IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-2010-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IGARSS52108.2023.10282563

CROSSADAPT: CROSS-SCENE ADAPTATION FOR MULTI-DOMAIN DEPTH ESTIMATION

Yu Zhang
University of Kentucky

M. Usman Rafique
Kitware, Inc.

*Gordon Christie**
BlackSky

Nathan Jacobs
Washington University in St. Louis

ABSTRACT

We address the task of monocular depth estimation in the multi-domain setting. Given a large dataset (source) with ground-truth depth maps, and a set of unlabeled datasets (targets), our goal is to create a model that works well on unlabeled target datasets across different scenes. This is a challenging problem when there is a significant domain shift, often resulting in poor performance on the target datasets. We propose to address this task with a unified approach that includes adversarial knowledge distillation and uncertainty-guided self-supervised reconstruction. We provide both quantitative and qualitative evaluations on four datasets: KITTI, Virtual KITTI, UAVid China, and UAVid Germany. These datasets contain widely varying viewpoints, including ground-level and overhead perspectives, which is more challenging than is typically considered in prior work on domain adaptation for single-image depth. Our approach significantly improves upon conventional domain adaptation baselines and does not require additional memory as the number of target sets increases.

1. INTRODUCTION

Conventional deep neural networks often generalize poorly to new domains, and *Domain Adaptation* (DA) methods aim to solve this problem by adapting a model trained on a label-rich source domain to a label-scarce target domain. Recently, most studies on domain adaptation have focused on the single target-domain setting [1], in which only one target domain is considered at a time. However, in many real-world scenarios, test data may be collected from various sources and domains. With the increasing prevalence of unmanned aerial vehicles (UAVs) and unmanned ground vehicles (UGVs), the need to adapt networks across such viewpoint shifts is increasingly important. Therefore, we consider the problem of domain adaptation between videos collected from UAVs and UGVs. For both UAVs and UGVs, monocular depth estimation [2] is an important fundamental task, but obtaining ground-truth depth annotations is difficult and expensive. Therefore, developing a method that can effectively adapt one source-trained model to multiple target datasets is important for both domain adaptation and monocular depth estimation tasks.

To exploit the temporal information stored in the video sequences, we propose an uncertainty-guided self-supervised reconstruction module and apply it to the unlabeled target imagery. This requires both depth estimates and relative pose estimates. Therefore, in addition to depth estimation, our network is also trained to predict the relative pose between two adjacent frames. This reconstruction loss does not require ground-truth depths or camera poses, making it easy to apply to new target domains. To further improve the reliability of the self-supervised reconstruction, we estimate the uncertainty map by computing the average reconstruction error map from four adjacent frames in the video sequence. Pixels with higher uncertainty will be down-weighted in the reconstruction loss during training.

Our contributions include: (1) an adversarial knowledge distillation framework that can bridge the domain gaps between the source and multiple targets without requiring additional memory as the number of targets increases; (2) an uncertainty-guided, self-supervised reconstruction loss that can be easily applied to unlabeled new domains; (3) evaluation on diverse datasets, which include real, synthetic, ground-level, and aerial images, and demonstrate that our model significantly reduces issues due to domain shift.

2. APPROACH

We introduce CrossAdapt (see Fig. 1 for an overview), an approach for training a monocular depth-estimation network in the multi-target domain adaptation (MTDA) setting.

2.1. Problem Statement

We are given a set of fully labeled source-domain samples $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ where $x_i^s \in \mathbb{R}^{H\times W\times 3}$ represents an image in the source domain and $y_i^s \in \mathbb{R}^{H\times W}$ the corresponding ground-truth depth map. In addition, we are given T sets of unlabeled samples $\mathcal{D}_{t,n} = \{x_i^{t,n}\}_{i=1}^{n_t}$, where $x_i^{t,n} \in \mathbb{R}^{H\times W\times 3}$ represents an image from the n-th target domain ($n \leq T$). Our goal is to train a robust monocular depth estimation model that can perform well on all of the target domains. This will require combining supervised training for depth estimation on the source domain and domain adaptation strategies capable of using the unlabeled target data.

*Work done while at the Johns Hopkins University Applied Physics Lab.

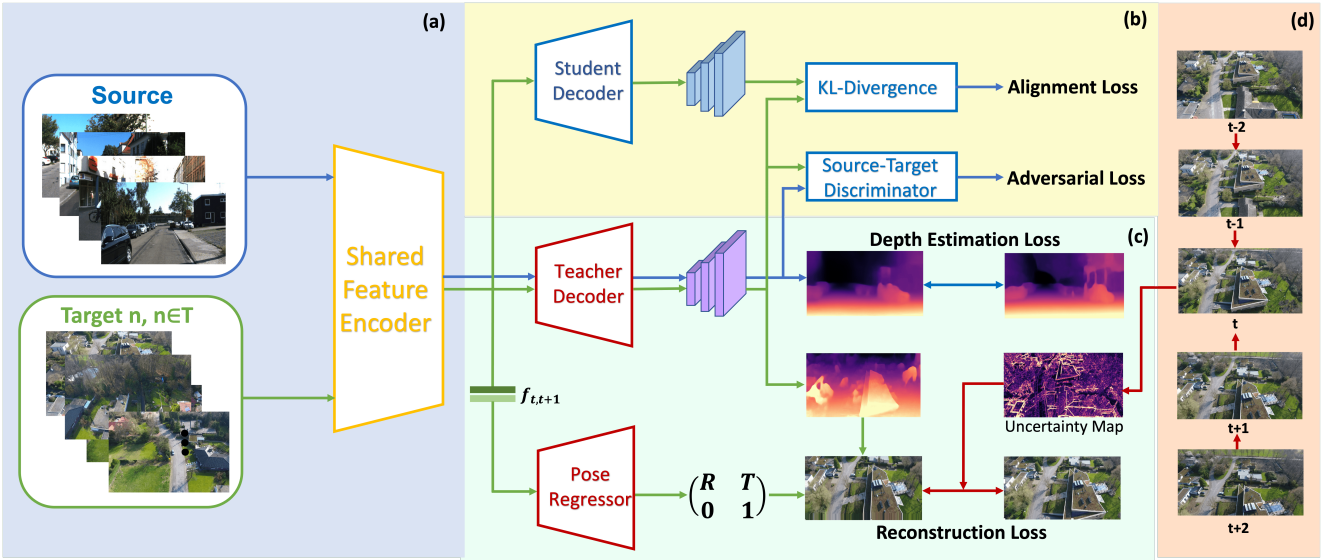

Fig. 1. Overview of CrossAdapt.

2.2. Approach Overview

During training, monocular video sequences from the source and target domains are passed into the shared encoder E (Fig. 1 (a)). The encoded features from both source and target are then passed into a depth estimation teacher decoder D_T, which yields estimated depth maps for both inputs, and the source-target discriminator, which encourages the network to learn domain-invariant features (Fig. 1 (c)). In addition, temporally adjacent target frames are passed into a pose regressor D_P and a relative camera pose is estimated. The estimated pose and depth are combined, along with the known camera intrinsic parameters and the uncertainty map (Fig. 1 (d)), to compute the reconstruction loss. We then minimize the KL divergence between the predictions of the student decoder D_S and the teacher decoder D_T (Fig. 1 (b)). This student decoder D_S is the final model to be used for evaluation. Our model is trained towards four objectives: supervised depth estimation, adversarial loss, alignment loss, and uncertainty-guided reconstruction loss. We describe them in the following sections.

2.3. Supervised Depth Estimation

For source imagery, where ground-truth depth is available, our proposed CrossAdapt framework is trained in a supervised manner using a depth estimation loss. The decoder takes as inputs the features extracted from the image x and predicts the depth map y^s.

Here, we minimize the ℓ_2 distance between the predicted depth $\tilde{y}^s$ and the ground-truth depth y^s:

$$\mathcal{L}_{supervised} = ||y^s - \tilde{y}^s||^2. \quad (1)$$

2.4. Adversarial Knowledge Distillation

One of the key goals of domain adaptation is to encourage the network to learn domain-invariant features. To achieve that goal, we use a source-target discriminator D_D to classify the output feature F_T that comes from the penultimate layer of D_T as either source (1) or target (0) by using binary cross entropy L_{BCE}.

$$\mathcal{L}_{dis} = L_{BCE}(F_T, 1)_{source} + L_{BCE}(F_T, 0)_{target}. \quad (2)$$

Our network has two goals: one is to predict accurate depth maps, and the other one is to fool the discriminator. To achieve the second goal, here we use the response from the discriminator in the subsequent loss and always encourage it to predict source (1) for all inputs. Note that the input feature of the discriminator comes from the penultimate layer of the teacher decoder.

$$\mathcal{L}_{extractor} = L_{BCE}(D_D(F_T), 1). \quad (3)$$

And the total adversarial loss is represented as:

$$\mathcal{L}_{adv} = \mathcal{L}_{dis} + \lambda_{adv}\mathcal{L}_{extractor}. \quad (4)$$

To distill knowledge from the teacher decoder the to student decoder, we use the output of the penultimate layer the teacher decoder F_T, and the student decoder F_S to compute the KL divergence.

$$\mathcal{L}_{align} = \mathcal{L}_{KL}(F_T, F_S). \quad (5)$$

We pass mini-batches of source-target pairs into the model, repeat the process mentioned above, and keep alternating between different targets during training. The student decoder gradually learns from the teacher decoder and tends to be able to represent features from all target sets in the end.

2.5. Uncertainty-Guided Reconstruction

To exploit the temporal information in the input video sequences in the unlabeled target sets. and learn domain-invariant features from various targets, we follow the state-of-the-art self-supervised depth estimation work [2] and reconstruct the appearance of a target image from the viewpoint of an adjacent image by combining predicted depth, pose, and known camera intrinsic parameters. We found that naively applying this will cause inaccurate predictions, especially for fast-changing pixels in the temporal sequences. To overcome this, we propose to estimate an uncertainty map by taking the average of the reconstruction error maps of an input frame and its 4 adjacent frames and using that to further guide the reconstruction loss.

The pose regressor in our model yields the relative pose $T_{t\to t'}$ for each source view image $I_{t'}$, with respect to the target image I_t, from a consecutive monocular video sequence, by taking a pair of features extracted from $(I_t, I_{t'})$ as the inputs. The depth estimation decoder predicts a dense depth map D_t simultaneously. Our goal is to minimize the reconstruction error L_r, where

$$L_r = \sum_{t'} ||I_t - I_{t'\to t}||. \tag{6}$$

The image reconstruction loss, in our case, is the ℓ_1 distance in pixel space. By using the source image $I_{t'}$, the predicted depth D_t, the relative pose $T_{t\to t'}$, and the camera intrinsic parameters K, we can reconstruct the target image I_t by:

$$I_{t'\to t} = I_{t'}\Big\langle proj(D_t, T_{t\to t'}, K)\Big\rangle. \tag{7}$$

where $proj()$ are the resulting 2D coordinates of the projected depths D_t in $I_{t'}$ and $\big\langle\big\rangle$ is the sampling operator.

To reduce noise in the prediction, we use edge-aware smoothness [3, 2]:

$$L_s = |\partial_x d_t^*|\, e^{-|\partial_x I_t|} + |\partial_y d_t^*|\, e^{-|\partial_y I_t|}, \tag{8}$$

where $d_t^* = d_t/\overline{d_t}$ is the mean-normalized inverse depth to discourage shrinking of the estimated depth. The complete self-supervised loss can be represented as:

$$\mathcal{L}'_{recon} = L_r + \lambda_s L_s \tag{9}$$

We propose to estimate uncertainty maps by computing the reconstruction error map generated from I_t and its N (in this case N=4 is used) adjacent frames I_{t+1}, I_{t+2}, I_{t-1}, I_{t-2}. Following Eq 7, the uncertainty map is estimated by taking the average of all the adjacent reconstruction error maps:

$$\mathcal{U}_t = \frac{1}{N}\sum_{i=0}^{N-1} ||I_{t_i\to t} - I_t||. \tag{10}$$

Note that the estimated uncertainty maps highlight drastically changing pixels (see Fig. 2 for illustration), e.g., edges of buildings. Therefore, those pixels with higher uncertainty are down-weighted in the reconstruction loss:

$$\mathcal{L}_{recon} = \frac{\lambda_r \mathcal{L}'_{recon}}{\mathcal{U}_t} + \mathcal{L}'_{recon}. \tag{11}$$

2.6. Overall Loss Function

The overall loss function of the proposed CrossAdapt framework is the weighted sum of the loss functions mentioned above and can be written as follows:

$$\mathcal{L} = \mathcal{L}_{supervised} + \alpha_1 \mathcal{L}_{adv} + \alpha_2 \mathcal{L}_{align} + \alpha_3 \mathcal{L}_{recon}. \tag{12}$$

3. EVALUATION

3.1. Implementation Details

We implement our model using PyTorch. We follow the MTDA protocols and report results on two adaptation scenarios: 1-source→2-target scenario and 1-source→3-target scenario. Following the existing state-of-the-art methods [2, 4], we use a similar U-Net style architecture and adopt the ResNet-18 as the feature extraction backbone to ensure a fair comparison. All networks are pre-trained on ImageNet. We follow the same training protocol used in Monodepth2 [2], with a learning rate of 10^{-4} for the first 15 epochs which is then dropped to 10^{-5} for the rest the training process. For hyperparameters, the adversarial term λ_{adv} is set to 1.0, the smoothness term λ_s is set to 0.001, and the reconstruction term λ_r is set to 0.01. For the overall loss function, we set α_1 and α_2 to 0.1 and α_3 0.01 to maintain a balance between each term during training. We evaluate on four diverse datasets: KITTI [5], Virtual KITTI [6], UAVid China [7], and UAVid Germany [7]. For data pre-processing, we resize all input images to 640×192.

3.2. Experimental Results

We compare our method with three state-of-the-art baselines, including a self-supervised depth estimation method Monodepth2 [2], a domain adaption method CoMoDA [4], and a multi-domain segmentation model MTKT [8]. We summarize the comparisons in Table 1 and 2. We report four metrics: ℓ_1 (p), ℓ_1 (n), S (p), and S (n). ℓ_1 (p)/(n) represents the ℓ_1 error between the target image and the reconstructed target image from the previous(p)/next(n) frame. S (p) and S (n) represent the similarity metric SSIM, which are reported as the SSIM loss (1-SSIM) in the tables. We use KITTI as the source for all the adaptation scenarios since it contains the most complete depth annotations. For the 1-source→2-targets scenario, we use both UAVid China and UAVid Germany as targets to evaluate the model's ability to handle extreme viewpoint changes. Both quantitative results (Table 1) and qualitative results (Fig. 2) demonstrate the effectiveness of our model.

Fig. 2. Illustration of the uncertainty maps. The 1st row shows the input images, the 2nd row shows the predicted depth maps, and the last row shows the estimated uncertainty maps, which mostly highlight rapidly-changing pixel regions including vehicles and building edges.

For the 1-source→3-targets scenario, we use UAVid China, UAVid Germany, and Virtual KITTI as targets. The results are listed in Table 2. We also conduct an ablation study for the first scenario (1→2), listed in Table 1. CrossAdapt (w/o r) shows the performance of our model without using the reconstruction loss, and CrossAdapt (w/o u) shows the results without using the uncertainty guidance. The experiments show that the self-supervised reconstruction loss significantly improved the performance and the uncertainty guidance slightly boosted the performance when it was applied together with the reconstruction loss.

Table 1. KITTI→UAVid China + UAVid Germany

Target	Method	ℓ_1 (p)	ℓ_1 (n)	S (p)	S (n)
China	Monodepth2 [2]	0.1230	0.1261	0.3181	0.3226
	CoMoDA [4]	0.1193	0.1042	0.2901	0.3009
	MTKT [8]	0.0812	0.0833	0.2216	0.2305
	CrossAdapt (w/o r)	0.0910	0.0907	0.2270	0.2299
	CrossAdapt (w/o u)	0.0629	0.0651	0.1876	0.1841
	CrossAdapt (Ours)	**0.0620**	**0.0513**	**0.1702**	**0.1788**
Germany	Monodepth2 [2]	0.1861	0.1873	0.3909	0.3981
	CoMoDA [4]	0.1741	0.1725	0.3676	0.3755
	MTKT [8]	0.1785	0.1680	0.3601	0.3761
	CrossAdapt (w/o r)	0.1801	0.1795	0.3644	0.3606
	CrossAdapt (w/o u)	0.1581	**0.1526**	0.3511	0.3537
	CrossAdapt (Ours)	**0.1468**	0.1541	**0.3488**	**0.3412**

Table 2. KITTI→UAVid China + UAVid Germany + Virtual KITTI

Target	Method	ℓ_1 (p)	ℓ_1 (n)	S (p)	S (n)
China	Monodepth2 [2]	0.1487	0.1401	0.3590	0.3574
	CoMoDA [4]	0.1386	0.1344	0.3067	0.3156
	MTKT [8]	0.1345	0.1509	0.2687	0.2459
	CrossAdapt (Ours)	**0.0918**	**0.0927**	**0.2141**	**0.2108**
Germany	Monodepth2 [2]	0.1762	0.1705	0.3921	0.3700
	CoMoDA [4]	0.1676	**0.1609**	0.3822	0.3850
	MTKT [8]	0.1887	0.1654	0.3709	0.3885
	CrossAdapt (Ours)	**0.1531**	0.1676	**0.3596**	**0.3677**
V KITTI	Monodepth2 [2]	0.1648	0.1732	0.3390	0.3371
	CoMoDA [4]	0.1731	0.1704	0.3219	0.3232
	MTKT [8]	0.1666	0.1796	0.3395	0.3368
	CrossAdapt (Ours)	**0.1634**	**0.1681**	**0.3183**	**0.3210**

4. CONCLUSION

We introduced a novel multi-target depth estimation framework. Our approach makes it possible to train a unified model for multiple scenarios simultaneously.

5. REFERENCES

[1] Adrian Lopez-Rodriguez and Krystian Mikolajczyk, "DESC: Domain adaptation for depth estimation via semantic consistency," *BMVC*, 2020.

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

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

[4] Yevhen Kuznietsov, Marc Proesmans, and Luc Van Gool, "CoMoDA: Continuous monocular depth adaptation using past experiences," in *WACV*, 2021.

[5] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets robotics: The kitti dataset," *IJRR*, 2013.

[6] Yohann Cabon, Naila Murray, and Martin Humenberger, "Virtual KITTI 2," *arXiv:2001.10773*, 2020.

[7] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang, "UAVid: A semantic segmentation dataset for UAV imagery," *ISPRS*, 2020.

[8] Antoine Saporta, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez, "Multi-target adversarial frameworks for domain adaptation in semantic segmentation," in *ICCV*, 2021.