Multi-Task Learning with Knowledge Distillation for Dense Prediction

Yangyang Xu[1] Yibo Yang[3] Lefei Zhang[1,2*]

[1] Institute of Artificial Intelligence and School of Computer Science, Wuhan University, Wuhan, China
[2] Hubei Luojia Laboratory, Wuhan, China
[3] King Abdullah University of Science and Technology, Jeddah, Saudi Arabia

yangyangxu@whu.edu.cn, yibo.yang@kaust.edu.sa, zhanglefei@whu.edu.cn

Abstract

While multi-task learning (MTL) has become an attractive topic, its training usually poses more difficulties than the single-task case. How to successfully apply knowledge distillation into MTL to improve training efficiency and model performance is still a challenging problem. In this paper, we introduce a new knowledge distillation procedure with an alternative match for MTL of dense prediction based on two simple design principles. First, for memory and training efficiency, we use a single strong multi-task model as a teacher during training instead of multiple teachers, as widely adopted in existing studies. Second, we employ a less sensitive Cauchy-Schwarz (CS) divergence instead of the Kullback–Leibler (KL) divergence and propose a CS distillation loss accordingly. With the less sensitive divergence, our knowledge distillation with an alternative match is applied for capturing inter-task and intra-task information between the teacher model and the student model of each task, thereby learning more "dark knowledge" for effective distillation. We conducted extensive experiments on dense prediction datasets, including NYUD-v2 and PASCAL-Context, for multiple vision tasks, such as semantic segmentation, human parts segmentation, depth estimation, surface normal estimation, and boundary detection. The results show that our proposed method decidedly improves model performance and the practical inference efficiency.

1. Introduction

Multi-task learning (MTL) has become an increasingly popular approach in the field of computer vision, where the objective is to train a single model to perform multiple tasks simultaneously. MTL can provide several advantages over traditional single-task learning, including improved efficiency and generalization. First, the shared feature representations can be learned more efficiently than task-specific representations, reducing the overall training time compared to training multiple models independently. Second, the shared feature representations learned across tasks can capture more generalizable information in the data, leading to improved performance on all tasks. For those reasons, the MTL approach has been used extensively in various machine learning problems, such as natural language processing [11], computer vision [3] and speech recognition [4]. Specifically, in this paper, we focus on dense (pixel-wise) prediction vision tasks [49, 15, 23, 41, 20, 21, 29, 48], such as semantic segmentation, instance segmentation, depth estimation, surface normal estimation, saliency estimation, object detection and boundary detection from images.

A march of works [32, 46, 9, 3, 52, 2, 56, 53] aims to develop novel MTL architectures and construct efficient shared representation in the multi-taskdense prediction field, which leverages the encoder-decoder architecture. In these frameworks, an encoder is considered to generate a shared feature and then use a decoder to perform multi-task of dense predictions. In the pursuit of better performance, current MTL models are often designed to be deeper and wider, resulting in increasingly larger models. [3, 56] and [53, 24] demonstrate that better performance can be obtained by utilizing a larger backbone network. However, such a heavy model can be more demanding for computation and storage. It has been challenging to design an effective MTL framework that can learn these tasks efficiently. In this paper, we pose and study the question, *how can the knowledge from a large size MTL model be transferred to a small MTL model without increasing its size?*

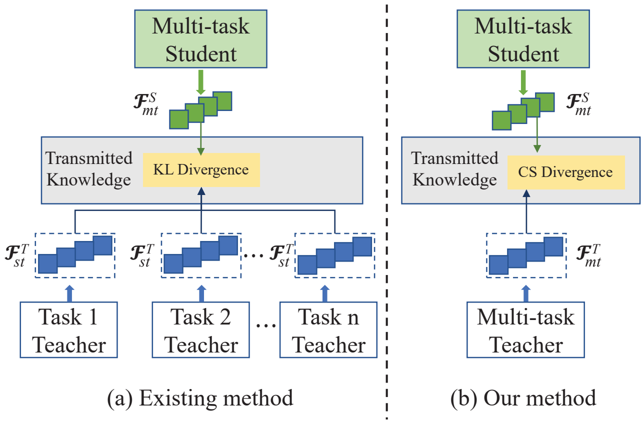
Recently, knowledge distillation (KD) [13] has been explored as a method to improve the MTL of dense prediction tasks, such as semantic segmentation, depth estimation, and surface normal prediction. Some studies [30, 12, 16, 1] leverage the unique logit of each task to provide task-specific information to the student model during knowledge transfer. The essence of knowledge distillation lies

[*] Corresponding author.

in translating knowledge from the teacher model (large model) to the student model (small model) by mimicking the teacher model's outputs. Typically, knowledge distillation methods use the logits matching by Kullback–Leibler (KL) divergence [13] between the probability distributions of the teacher and the probability distributions of the student. In this way, during training, the student model can be guided by more valuable information signals from the teacher model, therefore, is expected to have a more promising performance than training alone. This approach has shown promising results in improving model performance and generalization and speeding up convergence in MTL. For instance, We found that some works [22, 35, 12, 30, 17] attempt to use knowledge distillation to transfer the knowledge from teacher to student for multiple vision tasks in MTL. We, in particular, identify two technical challenges. 1) Most previous approaches [22, 35, 17] must train a task-specific model for each task, then load the trained task-specific models. 2) Exact logit matching of teacher and student predictions with KL divergence can interfere with the training of the student model and is sensitive to outliers, leading to less effective knowledge distillation.

To address these challenges, we explore knowledge distillation in multi-task learning of dense prediction tasks. We present a novel framework that leverages task-specific guidance to enable effective knowledge transfer. In Figure 1, we show the difference between the existing and our methods. Since using multiple single-task models as teacher models would require an inordinate amount of memory, we only load a single multi-task model as a teacher model instead of loading multiple teacher models. We opt for one strong teacher model to reduce computational and memory costs instead of using multiple teacher models that perform a load of each task separately.

We take inspiration from mathematical statistics methods from other domains to reduce the uncertainty in students' prediction when using KL divergence (see Figure 1). We propose a novel knowledge distillation with Cauchy-Schwarz (CS) divergence to replace the KL divergence. In addition, during the logit matching, inter-task and intra-task information are transferred from teacher to student. Specifically, we gather the corresponding predicted probabilities in a batch for all tasks, then transfer the inter-task information from teacher to student. For each task, we gather the corresponding predicted probabilities of all classes in a batch, then transfer the intra-task information from teacher to student. Our proposed knowledge distillation with the alternative match is less sensitive to small probabilities and can account for uncertainty in the student model predictions. To further close the computational and memory cost gap, we only use one multi-task teacher model during training. This MTL knowledge distillation procedure provides multiple training settings.



(a) Existing method

(b) Our method

Figure 1: Difference between existing KD methods and our KD method. $\mathcal{F}^T_{st}$ means the single-task (st) teacher (T) model output. $\mathcal{F}^T_{mt}$ means the multi-task (mt) teacher model output. $\mathcal{F}^S_{mt}$ means the multi-task (mt) student (S) model output. n denotes the number of tasks.

Through extensive experiments on several publicly available dense prediction datasets (*i.e.,* NYUD-v2 and PASCAL-Context), we compare our results with state-of-the-art methods to demonstrate the effectiveness of our framework.

In summary, our work makes the following contributions:

- We introduce a new procedure based on knowledge distillation with an alternative match named KDAM, which leverages inter-task and intra-task information. It is less sensitive to small probabilities and can account for uncertainty in the student model predictions.
- Our new distillation procedure aims at reproducing computational and memory costs by loading a strong multi-task model as a teacher to guide student learning.
- We conduct experiments using our KDAM on two MTL of dense prediction datasets, showing the superiority under different experiment settings.

2. Related Work

Multi-task Learning of Dense Prediction. Multi-task learning for dense prediction [37, 50, 58, 31, 10, 9, 3, 52, 56, 53, 33, 28, 25, 57, 26, 27] has been an active research area in the computer vision community, with many studies exploring various aspects of joint learning for tasks such as semantic segmentation, depth estimation, surface normal estimation, saliency estimation and boundary detection. A survey work [45] shows that MTL is mainly divided into two categories: encoder-based and decoder-based architectures. The encoder-based methods, such as [19, 44, 54], use a shared representation from an encoder to input task-specific heads and perform multiple task predictions. NDDR-CNN Network [10] is an encoder-focused MTL model which enables automatic feature fusing at every layer from different tasks and reduces channel dimension by processing the features with a 1×1 convolutional layer before feeding the result to the next layer. Their success heavily depends on the encoder (*i.e.,* backbone network) to learn a strong shared representation [47, 59]. However, the encoder-based methods lack task interaction information, failing to capture commonalities and differences among tasks. The decoder-based methods, such as [50, 31, 3, 56, 2], focus on improving each task by repeatedly refining the predictions through cross-task interaction and improving each task's performance. Unlike encoder-based, decoder-based architectures also exchange information during the decoding. MTFormer [51] performs cross-task reasoning and designs the cross-task attention mechanism to achieve effective feature propagation among tasks, resulting in performance improvement in MTL. [52, 56] introduces new techniques for training deeper and wider via Transformers in MTL, allowing for more efficient and accurate MTL performance. Their success depends on a strong backbone network and all relevant task interactions from different dimensions. In addition, [40] introduces a controllable dynamic multi-task architecture for dynamically adjusting the weighting of loss of each task, which allows matching the desired task preference as well as the resource constraints. We propose an KDAM, which uses a strong multi-task model as a teacher to guide the student's multi-task model learning.

Knowledge Distillation. The objective of the Knowledge distillation [13, 1] is mainly to distill the logits from certain outputs of a *teacher* to a *student* by minimizing the KL divergence [13], where the temperature $\mathcal{T}$ factor is applied to soften the output logits. Knowledge distillation is also introduced into vision tasks, such as image classification [30, 6, 5, 18], segmentation [39, 14, 38] and detection [55, 36, 48]. Knowledge distillation has emerged as an effective technique for improving multi-task learning of dense prediction tasks. Several recent studies [22, 35, 12, 17] have explored various aspects of knowledge distillation in the context of multi-task learning, with a focus on improving model performance and efficiency. The work [8] is to distill representation from a full image to the representation predicted from a masked image to perform multiple tasks. [17] designs the selective training layers for each task using an adaptive feature distillation loss with an online task weighting scheme. This task-based feature distillation allows MTL networks to be trained in a similar manner to single-task networks. Their success depends on multiple strong teacher models to guide student model training, which can get more information from the teacher model during training. However, these distillation MTL models' exact logits matching of teacher and student predictions with KL divergence can interfere with student model training and become more sensitive to outliers, leading to poor knowledge distillation. In this paper, we develop an efficient distillation procedure specific to MTL of dense prediction. We use an alternative knowledge distillation strategy that leverages inter-task and intra-task information. It provides less sensitivity to small probabilities and can account for uncertainty in the student model predictions.

3. Method

3.1. Notations

The input image $x_i \in \mathbb{R}^{H\times W\times C}$, where H, W, and C are the height, width, and channel of the image feature, respectively. We use n to denote the task number, with $n \in \{1, 2, ..., N\}$, st to denote the single-task and mt to denote the multi-task. Define $\mathcal{F}^T_{mt}$ as the multi-task teacher model and $\mathcal{F}^T_{st}$ as the single-task teacher model.

3.2. Formulation

The clustering regularizer loss is defined as the Kullback-Leibler (KL) [13] divergence between soft assignment y_{pred} and auxiliary target distribution y_{true}:

$$\begin{aligned}\mathcal{L}_{\text{KL-div}} &:= KL(y_{\text{pred}},\ y_{\text{true}}) = y_{\text{true}} \cdot \log \frac{y_{\text{true}}}{y_{\text{pred}}} \\ &= y_{\text{true}} \cdot (\log y_{\text{true}} - \log y_{\text{pred}}),\end{aligned} \tag{1}$$

where y_{pred} is the output of the model and y_{true} is the observation labels in the dataset. KL divergence is leveraged in plain KD to transform knowledge from the strong teacher model to the student model, where y_{pred} and y_{true} are the outputs of the student model and the teacher model. The plain KD loss is represented as:

$$\mathcal{L}_{plainKD} = \mathcal{T}^2 \cdot \mathcal{L}_{\text{KL-div}}(P_S, P_T), \tag{2}$$

where $\mathcal{T}$ is the distillation temperature factor to control the softness of logits of the softmax output. P_S is the softmax output of student model. P_T is the softmax output of the teacher model. The student and teacher generate the logits $\mathcal{F}^S$ and $\mathcal{F}^T$, respectively.

$$P_S = \text{softmax}(\mathcal{F}^S/\mathcal{T}), \quad P_T = \text{softmax}(\mathcal{F}^T/\mathcal{T}), \tag{3}$$

3.3. KD Alternative Match (KDAM)

KL divergence is a distance metric that measures the difference between the teacher's and student's probability distributions. However, it is sensitive to outliers in the data, which can result in unstable training and poor generalization.

There are two challenges using KL divergence in plain KD. First, one issue with KL divergence in knowledge distillation is that it can be unstable and lead to vanishing

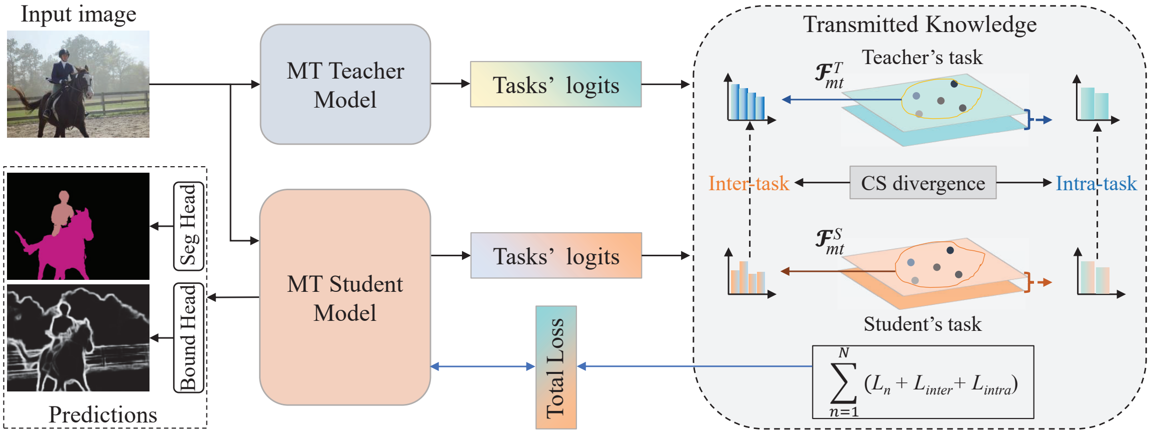


Figure 2: Illustration of our knowledge distillation framework in MTL. Inter-task information: transmitted knowledge between the predicted probabilistic distributions on all tasks' channels of teacher and student. Intra-task information: transmitted knowledge of the probabilities of all the instances on each task. For simple purposes, We set two tasks in this figure. N denotes the task number. $\mathcal{F}^T_{mt}$ and $\mathcal{F}^S_{mt}$ means the multi-task (mt) teacher and student model output, respectively.

gradients during training. This is because KL divergence involves taking the logarithm of probabilities, which can cause numerical instability when dealing with small or zero probabilities. It can make optimizing models that use KL divergence in knowledge distillation difficult. Second, KL divergence does not consider the uncertainty in the student predictions, which can lead to overconfidence and poor generalization in certain cases. This is particularly relevant in scenarios where the student model is less complex than the teacher model, as the student may not have the capacity to capture all of the information contained in the teacher's soft targets.

To address these issues, alternative divergence measures have been proposed, named Cauchy-Schwarz (CS) divergence, which is less sensitive to small probabilities and can account for uncertainty in the student model predictions. In this way, the alternative KD tries to save more task-specific information between the teacher and student on the probabilistic distribution.

In knowledge distillation, let $\mathcal{DS} = (x_i, y_i)$ denote the training dataset of the student network, where x_i and y_i are the input and the corresponding ground truth label, respectively. Let $\mathcal{DT} = (x_i, P_{T_i})$ denote the training dataset of the teacher network, where P_{T_i} is the predicted probability distribution over the class labels by the teacher network for the input x_i. Let $\mathcal{F}^S : \mathcal{X} \rightarrow \mathcal{Y}$ and $\mathcal{F}^T : \mathcal{X} \rightarrow \Delta^{C-1}$ denote the student and teacher networks, respectively, where Δ^{C-1} is the C-dimensional probability simplex.

Knowledge distillation aims to train the student network $\mathcal{F}^S$ to match the predictions of the teacher network $\mathcal{F}^T$ for the same input data. To achieve this, we minimize the CS divergence between the soft labels predicted by the teacher network and those predicted by the student network. The CS divergence between the soft label distributions P_{T_i} and P_{S_i} predicted by the teacher and student models, respectively, for the input x_i is defined as:

$$\mathcal{L}_{CS} := D_{CS}(P_S, P_T) = -\ln \frac{\int P_{T_i}(x)P_{S_i}(x)dx}{\sqrt{\int P_{T_i}(x)dx \ \int P_{S_i}(x)dx}}. \tag{4}$$

Note that $D_{CS} = 0$ only for $P_S = P_T$. Eq. 4 can improve numerical stability and the ability to account for uncertainty in student predictions.

The inter-task information for each task. We develop an alternative inter-task comparison strategy in which the distribution of teacher and student of all tasks. The inter-task information between teacher and student is defined as:

$$\mathcal{L}_{inter} = \frac{1}{B}\sum_{i=1}^{B} D_{CS}(P_{Si,:}, P_{T_{i,:}}), \tag{5}$$

where the B denotes batch size.

The intra-task information for each task. For intra-class information, we discuss why naive hard example mining cannot handle noise/outliers and propose a simple and effective balancing strategy for fast and robust hard example mining. As shown in Figure 2, the intra-task is defined as:

$$\mathcal{L}_{intra} = \frac{1}{C}\sum_{j=1}^{C} D_{CS}(P_{S:,j}, P_{T_{:,j}}), \tag{6}$$

where C denotes the channel.

Finally, we wrap task-independent knowledge distillation strategies into MTL. It can be formulated as follows:

$$\mathcal{L}_{CSKD} = \sum_{n=1}^{N} (\mathcal{L}_{inter} + \mathcal{L}_{intra}), \quad (7)$$

where the N is the task number. We sum up the $\mathcal{L}_{inter}$ and $\mathcal{L}_{intra}$ to improve the distillation performance for MTL.

3.4. Training objective

As depicted in Figure 2, our method consists of a teacher MT model, a student MT model, and a new strategy of KD. Concretely, we first train a strong MT model as a teacher model. Then we introduce a trained teacher model for the student MT model training. The plain multi-task loss function is given as:

$$\mathcal{L}_{mt} := \mathcal{L}(y_{\text{true}}, P_S) = \sum_{n=1}^{N} \lambda_n \mathcal{L}_n, \quad (8)$$

where λ_n is a hyperparameter factor of the task number n.

In the multi-task knowledge distillation setting, there are a trained multi-task teacher model (*i.e.,* T^{mt}) and a student model (*i.e.,* S^{mt}). As a result, the KDAM overall training loss function is typically a weighted sum of the MTL loss ($\mathcal{L}_{mt}$) and the knowledge distillation loss.

$$\mathcal{L}_{total} = \alpha \cdot \mathcal{L}_{mt} + \beta \cdot \mathcal{L}_{CSKD}(P_{S^{mt}}, P_{T^{mt}}), \quad (9)$$

where α and β are the scale factor for balancing the losses. In this way, via the $\mathcal{L}_{CSKD}$ loss, the student to match the strong teacher network's output.

4. Experiments

We validate the superiority of our KDAM framework under two representative datasets: NYUD-v2 [42] and PASCAL-Context [7]. In Section 4.1, we provide the experimental setup and evaluation details we used for experiments. Section 4.2 presents the main qualitative and quantitative MTL results. Finally, in Section 4.3, we perform ablation studies to verify the effect of components and efficiencies of our method.

4.1. Experiment setup

NYUD-v2 dataset and metrics. NYUD-v2 comprises RGB and Depth frames 795 images are used for training and 654 images for testing. NYUD-v2 is adopted for semantic segmentation (SemSeg), depth estimation (Depth), surface normal estimation (Normal) and boundary detection (Bound) tasks by providing dense labels. The semantic segmentation labels classify each pixel in the RGB images into one of 40 object categories. Four evaluation metrics are available to evaluate the MTL model performance, which includes mean Intersection over Union (mIoU) for the SemSeg task, root mean square error (rmse) for the Depth task, mean Error (mErr) for the Normal task, and optimal dataset scale F-measure (odsF) for the Bound task.

PASCAL-Context dataset and metrics. PASCAL-Context consists of 10,103 images with complex scenes, covering 400 object categories and 59 background regions. PASCAL-Context is adopted for semantic segmentation (SemSeg), human parts segmentation (PartSeg), saliency estimation (Sal), surface normal estimation (Normal), and boundary detection (Bound) tasks by pixel-level semantic labels for each image. Five evaluation metrics are available to evaluate the MTL model performance, which includes mean Intersection over Union (mIoU) for the SemSeg and PartSeg tasks, mean Error (mErr) for the Normal task, optimal dataset scale F-measure (odsF) for the Bound task, and maximum F-measure (maxF) for the Sal task. The average per-task performance drop (Δ_{mt}) is used to quantify multi-task performance. $\Delta_{mt} = \frac{1}{N}\sum_{i=1}^{N}(F_{mt,n} - F_{st,n})/F_{st,n} \times 100\%$, where mt, st and N mean multi-task model, single-task baseline and task numbers. Δ_{mt}: higher is the better.

Setting. To be universal, we perform experiments on the NYUD-v2 [42] and PASCAL-Context [7] datasets using different teacher-student pairs. The "Plain KD vs. DIST vs. Our KDAM" experiment is conducted using Swin-L [34] as a teacher model, Swin-T [34] and Swin-S [34] as student models on the NYUD-v2. On PASCAL-Context, we adopt Swin-L [34] and Swin-B [34] as teacher models, HRNet18 [43], Swin-T [34] and Swin-S [34] as student models.

Loss weights. In our experiments, we use the algorithm in [3] to learn weights for each task over the course of training. For a fair comparison, we search the optimal hyperparameters (*i.e.,* the loss ratio α, β and the temperature $\mathcal{T}$) for each teacher-student pair. On NYUD-v2 and PASCAL-Context, we set $\alpha = 1.0$, $\beta = 1.0$ in Eq. 9 using our TKD method. For plain KD [13], we set $\alpha = 1.0$, $\beta = 1.0$ in Eq. 9 and use a default temperature $\mathcal{T} = 3$ in Eq. 2. In addition, For a fair comparison, we search the optimal hyperparameters *i.e.,* β and $\mathcal{T}$. We choose a temperature set $\mathcal{T} \in [0.1, 0.5, 1, 3, 4, 8, 16]$ for an ablation study on NYUD-v2 dataset (see Table 5a & 5b).

Baselines. Multi-task baseline uses a shared backbone network in conjunction with task-specific heads to perform the predictions for every task. We can choose different networks as backbones, such as HRNet [43] and Swin Transformer [34]. Single-task baseline uses a backbone network with a task head to conduct the predictions for a task.

4.2. Results

NYUD-v2 dataset. We investigate the effectiveness of our KDAM on NYUD-v2 dataset. We first train a multi-task baseline model of Swin-L for our knowledge distillation procedure. Then we load the well trained Swin-L model

Table 1: We report the comparison of the MTL models with the state-of-the-art on NYUD-v2 dataset. '↓': lower is better. '↑': higher is better. Δ_{mt} denotes the average per-task performance drop. Swin-$\diamond$ indicates that the specific Swin model is uncertain. Gray blocks mean the multi-task baseline using our knowledge distillation method.

Model	Backbone	Params (M)	GFLOPs (G)	SemSeg (mIoU)↑	Depth (rmse)↓	Normal (mErr)↓	Bound (odsF)↑	Δ_{mt}[%]↑
single-task (ST) baseline	HRNet18	16.09	40.93	38.02	0.6104	20.94	76.22	0.00
multi-task (MT) baseline	HRNet18	4.52	17.59	36.35	0.6284	21.02	76.36	-1.89
MT+KDAM (Ours)	HRNet18	4.52	17.59	36.85	0.6250	21.04	76.10	-0.73
Cross-Stitch[37]	HRNet18	4.52	17.59	36.34	0.6290	20.88	76.38	-1.75
Pad-Net[50]	HRNet18	5.02	25.18	36.70	0.6264	20.85	76.50	-1.33
PAP[58]	HRNet18	4.54	53.04	36.72	0.6178	20.82	76.42	-0.95
PSD[31]	HRNet18	4.71	21.10	36.69	0.6246	20.87	76.42	-1.30
NDDR-CNN[10]	HRNet18	4.59	18.68	36.72	0.6288	20.89	76.32	-1.51
MTI-Net[46]	HRNet18	12.56	19.14	36.61	0.6270	20.85	76.38	-1.44
ATRC[3]	HRNet18	5.06	25.76	38.90	0.6010	20.48	76.34	1.56
ATRC+KDAM (Ours)	HRNet18	5.06	25.76	39.30	0.5919	20.72	76.91	2.05
single-task (ST) baseline	Swin-T	115.08	161.25	42.92	0.6104	20.94	76.22	0.00
multi-task (MT) baseline	Swin-T	32.50	96.29	38.78	0.6312	21.05	75.60	-3.74
MT+KDAM (Ours)	Swin-T	32.50	96.29	44.34	0.601	21.03	76.4	1.1
single-task (ST) baseline	Swin-S	200.33	242.63	48.92	0.5804	20.94	77.20	0.00
multi-task (MT) baseline	Swin-S	53.82	116.63	47.90	0.6053	21.17	76.90	-1.96
MTFormer[51]	Swin-$\diamond$	64.03	117.73	50.56	0.4830	-	-	4.12
MT+KDAM (Ours)	Swin-S	53.82	116.63	49.41	0.564	20.60	77.3	1.38

Table 2: Comparison results of DeMT and InvPT MTL model with our KDAM method. '+' means performance increase.

Network structure		Accuracy multi-task student models				Accuracy of our KDAM			
Teacher	Student	SemSeg↑	Depth↓	Normal↓	Bound↑	SemSeg↑	Depth↓	Normal↓	Bound↑
Swin-L	InvPT-T [56]	44.27	0.5589	20.46	76.10	44.93 (+ 0.66)	0.5577 (+ 0.0012)	20.27 (+ 0.19)	76.2 (+ 0.1)
Swin-L	DeMT-T[53]	46.36	0.5871	20.65	76.90	47.07 (+ 0.71)	0.5855 (+ 0.0016)	20.62 (+ 0.03)	76.9 (+ 0.0)
Swin-L	DeMT-S[53]	51.50	0.5474	20.02	78.10	51.91 (+ 0.41)	0.5512 (− 0.0038)	20.01 (+ 0.01)	78.1 (+ 0.0)

as a teacher model for student model training. In Table 1, ours surpasses the HRNet18, Swin-T and Swin-S multi-task baselines by more than 1.16%, 4.84% and 3.34% (average per-task performance drop Δ_{mt}), respectively. The baseline model coupled with the improvement in Δ_{mt} metric by our distillation method demonstrates the effectiveness of distillation. Note that These results are only conducted on the outputs of the multi-task baseline model and have a similar computational cost as the baseline model. Nevertheless, it even achieves better performance compared to those carefully designed methods. As shown in Table 1, although ATRC achieves better performance, it brings much more parameters and computation compared to the MT baseline. By contrast, our method based on distillation is cost-free. We also conduct experiments based on DeMT [53] and InvPT [56] and compare using our KD strategy with the DeMT-T, DeMT-S and InvPT-T. As shown in Table 2, our KDAM is outperformed by DeMT-T +0.71 mIoU and DeMT-S by +0.41 mIoU on SemSeg task. Applying our distillation recipe in combination with the DeMT configuration to prove our method is effective. We can see that all the added KDAM, regardless of DeMT or InvPT, improves the performance. We observe that our distillation with an alternative match leads the state-of-the-art performance.

PASCAL-Context. We also use the multi-task baseline model of Swin-L as a teacher model. We further investigate the effectiveness of our KDAM for dense predictions on PASCAL-Context dataset. We train HRNet18, Swin-T and Swin-S baseline with our distillation method. As shown in Table 3, our KDAM outperforms all the previous MTL methods on five dense predictions and the average per-task performance drops Δ_{mt}. It is worth mentioning that using a strong Swin-L teacher fails to give HRNet18 baseline an intuitive performance increase. We find that the performance of SemSeg task tends to get better as the student's backbone increases, which shows the effectiveness of our approach in segmentation tasks. Results suggest that our method can learn more semantic features for dense predictions.

Table 3: We report a comparison of the MTL models with the state-of-the-art on PASCAL-Context dataset. '↓': lower is better. '↑': higher is better. Δ_m denotes the average per-task performance drop (higher is better). Gray blocks mean the multi-task baseline using our knowledge distillation method.

Model	Backbone	SemSeg (mIoU)↑	PartSeg (mIoU)↑	Sal (maxF)↑	Normal (mErr)↓	Bound (odsF)↑	$\Delta_{mt}[\%]$↑
single-task (ST) baseline	HRNet18	62.23	61.66	85.08	13.69	73.06	0.00
multi-task (MT) baseline	HRNet18	51.48	57.23	83.43	14.10	69.76	-6.77
MT+KDAM (Ours)	HRNet18	51.91	57.63	83.92	13.90	70.1	-5.06
PAD-Net [50]	HRNet18	53.60	59.60	65.80	15.3	72.50	-4.41
ATRC [3]	HRNet18	57.89	57.33	83.77	13.99	69.74	-4.45
MQTransformer[52]	HRNet18	58.91	57.43	83.78	14.17	69.80	-4.20
ATRC+KDAM (Ours)	HRNet18	58.92	57.51	83.87	13.97	69.75	-3.94
single-task (ST) baseline	Swin-T	67.81	56.32	82.18	14.81	70.90	0.00
multi-task (MT) baseline	Swin-T	64.74	53.25	76.88	15.86	69.00	-3.23
MT+KDAM (Ours)	Swin-T	64.81	53.80	81.72	15.06	69.5	-2.3
single-task (ST) baseline	Swin-S	70.83	59.71	82.64	15.13	71.20	0.00
multi-task (MT) baseline	Swin-S	68.10	56.20	80.64	16.09	70.20	-3.97
MT+KDAM (Ours)	Swin-S	68.80	56.62	81.91	15.61	70.4	-2.54

Table 4: We report a comparison of the multi-task (MT) baseline models using different KD strategies on NYUD-v2 dataset. MT and ST denote multi-task and single-task respectively. Gray blocks mean the MT baseline using our method.

Teacher (Student)	Metrics	Accuracy ST models		Accuracy MT models		Plain KD	DIST KD	Our KDAM
		Teacher	Student	Teacher	Student	Student (MT)	Student (MT)	Student(MT)
Swin-L (Swin-T)	SemSeg↑	56.46	42.92	54.53	38.78	44.48	43.27	44.54
	Depth↓	0.508	0.610	0.532	0.631	0.604	0.599	0.601
	Normal↓	19.38	20.94	19.51	21.05	21.03	21.02	21.03
	Bound↑	78.8	76.22	78.3	75.60	76.80	76.50	76.40
	$\Delta_{mt}[\%]$↑	-	-	-2.36	-3.74	1.08	0.58	1.10
Swin-L (Swin-S)	SemSeg↑	56.46	48.92	54.53	47.90	49.20	49.39	49.41
	Depth↓	0.508	0.580	0.532	0.605	0.571	0.560	0.564
	Normal↓	19.38	20.94	19.51	21.17	20.58	20.51	20.60
	Bound↑	78.8	77.20	78.3	76.90	77.20	77.2	77.3
	$\Delta_{mt}[\%]$↑	-	-	-2.36	-1.96	0.97	1.6	1.38
Swin-B (Swin-T)	SemSeg↑	53.01	42.92	51.44	38.78	43.84	44.06	44.52
	Depth↓	0.552	0.610	0.581	0.631	0.603	0.596	0.592
	Normal↓	19.34	20.94	20.44	21.05	21.19	21.08	21.09
	Bound↑	78.00	76.22	77.80	75.60	76.60	76.70	76.6
	$\Delta_{mt}[\%]$↑	-	-	-3.2	-3.74	0.49	1.07	1.47

4.3. Ablation Studies

Distillation with alternative match helps minimize the performance gap between teacher and student. In Table 4, we first conduct experiments to compare our KDAM with plain KD [13] and DIST KD [16] at different student and teacher model sizes. We compare three types of teacher-student pairs: Swin-L:Swin-T, Swin-L:Swin-S and Swin-B: Swin-T. As shown in Table 4 (Swin-T (Swin-L) and Swin-T (Swin-B) rows), when the teacher goes larger, the SWin-T students MTL performance Δ_{mt} perform even worse than that with a medium-sized Swin-B teacher. We observe that it is not the fact that a larger teacher model leads to better performance. The plain KD using KL divergence loss can help minimize the performance gap between strong teachers and students. In addition, our KDAM showed an increasing trend with strong teachers and a more significant improvement compared to the other KD strategies, suggesting that our KDAM better handles the differences between student and strong teacher performance. Our KDAM can significantly outperform existing knowledge distillation methods on MTL of dense predictions. The results in Table 4 highlight that our KDAM is applied for capturing inter-task and intra-task information between the teacher model and the student model of each task, and thus learns more "dark knowledge" for effective distillation.

Ablation on hyperparameters $\mathcal{T}\&\beta$. We compare the ef-

Table 5: Ablation experiments of the optimal hyperparameters temperature factor $\mathcal{T}$ and the loss scale factor β of the plain KD (Eq. 2) using a Swin-L as teacher model and Swin-T as student model on NYUD-v2 dataset.

(a) Ablation on β in Eq. 9 using $\mathcal{T}$=3 and α=1.

KD style	β	Accuracy multi-task student models			
		SemSeg↑	Depth↓	Normal↓	Bound↑
Plain KD	0.1	44.42	0.6001	21.16	76.2
	0.5	**44.90**	0.6015	21.09	**76.3**
	1	44.47	0.6044	21.03	76.2
	2	44.61	**0.5975**	21.00	76.2
	4	44.09	0.6017	21.06	76.1
	8	44.42	0.6018	21.07	76.2
	16	43.53	0.6104	**21.34**	**76.3**

(b) Ablation on temperatures $\mathcal{T}$ in Eq. 2 using α=1 and β=1.

KD style	$\mathcal{T}$	Accuracy multi-task student models			
		SemSeg↑	Depth↓	Normal↓	Bound↑
Plain KD	0.1	44.27	0.6076	21.15	76.1
	0.5	44.52	0.6054	21.16	76.2
	1	44.83	**0.5969**	21.10	76.3
	3	44.47	0.6044	21.03	76.2
	4	44.21	0.6014	21.05	76.1
	8	**45.35**	0.6012	**21.04**	**76.5**
	16	44.69	0.6050	21.08	76.1

Table 6: Ablation on part of the loss. The student and teacher models are Swin-T and Swin-L, respectively. ”w/” indicates ”with”.

Method	Accuracy multi-task student model			
	SemSeg↑	Depth↓	Normal↓	Bound↑
MT baseline	38.78	0.6312	21.05	75.60
Ours w/ intra	43.31	0.6068	21.47	76.00
Ours w/ inter	44.06	0.5964	21.08	76.50
Ours w/ inter&intra	44.34	0.6012	21.03	76.70

Table 7: Ablation on the training epochs. The student and teacher models are Swin-T and Swin-L, respectively.

Model	epoch	Accuracy multi-task student model			
		SemSeg↑	Depth↓	Normal↓	Bound↑
KDAM	100	34.13	0.6715	23.26	75.3
	200	42.33	0.6285	21.86	76.0
	300	44.14	0.6156	21.28	76.2
	400	44.34	0.6010	21.03	76.2
	600	45.00	0.5955	20.69	76.4

fect of the number of temperature factor $\mathcal{T}$ and the distillation loss scale factor β in Table 5a & 5b. We find that having a larger temperature of plain KD improves the performance slightly, whereas increasing the number of distillation loss scale factors does not improve the performance.

Ablation on part of the loss. We compare three different methods named ”Ours w/ inter,” ”Ours w/ intra” and ”Ours w/ inter & intra” in Table 6. To validate the effectiveness of each loss, we conduct experiments to train students with these methods separately. For all tasks, the intrinsic inter-task and intra-task variance of the semantic similarities is actually also informative. We can see that our KDAM can achieve a performance improvement, which proves the effectiveness of the inter-task and intra-task. Our knowledge distillation with an alternative match is applied to capture inter-task and intra-task information between the teacher and the student. The results show our distillation method can learn more “dark knowledge” for effective distillation.

Ablation on training epochs. The default setting for the number of training epochs for teachers and students is 400. We independently train teacher models and pick them up at epoch 200^{th} and 400^{th}. In Table 7, we conduct ablation experiments of the training epochs using trained Swin-L as a teacher on NYUD-v2 dataset. The teacher model is trained at the 400^{th} epoch. The student model of Swin-T are selected at the 50^{th}, 100^{th}, 200^{th}, 300^{th}, 400^{th}, 500^{th} and 600^{th} epochs. Our multi-task student model reaches 45.00 (mIoU, Semseg) and 0.5955 (rmse, Depth) after 600 epochs. We observe that the multi-task student's accuracy increases monotonically with the increase of training epoch. Noticeably, the trained strong teacher model is used for improving student efficiency, while when the student model is not trained still enough leads to poor performance.

5. Conclusion

This paper introduces a new knowledge distillation procedure with an alternative match (KDAM) for MTL of dense prediction based on two simple design principles. For memory and training efficiency, we use a single strong multi-task model as a teacher during training instead of multiple teachers, as widely adopted in existing studies. Furthermore, we employ a less sensitive CS divergence instead of the KL divergence and propose a CS distillation loss accordingly. This technique can significantly improve the MTL model's distillation performance and reduce the training cost. The Extensive experiment results show that our proposed method significantly improves the model performance and the practical inference efficiency.

Acknowledgements. This work was supported by the National Natural Science Foundation of China under Grant 62122060 and the Special Fund of Hubei Luojia Laboratory under Grant 220100014.

References

[1] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *CVPR*, pages 10925–10934, 2022. 1, 3

[2] Deblina Bhattacharjee, Tong Zhang, Sabine Süsstrunk, and Mathieu Salzmann. Mult: An end-to-end multitask learning transformer. In *CVPR*, pages 12031–12041, 2022. 1, 3

[3] David Bruggemann, Menelaos Kanakis, Anton Obukhov, Stamatios Georgoulis, and Luc Van Gool. Exploring relational context for multi-task dense prediction. In *ICCV*, pages 15869–15878, 2021. 1, 2, 3, 5, 6, 7

[4] Xingyu Cai, Jiahong Yuan, Renjie Zheng, Liang Huang, and Kenneth Church. Speech emotion recognition with multi-task learning. In *Interspeech*, volume 2021, pages 4508–4512, 2021. 1

[5] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *CVPR*, pages 11933–11942, 2022. 3

[6] Xianing Chen, Qiong Cao, Yujie Zhong, Jing Zhang, Shenghua Gao, and Dacheng Tao. Dearkd: data-efficient early knowledge distillation for vision transformers. In *CVPR*, pages 12052–12062, 2022. 3

[7] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, pages 1971–1978, 2014. 5

[8] Xiaoyi Dong, Yinglin Zheng, Jianmin Bao, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. *arXiv preprint arXiv:2208.12262*, 2022. 3

[9] Yuan Gao, Haoping Bai, Zequn Jie, Jiayi Ma, Kui Jia, and Wei Liu. Mtl-nas: Task-agnostic neural architecture search towards general-purpose multi-task learning. In *CVPR*, pages 11543–11552, 2020. 1, 2

[10] Yuan Gao, Jiayi Ma, Mingbo Zhao, Wei Liu, and Alan L Yuille. Nddr-cnn: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction. In *CVPR*, pages 3205–3214, 2019. 2, 3, 6

[11] Zhichao Geng, Hang Yan, Xipeng Qiu, and Xuanjing Huang. fasthan: A bert-based multi-task toolkit for chinese nlp. In *ACL*, pages 99–106, 2021. 1

[12] Golnaz Ghiasi, Barret Zoph, Ekin D. Cubuk, Quoc V. Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. In *ICCV*, pages 8856–8865, 2021. 1, 2, 3

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2, 3, 5, 7

[14] Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *CVPR*, pages 8479–8488, 2022. 3

[15] Shihua Huang, Zhichao Lu, Ran Cheng, and Cheng He. Fapn: Feature-aligned pyramid network for dense image prediction. In *ICCV*, 2021. 1

[16] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In *NeurIPS*, 2022. 1, 7

[17] Geethu Miriam Jacob, Vishal Agarwal, and Björn Stenger. Online knowledge distillation for multi-task learning. In *WACV*, pages 2359–2368, 2023. 2, 3

[18] Minsoo Kang, Jaeyoo Park, and Bohyung Han. Class-incremental learning by knowledge distillation with adaptive feature consolidation. In *CVPR*, pages 16071–16080, 2022. 3

[19] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, pages 7482–7491, 2018. 2

[20] Meng Lan, Jing Zhang, Lefei Zhang, and Dacheng Tao. Learning to learn better for video object segmentation. In *AAAI*, volume 37, pages 1205–1212, 2023. 1

[21] Youngwan Lee, Jonghee Kim, Jeffrey Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction. In *CVPR*, pages 7287–7296, 2022. 1

[22] Wei-Hong Li and Hakan Bilen. Knowledge distillation for multi-task learning. In *ECCV Workshops*, pages 163–176, 2020. 2, 3

[23] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Learning multiple dense prediction tasks from partially annotated data. In *CVPR*, pages 18879–18889, 2022. 1

[24] Xiangtai Li, Henghui Ding, Wenwei Zhang, Haobo Yuan, Guangliang Cheng, Pang Jiangmiao, Kai Chen, Ziwei Liu, and Chen Change Loy. Transformer-based visual segmentation: A survey. *arXiv pre-print*, 2023. 1

[25] Xiangtai Li, Shilin Xu, Yibo Yang, Haobo Yuan, Guangliang Cheng, Yunhai Tong, Zhouchen Lin, and Dacheng Tao. Panopticpartformer++: A unified and decoupled view for panoptic part segmentation. *arXiv preprint arXiv:2301.00954*, 2023. 2

[26] Zekun Li, Zhengyang Geng, Zhao Kang, Wenyu Chen, and Yibo Yang. Eliminating gradient conflict in reference-based line-art colorization. In *ECCV*, pages 579–596, 2022. 2

[27] Hanxue Liang, Zhiwen Fan, Rishov Sarkar, Ziyu Jiang, Tianlong Chen, Kai Zou, Yu Cheng, Cong Hao, and Zhangyang Wang. M ^{3} vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. In *NeurIPS*, 2022. 2

[28] Xiwen Liang, Yangxin Wu, Jianhua Han, Hang Xu, Chunjing Xu, and Xiaodan Liang. Effective adaptation in multi-task co-training for unified autonomous driving. In *NeurIPS*, 2022. 2

[29] Fanqing Lin, Brian Price, and Tony Martinez. Generalizing interactive backpropagating refinement for dense prediction networks. In *CVPR*, pages 773–782, 2022. 1

[30] Sihao Lin, Hongwei Xie, Bing Wang, Kaicheng Yu, Xiaojun Chang, Xiaodan Liang, and Gang Wang. Knowledge distillation via the target-aware transformer. In *CVPR*, pages 10915–10924, 2022. 1, 2, 3

[31] Zhou Ling, Cui Zhen, Xu Chunyan, Zhang Zhenyu, Wang Chaoqun, Zhang Tong, and Yang Jian. Pattern-structure diffusion for multi-task learning. In *CVPR*, pages 4514–4523, 2020. 2, 3, 6

[32] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *CVPR*, pages 1871–1880, 2019. 1

[33] Yen-Cheng Liu, Chih-Yao Ma, Junjiao Tian, Zijian He, and Zsolt Kira. Polyhistor: Parameter-efficient multi-task adaptation for dense vision tasks. In *NeurIPS*, 2022. 2

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 5

[35] Sihui Luo, Wenwen Pan, Xinchao Wang, Dazhou Wang, Haihong Tang, and Mingli Song. Collaboration by competition: Self-coordinated knowledge amalgamation for multi-talent student learning. In *ECCV*, pages 631–646, 2020. 2, 3

[36] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *CVPR*, pages 14074–14083, 2022. 3

[37] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *CVPR*, pages 3994–4003, 2016. 2, 6

[38] Minh Hieu Phan, Son Lam Phung, Long Tran-Thanh, Abdesselam Bouzerdoum, et al. Class similarity weighted knowledge distillation for continual semantic segmentation. In *CVPR*, pages 16866–16875, 2022. 3

[39] Minh Hieu Phan, The-Anh Ta, Son Lam Phung, Long Tran-Thanh, and Abdesselam Bouzerdoum. Class similarity weighted knowledge distillation for continual semantic segmentation. In *CVPR*, pages 16866–16875, 2022. 3

[40] Dripta S Raychaudhuri, Yumin Suh, Samuel Schulter, Xiang Yu, Masoud Faraki, Amit K Roy-Chowdhury, and Manmohan Chandraker. Controllable dynamic multi-task architectures. In *CVPR*, pages 10955–10964, 2022. 3

[41] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *CVPR*, pages 16846–16855, 2022. 1

[42] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760, 2012. 5

[43] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 5

[44] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *IV*, pages 1013–1020, 2018. 2

[45] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE TPAMI*, 44(7), 2022. 2

[46] Simon Vandenhende, Stamatios Georgoulis, Luc Van Gool, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *ECCV*, pages 527–543, 2020. 1, 6

[47] Matthew Wallingford, Hao Li, Alessandro Achille, Avinash Ravichandran, Charless Fowlkes, Rahul Bhotika, and Stefano Soatto. Task adaptive parameter sharing for multi-task learning. In *CVPR*, pages 7561–7570, 2022. 3

[48] Ruibin Wang, Yibo Yang, and Dacheng Tao. Art-point: Improving rotation robustness of point cloud classifiers via adversarial rotation. In *CVPR*, pages 14371–14380, 2022. 1, 3

[49] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, pages 568–578, 2021. 1

[50] 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. In *CVPR*, pages 675–684, 2018. 2, 3, 6, 7

[51] Xiaogang Xu, Hengshuang Zhao, Vibhav Vineet, Ser-Nam Lim, and Antonio Torralba. Mtformer: Multi-task learning via transformer and cross-task reasoning. In *ECCV*, pages 304–321, 2022. 3, 6

[52] Yangyang Xu, Xiangtai Li, Haobo Yuan, Yibo Yang, and Lefei Zhang. Multi-task learning with multi-query transformer for dense prediction. *IEEE TCSVT*, 2023. 1, 2, 3, 7

[53] Yangyang Xu, Yibo Yang, and Lefei Zhang. DeMT: Deformable mixer transformer for multi-task learning of dense prediction. In *AAAI*, 2023. 1, 2, 6

[54] Yangyang Xu and Lefei Zhang. DGMLP: Deformable gating mlp sharing for multi-task learning. In *CICAI*, pages 117–128, 2022. 2

[55] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *CVPR*, pages 4643–4652, 2022. 3

[56] Hanrong Ye and Dan Xu. Inverted pyramid multi-task transformer for dense scene understanding. In *ECCV*, 2022. 1, 2, 3, 6

[57] Lijun Zhang, Xiao Liu, and Hui Guan. Automtl: A programming framework for automating efficient multi-task learning. In *NeurIPS*, 2021. 2

[58] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*, pages 4106–4115, 2019. 2, 6

[59] Qingping Zheng, Jiankang Deng, Zheng Zhu, Ying Li, and Stefanos Zafeiriou. Decoupled multi-task learning with cyclical self-regulation for face parsing. In *CVPR*, pages 4156–4165, 2022. 3