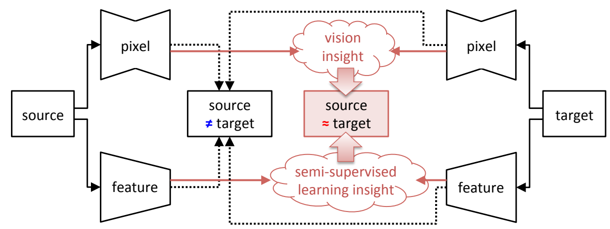
Gotta Adapt 'Em All: Joint Pixel and Feature-Level Domain Adaptation for Recognition in the Wild

Luan Tran[1*] Kihyuk Sohn[2] Xiang Yu[2] Xiaoming Liu[1] Manmohan Chandraker[2,3]
[1]Michigan State University [2]NEC Labs America [3]UC San Diego

Abstract

Recent developments in deep domain adaptation have allowed knowledge transfer from a labeled source domain to an unlabeled target domain at the level of intermediate features or input pixels. We propose that advantages may be derived by combining them, in the form of different insights that lead to a novel design and complementary properties that result in better performance. At the feature level, inspired by insights from semi-supervised learning, we propose a classification-aware domain adversarial neural network that brings target examples into more classifiable regions of source domain. Next, we posit that computer vision insights are more amenable to injection at the pixel level. In particular, we use 3D geometry and image synthesis based on a generalized appearance flow to preserve identity across pose transformations, while using an attribute-conditioned CycleGAN to translate a single source into multiple target images that differ in lower-level properties such as lighting. Besides standard UDA benchmark, we validate on a novel and apt problem of car recognition in unlabeled surveillance images using labeled images from the web, handling explicitly specified, nameable factors of variation through pixel-level and implicit, unspecified factors through feature-level adaptation.



Feature \ Pixel	–	CycleGAN	**MKF+AC-CGAN (ours)**
–	55.0	64.3	**79.7**
DANN	60.4	64.8	78.0
DANN-CA (ours)	**75.8**	77.7	**84.2**

Table 1: Our framework for unsupervised domain adaptation at multiple semantic levels: at *feature-level*, we bring insights from semi-supervised learning to obtain highly discriminative domain-invariant representations; at *pixel-level*, we leverage complementary domain-specific vision insights e.g., geometry and attributes. Our joint pixel and feature-level DA demonstrates significant improvement over individual adaptation counterparts as well as other competing methods such as CyCADA (CycleGAN+DANN) [18] on car recognition in surveillance domain under UDA setting. Please see Section 5 for complete experimental analysis.

1. Introduction

Deep learning has made an enormous impact on many applications in computer vision such as generic object recognition [22, 44, 48, 17], fine-grained categorization [59, 21, 41], object detection [26, 27, 28, 36, 37], semantic segmentation [6, 42] and 3D reconstruction [53, 52]. Much of its success is attributed to the availability of large-scale labeled training data [8, 15]. However, this is hardly true in many practical scenarios: since annotation is expensive, most data remains unlabeled. Consider car recognition problem from surveillance images, where factors such as camera angle, distance, lighting or weather condition are different across locations. It is not feasible to exhaustively annotate all these images. Meanwhile, there exists abundant labeled data from web domain [21, 62, 12], but with very different image characteristics that precludes direct transfer of discriminative CNN-based classifiers. For instance, web images might be from catalog magazines with professional lighting and ground-level camera poses, while surveillance images can originate from cameras atop traffic lights with challenging lighting and weather conditions.

Unsupervised domain adaptation (UDA) is a promising tool to overcome the lack of labeled training data problem in target domains. Several approaches aim to match distributions between source and target domains at different levels of representations, such as feature [57, 56, 11, 45, 31] or pixel levels [49, 43, 66, 3]. Certain adaptation challenges are better handled in the feature space, but feature-level DA is a black-box algorithm for which adding domain-specific insights during adaptation is more difficult than in pixel space. On the contrary, pixel space is much higher-dimensional and the optimization problem is under-determined. How to

*This work is done when L. Tran was an intern at NEC Labs America.

effectively combine them has become an open challenge.

In this work we address this challenge by leveraging complementary tools that are better-suited at each level (see figure in Table 1). Specifically, we posit that feature-level DA is more amenable to techniques from semi-supervised learning (SSL), while pixel-level DA allows domain-specific insights from computer vision. In Section 3, we present our feature-level DA method called classification-aware domain adversarial neural network (DANN-CA) that jointly parameterizes the classifier and domain discriminator inspired by an instance of SSL algorithm [40]. We show this to be a generalization of DANN [11] to incorporate constraints (Fig. 1) that guide discriminator to easily find major modes corresponding to classes in the feature space, and in turn put target examples into more classifiable regions via adversarial loss.

A challenge for pixel-level DA is to simultaneously transform source image properties at multiple semantic levels. In Section 4, we present pixel-level DA by image transformations that make use of vision concepts to deal with different variation factors, such as photometric or geometric transformations (Fig. 2),[1] for recognition in surveillance domain. To handle low-level transformations, we propose an attribute-conditioned CycleGAN (AC-CGAN) that extends [66] to generate multiple target images with different attributes. To handle high-level identity-preserving pose transformations, we use an appearance flow (AF) [65], an warping-based image synthesis tool. To reduce semantic gaps between synthetic and real images, we propose a generalization of AF with 2D keypoints [25] as a domain bridge.

In Section 5, we evaluate our framework on car recognition in surveillance images from the comprehensive cars (CompCars) dataset [62]. We define an experimental protocol with web images as labeled source domain and surveillance images as unlabeled target domain. We explicitly handle nameable factors of variation such as pose and lighting through pixel-level DA, while other nuisance factors are handled by feature-level DA. As in Table 1, we achieve $\mathbf{84.20\%}$ accuracy, reducing error by $\mathbf{64.9\%}$ from a model trained only on the source domain. We present ablation studies to demonstrate the importance of each adaptation component by extensively evaluating performances with various mixtures of components. We further validate the effectiveness of our proposed feature-level DA methods on standard UDA benchmarks, namely digits and traffic signs [11] and office-31 [38], achieving state-of-the-art recognition performance.

In summary, the contributions of our work are:

- A novel UDA framework that adapts at multiple semantic levels from feature to pixel, with complementary insights for each type of adaptation.
- For feature-level DA, a connection of DANN to a semi-supervised variant, motivating a novel regularization via classification-aware domain adversarial neural network.
- For pixel-level DA, an attribute-conditioned CycleGAN to translate a source image into multiple target images with different attributes, along with an warping-based image synthesization for identity-preserving pose translations via a keypoint-based appearance flow.
- A new experimental protocol on car recognition in surveillance domain, with detailed analysis of various modules and efficacy of our UDA framework.
- State-of-the-art performance on standard UDA benchmarks, such as office-31 and digits, traffic signs adaptation tasks, with our feature-level DA method.

Due to a large volume of our work, we put additional detail in Section S1–S6 of the supplementary material at www.nec-labs.com/~mas/jointDA.

[1] Our framework is unsupervised DA in the sense that we don't require recognition labels from the target domain for training, but it uses side annotations to inject insights from vision concepts for pixel-level adaptation.

2. Related Work

Unsupervised Domain Adaptation. Following theoretical developments of domain adaptation [2, 1], a major challenge is to define a proper metric measuring the domain difference. The maximum mean discrepancy [29, 57, 9, 56, 47], which measures the difference based on kernels, and the domain adversarial neural network [11, 4, 3, 45, 46], which measures the difference using discriminator, have been successful. Noticing the similarity in problem settings between UDA and SSL, there have been attempts to combine ideas from SSL. For example, entropy minimization [14] has been used in addition to domain adversarial loss [30, 31]. Our feature-level DA is built on DANN by resolving issues of discriminator in discovering modes in the feature space. Our formulation also connects tightly to SSL and we explain why entropy minimization is essential for DANN.

Perspective Transformation. Previous works [61, 23, 51] propose encoder-decoder networks to generate output images of target viewpoint. Adversarial learning for perspective transformation [54, 55, 63] has demonstrated good performance on disentangling viewpoint from other appearance factors, but there are still concept (e.g., class label) switches in unpaired settings. Rather than learning the output distribution, [65, 34] propose an warping-based viewpoint synthesization by estimating a pixel-level flow field. We extend it to improve generalization to real images using synthetic-to-real domain invariant representations such as 2D key points [25].

Image-to-image Translation. With the success of GAN on image generation [13, 35], conditional variants of GAN [32] have been successfully adopted to image-to-image translation problems in both paired [19] and unpaired [43, 49, 66] training settings. Our model extends the work of [66] for image translation in unpaired settings using a control variable or visual attribute [60] to generate multiple outputs.

Multi-level UDA. A combination of pixel and feature level adaptation has been attempted in [18], however, we differ in a few important ways. Specifically, we go further in using insights from SSL that allows novel regularization for feature-level DA, while exploiting 3D geometry and attribute-based conditioning in GANs to simultaneously handle high-level pose and low-level lighting variations. Our experiments include a detailed study of the complementary benefits, as well as the effectiveness of various adaptation modules. While [18] consider problems such as semantic segmentation, we study a car recognition problem that highlights the need for adaptation at various levels. We also demonstrate state-of-the-art results on standard UDA benchmarks.

3. Domain Adversarial Feature Learning

This section describes a classification-aware domain adversarial neural network (Fig. 1(b)) that improves upon a domain adversarial neural network [11] by joint parameterization of classifier and discriminator.

Notation. Let $\mathcal{X}_\mathrm{S}, \mathcal{X}_\mathrm{T} \subset \mathcal{X}$ be source and target datasets and $\mathcal{Y} = \{1, ..., N\}$ be the set for class label. Let $f : \mathcal{X} \to \mathbb{R}^K$ be the feature generator, e.g., CNN, with parameters θ_f that maps input $x \in \mathcal{X}$ into a K-dimensional vector.

3.1. Recap: Domain Adversarial Neural Network

Domain adversarial training [11] aims to adapt classifier learned from the labeled source domain to unlabeled target domain by making feature distributions of the two domains indistinguishable. This is achieved through a domain discriminator $D : \mathbb{R}^K \to (0, 1)$ that tells whether features from the two domains are still distinguishable. Then, f is trained to confuse D while classifying the source data correctly:

$$\max_{\theta_c} \{\mathcal{L}_\mathrm{C} = \mathbb{E}_{\mathcal{X}_\mathrm{S}} \log C(f, y)\} \tag{1}$$

$$\max_{\theta_d} \{\mathcal{L}_\mathrm{D} = \mathbb{E}_{\mathcal{X}_\mathrm{S}} \log(1 - D(f)) + \mathbb{E}_{\mathcal{X}_\mathrm{T}} \log D(f)\} \tag{2}$$

$$\max_{\theta_f} \{\mathcal{L}_\mathrm{F} = \mathcal{L}_\mathrm{C} + \lambda \mathbb{E}_{\mathcal{X}_\mathrm{T}} \log(1 - D(f))\} \tag{3}$$

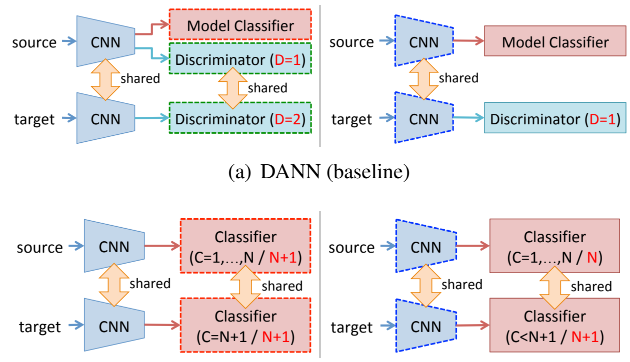
$C : \mathbb{R}^K \times \mathcal{Y} \to (0, 1)$ is a class score function that outputs the probability of an input x being a class y among N categories, i.e., $C(f(x), y) = P(y|f(x); \theta_c)$. λ balances classification and domain adversarial losses. The parameters $\{\theta_c, \theta_d\}$ and $\{\theta_f\}$ are updated in turn using stochastic gradient descent.

3.2. Classification-Aware Adversarial Learning

We note that the problem setup of unsupervised domain adaptation is not different from that of semi-supervised learning once we remove the notion of domains. Inspired by the semi-supervised learning formulation of GANs [40, 7], we propose a new domain adversarial learning objective that jointly parameterizes classifier and discriminator as follows:

$$\max_{\theta_c} \{\overline{\mathcal{L}}_\mathrm{C} = \mathbb{E}_{\mathcal{X}_\mathrm{S}} \log \overline{C}(y) + \mathbb{E}_{\mathcal{X}_\mathrm{T}} \log \overline{C}(N{+}1)\} \tag{4}$$

$$\max_{\theta_f} \{\overline{\mathcal{L}}_\mathrm{F} = \mathbb{E}_{\mathcal{X}_\mathrm{S}} \log \overline{C}(y|\mathcal{Y}) + \lambda \mathbb{E}_{\mathcal{X}_\mathrm{T}} \log(1 - \overline{C}(N{+}1))\} \tag{5}$$



(a) DANN (baseline)

(b) DANN-CA

Figure 1: (a) DANN and (b) classification-aware DANN (DANN-CA) with $(N{+}1)$-way joint parameterization of classifier and discriminator. CNN and classifiers are updated in turn (dotted boxes) while fixing the others (solid boxes).

where we omit $f(x)$ from $\overline{C}(f(x), y)$ for presentation clarity. The score function $\overline{C}$ is defined on $\mathbb{R}^K \times \{1, ..., N{+}1\}$ and the conditional score $\overline{C}(y|\mathcal{Y})$ is written as follows:

$$\overline{C}(y|\mathcal{Y}) = \frac{\overline{C}(y)}{1 - \overline{C}(N{+}1)}, \forall y \leq N, \ \overline{C}(N{+}1|\mathcal{Y}) = 0 \tag{6}$$

The formulation no more has a discriminator, but classifier has one additional output entry for the target domain. We call our model a classification-aware DANN or DANN-CA as it allows discriminator to access to classifier directly. While [40] has demonstrated an effectiveness of joint parameterization in semi-supervised GANs, it is not clearly explained why it is better. In the following, we aim to explain the advantage of DANN-CA in the context of feature-level UDA.

Discriminator Should Know Classification Boundary. Mode collapse is a critical issue in adversarial learning. To prevent it, discriminator needs to discover as many modes in data distribution as possible. While it is difficult to describe the modes in the input space for generative modeling [13], it is relatively easy to characterize the modes in the feature space: there are N major modes, each of which corresponds to each output class, and the discriminator is demanded for discovering these modes in the feature space. Unfortunately, the discriminator of DANN is trained with binary supervision, implying that the mode discovery is done unsupervisedly. On the other hand, the modes are already embedded in the discriminator of DANN-CA via joint parameterization and the adversarial learning can be made easier.

We further investigate the gradient of adversarial loss in (3) and (5) with respect to f. For the ease of presentation, we assume linear classifier and discriminator. Complete derivation including non-linear version is in Section S1.

$$\frac{\partial \log(1 - D(f))}{\partial f} = -\, D(f) w_d \tag{7}$$

$$\frac{\partial \log(1 - \overline{C}(N{+}1))}{\partial f} = -\, \overline{C}(N{+}1)(w_{N+1} - \textstyle\sum_{y=1}^{N} w_y \overline{C}(y|\mathcal{Y}))$$

where $w_d, w_y \in \mathbb{R}^K, y \in \{1, ..., N+1\}$ are discriminator and classifier weights, respectively. As is evident from (7), the adversarial loss of DANN cannot capture multiple modes as all target examples induce the gradient of the same direction. Even if we use MLP discriminator in practice, it still demands to discover modes correspond to classes without supervision. The joint parameterization allows not only to push features away from the target domain, but also guides them to be pulled close to classes based on the conditional score $\overline{C}(y|\mathcal{Y})$ of individual target examples.

Relation to DANN [11].
Besides parameterization, the learning objectives are tightly linked to those of DANN [11]. For instance, $\overline{\mathcal{L}}_{\text{F}} = \mathcal{L}_{\text{F}}$ with $D = \overline{C}(N+1)$ and $C(y) = \overline{C}(y|\mathcal{Y})$. It is also easy to show $\overline{\mathcal{L}}_{\text{C}} = \mathcal{L}_{\text{C}} + \mathcal{L}_{\text{D}}$ by rewriting $\overline{C}(y)$ using (6) as follows:

$$\overline{\mathcal{L}}_{\text{C}} = \mathbb{E}_{\mathcal{X}_{\text{S}}} \log \overline{C}(y|\mathcal{Y}) + \mathbb{E}_{\mathcal{X}_{\text{S}}} \log(1-\overline{C}(N+1)) + \mathbb{E}_{\mathcal{X}_{\text{T}}} \log \overline{C}(N+1) \quad (8)$$

Relation to Maximum Classifier Discrepancy [39].
We also relate our proposed DANN-CA to recently proposed maximum classifier discrepancy (MCD) learning for UDA. MCD learns shared feature extractor by reducing the prediction discrepancy between two (or more) maximally different classifiers. We show that our DANN-CA can be understood as MCD with choices of classifiers and the divergence. Following [39], we define the two classification distributions:

$$p_1(y|x_t) = \overline{C}(y|\mathcal{Y}),\ p_2(y|x_t) = \overline{C}(y),\ y \leq N+1 \quad (9)$$

Note that two classifiers F_1 and F_2 in [39] are both represented as $(N+1)$-way classifier. Using KL divergence, we obtain following discrepancy loss:

$$-\,\text{KL}(p_1 \| p_2) = \log(1-\overline{C}(N+1)) \quad (10)$$

which is equivalent to the adversarial loss in (5). This analysis provides a unified view of DANN, MCD and more general class of consistency-based SSL algorithms [24, 50, 10]. A theoretical comparison of UDA algorithms is important as empirical comparison could sometimes be misleading [33]. A full derivation of (10) and analysis are in Section S2.

4. Pixel-level Cross-Domain Image Translation

As is common for neural networks, DANN is a black-box algorithm and adding domain-specific insight is non-trivial. On the other hand, certain challenges in DA can be better handled in image space. In this section, we introduce complementary tools to deal with nameable factors of variation, such as photometric or perspective transformations, at the pixel level. To achieve this, we propose extensions to prior works on CycleGAN [66] and appearance flows [65]. We describe with an illustrative application of car recognition in surveillance domain where the only labeled data is from web domain. The pipeline of our system is in Fig. 2.

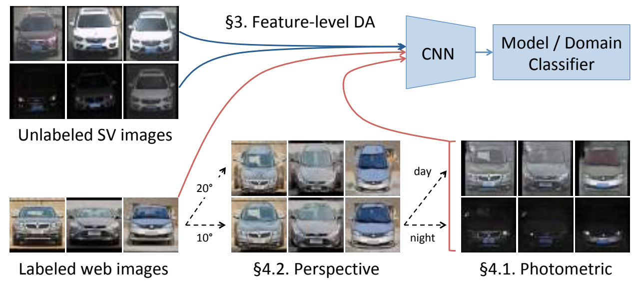


Figure 2: Overview of our car recognition system using labeled web and unlabeled surveillance (SV) images. Images taken by SV cameras are different from web images in nameable factors, such as viewpoint or lighting conditions as well as other nuisance factors. We integrate pixel-level DA for perspective and photometric transformations and feature-level DA for other nuisance factors.

4.1. Photometric Transformation by CycleGAN

As noticed from Fig. 2, images from surveillance domain have disparate color statistics from web images as they might be acquired outdoors at different times with significant lighting variations. CycleGAN [66] is proposed as a promising tool for image translation by disentangling low-level statistics from geometric structure. A limitation, however, is that it generates a single output when there could be multiple output styles. We propose an attribute-conditioned CycleGAN (AC-CGAN) that generates diverse output images with the same geometric structure by incorporating a conditioning variable into generators.

Let $\mathcal{A}$ be a set of attributes in the target domain (day or night). We learn a generator $G : \mathcal{X}_{\text{S}} \times \mathcal{A} \rightarrow \mathcal{X}_{\text{T}}$ that translates an image with certain style $a \in \mathcal{A}$ by fooling an attribute-specific discriminator D_a. The learning objectives are:

$$\max_{\theta_{d_a}} \{\mathcal{L}_{D_a} = \mathbb{E}_{\mathcal{X}_{\text{T}_a}} \log D_a(x) + \mathbb{E}_{\mathcal{X}_{\text{S}}} \log(1-D_a(G(x, a)))\} \quad (11)$$

$$\max_{\theta_g} \{\mathcal{L}_G = \mathbb{E}_{\mathcal{X}_{\text{S}}} \mathbb{E}_{\mathcal{A}} \log D_a(G(x, a))\} \quad (12)$$

We use multiple discriminators to prevent competition between different attribute configurations, but it is feasible to have one discriminator with $(|\mathcal{A}|+1)$-way domain classification loss [49]. Also, one might afford to have multiple generators per attribute without sharing parameters.[2] Following [66], we add cycle consistency loss as follows:

$$\mathbb{E}_{\mathcal{X}_{\text{S}}} \|F(G(x,a),a)-x\|_1 + \mathbb{E}_{\mathcal{X}_{\text{T}_a}} \|G(F(x,a),a)-x\|_1 \quad (13)$$

where an inverse generator F maps outputs back to source domain $F(G(x, a), a) = x$. We also use patchGAN [19, 66] for discriminators that makes real or fake decisions from local patches and UNet [19] for generators, each of which contributes to preserve geometric structure of an input image.

[2] Empirically, using two separate generators for day and night performs slightly better than a single generator. Please see Section S6 for results.

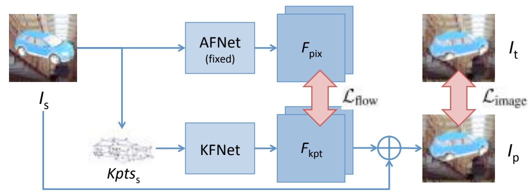


Figure 3: Training framework of keypoint-based appearance flow network (KFNet) by distilling knowledge from pretrained AFNet.

4.2. Perspective Synthesis by Appearance Flow

Besides color statistics, we observe significant differences in camera perspective (Fig. 2). In this section, we deal with perspective transformation using an image warping based on a pixel-wise dense flow called appearance flow (AF) [65]. Specifically, we propose to improve the generalization of AF estimation network (AFNet) trained on 3D CAD rendered images to real images by utilizing a robust representation across synthetic and real domains, i.e. 2D keypoints.

Appearance Flow.
Zhou et al. [65] propose to estimate a pixel-level dense flow from an input image with target viewpoint and synthesize an output by reorganizing pixels using bilinear sampling [20]:

$$I_p^{i,j} = \sum_{(h,w)\in N} I_s^{h,w}(1-|F_y^{i,j}-h|)(1-|F_x^{i,j}-w|), \quad (14)$$

where I_s, I_p are input and output, (F_x, F_y) is a pixel-level flow field in horizontal and vertical axes called appearance flow (AF), estimated by an AF estimation network (AFNet). N denotes 4-pixel neighborhood of $(F_x^{i,j}, F_y^{i,j})$. In contrast to neural network based image synthesization methods [51], AF-based transformation may have a better chance of preserving object identity since all pixels of an output image are from an input image and no new information, such as learned priors in the decoder network, is introduced.

Keypoint-based Robust Estimation of AF.
AFNet requires image pairs (I_s, I_t) with perspective being the only factor of variation for training. Since it is infeasible to collect precisely controlled dataset of real images at large-scale, rendered images from 3D CAD models are used. However, this induces a generalization issue when applied to real images at test time.

To make AFNet generalizable, we propose sparse 2D keypoints in replace of an RGB image as an input to AFNet both at train and test times. Although sparse, for objects like cars, we argue that 2D keypoints contain sufficient information to reconstruct (rough) geometry of an entire object, while being invariant across rendered and real domains. Besides, keypoint estimation can be done robustly across synthetic and real domains even when the keypoint localization network is trained only on the synthetic data [25]. To this end, we propose a 2D keypoint-based AFNet (KFNet) that takes estimated 2D keypoints and the target viewpoint as an input pair to generate flow fields F for synthesization.

The KFNet is trained using rendered image pairs. Moreover, we leverage pretrained AFNet that produces a robust AF representation for rendered images to train the KFNet by distillation. The learning objective is as follows:

$$\min\{\mathcal{L} = \|F_{\text{kpt}} - F_{\text{pix}}\|_1 + \lambda\|I_p(F_{\text{kpt}}, I_s) - I_t\|_1\} \quad (15)$$

where F_{kpt} is an estimated appearance flow by KFNet and F_{pix} is that by AFNet. Here, $I_p(F, I_s)$ is the predicted image from I_s using F based on (14). The training framework by distillation is visualized in Fig. 3.

5. Experiments

We strive for providing empirical evidence for the effectiveness of individual components of our proposed framework as well as their complementarity by conducting extensive experiments on car recognition in surveillance domain. For feature-level adaptation, we also provide performance comparison on standard benchmarks, namely digits and traffic signs [11] and office-31 [38].

5.1. Car Recognition in Surveillance Domain

Dataset. CompCars dataset [62] offers two datasets, one from the web and the other from the surveillance (SV) domains. It contains $52,083$ web images across 431 car models and $44,481$ SV images across 181 car models. Samples are in Fig. 2. The SV test set contains $9,630$ images across 181 car models, of which $6,404$ images are in day condition.[3]

To train an appearance flow estimation network, based on emperical distribution of web images, we render car images at multiple elevation ($0°\sim30°$) and azimuth variations ($\pm15°$) from ShapeNet [5]. We apply pixel-level adaptation to $5,508$ web images of frontal view.

Training. The task is to train a classifier that works well on SV images using labeled web (source) and unlabeled SV (target) images. We use ResNet-18 [17] fine-tuned on web images as our baseline. Then, we train models with different integration of pixel and feature-level DA components. Note that synthesized images by pixel-level adaptation are considered as labeled training examples. Furthermore, we use data augmentation, such as translation, horizontal flip or chromatic jitter, for all models by default. We refer to Section S4.3 for more training details.

Model Selection. While it is desirable to do a model selection without labeled examples from the target domain, to our knowledge, there does not exist an unsupervised evaluation measure that is highly correlated with the supervised performance [4]. To allow more meaningful and interpretable comparisons across different methods, we report our results based on a supervised model selection [4] using a small validation set containing approximately 5 labeled examples per

[3] We provide a binary label (day or night) for images from surveillance domain by computing the mean pixel-intensity.

ID	Perspective Transformation	SV	Day	Night
M1	Baseline (web only)	54.98	72.67	19.87
M3	Appearance Flow (AF)	59.73	75.78	27.87
M4	Keypoint-based AF (KF)	61.55	77.98	28.92
M5	KF with mask (MKF)	64.30	78.62	35.87

Table 2: Accuracy on SV test set with different perspective transformation methods: appearance flow (AF), keypoint-based AF (KF) and with mask (MKF).

ID	Photometric Transformation	SV	Day	Night
M1	Baseline (web only)	54.98	72.67	19.87
M6	CycleGAN	64.32	77.01	39.12
M7	AC-CGAN	67.30	78.20	45.66
M8	MKF+CycleGAN	71.21	81.54	50.68
M9	MKF+AC-CGAN	79.71	84.10	70.99

Table 3: Accuracy on SV test set with different photometric transformation methods: CycleGAN [66], attribute-conditioned CycleGAN (AC-CGAN), and combinations with MKF.

ID	Pixel	Feature	SV	Day	Night
M1	Baseline (web only)		54.98	72.67	19.87
M2	Supervised (web+SV)		98.63	98.92	98.05
M10	–	DANN	60.40	75.56	30.31
[18]	CycleGAN	DANN	64.82	76.35	41.93
M11	–	DANN-CA	75.83	76.73	74.05
M12	MKF	DANN-CA	80.40	82.50	76.22
M13	AC-CGAN	DANN-CA	80.24	82.15	76.44
M14	MKF+AC-CGAN	DANN-CA	**84.20**	**85.77**	**81.10**

Table 4: Accuracy on SV test set with pixel and feature-level DA components. We consider an MKF for perspective and attribute-conditioned CycleGAN (AC-CGAN) for photometric transformations for pixel-level DA, and DANN-CA for feature-level DA.

class from the target domain. We provide a comprehensive comparison to unsupervised model selection using a variant of reverse validation [64, 11] in Section S3.

5.2. Summary Results

We report the classification accuracy on the surveillance test set in Tables 2 to 4. Noticing a huge accuracy drop on night images, we also report accuracy of individual day and night sets. We present t-SNE [58] plots of web (blue), day (red) and night (green) images in Fig. 4 and Fig. 8.

Firstly, although achieving state-of-the-art accuracy on the web test set (96.4% vs 91.2% [62]), the baseline model trained only on web images suffers from generalization to SV images, resulting in only 54.98% accuracy. Comparing to the performance of the model trained with target domain supervision (98.65% in Table 4) provides a sense of how different two domains are. While the baseline adaptation model, DANN (M10 in Table 4), achieves only 58.80%, the proposed joint pixel and feature-level adaptation method achieves **84.20**%, reducing the error by **64.9**% from the baseline M1. While the use of baseline pixel (CycleGAN) and feature-level (DANN) DA methods as in [18] demonstrates moderate improvement (64.82%) over the baseline, this is far below our proposed DA framework. In the following, we present comprehensive studies on the contribution of individual components and their complementarity.

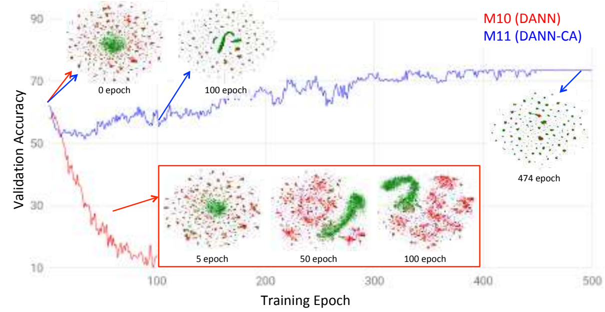


Figure 4: Accuracy of DANN (M10) and DANN-CA (M11) on SV validation set over training. We also visualize t-SNE plots of each model at different training epochs.

Method	M→MM	S→S	S→M	M→S	S→G
Source only	67.90	87.05	63.74	62.44	94.53
DANN	**98.00**	92.24	88.70	82.30	97.38
DANN-CA	**98.03**	**94.47**	**96.23**	**87.48**	**98.70**

Table 5: Evaluation on UDA tasks [11], such as MNIST to MNIST-M (M→MM), Synthetic Digits to SVHN (S→S), SVHN to MNIST (S→M), MNIST to SVHN (M→S), or Synthetic Signs to GTSRB (S→G). Test set accuracy averaged over 10 runs is reported. The best performers and the ones within standard error are bold-faced.

Method	A→W	D→W	W→D	A→D	D→A	W→A
Source only	76.42	96.76	97.99	79.81	60.44	59.53
DANN	85.97	96.87	97.94	84.12	67.63	66.78
DANN-CA	**91.35**	**98.24**	**99.48**	**89.94**	**69.63**	**68.76**

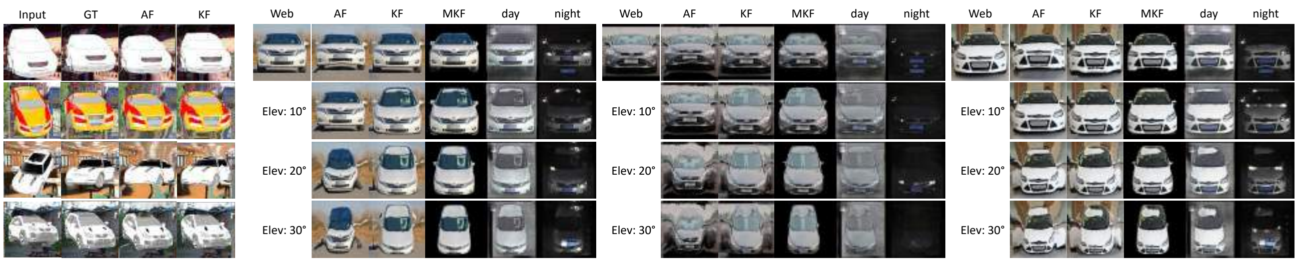
Table 6: Evaluation on office-31 benchmark [38] between Amazon (A), DSLR (D), and Webcam (W) domains using ResNet-50. Target domain accuracy averaged over 5 runs is reported. The best performers and the ones within standard error are bold-faced.

5.3. Analysis on Pixel-level Adaptation

This section contributes to the analysis of our pixel-level DA on dealing with perspective and photometric transformations, typical factors of variation introduced in SV domain.

Perspective Transformation with CycleGAN [66]. The success of CycleGAN on image translation is attributed by few factors, such as cycle consistency loss, patch-based discriminator, or generator with skip connection. However, these constraints may be too strong to translate viewpoint. As is evident from Fig. 6, the output of CycleGAN (second row) maintains the geometric structure of the input (first row) faithfully but fails at adapting to the viewpoint of SV domain. Relaxing constraints, such as removing skip connections of generator and increasing receptive field size of patch-based discriminator, allows perspective adaptation possible (third row), but we lose many details crucial for recognition tasks.

Our approach solves the challenge by translating images



(a) Persp. on rendered data

(b) Perspective ($0^\circ \sim 30^\circ$) and photometric (day, night) transformations on real data

Figure 5: Synthesized images by (a) perspective on rendered images of 3D CAD models and (b) perspective and photometirc transformations on real images from CompCars dataset. (a) From left to right: input, GT of target view, and perspective transformed images using AFNet and sparse 2D keypoint-based AFNet (KFNet). (b) From left to right: for each web image, perspective transformed images using AFNet, KFNet and its masked output (MKF), followed by photometric transformation into day and night by AC-CGAN.

Figure 6: Web to SV (day) translation using CycleGAN (second) and its variant (third) by removing skip connection from generator and increasing receptive filed size for patch discriminator. On the right, we overlay left half of translated images with SV image to highlight the impact of constraints on perspective transformation.

Figure 7: Continuous interploation of latent code of AC-CGAN.

in two steps, resulting in high-quality image synthesis from web to SV domain as in Fig. 5(b). The conclusion from our visual investigation aligns with the recognition performance, where combined perspective transformation and CycleGAN (M8) achieves 71.21%, which improves upon a model without perspective transformation (M6, 64.32%) in Table 3 or a model without CycleGAN (M5, 64.30%) in Table 2.

Disentangling Illumination via AC-CGAN.
The AC-CGAN fixes the unimodal translation nature of CycleGAN with a latent code [60]. This allows learning disentangled representation from an attribute, which in our case the illumination, and as a result, we can synthesize images of the same car with different illumination conditions, as in Fig. 5(b). Moreover, the continuous interpolation of latent code allows to generate continuous change in illumination factor (e.g., color tone, pixel intensity of headlight) without changing the shape and appearance of each car, as in Fig. 7.

Generating images with diverse illumination conditions improves the recognition accuracy as in Table 3, especially on the night images of SV domain. The AC-CGAN (M7) improves by 2.98% upon the CycleGAN (M6). Moreover, when combined with perspective transformation (M8 and M9), we observe a larger increase in improvement of 8.50%.

Comparison between AFNet and KFNet.
KFNet is developed to improve the generalization of AFNet to real images. Before comparing these models on them, we evaluate KFNet on rendered images from 3D CAD models to demonstrate comparable performance to AFNet. We show inputs, output targets and transformed images by AFNet and KFNet in Fig. 5(a). We observe reliable estimation of appearance flow by KFNet. Furthermore, we obtain 0.072 per-pixel L1 reconstruction error between rendered output images and perspective transformed images at four elevations (0° to 30°) using KFNet, which is comparable to 0.071 error of AFNet (pixel values are normalized to $[0, 1]$).

Now, we show results on real images in Fig. 5(b). AFNet struggles to generalize on real images and generates distorted images with incorrect target elevation. Although sparse, 2D keypoints are more robust to domain shift from synthetic to real and are sufficient to preserve the object geometry and correctly transform to the target perspective. Finally, better recognition performance on SV domain of the network trained with source and the perspective transformed images ($59.73\% \rightarrow 61.55\%$ from M3 to M4 in Table 2) implies the superiority of the proposed KFNet.

5.4. Analysis on Feature-level Adaptation

We demonstrate the superiority of the proposed DANN-CA to the DANN on car recognition and other UDA tasks.

Evaluation on Car Recognition in SV Domain.
Note that, on top of 512-dim features, the linear classifier ($512-431/432$) is used for both models, while we use the 3-layer MLP ($512-320-320-1$) for the discriminator of DANN after trying several discriminator architectures with

different depths. As in Table 4, the improvement of DANN-CA is larger than that of DANN, confirming the superiority of the proposed method. We further investigate the behavior of these methods from training curves in Fig. 4. The DANN starts to drop significantly after few epochs of adversarial training, remaining with a few collapsed modes in the end. While it shows some fluctuations at the beginning of training, DANN-CA shows clear progression over training and finally reaches at convergence.

Evaluation on UDA Benchmarks.
We also evaluate the performance of DANN and our DANN-CA on UDA benchmarks. For digits and traffic signs tasks, we use data augmentation as in [16]. Due to space constraint, we provide more details on experimental setting and comparison to other methods in Section S5. As we see in the summary results of Table 5 and 6, our proposed DANN-CA outperforms the DANN on all tasks and sometimes by a huge margin. We remind that the only difference between the two methods is the parameterization of the classifier and discriminator, and it clearly shows the importance of joint parameterization in adversarial domain adaptation.

5.5. Analysis on Joint PnF Adaptation

Finally, we provide an empirical analysis on the proposed joint pixel and feature-level (PnF) adaptation. In the joint framework, we train models with feature-level adaptation methods using unlabeled target domain and expanded labeled source domain including original source images and synthesized images by pixel-level DA.

Improved Domain Alignment with Feature-level DA.
While it allows high-fidelity generation, constraints in the pixel-level DA make it hard to faithfully adapt to the target domain. It is evident from Fig. 8 where t-SNE plot of M9 is less clean than that of M11. This implies that the role of feature-level DA in joint DA framework is to learn remaining factors not yet discovered by the pixel-level DA.

Improved Training Stability with Pixel-level DA.
We delve deeper into understanding the interplay between pixel and feature-level DAs. Fig. 8 shows accuracy curves of pixel-level (M9), feature-level (M11) and joint (M14) DA models on day (dotted) and night (solid) of SV validation sets. While the accuracy on days are stable for all models, we observe a large up-and-down for curve on nights of M11. Note that the fluctuation in the night curve of M9 is not as significant. This is due to many constraints (e.g., warp-based viewpoint synthesis, cycle-consistency or UNet architecture) imposed on the training of pixel-level DA, allowing high-fidelity translation of perspective and illumination variations whose outputs are closer to the target domain than the source examples. Consequently, M14 shows significantly less fluctuation during the training than M11.

We further study the training stability from the mode coverage perspective. Assuming modes correspond to classes in the feature space, the number of classes that are not assigned as top-1 prediction by any of SV test set images is used as a proxy to mode coverage. We provide results in Table 7. While M11 has 29.6 classes on average over 5 runs with no assigned SV image, only 2 classes are missing for M9. The pixel-level DA effectively complements the mode collapse of adversarial learning in the feature-level DA, reducing the number of missing modes to 10.4 for M14.

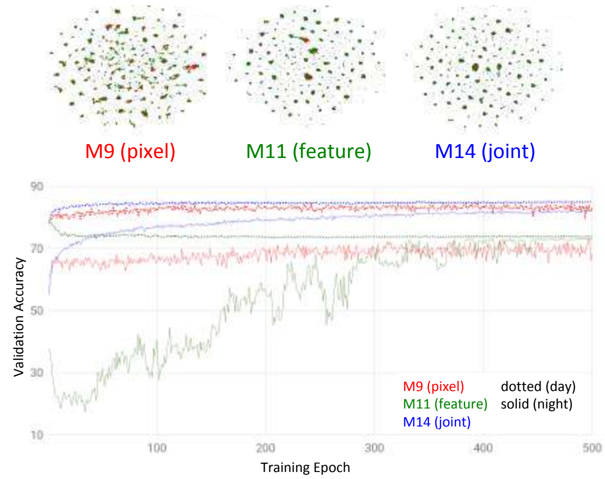


Figure 8: Accuracy curves on day (dotted) and night (solid) SV validation set over training and t-SNE plots of pixel-level (M9), feature-level (M11) and joint (M14) DA models.

	M9 (pixel)	M11 (feature)	M14 (joint)
# missing modes	2	29.6±1.1	10.4±0.6

Table 7: Number of missing modes (classes) out of 181 classes.

Complementarity of Components.
To summarize, each module has its own disadvantage, such as training instability for feature-level DA and the lack of adaptation flexibility for pixel-level DA. Our empirical analysis suggests that these shortages can be complemented when combined in a unified framework, improving the accuracy by 4.49% and 8.37% upon individual modules, respectively.

6. Conclusion

With an observation that certain adaptation challenges are better handled in feature space and others in pixel space, we propose a joint UDA framework by leveraging complementary tools that are better-suited for each type of adaptation challenge. Importance and complementarity of each component are demonstrated through extensive experiments on a novel application of car recognition in surveillance domain. We also demonstrate state-of-the-art performance on UDA benchmarks with our proposed feature-level DA methods.

Acknowledgement

We thank Paul Vernaza, Gaurav Sharma, Wongun Choi, and Wenling Shang for helpful discussions.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010. 2

[2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NeurIPS*, 2007. 2

[3] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, July 2017. 1, 2

[4] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *NeurIPS*, 2016. 2, 5

[5] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 5

[6] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *TPAMI*, 2017. 1

[7] Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Ruslan Salakhutdinov. Good semi-supervised learning that requires a bad gan. In *NeurIPS*, 2017. 3

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[9] Basura Fernando, Tatiana Tommasi, and Tinne Tuytelaars. Joint cross-domain classification and subspace learning for unsupervised adaptation. *Pattern Recognition Letters*, 65:60–66, 2015. 2

[10] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *ICLR*, 2018. 4

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. 1, 2, 3, 4, 5, 6

[12] Timnit Gebru, Jonathan Krause, Jia Deng, and Li Fei-Fei. Scalable annotation of fine-grained categories without experts. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1877–1881. ACM, 2017. 1

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2, 3

[14] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *NeurIPS*, 2005. 2

[15] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 1

[16] Philip Haeusser, Thomas Frerix, Alexander Mordvintsev, and Daniel Cremers. Associative domain adaptation. In *ICCV*, 2017. 8

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 5

[18] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017. 1, 3, 6

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2, 4

[20] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NeurIPS*, 2015. 5

[21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *CVPR Workshop*, 2013. 1

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 1

[23] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *NeurIPS*, 2015. 2

[24] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017. 4

[25] Chi Li, M Zeeshan Zia, Quoc-Huy Tran, Xiang Yu, Gregory D Hager, and Manmohan Chandraker. Deep supervision with shape concepts for occlusion-aware 3d object parsing. In *CVPR*, 2017. 2, 5

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. In *CVPR*, 2017. 1

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *ICCV*, 2017. 1

[28] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. In *ECCV*, 2016. 1

[29] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *ICCV*, 2013. 2

[30] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *NeurIPS*, 2016. 2

[31] Zelun Luo, Yuliang Zou, Judy Hoffman, and Li Fei-Fei. Label efficient learning of transferable representations across domains and tasks. In *NeurIPS*, 2017. 1, 2

[32] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. In *NeurIPS Workshop*, 2014. 2

[33] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D Cubuk, and Ian J Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *arXiv preprint arXiv:1804.09170*, 2018. 4

[34] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *CVPR*, 2017. 2

[35] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 2

[36] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*, 2016. 1

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*, 2015. 1

[38] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 2, 5, 6

[39] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, June 2018. 4

[40] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, 2016. 2, 3

[41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 1

[42] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *TPAMI*, 39(4):640–651, 2017. 1

[43] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, 2017. 1, 2

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[45] Kihyuk Sohn, Sifei Liu, Guangyu Zhong, Xiang Yu, Ming-Hsuan Yang, and Manmohan Chandraker. Unsupervised domain adaptation for face recognition in unlabeled videos. In *ICCV*, 2017. 1, 2

[46] Kihyuk Sohn, Wenling Shang, Xiang Yu, and Manmohan Chandraker. Unsupervised domain adaptation for distance metric learning. In *ICLR*, 2019. 2

[47] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV Workshop*, 2016. 2

[48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 1

[49] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *ICLR*, 2017. 1, 2, 4

[50] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 4

[51] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network. In *ECCV*, 2016. 2, 5

[52] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3D face morphable model. In *CVPR*, June 2019. 1

[53] Luan Tran and Xiaoming Liu. Nonlinear 3D face morphable model. In *CVPR*, June 2018. 1

[54] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, July 2017. 2

[55] Luan Tran, Xi Yin, and Xiaoming Liu. Representation learning by rotating your faces. *TPAMI*, September 2018. 2

[56] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, 2015. 1, 2

[57] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 1, 2

[58] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of machine learning research*, 15(1):3221–3245, 2014. 6

[59] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset, 2011. 1

[60] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *ECCV*, 2016. 2, 7

[61] Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *NeurIPS*, 2015. 2

[62] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, 2015. 1, 2, 5, 6

[63] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017. 2

[64] Erheng Zhong, Wei Fan, Qiang Yang, Olivier Verscheure, and Jiangtao Ren. Cross validation framework to choose amongst models and datasets for transfer learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 547–562. Springer, 2010. 6

[65] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *ECCV*, 2016. 2, 4, 5

[66] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1, 2, 4, 6