Generative multi-view human action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6212–6221.
Wang, Q., Gao, J., Lin, W., Li, X., 2020b. NWPU-crowd: A large-scale benchmark for crowd counting and localization. IEEE Trans. Pattern Anal. Mach. Intell. 43 (6), 2141–2149.
Wang, Q., Gao, J., Lin, W., Yuan, Y., 2019c. Learning from synthetic data for crowd counting in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8198–8207.
Wang, Q., Gao, J., Lin, W., Yuan, Y., 2021a. Pixel-wise crowd understanding via synthetic data. Int. J. Comput. Vis. 129 (1), 225–245.
Wang, P., Gao, C., Wang, Y., Li, H., Gao, Y., 2020c. MobileCount: An efficient encoder-decoder framework for real-time crowd counting. Neurocomputing 407, 292–299.
Wang, Q., Han, T., Gao, J., Yuan, Y., 2021b. Neuron linear transformation: Modeling the domain shift for crowd counting. IEEE Trans. Neural Netw. Learn. Syst..
Wang, L., Shao, W., Lu, Y., Ye, H., Pu, J., Zheng, Y., 2018b. Crowd counting with density adaption networks. arXiv preprint arXiv:1806.10040.
Wang, L., Shi, J., Song, G., Shen, I.-f., et al., 2007. Object detection combining recognition and segmentation. In: Asian Conference on Computer Vision. Springer, pp. 189–199.
Wang, C., Yang, G., Papanastasiou, G., Tsaftaris, S.A., Newby, D.E., Gray, C., Macnaught, G., MacGillivray, T.J., 2021c. DiCyc: GAN-based deformation invariant cross-domain information fusion for medical image synthesis. Inf. Fusion 67, 147–160.
Wang, P., Zhang, J., Zhu, H., 2021d. Fire detection in video surveillance using superpixel-based region proposal and ESE-ShuffleNet. Multimedia Tools Appl. 1–28.
Wei, P., Ke, Y., Goh, C.K., 2018a. A general domain specific feature transfer framework for hybrid domain adaptation. IEEE Trans. Knowl. Data Eng. 31 (8), 1440–1451.
Wei, H., Kehtarnavaz, N., 2019. Semi-supervised faster RCNN-based person detection and load classification for far field video surveillance. Mach. Learn. Knowl. Extr. 1 (3), 756–767.
Wei, H., Laszewski, M., Kehtarnavaz, N., 2018b. Deep learning-based person detection and classification for far field video surveillance. In: 2018 IEEE 13th Dallas Circuits and Systems Conference. DCAS, IEEE, pp. 1–4.
Weiss, K., Khoshgoftaar, T.M., Wang, D., 2016. A survey of transfer learning. J. Big Data 3 (1), 1–40.
Wilie, B., Cahyawijaya, S., Adiprawita, W., 2018. CountNet: End to end deep learning for crowd counting. In: 2018 5th International Conference on Electrical Engineering, Computer Science and Informatics. EECSI, IEEE, pp. 128–132.
Wu, C., Shao, S., Tunc, C., Satam, P., Hariri, S., 2022a. An explainable and efficient deep learning framework for video anomaly detection. Cluster Comput. 25 (4), 2715–2737.
Wu, H., Song, C., Yue, S., Wang, Z., Xiao, J., Liu, Y., 2022b. Dynamic video mix-up for cross-domain action recognition. Neurocomputing 471, 358–368.
Wu, Q., Wan, J., Chan, A.B., 2021. Dynamic momentum adaptation for zero-shot cross-domain crowd counting. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 658–666.
Wu, Q., Wu, H., Zhou, X., Tan, M., Xu, Y., Yan, Y., Hao, T., 2017. Online transfer learning with multiple homogeneous or heterogeneous sources. IEEE Trans. Knowl. Data Eng. 29 (7), 1494–1507.
Xiao, Y., Liu, B., Yu, P.S., Hao, Z., 2015. A robust one-class transfer learning method with uncertain data. Knowl. Inf. Syst. 44 (2), 407–438.
Xie, H., Du, Y., Yu, H., Chang, Y., Xu, Z., Tang, Y., 2018. Open set face recognition with deep transfer learning and extreme value statistics. Int. J. Wavelets, Multiresolut. Inf. Process. 16 (04), 1850034.
Xiong, F., Shi, X., Yeung, D.-Y., 2017. Spatiotemporal modeling for crowd counting in videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5151–5159.
Xu, R., Chen, Z., Zuo, W., Yan, J., Lin, L., 2018. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3964–3973.
Xu, D., Ouyang, W., Ricci, E., Wang, X., Sebe, N., 2017. Learning cross-modal deep representations for robust pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5363–5371.
Xu, C., Qiu, K., Fu, J., Bai, S., Xu, Y., Bai, X., 2019. Learn to scale: Generating multipolar normalized density maps for crowd counting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8382–8390.
Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., Zuo, W., 2017. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2272–2281.
Yan, Z., Sun, L., Duckctr, T., Bellotto, N., 2018. Multisensor online transfer learning for 3d lidar-based human detection with a mobile robot. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 7635–7640.
Yang, L., Balaji, Y., Lim, S.-N., Shrivastava, A., 2020. Curriculum manager for source selection in multi-source domain adaptation. In: European Conference on Computer Vision. Springer, pp. 608–624.
Yang, B., Cao, J.-M., Wang, N., Zhang, Y.-Y., Cui, G.-Z., 2018. Cross-scene counting based on domain adaptation-extreme learning machine. IEEE Access 6, 17029–17038.
Yang, B., Lee, C.-G., Lei, Y., Li, N., Lu, N., 2021. Deep partial transfer learning network: A method to selectively transfer diagnostic knowledge across related machines. Mech. Syst. Signal Process. 156, 107618.
Yao, H., Han, K., Wan, W., Hou, L., 2017. Deep spatial regression model for image crowd counting. arXiv preprint arXiv:1710.09757.
Yi, D., Lei, Z., Liao, S., Li, S.Z., 2014. Learning face representation from scratch. arXiv preprint arXiv:1411.7923.
Yi, Z., Zhang, H., Tan, P., Gong, M., 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2849–2857.
Yu, Z., Wang, G., Chen, L., Raschka, S., Luo, J., 2021. Few-shot learning for video object detection in a transfer-learning scheme. arXiv preprint arXiv:2103.14724.
Yu, X., Wang, J., Hong, Q.-Q., Teku, R., Wang, S.-H., Zhang, Y.-D., 2022. Transfer learning for medical images analyses: A survey. Neurocomputing 489, 230–254.
Yu, F., Wu, X., Chen, J., Duan, L., 2019. Exploiting images for video recognition: Heterogeneous feature augmentation via symmetric adversarial learning. IEEE Trans. Image Process. 28 (11), 5308–5321.
Yuan, Y., Zhao, Y., Wang, Q., 2018. Action recognition using spatial-optical data organization and sequential learning framework. Neurocomputing 315, 221–233.
Zaidi, S.S.A., Ansari, M.S., Aslam, A., Kanwal, N., Asghar, M., Lee, B., 2022. A survey of modern deep learning based object detection models. Digit. Signal Process. 103514.
Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. Springer, pp. 818–833.
Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R., 2010. Deconvolutional networks. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, pp. 2528–2535.
Zeng, X., Wu, Y., Hu, S., Wang, R., Ye, Y., 2020. DSPNet: Deep scale purifier network for dense crowd counting. Expert Syst. Appl. 141, 112977.
Zhang, Q., Chan, A.B., 2019. Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8297–8306.
Zhang, L., Germain, P., Kessaci, Y., Biernacki, C., 2022a. Interpretable domain adaptation for hidden subdomain alignment in the context of pre-trained source models. In: 36th AAAI Conférence on Artificial Intelligence.
Zhang, D., Li, J., Li, X., Du, Z., Xiong, L., Ye, M., 2021a. Local–global attentive adaptation for object detection. Eng. Appl. Artif. Intell. 100, 104208.
Zhang, W., Li, X., Ma, H., Luo, Z., Li, X., 2021b. Open-set domain adaptation in machinery fault diagnostics using instance-level weighted adversarial learning. IEEE Trans. Ind. Inform. 17 (11), 7445–7455.

Zhang, W., Li, X., Ma, H., Luo, Z., Li, X., 2021c. Universal domain adaptation in fault diagnostics with hybrid weighted deep adversarial learning. IEEE Trans. Ind. Inform. 17 (12), 7957–7967.

Zhang, C., Li, G., Su, L., Zhang, W., Huang, Q., 2020a. Video anomaly detection using open data filter and domain adaptation. In: 2020 IEEE International Conference on Visual Communications and Image Processing. VCIP, IEEE, pp. 395–398.

Zhang, C., Li, H., Wang, X., Yang, X., 2015. Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 833–841.

Zhang, Q., Lin, W., Chan, A.B., 2021d. Cross-view cross-scene multi-view crowd counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 557–567.

Zhang, D., Ye, M., Liu, Y., Xiong, L., Zhou, L., 2022b. Multi-source unsupervised domain adaptation for object detection. Inf. Fusion 78, 138–148.

Zhang, X., Zhang, H., Song, Z., 2021e. Feature-aligned stacked autoencoder: A novel semi-supervised deep learning model for pattern classification of industrial faults. IEEE Trans. Artif. Intell..

Zhang, C., Zhao, Q., Wang, Y., 2020b. Hybrid adversarial network for unsupervised domain adaptation. Inform. Sci. 514, 44–55.

Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y., 2016. Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 589–597.

Zhang, X., Zhou, X., Lin, M., Sun, J., 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6848–6856.

ZhanLi, L., JiaWei, Y., 2019. Abnormal behavior recognition based on transfer learning. J. Phys.: Conf. Ser. 1213, 022007.

Zhao, S., Li, B., Xu, P., Keutzer, K., 2020. Multi-source domain adaptation in the deep learning era: A systematic survey. arXiv preprint arXiv:2002.12169.

Zhao, S., Li, B., Yue, X., Gu, Y., Xu, P., Hu, R., Chai, H., Keutzer, K., 2019. Multi-source domain adaptation for semantic segmentation. Adv. Neural Inf. Process. Syst. 32.

Zheng, Y., 2015. Methodologies for cross-domain data fusion: An overview. IEEE Trans. Big Data 1 (1), 16–34.

Zhou, X., Wang, D., Krähenbühl, P., 2019. Objects as points. arXiv preprint arXiv:1904.07850.

Zhou, Y., Yang, J., Li, H., Cao, T., Kung, S.-Y., 2020. Adversarial learning for multiscale crowd counting under complex scenes. IEEE Trans. Cybern. 51 (11), 5423–5432.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232.

Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., Ling, H., 2020. Detection and tracking meet drones challenge. arXiv preprint arXiv:2001.06303.

Zhu, Y., Zhuang, F., Wang, D., 2019. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, (no. 01), pp. 5989–5996.

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

Zhuo, T., Cheng, Z., Zhang, P., Wong, Y., Kankanhalli, M., 2019. Explainable video action reasoning via prior knowledge and state transitions. In: Proceedings of the 27th Acm International Conference on Multimedia. pp. 521–529.

Zou, Z., Qu, X., Zhou, P., Xu, S., Ye, X., Wu, W., Ye, J., 2021. Coarse to fine: Domain adaptive crowd counting via adversarial scoring network. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2185–2194.

Zunino, A., Bargal, S.A., Volpi, R., Sameki, M., Zhang, J., Sclaroff, S., Murino, V., Saenko, K., 2021. Explainable deep classification models for domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3233–3242.