Student Becoming the Master: Knowledge Amalgamation for Joint Scene Parsing, Depth Estimation, and More
– *Supplementary Material* –

Jingwen Ye[1], Yixin Ji[1], Xinchao Wang[2], Kairi Ou[3], Dapeng Tao[4], Mingli Song[1]
[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[2]Department of Computer Science, Stevens Institute of Technology, New Jersey, United States
[3]Alibaba Group, Hangzhou, China
[4]School of Information Science and Engineering, Yunnan University, Kunming, China
{yejingwen, jiyixin, brooksong}@zju.edu.cn, xinchao.w@gmail.com,
suzhe.okr@taobao.com, dptao@ynu.edu.cn

In this document, we provide more implementation details of the proposed approach and additional results, including the results obtained by training TagetNet-3 offline, the results on two more datasets, KITTI [3] and Taskonomy [7], and more visualizations. Also, to show the proposed method's generalization ability, we apply our approach to Deeplab v3 with ResNet50.

Please note that, though not explicitly stated, the results of TargetNet-3 presented in the main manuscript are obtained by training TargetNet-3 *online*. The details of the online and offline training of TargetNet-3 are discussed in Sec. 5.2 of the main manuscript.

1. Implementation Details

We provide here more details on our parameter settings and branch-out strategy.

1.1. Parameter Settings

Channel coding. As discussed in Sec. 5.1.1 of the main manuscript, the channel coding consists of a global pooling layer and two fully connected layers. Let c denote the number of channels of the feature maps in block n and thus also the number of channels fed to the channel coding blocks. Within each channel coding block, the first fully connected layer reduces the channel number to c/r, and then the second fully connected layer reverts the number to c.

For the *S-Channel Coding* and the *D-Channel Coding* as depicted in Fig. 2 in the manuscript, we set $r = 8$, while for the *N-Channel Coding* and the *U-Channel Coding* of TargetNet-2 as in Fig. 4 of the manuscript, we set $r = 4$.

Weights of the losses. We use the loss function

$$\mathcal{L}_{\text{u}} = \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}(D, \hat{D}) + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}(S, \hat{S}), +\lambda_{\text{norm}}\mathcal{L}_{\text{norm}}(M, \hat{M}) \tag{1}$$

for offline training, and use

$$\mathcal{L}_{\text{u}} = \lambda_{\text{u}}\mathcal{L}_{\text{u2}} + \lambda_{\text{norm}}\mathcal{L}_{\text{norm}}(M, \hat{M}) \tag{2}$$

for online training, with λ_{norm} and $\mathcal{L}_{\text{norm}}$ removed when training TargetNet-2. The settings for the balancing weights λ_{depth}, λ_{seg}, λ_{norm} and λ_{u} are shown in Tab. 1 below.

1.2. Details for Branch Out

As described in Sec. 5.1.3 of the main manuscript, we choose the branch-out block based on the final loss. In our implementation, for each task we collect the loss values of the final 100 iterations at each block of the decoder, and then compute the average loss values, upon which we choose the branch-out blocks.

Table 1. Balancing weights for the NYUDv2, Cityscape and KITTI dataset.

Balancing Weights	NYUDv2			Cityscape	KITTI
	TargetNet-2	TargetNet-3 (offline)	TargetNet-3 (online)		
λ_{depth}	1	$\lambda_u = 1$	1	1	1
λ_{seg}	1.5		1.5	2	2
λ_{norm}	–	3	3	–	–

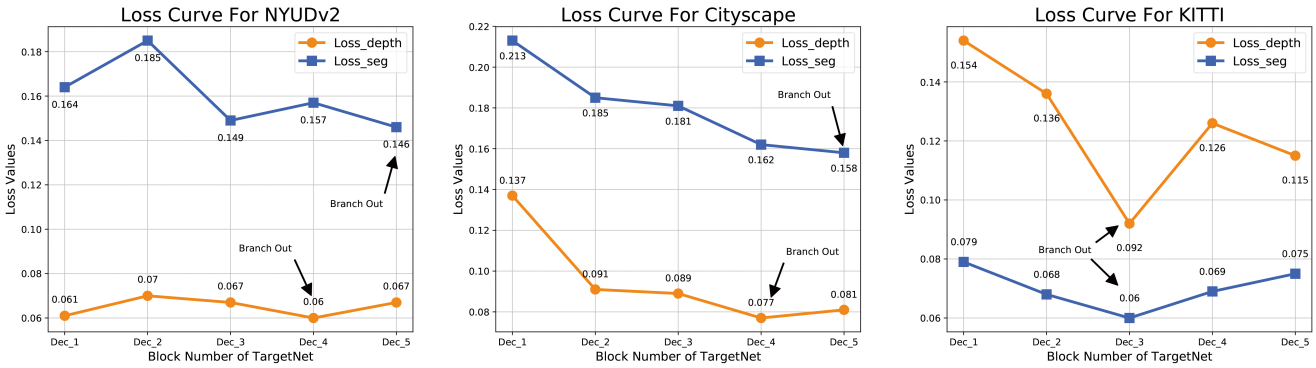

Figure 1. The loss values for scene parsing and depth estimation at each block of TargetNet's decoder.

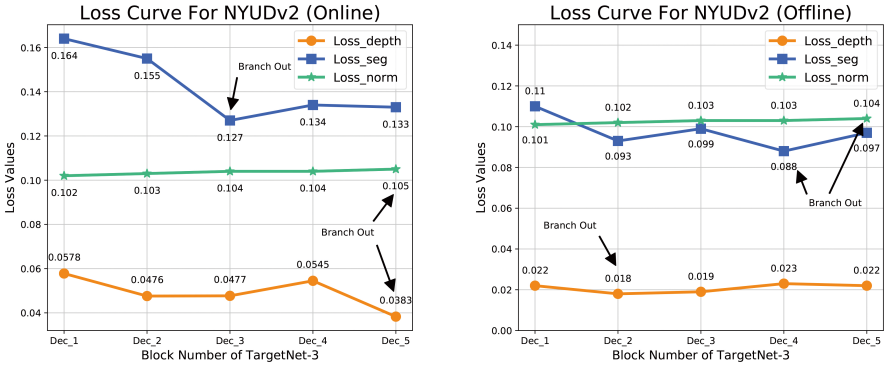

Figure 2. The loss values for scene parsing, depth estimation, and surface normal prediction at each decoder block of TargetNet-3, trained online (left) and offline (right).

We show in Fig. 1 the average loss values for different decoder blocks and for the two tasks, scene parsing and depth estimation, where we highlight the blocks for branching out. We show in Fig. 2 the average loss values for both the online and offline amalgamation of TargetNet-3 that handles three tasks, scene parsing, depth estimation, and surface normal prediction. Since the loss values for surface normal are negative, i.e., $\mathcal{L}_{\text{norm}} < 0$, we show their absolute values.

2. Additional Results

In what follows, we provide the results of TargetNet-3 trained offline, the results on the KITTI dataset, and additional visual results on the NYUDv2 dataset. The results of TargetNet-3 reported in the main manuscript are obtained by online training. And also, we utilize another framework, ResNet on the NYUDv2 dataset for joint scening parsing and depth estimation.

2.1. TargetNet-3 with Offline Training

Tab. 2 shows the quantitative results of TargetNet-3 that branches out at different decoding blocks when trained offline, as well as the results of the teacher networks and TargetNet-2. TargetNet-3 here is trained by treating NormNet and TargetNet-2 as the teachers. As can be seen from Tab. 2, TargetNet-3 yields results better than those of TargetNet-2 on both scene parsing and depth estimation, and better than the ones of NormNet on surface normal prediction.

Table 2. Comparative results of the teacher networks, TargetNet-2, and TargetNet-3 trained offline on NYUDv2. The best results are marked in bold.

Method		Params	Parsing		Depth Estimation		Surface Normal	
			mean IoU	Pixel Acc.	abs rel	sqr rel	Mean Angle	Median Angle
Teacher for TargetNet-2	SegNet	∼83.4M	0.448	0.684	–	–	–	–
	DepthNet		–	–	0.339	0.287	–	–
Teacher for TargetNet-3	NormNet		–	–	–	–	37.88	36.96
	TargetNet-2	∼27.8M	**0.458**	0.687	0.256	0.266	–	–
TargetNet-3	Decoder_b1	∼46.1M	0.452	0.680	0.258	0.269	37.4	32.7
	Decoder_b2	∼34.2M	0.455	0.685	0.254	0.258	36.9	32.0
	Decoder_b3	∼28.8M	0.457	0.687	**0.253**	**0.254**	36.3	31.4
	Decoder_b4	∼28.1M	**0.458**	**0.688**	0.258	0.271	36.1	31.0
	Decoder_b5	∼27.9M	0.457	0.683	0.261	0.278	**35.5**	**30.3**

2.2. Results on KITTI

The KITTI dataset comprises 200 semantically annotated images, of which 100 are used to train SegNet. For fair comparisons on the segmentation performances, we follow the work of [6] and keep only 6 well-represented classes; for comparisons on depth estimation, we follow [1] and employ 20,000 images with depth annotations for training and 697 images for testing. As KITTI features similar scenarios as Cityscape does, we adopt the SegNet trained on Cityscape and finetune it on KITTI.

We show the comparative results of the teachers and the student network, TargetNet, with different branch-out blocks in Tab. 3. TargetNet again consistently outperforms the teachers on all the evaluation metrics. Also, we compare in Tab. 4 the performances of TargetNet with several recent high-ranking methods on KITTI, where TargetNet yields results superior to the advanced scene parsing and depth estimation models, confirming the validity of the proposed knowledge amalgamation approach. In Figs. 3 and 4, we provide some visualizations comparing the teacher models and TargetNet, for which the results are denoted by Target-P and Target-D respectively for scene parsing and depth estimation. From the figures we can see that the results of TargetNet are more visually plausible than those of the teachers.

Table 3. Comparative results of the teacher networks and the student (TargetNet) with different branch-out blocks on the KITTI dataset. Decoder_bn denotes the TargetNet that branches out at the block n of the decoder.

Method	mean IOU	Pixel Acc.	abs rel	sqr rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
SegNet	0.842	0.942	–	–	–	–	–
DepthNet	–	–	0.156	1.18	0.784	0.936	0.978
Decoder_b1	0.824	0.936	0.159	1.20	0.777	0.938	0.980
Decoder_b2	0.843	0.943	0.161	1.21	**0.782**	0.939	0.980
Decoder_b3	**0.852**	**0.947**	**0.155**	**1.16**	0.785	**0.941**	**0.981**
Decoder_b4	0.840	0.942	0.163	1.24	0.769	0.934	0.980
Decoder_b5	0.848	0.946	**0.155**	1.17	0.778	0.939	**0.981**

Table 4. Comparative results of TargetNet and recent high-ranking methods on scene parsing and depth estimation on KITTI. The TargetNet here is branched out at different blocks for depth estimation and scene parsing.

Method	mean IOU	RMSE (lin),m	RMSE (log)
Zuo et al. [8]	0.849	–	–
Ren et al. [5]	0.719	–	–
Wang et al. [6]	0.748	–	–
Eigen et al. [1]	–	7.156	0.270
Garg et al. [2]	–	5.104	0.273
Goddard et al. [4]	–	4.471	0.232
TargetNet	**0.852**	**4.468**	**0.198**

2.3. Results on Taskonomy

The Taskonomy dataset includes over 4.5 million images from over 500 buildings. Each image has annotations for every one of the 2D, 3D, and semantic tasks in Taskonomy's dictionary. Here we only a small fraction of the dataset provided by the author. And we choose three tasks from the dictionary for experiments, which are scene parsing, depth estimation and surface normal prediction.

Table 5. Comparative results of the teachers (SegNet, DepthNet and NormNet) and TargetNet-3 on Taskonomy dataset.

Method	**mIOU**	**abs rel**	**sqr rel**	**Mean Angle**	**Median Angle**
SegNet	0.574	-	-	-	-
DepthNet	-	0.220	0.238	-	-
NormNet	-	-	-	25.3	24.1
TargetNet-3	0.587	0.193	0.202	24.8	23.5

In Tab. 5, we show the comparative results of the teachers (SegNet, DepthNet and NormNet) and the student network, TargetNet-3 trained in the offline manner. The learned TargetNet-3 outperforms all the teachers in their specified tasks, which prove the effectiveness of our methods further.

2.4. Additional Results on NYUDv2

2.4.1 Additional Results on NYUDv2 in ResNet architecture

To prove the effectiveness and generalization ability of the proposed method, we change the basic architecture of SegNet to the architecture of Deeplab v3, which utilizes ResNet-50 framework. The comparative results are shown in Tab. 6, in which, Target-P and Target-D both get better accuracies than the teachers.

Table 6. Comparative results of TargetNet (Target-P and Target-D) and Teachers (SegNet and DepthNet) with ResNet50 on NYUDv2. Target-P is the TargetNet branching out for the best parsing performance and Target-D is for best depth estimation performance.

Method	**mean IOU**	**abs rel**	**sqr rel**
SegNet	0.431	-	-
DepthNet	-	0.136	0.164
Target-P	**0.447**	0.109	0.122
Target-D	0.439	**0.108**	**0.115**

2.4.2 Additional Visual Results on NYUDv2

We also provide some additional visual results in Fig. 5, comparing the scene parsing, depth estimation, and normal perdition results of TargetNet-3 with those of the teachers on the NYUDv2 dataset. The ones obtained by TargetNet-3 are visually closer to the ground truths.

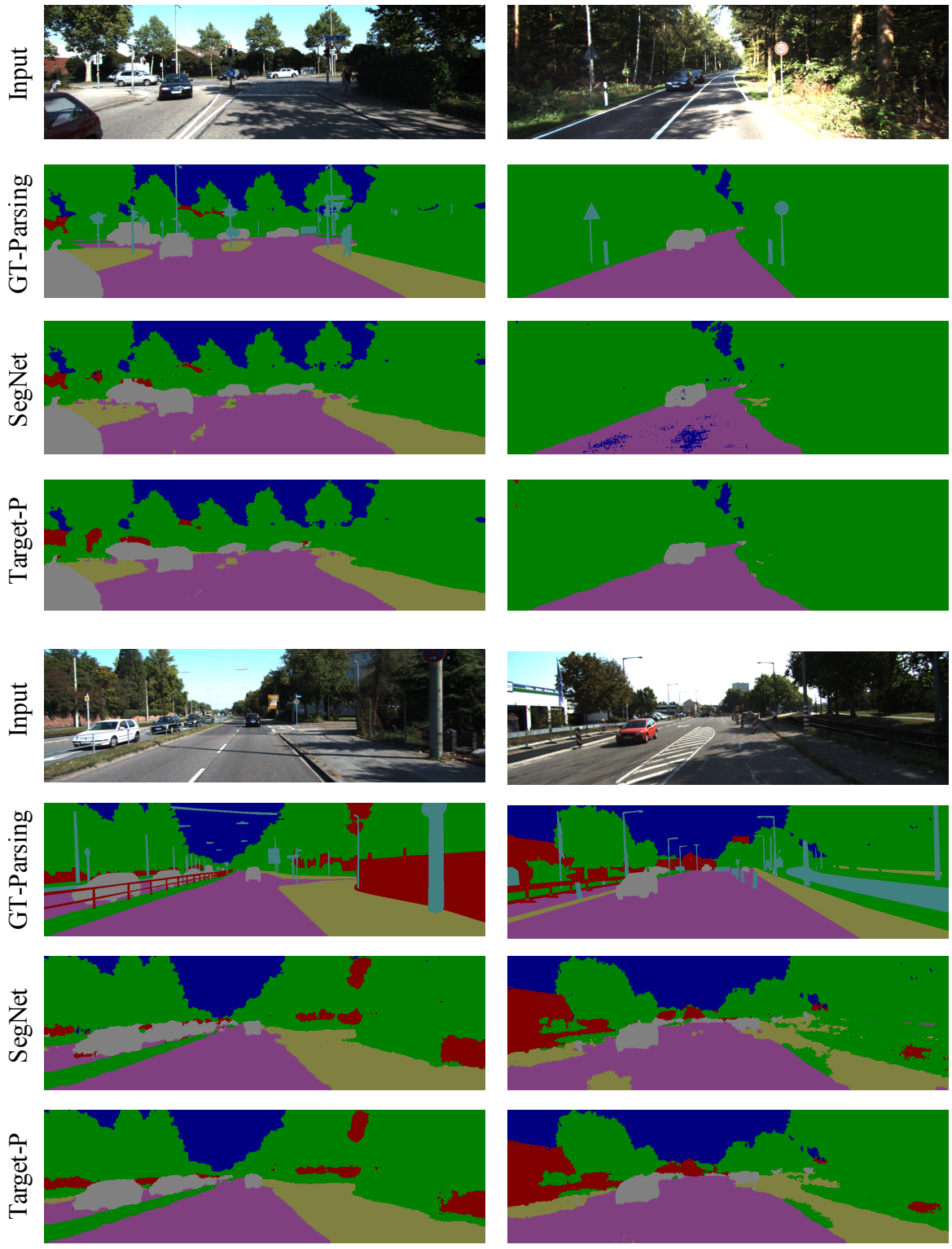

Figure 3. Qualitative results on outdoor scene parsing on KITTI. We compare the results of the teacher (SegNet) with those of the student (Target-P).

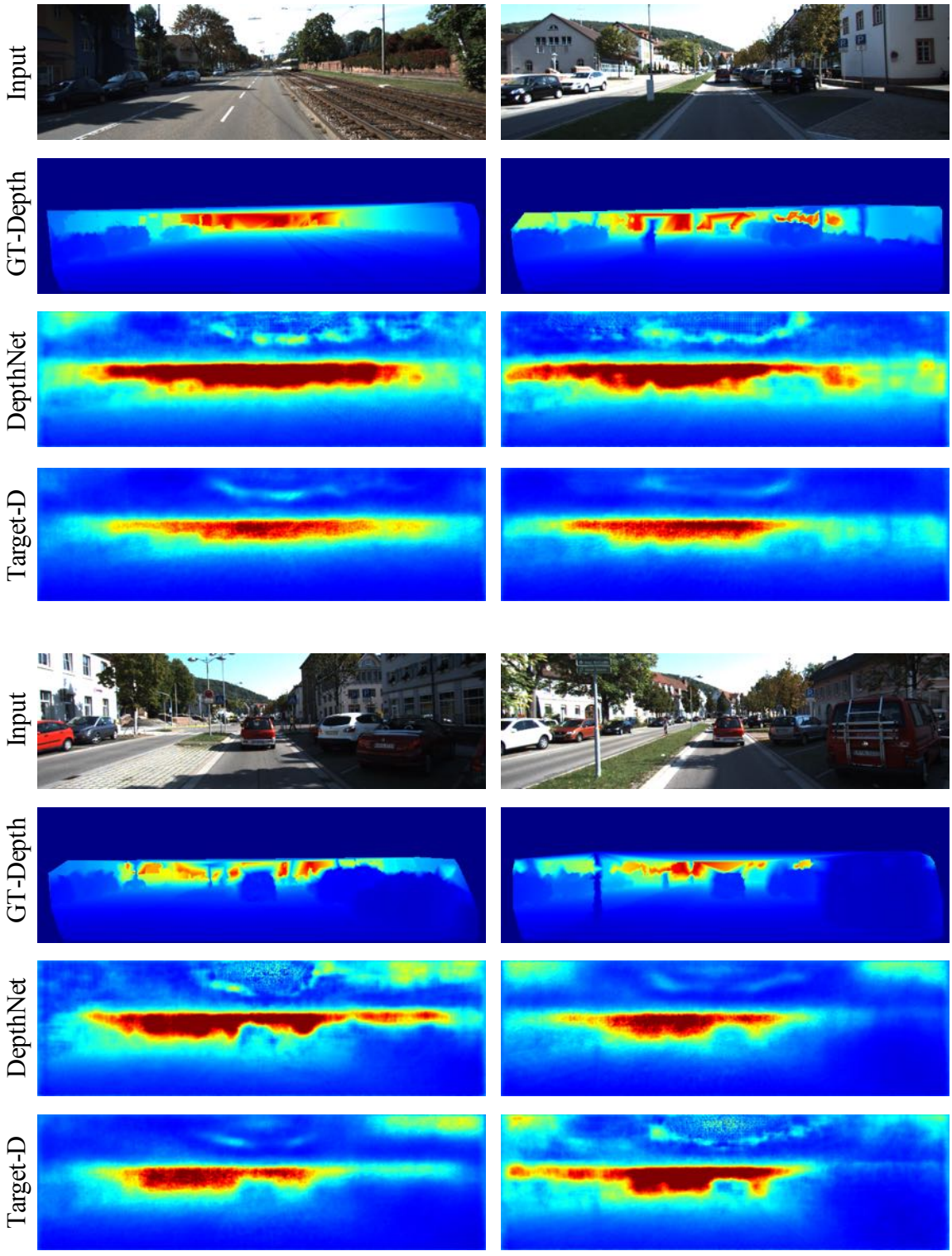

Figure 4. Qualitative results on depth estimation on KITTI. We compare the results of the teacher (DepthNet) with those of the student (Target-D).

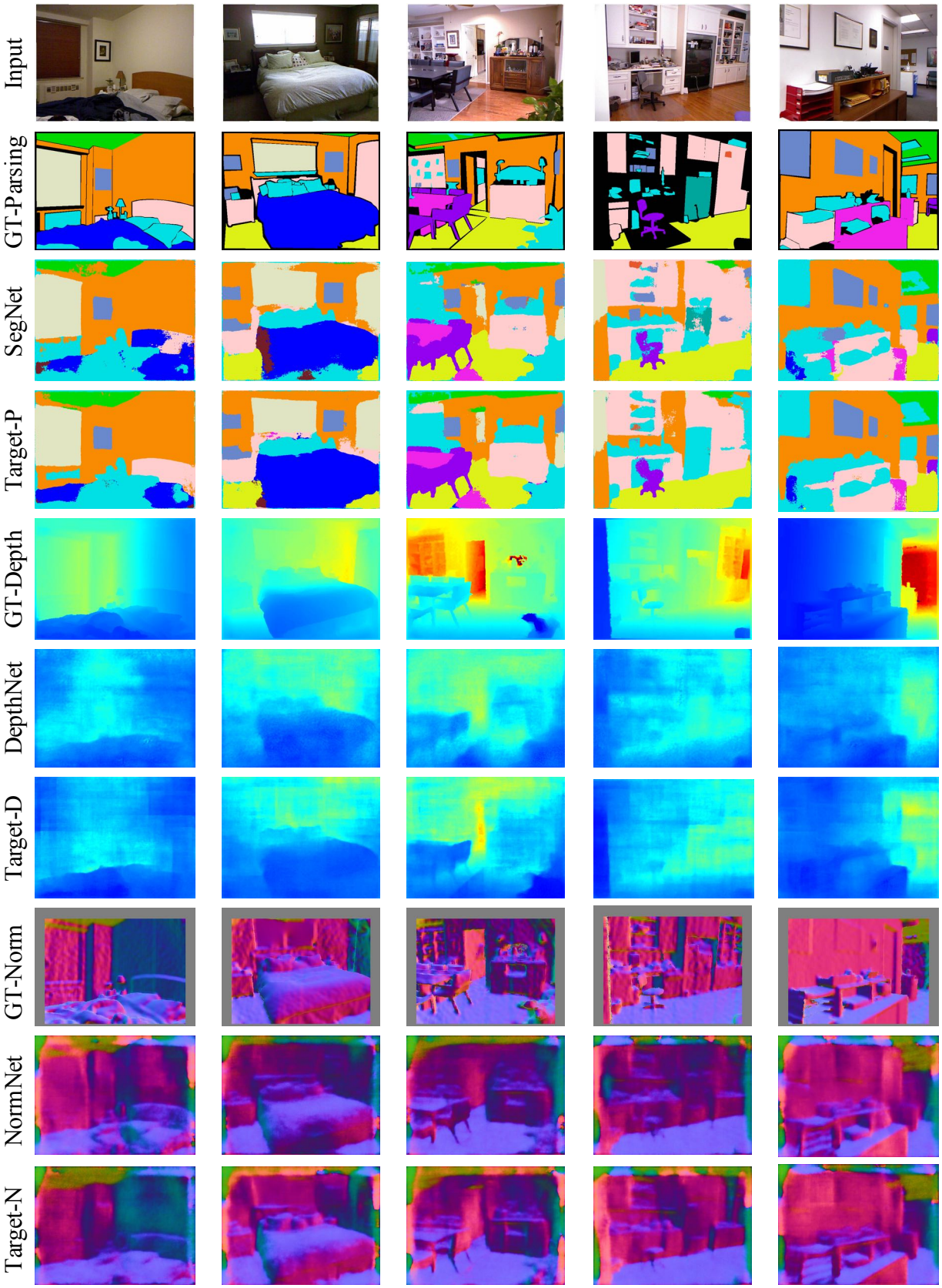

Figure 5. Qualitative results on depth estimation on NYUDv2. We compare the results of the teachers (SegNet, DepthNet, and NormNet) with those of the student (Target-P, Target-D, and Target-N).

References

[1] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Neural Information Processing Systems*, pages 2366–2374, 2014. 3

[2] Ravi Garg, B G Vijay Kumar, Gustavo Carneiro, and Ian D Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. *European Conference on Computer Vision*, pages 740–756, 2016. 3

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 1

[4] Clement Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. *Computer Vision and Pattern Recognition*, pages 6602–6611, 2017. 3

[5] Xiaofeng Ren, Liefeng Bo, and Dieter Fox. Rgb-(d) scene labeling: Features and algorithms. *Computer Vision and Pattern Recognition*, pages 2759–2766, 2012. 3

[6] Shenlong Wang, Sanja Fidler, and Raquel Urtasun. Holistic 3d scene understanding from a single geo-tagged image. *Computer Vision and Pattern Recognition*, pages 3964–3972, 2015. 3

[7] Amir R. Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Computer Vision and Pattern Recognition*, 2018. 1

[8] Yan Zuo and Tom Drummond. Fast residual forests: Rapid ensemble learning for semantic segmentation. *Annual Conference on Robot Learning*, pages 27–36, 2017. 3