

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

Supplementary material

In this supplementary, we present additional implementation details about the joint discriminator employed in *UM-Adapt-B* followed by architectural details of the *task-transfer* networks employed in the proposed Cross-task Distillation module. We also present qualitative results on the outdoor KITTI and Cityscapes dataset for the experimental setup mentioned in the main paper.

## UM-Adapt-B

Existing literature shows efficacy of simultaneous adaptation at hierarchical feature levels, while minimizing domain discrepancy for multi-layer deep architectures. Long *et al.* [5] suggested to apply moment matching at multiple levels of feature hierarchy to bridge both marginal and conditional distribution for improved domain adaptation performance. Recently Kundu *et al.* [4] proposed to use two different discriminators; one at the latent representation and the other at the final output prediction to yield efficient adaptation for spatially-structured depth prediction task. Motivated by this, we design a 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 1, the predicted joint distribution denoted as  $P(M_t(X_t), \hat{Y}_t^{depth}, \hat{Y}_t^{normal}, \hat{Y}_t^{seg})$  is matched with actual true distribution denoted as  $P(M_s(X_s), Y_s^{depth}, Y_s^{normal}, Y_s^{seg})$ . To realize this, we employ a discriminator with initial task specific convolutional layers after late-fusion

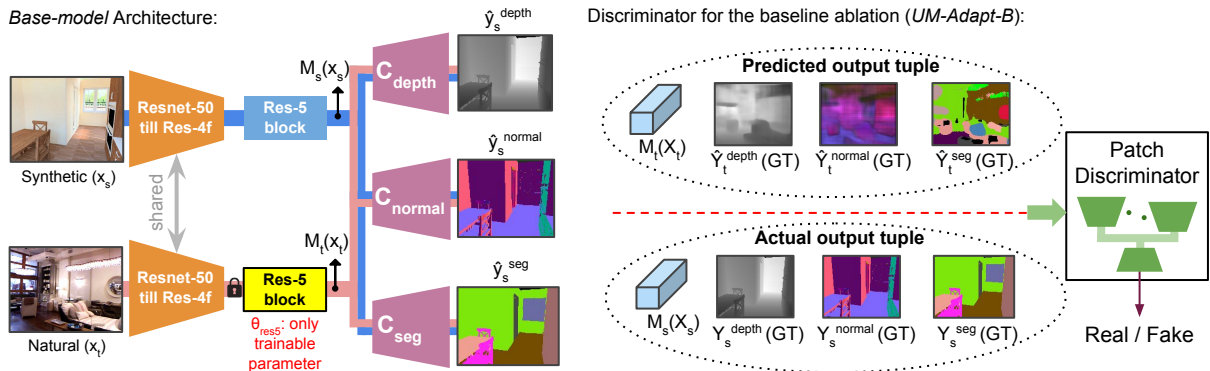


Figure 1: An overview of the proposed *UM-Adapt* architecture for multi-task adaptation. The blue and orange background line indicates data flow for synthetic and natural image respectively. Input set of the joint discriminator used in the ablation *UM-Adapt-B* is shown in dotted ellipses on right.

Table 1: Network architecture of the discriminator used in *UM-Adapt-B* baseline variant. Conv\* denotes standard convolutional layers followed by a batch-normalization with leaky-ReLU non-linearity. Here,  $jj$  denotes channel-wise concatenation.

Layer	input	Type	Filter	Stride	Output Size
$C1_d$	depth-map	Conv	3 3, 64	1	128 160 64
$C2_d$	$C1_d$	Conv*	3 3, 128	2	64 80 128
$C1_n$	normal-map	Conv	3 3, 64	1	128 160 64
$C2_n$	$C1_n$	Conv*	3 3, 128	2	64 80 128
$C3_{n+d}$	$C2_njjC2_d$	Conv	3 3, 256	2	32 40 256
$C1_m$	$M_s(x_s)$ or $M_t(x_t)$	Deconv	1 1, 512	1	8 10 512
$C2_m$	$C1_m$	Deconv*	3 3, 512	1	16 20 512
$C3_m$	$C1_m$	Deconv*	3 3, 256	2	32 40 256
$C1_s$	seg-map	Conv	3 3, 64	1	128 160 64
$C2_s$	$C1_s$	Conv*	3 3, 128	2	64 80 128
$C3_s$	$C2_s$	Conv*	3 3, 256	2	32 40 256
$C1_c$	$C3_sjjC3_{n+d}jjC3_m$	Conv	1 1, 512	1	32 40 512
$C2_c$	$C1_c$	Conv*	3 3, 1024	2	16 20 1024
$C3_c$	$C2_c$	Conv*	3 3, 512	2	8 10 512
$C4_c$	$C3_c$	Conv	1 1, 1	1	8 10 1

concatenation with the latent feature map followed by some deconvolutional layer to form an hour-glass like patch discriminator [3]. The discriminator architecture for *UM-Adapt-B* is presented in Table 1. Here the discriminator is trained following the improved techniques of Wasserstein GAN as proposed by Gulrajani *et al.* [2].

## Architecture of Task-transfer networks

As mentioned in the main paper, for NUYD we take a combined input of depth-map and normal-map for the baseline variant *UM-Adapt-(noAdv.)-1* (See Table 2). Whereas in *UM-Adapt-(noAdv.)* and *UM-Adapt-(Adv.)*, we use an additional task-transfer network which predicts depth-map output with segmentation map as the input representation. Note that, all the task-transfer networks are independently trained on synthetic depth, normal and segmentation tuples and kept frozen throughout the adaptation process for both *UM-Adapt-(noAdv.)-1* and *UM-Adapt-(noAdv.)*. Whereas for *UM-Adapt-(Adv.)*, we also update the parameters of the task-transfer networks as shown in Algorithm 2 (main paper). A similar task-transfer architecture is employed for GTA5 to KITTI+Cityscapes adaptation, where the input resolution is taken as 128 256.

Table 2: Architectural details of Task-transfer networks. Conv\* denotes standard convolutional layers followed by a batch-normalization with leaky-ReLU non-linearity. Here,  $jj$  denotes channel-wise concatenation.

Layer	input	Type	Filter	Stride	Output Size
$C1_d$	depth-map	Conv	3 3, 64	1	128 160 64
$C2_d$	$C1_d$	Conv*	3 3, 128	2	64 80 128
$C1_n$	normal-map	Conv	3 3, 64	1	128 160 64
$C2_n$	$C1_n$	Conv*	3 3, 128	2	64 80 128
$C3_m$	$C2_n/jjC2_d$	Conv	3 3, 256	2	32 40 256
$C3_m$	$C2_m$	Conv*	3 3, 512	2	16 20 512
$C4_m$	$C3_m$	Conv*	3 3, 512	1	16 20 512
$EN_{depth}(y^{depth})$	$C4_m$	Conv	3 3, 1024	2	8 10 1024
$C5_m$	$EN_{depth}(y^{depth})$	Deconv*	3 3, 512	2	16 20 512
$C6_m$	$C5_m$	Deconv*	3 3, 512	2	32 40 512
$C7_m$	$C6_m$	Deconv*	3 3, 256	2	64 80 256
$C8_m$	$C7_m$	Deconv*	3 3, 128	2	128 160 128
$C9_{seg}$	$C8_m$	Conv(softmax)	1 1, 40	1	128 160 40

## Qualitative Results on KITTI and Cityscapes

We provide qualitative results on KITTI and Cityscapes dataset for both depth estimation and semantic segmentation tasks (See Figure 2). Note that in the semi-supervised setting we have only depth-map ground-truth for KITTI and only segmentation map ground-truth for Cityscapes dataset. We treat the mixed image samples from KITTI and Cityscapes together as the target domain input ignoring the cross-city [1] input discrepancy.

## Other Implementation Details

In contrast to depth-map or normal-map, the segmentation map ground-truth has a specific structure as it is represented as a one-hot vector for each pixel-location. Discriminating such one-hot real segmentation maps (synthetic data) from the output of *UM-Adapt* segmentation after softmax non-linearity becomes an easier task for the *task-transfer* networks or the discriminator, taking these representations as input at different iterations. To alleviate this, we create a random *pseudo-softmax* representation of the ground-truth one-hot maps by keeping higher probability for the true-class (sampled from a truncated Normal pdf with mean 1 and standard-deviation 0.075) in the randomly generated vector satisfying sum one and values in the range of zero to one criteria. We provide 50% of direct one-hot and 50% of such *pseudo-softmax* (at each batch iteration) for the true segmentation map distribution. Besides this, to address the problem of class-balancing, we use a randomly generated gradient mask layer (50% of the batch without masking) in-between the UM-Adapt prediction and the corresponding Task-transfer or the discriminator network.

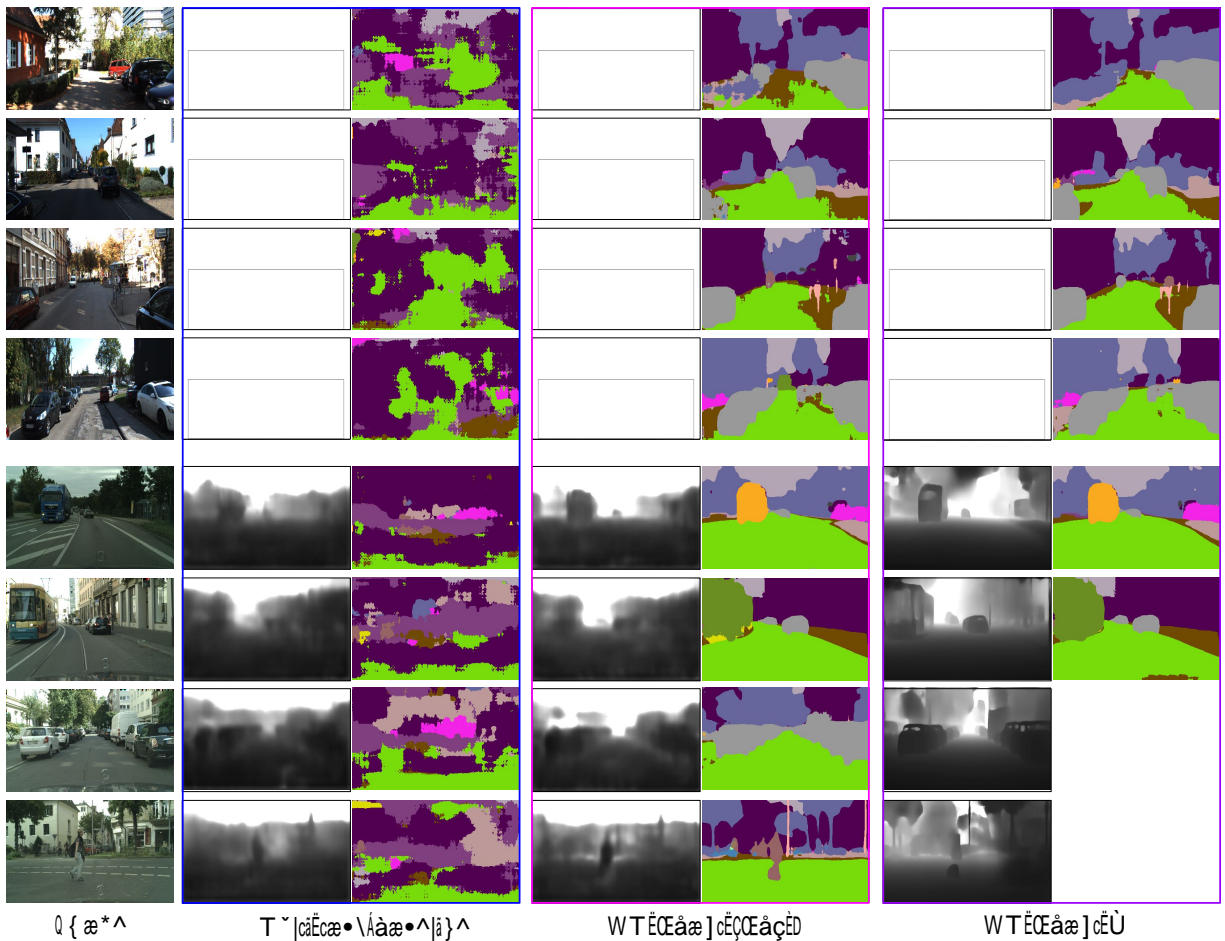


Figure 2: Qualitative comparison of different ablations of *UM-Adapt*. First 4 rows contains images from KITTI test set (Eigen-spilt), whereas the last 4 rows contains images from Cityscapes validation-set. We adapt a single multi-task model on combined training set from both the datasets.

## References

- [1] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *ICCV*, 2017.
- [2] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NeurIPS*, pages 5767–5777, 2017.
- [3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.
- [4] Jogendra Nath Kundu, Phani Krishna Uppala, Anuj Pahuja, and R Venkatesh Babu. Adadepth: Unsupervised content congruent adaptation for depth estimation. In *CVPR*, 2018.
- [5] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *NeurIPS*, 2016.