This WACV 2022 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Semi-supervised Multi-task Learning for Semantics and Depth

Yufeng Wang[1], Yi-Hsuan Tsai[2], Wei-Chih Hung[3], Wenrui Ding[1], Shuo Liu[1], Ming-Hsuan Yang[4]

[1]Beihang University, [2]Phiar Technologies, [3]Waymo, [4]University of California at Merced

{wyfeng, ding, liush}@buaa.edu.cn, wasidennis@gmail.com, hungwayne@waymo.com, mhyang@ucmerced.edu

Abstract

Multi-Task Learning (MTL) aims to enhance the model generalization by sharing representations between related tasks for better performance. Typical MTL methods are jointly trained with the complete multitude of ground-truths for all tasks simultaneously. However, one single dataset may not contain the annotations for each task of interest. To address this issue, we propose the Semi-supervised Multi-Task Learning (SemiMTL) method to leverage the available supervisory signals from different datasets, particularly for semantic segmentation and depth estimation tasks. To this end, we design an adversarial learning scheme in our semi-supervised training by leveraging unlabeled data to optimize all the task branches simultaneously and accomplish all tasks across datasets with partial annotations. We further present a domain-aware discriminator structure with various alignment formulations to mitigate the domain discrepancy issue among datasets. Finally, we demonstrate the effectiveness of the proposed method to learn across different datasets on challenging street view and remote sensing benchmarks.

1. Introduction

Multi-Task Learning (MTL) aims to leverage information contained in multiple related tasks to improve the performance of each single task [66]. The potential advantages of MTL over separate learning of each task can be attributed to the generalization ability by sharing representations among related tasks as well as the benefit of multiple sources with supervision. It has been widely used in numerous tasks in computer vision [28], natural language processing [12], and speech recognition [29], to name a few.

Recently, deep convolutional neural networks (CNNs) have been successfully applied to dense prediction tasks such as semantic segmentation [37, 7] and depth estimation [14, 16]. One of the reasons for this success is the construction of large-scale and diverse datasets with pixel-wise annotations. Typical MTL methods train all tasks simultaneously within one dataset that contains the complete multitude of ground-truths. However, in the real-world scenario, a single dataset usually does not contain all necessary ground-truths. In addition, annotating the dataset for missing tasks entails significant effort and time, especially for the dense prediction tasks. To tackle this issue, one can leverage different datasets that contain the corresponding annotations for each task. Therefore, it is of great interest and importance to enable the network to leverage different supervision information from diverse datasets in the MTL framework [28, 35].

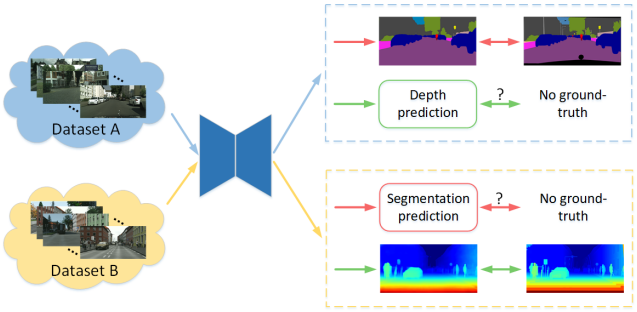

Figure 1. **Problem illustration.** Given two datasets $\mathcal{A}$ and $\mathcal{B}$, each of which only provides the annotation for partial tasks, we exploit the multi-tasking across datasets and mitigate the domain discrepancy to learn a more generalized model for all tasks on both datasets.

We consider the setting where only disjoint datasets are at our disposal with partial ground-truths, *e.g.*, dataset $\mathcal{A}$ developed for semantic segmentation and dataset $\mathcal{B}$ collected for depth estimation, as shown in Figure 1. It is in line with our intuition that a model can learn generalized representations from different datasets while ensuring at least one reliable supervision to train each task. To this end, one straightforward approach is to learn from one dataset for one task at a time, and alternatively train the joint model for MTL [28]. Nevertheless, such a learning scheme does not consider domain gaps across datasets [56, 44].

In this work, we propose a Semi-supervised Multi-Task Learning (SemiMTL) method to expand the MTL setting for training the joint model across different datasets with partial annotations. One challenge is how we train a multi-task model from diverse datasets, where each may only contain partial ground-truths for one task. Here, the problem can be treated within the semi-supervised learning (SSL) paradigm where the labeled data in one dataset is fully supervised and the unlabeled data in other datasets is used in a semi-supervised manner. Specifically, we use the adversarial learning [18] in the structured prediction space for semi-supervised training and iterate a similar scheme on all datasets and tasks. As such, all tasks are accomplished in the MTL setting. The second challenge arises from the domain discrepancy across diverse datasets as the data distributions often vary significantly. Therefore, it is important to align features across domains to learn a model that generalizes to different datasets. However, unlike the common setting using adversarial alignment [8, 56] that considers two domains to distinguish the real-fake distributions, we have multiple prediction distributions from diverse datasets. Therefore, we propose a domain-aware discriminator structure and analyze various learning modes to mitigate the domain discrepancy. The strategy differentiates the prediction distributions from multiple domains, which in turn enforces the generator to produce more plausible results to confuse the discriminator.

In practice, we focus on two fundamental yet challenging pixel-wise prediction tasks: semantic segmentation [37, 7] and depth/height estimation [14, 16]. These two tasks learn the semantic and geometric properties for scene understanding, where their correlation has been explored by joint training [67, 2]. We validate the effectiveness of our SemiMTL method in the challenging autonomous driving and remote sensing scenes, under various settings including the cross-city, cross-dataset, and cross-domain scenarios.

The contributions of this work are summarized as follows: 1) We propose a multi-task learning setting that leverages supervisory signals from diverse datasets in a semi-supervised paradigm; 2) We introduce a domain-aware adversarial learning approach to address the domain discrepancy problem during training across different datasets; 3) We demonstrate the effectiveness of our proposed method for semi-supervised multi-tasking across datasets in challenging street view and remote sensing scenarios.

2. Related Work

Multi-task Learning. MTL has been widely used in vision tasks, such as instance segmentation [17], semantic segmentation [36, 70], and face analysis [23]. As discussed in [52], MTL is typically conducted with either hard or soft parameter sharing of hidden layers in the context of deep learning [42, 28, 61, 41, 68, 71, 58]. On the other hand, several approaches explore to adaptively calibrate the relative losses of different tasks instead of a naive weighted summation [9, 53, 26, 63, 4, 59]. MTL can also be integrated with other learning paradigms, including unsupervised [72], self-supervised [51], and transfer learning [64, 49, 6], to either improve the performance of supervised MTL via additional information or use MTL to facilitate other paradigms [66]. A more comprehensive discussion of deep MTL methods can be found in [57]. Note that this study aims to solve a new learning paradigm for MTL rather than designing specific MTL architectures, as our method is compatible with other general MTL networks.

A few semi-supervised MTL methods have been developed [38, 10], but do not address the challenging pixel-level tasks to train models across different datasets with the absence of ground-truths. While some efforts have been made [35, 28, 46, 49, 6], these approaches do not conduct synchronous MTL on different datasets. To optimize jointly with the new labeled task, [35] preserves the models trained on old tasks to provide pseudo-ground-truth for these unlabeled tasks, where the joint training strategy can be seen as an upper-bound of their performance. The UberNet [28] model is proposed to update network parameters after observing sufficient samples to simulate asynchronous joint training. However, it does not account for domain gaps across diverse datasets. On the other hand, several recent approaches [46, 49, 6] mainly tackle knowledge transfer across tasks rather than across datasets with partial annotations.

Semi-supervised Learning. SSL methods leverage the vast amount of unlabeled data for classification and regression problems. Perturbation-based methods [45, 39] aim to utilize a teacher model to teach a student module whose predictions should be consistent. Similarly, several approaches exploit the fusion strategy by stochastic feature selection [33] or learning from multiple regressors [34]. Another line of research in SSL encourages the model to generate confident predictions on unlabeled data, *e.g.*, entropy minimization [19] and pseudo-labeling [24]. Auxiliary tasks can also be applied for SSL to integrate supervised and un-/weakly-supervised learning [50, 22]. More recently, several models in the adversarial setting [18] have been developed to either generate realistic samples for better discrimination [54] or distinguish directly the prediction for better generation [24, 31]. Nevertheless, the methods based on adversarial learning are not designed within the MTL framework.

Dense Prediction for Scene Understanding. Scene understanding involves a group of regression and classification tasks, and we discuss the most related work for semantic segmentation and depth estimation. Semantic segmentation can be treated as a pixel-wise classification problem tackled via deep models, such as the FCN [37] and Deeplab [7] networks. The recent methods mainly emphasize on learning and assembling features from multiple scales [69, 20] or multiple layers [3, 55], or leveraging global context in-

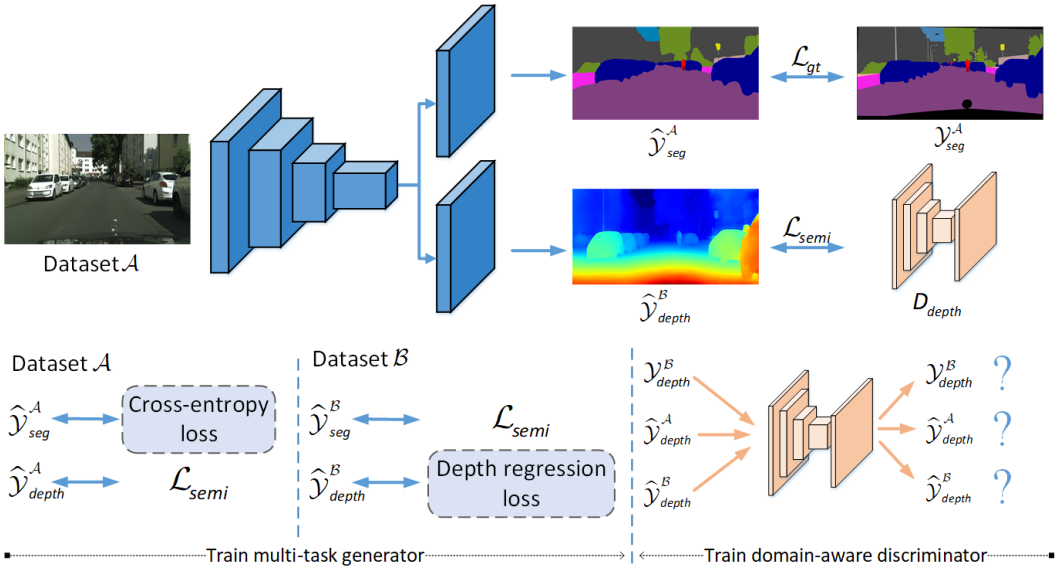

Figure 2. **Overview of the proposed SemiMTL method.** We design the SemiMTL algorithm based on an adversarial learning framework. During the training process within each dataset, we leverage the ground-truth to supervise the labeled tasks and construct one task-specific discriminator for each unlabeled task to provide semi-supervisory signals. We optimize the MTL network simultaneously for all task branches over all datasets. The discriminators can be updated after observing the ground-truth and predictions from different datasets to learn the domain knowledge.

formation [25, 65]. Similarly, deep models have been successfully applied to depth estimation [14, 16], and numerous algorithms have been developed through supervised [62], semi-/self-supervised [48, 15], unsupervised [16, 44] , and multi-tasking [30, 11] settings. Since these two tasks are closely related to learning the semantics and geometry of scenes, it is of great interest to accomplish them in a unified framework [67, 2].

3. SemiMTL

In this work, we treat the labeled and unlabeled tasks as supervised and semi-supervised problems respectively during the training process within one dataset, which allows us to leverage unlabeled data to further train the task branches without annotations. We propose the SemiMTL method within the adversarial learning framework [24, 56], where the discriminator and adversarial loss play the role of training signals when the annotation is not available for the training samples. In addition, we present a domain-aware structure for the discriminator and analyze different alignment patterns to address the domain gap in multi-tasking across datasets. This alleviates the domain discrepancy issue while leveraging the supervision signals from diverse datasets. In the remainder of this section, we formulate the proposed SemiMTL framework for dense prediction tasks in scene understanding.

3.1. Approach Overview

Problem Definition. We start with considering a typical MTL problem over an input space $\mathcal{X}$ and a collection of task output spaces $\{\mathcal{Y}_t\}_{t\in\mathcal{T}}$, where $\mathcal{T}$ is the total task set. Given such a dataset, we wish to learn the prediction model per task as $G_t(x;\theta^{sh},\theta^t) : \mathcal{X} \rightarrow \mathcal{Y}_t$, where θ^{sh} are the shared parameters among tasks and θ^t are the task-specific parameters. In this work, we address a different MTL setting for training across datasets, where each dataset only contains annotations for partial tasks. Therefore, we extend the setting in [28, 64] and denote the input spaces $\{\mathcal{X}^k\}_{k\in\mathcal{S}}$, where $\mathcal{S}$ is the set of all datasets. Here we assume to have two datasets ($\mathcal{A}$ and $\mathcal{B}$) and two tasks ($\mathcal{T}_{seg}$ and $\mathcal{T}_{depth}$) as semantic segmentation and depth estimation. However, each dataset only has some supervision, *i.e.*, $\mathcal{Y}^{\mathcal{A}}_{seg}$ for task $\mathcal{T}_{seg}$ in dataset $\mathcal{A}$ and $\mathcal{Y}^{\mathcal{B}}_{depth}$ for task $\mathcal{T}_{depth}$ in dataset $\mathcal{B}$, as shown in Figure 2.

Baseline: Joint Training. We first apply a joint training baseline method [28] that iteratively explores each dataset and updates the model (*i.e.*, θ^{sh} and θ^t) after observing sufficient annotated samples for each task. However, leveraging only partial annotations from each dataset may produce bias in the shared encoder as it does not observe gradients from unlabeled data. As a result, the model may perform well on the labeled tasks while generalizing poorly on the unlabeled tasks in each dataset.

Proposed Method. To address the above-mentioned issue, we propose to train the model on one domain $\mathcal{X}^k$ by minimizing the supervised loss for labeled task $\mathcal{T}_t$ with annotated samples (x^k, y^k_t) in $(\mathcal{X}^k, \mathcal{Y}^k_t)$, as well as the semi-supervised loss for unlabeled tasks $\mathcal{T} \setminus \mathcal{T}_t$ with identical samples x^k in $\mathcal{X}^k$ that do not have corresponding annotations. To consider all the input domains $\mathcal{X}$, we iteratively apply the above training scheme over each dataset $\mathcal{X}^k$ to fully leverage the

supervisory signals for each task.

With this formulation, our model is able to optimize all the task-specific decoders $\{F_t; \theta^t\}_{t\in\mathcal{T}}$ simultaneously with supervisions either from the supervised loss or the semi-supervised one on unlabeled data. Therefore, the shared encoder $\{E; \theta^{sh}\}$ can also update with gradients accumulated from the supervision of all tasks on both labeled and unlabeled data, which avoids the bias only on labeled data. We will describe the details about our proposed framework and semi-supervised loss in the following sections.

3.2. Objective Function

We formulate our SemiMTL problem as an adversarial learning framework, which consists of two modules: the generator G and the discriminators $\{D_t\}_{t\in\mathcal{T}}$. The generator G is a multi-task network that contains a shared encoder E parameterized by θ^{sh} and task-specific decoders $\{F_t\}_{t\in\mathcal{T}}$ parameterized by θ^t.

Typical MTL Loss. Given an input image $x \in \mathbb{R}^{H\times W\times 3}$, the typical MTL loss function is defined by the task-specific supervised loss $\mathcal{L}^t_{gt}$ for task t, with a weight w_t to balance the loss functions among tasks:

$$\mathcal{L}_{mtl} = \sum_{t=1}^{|\mathcal{T}|} w_t \mathcal{L}^t_{gt}(F_t(E(x)), y_t), \tag{1}$$

where y_t is the ground-truth for task t.

Discriminator Module. In our setting, each dataset may not contain the label for all the tasks, indicating that the task-specific branches cannot be supervised by unlabeled data via Eq. (1). Here, our goal is to accomplish all tasks to simultaneously update both the shared encoder and task-specific decoders using both labeled and unlabeled data. To this end, we utilize adversarial learning to construct a semi-supervised objective for the data without ground-truths. Our approach is motivated by the observation that the output space is structured in dense prediction tasks such as semantic segmentation and depth estimation [24, 56]. For example, the street-view images might have significant differences in appearance, but their outputs share many similarities such as spatial layout and local context.

In [56], they introduce a discriminator to distinguish whether the distribution is from the ground-truth or the prediction of unlabeled data. Differently, our method deals with the SemiMTL setting that contains labeled data from one domain and unlabeled data from other domains. Thus, we introduce a discriminator module that can tell which domain that the prediction comes from, *i.e.*, either the ground-truth or the prediction of the domain $\mathcal{X}^k$. Specifically, we first forward the input image x to the generator network $G = \{E; F_t\}$ and produce the task-specific prediction $\hat{y}_t = F_t(E(x))$ for task t, which is then taken as the input to the discriminator. We minimize the cross-entropy loss $\mathcal{L}^t_D$ for the task-specific discriminator D_t:

$$\mathcal{L}^t_D = \mathcal{L}_{ce}(z_t, c) = -\sum_{h,w} c \log(D_t(z_t)^{(h,w,c)}), \tag{2}$$

where c is the one-hot domain label and z_t denotes the input to the discriminator, which could be the ground-truth y_t ($c = 0$) or the prediction $\hat{y}_t$ from the c-th dataset (domain). In this paper, c is a 3-dimensional one-hot vector, in which a three-way classifier is utilized in the discriminator to tell whether the input is from the ground-truth or the prediction from dataset $\mathcal{A}$ or $\mathcal{B}$.

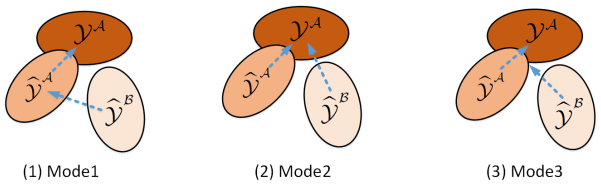

Figure 3. **Illustration of alignment strategies.** In our setting, assume only dataset $\mathcal{A}$ contains the annotations $\mathcal{Y}^{\mathcal{A}}$ for task $\mathcal{T}_{seg}$. The predictions $\hat{\mathcal{Y}}^{\mathcal{A}}$ and $\hat{\mathcal{Y}}^{\mathcal{B}}$ are obtained from the labeled dataset $\mathcal{A}$ and unlabeled dataset $\mathcal{B}$ respectively. For the alignment of this task, we directly enforce $\hat{\mathcal{Y}}^{\mathcal{A}}$ as close to ground-truth distribution as possible, while align $\hat{\mathcal{Y}}^{\mathcal{B}}$ with different strategies, that is, to the distributions of (1) prediction $\hat{\mathcal{Y}}^{\mathcal{A}}$, (2) ground-truth $\mathcal{Y}^{\mathcal{A}}$, or (3) their intersection $\hat{\mathcal{Y}}^{\mathcal{A}} \cup \mathcal{Y}^{\mathcal{A}}$.

Adversarial Loss. Based on the discriminator, our goal is to provide training signals for unlabeled data and enforce the prediction to be close to the ground-truth distribution. However, it is not trivial to directly apply conventional adversarial alignment [24, 56] as our setting involves predictions from multiple datasets. Here, we investigate several alignment strategies that can achieve the desired goal, as shown in Figure 3. Suppose the dataset $\mathcal{A}$ is labeled for task t but not for the dataset $\mathcal{B}$, we note that there are three types of distributions for each task: the ground-truth $y^{\mathcal{A}}_t$, the prediction from the labeled dataset $\hat{y}^{\mathcal{A}}_t$, and the prediction from the unlabeled dataset $\hat{y}^{\mathcal{B}}_t$. We can directly align the distribution of $\hat{y}^{\mathcal{A}}_t$ to $y^{\mathcal{A}}_t$ as they are from the same domain $\mathcal{A}$ (intra-domain loss):

$$\mathcal{L}^t_{intra} = -\sum_{h,w} \log(D_t(\hat{y}^{\mathcal{A}}_t)^{(h,w,0)}), \tag{3}$$

where label 0 indicates the ground-truth. For the inter-domain loss, we exploit different training modes and construct the corresponding loss function. Here, we can choose to align the distribution of prediction $\hat{y}^{\mathcal{B}}_t$ from the unlabeled data to the one from the labeled domain, *i.e.*, $\hat{y}^{\mathcal{A}}_t$ (labeled as 1) as Mode 1:

$$\mathcal{L}^t_{inter} = -\sum_{h,w} \log(D_t(\hat{y}^{\mathcal{B}}_t)^{(h,w,1)}), \tag{4}$$

or the ground-truth $y^{\mathcal{A}}_t$ as Mode 2:

$$\mathcal{L}^t_{inter} = -\sum_{h,w} \log(D_t(\hat{y}^{\mathcal{B}}_t)^{(h,w,0)}), \tag{5}$$

Algorithm 1 Training procedure of the SemiMTL method.

for iteration $i = 1$ to N **do**
  for dataset $k \in \{1, ..., K\}$ **do**
    {Construct mini-batch}
    for task $t = 1, ..., m$ with ground-truth **do**
      {calculate gradients for E and F_t}
      $\mathbf{L}_G^t \leftarrow w_t \mathcal{L}_{gt}^t + \lambda_{intra} \mathcal{L}_{intra}^t$
      {calculate gradients for D_t}
      $\mathbf{L}_D^{t,k} \leftarrow \mathcal{L}_D^t(y_t, 0) + \mathcal{L}_D^t(\hat{y}_t, k)$
    end for
    for task $t = m + 1, ..., |\mathcal{T}|$ without ground-truth **do**
      {calculate gradients for E and F_t}
      $\mathbf{L}_G^t \leftarrow \lambda_{inter} \mathcal{L}_{inter}^t$
      {calculate gradients for D_t}
      $\mathbf{L}_D^{t,k} \leftarrow \mathcal{L}_D^t(\hat{y}_t, k)$
    end for
    {Freeze $\{D_t\}$, and update G with}
    $\mathbf{L}_G = \sum_{t=1}^{|\mathcal{T}|} \mathbf{L}_G^t$
  end for
  {Freeze G, and update $\{D_t\}$ with}
  $\mathbf{L}_D^t = \sum_{k=1}^{K} \mathbf{L}_D^{t,k}$
end for

or the joint distribution of $y_t^{\mathcal{A}}$ and $\hat{y}_t^{\mathcal{A}}$ as Mode 3:

$$\mathcal{L}_{inter}^t = -\sum_{h,w} \log(1 - D_t(\hat{y}_t^{\mathcal{B}})^{(h,w,2)}), \tag{6}$$

where label 2 denotes the prediction of unlabeled data $\hat{y}_t^{\mathcal{B}}$. Hence we define the semi-supervised loss for task t as

$$\mathcal{L}_{semi}^t = \lambda_{intra} \mathcal{L}_{intra}^t + \lambda_{inter} \mathcal{L}_{inter}^t, \tag{7}$$

where λ_{intra} and λ_{inter} indicate the weight for the intra- and inter-domain adversarial losses, respectively.

We denote the SemiMTL framework applying these alignment modes as SemiMTL (M1), SemiMTL (M2), and SemiMTL (M3), respectively. We utilize the SemiMTL (M2) mode as our default implementation. Their effect on the performance of tasks is illustrated in Section 4.3. Note that we utilize the same adversarial training scheme to the task $\mathcal{T}_{seg}$ and $\mathcal{T}_{depth}$, which is also applicable to other similar tasks.

3.3. Optimization

We apply the synchronous SGD for joint training [28] as the baseline training method, where we extract mini-batches from every dataset iteratively and optimize all parameters synchronously using (1) after observing labeled samples for each task. With the baseline training approach, we construct our SemiMTL model, where the unlabeled tasks guided by the discriminators can be optimized simultaneously with the labeled ones. The main steps of this process are summarized in Algorithm 1. Within each training iteration, we minimize the overall objective functions for the generator:

$$\mathbf{L}_G = \sum_{t=1}^{|\mathcal{T}|} w_t \mathcal{L}_{gt}^t + \mathcal{L}_{semi}^t, \tag{8}$$

and for each discriminator D_t:

$$\mathbf{L}_D^t = \mathcal{L}_D^t(y_t, 0) + \sum_{k=1}^{K} \mathcal{L}_D^t(\hat{y}_t, k), \tag{9}$$

where K is the number of datasets/domains. The SemiMTL model is iteratively trained in a way similar to the GAN [18] method: the discriminator D_t aims to classify the ground-truth/predictions from different domain distributions, while the generator G attempts to fool D_t by producing predictions that are as indistinguishable to the ground-truth as possible.

3.4. Network Architecture

Encoder and Decoder Networks. The proposed SemiMTL framework can integrate any type of deep MTL architectures. Here we adopt the commonly utilized encoder-decoder MTL model that consists of a shared encoder coupled with two task-specific decoders to estimate segmentation and depth tasks. We leverage the ResNet101 [21] backbone for the shared encoder to obtain deep feature representations, which are passed to two parallel branches for independent task decoding. The segmentation decoder is built upon the PSP module [69] to increase contextual information for semantics, followed by a softmax layer to predict semantic classes. The depth decoder is constructed with several convolutional layers and up-sampling operations to produce detailed depth features and a regression layer to estimate depth. Finally, we apply an up-sampling layer to the output maps for both tasks to match the input image size. To optimize the network, we adopt the cross-entropy loss for semantic segmentation and the BerHu loss [32] for depth estimation.

Discriminator Networks. The structure of the discriminator network is similar to that in [47]. It consists of 5 convolution layers with 4×4 kernel and $\{64, 128, 256, 512, K\}$ channels in the stride of 2. The first four convolution layers are all followed by a spectral normalization layer [43] to stabilize the training process and a leaky ReLU [40] unit parameterized by 0.2. We implement an up-sampling layer to transform the output to the input size. The discriminators for both tasks share the same architecture except for the input layer which takes the segment and depth maps respectively.

4. Experimental Results and Analysis

To demonstrate the effectiveness of the SemiMTL method, we carry out experiments on several publicly available datasets for scene understanding, including the ISPRS Potsdam and Vaihingen [1] remote sensing datasets, and the real-world Cityscapes [13] and synthetic Synscapes [60] street-view datasets. In particular, we evaluate the algorithms in three scenarios, including the cross-city, cross-dataset, and cross-domain settings with various datasets. In the following, we will describe the experimental details.

4.1. Experimental Setup

Datasets. The Cityscapes dataset contains high-resolution outdoor images for urban scene understanding, which is collected from 50 diverse cities. It is annotated with pixel-wise semantic labels, associated with pre-computed disparity maps that can serve as pseudo depth ground-truth. The Cityscapes-depth dataset is an additional train-extra set of Cityscapes with disparity ground-truths, which is collected from different cities with the Cityscapes. The Synscapes dataset is generated by photorealistic rendering techniques to parse synthetic street scenes, containing 25K RGB images together with accurate pixel-wise class and depth annotations. Since this dataset does not provide an official split, we take the consequent 20K/1K/4K samples as our training/validation/test sets respectively. We estimate the inverse depth to represent points at an infinite distance like the sky as zero. As the images of these datasets are of high resolution, we resize the images to 512×1024 for experiments.

The ISPRS Potsdam and Vaihingen datasets were acquired by flight campaigns over German cities, accompanied by digital images, semantic labels, and digital surface model (DSM) height data. The digital images were captured by the airborne color-infrared camera in different channels: near-infrared (NIR)/infrared (IR), red (R), and green (G). The DSM data was acquired by LiDAR and the normalized DSM (nDSM) data was also made available, which is normalized to the range [0,1] in our experiments. The images in Potsdam and Vaihingen are composed of the IRRG bands at a ground sample distance (GSD) of 5cm and the NIRRG bands at a GSD of 9cm, respectively. We follow the official split for training and testing set as in [1]. In our experiments, we resize the image data in Potsdam to the GSD of 9cm to match with the Vaihingen dataset. We then extract patches of size 512×512 from the raw high-resolution images using a 50% overlapped sliding window along both the row and column.

Evaluation Metrics. For the evaluation of segmentation, we use the mean pixel accuracy (pAcc) and mean Intersection over Union (mIoU) metrics. The pAcc indicates the total accuracy of pixels regardless of classes while the mIoU is computed by averaging the Jaccard scores over all predicted categories. To evaluate the depth task, we adopt several quantitative metrics following [14, 16], including (a) abs relative error (AbR), (b) root mean squared error (RMSE), and (c) accuracy with thresholds: % of $\hat{y}_n$ s.t. $\max(\frac{\hat{y}_n}{y_n}, \frac{y_n}{\hat{y}_n}) = \delta_i < 1.25^i$ $(i \in [1, 2, 3])$, where $\hat{y}_n$ and y_n denote the prediction and ground-truth of depth at the n-th pixel. We also measure the multi-task performance ΔM [41], i.e. the average per-task performance gain of multi-task model m compared with the single-task baseline b :

$$\Delta M = \frac{1}{|\mathcal{T}|} \sum_{t=1}^{|\mathcal{T}|} (-1)^{l_t} (M_{m,t} - M_{b,t})/M_{b,t}, \qquad (10)$$

Table 1. **Quantitative results on the Cityscapes dataset.** We train all methods with the training and train-extra sets, and evaluate them on the validation set for both tasks. The cyan metrics indicate lower is better while pink ones mean higher is better.

Method	Segmentation		Depth					MTL
	pAcc	mIoU	AbR	RMSE	δ_1	δ_2	δ_3	$\Delta M(\%)$
STL_Seg	94.8	71.4	-	-	-	-	-	0.0
STL_Depth	-	-	0.414	6.744	67.6	84.5	92.0	0.0
JTL [28]	94.8	71.4	0.329	5.469	76.6	91.2	95.7	+9.4
SemiMTL	**94.9**	**71.9**	**0.287**	**5.234**	**79.3**	**92.6**	**96.3**	**+11.5**

where M indicates the representative measure for each task and we adopt the mIoU and RMSE metric as in [41]. $l_t = 1$ if a lower M means a better performance, and 0 otherwise.

Implementation Details. The SemiMTL method is implemented with PyTorch using Nvidia Titan RTX GPUs. We initiate the encoder backbone parameters with the ResNet101 [21] model pre-trained on ImageNet, and the decoders and discriminators are randomly initialized. We perform the data augmentation on the fly following [65] and fix the crop size during the training process. The MTL network is trained by the standard SGD optimizer [5] with momentum 0.9 and weight decay 10^{-4}. The learning rate is initialized by 0.01 and decreased using the polynomial decay with power 0.9. We adopt the Adam optimizer [27] for training the discriminators with learning rate as 10^{-4} and momentum as (0.9, 0.99). In all experiments, we fix the task weights as $w_{seg} = 1.0$ and $w_{depth} = 0.01$ in the MTL loss and set $\lambda_{intra} = 0.001$ and $\lambda_{inter} = 0.0001$ to balance the semi-supervised adversarial losses. We use the same hyper-parameters among all methods for fair comparisons.

4.2. Evaluation of SemiMTL Framework

We compare the experimental results of the SemiMTL approach with different baselines. We first build the models for each task with identical encoder structure and task-specific decoder head, termed as single task learning (STL). Then we utilize the joint training algorithm [28], named joint task learning (JTL), to train the MTL model across datasets. We also apply a domain adaptation algorithm [56] to both STL and JTL schemes, which does not consider the prediction distributions from different domains. Extensive experiments demonstrate the effectiveness of our method to leverage semi-supervised information during multi-tasking across datasets.

Across Cities. In this setting, we conduct experiments on the Cityscapes and Cityscapes-depth datasets where the former and latter only provide segmentation and depth ground-truths, respectively. They are captured from different European cities at different seasons, which can verify our method in a small domain gap scenario. We train the methods on the training sets of two datasets while evaluating them on the Cityscapes validation set containing the labels for both tasks. We fix the crop size as 256×256 during the training step.

Table 1 shows the evaluation results of our proposed algorithm against baseline methods, where our method achieves

Table 2. **Quantitative results on the Potsdam and Vaihingen datasets.** We train all the methods with the segmentation ground-truth from Potsdam and depth ground-truth from Vaihingen, while evaluating the performance of each task on both datasets to validate whether the above trained models generalize well across datasets.

	Method	Segmentation		Depth					MTL
		pAcc	mIoU	AbR	RMSE	δ_1	δ_2	δ_3	$\Delta M(\%)$
Potsdam	STL_Seg	89.5	79.7	-	-	-	-	-	0.0
	STL_Depth	-	-	6.926	4.686	16.1	24.8	33.7	0.0
	DA_Depth [56]	-	-	4.985	4.677	28.5	36.3	46.4	-
	JTL [28]	90.0	80.7	2.517	4.430	31.2	41.7	52.3	+3.4
	SemiSD [56]	90.2	80.8	2.695	4.322	34.1	44.7	55.3	+4.6
	SemiMTL	**90.5**	**81.4**	**2.420**	**4.217**	**38.4**	**47.0**	**57.3**	**+6.1**
Vaihingen	STL_Seg	64.3	42.4	-	-	-	-	-	0.0
	DA_Seg [56]	68.8	47.5	-	-	-	-	-	-
	STL_Depth	-	-	1.324	1.899	48.7	68.9	**78.5**	0.0
	JTL [28]	79.3	62.2	1.338	1.949	48.8	68.1	78.1	+22.1
	SemiSD [56]	81.4	63.8	1.432	2.088	49.6	67.5	78.1	+20.4
	SemiMTL	**81.6**	**64.9**	**1.316**	**1.802**	**50.9**	**69.4**	78.4	**+29.2**

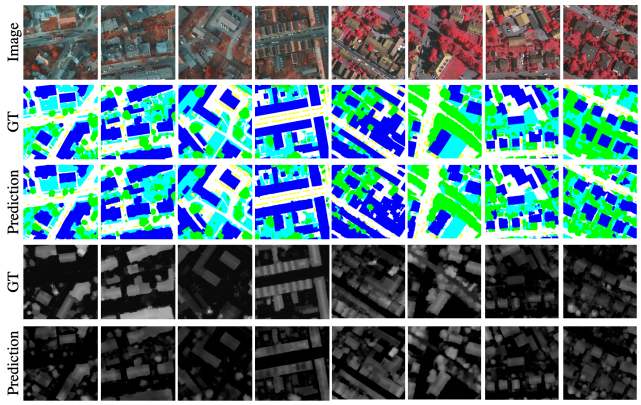

Figure 4. **Qualitative examples of the SemiMTL method on the remote sensing scenario.** The first and last four columns are the examples from the Potsdam and Vaihingen, respectively.

the best performance on both tasks. It is worth noting that compared with the separate training of each task, the joint training of both tasks achieves identical results for the segmentation task but obtains significant improvements on the depth task. The results indicate that the high-level segmentation task facilitates more on the low-level depth task. The proposed SemiMTL method improves further on all tasks compared with the JTL scheme.

Across Datasets. We conduct the cross-dataset experiments on the Potsdam and Vaihingen remote sensing datasets, which are collected with significantly different conditions such as imaging sensors, GSD, and even color channel compositions. Only the ground-truths of Potsdam segmentation and Vaihingen height are available during the training step. We then evaluate each task on the validation set of both datasets to verify the generalization ability, as the performance boost of all tasks on all datasets is preferred rather than improving only the specific tasks with supervision in each dataset. We fix the crop size as 384×384 for training.

The quantitative results of each task on both datasets are shown in Table 2. We adopt the domain adaptation (DA) method [56] to the STL of each task, namely, DA_Seg and DA_Depth. This method is originally proposed for semantic segmentation, however, it is also applicable to other dense prediction tasks with structured output space such as depth estimation. In the STL experiments, the models perform well on the supervised dataset but obtain poor results on the unsupervised dataset, whereas the DA method is able to improve the performance of every single task. The JTL [28] algorithm leverages more comprehensive supervision from related tasks in each dataset, which can learn useful complementary features in the shared encoder. As such, it achieves significant improvement for the task without supervision in each dataset, *e.g.*, the depth task in Potsdam and the segmentation task in Vaihingen.

Based on the above observation, we propose the SemiMTL strategy to further improve the performance of the unsupervised tasks in each dataset. We again employ the adversarial algorithm [56, 24] within our framework, where a task-specific discriminator provides the semi-supervisory signals for each task, termed as SemiSD. With the help of semi-supervision, the unsupervised tasks in each dataset are boosted against the JTL scheme. However, this scheme only transfers the features from the source dataset to target one, namely, aligning the prediction and ground-truth distributions. As such, we further consider the multi-domain issue and design the domain-aware discriminator to better differentiate the predictions, which in turn forces the generator to produce more realistic results to confuse the discriminator. As a result, our SemiMTL method not only improves the semi-supervised tasks with a large margin (4.0% mIoU gain in Vaihingen and 4.8% RMSE gain in Potsdam against the JTL baseline) but also achieves performance gains for the fully supervised tasks. It is worth noting that the performance of unlabeled tasks in each dataset is improved more significantly, which is in line with our motivation to leverage the semi-supervisory signals to improve the unlabeled tasks.

Across Domains. We further carry out experiments on the real-world Cityscapes and synthetic Synscapes datasets. The training on them is much more challenging to suffer from not only larger domain discrepancy across datasets but also differences between real and synthetic scenes. We perform the experiments similarly by training with the ground-truths of segmentation in Cityscapes and depth in Synscapes while evaluating each task on the validation set of both datasets.

The quantitative results of each task on both datasets are shown in Table 3. In the supervised experiments (segmentation of Cityscapes and depth of Synscapes), the JTL method improves the depth task similar to the cross-city setting, while degrading the segmentation mIoU due to the large domain gap among real-synthetic datasets. In contrast, our SemiMTL framework achieves mIoU performance gain by 1.2% and 1.8% against the STL and JTL baselines respectively, and also improves all metrics in the depth task. In the

Table 3. **Quantitative results on the Cityscapes and Synscapes datasets.** We train all methods with the ground-truths of segmentation from Cityscapes and depth from Synscapes, while evaluating each task on both datasets to validate the generalization ability.

	Method	Segmentation		Depth					MTL
		pAcc	mIoU	AbR	RMSE	δ_1	δ_2	δ_3	$\Delta M(\%)$
Cityscapes	STL_Seg	95.7	76.0	-	-	-	-	-	0.0
	STL_Depth	-	-	0.694	14.36	46.9	70.8	81.5	0.0
	JTL [28]	95.6	75.5	0.372	8.646	56.4	81.1	91.0	+19.6
	SemiSD [56]	95.6	75.8	0.349	7.893	59.3	81.6	91.5	+22.3
	SemiMTL (M1)	95.7	76.2	0.356	7.959	58.8	81.4	91.3	+22.4
	SemiMTL (M3)	95.7	76.4	0.341	7.645	59.5	81.9	91.7	+23.7
	SemiMTL	**95.8**	**76.9**	**0.334**	**7.558**	**61.4**	**83.0**	**91.9**	**+24.3**
Synscapes	STL_Seg	90.9	61.9	-	-	-	-	-	0.0
	STL_Depth	-	-	0.505	6.214	85.4	94.1	96.8	0.0
	JTL [28]	91.3	63.4	0.486	5.307	87.4	95.8	96.8	+8.5
	SemiSD [56]	90.9	62.8	0.449	5.183	88.1	95.8	97.8	+9.0
	SemiMTL (M1)	91.5	**65.8**	0.411	5.153	86.8	94.1	96.9	+11.7
	SemiMTL (M3)	**91.6**	65.5	0.407	5.157	87.7	95.2	96.3	+11.2
	SemiMTL	91.4	65.4	**0.380**	**5.056**	**88.5**	**95.9**	**97.9**	**+12.1**

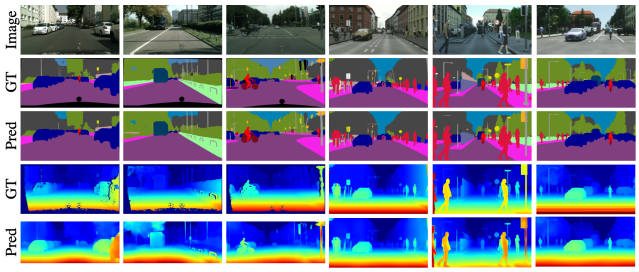

Figure 5. **Qualitative examples of the SemiMTL method on the street-view scenario.** The first and last three columns are the examples from the Cityscapes and Synscapes, respectively.

semi-supervised experiments (segmentation of Synscapes and depth of Cityscapes), the JTL method improves the performance of both tasks significantly, indicating that the observation of cross-domain samples can help the network to learn more generalized features. Compared with the JTL baseline, our SemiMTL framework further facilitates the segmentation task with a performance gain of 3.2% in mIoU and improves the depth task by 12.6% in RMSE. Figure 5 shows the qualitative results of the proposed SemiMTL algorithm. We also provide the comparisons for different methods in Figure 6, which indicates that the proposed method predicts more accurately for segmentation and estimates more sharply along boundaries and smoothly within regions for depth.

4.3. Ablation Study

To analyze the proposed approach thoroughly, we present the ablation study on the cross-domain setting in Table 3. We consider two baselines (JTL [28] and SemiSD [56, 24]) and different variants of the SemiMTL approach. As stated in Section 3.2 and Figure 3, the alternatives include (i) SemiMTL (M1): aligning the task output in unlabeled datasets to the prediction distribution in labeled datasets; (ii) SemiMTL (M2): encouraging the output to be similar to the ground-truth distribution, which is our default mode denoted as SemiMTL; (iii) SemiMTL (M3): enforcing the output to be close to the joint distribution of labeled prediction and ground-truth.

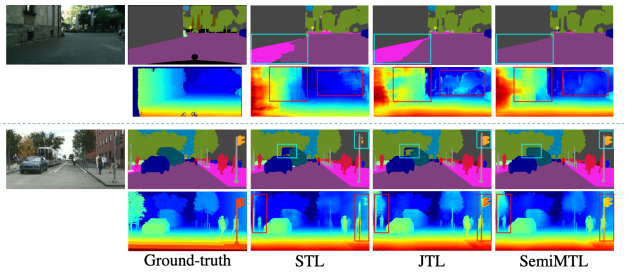

Figure 6. **Qualitative comparison for different methods.** The upper and lower examples are from the Cityscapes and Synscapes datasets respectively. The improvements are highlighted with cyan and red rectangles for segmentation and depth tasks respectively.

Effect of Adversarial Training. We first analyze the effect of directly utilizing the common adversarial learning method into the SemiMTL framework. We apply the adversarial scheme [56, 24] to construct the discriminators of both tasks for semi-supervision, denoted as SemiSD, which only distinguishes the ground-truth and prediction distributions without considering the domain gap problem. Table 3 shows that the SemiSD scheme performs better than the JTL baseline on both tasks of the Cityscapes dataset, but decreasing the segmentation IoU by 0.6% on the Synscapes dataset. These results show that a direct adversarial training lacks the generalization ability for tasks across domains.

Effect of Domain-aware Module. We further evaluate the effect of three different domain-aware modules which incorporate the domain information into training the discriminators. Table 3 illustrates that these variants of the SemiMTL model all perform better than the JTL and SemiSD methods on the segmentation task of both datasets. There is a slight performance loss in the SemiMTL (M1) model for the depth task, which shows that an ambiguous alignment to non-ground-truth distributions may not an effective way for the low-level tasks. However, the SemiMTL model performs consistently better than all baseline methods for all metrics on both datasets, indicating that the proposed semi-supervised MTL framework and domain-aware discriminators can learn more effective features and improve the performance of both tasks across domains.

5. Conclusions

In this paper, we present a new SemiMTL setting to address the multi-tasking across datasets. The proposed method is able to leverage the supervisory information from different domains and optimize all tasks simultaneously in a MTL model across datasets. We then introduce a domain-aware adversarial learning approach and various alignment modes to alleviate the domain discrepancy issue among datasets. We apply our SemiMTL model to two dense prediction tasks (semantic segmentation and depth estimation) on different challenging benchmarks. Experimental results demonstrate the proposed SemiMTL method performs favorably against the state-of-the-art approaches.

References

[1] 2d semantic labeling contest. http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html. Accessed July 1, 2020. 5, 6

[2] Amir Atapour-Abarghouei and Toby P. Breckon. Veritatem dies aperit - temporally consistent depth prediction enabled by a multi-task geometric and semantic scene understanding approach. In *CVPR*, 2019. 2, 3

[3] Piotr Bilinski and Victor Prisacariu. Dense decoder shortcut connections for single-pass semantic segmentation. In *CVPR*, 2018. 2

[4] Shubhankar Borse, Ying Wang, Yizhe Zhang, and Fatih Porikli. Inverseform: A loss function for structured boundary-aware segmentation. In *CVPR*, 2021. 2

[5] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, 2010. 6

[6] Ruchika Chavhan, Ankit Jha, Biplab Banerjee, and Subhasis Chaudhuri. Ada-at/dt: An adversarial approach for cross-domain and cross-task knowledge transfer. In *WACV*, 2021. 2

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018. 1, 2

[8] Yu Chen, Chunhua Shen, Hao Chen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial learning of structure-aware fully convolutional networks for landmark localization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 2

[9] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018. 2

[10] Yanhua Cheng, Xin Zhao, Rui Cai, Zhiwei Li, Kaiqi Huang, and Yong Rui. Semi-supervised multimodal deep learning for rgb-d object recognition. In *IJCAI*, 2016. 2

[11] Cheng Chi, Qingjie Wang, Tianyu Hao, Peng Guo, and Xin Yang. Feature-level collaboration: Joint unsupervised learning of optical flow, stereo depth and camera motion. In *CVPR*, 2021. 3

[12] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, 2008. 1

[13] 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. 5

[14] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014. 1, 2, 3, 6

[15] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 3

[16] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 1, 2, 3, 6

[17] Kratarth Goel, Praveen Srinivasan, Sarah Tariq, and James Philbin. Quadronet: Multi-task learning for real-time semantic depth aware instance segmentation. In *WACV*, 2021. 2

[18] 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, 5

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

[20] Junjun He, Zhongying Deng, and Yu Qiao. Dynamic multi-scale filters for semantic segmentation. In *ICCV*, 2019. 2

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6

[22] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *NeurIPS*, 2015. 2

[23] Zhizhong Huang, Junping Zhang, and Hongming Shan. When age-invariant face recognition meets face age synthesis: A multi-task learning framework. In *CVPR*, 2021. 2

[24] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *BMVC*, 2018. 2, 3, 4, 7, 8

[25] Wei-Chih Hung, Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Scene parsing with global context embedding. In *ICCV*, 2017. 3

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

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 6

[28] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*, 2017. 1, 2, 3, 5, 6, 7, 8

[29] Seltzer Michael L. and Jasha Droppo. Multi-task learning in deep neural networks for improved phoneme recognition. In *ICASSP*, 2013. 1

[30] Hsueh-Ying Lai, Yi-Hsuan Tsai, and Wei-Chen Chiu. Bridging stereo matching and optical flow via spatiotemporal correspondence. In *CVPR*, 2019. 3

[31] Wei-Sheng Lai, Jia-Bin Huang, and Ming-Hsuan Yang. Semi-supervised learning for optical flow with generative adversarial networks. In *NeurIPS*, 2017. 2

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

[33] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *CVPR*, 2019. 2

[34] Yu-Feng Li, Han-Wen Zha, and Zhi-Hua Zhou. Learning safe prediction for semi-supervised regression. In *AAAI*, 2017. 2

[35] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947, 2018. 1, 2

[36] Shuo Liu, Wenrui Ding, Chunhui Liu, Yu Liu, Yufeng Wang, and Hongguang Li. ERN: Edge loss reinforced semantic segmentation network for remote sensing images. *Remote Sensing*, 10(9):1339, 2018. 2

[37] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1, 2

[38] Xiaoqiang Lu, Xuelong Li, and Lichao Mou. Semi-supervised multitask learning for scene recognition. *IEEE Trans. Cybernetics*, 45(9):1967–1976, 2014. 2

[39] Yucen Luo, Jun Zhu, Mengxi Li, Yong Ren, and Bo Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *CVPR*, 2018. 2

[40] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013. 5

[41] Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. Attentive single-tasking of multiple tasks. In *CVPR*, 2019. 2, 6

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

[43] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv:1802.05957*, 2018. 5

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

[45] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*, 2018. 2

[46] Anastasia Pentina and Christoph H Lampert. Multi-task learning with labeled and unlabeled tasks. In *ICML*, 2017. 2

[47] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015. 5

[48] Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Geometry meets semantics for semi-supervised monocular depth estimation. *arXiv:1810.04093*, 2018. 3

[49] Pierluigi Zama Ramirez, Alessio Tonioni, Samuele Salti, and Luigi Di Stefano. Learning across tasks and domains. In *ICCV*, 2019. 2

[50] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *NeurIPS*, 2015. 2

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

[52] Ruder Sebastian. An overview of multi-task learning in deep neural networks. *arXiv:1706.05098*, 2017. 2

[53] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *NeurIPS*, 2018. 2

[54] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *ICCV*, 2017. 2

[55] Naoya Takahashi and Yuki Mitsufuji. Densely connected multi-dilated convolutional networks for dense prediction tasks. In *CVPR*, 2021. 2

[56] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 1, 2, 3, 4, 6, 7, 8

[57] Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *arXiv:2004.13379*, 2020. 2

[58] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *ECCV*, 2020. 2

[59] Pavan Kumar Anasosalu Vasu, Shreyas Saxena, and Oncel Tuzel. Instance-level task parameters: A robust multi-task weighting framework. *arXiv:2106.06129*, 2021. 2

[60] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv:1810.08705*, 2018. 5

[61] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediciton-and-distillation network for simultaneous depth estimation and scene parsing. In *CVPR*, 2018. 2

[62] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *CVPR*, 2018. 3

[63] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *NeurIPS*, 2020. 2

[64] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 2, 3

[65] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018. 3, 6

[66] Yu Zhang and Qiang Yang. A survey on multi-task learning. *arXiv:1707.08114*, 2017. 1, 2

[67] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *ECCV*, 2018. 2, 3

[68] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*, 2019. 2

[69] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2, 5

[70] Mingmin Zhen, Jinglu Wang, Lei Zhou, Shiwei Li, Tianwei Shen, Jiaxiang Shang, Tian Fang, and Long Quan. Joint semantic segmentation and boundary detection using iterative pyramid contexts. In *CVPR*, 2020. 2

[71] Ling Zhou, Zhen Cui, Chunyan Xu, Zhenyu Zhang, Chaoqun Wang, Tong Zhang, and Jian Yang. Pattern-structure diffusion for multi-task learning. In *CVPR*, 2020. 2

[72] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *ECCV*, 2018. 2