Beyond Supervised Learning: A Computer Vision Perspective

Lovish Chum, Anbumani Subramanian, Vineeth N. Balasubramanian & C. V. Jawahar

Journal of the Indian Institute of Science
A Multidisciplinary Reviews Journal

ISSN 0970-4140

J Indian Inst Sci
DOI 10.1007/s41745-019-0099-3

Springer

J. Indian Inst. Sci.
A Multidisciplinary Reviews Journal
ISSN: 0970-4140 Coden-JIISAD

REVIEW ARTICLE

Beyond Supervised Learning: A Computer Vision Perspective

Lovish Chum[1], Anbumani Subramanian[2], Vineeth N. Balasubramanian[3] and C. V. Jawahar[1]*

Abstract | Fully supervised deep learning-based methods have created a profound impact in various fields of computer science. Compared to classical methods, supervised deep learning-based techniques face scalability issues as they require huge amounts of labeled data and, more significantly, are unable to generalize to multiple domains and tasks. In recent years, a lot of research has been targeted towards addressing these issues within the deep learning community. Although there have been extensive surveys on learning paradigms such as semi-supervised and unsupervised learning, there are a few timely reviews after the emergence of deep learning. In this paper, we provide an overview of the contemporary literature surrounding alternatives to fully supervised learning in the deep learning context. First, we summarize the relevant techniques that fall between the paradigm of supervised and unsupervised learning. Second, we take autonomous navigation as a running example to explain and compare different models. Finally, we highlight some shortcomings of current methods and suggest future directions.

Keywords: *Deep learning, Synthetic data, Domain adaptation, Weakly supervised learning, Few-shot learning, Self-supervised learning*

1 Introduction

Distilling useful information from prior experience is one of the primary research problems in computer science. Past information contained in the training data is extracted as a model and used to predict future outcomes in machine learning. In the past few years, the advent of deep learning techniques has greatly benefited the areas of computer vision, speech, and Natural Language Processing (NLP). However, supervised deep learning-based techniques require a large amount of human-annotated training data to learn an adequate model. Although data have been painstakingly collected and annotated for problems such as image classification[120,186], image captioning[115], instance segmentation[134], visual question answering[81], and other tasks, it is not viable to do so for every domain and task. Particularly, for problems in health care and autonomous navigation, collecting an exhaustive data set is either very expensive or all but impossible.

Even though supervised methods excel at learning from a large quantity of data, results show that they are particularly poor in generalizing the learned knowledge to new task or domain[221]. This is because a majority of learning techniques assume that both the train and test data are sampled from the same distribution. However, when the distributions of the train and test data are different, the performance of the model is known to degrade significantly[201,221]. For instance, take the example of autonomous driving. The roadside environment for a city in Europe is significantly different from a city in South Asia. Hence, a model trained with input video frames from the former suffers a significant degradation in performance when tested on the latter. This is in direct contrast to living

This article belongs to the Special issue–Recent Advances in Machine Learning.

[1] *CVIT, IIIT Hyderabad, Hyderabad, India.*
[2] *Intel, Bangalore, India.*
[3] *IIT Hyderabad, Hyderabad, India.*
**lovish1234@gmail.com*

organisms which perform a wide variety of tasks in different settings without receiving direct supervision[168,237].

This survey is targeted towards summarizing the recent literature that addresses two bottlenecks of fully supervised deep learning methods—(1) lack of labeled data in a particular domain; (2) unavailability of direct supervision for a particular task in a given domain. Broadly, we can categorize the methods which aim to tackle these problems into three sets—(1) data-centric techniques which solve the problem by generating a large amount of data similar to the one present in the original data set; (2) algorithm-centric techniques which tweak the learning method to harness the limited data efficiently through various techniques like on-demand human intervention, exploiting the inherent structure of data, capitalizing on freely available data on the web or solving for an easier but related surrogate task; (3) hybrid techniques which combine ideas from both the data and algorithm-centric methods.

Data-centric techniques include data augmentation which involves tweaking the data samples with some pre-defined transformations to increase the overall size of the data set. For images, this involves affine transformations such as shifting, rotation, shearing, flipping, and distortion of the original image[116]. Some recent papers also advocate adding Gaussian noise to augment the images in the data set. Ratner et al.[171] recommend learning these transforms instead of hard-coding them before training. Another method is to use techniques borrowed from computer graphics to generate synthetic data which is used along with the original data to train the model. In the case when data are in the form of time-series, window slicing and window warping can be used for augmentation purposes[126].

Algorithm-centric techniques try to relax the need of perfectly labeled data by altering the model requirements to acquire supervision through inexact[248], inaccurate[148], and incomplete labels[24]. For most of the tasks, these labels are cheaper and relatively easy to obtain than full-fledged task-pertinent annotations. Techniques involving on-demand human supervision have also been used to label selective instances from the data set[220]. Another set of methods exploit the knowledge gained while learning from a related domain or task by efficiently transferring it to the test environment[189].

Hybrid methods incorporate techniques which focus on improving the performance of the model at both the data and algorithm level. For instance, in urban scene understanding task, researchers often use a synthetically generated data set along with the real data for training. This proves to be greatly beneficial as real-world data set may not cover all the variations encountered during the test time i.e. different lighting conditions, seasons, camera angles etc. However, a model trained using synthetic images suffers a significant decrease in performance when tested on real images due to domain shift. This issue is algorithmically addressed by making the model "adapt" to the real-world scenario[259]. Most of the methods discussed in this survey fall under this category.

In this paper, we discuss some of these methods along with describing their qualitative results. We use tasks associated with autonomous navigation as a case study to explain each paradigm. As a preliminary step, we introduce some common notations used in the paper. We follow this by mentioning the radical improvement brought by supervised deep learning methods in computer vision tasks briefly in Sect. 1.2. Section 2 contains an overview of work which involves the use of synthetic data for training. Various techniques for transfer learning are compared in Sect. 3. Methods for weak and self-supervision are discussed in Sects. 4 and 6, respectively. Methods which address the task of learning an adequate model from a few instances are discussed in Sect. 5. Finally, we conclude the paper discussing the promises, challenges, and open research frontiers beyond supervised learning in Sect. 7. Figure 1 gives a brief overview of the survey in the context of semantic segmentation task for autonomous navigation.

1.1 Notations and Definitions

In this section, we introduce some notations which aid the explanation of the paradigms surveyed in the paper. Let $\mathcal{X}$ and $\mathcal{Y}$ be the input and label space, respectively. In any machine learning problem, we assume to have N objects from which we wish to learn the representation of the data set. We extract features from these objects $X = (x_1, x_2, \ldots, x_N)$ to train our model. Let $P(X)$ be the marginal probability over X. In a fully supervised setting, we also assume to have labels $Y = (y_1, y_2, \ldots, y_N)$ corresponding to each of these feature sets. A learning algorithm seeks to find a function $f : \mathcal{X} \longrightarrow \mathcal{Y}$ in the hypothesis space $\mathcal{F}$. To measure the suitability of the function f, a loss function $l : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}^{\geq 0}$ is defined over space $\mathcal{L}$. A machine learning algorithm tries to minimize the risk R associated with wrong predictions:

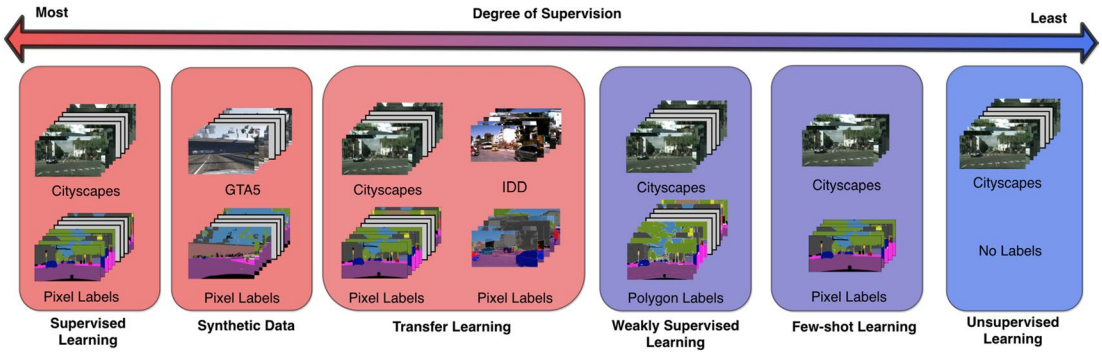


Figure 1: Learning paradigms arranged in decreasing order of supervision signal. Semantic segmentation of outdoor scene is taken as an example task (1) Fully supervised learning requires a lot of annotated data to learn a viable model[35]. (2) Synthetically generated instances can be used to compensate for the lack of real-world data[179]. (3) Knowledge from one real-world data set can be transferred to another data set which does not contain the sufficient amount of instances. For instance, a model trained on Cityscapes can be fine-tuned with the data from the Indian Driving Data set (IDD)[231]. (4) In case pixel-level labels are expensive to obtain, inexact supervision from polygon labels can be exploited to accomplish the task. (5) If only a few instances are available along with their labels, few-shot learning techniques can be employed to learn a generalizable model. (6) Finally, unsupervised learning exploits the inherent structure of the unlabelled data instances.

$$R = \frac{1}{N}\sum_{n=0}^{N} l(y_i, f(x_i)).$$

Use of synthetic data has become mainstream in computer vision literature. Note that even though synthetic data may appear to contain the same entities, we cannot assume that it has been generated from the same distribution. Hence, we denote that it is input space as $\mathcal{X}_{\text{synth}}$ instead of $\mathcal{X}$. However, the label space remains the same. To elaborate, we have a new domain $\mathcal{D}_{\text{synth}} = \{X_{\text{synth}}, P(X_{\text{synth}})\}$ which is different from the real domain $\mathcal{D} = \{X, P(X)\}$ as both their input feature space and marginal distributions are different. Hence, we cannot use the objective predicting function $f_{\text{synth}} : \mathcal{X}_{\text{synth}} \longrightarrow \mathcal{Y}$ for mapping $\mathcal{X}$ to $\mathcal{Y}$.

Transfer learning, a term interchangeably used with domain adaptation (DA), aims to solve this problem. However, the term is not only used to transfer knowledge between different domains but also between distinct tasks. We define a task as containing the label space $\mathcal{Y}$ and the conditional distribution $P(Y|X)$, as $\mathcal{T} = \{\mathcal{Y}, P(Y|X)\}$. Building on the above notations, we define domain shift ($\mathcal{D}_s \neq \mathcal{D}_t$) and label space shift ($\mathcal{T}_s \neq \mathcal{T}_t$), where $\mathcal{D}_s$ and $\mathcal{D}_t$ are source and target domains, respectively. As an example, using synthetic data and then adapting the learned objective to real domain fall under domain shift as $\mathcal{D} \neq \mathcal{D}_{\text{synth}}$. Within the domain adaptation literature, methods have been categorized into homogeneous and heterogeneous settings. Homogeneous domain adaptation methods assume that the input feature space for both the source and target input distribution is the same, i.e., $X_s = X_t$. Heterogeneous domain adaptation techniques relax this assumption. As a result, heterogeneous DA is considered a more challenging problem than homogeneous DA.

Although supervised learning considers that all the feature sets x_i have a corresponding label y_i available at the time of training, the labels can be *inaccurate*, *inexact*, or *incomplete* in a real-world scenario. These scenarios collectively fall under the paradigm of weakly supervised learning. These conditions are particularly true if the training data has been obtained from web. Formally, we define the feature set for incomplete label scenario as $X = (x_1, x_2, \ldots, x_l, x_{l+1}, \ldots x_n)$ where $X_{\text{labeled}} = (x_1, x_2, \ldots, x_l)$ have corresponding labels $Y_{\text{labeled}} = (y_1, y_2, \ldots y_l)$ available while training, but the rest of the feature sets $X_{\text{unlabeled}} = (x_{l+1}, \ldots, x_n)$ do not have any labels associated with them.

Other interesting weakly supervised models encompass cases where each instance has multiple labels or a bag of instances have a single label assigned to it. To formalize for multiple-instance single-label scenario, we assume that each feature set x_i is composed of many sub-feature sets $(x_{i,1}, x_{i,2}, \ldots, x_{i,m})$. Here, x_i is called a "bag" of features and the paradigm is known as multiple-instance learning. A bag is labeled positive if at least one item $x_{i,j}$ is positive otherwise negative.

Although the above paradigms correspond to a varied amount of supervision, they always assume a huge number of instances X available at the time of training the model. This assumption breaks down when some classes do not have sufficient instances.

Few-shot learning entail the scenario when only a few (usually not more than 10) instances per class are available at the time of training. Zero-shot learning (ZSL) is an extreme scenario which arises when no instance is available for some classes during training. Given the training set with features $X = (x_1, x_2, \ldots, x_n)$ and labels $Y_{\text{train}} = (y_1, y_2, \ldots, y_n)$, the test instances belong to previously unseen classes $Y_{\text{test}} = (y_{n+1}, y_{n+2}, \ldots, y_m)$. Recently, some papers address a generalized ZSL scenario where the test classes have both seen or unseen labels.

When no supervision signal is available, the inherent structure of the instances is utilized to train the model. Let X and Y be the feature and label set, respectively; as we do not have $P(Y|X)$, we cannot define the task $\mathcal{T} = \{\mathcal{Y}, P(Y|X)\}$. Instead, we define a proxy task $\mathcal{T}_{\text{proxy}} = \{Z, P(Z|X)\}$ whose label set Z can be extracted within the data itself. For computer vision problems, proxy tasks have been defined based on spatial and temporal alignment, color, and motion cues.

1.2 Success of Supervised Learning

Over the past few years, supervised learning methods have enabled computer vision researchers to train more and more accurate models. For several tasks, these models have achieved state-of-the-art performance which is comparable to humans. In the visual domain, accuracy for both structure and unstructured prediction tasks such as image classification[91,96,116,203,214], object detection[75,76,136,174,178], semantic segmentation[12,27,92,133,138,182,260], pose estimation[23,222], action recognition[46,58,74,104,223], video classification[110], and optical flow estimation[47] has consistently increased allowing for their large-scale deployment. Apart from computer vision, problems in other domains such as speech recognition[82,83,190], speech synthesis[229], machine translation[13,84,213,244], and machine reading[170] have also seen a significant improvement in their performance metrices.

Despite their success, supervised learning-based models have a fair share of issues. First of all, they are data hungry requiring a huge amount of instance-label pairs. To add, a majority of large data sets required to train these models are proprietary as they provide an advantage to the owner in training a supervised model for a particular task and domain. Second, when applying a machine learning model in the wild, it encounters a multitude of conditions which are not observed in the training data. In these situations, fully supervised methods, despite the super-human-level performance on a particular domain suffer drastic degradation in performance on a real-world test set as they are biased towards the training data set.

2 Effectiveness of Synthetic Data

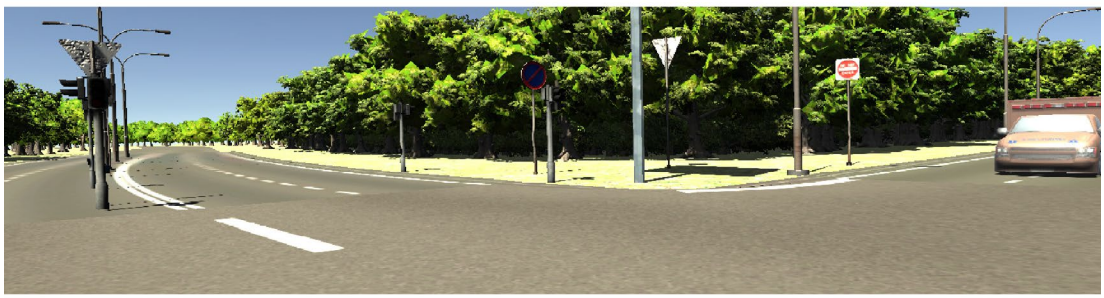
A much better degree of photo-realism, easy-to-use graphics tools such as game engines, large libraries of 3D models, and appropriate hardware have made it is possible to simulate virtual visual environments which can be used to construct synthetic data sets which are exponentially larger than real-world data sets. One primary advantage of using synthetic data is that the precise ground truth is often available for free. On the other hand, collecting and annotating data for a large number of problems is not only a tedious process but also prone to human errors. To add, one can easily vary factors such as viewpoint, lighting, and material properties earning full control over configurations and visual challenges to be introduced in the data set. This presents a major advantage for computer vision researchers as real-world data sets tend to be non-exhaustive, redundant, heavily biased, and partly representative of the complexity of natural images[221]. Moreover, some situations are not possible to be arranged in a real-world setting because of safety issues, e.g., a head-on collision in an urban scene understanding data set. Last but not least, having a few high-profile real-world data sets bias the research community towards the tasks for which annotations have been provided with these data sets. Thus, graphically generated synthetic data sets have become a norm in the computer vision community, particularly for tasks such as medical imaging and autonomous navigation.

In the visual domain, synthetic data have been used mainly for two purposes: (1) evaluation of the generalizability of the model due to the large variability of synthetic test examples, and (2) aiding the training through data augmentation for tasks where it is difficult to obtain ground truth, e.g., optical flow or depth perception. A virtual test bed for design and evaluation of surveillance systems is proposed in Taylor et al.[217]. Kaneva et al.[108] and Aubry and Russell[10] use synthetic data to evaluate hand-crafted and deep features,

respectively. Butler et al.[22] propose MPI Sintel Flow data set, a synthetic benchmark for optical flow estimation. Handa et al.[88] introduce ICL-NUIM, a data set for evaluation of visual odometry.

More significantly, synthetic data are utilized for gathering additional training instances, mainly beneficial due to the availability of precise ground truth. There are various data generation strategies, from real-world images combined with 3D models to full rendering of dynamic visual scenes. Figure 2 illustrates two common methods for synthetic data generation. Vaquez et al.[232] learn the appearance models of pedestrians in a virtual world and use the learned model for detection in the real-world scenario. A similar technique is described for pose estimation[11,162], indoor scene understanding[89], action recognition[41], and variety of other tasks. Instead of rendering the entire scene, Gupta et al.[85] overlay text on natural images consistent with the local 3D scene geometry to generate data for text localization task. A similar method is used for object detection[52] and semantic segmentation[177] where real images of both the objects and backgrounds are composed to synthetically generate a new scene. One drawback of using synthetic data for training a model is that it gives rise to "sim2real" domain gap. Recently, a stream of works in domain randomization[188,219,224] claims

(a) Example image from KITTI dataset[71]



(b) Example image from Virtual KITTI dataset[64]

(c) Real cars augmented to the KITTI dataset[5]

Figure 2: Data collected in real-world setting may not have sufficient diversity in terms of illumination, viewpoints, etc. Synthetic data produced through virtual visual models help to get around this bottleneck. Another way to create additional data for training is to paste real or virtual objects to real scenes. One advantage of this approach is that the domain gap between real and synthetically generated data is lesser leading to better performance on the real data set.

to generate synthetic data with sufficient variations, such that the model views real data as just another variation of the synthetic data set.

Modern game engines are a popular method to extract synthetic data along with the annotation due to their photo-realism and realistic physics simulation. Gaidon et al.[64] present the Virtual KITTI data set and conduct experiments on multi-object tracking. SYNTHIA[183] and GTA [179] provide urban scene understanding data along with semantic segmentation benchmarks. UnrealCV[167] provides a simple interface for researchers to build a virtual world without worrying about the game's API.

Synthetic data for Autonomous Navigation

Autonomous Navigation has greatly benefited from the use of synthetic data sets as pixel-level ground truth can be obtained easily and cheaply using label propagation from frame to frame. As a result, several synthetic data sets have been curated particularly for visual tasks pertaining to autonomous navigation[64,129,179,180,183,191]. Alhaija et al.[5] propose a method to augment virtual objects to real road scene for creating additional data to be used during training the model. Apart from training the models, racing simulators have also been used to evaluate the performance of different approaches to autonomous navigation[26,48]. Janai et al.[102] offer a comprehensive survey of literature pertinent to autonomous driving

One of the major challenges in using synthetic data for training is the domain gap between real and synthetic data sets. Transfer learning discussed in Sect. 3 offers a solution to this problem. Eventually, through the use of synthetic data, we would like to replace the expensive data acquisition process and manual labeling of ground truth into a generic problem of training with unlimited computer-generated data and testing in the real-world scenario without any degradation in performance.

3 Domain Adaptation and Transfer Learning

As stated in Sect. 2, a model trained on source domain does not perform well on a target domain with different distribution. Domain adaptation (DA) is a technique which addresses this issue by reusing the knowledge gained through the source domain for the target domain. DA techniques have been categorized according to three criteria: (1) distance between domains; (2) presence of supervision in the source and target domain; (3) type of domain divergences. Most of the DA techniques assume that the source and target domain are "nearer" to each other, in the sense that the instances are directly related. In these cases, single-step adaptation is sufficient to align both the domains. However, if this assumption does not hold true, multi-step adaptation is used where a set of intermediate domains is used to align the source and target domains. Prevalent literature also classifies DA in supervised, semi-supervised, and unsupervised setting according to the presence of labels in source and target domain. Nevertheless, there are inconsistencies in the definition within the literature; while some papers refer to the absence of target labels as unsupervised DA, others define it as an absence of both the source and target labels. Hence, in this section, we categorize the DA techniques with respect to the type of domain divergences. Section 1.1 gives out the formal notation and formulations for DA setting.

Earlier works categorized the domain adaptation problem into homogeneous and heterogeneous settings. Homogeneous domain adaptation deals with the situation when both the source and target domains share a common feature space $\mathcal{X}$ but different data distributions $P(X)$ or $P(Y|X)$. Some traditional methods for homogeneous domain adaptation include instance re-weighting[25], feature transformations[39,97], or kernel-based techniques that learn an explicit transform from source to target domain[50,78,154]. Figure 3 pictorially presents the traditional domain adaptation methods. All the techniques addressing this problem aim to correct the differences

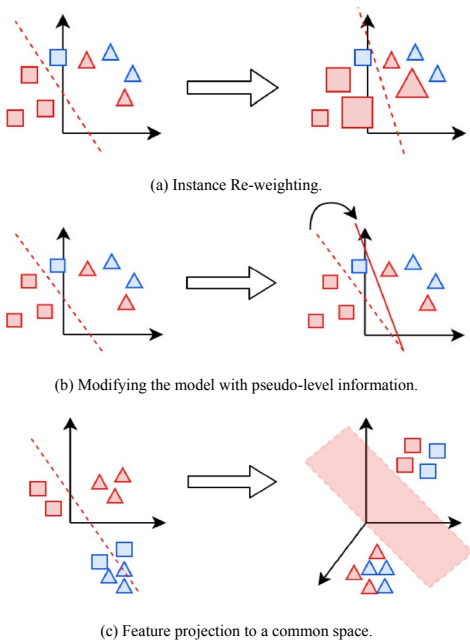


Figure 3: Conventional techniques for domain adaptation. The original model is trained to classify ■ and ▲. However, it is able to classify ■ and ▲ only after applying appropriate DA techniques.

between conditional and marginal distributions between the source and target domain. Heterogeneous domain adaptation pertains to the condition when the source and target domains are represented in different feature space. This is particularly important for problems in the visual domain such as image recognition[80,117,263], object detection, semantic segmentation[128], and face recognition as different environments, background, illumination, viewpoint, sensor, or post-processing can cause a shift between the train and test distributions. Moreover, a difference between the tasks also demands the model to be adapted to the target domain task. Manifold alignment[238] and feature augmentation[49,130] are some of the techniques used for aligning feature spaces in heterogeneous adaptation. A detailed survey of traditional adaptation techniques is out of the scope of this survey. We direct readers to Ben-David et al.[16] and Pan et al.[153] for a summary of homogeneous and Day and Khoshgoftaar[40] and Weiss et al.[242] for a detailed overview of heterogeneous adaptation techniques. Patel et al.[158], Shao et al.[199], and Csurka[37] provide an overview of shallow domain adaptation methods on visual tasks. In this paper, we briefly state recent advances in deep domain adaptation techniques pertaining computer vision tasks.

Taking a cue from the success of deep neural networks for learning a feature representation, recent DA methods use them to learn representations invariant to the domain; thus inserting the DA framework within the deep learning pipeline. Earlier work using deep neural networks only used the features extracted from the deep network for feature augmentation[149] or subspace alignment[139,169] of two distinct visual domains. Although these methods perform better than state-of-the-art traditional DA techniques, they do not leverage neural networks to directly learn a semantically meaningful and domain-invariant representation.

Contemporary methods use discrepancy-based or adversarial approaches for domain adaptation. Discrepancy-based methods posit that fine-tuning a deep network with target domain data can alleviate the shift between domain distributions[45,151,253]. Labels or attribute information[70,227], Maximum Mean Discrepancy (MMD)[226,249], correlation alignment[212], statistical associations[87], and batch normalization[131] are some of the criterion used while fine-tuning the model.

Adversarial methods encompass a framework which consists of a label classifier trained adversarially to the domain classifier. This formulation aids the network in learning features which are discriminative with respect to the learning task but indiscriminate with respect to the domain. Ganin et al.[68] introduced DANN architecture which uses a gradient reversal layer to ensure that feature distributions over the two domains are aligned. Liu and Tuzel[136] introduce a GAN-based framework in which the generator tries to convert the source domain instances to those from the target domain and the discriminator tries to distinguish between transformed source and target domain instances. Bousmalis et al.[20], Hoffman et al.[95], Shrivastava et al.[202] and Yoo et al.[252] also focus on generating synthetic target data using adversarial loss, albeit using it in pixel space instead of embedding space. Sankaranarayanan et al.[193] use a GAN only to obtain the gradient information for learning a domain-invariant embedding, noting that successful domain alignment does not strictly depend on image generation. Tzeng et al.[228] propose a unified framework for adversarial methods summarizing the type of adversary, loss function, and weight sharing constraint to be used during training.

Generative Adversarial Network (GAN)

GAN[79] consists of two neural networks; a generator that creates samples using noise and a discriminator which receives samples from both the generator and real data set and classifies them. The two networks are trained simultaneously with the intention that the generated samples are indistinguishable from real data at equilibrium. Apart from producing images, text, sound, and other forms of structured data, GANs have been instrumental in driving research in machine learning; particularly in the cases where data availability is limited. Data augmentation[7,62] using GANs has resulted in higher performing models than those which use affine transformations. Adversarial adaptation, a paradigm inspired by GAN framework, is used to transfer the data from the source to the target domain. Other notable applications of GANs include data manipulation[140], adversarial training[119], anomaly detection[195], and adversarial cryptography[1]

Reconstruction-based techniques try to construct a shared representation between the source and target domains while maintaining the individual characteristics of both the domains intact. Ghifary et al.[72] use an encoder which is trained simultaneously to accomplish source-label prediction along with target data reconstruction. Bousmalis et al.[19] train separate encoders to account for domain-specific and domain-invariant features. In addition, it uses domain-invariant features for classification while using both kinds of features for reconstruction. Methods based on adversarial reconstruction are proposed in Kim et al.[112], Russo et al.[187], Yi et al.[251], Zhu et al.[262] which use a cyclic consistency loss as the

reconstruction loss along with the adversarial loss to align two different domains.

Optimal transport is yet another technique used for deep DA[38,173]. Courty et al.[36] assign pseudo-labels to the target data using the source classifier. Furthermore, they transport the source data points to the target distribution minimizing the distance traveled and changes in labels while moving the points.

Visual adaptation has been studied for problems such as cross-modal face recognition[137,207], object detection[31,94], semantic segmentation[32,225,259], person re-identification[42], and image captioning[29]. Although deep DA has achieved considerable improvement over the traditional techniques, much of the work in the visual domain has focused on addressing homogeneous DA problems. Recently, heterogeneous domain adaptation problems such as face-to-emoji[215] and text-to-image synthesis[176,254] have also been addressed using adversarial adaptation techniques. Another interesting direction of work pertains open set DA[21,23,255] which loosens the assumption that output sets of both the source and target class must exactly be the same. Tan et al.[216] address the problem of distant domain supervision transferring the knowledge from source to target via intermediate domains. An in-depth survey of deep domain adaptation techniques is presented in Wang and Deng[239].

4 Weakly Supervised Learning

Weakly supervised learning is an umbrella term covering the predictive models which are trained under incomplete, inexact, or inaccurate labels. Incomplete supervision encompasses the situation when the annotation is only available for a subset of training data. As an example, take the problem of image classification with the ground truth being provided through human annotation. Although it is possible to get a huge number of images from the internet, only a subset of these images can be annotated due to the cost associated with labeling. Inexact supervision pertains to the use of related, often coarse-level annotations. For instance, a fully supervised object localization requires to delineate the bounding boxes; however, usually, we only have image-level labels. Finally, noisy or non-ground truth labels can be categorized as inaccurate supervision. Collaborative image tags on social media websites can be considered as noisy supervision. Apart from saving annotation cost and time, weakly supervised methods have proven to be robust to change in the domain during testing.

4.1 Incomplete Supervision

Weakly supervised techniques pertaining incomplete labels make use of either semi-supervised or active learning methods. The conventional semi-supervised approaches include self-training, co-training[18,165], and graph-based methods[51]. A discussion on these is out of the scope of this survey. Interested readers are directed to Chapelle et al.[24] for a detailed overview of semi-supervised learning.

Active learning methods are used in computer vision to reduce labeling efforts in problems such as image annotation[109], recognition[235], object detection[250], segmentation[234], and pose estimation[135]. In this paradigm, unlabeled observations are optimally selected from the data set to query at the training time. For instance, localizing a car occluded by a tree is more difficult than another non-occluded car. Thus, the human annotator could be asked to assign ground truth for the former case which may lead to improved performance for the latter case. A typical active learning pipeline alternates between picking the most relevant unlabeled examples as queries to the oracle and updating the prior on the data distribution with the response[34]. Some common query formulation strategies include maximizing the label change[61], maximizing the diversity of selected samples[53], reducing the expected error of classifier[184], or uncertainty sampling[194]. A survey by Settles[198] gives insight into various active learning techniques.

Although both semi-supervised and active learning techniques have been used to address different problems in the visual domain, there has been an increased interest towards the latter after the emergence of deep learning-based methods. Sener and Savarese[197] and Gal et al.[65] present an effective method to train a CNN using active learning heuristics. An approach to synthesize query examples using GAN is proposed by Bento and EDU[261]. Fang et al.[55] reframe active learning as a reinforcement learning problem. In addition, deep active learning methods have been used to address vision tasks such as object detection in Roy et al.[185].

4.2 Inexact Supervision

Apart from dealing with partially labeled data sets, weakly supervised techniques also help relax the degree of annotation needed to solve a structured prediction problem. Full annotation is tedious and time-consuming process—contemporary vision data sets reflect this fact. For example, in Imagenet[186], while 14 million images are provided

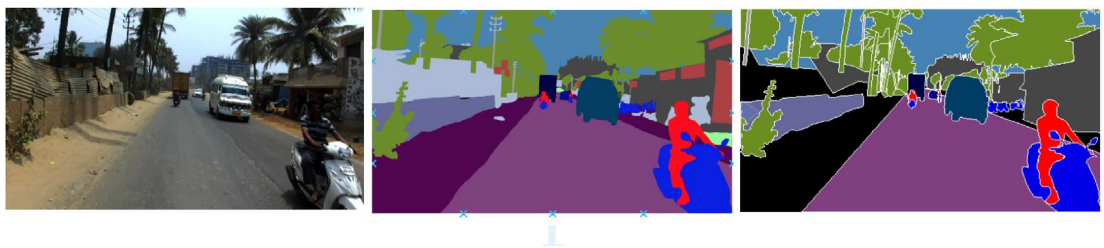
with image-level labels and 500,000 are annotated with bounding boxes; only 4460 images have pixel-level object category labels. Thus, the development of training regimes which learn complex concepts from light labels is instrumental in improving the performance of several tasks.

A popular approach to harness inexact labels is to formulate the problem in multiple-instance learning (MIL) framework. In MIL, the image is interpreted as a bag of patches. If one of the patches within the image contains the object of interest, the image is labeled as a positive instance, otherwise negative. Learning scheme alternates between estimating object appearance model and predicting the patches within positive images. As this setup results in a non-convex optimization objective, several works suggest initialization[209], regularization[208], and curriculum learning[118] techniques to alleviate the issue. Recent works[100,243] embed the MIL framework within a deep neural network to exploit the weak supervision signal.

Structured prediction problems such as weakly supervised object detection (WSOD) and semantic segmentation have garnered a lot of attention in the recent years. Bilen and Vedaldi[17] propose an end-to-end WSOD framework for object detection using image-level labels. Several other techniques have been employed as supervision signal for WSOD such as object size[200] and count[69], click supervision[156,157], and human verification[155]. Similar methods have also been proposed for weakly supervised semantic segmentation problems[15,98,111,132,142,163]. Figure 4 depicts some weak supervision signals used for semantic segmentation problem.

4.3 Inaccurate Supervision

As curating large-scale data sets is an expensive process, building a machine learning model which uses web data sets such as YouTube8m[2], YFCC100M[218], and Sports-1M[110] is one of the pragmatic ways to leverage the almost infinite amount of visual data. However, labels in these data sets are noisy and pose a challenge for the learning algorithm. Several studies have investigated the effect of noisy instances or labels on the performance of the machine learning algorithm. Broadly, we categorize the techniques into two sets—the first approach resorts to treating the noisy instances as outliers and discard them during training[54,211]. Nevertheless, noisy instances may not be outliers and occupy a significant portion of the training data. Moreover, algorithms



(a) Original Image (b) Pixel level supervision[92,28] (c) Polygon around the instances

(d)Scribbles[132] (e) Image-level supervision[163] (f) Supervision via collaborative image tags

Figure 4: An example of the varying degree of supervision for semantic segmentation problem. Although pixel-level labels provide strong supervision, they are relatively expensive to obtain. Thus, the recent literature suggests techniques which exploit polygon labels, scribbles, image-level labels, or even collaborative image tags from social media platforms (note that hashtags are not only inexact but also an inaccurate form of supervision).

pursuing this approach find it difficult to distinguish between noisily labeled and hard training examples. Hence, methods in this set often use a small set of perfectly labeled data. Another stream of methods focus on building algorithms robust to noise[107,146,175,230] by devising noise-tolerant loss functions[73] or adding appropriate regularization terms[9]. For a comprehensive overview of learning algorithms robust to noise, we refer to Frénay and Verleysen[60].

Consequently, a plethora of techniques have been proposed to harness the deep neural networks in a "webly" supervised scenario. As most of the data on the web is contributed by non-experts, it is bound to be inaccurately labeled. Hence, techniques used to address noisy annotations can be applied if the training data are collected from the web. Chen and Gupta[30] propose a two-stage curriculum learning technique on easier examples before adapting it to web images. Xiao et al.[247] predict the type of noise in each of the instances and attempt to remove it. Webly supervised methods have been proposed for many tasks in visual domain such as learning visual concepts[43,67], image classification[233], video recognition[66], and object localization[264].

5 k-Shot Learning

One of the distinguishing characteristics of human visual intelligence is the ability to acquire an understanding of novel concepts from very few examples. Conversely, a majority of current machine learning techniques show a precipitous decrease in performance if there are an insufficient number of labeled examples pertaining to a certain class. Few-shot learning techniques attempt to adapt the current machine learning methods to perform well under a scenario where only a few training instances are available per class. This is of immense practical importance—for instance, collecting a traffic data set might result in only a few instances of auto-rickshaws. However, during testing, we would like the model to recognize auto-rickshaws with various scales, angles and other variations which might not be present in the training set. Earlier methods such as Fei-Fei et al.[57] use Bayesian learning-based generative framework with the assumption that the prior built from previously learned classes can be used to bootstrap learning for novel categories. Lake et al.[121] built a Hierarchical Bayesian model which performs similarly to humans on few-shot alphabet recognition tasks. However, their method is shown to work only for simple data sets such as Omniglot[122]. Wang and Hebert[241] learn to regress from parameters of the classifier trained on a few images to the parameters of the classifier trained on a large number of images. More recent efforts into a few-shot learning techniques can be broadly categorized into metric-learning and meta-learning-based methods.

Metric learning aims to design techniques for embedding the input instances to a feature space beneficial to few-shot settings. A common approach is to find a good similarity metric in the new feature space applicable to novel categories. Koch et al.[114] use a deep learning model based on computing the pairwise distance between the samples based on Siamese networks following which the learned distance is used to solve a few-shot problems through k-nearest-neighbor classification. Vinyals et al.[236] propose an end-to-end trainable one-shot learning technique based on cosine distance. Other loss functions used for deep metric learning include triplet loss Schroff et al.[196] and adaptive density estimation Rippel et al.[181]. Mehrotra and Dukkipati[143] approximate the pairwise distance by training a deep residual network in conjunction with a generative model.

Meta-learning entails a class of approaches which quickly adapt to a new task using only a few data instances and training iterations. To achieve this, the model is trained on a set of tasks, such that it transfers the "learning ability" to a novel task. In other words, meta-learners treat the tasks as training examples. Finn et al.[59] propose a model agnostic meta-learning technique which uses gradient descent to train a classification model such that it is able to generalize well on any novel task given very few instances and training steps. Ravi and Larochelle[172] also introduce a meta-learning framework employing LSTM updates for a given episode. Recently, a method proposed by Mishra et al.[145] also exploits contextual information within the tasks using temporal convolutions.

Another set of methods for few-shot learning relies on efficient regularization techniques to avoid over-fitting on the small number of instances. Hariharan and Girshick[90] suggest a gradient magnitude regularization technique for training a classifier along with a method to hallucinate additional examples for a few-shot classes. Yoo et al.[252] also regularize the dimensionality of parameter search space through efficiently clustering them ensuring the intra-cluster similarity and inter-cluster diversity.

Literature pertaining to Zero-Shot Learning (ZSL) focuses on finding the representation of a novel category without any instance. Although it has a strong semblance to few-shot learning

paradigm, methods used to address ZSL are distinct from few-shot learning. A major assumption taken in this setting is that the classes observed by model during training are semantically related to the unseen classes encountered during testing. This semantic relationship is often captured through class-attributes containing shape, color, pose, etc., of the object which are either labeled by experts or obtained through knowledge sources such as Wikipedia, Flickr, etc. Lampert et al.[123] were first to propose a zero-shot recognition model which assumes independence between different attributes and estimates the test class by combining the attribute prediction probabilities. However, most of the subsequent work takes attributes as the semantic embedding of classes and tackles it as a visual semantic embedding problem[4,56,124,245]. More recently, word embeddings[206,256] and image captions[176] have also been used in place attributes as a semantic space. Figure 5 compares the two common approaches to ZSL with supervised learning.

In ZSL, a joint embedding space is learned during training where both the visual features and semantic vectors are projected. During testing on unseen classes, nearest-neighbor search is performed in this embedding space to match the projection of visual feature vector against a novel object type. A pairwise ranking formula is used to learn parameters of a bi-linear model in Akata et al.[4] and Frome et al.[63]. Recently, Zhang et al.[256] argue to use the visual space as the embedding space to alleviate the hubness problem when performing nearest-neighbor search in semantic space. We refer the readers to Xian et al.[246] for detailed evaluation and comparison of contemporary ZSL methods.

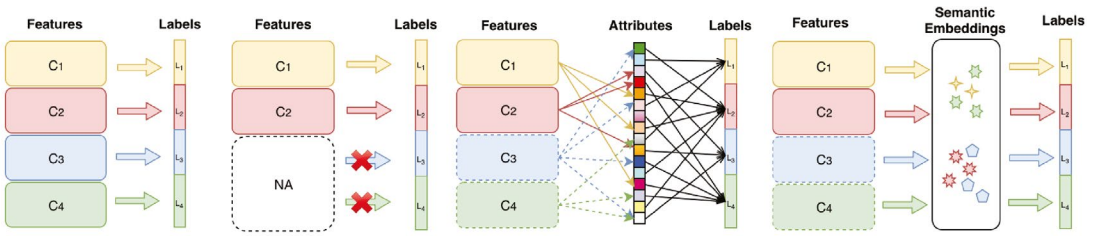
Some other tasks which have shown promising results in a zero-shot setting are video event detection[86], object detection[14], action recognition[166], and machine translation[106].

6 Self-supervised Learning

In self-supervised learning, we obtain feature representation for semantic understanding tasks such as classification, detection, and segmentation without any external supervision. Explicit annotation pertaining to the main task is avoided by defining an auxiliary task that provides a supervisory signal in self-supervised learning. The assumption is that successful training of the model on the auxiliary task will inherently make it learn semantic concepts such as object classes and boundaries. This makes it possible to share knowledge between two tasks. Self-supervision has a semblance to transfer learning where knowledge is shared between two different but related domains. However, unlike transfer learning, it does not require a large amount of annotated data from another domain or task. Figure 6 illustrates the difference between both the paradigms in the context of vehicle detection.

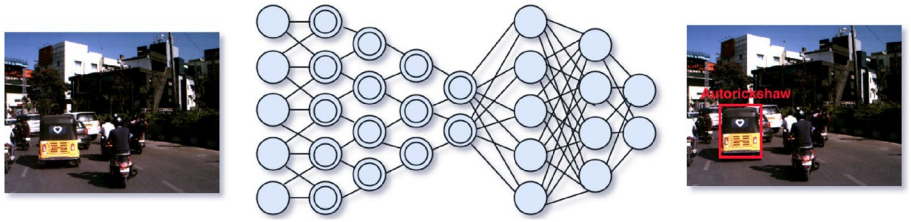
Before the advent of deep learning-driven self-supervision models, a significant work was carried out in unsupervised learning of image representations using hand-crafted[205] or mid-level features[204]. This was followed by deep learning-based methods like autoencoders[93], Boltzmann machines[192], and variational methods[113] which learn by estimating latent parameters which help to reconstruct the data.

The existing literature pertaining self-supervision relies on using the spatial and temporal context of an entity for "free" supervision signal. A prime example of this is Word2Vec[144] which predicts the semantic embedding of a particular word based on the surrounding words. In the visual domain, context is efficiently used by Doersch et al.[44] to predict the relative location of two image patches as a pretext task. The same notion is extended in Noroozi and Favaro[150] by predicting the order of shuffled image patches.

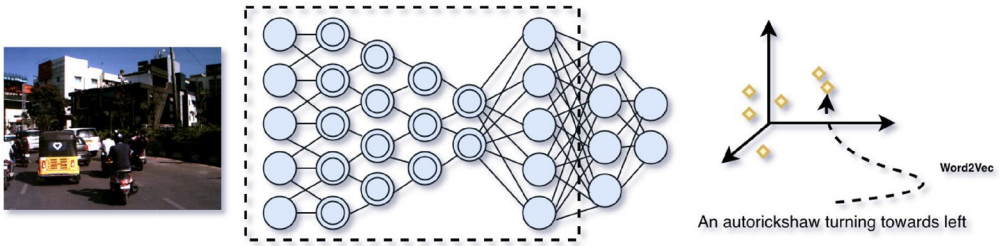


(a) Supervised Learning (b) Zero-Shot Case (c) Direct Attribute Prediction (d) Visual Semantic based Embedding

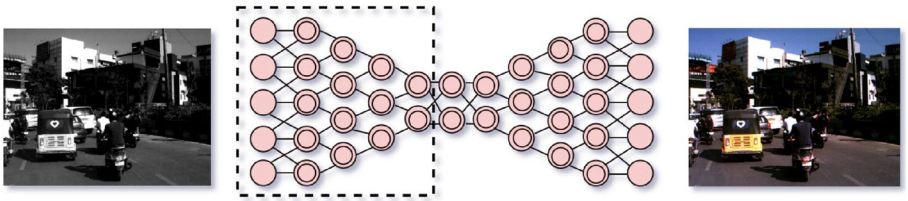
Figure 5: Comparison of supervised learning with ZSL. Features are not available for C_3 and C_4 at the time of training. However, the availability of attributes or semantic embeddings for both the train and test classes aid the training of ZSL framework.



(a) Supervised learning needs the bounding box information corresponding to the object at the time of training.



(b) Weakly Supervised Learning using image level captions use semantic embeddings of the sentece to pre-train the neural network.



(c) Self-supervised Learning uses a pretext task to learn the representation of the objects. In this case, the task is to learn to generate the luminance value of each pixel in the image given the intensity values.

Figure 6: Strong supervision vs. weak supervision vs. self-supervision. ○ and ◎ depict fully connected and convolutional layers, respectively.

Apart from spatial context based auxiliary tasks, predicting color channel from luminance values[125,257] and regressing to a missing patch in an image using generative models[159] have also been used to learn useful semantic information in images. Other modalities used for feature learning in images include text[77], motion[160,164], and cross-channel prediction[258]. Recently, Huh et al.[99] take advantage of EXIF metadata embedded in the image as a supervisory signal to determine if it has been formed by splicing different images.

For videos, temporal coherence serves as an intrinsic underlying structure: two consecutive image frames are likely to contain semantically similar content. Each object within the frame is expected to undergo some transformations in the subsequent frames. Wang and Gupta[240] use relationships between the triplet of image patches obtained from tracking. Misra et al.[147] train a network to guess whether a given sequence of frames from a video are in chronological order. Lee et al.[127] make the network predict the correct sequence of frames given a shuffled set. Apart from temporal context, estimating camera motion[103], ego-motion[3], and predicting the statistics of ambient sound[8,152] have also been used as a proxy task for video representation learning.

Self-supervision for Urban Scene Understanding

As solving autonomous navigation takes centre stage in both vision and robotics community, urban scene understanding has become a problem of utmost interest. More often than not, annotating each frame for training is a tedious job. As self-supervision gives the flexibility to define an implicit proxy task which may or may not require annotation, it is one of the preferred methods for addressing problems such as urban scene understanding. Earlier work in this area includes Stavens and Thrun[210] where authors estimate the terrain roughness based on the "shocks" which the vehicle receives while passing over it. Jiang et al.[105] show that predicting relative depth is an effective proxy task for learning visual representations. Ma et al.[141] propose a multi-modal self-supervised algorithm for depth completion using LiDAR data along with a monocular camera

7 Conclusion and Discussions

In the past decade, computer vision has benefited greatly from the fact that neural networks act as universal approximator of functions. Integrating these networks in the pre-existing machine learning paradigms and optimizing through backpropagation have consistently improved performance for different visual tasks. In this survey paper, we reviewed recent work pertaining to the paradigms which fall between fully supervised and unsupervised learning. Although most of our references lie in the visual domain, the same paradigms have been prevalent in related fields such as NLP, speech, and robotics.

The space between fully supervised and unsupervised learning can be qualitatively divided on the basis of the degree of supervision needed to learn the model. While synthetic data are cost effective and flexible alternative to real-world data sets, the models learned using it still need to be adapted to the real-world setting. Transfer learning techniques address this issue by explicitly aligning different domains through discrepancy-based or adversarial approaches. However, both of these techniques require "strict" annotation pertaining to the task which hinders the generalization capability of the model. Weakly supervised algorithms relax the need of exact supervision by making the learning model tolerant of incomplete, inexact, and inaccurate supervision. This helps the model to harness the huge amount of data available on the web. Even when a particular domain contains an insufficient number of instances, methods in k-shot learning try to build a reasonable model using parameter regularization or meta-learning techniques. Finally, self-supervised techniques completely eliminate the need of annotation as they define a proxy task for which annotation is implicit within the data instances.

These techniques have been successfully applied in both structured and unstructured computer vision applications such as image classification, object localization, semantic segmentation, action recognition, image super-resolution, image caption generation, and visual question answering. Despite their success, recent models weigh heavily on deep neural networks for their performance. Hence they carry both the pros and cons of using these models; cons being lack of interpretability and outcomes which largely depend on hyperparameters. Addressing these topics may attract increasingly more attention in the future.

Some very recent work combines ideas from two or more paradigms to obtain results in a very specialized setting. Peng et al.[161] address the domain adaptation problem when no task-relevant data are present in the target domain. Inoue et al.[101] leverage the full supervision in source and inaccurate supervision in the target domain to perform transfer learning for object localization task.

In the coming years, the other learning paradigms inspired by human reasoning and abstraction such as meta-learning[6,59], lifelong learning[33], and evolutionary methods may also provide interesting avenues in research. We hope that this survey helps researchers by easing the understanding of the field and encourage research in the field.

Publisher's Note

Received: 2 December 2018 Accepted: 16 January 2019
Published online: 18 February 2019

References

1. Abadi M, Andersen DG (2016) Learning to protect communications with adversarial neural cryptography. CoRR. arXiv:1610.06918
2. Abu-El-Haija S, Kothari N, Lee J, Natsev AP, Toderici G, Varadarajan B, Vijayanarasimhan S (2016) Youtube-8m: a large-scale video classification benchmark. arXiv :1609.08675v1
3. Agrawal P, Carreira J, Malik J (2015) Learning to see by moving. In: International conference on computer vision (CVPR), Boston, MA, USA
4. Akata Z, Perronnin F, Harchaoui Z, Schmid C (2013) Label-embedding for attribute-based classification. In: Computer vision and pattern recognition (CVPR), Portland, OR, USA
5. Alhaija H, Mustikovela S, Mescheder L, Geiger A, Rother C (2018) Augmented reality meets computer vision: efficient data generation for urban driving scenes. Int J Comput Vis 126(9):961–972
6. Andrychowicz M, Denil M, Gomez S, Hoffman MW, Pfau D, Schaul T, Shillingford B, De Freitas N (2016) Learning to learn by gradient descent by gradient descent. In: Advances in neural information processing systems (NIPS), Barcelona, Spain
7. Antoniou A, Storkey A, Edwards H (2017) Data augmentation generative adversarial networks. CoRR. arXiv:1711.04340
8. Arandjelovic R, Zisserman A (2017) Look, listen and learn. In: International conference on computervision (ICCV), Venice, Italy
9. Arpit D, Jastrzębskis S, Ballas N, Krueger D, Bengio E, Kanwal MS, Maharaj T, Fischer A, Courville A, Bengio

Y, et al (2017) A closer look at memorization in deep networks. In: International conference on machine learning (ICML), Sydney, Australia

10. Aubry M, Russell BC (2015) Understanding deep features with computer-generated imagery. In:International conference on computer vision (ICCV), Santiago, Chile
11. Aubry M, Maturana D, Efros AA, Russell BC, Sivic J (2014) Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of cad models. In: Computer vision and pattern recognition (CVPR), Columbus, OH, USA
12. Badrinarayanan V, Kendall A, Cipolla R (2017) SegNet: a deep convolutional encoder-decoder architecture for image segmentation. Trans Pattern Anal Mach Intell 39(12):2481–2495
13. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: International conference on learning representations (ICLR), San Diego, CA, USA
14. Bansal A, Sikka K, Sharma G, Chellappa R, Divakaran A (2018) Zero-shot object detection. In: European conference on computer vision (ECCV), Munich, Germany
15. Bearman A, Russakovsky O, Ferrari V, Fei-Fei L (2016) Whats the point: Semantic segmentation with pointsupervision. In: European conference on computer vision (ECCV), Amsterdam, Netherlands
16. Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW (2010) A theory of learning from different domains. Mach Learn 79(1–2):151–175
17. Bilen H, Vedaldi A (2016) Weakly supervised deep detection networks. In: Computer vision and pattern recognition (CVPR), Las Vegas, NV, USA
18. Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training. In: Computational learning theory (CoLT), Madison, Wisconsin, USA
19. Bousmalis K, Trigeorgis G, Silberman N, Krishnan D, Erhan D (2016) Domain separation networks. In: Advances in neural information processing systems (NIPS), Barcelona, Spain
20. Bousmalis K, Silberman N, Dohan D, Erhan D, Krishnan D (2017) Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
21. Busto PP, Gall J (2017) Open set domain adaptation. In: International conference on computer vision (ICCV), Venice, Italy
22. Butler DJ, Wulff J, Stanley GB, Black MJ (2012) A naturalistic open source movie for optical flow evaluation. In: European conference on computer vision (ECCV), Firenze, Italy
23. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2D pose estimation using part affinity fields. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
24. Chapelle O, Scholkopf B, Zien A (2009) Semi-supervised learning (Chapelle O. et al., eds.; 2006) [book reviews]. IEEE Trans Neural Netw 20(3):542
25. Chattopadhyay R, Sun Q, Fan W, Davidson I, Panchanathan S, Ye J (2012) Multi-source domain adaptation and its application to early detection of fatigue. Trans Knowl Discov Data (TKDD) 6(4):18
26. Chen C, Seff A, Kornhauser A, Xiao J (2015) Deepdriving: learning affordance for direct perception in autonomous driving. In: International conference on computer vision (ICCV), Santiago, Chile
27. Chen LC, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation. CoRR. arXiv:1706.05587
28. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018) DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. Pattern Anal Mach Intell 40(4):834–848
29. Chen TH, Liao YH, Chuang CY, Hsu WT, Fu J, Sun M (2017) Show, adapt and tell: adversarial training of cross-domain image captioner. In: International conference on computer vision (ICCV), Venice, Italy
30. Chen X, Gupta A (2015) Webly supervised learning of convolutional networks. In: International conference on computer vision (ICCV), Santiago, Chile
31. Chen Y, Li W, Sakaridis C, Dai D, Van Gool L (2018) Domain adaptive faster R-CNN for object detection in the wild. In: Computer vision and pattern recognition (CVPR), Salt Lake City, UT, USA
32. Chen YH, Chen WY, Chen YT, Tsai BC, Wang YCF, Sun M (2017) No more discrimination: cross city adaptation of road scene segmenters. In: International conference on computer vision (ICCV), Venice, Italy
33. Chen Z, Liu B (2016) Lifelong machine learning. Synth Lect Artif Intell Mach Learn 10(3):1–145
34. Cohn DA, Ghahramani Z, Jordan MI (1996) Active learning with statistical models. J Artif Intell Res 4:129–145
35. Cordts M, Omran M, Ramos S, Scharwächter T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2015) The cityscapes dataset. In: CVPR workshop on the future of datasets in vision (CVPRW), Boston, MA, USA
36. Courty N, Flamary R, Habrard A, Rakotomamonjy A (2017) Joint distribution optimal transportation for domain adaptation. In: Advances in neural information processing systems (NIPS), Long Beach, CA, USA
37. Csurka G (2017) Domain adaptation for visual applications: a comprehensive survey. CoRR. arXiv:1702.05374
38. Damodaran BB, Kellenberger B, Flamary R, Tuia D, Courty N (2018) Deepjdot: deep joint distribution optimal transport for unsupervised domain adaptation. In: European conference on computer vision (ECCV), Munich, Germany

39. Daumé III H (2007) Frustratingly easy domain adaptation. In: Association of computational linguistics (ACL), Prague, Czech Republic
40. Day O, Khoshgoftaar TM (2017) A survey on heterogeneous transfer learning. J Big Data 4(1):29
41. De Souza CR, Gaidon A, Cabon Y, Peña AML (2017) Procedural generation of videos to train deep action recognition networks. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
42. Deng W, Zheng L, Kang G, Yang Y, Ye Q, Jiao J (2018) Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In: Computer vision and pattern recognition (CVPR), Salt Lake City, UT, USA
43. Divvala SK, Farhadi A, Guestrin C (2014) Learning everything about anything: webly-supervised visual concept learning. In: Computer vision and pattern recognition (CVPR), Columbus, OH, USA
44. Doersch C, Gupta A, Efros AA (2015) Unsupervised visual representation learning by context prediction. In: International conference on computer vision (ICCV), Santiago, Chile
45. Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T (2014) DeCAF: a deep convolutional activation feature for generic visual recognition. In: International conference on machine learning (ICML), Beijing, China
46. Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Computer vision and pattern recognition (CVPR), Boston, MA, USA
47. Dosovitskiy A, Fischer P, Ilg E, Hausser P, Hazirbas C, Golkov V, Van Der Smagt P, Cremers D, Brox T (2015) FlowNet: learning optical flow with convolutional networks. In: International conference on computer vision (ICCV), Santiago, Chile
48. Dosovitskiy A, Ros G, Codevilla F, Lopez A, Koltun V (2017) CARLA: an open urban driving simulator. In: Conference on robot learning (CoRL), Mountain View, California, USA
49. Duan L, Xu D, Tsang I (2011) Learning with augmented features for heterogeneous domain adaptation. In: International conference on machine learning (ICML), Edinburgh, Scotland
50. Duan L, Tsang IW, Xu D (2012) Domain transfer multiple kernel learning. Trans Pattern Anal Mach Intell 34(3):465–479
51. Duchenne O, Audibert JY, Keriven R, Ponce J, Ségonne F (2008) Segmentation by transduction. In: Computer vision and pattern recognition (CVPR), Anchorage, AL, USA
52. Dwibedi D, Misra I, Hebert M (2017) Cut, paste and learn: surprisingly easy synthesis for instance detection. In: International conference on computer vision (ICCV), Venice, Italy
53. Elhamifar E, Sapiro G, Yang A, Shankar Sasrty S (2013) A convex optimization framework for active learning. In: International conference on computer vision (ICCV), Sydney, Australia
54. Fan J, Shen Y, Zhou N, Gao Y (2010) Harvesting large-scale weakly-tagged image databases from the web. In: Computer vision and pattern recognition (CVPR), San Francisco, CA, USA
55. Fang M, Li Y, Cohn T (2017) Learning how to active learn: a deep reinforcement learning approach. In: Association of computational linguistics (ACL), Vancouver, Canada
56. Farhadi A, Endres I, Hoiem D, Forsyth D (2009) Describing objects by their attributes. In: Computer vision and pattern recognition (CVPR), Miami, FL, USA
57. Fei-Fei L, Fergus R, Perona P (2006) One-shot learning of object categories. Trans Pattern Anal Mach Intell 28(4):594–611
58. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In: Computer vision and pattern recognition (CVPR), Las Vegas, NV, USA
59. Finn C, Abbeel P, Levine S (2017) Model-agnostic meta-learning for fast adaptation of deep networks. In: International conference of machine learning (ICML), Sydney, Australia
60. Frénay B, Verleysen M (2014) Classification in the presence of label noise: a survey. Trans Neural Netw Learn Syst 25(5):845–869
61. Freytag A, Rodner E, Denzler J (2014) Selecting influential examples: active learning with expected model output changes. In: European conference on computer vision (ECCV), Zurich, Switzerland
62. Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H (2018) GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. CoRR. arXiv :1803.01229
63. Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Mikolov T et al (2013) Devise: a deep visual-semantic embedding model. In: Advances in neural information processing systems (NIPS), Stateline, NA, USA
64. Gaidon A, Wang Q, Cabon Y, Vig E (2016) Virtual worlds as proxy for multi-object tracking analysis. In: Computer vision and pattern recognition (CVPR), Las Vegas, NV, USA
65. Gal Y, Islam R, Ghahramani Z (2017) Deep Bayesian active learning with image data. In: Advances in neural information processing systems workshops, Long Beach, CA, USA
66. Gan C, Sun C, Duan L, Gong B (2016) Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In: European conference on computer vision (ECCV), Amsterdam, Netherlands

67. Gan C, Yao T, Yang K, Yang Y, Mei T (2016) You lead, we exceed: labor-free video concept learning by jointly exploiting web videos and images. In: Computer vision and pattern recognition (CVPR), Las Vegas, NV, USA
68. 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):2096–2030
69. Gao M, Li A, Yu R, Morariu VI, Davis LS (2018) C-WSL: count-guided weakly supervised localization. In: Europeanconference on computer vision (ECCV), Munich, Germany
70. Gebru T, Hoffman J, Fei-Fei L (2017) Fine-grained recognition in the wild: a multi-task domain adaptation approach. In: International conference on computer vision (ICCV), Venice, Italy
71. Geiger A, Lenz P, Stiller C, Urtasun R (2013) Vision meets robotics: the KITTI dataset. Int J Robot Res 32(11):1231–1237
72. Ghifary M, Kleijn WB, Zhang M, Balduzzi D, Li W (2016) Deep reconstruction-classification networks for unsupervised domain adaptation. In: European conference on computer vision (ECCV), Amsterdam, Netherlands
73. Ghosh A, Kumar H, Sastry P (2017) Robust loss functions under label noise for deep neural networks. In: AAAI, San Francisco, CA, USA
74. Girdhar R, Ramanan D, Gupta A, Sivic J, Russell B (2017) ActionVLAD: learning spatio-temporal aggregation for action classification. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
75. Girshick R (2015) Fast R-CNN. In: International conference on computer vision (ICCV), Santiago, Chile
76. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Computer vision and pattern recognition (CVPR), Columbus, OH, USA
77. Gomez L, Patel Y, Rusiñol M, Karatzas D, Jawahar C (2017) Self-supervised learning of visual features through embedding images into text topic spaces. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
78. Gong B, Shi Y, Sha F, Grauman K (2012) Geodesic flow kernel for unsupervised domain adaptation. In: Computer vision and pattern recognition (CVPR), Providence, RI, USA
79. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems (NIPS), Montreal, Canada
80. Gopalan R, Li R, Chellappa R (2011) Domain adaptation for object recognition: an unsupervised approach. In: International conference on computer vision (ICCV), Barcelona, Spain
81. Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D (2017) Making the V in VQA matter: elevating the role of image understanding in Visual Question Answering. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
82. Graves A (2013) Generating sequences with recurrent neural networks. CoRR. arXiv:1308.0850
83. Graves A, Jaitly N (2014) Towards end-to-end speech recognition with recurrent neural networks. In: International conference on machine learning (ICML), Beijing, China
84. Gu J, Neubig G, Cho K, Li VO (2017) Learning to translate in real-time with neural machine translation. In: Association of computational linguistics (ACL), Vancouver, Canada
85. Gupta A, Vedaldi A, Zisserman A (2016) Synthetic data for text localisation in natural images. In: Computer vision and pattern recognition (CVPR), Las Vegas, NV, USA
86. Habibian A, Mensink T, Snoek CG (2014) Composite concept discovery for zero-shot video event detection. In: International conference on multimedia retrieval (ICMR), Glasgow, UK
87. Haeusser P, Frerix T, Mordvintsev A, Cremers D (2017) Associative domain adaptation. In: International conference on computer vision (ICCV), Venice, Italy
88. Handa A, Whelan T, McDonald J, Davison AJ (2014) A benchmark for RGB-D visual odometry, 3D reconstruction and slam. In: International conference on robotics and automation (ICRA), Hong Kong
89. Handa A, Patraucean V, Badrinarayanan V, Stent S, Cipolla R (2016) Understanding real world indoor scenes with synthetic data. In: Computer vision and pattern recognition (CVPR), Las Vegas, NV, USA
90. Hariharan B, Girshick RB (2017) Low-shot visual recognition by shrinking and hallucinating features. In: International conference on computer vision (ICCV), Venice, Italy
91. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Computer vision and pattern recognition (CVPR), Las Vegas, NV, USA
92. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask R-CNN. In: International conference on computer vision (ICCV), Honolulu, HI, USA
93. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507
94. Hoffman J, Gupta S, Leong J, Guadarrama S, Darrell T (2016) Cross-modal adaptation for RGB-D detection. In: International conference on robotics and automation (ICRA), Stockholm, Sweden
95. Hoffman J, Wang D, Yu F, Darrell T (2016) FCNs in the wild: pixel-level adversarial and constraint-based adaptation. CoRR. arXiv:1612.02649
96. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA

97. Huang J, Gretton A, Borgwardt KM, Schölkopf B, Smola AJ (2007) Correcting sample selection bias by unlabeled data. In: Advances in neural information processing systems (NIPS), Vancouver, Canada
98. Huang Z, Wang X, Wang J, Liu W, Wang J (2018) Weakly-supervised semantic segmentation network with deep seeded region growing. In: Computer vision and pattern recognition (CVPR), Salt Lake City, UT, USA
99. Huh M, Liu A, Owens A, Efros AA (2018) Fighting fake news: image splice detection via learned self-consistency. In: European conference on computer vision (ECCV), Munich, Germany
100. Ilse M, Tomczak JM, Welling M (2018) Attention-based deep multiple instance learning. In: International conference on machine learning (ICML), New Orleans, LA, USA
101. Inoue N, Furuta R, Yamasaki T, Aizawa K (2018) Cross-domain weakly-supervised object detection through progressive domain adaptation. In: Computer vision and pattern recognition (CVPR), Salt Lake City, UT, USA
102. Janai J, Güney F, Behl A, Geiger A (2017) Computer vision for autonomous vehicles: problems, datasets and state-of-the-art. CoRR. arXiv:1704.05519
103. Jayaraman D, Grauman K (2015) Learning image representations tied to ego-motion. In: International conference on computer vision (CVPR), Boston, MA, USA
104. Ji S, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. Trans Pattern Anal Mach Intell 35(1):221–231
105. Jiang H, Larsson G, Maire M, Shakhnarovich G, Learned-Miller E (2018) Self-supervised relative depth learning for urban scene understanding. In: European conference on computer vision (ECCV), Amsterdam, Netherlands
106. Johnson M, Schuster M, Le QV, Krikun M, Wu Y, Chen Z, Thorat N, Viégas F, Wattenberg M, Corrado G et al (2017) Google's multilingual neural machine translation system: enabling zero-shot translation. In: Association of computational linguistics (ACL), Vancouver, Canada
107. Joulin A, van der Maaten L, Jabri A, Vasilache N (2016) Learning visual features from large weakly supervised data. In: European conference on computer vision (ECCV), Amsterdam, Netherlands
108. Kaneva B, Torralba A, Freeman WT (2011) Evaluation of image features using a photorealistic virtual world. In: International conference on computer vision (ICCV), Barcelona, Spain
109. Kapoor A, Hua G, Akbarzadeh A, Baker S (2009) Which faces to tag: adding prior constraints into active learning. In: International conference on computer vision (ICCV), Kyoto, Japan
110. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: Computer vision and pattern recognition (CVPR), Columbus, OH, USA
111. Khoreva A, Benenson R, Hosang JH, Hein M, Schiele B (2017) Simple does it: weakly supervised instance and semantic segmentation. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
112. Kim T, Cha M, Kim H, Lee JK, Kim J (2017) Learning to discover cross-domain relations with generative adversarial networks. In: International conference on machine learning (ICML), Sydney, Australia
113. Kingma DP, Welling M (2013) Auto-encoding variational Bayes. In: International conference on learning representations (ICLR), Scottsdale, AZ, USA
114. Koch G, Zemel R, Salakhutdinov R (2015) Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop, Lille, France
115. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma DA et al (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis 123(1):32–73
116. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems (NIPS), Stateline, NV, USA
117. Kulis B, Saenko K, Darrell T (2011) What you saw is not what you get: domain adaptation using asymmetric kernel transforms. In: Computer vision and pattern recognition (CVPR), Colorado Springs, CO, USA
118. Kumar MP, Packer B, Koller D (2010) Self-paced learning for latent variable models. In: Advances in neural information processing systems (NIPS), Vancouver, Canada
119. Kurakin A, Goodfellow I, Bengio S (2015) Adversarial examples in the physical world. In: International conference on learning representations (ICLR), San Diego, CA, USA
120. Kuznetsova A, Rom H, Alldrin N, Uijlings J, Krasin I, Pont-Tuset J, Kamali S, Popov S, Malloci M, Duerig T, Ferrari V (2018) The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale. CoRR. arXiv:1811.00982
121. Lake BM, Salakhutdinov RR, Tenenbaum J (2013) One-shot learning by inverting a compositional causal process. In: Advances in neural information processing systems (NIPS), Stateline, NA, USA
122. Lake BM, Salakhutdinov R, Tenenbaum JB (2015) Human-level concept learning through probabilistic program induction. Science 350(6266):1332–1338
123. Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. In: Computer vision and pattern recognition, 2009 (CVPR), Miami, FL, USA
124. Lampert CH, Nickisch H, Harmeling S (2014) Attribute-based classification for zero-shot visual

object categorization. Trans Pattern Anal Mach Intell 36(3):453–465

125. Larsson G, Maire M, Shakhnarovich G (2017) Colorization as a proxy task for visual understanding. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
126. Le Guennec A, Malinowski S, Tavenard R (2016) Data augmentation for time series classification using convolutional neural networks. In: ECML/PKDD workshop on advanced analytics and learning on temporal data, Riva del Garda, Italy
127. Lee HY, Huang JB, Singh M, Yang MH (2017) Unsupervised representation learning by sorting sequences. In: International conference on computer vision (ICCV), Venice, Italy
128. Levinkov E, Fritz M (2013) Sequential Bayesian model update under structured scene prior for semantic road scenes labeling. In: International conference on computer vision (ICCV), Sydney, Australia
129. Li K, Li Y, You S, Barnes N (2017) Photo-realistic simulation of road scene for data-driven methods in bad weather. In: Conference on computer vision and pattern recognition workshop (CVPRW), Honolulu, HI, USA
130. Li W, Duan L, Xu D, Tsang IW (2014) Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. Trans Pattern Anal Mach Intell 36(6):1134–1148
131. Li Y, Wang N, Shi J, Liu J, Hou X (2016) Revisiting batch normalization for practical domain adaptation. In: International conference on learning representations workshops, Toulon, France
132. Lin D, Dai J, Jia J, He K, Sun J (2016) ScribbleSup: scribble-supervised convolutional networks for semantic segmentation. In: Computer vision and pattern recognition (CVPR), Las Vegas, NV, USA
133. Lin G, Milan A, Shen C, Reid ID (2017) RefineNet: multi-path refinement networks for high-resolution semantic segmentation. In: Conference on computer vision and pattern recognition (CVPR), Honolulu, HI, USA
134. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: European conference on computer vision (ECCV), Zurich, Switzerland
135. Liu B, Ferrari V (2017) Active learning for human pose estimation. In: International conference on computer vision (ICCV), Venice, Italy
136. Liu MY, Tuzel O (2016) Coupled generative adversarial networks. In: Advances in neural information processing systems (NIPS), Barcelona, Spain
137. Liu X, Song L, Wu X, Tan T (2016) Transferring deep representation for NIR-VIS heterogeneous face recognition. In: International conference on biometrics (ICB), Halmstad, Sweden
138. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Computer vision and pattern recognition (CVPR), Boston, MA, USA
139. Lu H, Zhang L, Cao Z, Wei W, Xian K, Shen C, van den Hengel A (2017) When unsupervised domain adaptation meets tensor representations. In: International conference on computer vision (ICCV), Venice, Italy
140. Lu Y, Tai YW, Tang CK (2018) Attribute-guided face generation using conditional CycleGAN. In: European conference on computer vision (ECCV), Munich, Germany
141. Ma F, Cavalheiro GV, Karaman S (2018) Self-supervised sparse-to-dense: self-supervised depth completion from LiDAR and monocular camera. In: International conference on robotics and automation (ICRA), Brisbane, Australia
142. Maninis KK, Caelles S, Pont-Tuset J, Van Gool L (2017) Deep extreme cut: from extreme points to object segmentation. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
143. Mehrotra A, Dukkipati A (2017) Generative adversarial residual pairwise networks for one shot learning. CoRR. arXiv:1703.08033
144. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: Advances in neural information processing systems (NIPS), Stateline, NA, USA
145. Mishra N, Rohaninejad M, Chen X, Abbeel P (2018) A simple neural attentive meta-learner. In: International conference on learning representations (ICLR), New Orleans, LA, USA
146. Misra I, Lawrence Zitnick C, Mitchell M, Girshick R (2016a) Seeing through the human reporting bias: visual classifiers from noisy human-centric labels. In: Computer vision and pattern recognition (CVPR), Las Vegas, NV, USA
147. Misra I, Zitnick CL, Hebert M (2016b) Shuffle and learn: unsupervised learning using temporal order verification. In: European conference on computer vision (ECCV), Amsterdam, Netherlands
148. Natarajan N, Dhillon IS, Ravikumar PK, Tewari A (2013) Learning with noisy labels. In: Advances in neural information processing systems (NIPS), Stateline, NA, USA
149. Nguyen HV, Ho HT, Patel VM, Chellappa R (2015) Dash-n: joint hierarchical domain adaptation and feature learning. IEEE Trans Image Process 24(12):5479–5491
150. Noroozi M, Favaro P (2016) Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision (ECCV), Amsterdam, Netherlands
151. Oquab M, Bottou L, Laptev I, Sivic J (2014) Learning and transferring mid-level image representations using

convolutional neural networks. In: Computer vision and pattern recognition (CVPR), Columbus, OH, USA
152. Owens A, Wu J, McDermott JH, Freeman WT, Torralba A (2016) Ambient sound provides supervision for visual learning. In: European conference on computer vision (ECCV), Amsterdam, Netherlands
153. Pan SJ, Yang Q et al (2010) A survey on transfer learning. Trans Knowl Data Eng 22(10):1345–1359
154. Pan SJ, Tsang IW, Kwok JT, Yang Q (2011) Domain adaptation via transfer component analysis. IEEE Trans Neural Netw 22(2):199–210
155. Papadopoulos DP, Uijlings JR, Keller F, Ferrari V (2016) We don't need no bounding-boxes: training object class detectors using only human verification. In: Computer vision and pattern recognition (CVPR), Las Vegas, NV, USA
156. Papadopoulos DP, Uijlings JR, Keller F, Ferrari V (2017) Extreme clicking for efficient object annotation. In: International conference on computer vision (ICCV), Venice, Italy
157. Papadopoulos DP, Uijlings JR, Keller F, Ferrari V (2017) Training object class detectors with click supervision. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
158. Patel VM, Gopalan R, Li R, Chellappa R (2015) Visual domain adaptation: a survey of recent advances. Signal Process Mag 32(3):53–69
159. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA (2016) Context encoders: feature learning by inpainting. In: Computer vision and pattern recognition (CVPR), Las Vegas, NV, USA
160. Pathak D, Girshick RB, Dollár P, Darrell T, Hariharan B (2017) Learning features by watching objects move. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
161. Peng KC, Wu Z, Ernst J (2018) Zero-shot deep domain adaptation. In: European conference on computer vision (ECCV), Munich, Germany
162. Peng X, Sun B, Ali K, Saenko K (2015) Learning deep object detectors from 3D models. In: International conference on computer vision (ICCV), Santiago, Chile
163. Pinheiro PO, Collobert R (2015) From image-level to pixel-level labeling with convolutional networks. In: Computer vision and pattern recognition (CVPR), Boston, MA, USA
164. Pinto L, Gandhi D, Han Y, Park YL, Gupta A (2016) The curious robot: learning visual representations via physical interactions. In: European conference on computer vision (ECCV), Amsterdam, Netherlands
165. Qiao S, Shen W, Zhang Z, Wang B, Yuille A (2018) Deep co-training for semi-supervised image recognition. In: European conference on computer vision (ECCV), Munich, Germany
166. Qin J, Liu L, Shao L, Shen F, Ni B, Chen J, Wang Y (2017) Zero-shot action recognition with error-correcting output codes. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
167. Qiu W, Yuille A (2016) UnrealCV: Connecting computer vision to unreal engine. In: European conference on computer vision (ECCV), Amsterdam, Netherlands
168. Rader N, Bausano M, Richards JE (1980) On the nature of the visual-cliff-avoidance response in human infants. Child Dev 51(1):61–68
169. Raj A, Namboodiri VP, Tuytelaars T (2015) Subspace alignment based domain adaptation for RCNN detector. In: British machine vision conference (BMVC), Swansea, UK
170. Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) Squad: 100,000+ questions for machine comprehension of text. In: Conference on empirical methods in natural language processing (EMNLP), Austin, TX, USA
171. Ratner AJ, Ehrenberg H, Hussain Z, Dunnmon J, Ré C (2017) Learning to compose domain-specific transformations for data augmentation. In: Advances in neural information processing systems, Long Beach, CA, USA
172. Ravi S, Larochelle H (2017) Optimization as a model for few-shot learning. In: International conference on learning representations (ICLR), Toulon, France
173. Redko I, Habrard A, Sebban M (2017) In: Theoretical analysis of domain adaptation with optimal transport. In: Joint European conference on machine learning and knowledge discovery in databases (ECML KDD), Skopje, Macedonia
174. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Computer vision and pattern recognition (CVPR), Las Vegas, NV, USA
175. Reed S, Lee H, Anguelov D, Szegedy C, Erhan D, Rabinovich A (2014) Training deep neural networks on noisy labels with bootstrapping. In: International conference on learning representations workshops, Banff, Canada
176. Reed S, Akata Z, Lee H, Schiele B (2016) Learning deep representations of fine-grained visual descriptions. In: Computer vision and pattern recognition (CVPR), Las Vegas, NV, USA
177. Remez T, Huang J, Brown M (2018) Learning to segment via cut-and-paste. In: European conference on computer vision (ECCV), Munich, Germany
178. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems (NIPS), Montreal, Canada
179. Richter SR, Vineet V, Roth S, Koltun V (2016) Playing for data: ground truth from computer games. In: European conference on computer vision (ECCV), Amsterdam, Netherlands
180. Richter SR, Hayder Z, Koltun V (2017) Playing for benchmarks. In: International conference on computer vision (ICCV), Venice, Italy

181. Rippel O, Paluri M, Dollar P, Bourdev L (2016) Metric learning with adaptive density discrimination. In: International conference on learning representations (ICLR), San Juan, Puerto Rico
182. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention (MICCAI), Munich, Germany
183. Ros G, Sellart L, Materzynska J, Vazquez D, Lopez AM (2016) The SYNTHIA dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In: The computer vision and pattern recognition (CVPR), Las Vegas, NV, USA
184. Roy N, McCallum A (2001) Toward optimal active learning through monte carlo estimation of error reduction. In: International conference on machine learning (ICML), Williamstown, MA, USA
185. Roy S, Unmesh A, Namboodiri VP (2018) Deep active learning for object detection. In: British machine vision conference (BMVC), Newcastle, UK
186. 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
187. Russo P, Carlucci FM, Tommasi T, Caputo B (2018) From source to target and back: symmetric bi-directional adaptive GAN. In: Computer vision and pattern recognition (CVPR), Salt Lake City, UT, USA
188. Sadeghi F, Levine S (2017) CAD2RL: real single-image flight without a single real image. In: Robotics science and systems (RSS), Boston, MA, USA
189. Saenko K, Kulis B, Fritz M, Darrell T (2010) Adapting visual category models to new domains. In: European conference on computer vision (ECCV), Crete, Greece
190. Sak H, Senior A, Beaufays F (2014) Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Conference of the international speech communication association (INTERSPEECH), Singapore
191. Sakaridis C, Dai D, Van Gool L (2018) Semantic foggy scene understanding with synthetic data. Int J Comput Vis 126:973–992
192. Salakhutdinov R, Larochelle H (2010) Efficient learning of deep Boltzmann machines. In: International conference on artificial intelligence and statistics (ICAIS), San Diego, CA, USA
193. Sankaranarayanan S, Balaji Y, Castillo CD, Chellappa R (2018) Generate to adapt: aligning domains using generative adversarial networks. In: Computer vision and pattern recognition (CVPR), Salt Lake City, UT, USA
194. Scheffer T, Decomain C, Wrobel S (2001) Active hidden Markov models for information extraction. In: International symposium on intelligent data analysis, Berlin, Heidelberg
195. Schlegl T, Seeböck P, Waldstein SM, Schmidt-Erfurth U, Langs G (2017) Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International conference on information processing in medical imaging (IPMI), Boone, NC, USA
196. Schroff F, Kalenichenko D, Philbin J (2015) FaceNet: a unified embedding for face recognition and clustering. In: Computer vision and pattern recognition (CVPR), Boston, MA, USA
197. Sener O, Savarese S (2018) Active learning for convolutional neural networks: a core-set approach. In: International conference on learning representations (ICLR), New Orleans, LA, USA
198. Settles B (2009) Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison
199. Shao L, Zhu F, Li X (2015) Transfer learning for visual categorization: a survey. IEEE Trans Neural Netw Learn Syst 26(5):1019–1034
200. Shi M, Ferrari V (2016) Weakly supervised object localization using size estimates. In: European conference on computer vision (ECCV), Amsterdam, Netherlands
201. Shimodaira H (2000) Improving predictive inference under covariate shift by weighting the log-likelihood function. J Stat Plan Inference 90(2):227–244
202. Shrivastava A, Pfister T, Tuzel O, Susskind J, Wang W, Webb R (2017) Learning from simulated and unsupervised images through adversarial training. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
203. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International conference on learning representations (ICLR), San Diego, CA, USA
204. Singh S, Gupta A, Efros AA (2012) Unsupervised discovery of mid-level discriminative patches. In: European conference on computer vision (ECCV), Firenze, Italy
205. Sivic J, Russell BC, Efros AA, Zisserman A, Freeman WT (2005) Discovering objects and their location in images. In: Computer vision and pattern recognition (CVPR), San Diego, CA, USA
206. Socher R, Ganjoo M, Manning CD, Ng A (2013) Zero-shot learning through cross-modal transfer. In: Advances in neural information processing systems (NIPS), Stateline, NA, USA
207. Sohn K, Liu S, Zhong G, Yu X, Yang MH, Chandraker M (2017) Unsupervised domain adaptation for face recognition in unlabeled videos. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
208. Song HO, Girshick R, Jegelka S, Mairal J, Harchaoui Z, Darrell T (2014) On learning to localize objects with minimal supervision. In: International conference on machine learning (ICML), Beijing, China
209. Song HO, Lee YJ, Jegelka S, Darrell T (2014) Weakly-supervised discovery of visual pattern configurations.

In: Advances in neural information processing systems (NIPS), Montreal, Canada

210. Stavens D, Thrun S (2006) A self-supervised terrain roughness estimator for off-road autonomous driving. In: Uncertainty in artificial intelligence (UAI), Cambridge, MA, USA
211. Sukhbaatar S, Bruna J, Paluri M, Bourdev L, Fergus R (2014) Training convolutional networks with noisy labels. In: International conference on learning representations workshops, Banff, Canada
212. Sun B, Saenko K (2016) Deep coral: correlation alignment for deep domain adaptation. In: European conference on computer vision (ECCV), Amsterdam, Netherlands
213. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Advances in neural information processing systems (NIPS), Montreal, Canada
214. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Computer vision and pattern recognition (CVPR), Boston, MA, USA
215. Taigman Y, Polyak A, Wolf L (2017) Unsupervised cross-domain image generation. In: International conference on learning representations (ICLR), Toulon, France
216. Tan B, Zhang Y, Pan SJ, Yang Q (2017) Distant domain transfer learning. In: AAAI, San Francisco, CA, USA
217. Taylor GR, Chosak AJ, Brewer PC (2007) OVVV: using virtual worlds to design and evaluate surveillance systems. In: Computer vision and pattern recognition (CVPR), Minneapolis, MN, USA
218. Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, Borth D, Li L (2016) Yfcc100m: the new data in multimedia research. Commun ACM 59:64–73
219. Tobin J, Fong R, Ray A, Schneider J, Zaremba W, Abbeel P (2017) Domain randomization for transferring deep neural networks from simulation to the real world. In: International conference on intelligent robots and systems (IROS), Vancouver, Canada
220. Tong S, Chang E (2001) Support vector machine active learning for image retrieval. In: ACM international conference on multimedia (MM), Ottawa, Canada
221. Torralba A, Efros AA (2011) Unbiased look at dataset bias. In: Computer vision and pattern recognition (CVPR), Colorado Springs, CO, USA
222. Toshev A, Szegedy C (2014) Deeppose: human pose estimation via deep neural networks. In: Computer vision and pattern recognition (CVPR), Columbus, OH, USA
223. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3D convolutional networks. In: International conference on computer vision (ICCV), Santiago, Chile
224. Tremblay J, Prakash A, Acuna D, Brophy M, Jampani V, Anil C, To T, Cameracci E, Boochoon S, Birchfield S (2018) Training deep networks with synthetic data: bridging the reality gap by domain randomization. In: Computer vision and pattern recognition workshops (CVPRW), Salt Lake City, UT, USA
225. Tsai YH, Hung WC, Schulter S, Sohn K, Yang MH, Chandraker M (2018) Learning to adapt structured output space for semantic segmentation. In: Computer vision and pattern recognition (CVPR), Salt Lake City, UT, USA
226. Tzeng E, Hoffman J, Zhang N, Saenko K, Darrell T (2014) Deep domain confusion: maximizing for domain invariance. In: Computer vision and pattern recognition (CVPR), Columbus, OH, USA
227. Tzeng E, Hoffman J, Darrell T, Saenko K (2015) Simultaneous deep transfer across domains and tasks. In: International conference on computer vision (ICCV), Santiago, Chile
228. Tzeng E, Hoffman J, Saenko K, Darrell T (2017) Adversarial discriminative domain adaptation. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
229. Van Den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior AW, Kavukcuoglu K (2016) WaveNet: a generative model for raw audio. CoRR. arXiv:1609.03499 (125)
230. Van Horn G, Branson S, Farrell R, Haber S, Barry J, Ipeirotis P, Perona P, Belongie S (2015) Building a bird recognition app and large scale dataset with citizen scientists: the fine print in fine-grained dataset collection. In: Computer vision and pattern recognition (CVPR), Boston, MA, USA
231. Varma G, Subramanian A, Namboodiri A, Chandraker M, Jawahar CV (2019) IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. In: IEEE Winter conference on applications of computer vision (WACV),Waikoloa, Hawaii
232. Vazquez D, Lopez AM, Marin J, Ponsa D, Geronimo D (2014) Virtual and real world adaptation for pedestrian detection. Trans Pattern Anal Mach Intell 36(4):797–809
233. Veit A, Alldrin N, Chechik G, Krasin I, Gupta A, Belongie SJ (2017) Learning from noisy large-scale datasets with minimal supervision. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
234. Vezhnevets A, Buhmann JM, Ferrari V (2012) Active learning for semantic segmentation with expected change. In: Computer vision and pattern recognition (CVPR), Providence, RI, USA
235. Vijayanarasimhan S, Grauman K (2014) Large-scale live active learning: training object detectors with crawled data and crowds. Int J Comput Vis 108(1–2):97–114
236. Vinyals O, Blundell C, Lillicrap T, Wierstra D et al (2016) Matching networks for one shot learning. In: Advances in neural information processing systems (NIPS), Barcelona, Spain

237. Vogt P, Smith ADM (2005) Learning color words is slow: a cross-situational learning account. Behav Brain Sci 28(4):509–510
238. Wang C, Mahadevan S (2011) Heterogeneous domain adaptation using manifold alignment. In: International joint conference on artificial intelligence (IJCAI), Barcelona, Spain
239. Wang M, Deng W (2018) Deep visual domain adaptation: a survey. Neurocomputing 312:135–153
240. Wang X, Gupta A (2015) Unsupervised learning of visual representations using videos. In: International conference on computer vision (ICCV), Santiago, Chile
241. Wang YX, Hebert M (2016) Learning to learn: model regression networks for easy small sample learning. In: European conference on computer vision (ECCV), Amsterdam, Netherlands
242. Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. J Big Data 3(1):9
243. Wu J, Yu Y, Huang C, Yu K (2015) Deep multiple instance learning for image classification and auto-annotation. In: Computer vision and pattern recognition (CVPR), Boston, MA, USA
244. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, ukasz Kaiser, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M, Dean J (2016) Google's neural machine translation system: bridging the gap between human and machine translation. CoRR. arXiv :1609.08144
245. Xian Y, Akata Z, Sharma G, Nguyen Q, Hein M, Schiele B (2016) Latent embeddings for zero-shot classification. In: Computer vision and pattern recognition (CVPR), Las Vegas, NV, USA
246. Xian Y, Schiele B, Akata Z (2017) Zero-shot learning-the good, the bad and the ugly. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
247. Xiao T, Xia T, Yang Y, Huang C, Wang X (2015) Learning from massive noisy labeled data for image classification. In: Computer vision and pattern recognition (CVPR), Boston, MA, USA
248. Xu J, Schwing AG, Urtasun R (2015) Learning to segment under various forms of weak supervision. In: Computer vision and pattern recognition (CVPR), Boston, MA, USA
249. 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: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
250. Yao A, Gall J, Leistner C, Van Gool L (2012) Interactive object detection. In: Computer vision and pattern recognition (CVPR), Providence, RI, USA
251. Yi Z, Zhang HR, Tan P, Gong M (2017) DualGAN: unsupervised dual learning for image-to-image translation. In: International conference on computer vision (ICCV), Venice, Italy
252. Yoo D, Fan H, Boddeti VN, Kitani KM (2018) Efficient k-shot learning with regularized deep networks. In: AAAI, New Orleans, LA, USA
253. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural. networks? In: Advances in neural information processing systems (NIPS), Montreal, Canada
254. Zhang H, Xu T, Li H, Zhang S, Huang X, Wang X, Metaxas D (2017a) StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks. In: International conference on computer vision (ICCV), Venice, Italy
255. Zhang J, Ding Z, Li W, Ogunbona P (2018) Importance weighted adversarial nets for partial domain adaptation. In: Computer vision and pattern recognition (CVPR), Salt Lake City, UT, USA
256. Zhang L, Xiang T, Gong S et al (2017b) Learning a deep embedding model for zero-shot learning. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
257. Zhang R, Isola P, Efros AA (2016) Colorful image colorization. In: European conference on computer vision (ECCV), Amsterdam, Netherlands
258. Zhang R, Isola P, Efros AA (2017c) Split-brain autoencoders: unsupervised learning by cross-channel prediction. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
259. Zhang Y, David P, Gong B (2017d) Curriculum domain adaptation for semantic segmentation of urban scenes. In: International conference on computer vision (ICCV), Venice, Italy
260. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA
261. Zhu JJ, Bento J (2017) Generative adversarial active learning. In: Advances in neural information processing systems workshops, Long Beach, CA
262. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: International conference on computer vision (ICCV), Venice, Italy
263. Zhu Y, Chen Y, Lu Z, Pan SJ, Xue GR, Yu Y, Yang Q (2011) Heterogeneous transfer learning for image classification. In: AAAI, San Francisco, California, USA
264. Zhuang B, Liu L, Li Y, Shen C, Reid ID (2017) Attend in groups: a weakly-supervised deep learning framework for learning from web data. In: Computer vision and pattern recognition (CVPR), Honolulu, HI, USA

Lovish Chum is currently a research fellow at Center for Visual Information Technology at IIIT Hyderabad. Earlier, he graduated from Indian Institute of Technology Kanpur. He is interested in exploring self-supervision for problems in computer vision.

Anbumani Subramanian is presently a Lead Architect at Intel in India. Prior to that, he was a Senior Research Scientist at Hewlett-Packard Labs. He was a postdoctoral researcher at Virginia Tech, where he also received his Ph.D. degree. Anbu began his career as a Software Engineer at IBM. He is a Senior Member of IEEE, has several peer-reviewed publications, and has nine patents granted. His expertise and interests include computer vision, machine learning, and human–computer interaction.

Vineeth N. Balasubramanian is an Associate Professor in the Department of Computer Science and Engineering at the Indian Institute of Technology, Hyderabad. His research interests include deep learning, machine learning, computer vision, non-convex optimization, and real-world applications in these areas. He has over 60 research publications in these areas including premier peer-reviewed venues, including CVPR, ICCV, KDD, ICDM, IEEE TPAMI, and ACM MM, 5 patent applications, and an edited book on a recent development in machine learning called Conformal Prediction. His Ph.D. dissertation at Arizona State University (completed in 2010) on the Conformal Predictions framework was nominated for the Outstanding Ph.D. Dissertation at the Department of Computer Science. He was also awarded the Gold Medals for Academic Excellence in the Bachelors program in Math in 1999, and for his Masters program in Computer Science in 2003. He is an active reviewer/contributor at many conferences such as ICCV, IJCAI, ACM MM, and ACCV, as well as journals including IEEE TPAMI, IEEE TNNLS, Machine Learning, and Pattern Recognition. He is a member of the IEEE and ACM, and currently serves as the Secretary of the AAAI India Chapter.

CV Jawahar is a professor at IIIT Hyderabad, India. He received his Ph.D. from IIT Kharagpur and has been with IIIT Hyderabad since Dec. 2000. At IIIT Hyderabad, Jawahar leads a group focusing on computer vision, machine learning, and multimedia systems. In the recent years, he has been looking into a set of problems that overlap with vision, language, and text. He is also interested in large-scale multimedia systems with a special focus on retrieval. He has more than 50 publications in top tier conferences in computer vision, robotics, and document image processing. He has served as a chair for the previous editions of ACCV, WACV, IJCAI, and ICVGIP. Presently, he is an area editor of CVIU and an associate editor of IEEE PAMI.