UM-Adapt: Unsupervised Multi-Task Adaptation Using Adversarial Cross-Task Distillation

Jogendra Nath Kundu   Nishank Lakkakula   R. Venkatesh Babu

Video Analytics Lab, Indian Institute of Science, Bangalore, India

`jogendrak@iisc.ac.in, nishank974@gmail.com, venky@iisc.ac.in`

Abstract

Aiming towards human-level generalization, there is a need to explore adaptable representation learning methods with greater transferability. Most existing approaches independently address task-transferability and cross-domain adaptation, resulting in limited generalization. In this paper, we propose UM-Adapt - a unified framework to effectively perform unsupervised domain adaptation for spatially-structured prediction tasks, simultaneously maintaining a balanced performance across individual tasks in a multi-task setting. To realize this, we propose two novel regularization strategies; a) Contour-based content regularization (CCR) and b) exploitation of inter-task coherency using a cross-task distillation module. Furthermore, avoiding a conventional ad-hoc domain discriminator, we re-utilize the cross-task distillation loss as output of an energy function to adversarially minimize the input domain discrepancy. Through extensive experiments, we demonstrate superior generalizability of the learned representations simultaneously for multiple tasks under domain-shifts from synthetic to natural environments. UM-Adapt yields state-of-the-art transfer learning results on ImageNet classification and comparable performance on PASCAL VOC 2007 detection task, even with a smaller backbone-net. Moreover, the resulting semi-supervised framework outperforms the current fully-supervised multi-task learning state-of-the-art on both NYUD and Cityscapes dataset.

1. Introduction

Deep networks have proven to be highly successful in a wide range of computer vision problems. They not only excel in classification or recognition based tasks, but also deliver comparable performance improvements for complex spatially-structured prediction tasks [8] like semantic segmentation, monocular depth estimation etc. However, generalizability of such models is one of the major concerns before deploying them in a target environment, since such
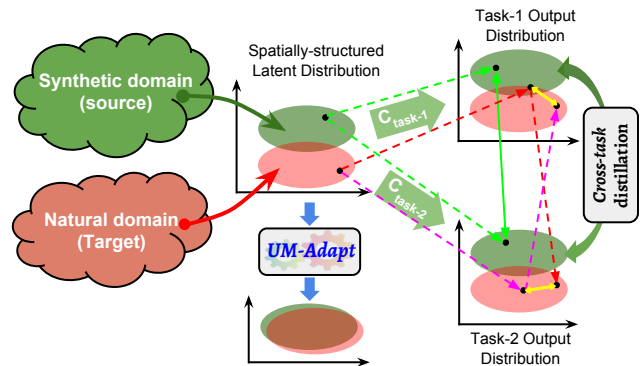

Figure 1. A schematic diagram to understand the implications of *cross-task* distillation. The green arrows show consistency in *cross-task* transfer for a source sample. Whereas, the red and purple arrows show a discrepancy (yellow arrows) in cross-task transfer for a target sample as a result of input domain-shift. *UM-Adapt* aims to minimize this discrepancy as a proxy to achieve adaptation at a spatially-structured common latent representation.

models exhibit alarming dataset or domain bias [4, 28]. To effectively address this, researchers have started focusing on unsupervised domain adaptation approaches [6]. In a fully-unsupervised setting without target annotations, one of the effective approaches [12, 57] is to minimize the domain discrepancy at a latent feature level so that the model extracts domain agnostic and task-specific representations. Although such approaches are very effective in classification or recognition based tasks [56], they yield suboptimal performance for adaptation of fully-convolutional architectures, which is particularly essential for spatial prediction tasks [62]. One of the major issues encountered in such scenarios is attributed to the spatially-structured high-dimensional latent representation in contrast to vectorized form [29]. Moreover, preservation of spatial-regularity, avoiding mode-collapse [49] becomes a significant challenge while aiming to adapt in a fully unsupervised setting.

While we aim towards human level performance, there is a need to explore scalable learning methods, which can yield generic image representation with improved transferability across both tasks and data domains. Muti-task learn-

ing [38, 24] is an emerging field of research in this direction, where the objective is to realize a task-agnostic visual representation by jointly training a model on several complementary tasks [17, 36]. In general, such networks are difficult to train as they require explicit attention to balance performance across each individual task. Also, such approaches only address a single aspect of the final objective (i.e. generalization across tasks) ignoring the other important aspect of generalization across data domains.

In this paper, we focus on multi-task adaptation of spatial prediction tasks, proposing efficient solutions to the specific difficulties discussed above. To effectively deliver optimal performance in both, generalization across tasks and input domain shift, we formulate a multitask adaptation framework called *UM-Adapt*. To effectively preserve the spatial-regularity information during the unlabelled adversarial adaptation [56] procedure, we propose two novel regularization techniques. Motivated by the fact that, the output representations share a common spatial-structure with respect to the input image, we first introduce a novel contour-based content regularization procedure. Additionally, we formalize a novel idea of exploiting cross-task coherency as an important cue to further regularize the multi-task adaptation process.

Consider a base-model trained on two different tasks, task-A and task-B. Can we use the supervision of task-A to learn the output representation of task-B and vice versa? Such an approach is feasible, particularly when the tasks in consideration share some common characteristics (i.e. consistency in spatial structure across task outputs). Following this reasoning, we introduce a *cross-task Distillation* module (see Figure 1). The module essentially constitutes of multiple encoder-decoder architectures called *task-transfer* networks, which are trained to get back the representation of a certain task as output from a combination of rest of the tasks as the input. The overarching motivation behind such a framework is to effectively balance performance across all the tasks in consideration, thereby avoiding domination of the easier tasks during training. To intuitively understand the effectiveness of *cross-task distillation*, let us consider a particular training state of the *base-model*, where the performance on task-A is much better than the performance on task-B. Here the *task-transfer* network (which is trained to output the task-B representation using the base-model task-A prediction as input) will yield improved performance on task-B, as a result of the dominant base task-A performance. This results in a clear discrepancy between the *base-task-B* performance and the task-B performance obtained through the *task-transfer* network. We aim to minimize this discrepancy which in turn acts as a regularization encouraging balanced learning across all the tasks.

In single-task domain adaptation approaches [56, 29], it is common to employ an ad-hoc discriminator to minimize the domain discrepancy. However, in presence of *cross-task distillation* module, this approach highly complicates the overall training pipeline. Therefore, avoiding such a direction, we propose to design a unified framework to effectively address both the diverse objectives i.e. a) to realize a balanced performance across all the tasks and b) to perform domain adaptation in an unsupervised setting. Taking inspirations from energy-based GAN [64], we re-utilize the *task-transfer* networks, treating the *transfer-discrepancies* as output of energy-functions, to adversarially minimize the domain discrepancy in a fully-unsupervised setting.

Our contributions in this paper are as follows:

- We propose a simplified, yet effective unsupervised multi-task adaptation framework, utilizing two novel regularization strategies; a) Contour-based content regularization (CCR) and b) exploitation of inter-task coherency using a *cross-task* distillation module.
- Further, we adopt a novel direction by effectively utilizing *cross-task* distillation loss as an energy-function to adversarially minimize the input domain discrepancy in a fully-unsupervised setting.
- *UM-Adapt* yields *state-of-the-art* transfer learning results on ImageNet classification, and comparable performance on PASCAL VOC 2007 detection task, even with a smaller backbone-net. The resulting semi-supervised framework outperforms the current fully-supervised multi-task learning *state-of-the-art* on both NYUD and Cityscapes dataset.

2. Related work

Domain adaptation. Recent adaptation approaches for deep networks focus on minimization of domain discrepancy by optimizing some distance function related to higher order statistical distributions [6]. Works following adversarial discriminative approaches [55, 56, 10, 11] utilize motivations from Generative Adversarial Networks [14] to bridge the domain-gap. Recent domain adaptation approaches, particularly targeting spatially-structured prediction tasks, can be broadly divided into two sub branches viz. a) pixel-space adaptation and b) feature space adaptation. In pixel-space adaptation [1, 19, 2] the objective is to train a image-translation network [67], which can transform an image from the target domain to resemble like an image from the source domain. On the other hand, feature space adaptation approaches focus on minimization of various statistical distance metrics [15, 36, 53, 21] at some latent feature-level, mostly in a fully-shared source and target domain parameter setting [45]. However, unshared setups show improved adaptation performance as a result of learning dedicated filter parameters for both the domains in consideration [56, 29]. But, a fully unshared setup comes with other difficulties such as *mode-collapse* due to inconsistent output in absence of paired supervision [22, 67]. Therefore,
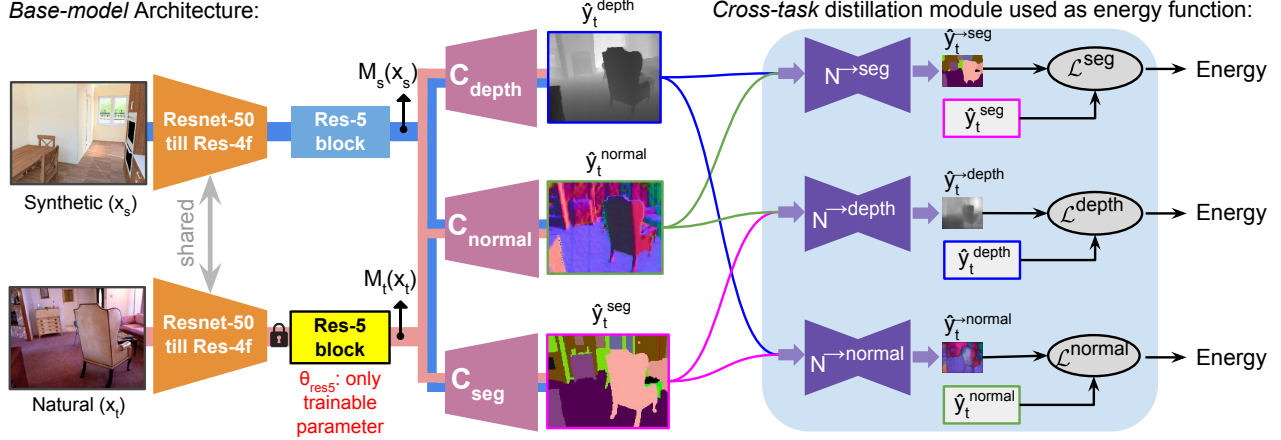
Figure 2. An overview of the proposed *UM-Adapt* architecture for multi-task adaptation. The blue and pink background wide-channel indicates data flow for synthetic and natural domain respectively. On the right we show an illustration of the proposed *cross-task* distillation module, which is later utilized as an energy-function for adversarial adaptation (Section 3.3.2).

an optimal strategy would be to adapt minimally possible parameters separately for the target domain in a partially-shared architecture setting [29].

Multi-task learning. Multi-task learning [3] has been applied in computer vision literature [38] for quite a long time for a varied set of tasks in consideration [27, 8, 9, 16, 43]. To realize this, a trivial direction is to formulate a multi-task loss function, which weighs the relative contribution of each task, enabling equal importance to individual task performance. It is desirable to formulate techniques which adaptively modify the relative weighting of individual tasks as a function of the current learning state or iteration. Kendall *et al.* [24] proposed a principled approach by utilizing a joint likelihood formulation to derive task weights based on the intrinsic uncertainty in individual tasks. Chen *et al.* [65] proposed a gradient normalization (GradNorm) algorithm that automatically balances training in deep multitask models by dynamically tuning gradient magnitudes. Another set of work, focuses on learning task-agnostic generalized visual representations utilizing the advances in multi-task learning techniques [61, 45].

3. Approach

Here we define the notations and problem setting for unsupervised multi-task adaptation. Consider source input image samples $x_s \in X_s$ with the corresponding outputs for different tasks being $y_s^{depth} \in Y_s^{depth}$, $y_s^{seg} \in Y_s^{seg}$ and $y_s^{normal} \in Y_s^{normal}$ for a set of three complementary tasks namely, monocular-depth, semantic-segmentation, and surface-normal respectively. We have full access to the source image and output pair as it is extracted from synthetic graphical environment. The objective of *UM-Adapt* is to estimate the most reliable task-based predictions for an unknown target domain input, $x_t \in X_t$. Considering nat-

ural images as samples from the target domain $P(X_t)$, to emulate an unsupervised setting, we restrict access to the corresponding task-specific outputs viz. $y_t^{depth} \in Y_t^{depth}$, $y_t^{seg} \in Y_t^{seg}$ and $y_t^{normal} \in Y_t^{normal}$. Note that the objective can be readily extended to a semi-supervised setting, considering availability of output annotations for only few input samples from the target domain.

3.1. UM-Adapt architecture

As shown in Figure 2, the base multi-task adaptation architecture is motivated from the standard CNN encoder-decoder framework. The mapping function from the source domain, X_s to a spatially-structured latent representation is denoted as $M_s(X_s)$. Following this, three different decoders with up-convolutional layers [30] are employed for the three tasks in consideration, (see Figure 2) i.e., $\hat{Y}_s^{depth} = C_{depth} \circ M_s(X_s)$, $\hat{Y}_s^{seg} = C_{seg} \circ M_s(X_s)$, and $\hat{Y}_s^{normal} = C_{normal} \circ M_s(X_s)$. Initially the entire architecture is trained with full supervision on source domain data. To effectively balance performance across all the tasks, we introduce the *cross-task Distillation* module as follows.

3.1.1 Cross-task Distillation module

This module aims to get back the representation of a certain task through a transfer function which takes a combined representation of all other tasks as input. Consider T as the set of all tasks, i.e. $T = \{t_1, t_2, ..., t_k\}$, where k is the total number of tasks in consideration. For a particular task t_i, we denote $\hat{Y}_s^{t_i}$ as the prediction of the *base-model* at the t_i output-head. We denote the task specific loss function as $\mathcal{L}^{t_i}(.,.)$ in further sections of the paper. Here, *task-transfer* network is represented as $N^{\rightarrow t_i}$, which takes a combined set $O_s^{t_i} = \{\hat{Y}_s^{t_j} \mid t_j \in T - \{t_i\}\}$ as the input representation and the corresponding output is de-

a) Contour-based Content Regularization (CCR)

b) Contour-map obtained from

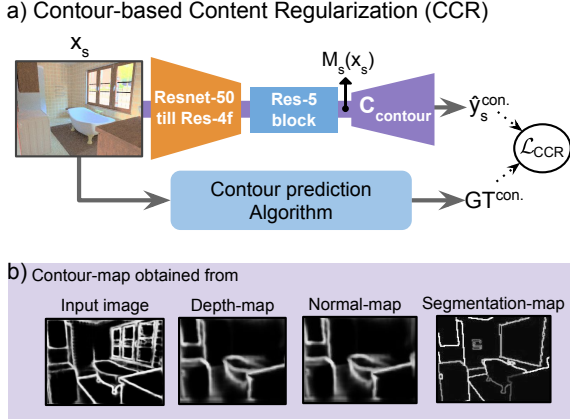Input image | Depth-map | Normal-map | Segmentation-map

Figure 3. An overview of the (a) proposed CCR framework with (b) evidence of consistency in contour-map computed on the output-map of each task to that of the input RGB image.

noted as $Y_s^{\to t_i} = N^{\to t_i}(O_s^{t_i})$. The parameters of $N^{\to t_i}$, $\theta_{N^{\to t_i}}^s$ are obtained by optimizing a task-transfer loss function, denoted by $\mathcal{L}^{t_i}(Y_s^{\to t_i}, Y_s^{t_i})$ and are kept frozen in further stages of training. However, one can feed the predictions of *base-model* through the *task-transfer* network to realize another estimate of task t_i represented as $\hat{Y}_s^{\to t_i} = N^{\to t_i}(\hat{O}_s^{t_i})$, where $\hat{O}_s^{t_i} = \{\hat{Y}_s^{t_j} \mid t_j \in T - \{t_i\}\}$. Following this, we define the *distillation-loss* for task t_i as $\mathcal{L}^{t_i}(\hat{Y}_s^{\to t_i}, Y_s^{t_i})$. While optimizing parameters of the *base-model*, this *distillation-loss* is utilized as one of the important loss components to realize an effective balance across all the task objectives (see Algorithm 1). Here, we aim to minimize the discrepancy between the direct and indirect prediction (via other tasks) of individual tasks. The proposed learning algorithm does not allow any single task to dominate the training process, since the least performing task will exhibit higher discrepancy and hence will be given more importance in further training iterations.

Compared to the general knowledge-distillation framework [18], one can consider $Y_s^{\to t_i} = N^{\to t_i}(O_s^{t_i})$ to be analogous to the output of a teacher network and $\hat{Y}_s^{\to t_i} = N^{\to t_i}(\hat{O}_s^{t_i})$ as the output of a student network. Here the objective is to optimize the parameters of the *base-model* by effectively employing the distillation loss which in turn enforces coherence among the individual task performances.

3.1.2 Architecture for target domain adaptation

Following a partially-shared adaptation setup, a separate latent mapping network is introduced specifically for the target domain samples, i.e. $M_t(X_t)$ (see Figure 2). Inline with AdaDepth [29], we initialize M_t using the pre-trained parameters of the source domain counterpart, M_s (*Resnet-50* encoder), in order to start with a good baseline initialization. Following this, only *Res-5* block (i.e. θ_{res5}) parameters are updated for the additional encoder branch, M_t. Note that

/*Initialization of parameters */
θ_{base}: *base-model* parameters $\{\theta_{M_s}, \theta_{T_{t_1}}, ..., \theta_{T_{t_k}}\}$
for m *iterations* **do**
$\quad$ **for** *task* t_i; $i = 1, 2, ..., k$ **do**
$\quad\quad \mathcal{L}_{dist.}^{t_i} = \mathcal{L}^{t_i}(\hat{Y}_s^{t_i}, Y_s^{t_i}) + \alpha \mathcal{L}^{t_i}(\hat{Y}_s^{\to t_i}, Y_s^{t_i})$
$\quad\quad \theta_{base}^* := \underset{\theta_{base}}{\text{argmin}} \ \mathcal{L}_{dist.}^{t_i}$

Algorithm 1: *Base-model* training algorithm on fully-supervised source data with cross-task distillation loss.

the learning algorithm for unsupervised adaptation does not update other layers in the task-specific decoders and the initial shared M_t layers till the *Res-4f* block.

3.2. Contour-based Content Regularization (CCR)

Spatial content inconsistency is a serious problem for unsupervised domain adaptation focused on pixel-wise dense-prediction tasks [19, 29]. In order to address this, [29] proposes a Feature Consistency Framework (*FCF*), where the authors employ a cyclic feature reconstruction setup to preserve the spatially-structured content such as semantic contours, at the *Res-4f* activation map of the frozen *Resnet-50* (till *Res-4f*) encoder. However, the spatial size of the output activation of *Res-4f* feature (i.e. 20×16) is inefficient to capture relevant spatial regularities required to alleviate the contour alignment problem.

To address the above issue, we propose a novel Contour-based Content Regularization (*CCR*) method. As shown in Figure 3a, we introduce a shallow (4 layer) contour decoder, $C_{contour}$ to reconstruct only the contour-map of the given input image, where the ground-truth is obtained using a standard contour prediction algorithm [60]. This content regularization loss (mean-squared loss) is denoted as $\mathcal{L}_{CCR}$ in further sections of the paper. Assuming that majority of the image based contours align with the contours of task-specific output maps, we argue that the encoded feature (*Res-5c* activation) must retain the contour information during the adversarial training for improved adaptation performance. This clearly makes *CCR* superior over the existing approaches of enforcing image reconstruction based regularization [1, 40, 50], by simplifying the additionally introduced decoder architecture devoid of the burden of generating irrelevant color-based appearance. $C_{contour}$ is trained on the fixed output transformation $M_s(X_s)$ and the corresponding ground-truth contour pair, i.e. $\mathcal{L}_{CCR}(\hat{Y}_s^{con.}, GT^{con.})$. However, unlike *FCF* regularization [29], the parameter of $C_{contour}$ is not updated during the adversarial learning, as the expected output contour map is independent of the M_s or M_t transformation. As a result, $\mathcal{L}_{CCR}$ is treated as output of an energy-function, which is later minimized for $M_t(X_t)$ to bridge the discrepancy between the distributions $P(M_s(X_s))$ and $P(M_t(X_t))$ during adaptation, as shown in Algorithm 2.

/*Initialization of parameters */
θ_{res5}: *Res5* parameters of M_t initialized from M_s
$\theta_{N \to t_i}$: parameters of fully trained $N^{\to t_i}$ (i.e. $\theta_{N \to t_i}^s$)
       on ground-truth task output-maps, $Y_s^{t_i}$

for n iterations **do**
     for m steps **do**
         for task t_i; $i = 1, 2, ..., k$ **do**
             /* Update trainable parameters of M_t by
              minimizing energy of target samples.*/
             $\mathcal{L}_G^{t_i} = \mathcal{L}^{t_i}(\hat{Y}_t^{\to t_i}, \hat{Y}_t^{t_i})$
             $\theta_{res5}^* :=$
             $\underset{\theta_{res5}}{\text{argmin}}\ \mathcal{L}_G^{t_i} + \lambda \mathcal{L}_{CCR}(\hat{Y}_t^{con.}, GT^{con.})$
     for task t_i; $i = 1, 2, ..., k$ **do**
         /* Update the energy function $N^{\to t_i}$ */
         $\mathcal{L}_D^{t_i} = \mathcal{L}^{t_i}(\hat{Y}_s^{\to t_i}, Y_s^{t_i}) - \mathcal{L}^{t_i}(\hat{Y}_t^{\to t_i}, \hat{Y}_t^{t_i})$
         $\theta_{N \to t_i}^* := \underset{\theta_{N \to t_i}}{\text{argmin}}\ \mathcal{L}_D^{t_i}$

Algorithm 2: Training algorithm of *UM-Adapt-(Adv.)* utilizing energy-based adversarial *cross-task* distillation. In *UM-Adapt-(noAdv.)* we do not update parameters of the task-transfer network, i.e. $\theta_{N \to t_i}^* = \theta_{N \to t_i}$ throughout the adaptation procedure (see Section 3.3.2).

3.3. Unsupervised Multi-task adaptation

In unsupervised adaptation, the overall objective is to minimize the discrepancy between the source and target input distributions. However, minimizing the discrepancy between $P(Y_s^{t_i})$ and $P(\hat{Y}_t^{t_i})$ can possibly overcome the differences between the ground-truth $Y_s^{t_i}$ and prediction $\hat{Y}_s^{t_i}$ better, when compared to matching $P(\hat{Y}_s^{t_i})$ with $P(\hat{Y}_t^{t_i})$ as proposed in some previous approaches [54]. Aiming towards optimal performance, *UM-Adapt* focuses on matching target prediction with the actual ground-truth map distribution, and the proposed *cross-task distillation* module provides a means to effectively realize such an objective.

3.3.1 UM-Adapt baseline (*UM-Adapt-B*)

Existing literature [37, 29] shows efficacy of simultaneous adaptation at hierarchical feature levels, while minimizing domain discrepancy for multi-layer deep architectures. Motivated by this, we design a single discriminator which can match the joint distribution of latent representation and the final task-specific structured prediction maps with the corresponding true joint distribution. As shown in Figure 2, the predicted joint distribution denoted by $P(M_t(X_t), \hat{Y}_t^{depth}, \hat{Y}_t^{normal}, \hat{Y}_t^{seg})$, is matched with true distribution denoted by $P(M_s(X_s), Y_s^{depth}, Y_s^{normal}, Y_s^{seg})$, following the usual adversarial discriminative strategy [29] (see Supplementary for more details). We will denote this framework as *UM-Adapt-B* in further sections of this paper.

3.3.2 Adversarial cross-task distillation

Aiming towards formalizing a unified framework to effectively address multi-task adaptation as a whole, we plan to treat the *task-transfer* networks, $N^{\to t_i}$ as energy functions to adversarially minimize the domain discrepancy. Following the analogy of Energy-based GAN [64], the *task-transfer* networks are first trained to obtain low-energy for the ground-truth task-based source tuples (i.e. $(O_s^{t_i}, Y_s^{t_i})$) and high-energy for the similar tuples from the target predictions (i.e. $(\hat{O}_t^{t_i}, \hat{Y}_t^{t_i})$). This is realized by minimizing $\mathcal{L}_D^{t_i}$ as defined in Algorithm 2. Conversely, the trainable parameters of M_t are updated to assign low energy to the predicted target prediction tuples, as enforced by $\mathcal{L}_G^{t_i}$ (see Algorithm 2). Along with the previously introduced CCR regularization, the final update equation for θ_{res5} is represented as $\mathcal{L}_G^{t_i} + \lambda \mathcal{L}_{CCR}$. We use different optimizers for energy functions of each task, t_i. As a result, θ_{res5} is optimized to have a balanced performance across all the tasks even in a fully unsupervised setting. We denote this framework as *UM-Adapt-(Adv.)* in further sections of this paper.

Note that the *task-transfer* networks are trained only on ground-truth output-maps under sufficient regularization due to the compressed latent representation, as a result of the encoder-decoder setup. This enables $N^{\to t_i}$ to learn a better approximation of the intended *cross-task* energy manifold, even in absence of negative examples (target samples) [64]. This analogy is used in Algorithm 1 to effectively treat the frozen *task-transfer* network as an energy-function to realize a balanced performance across all the tasks on the fully-supervised source domain samples. Following this, we plan to formulate an ablation of *UM-Adapt*, where we restrain the parameter update of $N^{\to t_i}$ in Algorithm 2. We denote this framework as *UM-Adapt-(noAdv.)* in further sections of this paper. This modification gracefully simplifies the unsupervised adaptation algorithm, as it finally retains only θ_{res5} as the minimal set of trainable parameters (with frozen parameter of $N^{\to t_i}$ as $\theta_{N \to t_i}^s$).

4. Experiments

To demonstrate effectiveness of the proposed framework, we evaluate on three different publicly available benchmark datasets, separately for indoor and outdoor scenes. Further, in this section, we discuss details of our adaptation setting and analysis of results on standard evaluation metrics for a fair comparison against prior art.

4.1. Experimental Setting

We follow the encoder-decoder architecture exactly as proposed by Liana *et al.* [30]. The decoder architecture is replicated three times to form C_{depth}, C_{normal} and C_{seg} respectively. However, the number of feature maps and non-linearity for the final task-based prediction layers is adopted

Table 1. Quantitative comparison of different ablations of *UM-Adapt* framework with comparison against prior arts for depth estimation on NYUD-v2. The second column indicates amount of supervised target samples used during training.

Method	sup.	Error↓			Accuracy↑ ($\gamma = 1.25$)		
		rel	log10	rms	$\delta < \gamma$	$\delta < \gamma^2$	$\delta < \gamma^3$
Saxena *et al.* [51]	795	0.349	-	1.214	0.447	0.745	0.897
Liu *et al.* [33]	795	0.230	0.095	0.824	0.614	0.883	0.975
Eigen *et al.* [9]	120K	0.215	-	0.907	0.611	0.887	0.971
Roy *et al.* [47]	795	0.187	0.078	0.744	-	-	-
Laina *et al.* [30]	96K	0.129	0.056	0.583	0.801	0.950	0.986
Simultaneous multi-task learning							
Multi-task baseline	0	0.27	0.095	0.862	0.559	0.852	0.942
UM-Adapt-B(FCF)	0	0.218	0.091	0.679	0.67	0.898	0.974
UM-Adapt-B(CCR)	0	0.192	0.081	0.754	0.601	0.877	0.971
UM-Adapt-(noAdv.)-1	0	0.181	0.077	0.743	0.623	0.889	0.978
UM-Adapt-(noAdv.)	0	0.178	**0.063**	0.712	0.781	0.917	0.984
UM-Adapt-(Adv.)	0	**0.175**	0.065	**0.673**	**0.783**	**0.92**	**0.984**
Wang *et al.* [58]	795	0.220	0.094	0.745	0.605	0.890	0.970
Eigen *et al.* [8]	795	0.158	-	0.641	0.769	0.950	0.988
Jafari *et al.* [23]	795	0.157	0.068	0.673	0.762	0.948	0.988
UM-Adapt-S	795	**0.149**	**0.067**	**0.637**	**0.793**	**0.938**	**0.983**

Table 2. Quantitative comparison of different ablations of *UM-Adapt* framework with comparison against prior arts for Surface-normal estimation on the standard test-set of NYUD-v2.

Method	sup.	Error ↓		Accuracy↑		
		mean	median	11.25°	22.5°	30°
Eigen *et al.* [8]	120k	22.2	15.3	38.6	64	73.9
PBRS [63]	795	21.74	14.75	39.37	66.25	76.06
SURGE [59]	795	20.7	12.2	47.3	68.9	76.6
GeoNet [42]	30k	**19.0**	**11.8**	**48.4**	71.5	79.5
Simultaneous multi-task learning						
Multi Task Baseline	0	25.8	18.73	29.65	61.69	69.83
UM-Adapt-B(FCF)	0	24.6	16.49	37.53	65.73	75.51
UM-Adapt-B(CCR)	0	23.8	14.67	42.08	69.13	77.28
UM-Adapt-(noAdv.)-1	0	22.3	15.56	43.17	69.11	78.36
UM-Adapt-(noAdv.)	0	22.2	15.31	**43.74**	70.18	78.83
UM-Adapt-(Adv.)	0	**22.2**	**15.23**	43.68	**70.45**	**78.95**
UM-Adapt-S	795	21.2	13.98	44.66	**72.11**	**81.08**

Table 3. Quantitative comparison of different ablations of *UM-Adapt* framework with comparison against prior arts for sematic segmentation on the standard test-set of NYUD-v2.

Method	Sup.	Mean IOU	Mean Accuracy	Pixel Accuracy
PBRS [63]	795	0.332	-	-
Long *et al.* [35]	795	0.292	0.422	0.600
Lin *et al.* [32]	795	0.406	0.536	0.700
Kong *et al.* [26]	795	**0.445**	-	0.721
RefineNet(Res50) [31]	795	0.438	-	-
Simultaneous multi-task learning				
Multi Task Baseline	0	0.022	0.063	0.067
UM-Adapt-B(FCF)	0	0.154	0.295	0.514
UM-Adapt-B(CCR)	0	0.163	0.308	0.557
UM-Adapt-(noAdv.)-1	0	0.189	0.345	0.603
UM-Adapt-(noAdv.)	0	0.214	0.364	0.608
UM-Adapt-(Adv.)	0	**0.221**	*0.366*	**0.619**
Eigen *et al.* [8]	795	0.341	0.451	0.656
Arsalan *et al.* [39]	795	0.392	0.523	0.686
UM-Adapt-S	795	**0.444**	**0.536**	**0.739**

according to the standard requirements. We use BerHu loss [30] as the loss function for the depth estimation task, i.e. $\mathcal{L}^{depth}(.,.)$. Following Eigen *et al.* [8], an inverse of element-wise dot product on unit normal vectors for each pixel location is consider as the loss function for surface-normal estimation, $\mathcal{L}^{normal}(.,.)$. Similarly for segmentation, i.e $\mathcal{L}^{seg}(.,.)$, classification based cross-entropy loss is implemented with a weighing scheme to balance gradients from different classes depending on their coverage.

We also consider a semi-supervised setting (*UM-Adapt-S*), where the training starts from the initialization of the trained unsupervised version, *UM-Adapt-(Adv.)*. For better generalization, alternate batches of labelled (optimize supervised loss, $\mathcal{L}^{t_i}_{dist.}$) and unlabelled (optimize unsupervised loss, $\mathcal{L}^{t_i}_G + \lambda \mathcal{L}_{CCR}$) target samples are used to update the network parameters (i.e. θ_{res5}).

Datasets. For representation learning on indoor scenes, we use the publicly available NYUD-v2 [52] dataset, which has been used extensively for supervised multi-task predic-

tion of depth-estimation, sematic segmentation and surface-normal estimation. The processed version of the dataset consists of 1449 sample images with a standard split of 795 for training and 654 for testing. While adapting in semi-supervised setting, we use the corresponding ground-truth maps of all the 3 tasks (795 training images) for the supervised loss. The CNN takes an input of size $228{\times}304$ with various augmentations of scale and flip following [9], and outputs three task specific maps, each of size $128{\times}160$. For the synthetic counterpart, we use 100,000 randomly sampled synthetic renders from PBRS [63] dataset along with the corresponding clean ground-truth maps (for all the three tasks) as the source domain samples.

To demonstrate generalizability of *UM-Adapt*, we consider outdoor-scene dataset for two different tasks, sematic segmentation and depth estimation. For the synthetic source domain, we use the publicly available GTA5 [46] dataset consisting of 24966 images with the corresponding depth and segmentation ground-truths. However, for real outdoor scenes, the widely used KITTI dataset does not have semantic labels that are compatible with the synthetic counterpart. On the other hand, the natural image Cityscapes dataset [5] does not contain ground-truth depth maps. Therefore to formulate a simultaneous multi-task learning problem and to perform a fair comparison against prior art, we consider the Eigen test-split on KITTI [8] for comparison of depth-estimation result and the Cityscapes validation set to benchmark our outdoor segmentation results in a single *UM-Adapt* framework. For the semi-supervised setting, we feed alternate KITTI and Cityscapes minibatches with the corresponding ground-truth maps for supervision. Here, input and output resolution for the network is considered to be $256{\times}512$ and $128{\times}256$ respectively.

Training details. We first train a set of task-transfer networks on the synthetic task label-maps separately for both indoor (PBRS) and outdoor (GTA5) scenes. For in-

Table 4. Quantitative comparison of ablations of *UM-Adapt* framework with comparison against prior arts for depth-estimation on the Eigen test-split [9] of KITTI dataset.

Method	Target image supervision	Error↓			
		rel	sq.rel	rms	rms(log10)
Eigen *et al.* [9]	Full	0.203	1.548	6.307	0.282
Godard *et al.* [13]	Binocular	0.148	1.344	5.927	0.247
zhou *et al.* [66]	Video	0.208	1.768	6.856	0.283
AdaDepth [29]	No	0.214	1.932	7.157	0.295
Simultaneous multi-task learning					
Multi-task baseline	No	0.381	2.08	8.482	0.41
UM-Adapt-(noAdv.)	No	0.28	1.99	**7.791**	0.346
UM-Adapt-(Adv.)	No	**0.27**	**1.98**	7.823	**0.336**
UM-Adapt-S	few-shot	**0.201**	**1.72**	**5.876**	**0.259**

Table 5. Quantitative comparison of ablations of *UM-Adapt* framework with comparison against prior arts for semantic segmentation on the validation set of Cityscapes dataset.

Method	Image supervision	Mean IOU
FCN-Wild [20]	0	0.271
CDA [62]	0	0.289
DC [55]	0	0.376
Cycada [19]	0	0.348
AdaptSegNet [54]	0	0.424
Simultaneous multi-task learning		
Multi Task Baseline	0	0.224
UM-Adapt-(noAdv.)	0	0.408
UM-Adapt-(Adv.)	0	**0.420**
UM-Adapt-S	500	**0.544**

door dataset, we train only the following two *task-transfer* networks; $N^{\rightarrow seg}(Y_s^{depth}, Y_s^{normal})$ and $N^{\rightarrow depth}(Y_s^{seg})$ considering the fact that surface-normal and depth estimation are more correlated, when compared to other pair of tasks. Similarly for outdoor, we choose the only two task-transfer possible combinations $N^{\rightarrow seg}(Y_s^{depth})$ and $N^{\rightarrow depth}(Y_s^{seg})$. Following this, two separate *base-model*s are trained with full-supervision on synthetic source domain using Algorithm 1 ($\alpha = 10$) with different optimizers (Adam [25]) for each individual task. After obtaining a frozen fully-trained source-domain network, the $C_{contour}$ network is trained as discussed in Section 3.2 and it remains frozen during its further usage as a regularizer.

4.2. Evaluation of the *UM-Adapt* Framework

We have conducted a thorough ablation study to establish effectiveness of different components of the proposed *UM-Adapt* framework. We report results on the standard benchmark metrics as followed in literature for each of the individual tasks, to have a fair comparison against state-of-the-art approaches. Considering the difficulties of simultaneous multi-task learning, we have clearly segregated prior art based on single-task or multi-task optimization approaches in all the tables in this section.

Ablation study of *UM-Adapt*. As a multi-task baseline, we report performance on the standard test-set of natural samples with direct inference on the frozen source-domain parameters without adaptation. With the exception of results on semantic-segmentation, baseline performance for the other two regression tasks (i.e. depth estimation and surface-normal prediction) are strong enough to support the idea of achieving first-level generalization using multi-task learning. However, the prime focus of *UM-Adapt* is to achieve the second-level of generalization through unsupervised domain adaptation. In this regard, to analyze effectiveness of the proposed *CCR* regularization (Section 3.2) against FCF [29], we conduct an experiment on the *UM-Adapt-B* framework defined in Section 3.3.1. The reported benchmark numbers for unsupervised adap-

tation from PBRS to NYUD (see Table 1, 2 and 3) clearly indicates the superiority of *CCR* for adaptation of structured prediction tasks. Following this inference, all later ablations (i.e. *UM-Adapt-(noAdv.)*, *UM-Adapt-(Adv.)* and *UM-Adapt-S*) use only *CCR* as content-regularizer.

Utilizing gradients from the frozen *task-transfer* network yields a clear improvement over *UM-Adapt-B* as shown in Tables 1, 2 and 3 for all the three tasks in NYUD dataset. This highlights significance of the idea to effectively exploit the inter-task correlation information for adaptation of a multi-task learning framework. To quantify the importance of multiple *task-transfer* network against employing a single such network, we designed another ablation setting denoted as *UM-Adapt-(noAdv.)-1*, which utilizes only $N^{\rightarrow seg}(Y_s^{depth}, Y_s^{normal})$ for adaptation to NYUD, as reported in Tables 1, 2 and 3. Next, we report a comparison between the proposed energy-based cross-task distillation frameworks (Section 3.3.2) i.e. a) *UM-Adapt-(Adv.)* and b) *UM-Adapt-(noAdv.)*. Though, *UM-Adapt-(Adv.)* shows minimal improvement over the other counterpart, training of *UM-Adapt-(noAdv.)* is found to be significantly stable and faster as it does not include parameter update of the *task-transfer* networks during the adaptation process.

Comparison against prior structured-prediction works. The final unsupervised multi-task adaptation result by the best variant of *UM-Adapt* framework, i.e. *UM-Adapt-(Adv.)* delivers comparable performance against previous fully-supervised approaches (see Table 1 and 3). One must consider the explicit challenges faced by *UM-Adapt* to simultaneously balance performance across multiple tasks in a unified architecture as compared to prior arts focusing on single-task at a time. This clearly demonstrates superiority of the proposed approach towards the final objective of realizing generalization across both tasks and data domains. The semi-supervised variant, *UM-Adapt-S* is able to achieve state-of-the-art multi-task learning performance when compared against other fully supervised approaches as clearly highlighted in Tables 1, 2 and 3.

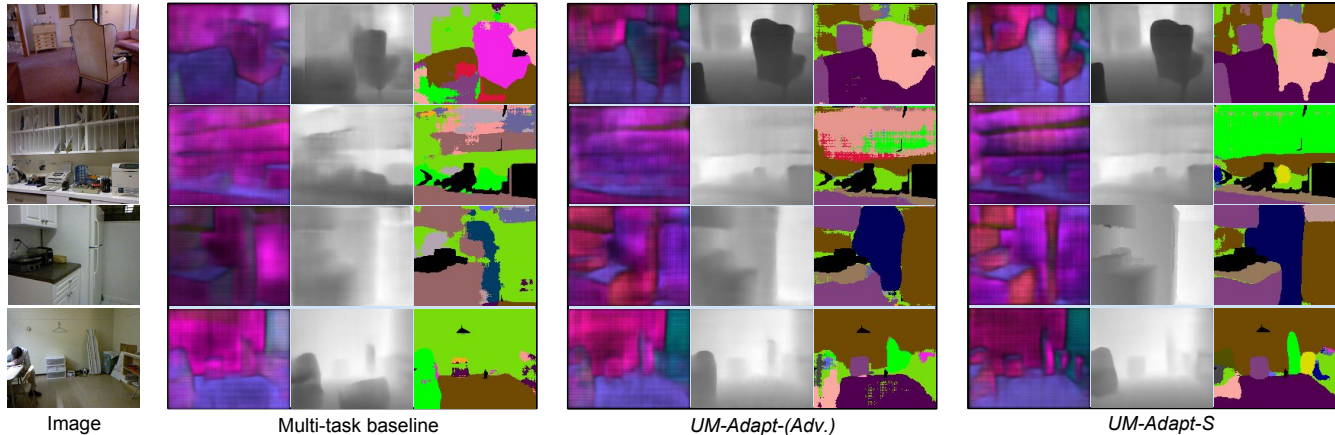Note that the adaptation of depth estimation from KITTI and semantic-segmentation from Cityscapes in a single *UM-*

Figure 4. Qualitative comparison of different ablations of *UM-Adapt*, i.e. a) Multi-task baseline, b) *UM-Adapt-(Adv.)*, and c) *UM-Adapt-S*.

Adapt framework is a much harder task due to the input domain discrepancy (cross-city [4]) along with the challenges in simultaneous multi-task optimization setting. Even in such a drastic scenario *UM-Adapt* is able to achieve reasonable performance in both depth estimation and sematic segmentation as compared to other unsupervised single task adaptation approaches reported in Tables 4 and 5.

Comparison against prior multi-task learning works. Table 6 and Table 7 present a comparison of *UM-Adapt* with recent multi-task learning approaches [24, 65] on NYUD test-set and CityScapes validation-set respectively. It clearly highlights *state-of-the-art* performance achieved by *UM-Adapt-S* as a result of the proposed *cross-task* distillation framework.

Table 6. Test error on NYUDv2 with ResNet as the base-model

Method	Sup.	Depth rms Err. (m)	Seg. Err. (100-IoU)	Normals Err. (1-|cos|)
Kendall *et al*. [24]	30k	0.702	-	0.182
GradNorm [65]	30k	0.663	67.5	0.155
UM-Adapt-S	795	**0.637**	**55.6**	**0.139**

Table 7. Validation mIOU on Cityscapes, where all the approaches are trained simultaneously for segmentation and depth estimation.

Method	Semantic (mean IOU)
Kendall *et al*. [24] (Uncert. Weights)	51.52
Liu *et al*. [34]	52.68
UM-Adapt-S	**54.4**

Transferability of the learned representation. One of the overarching goals of *UM-Adapt* is to learn general-purpose visual representation, which can demonstrate improved transferability across both tasks and data-domains. To evaluate this, we perform experiments on large-scale representation learning benchmarks. Following evaluation protocol by Doersch *et al*. [7], we setup *UM-Adapt* for transfer-learning on ImageNet [48] classification and PASCAL VOC 2007 Detection tasks. The base trunk till *Res5*

Table 8. Transfer learning results on novel unseen tasks.

Method	Backbone	Classification ImageNet top5	Detection PASCAL 2007
Motion Seg. [41]	ResNet-101	48.29	61.13
Exemplar [7]	ResNet-101	53.08	60.94
RP+Col+Ex+MS [7]	ResNet-101	69.30	**70.53**
UM-Adapt-S	**ResNet-50**	**69.51**	70.02

block is initialized from our *UM-Adapt-(Adv.)* variant (the adaptation of PBRS to NYUD) for both classification and detection task. For ImageNet classification, we train randomly initialized fully connected layers after the output of *Res5* block. Similarly, for detection we use *Faster-RCNN* [44] with 3 different output heads for object proposal, classification, and localization after the *Res4* block. We finetune all the network weights separately for classification and detection [7]. The results in Table 8 clearly highlight superior transfer learning performance of our learned representation even for novel unseen tasks with a smaller backbone-net.

5. Conclusion

The proposed *UM-Adapt* framework addresses two important aspects of generalized feature learning by formulating the problem as a multi-task adaptation approach. While the multi-task training ensures learning of task-agnostic representation, the unsupervised domain adaptation method provides domain agnostic representation projected to a common spatially-structured latent representation. The idea of exploiting cross-task coherence as an important cue for preservation of spatial regularity can be utilized in many other scenarios involving fully-convolutional architectures. Exploitation of auxiliary task setting is another direction that remains to be explored in this context.

References

[1] Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *CVPR*, 2018. 2, 4

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

[3] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 3

[4] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *ICCV*, 2017. 1, 8

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 6

[6] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017. 1, 2

[7] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *ICCV*, 2017. 8

[8] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015. 1, 3, 6

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, pages 2366–2374, 2014. 3, 6, 7

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 2

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 2

[12] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, 2015. 1

[13] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 7

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

[15] A. Gretton, AJ. Smola, J. Huang, M. Schmittfull, KM. Borgwardt, and B. Schölkopf. *Covariate shift and local learning by distribution matching*. MIT Press. 2

[16] Hu Han, Anil K Jain, Shiguang Shan, and Xilin Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE transactions on pattern analysis and machine intelligence*, 2017. 3

[17] Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. A joint many-task model: Growing a neural network for multiple nlp tasks. *arXiv preprint arXiv:1611.01587*, 2016. 2

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. 2014. 4

[19] 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. In *ICML*, 2018. 2, 4, 7

[20] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 7

[21] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *CVPR*, 2018. 2

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

[23] Omid Hosseini Jafari, Oliver Groth, Alexander Kirillov, Michael Ying Yang, and Carsten Rother. Analyzing modular cnn architectures for joint depth prediction and semantic segmentation. In *ICRA*, 2017. 6

[24] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 2, 3, 8

[25] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[26] Shu Kong and Charless Fowlkes. Recurrent scene parsing with perspective understanding in the loop. *arXiv preprint arXiv:1705.07238*, 2017. 6

[27] Srinivas SS Kruthiventi, Vennela Gudisa, Jaley H Dholakiya, and R Venkatesh Babu. Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. In *CVPR*, 2016. 3

[28] Jogendra Nath Kundu, Maharshi Gor, Phani Krishna Uppala, and Venkatesh Babu Radhakrishnan. Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In *WACV*, 2019. 1

[29] Jogendra Nath Kundu, Phani Krishna Uppala, Anuj Pahuja, and R Venkatesh Babu. Adadepth: Unsupervised content congruent adaptation for depth estimation. In *CVPR*, 2018. 1, 2, 3, 4, 5, 7

[30] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016. 3, 5, 6

[31] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 6

[32] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016. 6

[33] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, 2015. 6

[34] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. *arXiv preprint arXiv:1803.10704*, 2018. 8

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 6

[36] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 2

[37] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *NIPS*, 2016. 5

[38] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *CVPR*, 2016. 2, 3

[39] Arsalan Mousavian, Hamed Pirsiavash, and Jana Košecká. Joint semantic segmentation and depth estimation with deep convolutional networks. In *3DV*. IEEE, 2016. 6

[40] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *CVPR*, 2018. 4

[41] Deepak Pathak, Ross Girshick, Piotr Dollar, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *CVPR*, 2017. 8

[42] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *CVPR*, 2018. 6

[43] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 8

[45] Zhongzheng Ren and Yong Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *CVPR*, 2018. 2, 3

[46] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 6

[47] Anirban Roy and Sinisa Todorovic. Monocular depth estimation using neural regression forest. In *CVPR*, 2016. 6

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 8

[49] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016. 1

[50] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *CVPR*, 2018. 4

[51] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840, 2009. 6

[52] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*. Springer, 2012. 6

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

[54] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 5, 7

[55] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, 2015. 2, 7

[56] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 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

[58] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Towards unified depth and semantic prediction from a single image. In *CVPR*, 2015. 6

[59] Xiaolong Wang, David Fouhey, and Abhinav Gupta. Designing deep networks for surface normal estimation. In *CVPR*, 2015. 6

[60] Saining "Xie and Zhuowen" Tu. Holistically-nested edge detection. In *ICCV*, 2015. 4

[61] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012. 3

[62] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *ICCV*, 2017. 1, 7

[63] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *CVPR*, 2017. 6

[64] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. In *ICLR*, 2017. 2, 5

[65] Chen-Yu Lee Zhao Chen, Vijay Badrinarayanan and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018. 3, 8

[66] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 7

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