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 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,
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].

![Base-model Architecture:](Image)

**Cross-task distillation module used as energy function:**

![Cross-task distillation module](Image)

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 $y_{s, depth}^i \in Y_{s, depth}^i$, $y_{s, seg}^i \in Y_{s, seg}^i$ and $y_{s, normal}^i \in Y_{s, normal}^i$ 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 natural 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. $y_{t, depth}^i \in Y_{t, depth}^i$, $y_{t, seg}^i \in Y_{t, seg}^i$ and $y_{t, normal}^i \in Y_{t, normal}^i$. 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., $Y_{depth}^i = C_{depth} \circ M_s(X_s)$, $Y_{seg}^i = C_{seg} \circ M_s(X_s)$, and $Y_{normal}^i = C_{normal} \circ M_s(X_s)$. 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 $Y_{t_i}$ as the prediction of the base-model at the $t_i$ output-head. We denote the task specific loss function as $L_{\cdot, t_i}(\cdot)$ in further sections of the paper. Here, task-transfer network is represented as $N^{-t_i}$, which takes a combined set $O_{s, t_i} = \{Y_{s, t_j} \mid t_j \in T - \{t_i\} \}$ as the input representation and the corresponding output is de-
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. \( \theta_{res5} \)) parameters are updated for the additional encoder branch, \( M_t \). Note that

\[
\theta_{base: base-model parameters} = \{ \theta_M, \theta_{T_1}, \ldots, \theta_{T_k} \}
\]

for \( m \) iterations do

\[
\text{for task } t_i; i = 1, 2, \ldots, k \text{ do}
\]

\[
L_{dist}^{t_i} = \mathcal{L}^{t_i}(\hat{Y}^{t_i}, Y^{t_i}) + \alpha \mathcal{L}^{t_i}(\tilde{Y}^{t_i}, Y^{t_i})
\]

\[
\theta_{base} := \arg\min \mathcal{L}_{dist}^{t_i}
\]

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. \( 20 \times 16 \)) 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. \( \mathcal{L}_{CCR}(Y_s^{con.}, GT^{con.}) \). 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 \( M_t \) or \( M_s \) 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.
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, $N \rightarrow t_i$ 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-source based samples (i.e. $(O_i^t, Y_i^s)$) and high-energy for the similar tuples from the target predictions (i.e. $(O_i^t, Y_i^t)$). This is realized by minimizing $L_Q^t$ 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 $L_Q^t$ (see Algorithm 2). Along with the previously introduced CCR regularization, the final update equation for $\theta_{res5}$ is represented as $L_Q^t + \lambda L_{CCR}$. We use different optimizers for energy functions of each task, $t_i$. As a result, $\theta_{res5}$ 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 $N \rightarrow t_i$ 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 $N \rightarrow t_i$ 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 $\theta_{res5}$ as the minimal set of trainable parameters (with frozen parameter of $N \rightarrow t_i$, as $\theta_{N \rightarrow t_i}$).

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 $P(Y_i^s)$ and $P(Y_i^t)$ can possibly overcome the differences between the ground-truth $Y_i^s$, and prediction $Y_i^t$, better, when compared to matching $P(Y_i^t)$ with $P(Y_i^s)$ 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 $P(M_i(X_i), Y_i^{depth}, Y_i^{normal}, Y_i^{seg})$, is matched with true distribution denoted by $P(M_i(X_i), Y_i^{depth}, Y_i^{normal}, Y_i^{seg})$, 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.

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>sup.</th>
<th>Error $\downarrow$</th>
<th>Accuracy $\uparrow$ ($\gamma = 1.25$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rel</td>
<td>(m)</td>
<td>($\delta &lt; \gamma$)</td>
</tr>
<tr>
<td>Saxena et al. [51]</td>
<td>795</td>
<td>0.349</td>
<td>1.214 0.447 0.745 0.891</td>
</tr>
<tr>
<td>Liu et al. [33]</td>
<td>795</td>
<td>0.230 0.095 0.824</td>
<td>0.614 0.883 0.975</td>
</tr>
<tr>
<td>Eigen et al. [9]</td>
<td>120k</td>
<td>0.215 0.097</td>
<td>0.611 0.887 0.971</td>
</tr>
<tr>
<td>Roy et al. [47]</td>
<td>795</td>
<td>0.187 0.078 0.744</td>
<td>- - -</td>
</tr>
<tr>
<td>Laina et al. [30]</td>
<td>96k</td>
<td>0.129 0.056 0.583</td>
<td>0.801 0.950 0.986</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>sup.</th>
<th>Error $\downarrow$</th>
<th>Accuracy $\uparrow$ (mean median)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rel</td>
<td>(m)</td>
<td>($\delta &lt; \gamma$) $\lambda &lt; \sigma$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$11.25^- 22.5^- 30^-$</td>
</tr>
<tr>
<td>Eigen et al. [8]</td>
<td>120k</td>
<td>22.2 15.3</td>
<td>38.6 64 73.9</td>
</tr>
<tr>
<td>PBRS [63]</td>
<td>795</td>
<td>21.74 14.75</td>
<td>39.37 66.25 76.06</td>
</tr>
<tr>
<td>SURGE [59]</td>
<td>795</td>
<td>20.7 12.2</td>
<td>47.3 68.9 76.6</td>
</tr>
<tr>
<td>GeoNet [42]</td>
<td>30k</td>
<td>19.0 11.8</td>
<td>48.4 71.5 79.5</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>sup.</th>
<th>Mean IOU</th>
<th>Mean Accuracy</th>
<th>Pixel Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBRS [63]</td>
<td>795</td>
<td>0.332</td>
<td>0.600</td>
<td></td>
</tr>
<tr>
<td>Long et al. [35]</td>
<td>795</td>
<td>0.292</td>
<td>0.700</td>
<td></td>
</tr>
<tr>
<td>Liu et al. [32]</td>
<td>795</td>
<td>0.406 0.536</td>
<td>0.721</td>
<td></td>
</tr>
<tr>
<td>Kong et al. [26]</td>
<td>795</td>
<td>0.445</td>
<td>-</td>
<td>0.721</td>
</tr>
<tr>
<td>RefineNet (Res50) [31]</td>
<td>795</td>
<td>0.438 0.619</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

To demonstrate generalizability of UM-Adapt, we consider outdoor-scene dataset for two different tasks, semantic 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 256×512 bins and 128×256 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 implementation of depth-estimation, semantic 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 228×304 with various augmentations of scale and flip following [9], and outputs three task specific maps, each of size 128×160. 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.
door dataset, we train only the following two task-transfer networks: $N \rightarrow \text{seg}(Y_s^{\text{depth}}, Y_s^{\text{normal}})$ and $N \rightarrow \text{depth}(Y_s^{\text{seg}})$ 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 $N \rightarrow \text{seg}(Y_s^{\text{depth}})$ and $N \rightarrow \text{depth}(Y_s^{\text{seg}})$. Following this, two separate base-models 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 $C_{\text{contour}}$ 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 adaptation 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 $N \rightarrow \text{seg}(Y_s^{\text{depth}}, Y_s^{\text{normal}})$ 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-
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 semantic 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

<table>
<thead>
<tr>
<th>Method</th>
<th>Sup.</th>
<th>Depth rms Err. (m)</th>
<th>Seg. Err. (100-IoU)</th>
<th>Normals Err. (1-IoU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kendall et al. [24]</td>
<td>30k</td>
<td>0.702</td>
<td>-</td>
<td>0.182</td>
</tr>
<tr>
<td>GradNorm [65]</td>
<td>30k</td>
<td>0.663</td>
<td>67.5</td>
<td>0.155</td>
</tr>
<tr>
<td>UM-Adapt-S</td>
<td>795</td>
<td>0.637</td>
<td>55.6</td>
<td>0.139</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Semantic (mean IOU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kendall et al. [24] (Uncert. Weights)</td>
<td>51.52</td>
</tr>
<tr>
<td>Liu et al. [34]</td>
<td>52.68</td>
</tr>
<tr>
<td>UM-Adapt-S</td>
<td>54.4</td>
</tr>
</tbody>
</table>

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 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.

Table 8. Transfer learning results on novel unseen tasks.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Classification</th>
<th>Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motion Seg. [41]</td>
<td>ResNet-101</td>
<td>48.29</td>
<td>61.13</td>
</tr>
<tr>
<td>Exemplar [7]</td>
<td>ResNet-101</td>
<td>53.08</td>
<td>60.94</td>
</tr>
<tr>
<td>RP+Col+Ex+MS [7]</td>
<td>ResNet-101</td>
<td>69.30</td>
<td>70.53</td>
</tr>
<tr>
<td>UM-Adapt-S</td>
<td>ResNet-50</td>
<td>69.51</td>
<td>70.02</td>
</tr>
</tbody>
</table>

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.

Acknowledgements. This work was supported by a CSIR Fellowship (Jogendra) and a grant from RBCCPS, IISc. We also thank Google India for the travel grant.
References

[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
[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
[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. 2014. 4
[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
[46] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In ECCV, 2016. 6
[53] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In ECCV Workshops, 2016. 2
[60] Saining "Xie and Zhuowen" Tu. Holistically-nested edge detection. In ICCV, 2015. 4