Learning to Project for Cross-Task Knowledge Distillation

Dylan Auty*
dylan.auty12@imperial.ac.uk
Roy Miles*
r.miles18@imperial.ac.uk
Benedikt Kolbeinsson*
benedikt.kolbeinsson15@imperial.ac.uk
Krystian Mikolajczyk
k.mikolajczyk@imperial.ac.uk

MatchLab
Imperial College London
South Kensington, UK

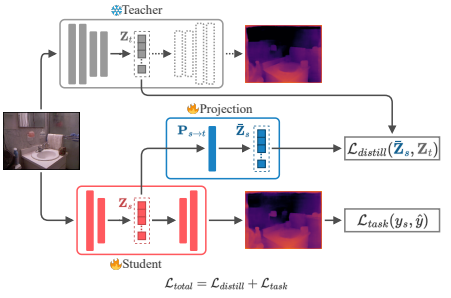
Abstract

Traditional knowledge distillation (KD) relies on a proficient teacher trained on the target task, which is not always available. In this setting, cross-task distillation can be used, enabling the use of any teacher model trained on a different task. However, many KD methods prove ineffective when applied to this cross-task setting. To address this limitation, we propose a simple modification: the use of an inverted projection. We show that this drop-in replacement for a standard projector is effective by learning to disregard any task-specific features which might degrade the student's performance. We find that this simple modification is sufficient for extending many KD methods to the cross-task setting, where the teacher and student tasks can be very different. In doing so, we obtain up to a 1.9% improvement in the cross-task setting compared to the traditional projection, at no additional cost. Our method can obtain significant performance improvements (up to 7%) when using even a randomly-initialised teacher on various tasks such as depth estimation, image translation, and semantic segmentation, despite the lack of any learned knowledge to transfer. To provide conceptual and analytical insights into this result, we show that using an inverted projection allows the distillation loss to be decomposed into a knowledge transfer and a spectral regularisation component. Through this analysis we are additionally able to propose a novel regularisation loss that allows teacher-free distillation, enabling performance improvements of up to 2.3% on ImageNet with no additional training costs.

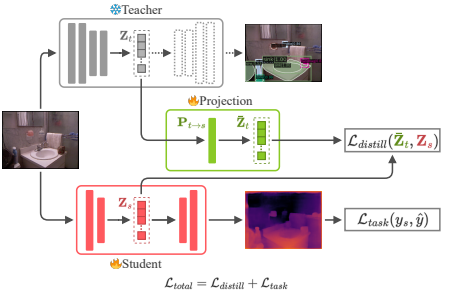
1 Introduction

Knowledge distillation (KD) has emerged as a very effective tool for training small and efficient models [5, 17, 37, 40, 43, 50]. It leverages the pre-trained knowledge of a much larger (teacher) model to guide and enhance the training process of a significantly smaller (student) model. Since its inception, KD has been applied to a wide variety of tasks in the

* The authors contributed equally to this paper



(a) Traditional same-task knowledge distillation that projects to the teacher feature space.



(b) Cross-task knowledge distillation using our inverted projection.

Figure 1: **Our cross-task knowledge distillation pipeline**, where a student model is trained on a target task with the aid of a frozen teacher that is pretrained on a **different** task. Compared to standard same-task feature distillation (fig. 1(a)), our cross-task approach uses an *inverted projector* (fig. 1(b)) which is able to discard irrelevant task-specific features from the different-task teacher. The loss comprises a feature distillation loss $\mathcal{L}_{distill}$ that matches the student features with the projected teacher features, and a task-specific supervised loss $\mathcal{L}_{task}$ applied only to the student model's output for the target task.

computer vision [7], audio [11] and language [56] domains, enabling the deployment of models across many embedded devices.

However, existing approaches for KD are often limited to the cases where the teacher shares the same task with the student [6, 7, 8, 23, 25, 60]. This is very restrictive since there are many applications where there is simply no suitable pretrained teacher available due to, for example, the lack of any large annotated training data. This problem commonly arises for tasks that require expensive human annotation [39], such as in robotics [26], or where the collection of data is prohibitive for other reasons, such as in the medical [51, 55] and aerial domains [19, 29, 63]. In these cases, it is not possible to train a suitable teacher for the target task, therefore we propose a **cross-task knowledge distillation**. In the cross-task KD setup, a teacher model trained for a **different** task can be used to improve the student performance. This setting is well-suited for tasks lacking a task-specific pretrained teacher, as it allows for any other off-the-shelf pretrained model to be used to improve the student model performance instead. It is also increasingly relevant as the compute and data costs to train large models increases. Similarly, data labelling may be cheaper for one task than another, e.g. training a model using a cheaply-labelled auxiliary task is very common in active learning [4, 21, 24] and federated learning [1].

We show that the traditional methods for same-task KD fail in this new and more general cross-task setting since they transfer domain-specific knowledge, which is associated with the *teacher's* task. Therefore, while they increase the student's performance in the traditional same-task setting, they degrade it in the cross-task scenario. We propose the use of an inverted projection to address this problem. We find that this modification is very effective in the cross-task setting due to its suppression of task-specific information. Most notably, we can obtain up to a 7.47% performance improvement by distilling from a teacher trained on various different tasks. We demonstrate that this simple drop-in replacement enables many KD methods to adapt to the cross-task setting, and we show consistent improvements across various tasks including depth estimation, segmentation, and image translation.

To obtain more insights into the underlying mechanism of the inverted projector, we explore the training dynamics of its weights. We find that the least-significant singular vectors of the teacher's features are suppressed in cases where there is a significant task gap, which indicates that these singular vectors tend to be more task-specific. Based on this observation we show that the suppression of singular vectors by the projector naturally leads to a decoupling of the distillation loss into a knowledge transfer and spectral regularisation component. This enables us to derive a **cheap spectral regularisation loss.** We describe this loss as a *teacher-free distillation* method since it explicitly exploits the emergent regularisation component from cross-task distillation. The new loss makes it possible to efficiently achieve performance competitive with many state-of-the-art KD methods without the need for any pre-trained teachers, with a 3.2% relative improvement over the baseline model on ImageNet-1K. In summary, our contributions are given as follows:

- We propose a simple modification to standard KD that enables cross-task distillation: a learnable inverted projection.
- We show consistent and substantial performance improvements in the cross-task setting of up to 7% through extensive experiments.
- By analysing the training dynamics of the projector weights, we are able to decouple a knowledge transfer and spectral regularisation component. We use this to derive a teacher-free regularisation loss that obtains up to 8% improvement over the baseline with no additional training cost.

2 Related Work

Knowledge distillation (KD) is a technique that involves transferring the knowledge from a large teacher model to a smaller student model, aiming to improve the performance of the student on the target task. It has become increasingly popular in recent years for the deployment of models on resource constrained devices, such as mobile phones, and has been applied in image classification [6, 23], semantic segmentation [35], video object segmentation [43], and natural language processing [57]. Existing literature has extensively explored various distillation pipelines [32, 38, 52] along with both empirical and theoretically motivated loss formulations [6, 23, 42, 72] that can facilitate the knowledge transfer process. However, these conventional methods still predominantly focus on same-task distillation [7, 60], wherein the student and teacher models are trained on the same task. There are many applications where there are no pre-trained off-the-shelf teacher models available, thus motivating the need to perform *cross-task distillation*. Some prior works have pursued cross-task distillation both as a generalisation of the knowledge transfer occurring in traditional knowledge distillation [67] and because of the observation that some tasks will naturally tend to share information [65, 69]. CrossDistil [65] was one of the first to partially explore this new setting by introducing a quadruplet loss, calibration term, and an error correction mechanism, however knowledge was distilled between the task-specific decoder heads of a multi-task model with shared encoder weights, rather than between two fully-separate models. ProC-KD [33] transfer local-level object knowledge for large-scale cross-task distillation, while [67] construct a relational embedding for the loss. [69] perform cross-task KD to augment text-to-image generation using image semantic encoders, but the proposed method is tightly coupled with the model architecture at each stage of the pipeline. In contrast to these works, we propose a

very simple extension to the typical feature distillation pipeline that enables the distillation of knowledge cross-task across a wide range of settings.

Transfer learning and domain adaptation are widely studied areas in machine learning that leverage the knowledge acquired by a pretrained model on one task to enhance the performance on a different, yet related task [73]. This paradigm has demonstrated significant success in various fields, especially in computer vision [16, 71] and natural language processing [50, 59], by reducing the training time and data requirements in the target domain. Most existing transfer learning or domain adaptation algorithms attempt to align the feature representations across the two domains. This can be achieved by minimising some statistical discrepancy between the two spaces [23] or introducing additional adversarial losses [14]. More recent works [10] have shown a spectral divide between the domain-specific and domain-agnostic features. This is where the large singular values of the features can generalise across domains, whereas the small singular values are domain-specific. This observation has led to follow-up works [9, 10, 47] by proposing spectral alignment and normalisation techniques. We take a similar approach for transferring knowledge between different tasks, but in the context of knowledge distillation, where an additional capacity gap between the source and target task models exists. This work also enables a more concrete bridge between the field of transfer learning and knowledge distillation.

Multi-task learning. There are many cases where jointly training on multiple tasks or modalities can improve not only the generality of models [48] but also the single-task performance. For example, monocular depth estimation has been shown to share knowledge with other tasks, such as semantic segmentation [2, 8, 27, 30, 51, 62, 64]. Intuitively, this follows for other task pairs; for instance, both semantic segmentation and classification target the semantics within an image. Unfortunately, multitask models are often too large and expensive to run on resource-constrained devices [28]. Additionally, jointly learning multiple tasks with a small model can degrade the downstream performance, as additional tasks or objectives can conflict with the target task when there is insufficient capacity in the student to optimise for both [18].

3 Method

3.1 Cross-task Feature Distillation

Cross-task distillation is motivated by the intuitive and demonstrated overlap in useful information between different tasks (see section 2). We use feature-space distillation, which aims to align the feature spaces of a student model and a teacher model. To do this, a learnable projection is used to map the features from one model into the feature space of the other. However, in the cross-task case, where the teacher model has been trained for a significantly different task to the student model's target task, there are specific issues to contend with that traditional KD methods do not address. We introduce a novel inverted projection that is well-suited to the cross-task setting, in contrast to the traditional projection [54, 60] which is better suited to the same-task setting.

3.2 Importance of Feature Projection

In the traditional same-task setting, the teacher model is already pre-trained for the student's target task, so it is desirable for the student to match features as closely as possible with those produced by the teacher. In this case, the task-specific knowledge is helpful in improving the

KD Method	Teacher task ⟶ / Projection type	Depth $\delta_1 \uparrow$	Depth Abs. ↓	Depth RMS ↓	Instance Seg. $\delta_1 \uparrow$	Instance Seg. Abs. ↓	Instance Seg. RMS ↓	Classification $\delta_1 \uparrow$	Classification Abs. ↓	Classification RMS ↓	Random $\delta_1 \uparrow$	Random Abs. ↓	Random RMS ↓
No teacher (baseline)		*0.845* ±0.007	*0.127* ±0.003	*0.440* ±0.005	*0.845* ±0.007	*0.127* ±0.003	*0.440* ±0.005	*0.845* ±0.007	*0.127* ±0.003	*0.440* ±0.005	*0.845* ±0.007	*0.127* ±0.003	*0.440* ±0.005
FitNets [34] *ICLR 2015*	Traditional	**0.868**	**0.117**	**0.406**	**0.855**	**0.122**	**0.425**	0.845	0.125	0.439	0.828	0.134	0.455
	Inverted (ours)	0.849	0.124	0.432	0.851	0.124	0.431	**0.850**	**0.124**	**0.434**	**0.851**	**0.124**	**0.431**
	Improvement	*-2.17%*	*-6.35%*	*-6.49%*	*-0.41%*	*-1.78%*	*-1.31%*	*0.50%*	*0.53%*	*1.34%*	*2.86%*	*7.47%*	*5.20%*
AT [41] *ICLR 2017*	Traditional	**0.856**	**0.122**	0.426	0.852	0.123	0.431	0.850	0.125	0.433	**0.857**	**0.121**	**0.428**
	Inverted (ours)	**0.856**	**0.122**	**0.425**	**0.855**	**0.121**	**0.429**	**0.853**	**0.123**	**0.430**	**0.857**	0.122	**0.428**
	Improvement	*-0.11%*	*-0.08%*	*0.02%*	*0.42%*	*1.38%*	*0.53%*	*0.35%*	*1.61%*	*0.79%*	*0.05%*	*-0.83%*	*0.09%*
PKT [30] *ECCV 2018*	Traditional	**0.854**	**0.122**	0.429	**0.857**	**0.123**	**0.427**	0.851	0.124	0.432	0.856	0.123	0.429
	Inverted (ours)	**0.854**	**0.122**	**0.427**	0.854	**0.123**	0.429	**0.853**	**0.123**	**0.431**	**0.858**	**0.122**	**0.426**
	Improvement	*0.04%*	*-0.16%*	*0.42%*	*-0.34%*	*-0.08%*	*-0.44%*	*0.25%*	*1.29%*	*0.30%*	*0.29%*	*1.22%*	*0.84%*
Ensemble [31] *NeurIPS 2022*	Traditional	**0.861**	**0.119**	**0.416**	**0.856**	**0.122**	**0.425**	**0.852**	**0.124**	**0.431**	0.835	0.128	0.446
	Inverted (ours)	0.849	0.124	0.433	0.848	0.124	0.435	0.847	0.125	0.437	**0.849**	**0.124**	**0.432**
	Improvement	*-1.46%*	*-4.64%*	*-4.11%*	*-0.95%*	*-1.64%*	*-2.16%*	*-0.63%*	*-0.89%*	*-1.30%*	*1.74%*	*2.75%*	*3.03%*

Table 1: **Cross-task distillation to a depth estimation student model using similar (■) and dissimilar (■) teacher tasks**, showing the **increasing** effect of our inverted projection as similarity between teacher and student tasks **decreases**. We use our inverted projector with four different KD methods to show its general applicability. The inverted projector outperforms traditional projections in the cross-task case for which it is designed, but always produces a performance improvement over the baseline (no distillation) regardless of the teacher task. ■ ■ ■ ■ colour map denotes *decreasing* student-teacher task similarity.

student performance. However, for the cross-task setting, the teacher is trained on a different task to the student. As detailed in section 2, there is likely at least some shared knowledge between different tasks. The issue in the cross-task knowledge distillation setting is how to extract only the *task-agnostic* knowledge and the knowledge *shared* between tasks, all while ignoring the irrelevant features produced by teacher. This last point is especially important for smaller student models as they do not have the capacity to effectively learn the union of two very different feature spaces. A projection layer is often used in knowledge distillation to match the student's feature dimensions with those of the teacher [54, 60]. Although recent works have highlighted the importance of the projector in improving the efficacy of same-task distillation [12, 41, 44], they have proven ineffective when there is cross-task information present. We propose a modification of the projection in which we instead map from the teacher space onto the student space. We describe this as *inverting* the projector, and we find that it enables the suppression of irrelevant (task-specific) features. We show that this inverted projector can effectively discard these irrelevant features if needed. If this were to be used in the traditional same-task setting, the discarding of these features would be detrimental, but in the cross-task setting, it is actively desirable.

3.3 Setup and Training loss

Our cross-task knowledge distillation pipeline is shown in figure 1(b). It consists of a trainable student model, which is to be trained on a given target task, and a frozen teacher model, which is pre-trained on a different task. This setup is in contrast to the traditional same-task knowledge distillation setting, which is shown in figure 1(a). Both the student and teacher models receive the same input image, and their respective encoders produce features $\mathbf{Z}_s$ and $\mathbf{Z}_t$ respectively. A learnable linear projection matrix $\mathbf{P}$ is used to project $\mathbf{Z}_t$ to the dimensions of $\mathbf{Z}_s$, giving $\bar{\mathbf{Z}}_t = \mathbf{Z}_t\mathbf{P}$. A distance function d is then used between the student features and the projected teacher features:

$$\mathcal{L}_{distill} = d(\mathbf{Z}_s, \mathbf{Z}_t\mathbf{P}), \tag{1}$$

where d can be any distance metric, such as the L2 loss used by FitNets [24] or the attention mapping described by AT [32]. In addition to this loss, we also use a task-specific supervision loss between the student model's output y_s and the ground truth labels y to ensure the student's output aligns with the target task. Since the teacher's output is not used, we only perform a forward pass through its encoder in order to reduce the training compute required. The final loss is given by:

$$\mathcal{L}_{total} = \mathcal{L}_{task}(y_s, y) + \mathcal{L}_{distill}(\bar{\mathbf{Z}}_t, \mathbf{Z}_s), \tag{2}$$

where $\mathcal{L}_{task}$ is the downstream target-task loss. For example, for depth estimation it will be a pixel-wise loss with the ground truth depth, and for image classification it will be a cross entropy term.

3.4 Decoupled Feature Distillation

To obtain analytical insights into the consequences of projecting the teacher features, we take the case where $\mathcal{L}_{distill}$ is a simple L2 loss between the student's features $\mathbf{Z}_s$ and the projected teacher features $\bar{\mathbf{Z}}_t$. We perform singular value decomposition on both, i.e. $\bar{\mathbf{Z}}_t = \bar{\mathbf{U}}\bar{\mathbf{\Sigma}}\bar{\mathbf{V}}^T$ and $\mathbf{Z}_s = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. The cross-task setting requires that our inverted projection learns to discard the irrelevant task-specific features from the teacher model. This can be implemented using a low-rank projection of the features. However, we observe that a low-rank projection naturally emerges in the cross-task setting when using our inverted projector. In fact, this emergence is even more prominent when there is a significant task gap (see section 4). Using this low-rank property, we can express $\bar{\mathbf{Z}}_t$ using a truncated SVD, i.e. keeping few non-zero singular values. Substituting this into our $\mathcal{L}_{distill}$ with an L2 loss, we can then decouple an upper bound into a knowledge transfer and a spectral regularisation component[1]:

$$\begin{aligned}
\mathcal{L}_{distill} &= d(\mathbf{Z}_s, \mathbf{Z}_t\mathbf{P}) \rightarrow \|\mathbf{Z}_s - \mathbf{Z}_t\mathbf{P}\|_2 && \text{e.g. FitNet loss} && (3)\\
&= \|\sum_{i\in k}(\bar{\sigma}_i\bar{\mathbf{u}}_i\bar{\mathbf{v}}_i^T - \sigma_i\mathbf{u}_i\mathbf{v}_i^T) + \sum_{i\notin k}\sigma_i\mathbf{u}_i\mathbf{v}_i^T\|_2 && \text{Low-rank projection} && (4)\\
&\leq \underbrace{\|\sum_{i\in k}\bar{\sigma}_i\bar{\mathbf{u}}_i\bar{\mathbf{v}}_i^T - \sigma_i\mathbf{u}_i\mathbf{v}_i^T\|_2}_{\text{knowledge transfer}} + \underbrace{\|\sum_{i\notin k}\sigma_i\mathbf{u}_i\mathbf{v}_i^T\|_2}_{\text{regularisation}}, && \text{Decoupled upper bound} && (5)
\end{aligned}$$

where k denotes the set of indices indexing the non-zero singular values of $\bar{\mathbf{Z}}_t$. In practice, any metric d that satisfies the triangle inequality has this decoupled upper bound in the cross-task setting. This result shows that the distillation loss can be decomposed into a knowledge transfer and an implicit spectral regularisation component. It explains how the inverted projection can help to improve performance even when there is little or no knowledge to transfer from the teacher: through a low-rank regularisation on the feature space. We empirically observe this emergent decoupling in figure 3. Here we see that an inverted projected is more effective at removing irrelevant task information, which is important in the cross-task KD setting.

3.5 Teacher-Free Distillation

The decoupled feature distillation in equation 5 allows us to introduce a novel spectral regularisation loss $\mathcal{L}_{spectral}$. This loss captures the regularisation effect of the cross-task distillation process without the use of any teacher, therefore we call this method *"teacher-free*

[1] For full details, please see the supplementary material.

Teacher Task	IoU ↑	Pix. Acc. ↑
No teacher	34.60 ± 0.36	0.759 ± 0.001
Random	**37.20**	**0.768**
Classif.	36.00	0.766
Seg.	36.50	0.767

Semantic segmentation

Teacher Task	PSNR ↑	FID ↓
No teacher	20.48 ± 0.04	65.77 ± 1.21
Random	20.99	63.23
Seg.	21.10	63.44
Classif	**21.27**	65.92
Depth	20.84	**62.60**

Colourisation

Teacher Task	PSNR ↑	FID ↓
No teacher	35.29 ± 0.18	67.43 ± 1.70
Random	35.28	72.54
Classif	**36.29**	**59.86**

Satellite-to-map conversion

Table 2: **Comparison for different cross-task settings**. We observe that our inverted projector is effective across many different task pairs and even for the same-task settings.

distillation" as is similarly done in other works [68]. Its objective is to minimise the least-significant singular vectors of the student model's features while keeping the most-significant. We define the spectral regularisation loss as follows. Assuming that the singular values/vectors are sorted from most to least significant, the loss can be given as follows:

$$\mathcal{L}_{spectral} = \| \sum_{i=r}^{rank} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \|_2, \tag{6}$$

where r is a hyperparameter expressing the strength of the regularisation loss and *rank* is the rank of the student features. More concretely, this hyperparameter defines the number of singular values being preserved. A smaller r will result in more singular values being suppressed, thus leading to a more aggressive regularisation of the feature space. In general, this loss effectively penalises the reconstruction of features by the least significant singular values. It suppresses the features that are overly task-specific, thus forcing the representation into a lower rank space, which leads to better generalisation. We perform experiments using this loss in section 4.5.

4 Experiments and Results

To validate the efficacy of our inverted projector in the cross-task setting, we perform experiments ablating across different distillation methods, task pairs, and architectures. We experiment with four target tasks: monocular depth estimation, semantic segmentation, image colourisation, and satellite-to-map translation. For each of these student tasks, teacher tasks are chosen that are either identical, similar, or different to them, thus demonstrating that our method is best-suited for the cross-task case where there are significant differences in the task specific knowledge learned by the teacher and student models.

Randomly-initialised teachers. An interesting question arises when we consider increasingly disparate student and teacher tasks: what happens if a *randomly-initialised* teacher is used? In this case, there is no knowledge shared between the teacher's task and the target task, however the random weights in the teacher may still produce diverse features. To investigate this question, we also distill from randomly initialised teacher models in our experiments.

4.1 Implementation details

The general framework used is described in section 3, and shown in figure 1(b). The model architectures used for the student and teacher vary depending on the task pairs. As an example, our depth estimation student is an encoder-decoder architecture with either a MobileNetV2, ResNet50, or EfficientNet-B0 backbone, and the frozen teacher model is a ViT-B/32 [15] trained for classification or the SwinV2-B [36] backbone of AiT [45] pretrained for instance

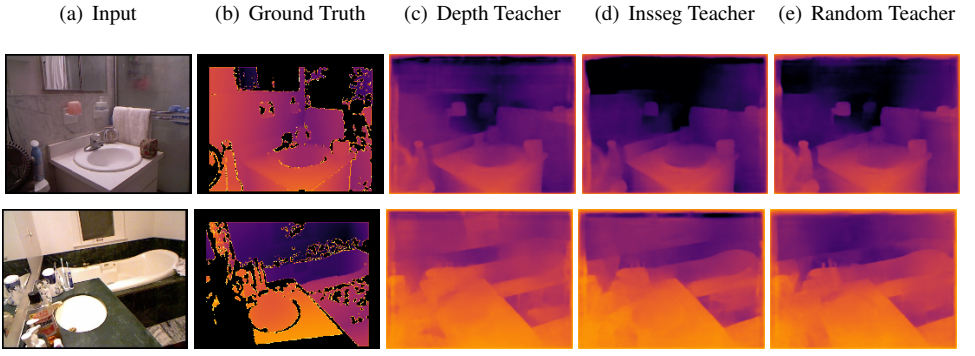


Figure 2: **Qualitative results on NYUv2 (depth) using different teacher tasks:** results from depth estimation, instance segmentation, and randomly-initialised teachers to a MobileNetV2 [20] student. In each case, we use the optimal projection type for the teacher task.

segmentation. All architectures used follow an encoder-decoder structure. The student and teacher features, $\mathbf{Z}_s$ and $\mathbf{Z}_t$, are extracted immediately after the encoder of the model in question. All decoders used require features with a spatial (height and width) dimension, therefore if the teacher model's encoder has a final pooling layer (as in the case of the classification teacher), this is removed.

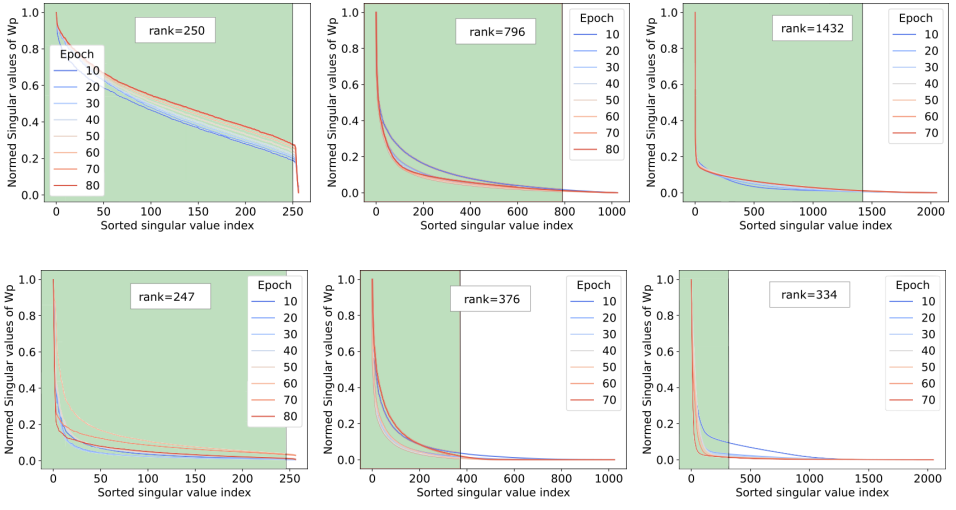
4.2 Monocular depth estimation

Monocular depth estimation is the task of inferring the depth, or distance to the camera, of every point shown in a single image. It is a challenging problem, as there is a many-to-one mapping from 3D scenes to a given 2D depth-image. When experimenting with depth estimation as a target task, we make use of teachers trained on tasks that range from similar to dissimilar to the target task of depth estimation, in terms of the overlap in knowledge. We use a depth teacher for the same-task distillation, and the instance segmentation and classification tasks for the increasingly dissimilar teacher tasks. Finally, we use a randomly initialised and frozen teacher for the most extreme cross-task setting. Experiments are run on the NYUv2 dataset [28]. Our results are shown in table 1. As expected (see section 3), the use of our inverted projection produces improvements in performance when using *dissimilar* teacher tasks (random and classification), and gives similar or worse performance when the teacher's task is similar to the target (instance segmentation and depth estimation teachers).

We use four different feature distillation methods from the same-task literature to show the utility of our method as a drop-in for use in the cross-task setting: FitNets [33], Attention Transfer (AT) [70], Probabilistic Knowledge Transfer (PKT) [46], and Ensemble (the projector ensemble method of [12]). The cross-task improvement using the inverted projection is most pronounced with FitNets and Ensemble, but there is some improvement with AT and PKT.

4.3 Semantic segmentation

Semantic segmentation is the task of labelling every pixel in the input image with a class. Our experiments are performed using MSCOCO [34], an 80-class segmentation dataset. We validate the effectiveness of our inverted projector using segmentation, classification,



(d) Seg. → Seg. (e) Class. → Seg. (f) Random → Seg.

Figure 3: **Evolution of singular values** of the projection matrix **P** under different cross-task settings and projector types. Green area highlights the rank of **P**. The projection tends towards a higher rank either when using the traditional projection or when using the same or similar-task teacher. The low-rank when using our inverted projection in the cross-task setting allows irrelevant features to be filtered out, if necessary for the task pair. **Top row**: traditional projection, **Bottom row**: our inverted projection.

or randomly initialised teachers. In all experiments, we use a simple L2 loss between the projected teacher features and the student features. Results shown in table 2. We are able to obtain significant improvement with all the teacher tasks considered, with the best improvements seen with the random teacher. This follows: both classification and semantic segmentation have significant overlap in knowledge, but the random teacher has significant task-irrelevant information that our inverted projection is able to discard. This further empirically validates the regularisation components described in equation 5.

4.4 Image-to-image translation

We experiment with two image-to-image translation tasks: transforming satellite images into maps, and colourisation of black-and-white images. These two student tasks are particularly important, as they do not have significant knowledge overlap with any of the teacher tasks used: classification, depth estimation, instance segmentation, and the randomly-initialised teacher. In contrast, segmentation and classification share a common goal of understanding semantic context of the world, while depth estimation and segmentation have been shown to aid one another (see section 2). Results are shown in table 2. Given the relative dissimilarity of the student tasks with all teacher tasks, unsurprisingly, our inverted projection performs well in all cases. We are able to use our inverted projection to produce a significant improvement over the baseline with all teacher tasks, including with the randomly initialised teacher.

Network	acc@1	#params
RegNety 160 [49]	82.6	84M
Methods using a pre-trained teacher		
DeiT-Ti⚗ [61]	74.5	6M
Co-advise [62]	74.9	6M
DearKD [8]	74.8	6M
USKD [56]	75.0	6M
*Methods **without** any teacher*		
DeiT-Ti [61]	72.2	5M
Ours: $\mathcal{L}_{spectral}(r=8)$	**74.5**	**5M**

Table 3: **Comparing our novel teacher-free spectral regularisation loss** to other state-of-the-art KD methods on ImageNet-1K [13]. Top row is the teacher model used by the KD methods that use a teacher. All methods use a DeiT-Ti [61] student model, with DeiT-Ti⚗ describing the distilled variant using distillation tokens.

4.5 Teacher-free distillation.

As shown in sections 4.2, 4.3, and 4.4, we are able to obtain significant performance improvements even when the teacher is randomly initialised and then frozen, thus containing no task-specific knowledge at all. This reinforces the conclusion reached in section 3.4: the distillation loss function $\mathcal{L}_{distill}$ may be comprised of a *knowledge transfer* component and a *spectral regularisation component*. In the case where there is no knowledge to transfer between the teacher and the student, only a regularising effect can explain the performance improvement over the baseline. To control for this, and to provide further evidence of the loss decoupling described in equation 5, we perform experiments using the *spectral regularisation loss* (equation 6). Experimenting with different r values in a depth estimation model trained on NYUv2 [58] without a teacher, we find that spectral regularisation significantly enhances performance across all r values, particularly at $r = 2$ (see appendix). This supports the decoupling of $\mathcal{L}_{distill}$ into knowledge transfer and regularisation terms (equation 5) and further validates using a randomly-initialised teacher. In table 3 we consider knowledge distillation on the large-scale ImageNet-1K dataset, and we observe that our simple regularisation loss achieves competitive performance with many state-of-the-art knowledge distillation methods.

5 Conclusion

In this paper we propose the inverted projector as a simple drop-in component for extending many knowledge distillation (KD) methods into cross-task settings, where the teacher's task differs from the student's. This inverted projector is able to suppress the irrelevant task-specific features from the teacher, which greatly improves the efficacy of cross-task distillation. We show consistent and substantial improvements across a number of cross-task pairs using our approach. Most notably, we achieve up to a 7.47% improvement for depth estimation by distilling across a significant task-gap. Through analysis, we provide a concrete interpretation and explanation for our results, leading to a natural decoupling of the objective into a knowledge transfer and a spectral regularisation component, and we extend this to demonstrate a novel drop-in teacher-free loss that achieves some of the benefits of knowledge distillation without the use of a teacher. In this work we have highlighted some of the limitations of KD in the cross-task setting, while also providing a step towards broadening its practical applicability in this new domain.

References

[1] Jin-Hyun Ahn, Kyungsang Kim, Jeongwan Koh, and Quanzheng Li. Federated active learning (f-al): an efficient annotation strategy for federated learning, 2022.

[2] Dylan Auty and Krystian Mikolajczyk. Monocular Depth Estimation Using Cues Inspired by Biological Vision Systems. In *International Conference on Pattern Recognition (ICPR) 2022*, 2022.

[3] Yucai Bai, Lei Fan, Ziyu Pan, and Long Chen. Monocular Outdoor Semantic Mapping with a Multi-task Network. *arXiv pre-print*, January 2019.

[4] Jason Baldridge and Miles Osborne. Active learning and the total cost of annotation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004.

[5] Kartikeya Bhardwaj, Naveen Suda, and Radu Marculescu. Dream distillation: A data-independent model compression framework. *ICML Joint Workshop on On-Device Machine Learning and Compact Deep Neural Network Representations (ODML-CDNNR)*, 2019.

[6] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein Contrastive Representation Distillation. *CVPR*, 2020.

[7] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling Knowledge via Knowledge Review. *CVPR*, 2021.

[8] Xianing Chen, Qiong Cao, Yujie Zhong, Jing Zhang, Shenghua Gao, and Dacheng Tao. Dearkd: Data-efficient early knowledge distillation for vision transformers. *CVPR*, 2022.

[9] Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, and Jianmin Wang. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. *NeurIPS*, 2019.

[10] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *PMLR*, 2019.

[11] Yanbei Chen, Yongqin Xian, A. Sophia Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. *CVPR*, 2021.

[12] Yudong Chen, Sen Wang, Jiajun Liu, Xuwei Xu, Frank de Hoog, and Zi Huang. Improved Feature Distillation via Projector Ensemble. *NeurIPS*, 2022.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[14] Zhun Deng, Linjun Zhang, Kailas Vodrahalli, Kenji Kawaguchi, and James Zou. Adversarial Training Helps Transfer Learning via Better Representations. *arXiv preprint*, 6 2021.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021.

[16] Utku Evci, Vincent Dumoulin, Hugo Larochelle, and Michael C Mozer. Head2Toe: Utilizing Intermediate Representations for Better Transfer Learning. In *ICML*. PMLR, 2022.

[17] Gongfan Fang, Jie Song, Xinchao Wang, Chengchao Shen, Xingen Wang, and Mingli Song. Contrastive Model Inversion for Data-Free Knowledge Distillation. *IJCAI*, 2021.

[18] Christopher Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *NeurIPS*, 2021.

[19] Michael Fonder and Marc Van Droogenbroeck. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[20] Michael H. Fox, Kyungmee Kim, and David Ehrenkrantz. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *CVPR*, 2018.

[21] Ruijiang Gao and Maytal Saar-Tsechansky. Cost-accuracy aware adaptive labeling for active learning. *AAAI*, 2020.

[22] Daniel Greenfeld and Uri Shalit. Robust Learning with the Hilbert-Schmidt Independence Criterion. *arXiv preprint*, 10 2019.

[23] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *NeurIPS*, 2015.

[24] Sheng-Jun Huang, Jia-Lve Chen, Xin Mu, and Zhi-Hua Zhou. Cost-effective active learning from diverse labelers. In *IJCAI*, 2017.

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

[26] Stephen James, Paul Wohlhart, Mrinal Kalakrishnan, Dmitry Kalashnikov, Alex Irpan, Julian Ibarz, Sergey Levine, Raia Hadsell, and Konstantinos Bousmalis. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. *CVPR*, 2019.

[27] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look Deeper into Depth: Monocular Depth Estimation with Semantic Booster and Attention-Driven Loss. In *ECCV*, 2018.

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *ICCV*, 2023.

[29] Benedikt Kolbeinsson and Krystian Mikolajczyk. DDOS: The Drone Depth and Obstacle Segmentation Dataset. *arXiv preprint arXiv:2312.12494*, 2023.

[30] Benedikt Kolbeinsson and Krystian Mikolajczyk. UCorr: Wire Detection and Depth Estimation for Autonomous Drones. In *International Conference on Robotics, Computer Vision and Intelligent Systems - ROBOVIS*, 2024.

[31] Piotr Komorowski, Hubert Baniecki, and Przemyslaw Biecek. Towards evaluating explanations of vision transformers for medical imaging. In *CVPR Workshops*, June 2023.

[32] Duong H. Le, Trung-Nhan Vo, and Nam Thoa. Paying more Attention to Snapshots of Iterative Pruning : Improving Model Compression via Ensemble Distillation. *BMVC*, 2020.

[33] Deng Li, Aming Wu, Yahong Han, and Qi Tian. Prototype-guided Cross-task Knowledge Distillation for Large-scale Models. *arXiv preprint*, 12 2022.

[34] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. *ECCV*, 2014.

[35] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured Knowledge Distillation for Semantic Segmentation. *CVPR*, 2019.

[36] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling Up Capacity and Resolution. *CVPR*, 2022.

[37] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *ICLR*, 2016.

[38] Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble Distribution Distillation. *ICLR*, 2020.

[39] John Mccormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. SceneNet RGB-D: Can 5M Synthetic Images Beat Generic ImageNet Pre-training on Indoor Segmentation? In *ICCV*, 2017.

[40] Roy Miles and Krystian Mikolajczyk. Cascaded channel pruning using hierarchical self-distillation. *BMVC*, 2020.

[41] Roy Miles and Krystian Mikolajczyk. Understanding the role of the projector in knowledge distillation. *AAAI*, 2024.

[42] Roy Miles, Adrian Lopez Rodriguez, and Krystian Mikolajczyk. Information Theoretic Representation Distillation. *BMVC*, 12 2022.

[43] Roy Miles, Mehmet Kerim Yucel, Bruno Manganelli, and Albert Saa-Garriga. MobileVOS: Real-Time Video Object Segmentation Contrastive Learning meets Knowledge Distillation. *CVPR*, 3 2023.

[44] Roy Miles, Ismail Elezi, and Jiankang Deng. V_kd: Improving knowledge distillation using orthogonal projections. *CVPR*, 2024.

[45] Jia Ning, Chen Li, Zheng Zhang, Zigang Geng, Qi Dai, Kun He, and Han Hu. All in Tokens: Unifying Output Space of Visual Tasks via Soft Token, January 2023. arXiv:2301.02229 [cs].

[46] Nikolaos Passalis and Anastasios Tefas. Learning Deep Representations with Probabilistic Knowledge Transfer. *ECCV*, 2018.

[47] Christoph Raab, Philipp Väth, Peter Meier, and Frank-Michael Schleif. Bridging Adversarial and Statistical Domain Transfer via Spectral Adaptation Networks. In *ACCV 2020*. Springer International Publishing, 2020.

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *PMLR*, 2021.

[49] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing Network Design Spaces. *CVPR*, 3 2020.

[50] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*, 2020.

[51] Michaël Ramamonjisoa and Vincent Lepetit. SharpNet: Fast and Accurate Recovery of Occluding Contours in Monocular Depth Estimation. *arXiv preprint*, 2019. arXiv: 1905.08598v1.

[52] Sucheng Ren, Zhengqi Gao, Tianyu Hua, Zihui Xue, Yonglong Tian, Shengfeng He, and Hang Zhao. Co-advise: Cross Inductive Bias Distillation. *CVPR*, 2022.

[53] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints For Thin Deep Nets. *ICLR*, 2015.

[54] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for Thin Deep Nets. *arXiv preprint*, March 2015. arXiv: 1412.6550.

[55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, 2015.

[56] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019.

[57] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019.

[58] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*, 2012.

[59] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. VL-ADAPTER: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks. In *CVPR*, 2022.

[60] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *ICLR*, 2019.

[61] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *PMLR*, 2021.

[62] Lijun Wang, Jianming Zhang, Oliver Wang, Zhe Lin, and Huchuan Lu. SDC-Depth: Semantic Divide-and-Conquer Network for Monocular Depth Estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 538–547, Seattle, WA, USA, June 2020. IEEE.

[63] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020.

[64] Daitao Xing, Jinglin Shen, Chiuman Ho, and Anthony Tzes. ROIFormer: Semantic-Aware Region of Interest Transformer for Efficient Self-Supervised Monocular Depth Estimation, December 2022. arXiv:2212.05729 [cs].

[65] Chenxiao Yang, Junwei Pan, Xiaofeng Gao, Tingyu Jiang, Dapeng Liu, and Guihai Chen. Cross-Task Knowledge Distillation in Multi-Task Recommendation. *AAAI*, 2 2022.

[66] Zhendong Yang, Ailing Zeng, Zhe Li, Tianke Zhang, Chun Yuan, and Yu Li. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. *ICCV*, 2023.

[67] Han Jia Ye, Su Lu, and De Chuan Zhan. Distilling cross-task knowledge via relationship matching. In *CVPR*, 2020.

[68] Li Yuan, Francis E.H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting Knowledge Distillation via Label Smoothing Regularization. *CVPR*, 2020.

[69] Mingkuan Yuan and Yuxin Peng. CKD: Cross-task knowledge distillation for text-to-image synthesis. *IEEE Transactions on Multimedia*, 8:1955–1968, 2020. Conference Name: IEEE Transactions on Multimedia.

[70] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2019.

[71] Amir Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling Task Transfer Learning. *arXiv:1804.08328 [cs]*, April 2018. arXiv: 1804.08328.

[72] Borui Zhao, Renjie Song, and Yiyu Qiu. Decoupled Knowledge Distillation. *CVPR*, 2022.

[73] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A Comprehensive Survey on Transfer Learning. *arXiv preprint*, 2020.