Collaboration by Competition: Self-coordinated Knowledge Amalgamation for Multi-talent Student Learning
– *Supplementary Material* –

In this document, we provide the additional details and results, which we cannot fit into the manuscript due to the page limit. In the first section, we demonstrate the architecture of the student network. In the second section, we provide details about the quantitative task transferability, which are utilized to decide if and which auxiliary pre-trained model (PTM) shall be introduced. In the third section, we show the results of student network with more target tasks, and comparison with the teahcer PTMs as well as knowledge distillation [2]. Besides, we also provide the qualitative performance of the multi-talent student. In the last section, we compare our multi-talent student model with multiple distilled students each specialized in a distinct task.

1 Model Architecture

The architecture of the student model utilized in the experiment is adapted from the taskonomy [3] teacher models. We used Tensorflow for our implementation. As for the shared encoder of student network, a modified ResNet-50 encoder with no average-pooling and replace the last stride 2 convolution with stride 1. This gives us an output shape of $16 \times 16 \times 2048$; we use a 3×3 convolution to transform the output to our final shared representation, which is of shape $16 \times 16 \times 8$. The shared representation is then fed to the multiple decoders for task specific inference.

In fact, the shape of the decoder depends on the task. For pixels-to-pixels prediction utilized in this paper, we use the 15-layer fully convolutional decoders, which consist of 5 convolutional layers, and followed by alternating convolution and convolution transpose layers. Besides, batchnorm is used for all layers (except output).

As done in [3], we also used a CGAN in addition to the objective distillation loss function for the pixels-to-pixels prediction tasks. The GAN training began after 25000 steps, and the discriminator was 5 layers with 10. the standard weight decay. The loss was a linear combination of the $0.996\times$ the soft target distillation loss and $0.004\times$ the GAN loss.

2 Quantitative Task Similarity

As aforementioned in the main text, we decide whether auxiliary PTM shall be introduced and which PTM is beneficial for multi-modal incorporated amalga-

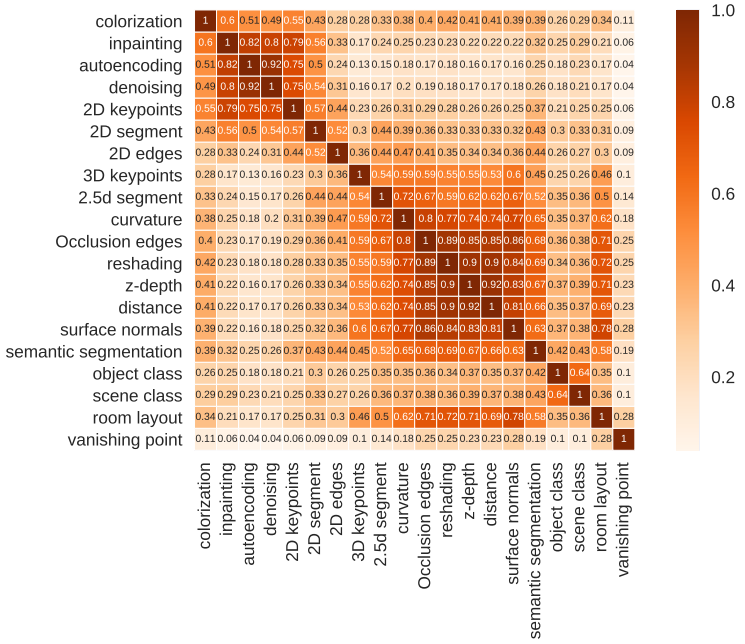

Fig. 1. Similarity matrix of the 20 Taskonomy task using RSA.

mation, by computing the quantitative task transferability on basis of representation similarity analysis (RSA) [1]. We show the task similarity matrix of the 20 taskonomy task using RSA in Fig. 1.

In our approach, we select as target task any pairwise combination of some pixels2pixels prediction tasks (a total of 4 tasks), which can be optimized using only L1/L2-metric loss, to learn the student network. The auxiliary model, on the other hand, is selected on basis of the similarity score in task similarity matrix and can be any model that is considered interconnected to the target ones from the 20 taskonomy pre-trained models. The hidden representation of the auxiliary model is then used to incorporate the teacher PTMs for guiding the representation learning of the student.

3 Knowledge Amalgamation from More Teachers

In the main text, we performed experiments that learn student network via the proposed SOKA-Net for any pairwise combination of the four pixes2pixels vision task. In fact, the proposed SOKA can be readily extended to amalgamate knowledge from more than two models.

Here we report the results when we handle more teachers specializing in different target tasks via the SOKA-Net. As demonstrated in Table. 1, student network of synchronous depth, edge2d, and surface normal are learnt via SOKA-Net in comparison with that of the teacher PTMs and traditional KD [2] method.

The student generated via the proposed SOKA-Net, in most cases, outperforms the KD baseline. Besides, the performance of the student network is comparable with that of its teachers on the surface normal as well as the edge 2D task, and outperforms the teacher on depth estimation, meanwhile being more compact in model size as compared to the teachers.

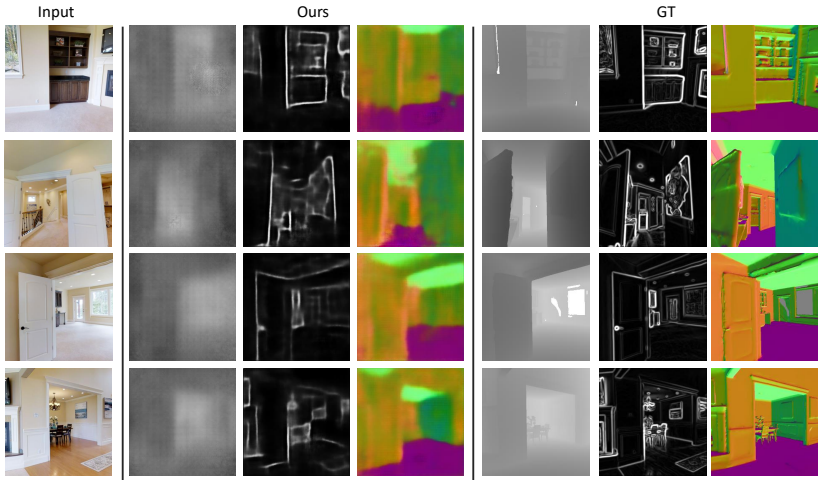

Fig. 2. Results of the student network that masters synchronous depth, edge2d, and surface normal prediction.

Table 1. Comparative results of student generated by SOKA against the teacher PTMs and knowledge distillation baseline under the task of synchronous depth, edge2d, and surface normal prediction. The performance of student on these task is measured by accuracy metric of $\sigma < 1.25, 1.25^2$ (*the higher the better*) and error metric of rmse or rel (*the lower the better for both*). Note: M is short for million.

Methods	#param	**Depth**			**Edge2D**		**Surface Normal**		
		rmse	rel	$\sigma < 1.25$	rmse	$\sigma < 1.25$	rmse	$\sigma < 1.25$	$\sigma < 1.25^2$
Teachers	$\sim 360.27M$	10.22	1.57	0.6687	6.07	0.4841	7.52	0.6767	0.7943
KD [2]	$\sim 221.96M$	10.41	1.36	0.4548	8.27	0.2453	9.68	0.4527	0.6178
Ours	$\sim 221.96M$	10.20	1.25	0.5543	6.78	0.4436	7.81	0.6213	0.7832

We also show the weight learning and the corresponding objective loss descending process during training of the student that handles the three tasks in Fig. 3. Besides, we provide more qualitative results of our method. Specially, the results of multi-talent student is shown in Fig. 2.

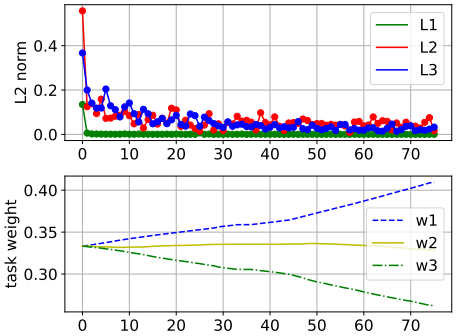

Fig. 3. Illustration of loss descending and weight changing over time during the training process of the student network that masters synchronous depth, edge2d, and surface normal prediction.

4 Multi-talent Student versus Multiple Single-task Students

We compare the proposed multiple teachers to one student (n-to-1) amalgamating approach with multiple teachers to multiple single-task students (n-to-n) distilling. These single-task students are trained via knowledge distillation (KD) [2]. For a fair comparison, the multiple KD students (with ResNet 26 backbones) have a total number of parameters similar to our multi-task student. Results are shown in Tab. 2.

Table 2. Comparison of ours one multi-talent student against multiple distilled single-task students.

Student	#params	**Depth**		**Edge2D**		**Surface Normal**		
		rmse	rel	rmse	$\sigma < 1.25$	rmse	$\sigma < 1.25$	$\sigma < 1.25^2$
KD 2-to-2	$\sim 177.74M$	10.38	1.47	7.24	0.2700	-	-	-
Ours 2-to-1	$\sim 175.75M$	10.21	1.22	6.40	0.4586	-	-	-
KD 3-to-3	$\sim 266.61M$	10.38	1.47	7.24	0.2700	8.41	0.5829	0.7296
Ours 3-to-1	$\sim 221.96M$	10.20	1.25	6.78	0.4436	7.81	0.6213	0.7832

As for efficiency, we record inference time under the same environment. The result is shown in Table. 3. As shown in Table. 2 and Table. 3, multi-talent student are not only superior in reference performance but also more efficient.

References

1. Dwivedi, K., Roig, G.: Representation similarity analysis for efficient task taxonomy & transfer learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12387–12396 (2019)

Table 3. Inference time (in sec) of different models with multiple tasks.

Model	2 tasks	3 tasks
Teachers	8.925	13.387
KD n-to-n student	7.724	11.545
Ours n-to-1 student	5.560	6.876

2. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
3. Zamir, A., Sax, A., Shen, W., Guibas, L., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3712–3722 (2018)