Noname manuscript No.
(will be inserted by the editor)

Beyond Supervised Learning: A Computer Vision Perspective

Lovish Chum · Anbumani Subramanian · Vineeth N Balasubramanian · C.V. Jawahar

Received: date / Accepted: date

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 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

Lovish Chum · C.V. Jawahar
CVIT, IIIT Hyderabad

Anbumani Subramanian
Intel, Bangalore

Vineeth N Balasubramanian
IIT Hyderabad

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 has been painstakingly collected and annotated for problems such as image classification (Russakovsky et al. (2015); Kuznetsova et al. (2018)), image captioning (Krishna et al. (2017)), instance segmentation (Lin et al. (2014)), visual question answering (Goyal et al. (2017)) 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 dataset 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 (Torralba and Efros (2011)). 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 (Shimodaira (2000); Torralba and Efros (2011)). 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 significant degradation in performance when tested on the latter. This is in direct contrast to living organisms which perform a wide variety of tasks in different settings without receiving direct supervision (Rader et al. (1980); Vogt and Smith (2005)).

This survey is targeted towards summarizing 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 dataset; (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 dataset. For images, this involves affine transformations such as shifting, rotation, shearing, flipping and distortion of the original image (Krizhevsky et al. (2012)). Some recent papers also advocate adding Gaussian noise to augment the images in the dataset. Ratner et al. (2017) 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 is in the form of time-series, window slicing and window warping can be used for augmentation purposes (Le Guennec et al. (2016)).

Algorithm-centric techniques try to relax the need of perfectly labeled data by altering the model requirements to acquire supervision through inexact (Xu et al.

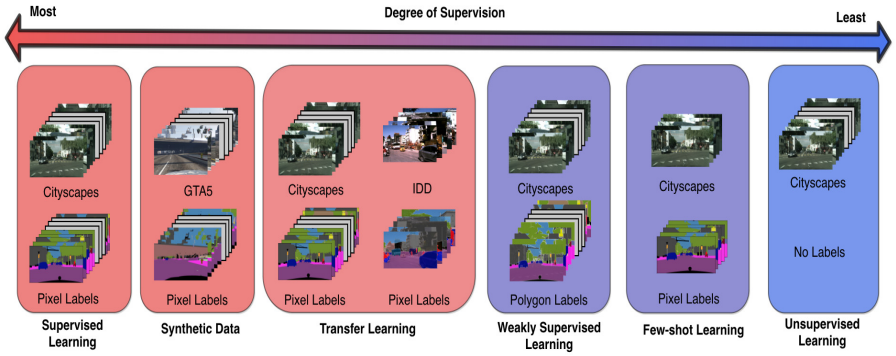


Fig. 1: Learning paradigms arranged in decreasing order of supervision signal. Semantic segmentation of outdoor scene is taken as an example task (1) Fully supervised learning requires a lot of annotated data to learn a viable model Cordts et al. (2015). (2) Synthetically generated instances can be used to compensate for the lack of real-world data Richter et al. (2016). (3) Knowledge from one real-world dataset can be transferred to another dataset 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 Dataset (IDD) Varma et al. (2018). (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
.

(2015)), inaccurate (Natarajan et al. (2013)) and incomplete labels (Chapelle et al. (2009)). 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 dataset (Tong and Chang (2001)). Another set of methods exploit the knowledge gained while learning from a related domain or task by efficiently transferring it to the test environment (Saenko et al. (2010)).

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 dataset along with the real data for training. This proves to be greatly beneficial as real-world dataset 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 (Zhang et al. (2017d)). 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 Section 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 Section 3. Methods for weak and self-supervision are discussed in Section 4 and 6 respectively. Methods which address the task of learning an adequate model from a few instances are discussed in Section 5. Finally, we conclude the paper discussing the promises, challenges and open research frontiers beyond supervised learning in Section 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 $\mathscr{X}$ and $\mathscr{Y}$ 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 dataset. 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 $f : \mathscr{X} \longrightarrow \mathscr{Y}$ in the hypothesis space $\mathscr{F}$. To measure the suitability of the function f, a loss function $l : \mathscr{Y} \times \mathscr{Y} \longrightarrow \mathbb{R}^{\geq 0}$ is defined over space $\mathscr{L}$. A machine learning algorithm tries to minimize the risk R associated with wrong predictions

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

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 it's input space as $\mathscr{X}_{\text{synth}}$ instead of $\mathscr{X}$. However, the label space remains the same. To elaborate, we have a new domain $\mathscr{D}_{\text{synth}} = \{X_{\text{synth}}, P(X_{\text{synth}})\}$ which is different from the real domain $\mathscr{D} = \{X, P(X)\}$ as both their input feature space and marginal distributions are different. Hence, we cannot use the objective predicting function $f_{\text{synth}} : \mathscr{X}_{\text{synth}} \longrightarrow \mathscr{Y}$ for mapping $\mathscr{X}$ to $\mathscr{Y}$.

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 $\mathscr{Y}$ and the conditional distribution $P(Y|X)$, as $\mathscr{T} = \{\mathscr{Y}, P(Y|X)\}$. Building on the above notations, we define domain shift ($\mathscr{D}_s \neq \mathscr{D}_t$) and label space shift ($\mathscr{T}_s \neq \mathscr{T}_t$) where $\mathscr{D}_s$ and $\mathscr{D}_t$ are source and target domains respectively. As an example, using synthetic data and then adapting the learned objective to real domain falls under domain shift as $\mathscr{D} \neq \mathscr{D}_{\text{synth}}$. 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 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 $\mathscr{T} = \{\mathscr{Y}, P(Y|X)\}$. Instead, we define a proxy task $\mathscr{T}_{\text{proxy}} = \{Z, P(Z|X)\}$ 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 (Krizhevsky et al. (2012); Szegedy et al. (2015); Simonyan and Zisserman (2015); He et al. (2016); Huang et al. (2017)), object detection (Girshick

et al. (2014); Girshick (2015); Ren et al. (2015); Redmon et al. (2016); Liu and Tuzel (2016)), semantic segmentation (Long et al. (2015); Ronneberger et al. (2015); Badrinarayanan et al. (2017); Lin et al. (2017); Zhao et al. (2017); Chen et al. (2017a); He et al. (2017)), pose estimation (Toshev and Szegedy (2014); Cao et al. (2017)), action recognition (Ji et al. (2013); Donahue et al. (2015); Tran et al. (2015); Feichtenhofer et al. (2016); Girdhar et al. (2017)), video classification (Karpathy et al. (2014)) and optical flow estimation (Dosovitskiy et al. (2015)) have consistently increased allowing for their large-scale deployment. Apart from computer vision, problems in other domains such as speech recognition (Graves (2013); Sak et al. (2014); Graves and Jaitly (2014)), speech synthesis (Van Den Oord et al. (2016)), machine translation (Sutskever et al. (2014); Bahdanau et al. (2015); Wu et al. (2016); Gu et al. (2017)) and machine reading (Rajpurkar et al. (2016)) have also seen a significant improvement in their performance metrices.

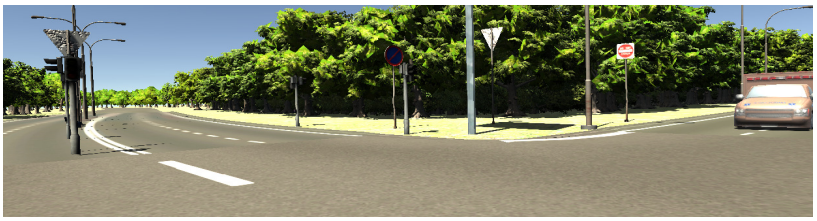
Despite their success, supervised learning-based models have a fair share of issues. First of all, they are data hungry requiring a huge amount of instance-label pairs. To add, a majority of large datasets required to train these models are proprietary as they provide an advantage to the owner in training a supervised model for a particular task and domain. Secondly, 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 dataset.

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 datasets which are exponentially larger than real-world datasets. 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 dataset. This presents a major advantage for computer vision researchers as real-world datasets tend to be non-exhaustive, redundant, heavily biased and partly representative of the complexity of natural images (Torralba and Efros (2011)). 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 dataset. Last, but not least, having a few high-profile real-world datasets bias the research community towards the tasks for which annotations have been provided with these datasets. Thus, graphically generated synthetic datasets 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 has been used mainly for two purposes: (1) evaluation of the generalizability of the model due to the large variability of synthetic

(a) Example image from KITTI dataset (Geiger et al. (2013))



(b) Example image from Virtual KITTI dataset (Gaidon et al. (2016))

(c) Real cars augmented to the KITTI dataset (Alhaija et al. (2018))

Fig. 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 dataset.

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. (2007). Kaneva et al. (2011) and Aubry and Russell (2015) use synthetic data to evaluate hand-crafted and deep features respectively. Butler et al. (2012) propose MPI Sintel Flow dataset, a synthetic benchmark for optical flow estimation. Handa et al. (2014) introduce ICL-NUIM, a dataset for evaluation of visual odometry.

More significantly, synthetic data is 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. Vazquez et al. (2014) 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 (Aubry et al. (2014); Peng et al. (2015)), indoor scene understanding (Handa et al. (2016)), action recognition (De Souza et al. (2017)) and variety of other tasks. Instead of rendering the entire scene, Gupta et al. (2016) 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 (Dwibedi et al. (2017)) and semantic segmentation (Remez et al. (2018)) 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 (Sadeghi and Levine (2017); Tobin et al. (2017); Tremblay et al. (2018)) claims to generate synthetic data with sufficient variations such that the model views real data as just another variation of the synthetic dataset.

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. (2016) present the Virtual KITTI dataset and conduct experiments on multi-object tracking. SYNTHIA (Ros et al. (2016)) and GTA (Richter et al. (2016)) provide urban scene understanding data along with semantic segmentation benchmarks. UnrealCV (Qiu and Yuille (2016)) provides a simple interface for researchers to build a virtual world without worrying about the game's API.

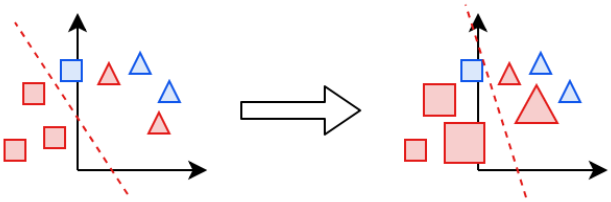
Synthetic data for Autonomous Navigation

Autonomous Navigation has greatly benefited from the use of synthetic datasets as pixel-level ground truth can be obtained easily and cheaply using label propagation from frame to frame. As a result, several synthetic datasets have been curated particularly for visual tasks pertaining to autonomous navigation (Gaidon et al. (2016); Ros et al. (2016); Richter et al. (2016, 2017); Li et al. (2017); Sakaridis et al. (2018)). Alhaija et al. (2018) 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 (Chen et al. (2015); Dosovitskiy et al. (2017)). Janai et al. (2017) offers a comprehensive survey of literature pertinent to autonomous driving.

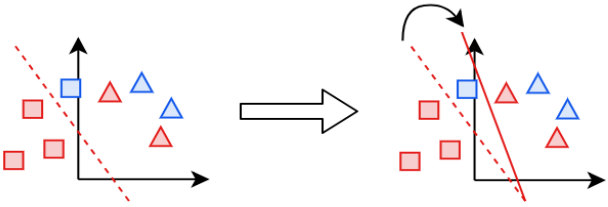
One of the major challenges in using synthetic data for training is the domain gap between real and synthetic datasets. Transfer learning discussed in Section 3 offers a solution to this problem. Eventually, through the use of synthetic data, we would like to replace the expensive data acquisition process 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

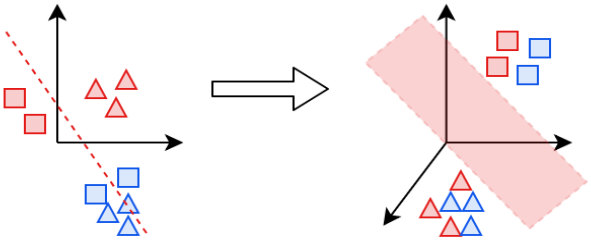
As stated in Section 2, a model trained on source domain does not perform well on a target domain with different distribution. Domain adaptation (DA) is a technique



(a) Instance Re-weighting.



(b) Modifying the model with pseudo-level information.



(c) Feature projection to a common space.

Fig. 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.

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 catego-

rize 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 $\mathscr{X}$ but different data distributions $P(X)$ or $P(Y|X)$. Some traditional methods for homogeneous domain adaptation include instance re-weighting (Chattopadhyay et al. (2012)), feature transformations (Huang et al. (2007); Daumé III (2007)) or kernel-based techniques that learn an explicit transform from source to target domain (Duan et al. (2012); Pan et al. (2011); Gong et al. (2012)). Figure 3 pictorially presents traditional domain adaptation methods. All the techniques addressing this problem aim to correct the differences 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 (Gopalan et al. (2011); Kulis et al. (2011); Zhu et al. (2011)), object detection, semantic segmentation (Levinkov and Fritz (2013)) 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 (Wang and Mahadevan (2011)) and feature augmentation (Duan et al. (2011); Li et al. (2014)) 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. (2010) and Pan et al. (2010) for a summary of homogeneous and Day and Khoshgoftaar (2017) and Weiss et al. (2016) for a detailed overview of heterogeneous adaptation techniques. (Patel et al. (2015); Shao et al. (2015); Csurka (2017)) 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 (Nguyen et al. (2015)) or subspace alignment (Raj et al. (2015); Lu et al. (2017)) 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 (Oquab et al. (2014); Yosinski et al. (2014); Donahue et al. (2014)). Labels or attribute information (Gebru et al. (2017); Tzeng et al. (2015)), Maximum Mean Discrepancy (MMD) (Tzeng et al. (2014); Yan et al. (2017)), correlation alignment (Sun and Saenko (2016)), statistical associations (Haeusser et al. (2017)), batch normalization (Li et al. (2016)) 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. (2016) introduced DANN architecture which uses a gradient reversal layer to ensure that feature distributions over the two domains are aligned. Liu and Tuzel (2016) 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. (Yoo et al. (2018); Shrivastava et al. (2017); Bousmalis et al. (2017); Hoffman et al. (2016b)) also focus on generating synthetic target data using adversarial loss, albeit using it in pixel space instead of embedding space. Sankaranarayanan et al. (2018) 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. (2017) 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 (Goodfellow et al. (????)) consists of two neural networks; a generator that creates samples using noise and a discriminator which receives samples from both the generator and real dataset 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 (Antoniou et al. (2017); Frid-Adar et al. (2018)) 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 (Lu et al. (2018)) , adversarial training (Kurakin et al. (2015)), anomaly detection (Schlegl et al. (2017)) and adversarial cryptography (Abadi and Andersen (2016)).

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. (2016) use an encoder which is trained simultaneously to accomplish source label prediction along with target data reconstruction. Bousmalis et al. (2016) train separate encoders to account for domain specific and domain invariant features. Additionally, it uses domain invariant features for classification while using both kinds of features for reconstruction. Methods based on adversarial reconstruction are proposed in (Zhu et al. (2017); Kim et al. (2017); Yi et al. (2017); Russo et al. (2018)) 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 (Redko et al. (2017); Damodaran et al. (2018)). Courty et al. (2017) assign pseudo-labels to the target data

using the source classifier. Further, they transport the source data points to the target distribution minimizing the distance traveled and changes in labels while moving the points.

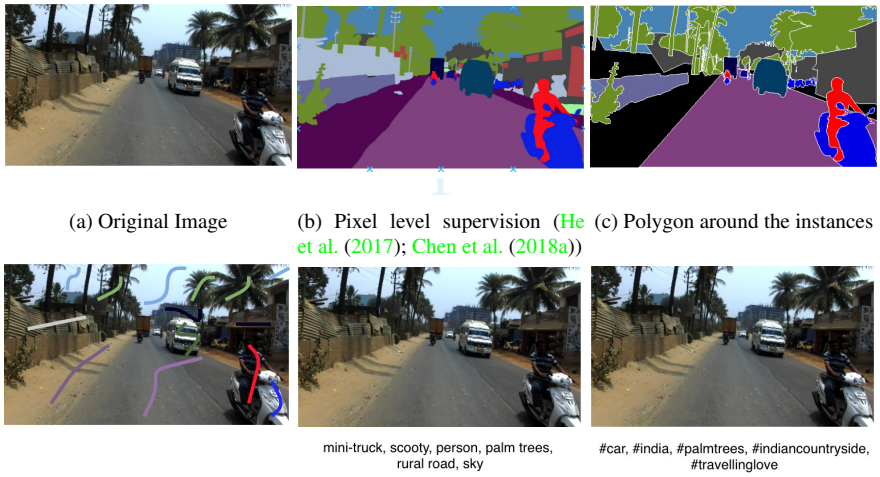
Visual adaptation has been studied for problems such as cross-modal face recognition (Liu et al. (2016); Sohn et al. (2017)), object detection (Hoffman et al. (2016a); Chen et al. (2018b)), semantic segmentation (Chen et al. (2017c); Zhang et al. (2017d); Tsai et al. (2018)), person re-identification (Deng et al. (2018)) and image captioning (Chen et al. (2017b)). Although deep DA has achieved considerable improvement over 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 (Taigman et al. (2017)) and text-to-image synthesis (Reed et al. (2016); Zhang et al. (2017a)) have also been addressed using adversarial adaptation techniques. Another interesting direction of work pertains open set DA (Busto and Gall (2017); Cao et al. (2017); Zhang et al. (2018)) which loosens the assumption that output sets of both the source and target class must exactly be the same. Tan et al. (2017) 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 (2018).

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. Lastly, 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. Conventional semi-supervised approaches include self-training, co-training (Blum and Mitchell (1998); Qiao et al. (2018)) and graph-based methods (Duchenne et al. (2008)). A discussion on these is out of the scope of this survey. Interested readers are directed to Chapelle et al. (2009) for a detailed overview of semi-supervised learning.



(a) Original Image (b) Pixel level supervision (He et al. (2017); Chen et al. (2018a)) (c) Polygon around the instances

(d) Scribbles (Lin et al. (2016)) (e) Image-level supervision (Pinheiro and Collobert (2015)) (f) Supervision via collaborative image tags

Fig. 4: An example of the varying degree of supervision for semantic segmentation problem. Although pixels-level labels provide strong supervision, they are relatively expensive to obtain. Thus, 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).

Active learning methods are used in computer vision to reduce labeling efforts in problems such as image annotation (Kapoor et al. (2009)), recognition (Vijayanarasimhan and Grauman (2014)), object detection (Yao et al. (2012)), segmentation (Vezhnevets et al. (2012)) and pose estimation (Liu and Ferrari (2017)). In this paradigm, unlabeled observations are optimally selected from the dataset 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 (Cohn et al. (1996)). Some common query formulation strategies include maximizing the label change (Freytag et al. (2014)), maximizing the diversity of selected samples (Elhamifar et al. (2013)), reducing the expected error of classifier (Roy and McCallum (2001)) or uncertainty sampling (Scheffer et al. (2001)). A survey by Settles (2009) 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 (2018) and Gal et al. (2017) present an effective method to train a CNN

using active learning heuristics. An approach to synthesize query examples using GAN is proposed by Zhu and Bento (2017). Fang et al. (2017) reframe active learning as a reinforcement learning problem. Also, deep active learning methods have been used to address vision tasks such as object detection in Roy et al. (2018).

4.2 Inexact Supervision

Apart from dealing with partially labeled datasets, 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 datasets reflect this fact. For example, in Imagenet (Russakovsky et al. (2015)), while 14 million images are provided with image-level labels and 500,000 are annotated with bounding boxes; only 4,460 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 (Song et al. (2014b)), regularization (Song et al. (2014a)) and curriculum learning (Kumar et al. (2010)) techniques to alleviate the issue. Recent works (Wu et al. (2015); Ilse et al. (2018)) 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 recent years. Bilen and Vedaldi (2016) 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 (Shi and Ferrari (2016)) and count (Gao et al. (2018)), click supervision (Papadopoulos et al. (2017b,a)) and human verification (Papadopoulos et al. (2016)). Similar methods have also been proposed for weakly supervised semantic segmentation problems (Khoreva et al. (2017); Lin et al. (2016); Bearman et al. (2016); Maninis et al. (2017); Pinheiro and Collobert (2015); Huang et al. (2018)). Figure 4 depicts some weak supervision signals used for semantic segmentation problem.

4.3 Inaccurate supervision

As curating large-scale datasets is an expensive process, building a machine learning model which uses web datasets such as YouTube8m (Abu-El-Haija et al. (2016)), YFCC100M (Thomee et al. (2016)) and Sports-1M (Karpathy et al. (2014)) is one of the pragmatic ways to leverage the almost infinite amount of visual data. However, labels in these datasets 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 (Fan et al. (2010); Sukhbaatar et al. (2014)). Nevertheless, noisy instances may not be outliers and occupy a significant portion of the training data. Moreover, algorithms 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 (Reed et al. (2014); Van Horn et al. (2015); Joulin et al. (2016); Misra et al. (2016a)) by devising noise-tolerant loss functions (Ghosh et al. (2017)) or adding appropriate regularization terms (Arpit et al. (2017)). For a comprehensive overview of learning algorithms robust to noise, we refer to Frénay and Verleysen (2014).

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 is collected from the web. Chen and Gupta (2015) propose a two-stage curriculum learning technique on easier examples before adapting it to web images. Xiao et al. (2015) 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 (Divvala et al. (2014); Gan et al. (2016b)), image classification (Veit et al. (2017)), video recognition (Gan et al. (2016a)) and object localization (Zhuang et al. (2017)).

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 dataset 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. (2006) 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. (2013) 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 datasets such as Omniglot (Lake et al. (2015)). Wang and Hebert (2016) learn to regress from parameters of the classifier trained on few images to the parameters of the classifier trained

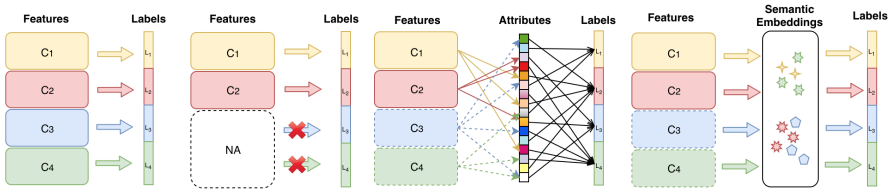
on a large number of images. More recent efforts into few-shot learning techniques can be broadly categorized into metric-learning and meta-learning based methods.

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. (2015) use a deep learning model based on computing the pair-wise distance between the samples based on Siamese networks following which the learned distance is used to solve few-shot problems through k-nearest neighbors classification. Vinyals et al. (2016) 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. (2015) and adaptive density estimation Rippel et al. (2016). Mehrotra and Dukkipati (2017) 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. (2017) 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 (2017) also introduce a meta-learning framework employing LSTM updates for a given episode. Recently, a method proposed by Mishra et al. (2018) also exploit contextual information within the tasks using Temporal Convolutions.

Another set of methods for few-shot learning rely on efficient regularization techniques to avoid over-fitting on the small number of instances. Hariharan and Girshick (2017) suggest a gradient magnitude regularization technique for training a classifier along with a method to hallucinate additional examples for few-shot classes. Yoo et al. (2018) also regularizes 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. (2009) 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 (Farhadi et al. (2009); Akata et al. (2013); Lampert et al. (2014); Xian et al. (2016)). More recently, word-embeddings (Socher et al. (2013); Zhang et al. (2017b)) and image captions (Reed et al. (2016)) have also been used in place



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

Fig. 5: A 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.

attributes as a semantic space. Figure 5 compares the two common approaches to ZSL with supervised learning.

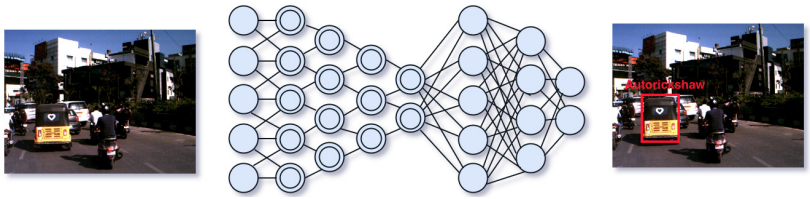
In ZSL, a joint embedding space is learned during training where both the visual features and semantic vectors are projected. During testing on unseen classes, nearest neighbor search is performed in this embedding space to match the projection of visual feature vector against a novel object type. A pairwise ranking formula is used to learn parameters of a bi-linear model in Akata et al. (2013) and Frome et al. (2013). Recently, Zhang et al. (2017b) 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. (2017) 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 (Habibian et al. (2014)), object detection (Bansal et al. (2018)), action recognition (Qin et al. (2017)), machine translation (Johnson et al. (2017)).

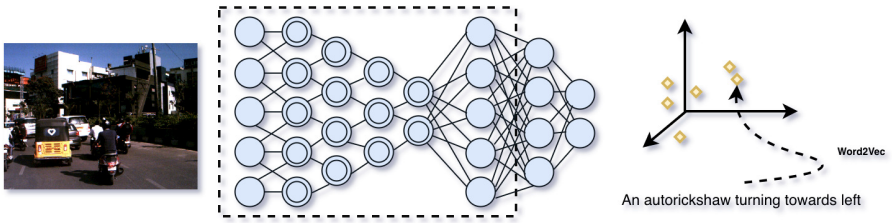
6 Self-supervised Learning

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.

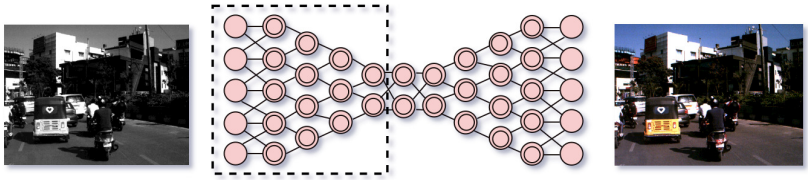
Before the advent of deep-learning driven self-supervision models, significant work was carried out in unsupervised learning of image representations using handcrafted (Sivic et al. (2005)) or mid-level features (Singh et al. (2012)). This was followed by deep learning-based methods like autoencoders (Hinton and Salakhutdinov



(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 sentence 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.

Fig. 6: Strong supervision vs. Weak Supervision vs. Self-supervision. ○ and ◎ depict fully connected and convolutional layers respectively.

(2006)), boltzmann machines (Salakhutdinov and Larochelle (2010)) and variational methods (Kingma and Welling (2013)) which learn by estimating latent parameters which help reconstruct the data.

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 (Mikolov et al. (2013)) 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. (2015) to predict the relative location of two image patches as a pretext task. The same notion is extended in Noroozi and Favaro (2016) by predicting the order of shuffled image patches. Apart from spatial context based auxiliary tasks, predicting color channel from luminance values (Zhang et al. (2016); Larsson et al. (2017)) and regressing to a missing patch in an image using generative models (Pathak et al. (2016)) have also been used to learn useful semantic information in images. Other modalities used for feature learning in images include text (Gomez et al. (2017)), motion (Pinto et al. (2016); Pathak et al. (2017)) and cross-channel prediction (Zhang et al. (2017c)). Recently, Huh et al. (2018) 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 (2015) authors use relationships between the triplet of image patches obtained from tracking. Misra et al. (2016b) train a network to guess whether a given sequence of frames from a video are in chronological order.Lee et al. (2017) make the network predict the correct sequence of frames given a shuffled set. Apart from temporal context, estimating camera motion (Jayaraman and Grauman (2015)), ego-motion (Agrawal et al. (2015)) and predicting the statistics of ambient sound (Owens et al. (2016); Arandjelovic and Zisserman (2017)) 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 (2006) where authors estimate the terrain roughness based on the ”shocks” the vehicle receives while passing over it. Jiang et al. (2018) show that predicting relative depth is an effective proxy task for learning visual representations. Ma et al. (2018) 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 has 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 is cost effective and flexible alternative to real-world datasets, 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 tech-

niques 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.* Peng et al. (2018) address the domain adaptation problem when no task-relevant data is present in the target domain. Inoue *et. al.* Inoue et al. (2018) 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, other learning paradigms inspired by human reasoning and abstraction such as meta-learning Andrychowicz et al. (2016); Finn et al. (2017), lifelong learning Chen and Liu (2016) 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.

References

Abadi M, Andersen DG (2016) Learning to protect communications with adversarial neural cryptography. CoRR abs/1610.06918 11

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. vol abs/1609.08675v1 14

Agrawal P, Carreira J, Malik J (2015) Learning to see by moving. In: International Conference on Computer Vision (CVPR), Boston, MA 19

Akata Z, Perronnin F, Harchaoui Z, Schmid C (2013) Label-embedding for attribute-based classification. In: Computer Vision and Pattern Recognition (CVPR), Portland, OR 16, 17

Alhaija H, Mustikovela S, Mescheder L, Geiger A, Rother C (2018) Augmented reality meets computer vision: Efficient data generation for urban driving scenes. International Journal of Computer Vision (IJCV) 126(9):961–972 7, 8

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 20

Antoniou A, Storkey A, Edwards H (2017) Data augmentation generative adversarial networks. CoRR abs/1711.04340 11

Arandjelovic R, Zisserman A (2017) Look, listen and learn. In: International Conference on Computer Vision (ICCV), Venice, Italy 19

Arpit D, Jastrzebski 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 15

Aubry M, Russell BC (2015) Understanding deep features with computer-generated imagery. In: International Conference on Computer Vision (ICCV), Santiago, Chile 7

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 8

Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: A deep convolutional encoder-decoder architecture for image segmentation. Transactions on Pattern Analysis and Machine Intelligence (PAMI) 39(12):2481–2495 6

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 6

Bansal A, Sikka K, Sharma G, Chellappa R, Divakaran A (2018) Zero-shot object detection. In: European Conference on Computer Vision (ECCV), Munich, Germany 17

Bearman A, Russakovsky O, Ferrari V, Fei-Fei L (2016) Whats the point: Semantic segmentation with point supervision. In: European Conference on Computer Vision (ECCV), Amsterdam, Netherlands 14

Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW (2010) A theory of learning from different domains. Machine Learning 79(1-2):151–175 10

Bilen H, Vedaldi A (2016) Weakly supervised deep detection networks. In: Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV 14

Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training. In: Computational Learning Theory (CoLT), Wisconsin, Madison 12

Bousmalis K, Trigeorgis G, Silberman N, Krishnan D, Erhan D (2016) Domain separation networks. In: Advances in Neural Information Processing Systems (NIPS), Barcelona, Spain 11

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 11

Busto PP, Gall J (2017) Open set domain adaptation. In: International Conference on Computer Vision (ICCV), Venice, Italy 12

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),

Firanze, Italy 7
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 6, 12
Chapelle O, Scholkopf B, Zien A (2009) Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. Transactions on Neural Networks 3, 12
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. Transactions on Knowledge Discovery from Data (TKDD) 6(4):18 10
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 8
Chen LC, Papandreou G, Schroff F, Adam H (2017a) Rethinking atrous convolution for semantic image segmentation. CoRR abs/1706.05587 6
Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018a) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. Pattern Analysis and Machine Intelligence (PAMI) 40(4):834–848 13
Chen TH, Liao YH, Chuang CY, Hsu WT, Fu J, Sun M (2017b) Show, adapt and tell: Adversarial training of cross-domain image captioner. In: International Conference on Computer Vision (ICCV), Venice, Italy 12
Chen X, Gupta A (2015) Webly supervised learning of convolutional networks. In: International Conference on Computer Vision (ICCV), Santiago, Chile 15
Chen Y, Li W, Sakaridis C, Dai D, Van Gool L (2018b) Domain adaptive faster r-cnn for object detection in the wild. In: Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT 12
Chen YH, Chen WY, Chen YT, Tsai BC, Wang YCF, Sun M (2017c) No more discrimination: Cross city adaptation of road scene segmenters. In: International Conference on Computer Vision (ICCV), Venice, Italy 12
Chen Z, Liu B (2016) Lifelong machine learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 10(3):1–145 20
Cohn DA, Ghahramani Z, Jordan MI (1996) Active learning with statistical models. Journal of Artificial Intelligence Research 4:129–145 13
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 3
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 11
Csurka G (2017) Domain adaptation for visual applications: A comprehensive survey. CoRR abs/1702.05374 10
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 11
Daumé III H (2007) Frustratingly easy domain adaptation. In: Association of Computational Lingustics (ACL), Prague, Czech Republic 10

Day O, Khoshgoftaar TM (2017) A survey on heterogeneous transfer learning. Journal of Big Data 4(1):29 10

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 8

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 12

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 15

Doersch C, Gupta A, Efros AA (2015) Unsupervised visual representation learning by context prediction. In: International Conference on Computer Vision (ICCV), Santiago, Chile 18

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 10

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 6

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 6

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 8

Duan L, Xu D, Tsang I (2011) Learning with augmented features for heterogeneous domain adaptation. In: International Conference on Machine Learning (ICML), Edinburgh, Scotland 10

Duan L, Tsang IW, Xu D (2012) Domain transfer multiple kernel learning. Transactions on Pattern Analysis and Machine Intelligence (PAMI) 34(3):465–479 10

Duchenne O, Audibert JY, Keriven R, Ponce J, Ségonne F (2008) Segmentation by transduction. In: Computer Vision and Pattern Recognition (CVPR), Anchorage, AL 12

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 8

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 13

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 15

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 14

Farhadi A, Endres I, Hoiem D, Forsyth D (2009) Describing objects by their attributes. In: Computer Vision and Pattern Recognition (CVPR), Miami, FL 16

Fei-Fei L, Fergus R, Perona P (2006) One-shot learning of object categories. transactions on Pattern Analysis and Machine Intelligence (PAMI) 28(4):594–611 15

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 6

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 16, 20

Frénay B, Verleysen M (2014) Classification in the presence of label noise: a survey. Transactions on Neural Networks and Learning Systems 25(5):845–869 15

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 13

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 abs/1803.01229 11

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 17

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 7, 8

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 13

Gan C, Sun C, Duan L, Gong B (2016a) Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In: European Conference on Computer Vision (ECCV), Amsterdam, Netherlands 15

Gan C, Yao T, Yang K, Yang Y, Mei T (2016b) 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 15

Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M, Lempitsky V (2016) Domain-adversarial training of neural networks. Journal of Machine Learning Research (JMLR) 17(1):2096–2030 11

Gao M, Li A, Yu R, Morariu VI, Davis LS (2018) C-wsl: Count-guided weakly supervised localization. In: European Conference on Computer Vision (ECCV), Munich, Germany 14

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 10

Geiger A, Lenz P, Stiller C, Urtasun R (2013) Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) 32(11):1231–1237 7

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 11

Ghosh A, Kumar H, Sastry P (2017) Robust loss functions under label noise for deep neural networks. In: AAAI, San Francisco, CA 15

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 6

Girshick R (2015) Fast r-cnn. In: International Conference on Computer Vision (ICCV), Santiago, Chile 6

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 5

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 18

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 10

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (????) Generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS), Montreal, Canada 11

Gopalan R, Li R, Chellappa R (2011) Domain adaptation for object recognition: An unsupervised approach. In: International Conference on Computer Vision (ICCV), Barcelona, Spain 10

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 2

Graves A (2013) Generating sequences with recurrent neural networks. CoRR abs/1308.0850 6

Graves A, Jaitly N (2014) Towards end-to-end speech recognition with recurrent neural networks. In: International Conference on Machine Learning (ICML), Beijing, China 6

Gu J, Neubig G, Cho K, Li VO (2017) Learning to translate in real-time with neural machine translation. In: Association of Computational Lingustics (ACL), Vancouver, Canada 6

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 8

Habibian A, Mensink T, Snoek CG (2014) Composite concept discovery for zero-shot video event detection. In: International Conference on Multimedia Retrieval (ICMR), Glasgow, United Kingdom 17

Haeusser P, Frerix T, Mordvintsev A, Cremers D (2017) Associative domain adaptation. In: International Conference on Computer Vision (ICCV), Venice, Italy 10

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 7

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 8

Hariharan B, Girshick RB (2017) Low-shot visual recognition by shrinking and hallucinating features. In: International Conference on Computer Vision (ICCV), Venice, Italy 16

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 5

He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: International Conference on Computer Vision (ICCV), Honolulu, HI 6, 13

Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507 17

Hoffman J, Gupta S, Leong J, Guadarrama S, Darrell T (2016a) Cross-modal adaptation for rgb-d detection. In: International Conference on Robotics and Automation (ICRA), Stockholm, Sweden 12

Hoffman J, Wang D, Yu F, Darrell T (2016b) Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. CoRR abs/1612.02649 11

Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Computer Vision and Pattern Recognition (CVPR), Honolulu, HI 5

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 10

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 14

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 18

Ilse M, Tomczak JM, Welling M (2018) Attention-based deep multiple instance learning. In: International Conference on Machine Learning (ICML), New Orleans, LA 14

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 20

Janai J, Güney F, Behl A, Geiger A (2017) Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. CoRR abs/1704.05519 8

Jayaraman D, Grauman K (2015) Learning image representations tied to ego-motion. In: International Conference on Computer Vision (CVPR), Boston, MA 19

Ji S, Xu W, Yang M, Yu K (2013) 3d convolutional neural networks for human action recognition. Transactions on Pattern Analysis and Machine Intelligence (PAMI) 35(1):221–231 6

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 19

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 17

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 15

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 7

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 13

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 6, 14

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 14

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 11

Kingma DP, Welling M (2013) Auto-encoding variational bayes. In: International Conference on Learning Representations (ICLR), Scottsdale, AZ 18

Koch G, Zemel R, Salakhutdinov R (2015) Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop, Lille, France 16

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. International Journal of Computer Vision (IJCV) 123(1):32–73 2

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS), Stateline, NV 2, 5

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 10

Kumar MP, Packer B, Koller D (2010) Self-paced learning for latent variable models. In: Advances in Neural Information Processing Systems (NIPS), Vancouver, Canada 14

Kurakin A, Goodfellow I, Bengio S (2015) Adversarial examples in the physical world. In: International Conference on Learning Representations (ICLR), San Diego, CA 11

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 abs/1811.00982 2

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 15

Lake BM, Salakhutdinov R, Tenenbaum JB (2015) Human-level concept learning through probabilistic program induction. Science 350(6266):1332–1338 15

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 16

Lampert CH, Nickisch H, Harmeling S (2014) Attribute-based classification for zero-shot visual object categorization. Transactions on Pattern Analysis and Machine Intelligence (PAMI) 36(3):453–465 16

Larsson G, Maire M, Shakhnarovich G (2017) Colorization as a proxy task for visual understanding. In: Computer Vision and Pattern Recognition (CVPR), Honolulu, HI 18

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 2

Lee HY, Huang JB, Singh M, Yang MH (2017) Unsupervised representation learning by sorting sequences. In: International Conference on Computer Vision (ICCV), Venice, Italy 19

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 10

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 8

Li W, Duan L, Xu D, Tsang IW (2014) Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. Transactions on Pattern Analysis and Machine Intelligence 36(6):1134–1148 10

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 10

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 13, 14

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 Recongition (CVPR), Honolulu, HI 6

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 2

Liu B, Ferrari V (2017) Active learning for human pose estimation. In: International Conference on Computer Vision (ICCV), Venice, Italy 13

Liu MY, Tuzel O (2016) Coupled generative adversarial networks. In: Advances in Neural Information Processing Systems (NIPS), Barcelona, Spain 6, 11

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 12

Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Computer Vision and Pattern Recognition (CVPR), Boston, MA 6

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 10

Lu Y, Tai YW, Tang CK (2018) Attribute-guided face generation using conditional cyclegan. In: European Conference on Computer Vision (ECCV), Munich, Germany 11

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 19

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 14

Mehrotra A, Dukkipati A (2017) Generative adversarial residual pairwise networks for one shot learning. CoRR abs/1703.08033 16

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 18

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 16

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 15

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 19

Natarajan N, Dhillon IS, Ravikumar PK, Tewari A (2013) Learning with noisy labels. In: Advances in Neural Information Processing Systems (NIPS), Stateline, NA 3

Nguyen HV, Ho HT, Patel VM, Chellappa R (2015) Dash-n: Joint hierarchical domain adaptation and feature learning. IEEE Transactions on Image Processing 24(12):5479–5491 10

Noroozi M, Favaro P (2016) Unsupervised learning of visual representations by solving jigsaw puzzles. In: European Conference on Computer Vision (ECCV), Amsterdam, Netherlands 18

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 10
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 19
Pan SJ, Yang Q, et al. (2010) A survey on transfer learning. Transactions on Knowledge and Data Engineering 22(10):1345–1359 10
Pan SJ, Tsang IW, Kwok JT, Yang Q (2011) Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks 22(2):199–210 10
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 14
Papadopoulos DP, Uijlings JR, Keller F, Ferrari V (2017a) Extreme clicking for efficient object annotation. In: International Conference on Computer Vision (ICCV), Venice, Italy 14
Papadopoulos DP, Uijlings JR, Keller F, Ferrari V (2017b) Training object class detectors with click supervision. In: Computer Vision and Pattern Recognition (CVPR), Honolulu, HI 14
Patel VM, Gopalan R, Li R, Chellappa R (2015) Visual domain adaptation: A survey of recent advances. Signal Processing Magazine 32(3):53–69 10
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 18
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 18
Peng KC, Wu Z, Ernst J (2018) Zero-shot deep domain adaptation. In: European Conference on Computer Vision (ECCV), Munich, Germany 20
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 8
Pinheiro PO, Collobert R (2015) From image-level to pixel-level labeling with convolutional networks. In: Computer Vision and Pattern Recognition (CVPR), Boston, MA 13, 14
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 18
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 12
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 17
Qiu W, Yuille A (2016) Unrealcv: Connecting computer vision to unreal engine. In: European Conference on Computer Vision (ECCV), Amsterdam, Netherlands 8
Rader N, Bausano M, Richards JE (1980) On the nature of the visual-cliff-avoidance response in human infants. Child Development 51(1):61–68 2

Raj A, Namboodiri VP, Tuytelaars T (2015) Subspace alignment based domain adaptation for rcnn detector. In: British Machine Vision Conference (BMVC), Swansea, United Kingdom 10

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 6

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 2

Ravi S, Larochelle H (2017) Optimization as a model for few-shot learning. In: International Conference on Learning Representations (ICLR), Toulon, France 16

Redko I, Habrard A, Sebban M (2017) Theoretical analysis of domain adaptation with optimal transport. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML KDD), Skopje, Macedonia 11

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 6

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 15

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 12, 16

Remez T, Huang J, Brown M (2018) Learning to segment via cut-and-paste. In: European Conference on Computer Vision (ECCV), Munich, Germany 8

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 6

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 3, 8

Richter SR, Hayder Z, Koltun V (2017) Playing for benchmarks. In: International conference on computer vision (ICCV), Venice, Italy 8

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 16

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 6

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 8

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 13

Roy S, Unmesh A, Namboodiri VP (2018) Deep active learning for object detection. In: Brisitsh Machine Vision Conference (BMVC), Newcastle, United Kingdom 14

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. International Journal of Computer Vision (IJCV) 115(3):211–252 2, 14

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 11

Sadeghi F, Levine S (2017) Cad2rl: Real single-image flight without a single real image. In: Robotics Science and Systems (RSS), Boston, MA 8

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 3

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 6

Sakaridis C, Dai D, Van Gool L (2018) Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision 126:973–992 8

Salakhutdinov R, Larochelle H (2010) Efficient learning of deep boltzmann machines. In: International Conference on Artificial Intelligence and Statistics (ICAIS), San Diego, CA 18

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 11

Scheffer T, Decomain C, Wrobel S (2001) Active hidden markov models for information extraction. In: International Symposium on Intelligent Data Analysis, Berlin, Heidelberg 13

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 11

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 16

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 13

Settles B (2009) Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison 13

Shao L, Zhu F, Li X (2015) Transfer learning for visual categorization: A survey. IEEE transactions on neural networks and learning systems 26(5):1019–1034 10

Shi M, Ferrari V (2016) Weakly supervised object localization using size estimates. In: European Conference on Computer Vision (ECCV), Amsterdam, Netherlands 14

Shimodaira H (2000) Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of statistical planning and inference 90(2):227–244 2

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 11

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR), San Diego, CA 5

Singh S, Gupta A, Efros AA (2012) Unsupervised discovery of mid-level discriminative patches. In: European Conference on Computer Vision (ECCV), Firanze, Italy 17

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 17

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 16

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 12

Song HO, Girshick R, Jegelka S, Mairal J, Harchaoui Z, Darrell T (2014a) On learning to localize objects with minimal supervision. In: International Conference on Machine Learning (ICML), Beijing, China 14

Song HO, Lee YJ, Jegelka S, Darrell T (2014b) Weakly-supervised discovery of visual pattern configurations. In: Advances in Neural Information Processing Systems (NIPS), Montreal, Canada 14

Stavens D, Thrun S (2006) A self-supervised terrain roughness estimator for off-road autonomous driving. In: Uncertainty in Artificial Intelligence (UAI), Cambridge, MA 19

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 15

Sun B, Saenko K (2016) Deep coral: Correlation alignment for deep domain adaptation. In: European Conference on Computer Vision (ECCV), Amsterdam, Netherlands 10

Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Advances in Neural Information Processing Systems (NIPS), Montreal, Canada 6

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 5

Taigman Y, Polyak A, Wolf L (2017) Unsupervised cross-domain image generation. In: International Conference on Learning Representations (ICLR), Toulon, France 12

Tan B, Zhang Y, Pan SJ, Yang Q (2017) Distant domain transfer learning. In: AAAI, San Francisco, CA 12

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 7
Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, Borth D, Li LJ (2016) Yfcc100m: The new data in multimedia research. Communuications of the ACM 59:64–73 14
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 8
Tong S, Chang E (2001) Support vector machine active learning for image retrieval. In: ACM International Conference on Multimedia (MM), Ottawa, Canada 3
Torralba A, Efros AA (2011) Unbiased look at dataset bias. In: Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO 2, 6
Toshev A, Szegedy C (2014) Deeppose: Human pose estimation via deep neural networks. In: Computer Vision and Pattern Recognition (CVPR), Columbus, OH 6
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 6
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 Woskshops (CVPRW), Salt Lake City, UT 8
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 12
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 10
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 10
Tzeng E, Hoffman J, Saenko K, Darrell T (2017) Adversarial discriminative domain adaptation. In: Computer Vision and Pattern Recognition (CVPR), Honolulu, HI 11
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 abs/1609.03499:125 6
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 15
Varma G, Subramanian A, Namboodiri A, Chandraker M, Jawahar C (2018) Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In: Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV 3
Vazquez D, Lopez AM, Marin J, Ponsa D, Geronimo D (2014) Virtual and real world adaptation for pedestrian detection. Transactions on Pattern Analysis and Machine

Intelligence (PAMI) 36(4):797–809 7
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 15
Vezhnevets A, Buhmann JM, Ferrari V (2012) Active learning for semantic segmentation with expected change. In: Computer Vision and Pattern Recognition (CVPR), Providence, RI 13
Vijayanarasimhan S, Grauman K (2014) Large-scale live active learning: Training object detectors with crawled data and crowds. International Journal of Computer Vision (IJCV) 108(1-2):97–114 13
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 16
Vogt P, Smith Adm (2005) Learning colour words is slow: a cross-situational learning account. Behavioral and Brain Sciences 28(4):509510 2
Wang C, Mahadevan S (2011) Heterogeneous domain adaptation using manifold alignment. In: International Joint Conference on Artificial Intelligence (IJCAI), Barcelona, Spain 10
Wang M, Deng W (2018) Deep visual domain adaptation: A survey. Neurocomputing 312:135–153 12
Wang X, Gupta A (2015) Unsupervised learning of visual representations using videos. In: International Conference on Computer Vision (ICCV), Santiago, Chile 19
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 15
Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. Journal of Big Data 3(1):9 10
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 14
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 abs/1609.08144 6
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 16
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 17
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 15

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 2

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 10

Yao A, Gall J, Leistner C, Van Gool L (2012) Interactive object detection. In: Computer Vision and Pattern Recognition (CVPR), Providence, RI 13

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 11

Yoo D, Fan H, Boddeti VN, Kitani KM (2018) Efficient k-shot learning with regularized deep networks. In: AAAI, New Orleans, LA 11, 16

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 10

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 12

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 12

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 16, 17

Zhang R, Isola P, Efros AA (2016) Colorful image colorization. In: European Conference on Computer Vision (ECCV), Amsterdam, Netherlands 18

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 18

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 3, 12

Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: Computer Vision and Pattern Recognition (CVPR), Honolulu, HI 6

Zhu JJ, Bento J (2017) Generative adversarial active learning. In: Advances in Neural Information Processing Systems Workshops, Long Beach, CA 14

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 11

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 10

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 15