Collaboration by Competition: Self-coordinated Knowledge Amalgamation for Multi-talent Student Learning

Sihui Luo[1], Wenwen Pan[1], Xinchao Wang[2], Dazhou Wang[3], Haihong Tang[3], and Mingli Song[1]

[1] Zhejiang University, Hangzhou, China
{sihuiluo829,wenwenpan,brooksong}@zju.edu.cn
[2] Stevens Institute of Technology, New Jersey, USA
xinchao.wang@stevens.edu
[3] Alibaba group, Hangzhou, China
{dazhou.wangdz,piaoxue}@alibaba-inc.com

Abstract. A vast number of well-trained deep networks have been released by developers online for plug-and-play use. These networks specialize in different tasks and in many cases, the data and annotations used to train them are not publicly available. In this paper, we study how to reuse such heterogeneous pre-trained models as teachers, and build a versatile and compact student model, without accessing human annotations. To this end, we propose a self-coordinate knowledge amalgamation network (SOKA-Net) for learning the multi-talent student model. This is achieved via a dual-step adaptive competitive-cooperation training approach, where the knowledge of the heterogeneous teachers are in the first step amalgamated to guide the shared parameter learning of the student network, and followed by a gradient-based competition-balancing strategy to learn the multi-head prediction network as well as the loss weightings of the distinct tasks in the second step. The two steps, which we term as the collaboration and competition step respectively, are performed alternatively until the balance of the competition is reached for the ultimate collaboration. Experimental results demonstrate that, the learned student not only comes with a smaller size but achieves performances on par with or even superior to those of the teachers.

Keywords: Knowledge Amalgamation · Competitive Collaboration

1 Introduction

Driven by the recent advances of deep learning, remarkable progress has been made in almost all the research areas of computer vision. The unprecedentedly prominent results, nevertheless, are made possible by the immense number of annotations and hundreds or even thousands of GPU hours spent to train the deep models. To save the reproducing effort, many researchers have, therefore, generously published their pre-trained models (PTMs) online. Yet in many cases,

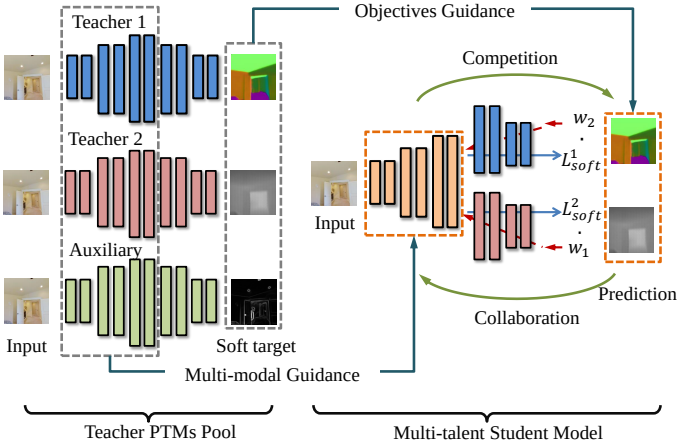


Fig. 1. Illustration of self-coordinated knowledge amalgamation pipeline.

such publicly available PTMs come without the training annotations, due to for example the data privacy issues.

In this paper, we study how to exploit PTMs that handle distinct tasks, to learn a multi-talent and compact student model, without accessing human annotations. Specifically, given a pool of heterogeneous PTMs, such as Taskonomy [34], we allow the user to pick any combination of the models from the same family, in our case autoencoder networks, and then customize a student that simultaneously tackles the distinct tasks of the selected teachers. The student training process is, again, free of human annotations. Once trained, the student not only comes with a size considerably smaller than the sum of the teachers, but also preserves and even at times surpasses performances of the teachers.

To this end, we introduce a novel strategy that treats the distinct tasks handled by the teachers as competing counterparts, and devise a *collaboration-by-competition* approach to amalgamating their heterogeneous knowledge and building the multi-talent student. In other words, different tasks compete for the student network resources to be allocated for themselves, in which process they also share features, collaborate with and benefit each other. Such collaboration-by-competition scheme leads to an adaptive loss function, of which the balance between the multiple tasks of interest is *learned* rather than handcrafted.

The proposed approach is therefore named as self-coordinated knowledge amalgamation (SOKA). The student training is achieved via a dual-step competitive-cooperation approach, where the correlated multi-modal information of the teacher models are in the first step amalgamated to guide the shared parameter learning of the student network, and followed by a competition balancing strategy to learn the multi-head prediction sub-network in the second step. The overall pipeline of the proposed SOKA is illustrated in Fig. 1.

Specifically, we adopt as teacher models some pixel-to-pixel task-specific PTMs, in our case the public available ones from Taskonomy. The intermediate representations of PTMs are utilized to guide the training of the features of the student, which are to be shared by the different tasks. The shared features are then used to train the final target task through a multi-head prediction subnetwork, in which the gradient based loss balancing method is used to balance the competition of the terminal target tasks. Both the network and the weight of the target task are learnable in the alternative learning process, until the final collaboration of the the task involved are reached.

As aforementioned, the customized network can be amalgamated from any combination of the models selected by the user from the same family, and thus the tasks involved may not be in strong interconnection. To this end, we seek for additional intermediate supervising information to guide the training of the student. We found that, as will be demonstrated in our experiments, cross modal information can serve as extra guidance in supervising the training of unlabeled data for learning more robust representation. Therefore, we introduce an auxiliary PTM, which specializes in an intermediate task close to both teachers, to providing extra-modal supervision for learning a robust shared representation of the student network. In the case of Fig. 1, for example, we take the edge 3D (occlusion edges) extracting task to be the auxiliary task to facilitate the training of surface normal and depth estimation.

Our contribution is thus summarized as follows.

- We propose a self-coordinate knowledge amalgamation method for generating a customized multi-talent student network, by reusing heterogeneous pretrained models without accessing the human annotations. This is achieved, specifically, via a novel competitive-collaboration strategy, in which both the parameters of the student network and the task-wise loss weightings are learnable.
- To bridge the potential semantic gaps between the heterogeneous tasks, we introduce an auxiliary model, which are inter-correlated to both target tasks of interest, to provide extra-modal supervision.
- We conduct a series of experiments on various combinations of PTMs set, from which we train multi-talent student models. Our results demonstrate that, the student model, which comes with a compact size, achieves results on par and at times even superior to those of the teachers.

2 Related work

Multi-task learning. Deep multi-task learning (MTL) [13, 4, 17, 10, 18] has been widely used in various computer vision problems, such as joint inference scene geometric and semantic [15], simultaneous depth estimation, surface normals and semantic segmentation [6]. It is typically conducted via hard or soft parameter sharing. In hard parameter sharing, a subset of the parameters is shared between tasks while other parameters are task-specic. In soft parameter sharing, all parameters are task-specic but they are jointly constrained via

Bayesian priors. However, most multi-task methods requires ground truth data which are either impractical or expensive to gather. Some researchers [15] have recently introduced competitive collaboration mechanism to unsupervised multi task learning for some complex geometric coupled vision tasks. Though the authors have demonstrated promising results, the balance of the task are driven by massive hand-crafted hyper-parameters. In contrast, our approach, which assume no manually labelled annotation are available, adopts gradient based loss balancing scheme in competition-collaboration training cycle, which is adaptive and requires much fewer hyperparameter.

Knowledge Distillation. Knowledge distillation (KD) [8, 9, 29, 35] adopts a teacher-guiding student strategy where a small student network learns to imitate the output of a large and deeper teacher network. In this way, the large teacher can transfer knowledge to the student with smaller model size, which is widely applied to model compression [32]. Following [8], some works are proposed to exploit the intermediate representation to optimize the learning of student network, such as FitNet [16], DK2PNet [27], AT [33] and NST [26]. Albeit many heuristics are found by these works, most knowledge distillation methods fall into single-teacher single-student manner, where the teacher and the student handles the same task. Recently, some researches [20, 31, 12, 30] started to investigate how to transfer knowledge from multiple trained models into a single one with unlabeled dataset. They generally adopted an auto-encoder architecture to amalgamate features from multiple single-task teachers in a layerwise manner. [21] customized the student network by generating component nets as byproducts for attribute learning tasks. [12] proposed a common feature learning approach for learning the student from heterogeneous-architecture teachers. These methods generally focus on designing better teacher-student learning architecture. In contrast, the proposed method aims at adaptive balancing of the target task learning and seek the collaboration of the target tasks by self-coordinated training strategy.

Model Transferability. Transfer learning [14, 3], is similar to multi-task learning in that solutions are learned for multiple tasks. Unlike multi-task learning, however, transfer learning methods rst learn a model for a source task and then adapt that model to a target task. Razavian *et al.* [19] demonstrated that features extracted from deep neural networks could be used as generic image representations to tackle the diverse range of visual tasks. Azizpour *et al.* [2] studied several factors affecting the transferability of deep features. Albeit many heuristics are found by these works, none of them explicitly quantify the transferability among different tasks and layers to provide a principled way for model selection. Recently, Taskonomy [34] aimed to find the underlying task relatedness by computing the transfer performance among tasks. This was followed by a number of recent works, which further analyzed task relationships [1, 5, 24, 17, 23, 22] for transfer learning. RSA [5] adopted representation similarity analysis to find relationship between visual tasks for efficient task taxonomy. [13] utilizes the mid-level representations from the labeled modality to supervise learning representations on

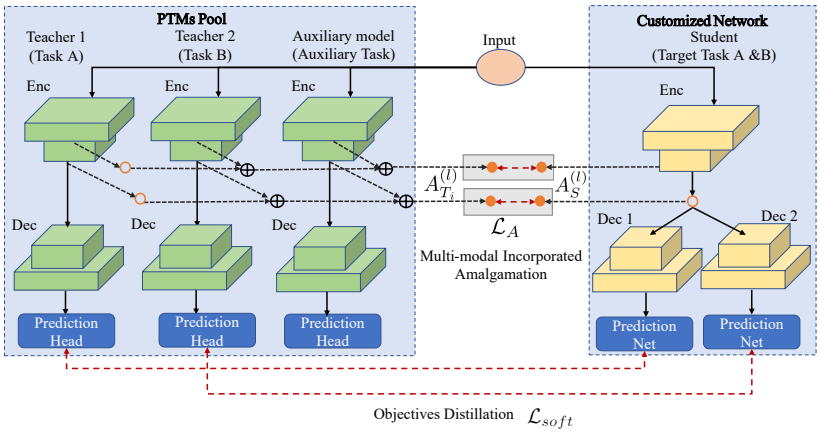


Fig. 2. Illustration of the proposed SOKA for customizing pixel2pixel multi-task network with unlabeled data. The student learns both the predictions and the intermediate representation from multiple teacher models in the PTMs pool. Auxiliary PTM, which is specialized at task strongly correlated to the target ones, is utilized to incorporate the teacher PTMs for providing multi-modal information in training more robust representation of the student network. Losses of representation transfer and objectives distillation are penalized to train the parameter of the student network.

the paired unlabeled modality. Model transferability provides the quantified evaluation of how connected the vision task are and it inspires us at introducing such evaluation to decide which auxiliary model can reinforce the target tasks via joint learning.

3 Self-coordinated Knowledge Amalgamation

In this section, we describe the proposed approach of Self-coordinated Knowledge Amalgamation (SOKA), which enables customizing a student network from coupled PTMs without accessing annotations. Specifically, we build SOKA by a multi-head encoder-decoder network, allowing for dense pixel-level prediction tasks, such as depth and surface-normal prediction. Our training, as will be demonstrated in the following section, is achieved via a novel strategy that learns the parameters of the student intertwined with those of the pre-trained teachers.

3.1 Architecture Design

As depicted in Fig. 2, the SOKA mainly consists of two parts, the PTMs pool and the customized network. The knowledge are transferred from the PTMs to the target network via two flow, the multi-modal incorporated feature amalgamation flow and the objectives distillation flow. For the former, we introduce a Multi-Modal Incorporated Amalgamation (MIA) scheme to transform multiple

teachers' expertise to student domain for computing the loss and thus updating the parameters of the shared encoder of the student network. For the latter, we propose an adaptive competition-collaboration training strategy, in which a gradient-based competition-balancing strategy is introduced to learn the multi-head prediction subnetwork as well as the loss weightings of the distinct tasks.

3.2 Multi Modal Incorporated Amalgamation

Consider a collection of PTM models term as PTMs pool, which consist of the teacher models and some auxiliary model. The auxiliary PTM, which is chosen according to the quantilized transferability of task taskonomy [34], is considered beneficial for providing multi-modal information for supervising the student network. To PTMs with the same CNN backbone, the regions with high activation from a neuron at the same depth may share some task related similarities, even though these similarities may not be intuitive for human interpretation.

Auxiliary PTM Selection We set a gate condition on basis of RDM correlation [25] to determine whether it is necessary to introduce an auxiliary PTM before training to guide the training of their student network. Representation dissimilarity matrices (RDMs) are generated by computing the pairwise dissimilarity (1 - Pearsons correlation) of each image pair in a subset of selected images. The similarity score $S(i,j)$ of task i and j are computed by Spearmans correlation of the low triangular RDMs of the two models.

Assume we have a set of PTM models $\{T_k\}_{k=1,\dots,m}$, denote T_i and T_j as the teacher PTMs of target task i and j respectively, and the pairwise similarity score [25] of them is $S_{(i,j)}$. Denote δ as the similarity score threshold that determines whether the two task are considered as correlated. If $S_{(i,j)} < \delta$, we search for a PTM term as T_x, which satisfy $\frac{(S_{(i,x)}-S_{(i,j)})^2}{S_{(i,x)}-S_{(i,j)}} + \frac{(S_{(i,x)}-S_{(i,j)})^2}{S_{(i,x)}-S_{(i,j)}} > 0$.

In our implementation, $\{T_k\}_{k=1,\dots,m}$ are the 21 single-task taskonomy models. The code of computing RDMs and similarity scores is available on the Internet[4]. More details are available in the supplementary document.

Feature Amalgamation with Multi-Modal Knowledge In iteration t, denote the activation map produced by the teacher network T_i at a particular layer l by $A^{(l)}_{T_i} \in \mathbb{R}^{c\times h\times w}$, where c is the number of output channels, and h and w are spatial dimensions. Let the activation map produced by the student network S at layer l be given by $A^{(l)}_{S} \in \mathbb{R}^{c\times h\times w}$. We note that as our student and its teachers share the same depth such that we compare the representation of them at the same depth. l can be a intermediate layer and the encoder output. Student mimic the target representation filtered from heterogeneous teachers by the Multi-modal Incorporated Amalgamation (MIA) module, we write

$$\sigma^{(l)}(t) = MIA\left(A^{(l)}_{T_1}(t), \cdots, A^{(l)}_{T_m}(t)\right). \tag{1}$$

[4] https://github.com/kshitijd20/RSA-CVPR19-release

The MIA takes as input the stacked representation of the teacher network and generate the target representation $\sigma^{(l)}(t)$ via passing message from the feature maps of other tasks as follows,

$$MIA\left(A_{T_1}^{(l)}(t), \cdots, A_{T_m}^{(l)}(t)\right) = C \odot \sum_{i}^{m} W_i^{(l)}(t) \otimes A_{T_i}^{(l)}(t), \tag{2}$$

where $\otimes$ denotes matrix multiplication operation, $\odot$ denotes the elementary multiplication and $W_i^{(l)}(t)$ denotes the parameters of the weight. Both C and $W_i^{(l)}(t)$ are learnable in the representation learning process.

To guide the student towards the activation correlations induced in the amalgamated activation of the multiple teachers, we define a representation transfer loss that penalizes differences in the L2-normalized outer products of the student's activation and the corresponding target activation $\sigma^{(l)}(t)$:

$$\mathcal{L}_A^{(l)}(t) = \|A_S^{(l)}(t) - \sigma^{(l)}(t)\|_2^2 \tag{3}$$

To this end, we define the total loss for transferring the knowledge induced in representation of the selected group of teachers to the student network as:

$$\mathcal{L}_A(t) = \sum_{l}^{L} \mathcal{L}_A^{(l)}(t), \tag{4}$$

where L is the number of layers, whose knowledge needs to be transferred from the teacher to the student network. In our implementation, we tile the representation similarity comparison in the 3rd block of convolution layer and the encoder output of the PTMs set and the student.

3.3 Objectives Distillation

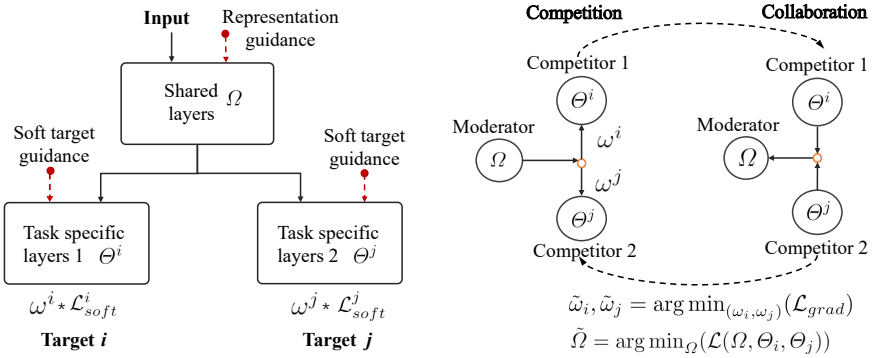
To imitate the predictions of teachers, we introduce a soft target loss between the predictions of teacher networks and that of the student. We write,

$$\mathcal{L}_{soft}^i(t) = \|\mathcal{F}_S^{score}(t) - \mathcal{F}_{T_i}^{score}(t)\|^2, \tag{5}$$

where $\mathcal{F}_S^{score}$ and $\mathcal{F}_{T_i}^{score}$ denote the prediction of the student and teachers. Our multi-task loss function is thus defined as:

$$\mathcal{L}_{soft}(t) = \sum_{i}^{n} \omega_g^i(t) \mathcal{L}_{soft}^i(t), \tag{6}$$

where $w_g^i(t)$ is the gradient based weighting function for balancing the target tasks. Let $G_W^{(i)}(t) = \| \bigtriangledown_W \omega^i(t)\mathcal{L}^i(t)\|_2^2$ be the L2 norm of the gradient of the weighted single-task loss $\omega^i(t)\mathcal{L}^i(t)$ with respect to the chosen weights ω. As in GradNorm [4], we use the relative inverse training rate of task i, $r_i(t)$, to balance our gradients of objectives distillation. $w_i^g(t)$ is designed to move gradient norms towards the target for each task, $\bar{G}(t) \times [r^i(t)]^\alpha$. GradNorm is then implemented



(a) Target network (b) Iterative competitive-collaboration

Fig. 3. The illustrative diagram of the adaptive competitive-collaboration training (ACCT) process. (a) is the parameterized target network. (b) demonstrate the training cycle of competition-collaboration. Task specific layers are considered as two competitor that are moderated by the moderator. The shared layers are considered as the moderator who controls the resource, which is the shared representation utilized for inference of task i and j. The task balance weight ω^i and ω^j are adaptively determined by the function of Ω.

as an L1 loss function $\mathcal{L}_{grad}$ between the actual and target gradient norms at each time step for each target task, summed over all tasks:

$$\mathcal{L}_{grad}(t; \omega^i_g(t)) = \sum_i \|G^i_W(t) - \bar{G}(t) \times [r^i(t)]^\alpha\|_1, \tag{7}$$

where the summation runs through all T tasks. When differentiating this loss $\mathcal{L}_{grad}$, we treat the target gradient norm $\bar{G}(t) \times [r^i(t)]^\alpha$ as a fixed constant to prevent loss weights $\omega^i_g(t)$ from spuriously drifting towards zero. $\mathcal{L}_{grad}$ is then differentiated only with respect to ω^i_g, as it directly control gradient magnitudes per task.

3.4 Adaptive Competitive-Collaboration Training Strategy

Competitive collaboration is typically formulated as a three-player game consisting of two counterparts competing for a resource that is regulated by a moderator. In the context of knowledge amalgamation for customizing a multi-task model, the moderator is the shared layers (the encoder) who map the input to some shared activation for inference. The two counterparts thus are the prediction subnetworks of the target tasks and compete for more inference-beneficial information in the shared representation to minimize their individual loss.

As depicted in Fig. 3, we use Ω, Θ_i, and Θ_j to denote the parameter of the shared encoder, and that of the two task specific prediction subnetwork respectively. The competing players Θ_i and Θ_j minimize their loss function $\mathcal{L}^i_{soft}$

and $\mathcal{L}_{soft}^{j}$ respectively such that each player optimizes for itself but not for the group. To resolve this problem, our training cycle consists of three phases. In the first phase which we term as competition step, we train the competitors by xing the moderator network parameter Ω and minimizing Eq. 5. In the second phase which we term as collaboration phase, the competitors(Θ_i, Θ_j) form a consensus and train the moderator Ω such that it correctly distributes the data in the next phase of the training cycle. In the third phase, task weights ω^i and ω^j are learnt adaptively by minimizing the gradient loss $\mathcal{L}_{grad}$. We note that the moderator and the competitors are initialized jointly before the training cycle to set the whole network to a good start point.

To summarize, Ω, Θ_i, and Θ_j are learnt through a multi-stage alternate training process as follows:

- **Step 1**: Randomly initialize the parameters of the student network.
- **Step 2**: Jointly initialize Ω, Θ_i and Θ_j with the input and the prediction of the teacher PTMs for 1000 steps. ω^i and ω^j are fixed to an initial value of 0.5 in this step.
- **Step 3**: Competition step. Freeze Ω, updating Θ_i and Θ_j by minimizing the soft target distilling loss $\mathcal{L}_{soft}$ with the weight w_i and w_j.
- **Step 4**: Collaborative step. Freeze Θ_i and Θ_j, training the shared layers Ω by Eq. 8 with representation transfer loss defined in Eq. 4 and the inference loss of the two competitors. The α in Eq. 8 is a hyperparameter to balance the magnitude level of $\mathcal{L}_A$ and $\mathcal{L}_{soft}$ and is set to be 0.05.

$$\tilde{\Omega} = \arg\min_{\Omega}(\alpha(\mathcal{L}_{soft}^{i}(x;\Omega,\Theta_i) + \mathcal{L}_{soft}^{j}(x;\Omega,\Theta_j)) + \mathcal{L}_A(x;\Omega)) \tag{8}$$

- **Step 5**: Adaptive weighting step. Update $w_i^g(t)$ with the task specific weight function by minimizing the gradient loss $\mathcal{L}_{grad}$ defined in Eq. 7:

$$\tilde{\omega}_i, \tilde{\omega}_j = \arg\min_{(\omega_i,\omega_j)}(\mathcal{L}_{grad}). \tag{9}$$

  After every update step, we renormalize the weights ω^i so that $\sum_i \omega_g^i(t) = 1$ in order to decouple gradient normalization from the global learning rate.
- **Step 6**: If the maximum training step is not reached, go to step 3 and continue the training loop. The maximum step in our case is set to be 6×10^5.

4 Experiment

We now describe a number of diagnostic experiments of the proposed approach carried out using taskonomy dataset [34] which provide various pre-trained vision models as well as extensive data. In the following the detailed description of our experimental evaluation is given. Besides, detailed experimental settings and more experimental results refer to our supplementary document.

4.1 Experimental Setup

Dataset Taskonomy dataset includes over 4 million indoor images from 500 buildings with annotations available for 26 image tasks. 21 of these tasks are single image tasks, and 5 tasks are multi-image tasks. For this work, we select one building (wiconisco) from taskonomy dataset, which contains 16749 images, to evaluate the proposed method. We divide them into 13749 training and 3000 validation images. For training, only RGB images are feed as the input to the student and teacher network. The student take the prediction of the task specific teacher network as supervision without accessing the annotations.

Pre-trained Teacher Models We adopt the taskonomy models[5], which consist of an encoder and decoder, as our pre-trained teacher models. The encoder for all the tasks is a ResNet-50 [7] model followed by convolution layer that compresses the channel dimension of the encoder output from 2048 to 8. The decoder is task-specific and varies according to the task. Among these models, we mainly select ones specialized for pixel prediction tasks. The decoder of these pixel prediction models consists of 15 layers (except colorization with 12 layers) consisting of convolution and deconvolution layers. The compressed output as long as representation of earlier layers of the teachers' encoder is also used as guidance to train the target network.

In addition to the representation of the target task specific PTMs, we also explore the representation of highly correlated tasks as auxiliary guidance for training the target ones. We perform this analysis to investigate how the features of models of interconnected tasks cooperate to reinforce the target network especially for loosely related target tasks.

Evaluation Metric For evaluating the performance of the vision tasks involved in this paper, we use several quantitative metrics following previous works [28]. For the pixel-wise prediction tasks involved in this paper, we adopt several metric including mean relative error (rel), root mean squared error (rmse) and the percentage of relative errors inside three thresholds (1.25,1.25^2,1.25^3).

Implementation Details The proposed method is implemented using TensorFlow with a NVIDIA M6000 with 24GB memory. We use the poly learning rate policy as done in [11], and set the base learning rate to 0.001, the power to 0.9, and the weight decay to 10^{-6}. The student take the prediction of the task specific teacher network as supervision without accessing the annotations.

4.2 Qualitative Evaluation

Given some sample queries, the results of student network built by proposed SOKA and the teacher PTMs are shown in Fig. 4 for visual perception. Synchronous

[5] publicly available at https://github.com/StanfordVL/taskonomy

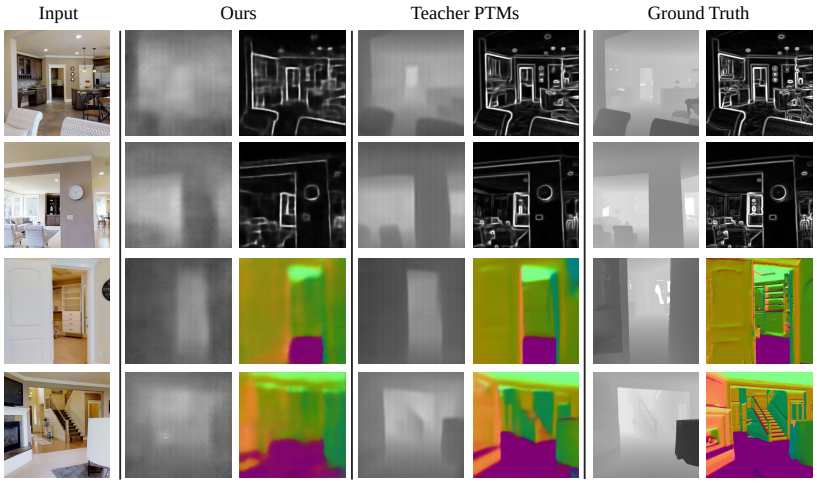


Fig. 4. Qualitative illustration of task specific outputs for the query (first column). Here, prediction of teacher PTMs, and proposed SOKA are compared. The first two rows are results of our synchronous depth and edge2d prediction in contrast with the task-specific teacher PTMs. The bottom two rows are results of our synchronous depth and surface normal prediction and that of the task-specific teacher PTMs.

depth estimation and surface normal, as well as synchronous edge prediction and depth estimation are compared in Fig. 4. It can be observed that though smaller in model size, the multi-talent student network, which are built via the proposed SOKA with limited unlabeled training data, achieves comparable visual performance with the teacher PTMs which are trained on the million-level training data with ground truth.

4.3 Quantitative Evaluation

Performance of student network learnt by SOKA We show in Table 1 the quantitative results of the teacher network and those of the student network that specialized in five group of target tasks. We performed tests on five group of pixel2pixel prediction tasks to evaluate the performance of SOKA. The five group are (Edge 2D, Depth estimation), (Edge 2D, Surface Norm), (Depth estimation, Surface Norm), (Depth estimation, Edge 3D), (Edge 2D, Edge 3D) and (Edge 3D, Surface Norm) respectively. Additionally, we collect the parameters of student network, teacher PTMs and a direct multi-task learning method with the same architecture as the student do. The number of their parameters are shown in Table. 2. It can be observed from Table 1 and Table 2 that the performance of multi-talent student networks learnt via proposed SOKA under different target task groups are generally on par or sometimes even better than the teacher PTMs and yet compact in model size.

Table 1. Comparative results on Depth Estimation, Surface Normal, Edge 2D, and Edge 3D task specific prediction of the single-task teacher PTMs and multi-talent student trained with different task groups. ($\sigma < 1.25, 1.25^2, 1.25^3$: *the higher the better*, rmse and rel: *the lower the better*)

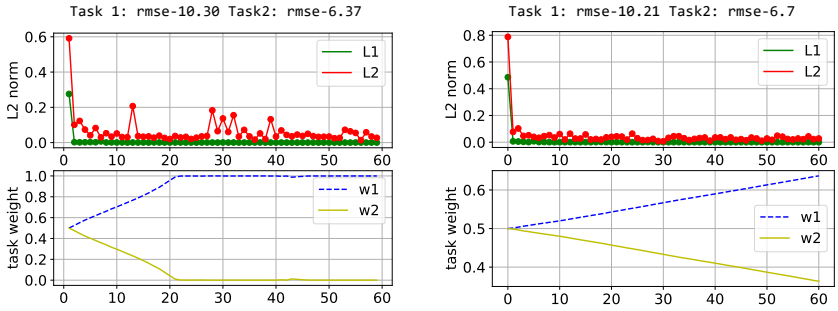
Model	**Depth**			Model	**Surface Normals**		
	rmse	rel	$\sigma < 1.25$		rmse	$\sigma < 1.25$	$\sigma < 1.25^2$
teacher	10.22	1.57	0.6687	teacher	7.52	0.6767	0.7943
depth-edge2d	10.21	1.22	0.6581	sfnorm-depth	7.56	0.6400	0.7283
depth-sfnorm	10.32	1.32	0.4656	sfnorm-edge2d	7.51	0.6587	0.7845
depth-edge3d	10.21	1.25	0.5841	sfnorm-edge3d	7.57	0.6323	0.7215
Model	**Edge 2D**			Model	**Edge 3D**		
	rmse	$\sigma < 1.25$	$\sigma < 1.25^2$		rmse	$\sigma < 1.25$	$\sigma < 1.25^2$
teacher	6.07	0.4841	0.7256	teacher	6.11	0.4840	0.7056
edge2d-depth	6.40	0.4586	0.7112	edge3d-depth	6.12	0.4706	0.6988
edge2d-sfnorm	6.51	0.4508	0.6997	edge3d-sfnorm	6.43	0.3913	0.6731
edge2d-edge3d	6.06	0.4774	0.7231	edge3d-edge2d	6.09	0.4763	0.7159

Table 2. Parameters of the teachers, student and MTL network. M is short for million.

Model	Teacher PTM	Student	MTL
#params	∼246.46 M	∼175.75 M	∼175.73 M

Robustness of The Task Weightings Learning As ω^i and ω^j are changing over time due to the alternative competition-collaboration training, we study the effect of ω^i and ω^j on performance of the target tasks of the learned student network when their value varies in a rather wide range, and show the results in Fig. 5. Under the initial learning rate of 0.001 in Fig. 5(a), the ratio ω^i/ω^j of the final ω^i and ω^j have eventually grown to hundreds while that under an initial learning rate of 0.0001 grew to about 2 times (Fig. 5(b)). It can be observed that no matter the weighting ratios is small to 2 or big to hundreds, parameter training of the whole network seem still goes the right way.

Comparative Results of SOKA Against Supervised Method Though we assume no manually labeled annotations but only some pre-trained PTMs are available in training the student network, we compare our method with supervised multi-task learning method. In particular, two alternative methods are compared. The first is direct multi-task learning (dMTL) that intuitively train the two target task together under the same student network architecture but with labeled data as ground truth. The second is similar to the first one but with GradNorm [4] method to balance the task weight during learning. The results of these methods on synchronous depth and edge2d prediction tasks are shown in Table. 3. As demonstrated in Table. 3, The performance of student network generated by SOKA generally outperforms the two multi-task learning methods on these two target tasks.



(a) initial lr: 0.001 (b) initial lr:0.0001

Fig. 5. Illustration of task training loss decay of the proposed SOKA method with different learning rate in the steps of task weightings learning. The corresponding rmse results of the final model on both tasks are also shown on the top of each subfigure.

Table 3. Comparative results of student generated by SOKA against supervised multi-task learning method under the task of synchronous depth and edge2d prediction.

Methods	**Depth**			**Edge2D**			
	rmse	rel	$\sigma < 1.25$	rmse	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
dMTL	10.26	1.23	0.4953	8.84	0.2002	0.3763	0.5037
GradNorm [4]	10.23	1.22	0.5581	7.48	0.4191	0.5270	0.6402
Ours	10.21	1.22	0.6176	6.40	0.4586	0.7112	0.7855

Ablation Studies In the basic mode, proposed SOKA takes unlabeled training image as input, and adopts knowledge distillation (KD) [8] method to impel the student to mimic the prediction of the teacher PTM. Due to the possible weak-interconnection of the customized task, we add multi-modal incorporated amalgamation (MIA) to the basic KD mode. Besides, to further achieve the collaboration of the tasks by balancing the competition, we introduce the adaptive competitive-collaboration training on basis of the KD and MIA.

In this section, ablation studies are conducted to investigate the effectiveness of the modules adopted in SOKA. We verify the effectiveness of each module by comparing the whole model to the model without the corresponding module. The additional compared method is KD, and KD with MIA. The results are shown in Table 5. Besides, we also analyze if correlated auxiliary model can enhance the performance of target task in Table 4. It can be observed from the two tables that both MIA and ACCT are beneficial for alleviating errors and enhancing the inference performance.

5 Conclusions

In this paper, we study how to reuse heterogeneous pre-trained models as teachers, and build a versatile and compact student model, without accessing hu-

Table 4. Comparative results of the teacher PTMs and the student of synchronous Depth estimation and Surface Normals with/without supervision of auxiliary PTM. For auxiliary supervision in multi-modal incorporated representation amalgamation, a weak connected auxiliary task, Vanishing Point, is compared against the strong connected auxiliary task Keypoint 3D.

Methods	**Depth**			**Edge2D**			
	rmse	rel	$\sigma < 1.25$	rmse	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
teacher	10.22	1.57	**0.6687**	**6.07**	**0.4841**	**0.7256**	**0.8148**
w/o auxiliary	10.26	1.23	0.4953	7.83	0.4002	0.5295	0.6132
vanishing Point	10.42	1.23	0.4737	8.86	0.2144	0.3227	0.5159
keypoint3d	**10.21**	**1.22**	0.6176	6.70	0.4586	0.7112	0.7981

Table 5. Ablation study of each component of the proposed SOKA under the task of synchronous depth and edge2d prediction.

Methods	**Depth**		**Edge2D**			
	rmse	rel	rmse	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
KD	10.26	1.23	7.83	0.2002	0.3905	0.5295
KD + MIA	10.22	1.25	6.91	0.3102	0.5216	0.7690
KD + MIA + ACCT	10.21	1.22	6.40	0.4586	0.7112	0.7855

man annotations. To this end, we introduce a novel strategy that treats the multiple tasks handled by the distinct teachers as competing counterparts, and devise a collaboration-by-competition approach to amalgamating their heterogeneous knowledge and building the multi-talent student. This collaboration-by-competition approach, which we call as SOKA, is achieved via a dual-step adaptive competitive-cooperation training approach, where the knowledge of the heterogeneous teachers are in the first step amalgamated to guide the shared parameter learning of the student network, and followed by a gradient-based competition-balancing strategy to learn the multi-head prediction subnetwork as well as the loss weightings of the distinct tasks in the second step. Experimental results demonstrate that, the learned student not only comes with a smaller size but all achieves performances on par with or even superior to those of the teachers.

Acknowledgments

This work is supported by National Key Research and Development Program (2018AAA0101503), National Natural Science Foundation of China (61976186), Key Research and Development Program of Zhejiang Province (2018C01004), and the Major Scientific Research Project of Zhejiang Lab (No. 2019KD0AC01).

References

1. Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C.C., Soatto, S., Perona, P.: Task2vec: Task embedding for meta-learning. In: IEEE International Conference on Computer Vision (ICCV). pp. 6430–6439 (2019)
2. Azizpour, H., Razavian, A.S., Sullivan, J., Maki, A., Carlsson, S.: Factors of transferability for a generic convnet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **38**(9), 1790–1802 (2015)
3. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning **79**(1-2), 151–175 (2010)
4. Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A.: Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: International Conference on Machine Learning (ICML). pp. 794–803 (2018)
5. 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)
6. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: IEEE International Conference on Computer Vision (ICCV) (2015)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
8. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
9. Huang, Z., Wang., N.: Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219 (2017)
10. Liu, S., Johns, E., Davison, A.J.: End-to-end multi-task learning with attention. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1871–1880 (2019)
11. Liu, W., Rabinovich, A., Berg, A.C.: Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579 (2015)
12. Luo, S., Wang, X., Fang, G., Hu, Y., Tao, D., Song, M.: Knowledge amalgamation from heterogeneous networks by common feature learning. International Joint Conference on Artificial Intelligence (IJCAI) (2019)
13. Misra, I., Shrivastava, A., Gupta, A., Hebert, M.: Cross-stitch networks for multi-task learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3994–4003 (2016)
14. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2009)
15. Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulff, J., Black, M.J.: Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12240–12249 (2019)
16. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: International Conference on Learning Representations (ICLR) (2015)
17. Ruder, S., Bingel, J., Augenstein, I., Søgaard, A.: Latent multi-task architecture learning. In: AAAI Conference on Artificial Intelligence (AAAI). vol. 33, pp. 4822–4829 (2019)

18. Sener, O., Koltun, V.: Multi-task learning as multi-objective optimization. In: Neural Information Processing Systems (NeurIPS). pp. 527–538 (2018)
19. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: CVPR Workshops. pp. 806–813 (2014)
20. Shen, C., Wang, X., Song, J., Sun, L., Song, M.: Amalgamating knowledge towards comprehensive classification. In: AAAI Conference on Artificial Intelligence (AAAI) (2019)
21. Shen, C., Xue, M., Wang, X., Song, J., Sun, L., Song, M.: Customizing student networks from heterogeneous teachers via adaptive knowledge amalgamation. In: IEEE International Conference on Computer Vision (ICCV). pp. 3504–3513 (2019)
22. Song, J., Chen, Y., Wang, X., Shen, C., Song, M.: Deep model transferability from attribution maps. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
23. Song, J., Chen, Y., Ye, J., Wang, X., Shen, C., Mao, F., Song, M.: Depara: Deep attribution graph for deep knowledge transferability. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
24. Standley, T., Zamir, A.R., Chen, D., Guibas, L., Malik, J., Savarese, S.: Which tasks should be learned together in multi-task learning? arXiv preprint arXiv:1905.07553 (2019)
25. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: IEEE International Conference on Computer Vision (ICCV). pp. 1365–1374 (2019)
26. Wang, H., Zhao, H., Li, X., Tan, X.: Progressive blockwise knowledge distillation for neural network acceleration. In: International Joint Conference on Artificial Intelligence (IJCAI). pp. 2769–2775 (2018)
27. Wang, Z., Deng, Z., Wang, S.: Accelerating convolutional neural networks with dominant convolutional kernel and knowledge pre-regression. In: European Conference on Computer Vision (ECCV). pp. 533–548 (2016)
28. Xu, D., Ouyang, W., Wang, X., Sebe, N.: Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 675–684 (2018)
29. Yang, Y., Qiu, J., Song, M., Tao, D., Wang, X.: Distilling knowledge from graph convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
30. Ye, J., Ji, Y., Wang, X., Gao, X., Song, M.: Data-free knowledge amalgamation via group-stack dual-gan. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
31. Ye, J., Ji, Y., Wang, X., Ou, K., Tao, D., Song, M.: Student becoming the master: Knowledge amalgamation for joint scene parsing, depth estimation, and more. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
32. Yu, X., Liu, T., Wang, X., Tao, D.: On compressing deep models by low rank and sparse decomposition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
33. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: International Conference on Learning Representations (ICLR) (2017)
34. Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3712–3722 (2018)

35. Zhao, Y., Xu, R., Wang, X., Hou, P., Tang, H., Song, M.: Hearing lips: Improving lip reading by distilling speech recognizers. In: AAAI Conference on Artificial Intelligence, (AAAI) (2020)