

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

Jingwen Ye¹, Yixin Ji¹, Xinchao Wang², Kairi Ou³, Dapeng Tao⁴, Mingli Song¹

¹College of Computer Science and Technology, Zhejiang University, Hangzhou, China

²Department of Computer Science, Stevens Institute of Technology, New Jersey, United States

³Alibaba Group, Hangzhou, China

⁴School 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

Abstract

In this paper, we investigate a novel deep-model reusing task. Our goal is to train a lightweight and versatile student model, without human-labelled annotations, that amalgamates the knowledge and masters the expertise of two pre-trained teacher models working on heterogeneous problems, one on scene parsing and the other on depth estimation. To this end, we propose an innovative training strategy that learns the parameters of the student intertwined with the teachers, achieved by “projecting” its amalgamated features onto each teacher’s domain and computing the loss. We also introduce two options to generalize the proposed training strategy to handle three or more tasks simultaneously. The proposed scheme yields very encouraging results. As demonstrated on several benchmarks, the trained student model achieves results even superior to those of the teachers in their own expertise domains and on par with the state-of-the-art fully supervised models relying on human-labelled annotations.

1. Introduction

Deep learning has accomplished unprecedentedly encouraging results in almost every dimension of computer vision applications. The state-of-the-art performances, however, often come at the price of the huge amount of annotated data trained on clusters for days or even weeks. In many cases, it is extremely cumbersome to conduct the training on personal workstations with one or two GPU cards, not to mention the infeasibility in the absence of training annotations.

This dilemma has been in part alleviated by the fact that many trained deep models have been released online by developers, enabling their direct deployment by the commu-

nity. As such, a series of work has been conducted to investigate reusing pre-trained deep models. Examples include the seminal work of knowledge distilling [11], which learns a compact student model using the soft targets obtained from the teachers. The more recent work of [39] makes one step further by training a student with faster optimization and enhanced performance. Despite the very promising results achieved, existing knowledge distilling methods focus on training a student to handle the same task as a teacher does or a group of homogeneous teachers do.

We propose in this paper to study a related yet new and more challenging model reusing task. Unlike the conventional setup of knowledge distilling where one teacher or an ensemble of teachers working in the same domain, like classification, are provided as input, the proposed task assumes that we are given a number of task-wise heterogeneous teachers, each of which works on a different problem. Our goal is then to train a versatile student model, with a size smaller than the ensemble of the teachers, that *amalgamates* the knowledge and learns the expertise of all teachers, again without accessing human-labelled annotations.

Towards this end, we look into two vital pixel-prediction applications, scene parsing and depth estimation. We attempt to learn a light student model that simultaneously handle both tasks from two pre-trained teachers, each of which specializes in one task only. Such a compact and dual-task student model finds its crucial importance in autonomous driving and robotics, where the model to be deployed should, on the one hand, be sufficiently lightweight to run on the edge side, and on the other hand, produce accurate segmentation and depth estimation for self-navigation.

To amalgamate knowledge from the two teachers in different domains, we introduce an innovative block-wise training strategy. By feeding unlabelled images to the teachers and the student, we learn the amalgamated features

within each block of the student intertwined with the teachers. Specifically, we “project” the amalgamated features onto each of the teachers’ domains to derive the transferred features, which then replace the features in the corresponding block of the teacher network for computing the loss. As the first attempt along this line, we assume for now that teacher models share the same architecture. This may sound a strong assumption but in fact not, as the encoder-decoder-like architecture has been showing state-of-the-art performance in many vision applications.

We also show that the proposed amalgamation method can be readily generalized for training students with three or more heterogeneous-task teachers. We introduce two such multi-teacher amalgamation strategies, one offline and one online, and demonstrate them with a third task of surface normal estimation.

The proposed training strategy for knowledge amalgamation yields truly promising results. As demonstrated on several benchmarks, the student model ends up with not only a compact size, but also desirable performances superior to those of the teachers on their own domains, indicating the knowledge aggregated from the heterogeneous domains may benefit each other’s task. Without accessing human-labelled annotations, the student model achieves results on par with the fully supervised state-of-the-art models trained with labelled annotations.

Our contribution is therefore an innovative knowledge amalgamation strategy for training a compact yet versatile student using heterogeneous-task teachers specializing in different domains. We start with the tasks of scene parsing and depth estimation, and show that the proposed strategy can be seamlessly extended to multiple tasks. Results on several benchmarks demonstrate the learned student is competent to handle all the tasks of the teachers with superior and state-of-the-art results, but comes with a lightweight size.

2. Related Work

We give a brief review here of the recent advances in scene parsing and depth estimation. We also discuss a related but different task, knowledge distillation, that aims to train a student model handling the same task as the teacher does.

Scene Parsing. Convolutional neural networks (CNNs) have recently achieved state-of-the-art scene parsing performances and have become the mainstream model. Many variants have been proposed based on CNNs. For example, PSPNet [40] takes advantage of pyramid pooling operation to acquire multi-scale features, RefineNet [18] utilizes a multi-path structure to exploit features at multiple levels, and FinerNet [38] cascades a series of networks to produce parsing maps with different granularities. SegNet [1], on the other hand, employs an encoder-decoder architecture fol-

lowed by a final pixelwise classification layer. Other models like the mask-based networks of [9, 26, 27] and the GAN-based ones of [14, 22, 35] have also produced very promising scene parsing results.

In our implementation, we have chosen SegNet as our scene-parsing teacher model due to its robust and state-of-the-art performance. However, the training strategy proposed in Sec. 5 is not limited to SegNet only, and other encoder-decoder scene parsers can be adopted as well.

Depth Estimation. Earlier depth estimation methods [29, 30, 31] rely on handcrafted features and graphical models. For example, the approach of [30] focuses on outdoor scenes and formulates the depth estimation as a Markov Random Field (MRF) labeling problem, where features are handcrafted. More recent approaches [16, 17, 19] apply CNNs to learn discriminant features automatically, having accomplished very encouraging results. For example, [5] introduces a multi-scale deep network, which first predicts a coarse global output followed by finer ones.

The approaches of [7, 25] handle depth estimation with other vision tasks like segmentation and surface normal prediction. The work of [36], on the other hand, proposes a multi-task prediction and distillation network (PAD-Net) structure for joint depth estimation and scene parsing in aim to improve both tasks. Unlike existing approaches, the one proposed in this paper aims to train a student model by learning from two pre-trained teachers working in different domains, without manually-labelled annotations.

Knowledge Distillation. Existing knowledge distillation methods focus on learning a student model from a teacher or a set of homogeneous teachers working on the same problem. The learned student model is expected to handle the same task as the teacher does, yet comes in a smaller size and preserves the teachers’ performance. The work of [11] shows knowledge distilling yields promising results with a regularized teacher model or an ensemble of teachers, when applied to classification tasks. [28] extends this idea to enable the training of a student that is deeper and thinner than the teacher by using both the outputs and the intermediate representations of the teacher.

Similar to knowledge distillation, [6] proposes a multi-teacher single-student knowledge concentration method to classify 100K object categories. The work of [32] proposes to train a student classifier that handles the comprehensive classification problem by learning from multiple teachers working on different classes.

Apart from classification, knowledge distillation has been utilized in other tasks [2, 13, 36]. The work of [2] resolves knowledge distilling on object detection and learns a better student model. The approach of [13] focuses on sequence-level knowledge distillation and has achieved encouraging results on speech applications.

Unlike knowledge distilling methods aiming to train

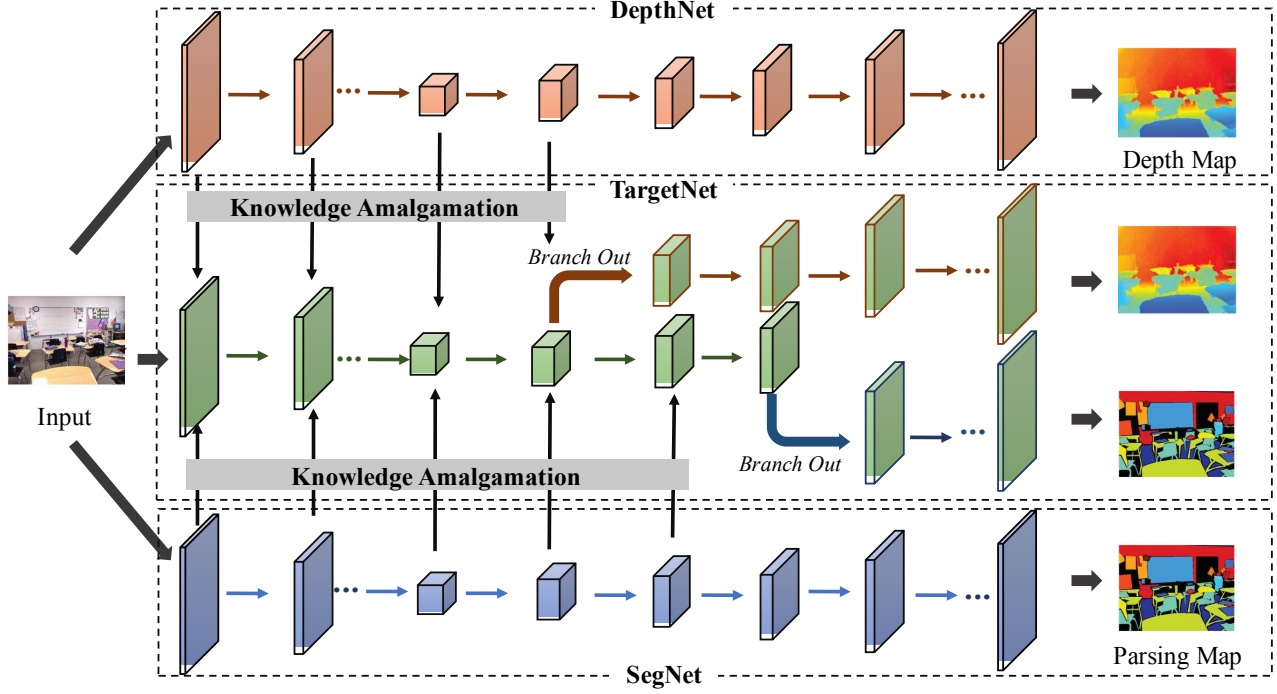


Figure 1. The proposed knowledge amalgamation approach for semantic segmentation and depth estimation with the encoder-decoder network. The *branch out* takes place in the decoder part of the target network.

a student model that works on the same problem as the teacher does, the proposed knowledge amalgamation methods learn a student model that acquires the super knowledge of all the heterogeneous-task teachers, each of which specializes in a different domain. Once trained, the student is therefore competent to simultaneously handle various tasks covering the expertise of all teachers.

3. Problem Definition

The problem we address here is to learn a compact student model, which we call TargetNet, without human-labeled annotations, that amalgamates knowledge and thus simultaneously handles several different tasks by learning from two or more teachers, each of which specializes in only one task. Specifically, we focus on two important pixel-prediction tasks, depth estimation and scene parsing, and describe the proposed strategy for knowledge amalgamation in Sec. 5. We also show that the proposed training strategy can be readily extended to train a student that handles three or even more tasks jointly.

For the rest of this paper, we take a *block* to be the part of a network bounded by two pooling layers, meaning that in each network all the features maps within a block have the same resolution. We use I to denote an input image, and use F_s^n and F_d^n to respectively denote the feature maps in the n -th block of the pre-trained segmentation teacher network (SegNet) and the depth prediction teacher network (DepthNet); we also let S denote the final pre-

diction of SegNet and D denote that of DepthNet, and let S_i and D_i denote the respective predictions at pixel i . In demonstrating the feasibility of amalgamation from more than two teachers, we look at a third pixel-wise prediction task, surface normal estimation, for which a surface-normal teacher network (NormNet) is pre-trained. We use M to denote its predicted normal map with M_i being its i -th pixel estimation.

4. Pre-Trained Teacher Networks

We describe here the pre-trained teacher networks, SegNet, DepthNet, and NormNet, based on which we train our student model, TargetNet.

Admittedly, we assume for now that the teacher networks share the same encoder-decoder architecture, but not restricted to any specific design. This assumption might sound somehow strong but in fact not, as many state-of-the-art pixel-prediction models deploy the encoder-decoder architecture. Knowledge amalgamation from multiple teachers with arbitrary architectures is left for future work.

Scene Parsing Teacher (SegNet). The goal of scene parsing is to assign a label denoting a category to each pixel of an image. In our case, we adopt the state-of-the-art SegNet [1], which has a encoder-decoder architecture, as the scene parsing teacher. The pixel-wise loss function is taken to be

$$\mathcal{L}_{\text{seg}}(S^{\text{gt}}, S^*) = \frac{1}{N} \sum_i \ell(S_i^*, S_i^{\text{gt}}) + \lambda R(\theta), \quad (1)$$

where S_i^* is the prediction for pixel i , S_i^{gt} is the ground truth, $\ell(\cdot)$ is the cross-entropy loss, N is the total pixel number of the input image, and R is the L2-norm regularization term.

Since we do not have access to the human-labeled annotations, we take the prediction S_i , converted from S_i^* by one-hot coding, as the supervision for training the TargetNet.

Depth Estimation Teacher (DepthNet). As another well-studied pixel-level task, depth estimation aims to predict for each pixel a value denoting the depth of an object with respect to the camera. The major difference between scene parsing and depth estimation is, therefore, the output of the former task is a discrete label while that of the latter is a continuous and positive number.

We transfer depth estimation to a classification problem, which has been proved to be effective [23], by quantizing the depth values to N_d bins, each of which has a length of l . For each bin b , the network predicts $p(b|x(i)) = \exp^{r_i} / \sum_b \exp^{r_{i,b}}$, the probability of having an object at the center of that bin, where $r_{i,b}$ is the response of network at pixel i and bin b . The continuous depth value D_i is then computed as:

$$D_i = \sum_{b=1}^{N_d} b \times l \times p(b|x(i)). \quad (2)$$

The loss function for depth estimation is computed taken to be:

$$\mathcal{L}_{\text{depth}}(D^{\text{gt}}, D) = \frac{1}{N} \sum_i (d_i)^2 - \frac{1}{2N^2} \left(\sum_i d_i \right)^2, \quad (3)$$

where $d = D^{\text{gt}} - D$, and N is the total number of valid pixels (we mask out pixels where the ground truth is missing). The prediction of the DepthNet, D , is utilized as the supervision to train the TargetNet.

Surface-Normal Prediction Teacher (NormNet). Given an input image, the goal of surface normal prediction is to estimate a surface-normal vector (x, y, z) for each pixel. The loss function to train a surface-normal-prediction teacher network (NormNet) is taken to be

$$\mathcal{L}_{\text{norm}}(M^{\text{gt}}, M) = -\frac{1}{N} \sum_i M_i \cdot M_i^{\text{gt}} = -\frac{1}{N} M \cdot M^{\text{gt}}, \quad (4)$$

where M , M^{gt} are respectively the prediction of NormNet and the ground truth, and M_i , M_i^{gt} are those at pixel i .

5. Proposed Method

In this section, we describe the proposed approach to learning a compact student network, TargetNet. We introduce a novel strategy that trains the student intertwined with the teachers. At the heart of our approach is a block-wise learning method shown in Fig. 1 that learns the parameters

of the student network, by “projecting” the amalgamated knowledge of the student to each teacher’s expertise domain for computing the loss and updating the parameters, as depicted in Fig. 2

In what follows, we start from the amalgamation of the SegNet and DepthNet, and then extend to the amalgamation of multiple networks including surface normal prediction networks (NormNet).

5.1. Learning from Two Teachers

Given the two pre-trained teacher networks, SegNet and DepthNet described in Sec. 4, we train a compact student model of a similar encoder-decoder architecture, except that the decoder part *eventually* comprises two streams, one for each task as shown in Fig. 1. In principle, all standard backbone CNNs like AlexNet [15], VGG [34] and ResNet [10] can be employed for constructing such encoder-decoder architecture. In our implementation, we choose the one of VGG, as done for [1].

We aim to learn a student model small enough in size so that it is deployable on edge systems, but not smaller as it is expected to master the expertise of both teachers. To this end, for each block of the TargetNet, we take its feature maps to be of the same size as those in the corresponding block of the teachers.

The encoder and the decoder of TargetNet play different roles in the tasks of joint parsing and depth estimation. The encoder part works as a feature extractor to derive discriminant features for both tasks. The decoder, on the hand other, is expected to “project” or transfer the learned features into each task domain so that they can be activated in different ways for the specific task flow.

Despite the TargetNet eventually has two output streams, it is initialized with the same encoder-decoder architecture as those of the teachers. We then train the TargetNet and finally branch out the two streams for the two tasks, after which the initial decoder blocks following the branch-out are removed. The overall training process is summarized as follows:

- **Step 1:** Initialize TargetNet with the same architecture as those of the teachers, as described in 5.1.1.
- **Step 2:** Train each block of TargetNet intertwined with the teachers, shown in Fig. 2 and described in Sec. 5.1.2.
- **Step 3:** Decide where to branch out on TargetNet, as described in Sec. 5.1.3.
- **Step 4:** Take the corresponding blocks from the teachers as the branched-out blocks of the student; remove all the initial blocks following the block where the last branch out takes place; fine-tune TargetNet.

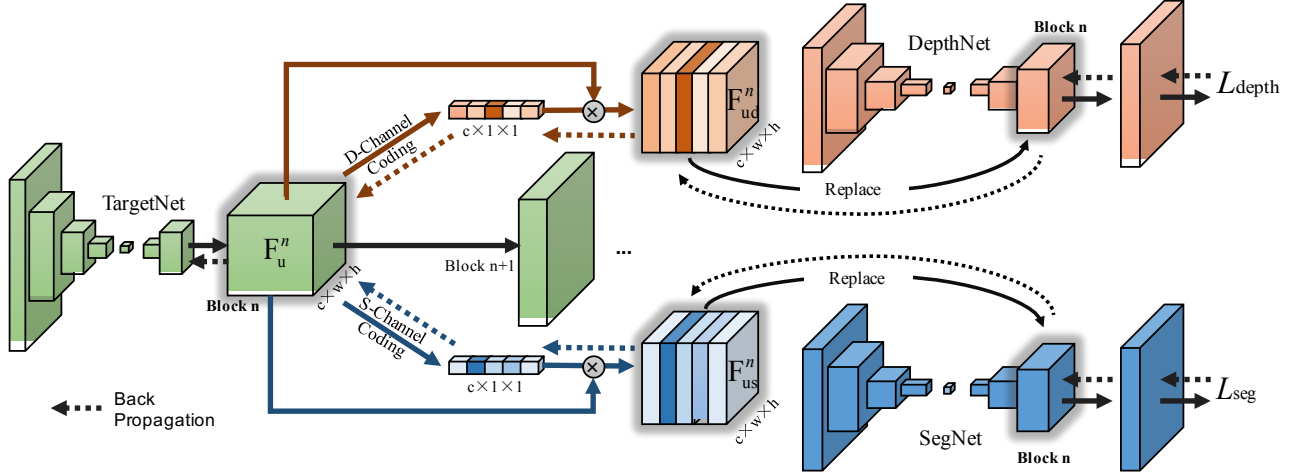


Figure 2. The proposed knowledge amalgamation module within block n of TargetNet. This operation is repeated for every block.

In what follows, we describe the architecture of the student network, its loss function and training strategy, as well as the branch out strategy.

5.1.1 TargetNet Architecture

The TargetNet is initialized with the same encoder-decoder architecture as those of the teachers, where the structures of the encoder and the decoder are symmetric, as shown in Fig. 1. Each block in the encoder comprises 2 or 3 convolutional layers with 3×3 kernel size, followed with a 2×2 non-overlapping max pooling layer, while in the decoder a pooling layer is replaced by an up-sampling one.

We conduct knowledge amalgamation for each block of TargetNet, by learning the parameters intertwined with the teachers. Let F_u^n denote the amalgamated feature maps at block n of TargetNet. We expect F_u^n to encode both the parsing and depth information, obtained from the two teachers. To allow the F_u^n to interact with the teachers and to be updated, we introduce two channel-coding branches, termed *D-Channel Coding* and *S-Channel Coding*, respectively for depth estimation and scene parsing as depicted in Fig. 2. Intuitively, F_u^n can be thought as a container of the whole set of features, and can be projected or transformed into the two task domains via the two channels. Here, we use F_{ud} and F_{us} to respectively denote the obtained features after passing F_u^n through the D-Channel and S-Channel Coding, in other words, the projected features in the two task domains.

The architecture of the two channel coding is depicted in Fig. 3. Modified from the channel attention module by [12], it consists of a global pooling layer and two fully connected layers, and is very light in size. In fact, it increases the total number of parameters by less than 4% only, leading to very low computation cost.

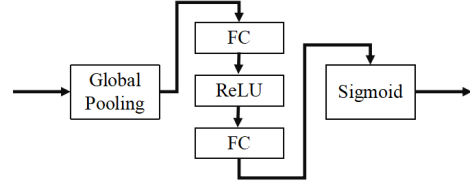


Figure 3. The architecture of the S-Channel Coding and D-Channel Coding.

5.1.2 Loss Function and Training Strategy

To learn the amalgamated features F_u^n at the n -th block of TargetNet, we feed the unlabelled samples into both teachers and TargetNet, in which way we obtain the features of two teachers and the initial ones of the Target. Let F_d^n , F_s^n denote the obtained features at the n -th block of the teacher networks, DepthNet and SegNet. One option to learn F_u^n is to minimize the loss between its projected features F_{ud}^n , F_{us}^n and the corresponding features F_d^n , F_s^n , as follows,

$$\mathcal{L}_u = \lambda_1 \|F_{ud}^n - F_d^n\|^2 + \lambda_2 \|F_{us}^n - F_s^n\|^2, \quad (5)$$

where λ_1 and λ_2 are weights that balance the depth estimation and scene parsing. With this loss function, parameter updates take place within block n and the connecting coding branches, during which process blocks 1 to $n-1$ remain unchanged.

The loss function of Eq. (5) is intuitively straightforward. However, using this loss to train each block in TargetNet turns out to be time- and effort-consuming. This is because for training each block, we have to adjust the weights λ_1 , λ_2 and the termination conditions, as the magnitudes of the feature maps and convergence rates vary block by block.

We therefore turn to another alternative that learns the features intertwined with the teachers, in aim to alleviate the cumbersome tuning process. For amalgamation in block n , we first obtain F_{us}^n by passing the amalgamated features F_u^n through the S-channel coding. We then *replace* the features

F_s^n at the n -th block of SegNet with F_{us}^n , and obtain the consequent prediction \hat{S} from the SegNet with F_{us}^n being its features. This process is repeated in the same way to get the predicted depth map \hat{D} from the DepthNet with F_{ud}^n being its features.

In this way, we are able to write the loss as a function comprising only the final predictions \hat{D} , \hat{S} and the predictions by original teachers D, S , as follows,

$$\mathcal{L}_u = \lambda_1 \mathcal{L}_{\text{depth}}(D, \hat{D}) + \lambda_2 \mathcal{L}_{\text{seg}}(S, \hat{S}), \quad (6)$$

where the λ_1, λ_2 are fixed for all the blocks in TargetNet during training, and $\mathcal{L}_{\text{depth}}(\cdot)$ and $\mathcal{L}_{\text{seg}}(\cdot)$ are respectively those defined in Eqs. (3) and (1).

5.1.3 Branch Out

As scene parsing and depth estimation are two closely related tasks, it is non-trivial to decide where to branch out the TargetNet into separate task-specific streams to achieve the optimal performances on both tasks simultaneously. Unlike conventional multi-branch models that choose to branch out at the boarder of encoder and decoder, we explore another option, which we find more effective. After training the N blocks of TargetNet using the loss function of Eq. (6), we also acquire the final losses for each block $\{\mathcal{L}_{\text{seg}}^1, \mathcal{L}_{\text{seg}}^2, \dots, \mathcal{L}_{\text{seg}}^N\}$ and $\{\mathcal{L}_{\text{depth}}^1, \mathcal{L}_{\text{depth}}^2, \dots, \mathcal{L}_{\text{depth}}^N\}$. The blocks for branching out, p_{seg} and p_{depth} , are taken to be:

$$\begin{aligned} p_{\text{seg}} &= \arg \min_n \mathcal{L}_{\text{seg}}^n \\ p_{\text{depth}} &= \arg \min_n \mathcal{L}_{\text{depth}}^n \end{aligned} \quad (7)$$

where we set $N/2 < n \leq N$ to allow the branching out take place only within the decoder.

Once the branch-out blocks p_{seg} and p_{depth} are decided, we remove those initial decoder blocks succeeding the last branch-out block. In the example shown in Fig. 1, scene parsing branches out later than depth estimation, and thus all the initial blocks after p_{seg} are removed. We then take the corresponding decoder blocks of the teachers as the branched-out blocks for TargetNet, as depicted by the upper- and lower-green stream in Fig. 1. In this way, we derive the final TargetNet architecture of one encoder and two decoders sharing some blocks, and fine-tune this model.

5.2. Learning from More Teachers

In fact, the proposed training scheme is flexible and not limited to amalgamating knowledge from two teachers only. We show here two approaches, an offline approach and an online one, to learn from three teachers. We take surface-normal prediction, another widely studied pixel-prediction task, as an example to illustrate the three-task amalgamation. The proposed two approaches can be directly applied to amalgamation from an arbitrary number of teachers.

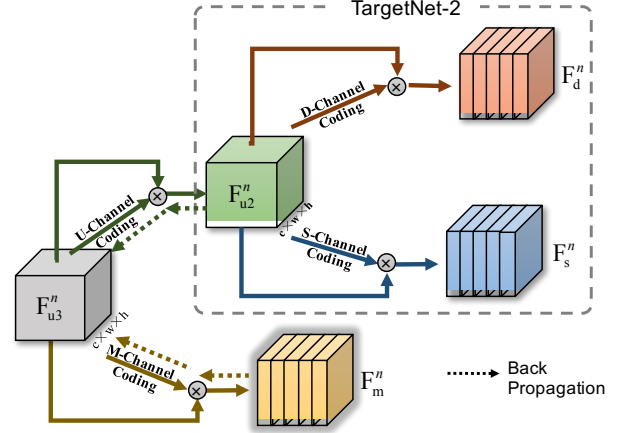


Figure 4. Knowledge amalgamation with TargetNet-2 and NormNet. The part in dash frame is from TargetNet-2 and keeps unchanged during training TargetNet-3.

The offline approach for three-task amalgamation is quite straightforward. The block-wise learning strategy described in Sec. 5.1.2 can be readily generalized here. Now instead of two coding channels we have three, wherein the third one M -channel coding transfers the amalgamated knowledge to the normal-prediction task. The loss function is then taken to be

$$\mathcal{L}_{u3} = \lambda_1 \mathcal{L}_{\text{depth}}(D, \hat{D}) + \lambda_2 \mathcal{L}_{\text{seg}}(S, \hat{S}) + \lambda_3 \mathcal{L}_{\text{norm}}(M, \hat{M}), \quad (8)$$

where λ_1, λ_2 and λ_3 are the balancing weights.

The online approach works in an incremental manner. Assume we have already trained a TargetNet for parsing and depth estimation, which we call TargetNet-2 now for clarity. We would like the student to amalgamate the normal-prediction knowledge as well, and we call this three-task student TargetNet-3. The core idea of this online approach is to treat TargetNet-2 itself as a pre-trained teacher, and then conduct amalgamation of TargetNet-2 and NormNet in the same way as done for the two-teacher amalgamation described in Sec. 5.1.2. Let F_{u2}^n denote the features at block n of TargetNet-2, and let $\{D_u, S_u\}$ denote its predictions of depths and segmentation. Also, let F_{u3}^n denote the features to be learned at the n -th block of TargetNet-3.

We connect F_{u3}^n and F_{u2}^n via the U -channel coding as depicted in Fig. 4. On the normal-prediction side, we pass F_{u3}^n through the M -channel coding to produce the projected features F_m^n , replace the corresponding features of NormNet, and obtain the norm prediction \hat{M} . On the TargetNet-2 side, we project F_{u3}^n through U -channel coding to obtain F_{u2}^n , repeat the same process and obtain predictions $\{\hat{S}_u, \hat{D}_u\}$. The loss function is then taken to be

$$\mathcal{L}_{u3} = \lambda_1 \mathcal{L}_{\text{norm}}(\hat{M}, M) + \lambda_2 \mathcal{L}_{u2}, \quad (9)$$

where λ_1 and λ_2 are the weights. As shown in Fig. 4, we freeze TargetNet-2 during this process since it is treated as

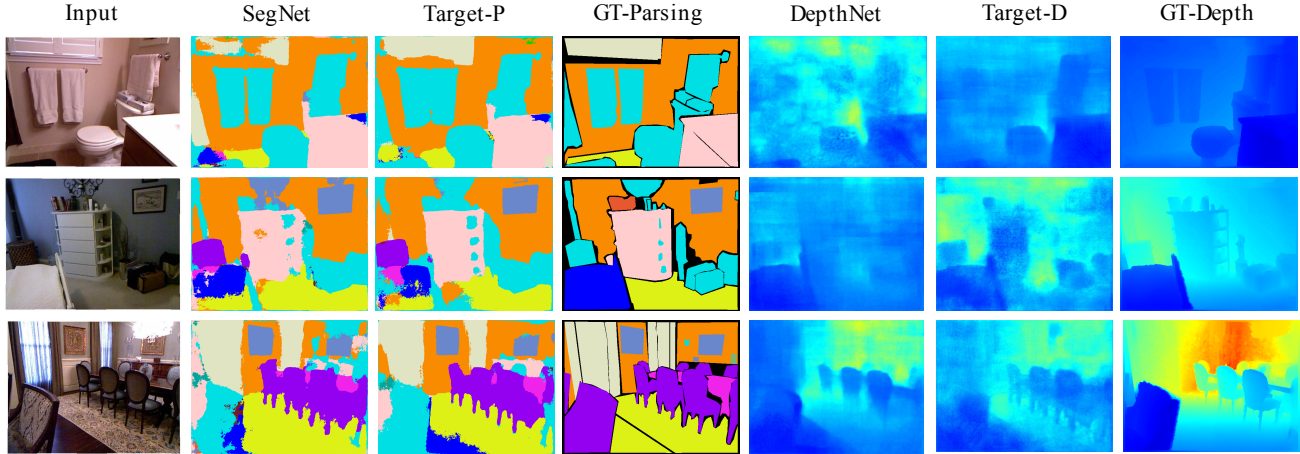


Figure 5. Qualitative results on scene parsing and depth estimation on NYUDv2. For both tasks, we compare the results of the teachers with those of the student, with Target-P and Target-D denoting the results of the student on scene parsing and depth estimation respectively.

a teacher. Thus, Eq. (9) will update the parameters of U-channel coding but not those within TargetNet-2.

6. Experiments and Results

Here we provide our experimental settings and results. More results can be found in our supplementary material.

6.1. Experimental Settings

Datasets. The NYUDv2 dataset [33] provides 1,449 labeled indoor-scene RGB images with both parsing annotations and Kinect depths, and 407,024 unlabeled ones, of which 10,000 are used to train TargetNet. Besides the scene parsing and depth estimation task, we also train TargetNet-3 that conducts surface normal prediction on this dataset. The surface normal ground truths are computed from the depth maps using the neighboring pixels’ cross products.

Cityscapes [3] is a large-scale dataset for semantic urban scene understanding. It provides RGB images and stereo disparity ones, collected over 50 different cities spanning several months, with overall 19 semantic classes being annotated. The finely annotated part consists of 2,975 images for training, 500 for validation, and 1,525 for test. Disparity images provided by Cityscape are used as depth maps, with bilinear interpolation filling the holes.

Evaluation Metrics. To qualitatively evaluate the depth-estimation performance, we use the standard metrics as done in [5], including absolute relative difference (abs rel), squared relative difference (sqr rel) and the percentage of relative errors inside a certain threshold t .

For scene parsing, we use both the pixel-wise accuracy (Pixel Acc.) and the mean Intersection over Union (mIoU).

For surface normal prediction, performance is measured by the same matrix as [4]: the mean and median angle from the groundtruth across all unmasked pixels, as well as the percent of vectors whose angles fall within three thresholds.

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 [20], and set the base learning rate to 0.005, the power to 0.9, and the weight decay to $4e - 6$. We perform data augmentation with random-size cropping to 640×400 for all datasets. Due to limited physical memory on GPU cards and to assure effectiveness, we set the batch size to be 8 during training. The encoder and decoder of TargetNet are initialized with parameters pre-trained on ImageNet. In total 2 epochs are used for training each block of the Target on both NYUDv2 and Cityscape.

6.2. Experimental Results

Performance on NYUDv2. We show in Tab. 1 the qualitative results of the TeacherNet (SegNet and DepthNet) and those of the student network with different branch-out blocks, denoted by Decoder_b1-5, on the testset of NYUDv2. For example, Decoder_b1 corresponds to the student network branched out at decoder block 1. Our final TargetNet branches out depth estimation at block 4 and scene parsing at block 5, which give us the best performances for the two tasks. As can be seen, the student networks consistently outperform the teachers in terms of all evaluation measures, which validates the effectiveness of the proposed method. As for the network scale, the more blocks of the features are amalgamated, the smaller TargetNet is, which means that branching out at a later stage of the decoder leads to smaller networks, as shown by the number of parameters of Decoder_b1-5. The final student network, which branches out at blocks 4 and 5 respectively for depth and scene, yields a size that is about half of the teachers.

The qualitative results are displayed in Fig. 5, where the results of student, Target-P and Target-D, are more visually plausible than those of the teacher networks.

For surface normal prediction, we train the TargetNet-

Table 1. Comparative results of the teacher networks and the student with different branch-out blocks on the NYUDv2 dataset. Decoder.*bn* denotes the TargetNet that branches out at the block *n* of the decoder, and TeacherNet contains SegNet and DepthNet. The final student network branches out the depth estimation at block 4, denoted by Target-D, and scene parsing at block 5 denoted by Target-P.

Method	Params	mean IoU	Pixel Acc.	abs rel	sqr rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
TeacherNet	~55.6M	0.448	0.684	0.287	0.339	0.569	0.845	0.948
Decoder.b1	~36.9M	0.447	0.684	0.276	0.312	0.383	0.753	0.939
Decoder.b2	~31.0M	0.451	0.684	0.259	0.275	0.448	0.799	0.952
Decoder.b3	~28.3M	0.451	0.684	0.260	0.277	0.448	0.796	0.951
Decoder.b4 (Target-D)	~28.0M	0.452	0.683	0.252	0.257	0.544	0.847	0.959
Decoder.b5 (Target-P)	~27.8M	0.458	0.687	0.256	0.266	0.459	0.810	0.956

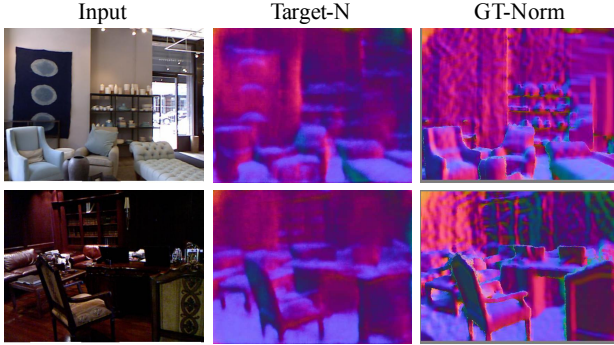


Figure 6. Qualitative results on surface normal prediction on NYUDv2. The Target-N column corresponds to the results of the proposed TargetNet-3.

Table 2. Comparative results of the TeacherNet (SegNet, DepthNet and NormNet), Target-2 and Target-3 on NYUDv2.

Method	mIOU	rel diff		Angle Dist.		With t° Deg.		
		abs	sqr	Mean	Median	11.25°	22.5°	30°
TeacherNet	0.448	0.287	0.339	37.88	36.96	0.236	0.450	0.567
TargetNet-2	0.458	0.252	0.257	—	—	—	—	—
TargetNet-3	0.459	0.243	0.255	35.45	34.88	0.237	0.448	0.585

Table 3. Comparative results of TargetNet and state-of-the-art methods on scene parsing and depth estimation on NYUDv2.

Method	Params	mIOU	PA	abs rel	sqr rel
Xu et al. [37]	—	—	—	0.121	—
FCN-VGG16 [21]	~134M	0.292	0.600	—	—
RefineNet-L152 [24]	~62M	0.444	—	—	—
RefineNet-101 [18]	~118M	0.447	0.686	—	—
Gupta et al. [8]	—	0.286	0.603	—	—
Arsalan et al. [23]	—	0.392	0.686	0.200	0.301
PADNet-Alexnet [36]	>50M	0.331	0.647	0.214	—
PADNet-ResNet [36]	>80M	0.502	0.752	0.120	—
TargetNet	~28M	0.459	0.688	0.124	0.203

3 described in Sec. 5.2 in the offline manner, and show the results in Tab. 2. With the surface normal information amalgamated in TargetNet-3, the accuracies of scene parsing and depth estimation increase even further and exceed those of the TargetNet-2. In particular, compared to the enhancement of scene parsing, that of depth estimation is more significant due to its tight relationship with surface normal. The visual results of the surface normal prediction are depicted in Fig. 6.

We also compare the performance of TargetNet with those of the state-of-the-art models. The results are shown in Tab. 3. TargetNet is trained with three tasks. With the smallest size and no access to human-labelled annotations,

Table 4. Comparative results of the TeacherNet (SegNet and DepthNet) and the student on the Cityscape dataset. Target-D and Target-P denote the results of TargetNet for parsing and depth estimation respectively.

Method	mean IOU	Pixel Acc.	abs rel	sqr rel
TeacherNet	0.521	0.875	0.289	5.803
Target-P	0.535	0.882	0.240	3.872
Target-D	0.510	0.882	0.224	3.509

TargetNet yields results better than all but one compared methods, PADNet-ResNet [36], for which the size is almost three times as that of TargetNet.

Performance on Cityscape. We also conduct the proposed knowledge amalgamation on the Cityscape dataset, and show the quantitative results in Tab. 4. We carry out the experiments in the same way as done for NYUDv2. Results on this outdoor dataset further validate the effectiveness of our approach, where the amalgamated network again both surpasses the teachers on their own domains.

7. Conclusion

In this paper, we investigate a novel knowledge amalgamation task, which aims to learn a versatile student model from pre-trained teachers specializing in different application domains, without human-labelled annotations. We start from a pair of teachers, one on scene parsing and the other on depth estimation, and propose an innovative strategy to learn the parameters of the student intertwined with the teacher. We then present two options to generalize the training strategy for more than two teachers. Experimental results on several benchmarks demonstrate that student model, once learned, outperforms the teachers in their own expertise domains. In our future work, we will explore knowledge amalgamation from teachers with different architectures, in which case the main challenge is to bridge the semantic gaps between the feature maps of the teachers.

Acknowledgement

This work is supported by National Key Research and Development Program (2016YFB1200203), National Natural Science Foundation of China (61572428, U1509206), Fundamental Research Funds for the Central Universities (2017FZA5014), Key Research and Development Program of Zhejiang Province (2018C01004), and the Startup Funding of Stevens Institute of Technology.

References

- [1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2481–2495, 2017.
- [2] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems*, pages 742–751, 2017.
- [3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.
- [4] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *International Conference on Computer Vision*, pages 2650–2658, 2015.
- [5] 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.
- [6] Jiyang Gao, Zijian Guo, Zhen Li, and Ram Nevatia. Knowledge concentration: Learning 100k object classifiers in a single cnn. *arXiv: Computer Vision and Pattern Recognition*, 2017.
- [7] Jean-Yves Guillemaut and Adrian Hilton. Space-time joint multi-layer segmentation and depth estimation. In *3D Imaging, Modeling, Processing, Visualization and Transmission*, pages 440–447, 2012.
- [8] Saurabh Gupta, Ross B Girshick, Pablo Arbelaez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. *European Conference on Computer Vision*, pages 345–360, 2014.
- [9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *International Conference on Computer Vision*, pages 2980–2988, 2017.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *European Conference on Computer Vision*, pages 630–645, 2016.
- [11] Geoffrey E Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *Neural Information Processing Systems*, 2015.
- [12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *Computer Vision and Pattern Recognition*, 2018.
- [13] Mingkun Huang, Yongbin You, Zhehuai Chen, Yanmin Qian, and Kai Yu. Knowledge distillation for sequence model. *Proc. Interspeech 2018*, pages 3703–3707, 2018.
- [14] Mateusz Kozinski, Loic Simon, and Frederic Jurie. An adversarial regularisation for semi-supervised training of structured output neural networks. *Neural Information Processing Systems*, 2017.
- [15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.
- [16] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. *International Conference on 3d Vision*, pages 239–248, 2016.
- [17] Bo Li, Yuchao Dai, and Mingyi He. Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference. *Pattern Recognition*, 83:328–339, 2018.
- [18] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. *Computer Vision and Pattern Recognition*, pages 5168–5177, 2017.
- [19] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. *Computer Vision and Pattern Recognition*, pages 5162–5170, 2015.
- [20] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv: Computer Vision and Pattern Recognition*, 2015.
- [21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.
- [22] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob J Verbeek. Semantic segmentation using adversarial networks. *Neural Information Processing Systems*, 2016.
- [23] Arsalan Mousavian, Hamed Pirsiavash, and Jana Kosecka. Joint semantic segmentation and depth estimation with deep convolutional networks. *International Conference on 3d Vision*, pages 611–619, 2016.
- [24] Vladimir Nekrasov, Chunhua Shen, and Ian D Reid. Lightweight refinenet for realtime semantic segmentation. *British Machine Vision Conference*, 2018.
- [25] Haesol Park and Kyoung Mu Lee. Joint estimation of camera pose, depth, deblurring, and super-resolution from a blurred image sequence. *International Conference on Computer Vision*, pages 4623–4631, 2017.
- [26] Pedro H O Pinheiro, Ronan Collobert, and Piotr Dollar. Learning to segment object candidates. *Neural Information Processing Systems*, pages 1990–1998, 2015.
- [27] Pedro H O Pinheiro, Tsungyi Lin, Ronan Collobert, and Piotr Dollar. Learning to refine object segments. *European Conference on Computer Vision*, pages 75–91, 2016.
- [28] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *International Conference on Learning Representations*, 2015.
- [29] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. Learning depth from single monocular images. In *Advances in Neural Information Processing Systems*, pages 1161–1168, 2006.
- [30] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Learning 3-d scene structure from a single still image. In *International Conference on Computer Vision*, pages 1–8, 2007.
- [31] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. *Computer Vision and Pattern Recognition*, 1:195–202, 2003.

- [32] Chengchao Shen, Xinchao Wang, Jie Song, Li Sun, and Mingli Song. Amalgamating knowledge towards comprehensive classification. In *AAAI Conference on Artificial Intelligence*, 2019.
- [33] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *European Conference on Computer Vision*, 2012.
- [34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.
- [35] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi and weakly supervised semantic segmentation using generative adversarial network. *arXiv: Computer Vision and Pattern Recognition*, 2017.
- [36] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. *Computer Vision and Pattern Recognition*, pages 675–684, 2018.
- [37] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. *Computer Vision and Pattern Recognition*, pages 161–169, 2017.
- [38] Jingwen Ye, Zunlei Feng, Yongcheng Jing, and Mingli Song. Finer-net: Cascaded human parsing with hierarchical granularity. In *International Conference on Multimedia and Expo*, pages 1–6, 2018.
- [39] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Computer Vision and Pattern Recognition*, 2017.
- [40] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.