APPENDIX

FANTASTIC GAINS AND WHERE TO FIND THEM: ON THE EXISTENCE AND PROSPECT OF GENERAL KNOWLEDGE TRANSFER BETWEEN ANY PRETRAINED MODEL

A IMPLEMENTATION DETAILS AND EXPERIMENTAL INSIGHTS

In this section, we describe the implementation details of our experiments to evaluate the effectiveness of different approaches and techniques for transferring complementary knowledge between pretrained expert models trained on the same dataset.

For our initial and exploratory experiments we use a 10% stratified subset of ImageNet (Deng et al., 2009) to reduce runtimes, in order to conduct a wider range of experiments across a large number of model pairs. In detail, we drew 130 samples per class using the standard ImageNet validation set for evaluation. All of our experiments utilize an SGD optimizer with momentum 0.9 and weight decay 1e-3. Further hyperparameters were individually tuned for each investigated transfer approach.

A.1 IMPLEMENTATION OF DISTILLATION-BASED KNOWLEDGE TRANSFER VARIANTS

To set the learning rate for our default knowledge distillation based transfer approach using KL divergence (as in Section 4.1), we conducted a parameter search over a set of 33 teacher-student pairs randomly selected from the `timm` Wightman (2019) library, with learning rates lr $\in \{1\text{e-}2, 1\text{e-}3, 1\text{e-}4, 1\text{e-}5\}$, for which we found a learning rate of 1e-4 to generally work best, albeit regardless of chosen values, the average transfer delta was consistently negative.

Following Section 4.1, we also extend KL distillation with a cross-entropy classification loss. In this case, hyperparameters were determined over a grid comprising learning rates lr $\in \{1\text{e-}2, 1\text{e-}3, 1\text{e-}4, 1\text{e-}5\}$, softmax temperatures $T \in \{0.1, 1, 4, 10\}$ and weightings $\lambda \in \{0.05, 0.1, 0.25, 0.5\}$. Again, we found that a learning rate of 1e-4 was the most effective on average, but found particular variance in the weighting λ, where we observed that a larger λ value - placing higher emphasis on distillation - is better suited for transferring knowledge from a stronger teacher to a weaker student, while a smaller λ seems to be preferable when transferring knowledge from a weaker teacher to a stronger student. This further highlights the trade-off between knowledge gain and retention, where for a weaker teacher, retention plays a much more crucial part to ensure overall high performance, as student knowledge is overwritten.

For the softmax temperature, we found that a small temperature of 0.1 limits the decrease in the student's performance when transfering from a weaker teacher model, but also limiting the knowledge transfer in general. This results in only small increases in the student's performance even when transferring from a stronger teacher model. Hinton et al. (2015) propose to use a larger temperature of 4 to match soft targets to better represent smaller probabilities in the output of a single sample. However, we do not find larger temperatures to benefit the transfer performance.

In general, we find that particularly the temperature and weighting parameter guide the aggressiveness of the distillation-based transfer approach, which is highly dependent on the observed teacher and student dynamic of the provided pair of pretrained expert models. The high variance across such arbitrary model pairs makes normal knowledge distillation, even paired with an additional classification loss for stability, *not well suited as a general knowledge transfer tool*.

A.2 IMPLEMENTATION OF CONTRASTIVE DISTILLATION KNOWLEDGE TRANSFER

While knowledge distillation approaches matching the soft targets of the teacher and student model remain popular, various recent approaches argue that more structural knowledge can be transferred by encouraging the student model to also match intermediate representations of the teacher model (Liu et al., 2019; 2020; Wu et al., 2021; Park and No, 2021). Thus, in this section, we highlight results of our exploration on the feasibility of using intermediate representations and their relations to transfer knowledge between pretrained experts.

We particularly follow Tian et al. (2020), who propose to extend the basic knowledge distillation approach of Hinton et al. (2015) by aligning the feature representations of the teacher and the student models. Here, the student is encouraged to provide feature representations close to the ones of the teacher for similar images while repelling the feature representation of dissimilar images. Unlike other existing distillation approaches operating on feature representations, such a contrastive approach puts less restrictions on the architectures of the teacher and the student model, particularly because the feature representations of both models can be cheaply projected into a common feature space using a learned projection layer for both models. This enables the distillation between models of different architectures, and allows us to explore an alternative to our utilized base KL Distillation objective for general knowledge transfer (Sec. 4.1).

To assess the feasibility of representation-matching for knowledge transfer between expert models, we implement two contrastive learning approaches. First, we utilize a simple approach that encourages the distances between the feature representations of a pair of images to be similar for both the teacher and the student model. Hence, if two images result in similar feature representations in the teacher's embedding space, the student is encouraged to also provide feature representations with close proximity in their respective embedding space. Such relative similarity-based matching has seen success in standard supervised contrastive learning, such as in (Roth et al., 2021; 2022). Using t and s to denote teacher and student respectively, this gives

$$\mathcal{L}_{\text{CD}} = \text{KL}\left(\sigma(S_s), \sigma(S_t)\right), \tag{6}$$

where S is a similarity matrix containing the cosine similarities of the normalized feature representations of the current batch ($S_{ij} = \cos\text{sim}(\text{norm}(s_i), \text{norm}(s_j)), \forall i, j \in 0, ..., n$). We denote this approach as *CD Distillation*.

Secondly, we implement the contrastive representation distillation approach (*CRD distillation*) of Tian et al. (2020). As noted, CRD distillation directly aligns representations by encouraging the student to be close to the teacher for positive pairs (different augmentations of the same image) while pushing apart feature representations of negative pairs (images of different classes). The respective objective is thus given as:

$$\mathcal{L}_{\text{CRD}} = \arg\max_{f_s} \max_{h} \mathbb{E}_{q(t,s|C=1)}[\log h(t,s)] + k\mathbb{E}_{q(t,s|C=1)}[\log(1 - h(t,s))], \tag{7}$$

where we utilize t, s as shorthand for respective teacher and student representations. In addition, we use $h : t, s \rightarrow [0, 1]$ to represent a discriminator estimating whether the feature representation t and s are drawn from the same joint distribution or from the respective marginal product. In this setup, k denotes the number of negative pairs drawn from the product of marginals.

Both contrastive distillation approaches compute the overall distillation loss $\mathcal{L}_{\text{dist}}$ as a weighted combination of the respective contrastive loss $\mathcal{L}_{\text{CD}}$ or $\mathcal{L}_{\text{CRD}}$ and a cross-entropy classification loss $\mathcal{L}_{\text{XE}}$ as also used in standard KL Divergence distillation objectives Beyer et al. (2022); Rajasegaran et al. (2020).

For CD distillation based knowledge transfer, we tested different weightings between the contrastive loss and the classification loss as well as different learning rates on a small set of teacher-student combinations. On a similar hyperparameter grid as noted in the previous section, we found an equal weighting of both losses in combination with a learning rate of 1e-4 to be most suitable on average, thought with a similar trade-off as depicted in Section A.1. For the CRD distillation transfer, we found the hyperparameters as provided in Tian et al. (2020) to work well.

A.3 Implementation of continual learning based transfer approaches

Finally, we describe hyperparameters and the corresponding hyperparameter studies utilized for our continual learning extension to distillation-based knowledge transfer (see Section 4.2), in particular the setup for *XE-KL-Dist+MCL transfer* and *KL-Dist+DP transfer*.

For *XE-KL+MCL transfer*, we conducted a parameter search on a learning rate grid with the same resolution as before. However, as there are several other parameters to validate, we only test lr $\in$ {1e-2, 1e-3}. In addition to that, we follow Stojanovski et al. (2022) and test the momentum for values in $\tau \in \{0.99, 0.999, 0.9999\}$) and the interpolation frequency $N \in \{2, 10, 50, 100\}$). For the weighting against the classification objective, λ, we test 0.5 and 0.7. We conducted the

Table 5: Selection of student and teacher models used for the experiments on the 10% ImageNet subset. Each set of models was selected to contain multiple architecture types and cover a wide range of model sizes and performance levels.

Student Models	Type	Acc.	# Param.
XCiT-Large-24-P16 El-Nouby et al. (2021)	Trafo	82.89	189.10
ViT-Base-P16 Dosovitskiy et al. (2021)	Trafo	84.53	86.57
PiT-B Heo et al. (2021)	Trafo	82.44	73.76
ViT-Relpos-Medium-P16	Trafo	82.46	38.75
PiT-XS Heo et al. (2021)	Trafo	78.19	11.00
PiT-XE-dist Heo et al. (2021)	Trafo	79.31	11.00
IG-ResNext101-32x16d Xie et al. (2017)	CNN	84.17	194.03
Gluon-SeNet154 He et al. (2018)	CNN	81.23	115.09
Wide-ResNet50-2 He et al. (2016a)	CNN	81.46	68.88
ResNet101 He et al. (2016a)	CNN	79.54	44.57
ResNetV2-50 He et al. (2016b)	CNN	80.40	25.55
ResNet34-v1b He et al. (2016a)	CNN	74.59	21.80
ResNetv2-50-dist He et al. (2016b)	CNN	82.80	25.55
Mixer-L16 Tolstikhin et al. (2021)	MLP	72.07	208.20
Mixer-B16-miil Tolstikhin et al. (2021)	MLP	82.30	59.88
Mixer-B16 Tolstikhin et al. (2021)	MLP	76.61	59.88
ResMLP-36 Touvron et al. (2021)	MLP	79.77	44.69
ResMLP-24 Touvron et al. (2021)	MLP	79.39	30.02
ResMLP-12 Touvron et al. (2021)	MLP	76.66	15.35
ResMLP-24-dist Touvron et al. (2021)	MLP	80.76	30.02

Teacher Models	Type	Acc.	# Param.
ConvNext Liu et al. (2022)	CNN	86.64	197.77
VOLO-D4 Yuan et al. (2021)	Trafo	85.88	192.96
RegNety-320 Radosavovic et al. (2020)	CNN	80.80	145.05
VGG13 Simonyan and Zisserman (2015)	CNN	71.60	133.05
RegNetx-320 Radosavovic et al. (2020)	CNN	80.24	107.81
TWINS Chu et al. (2021)	Trafo	83.68	99.27
SWSL-ResNext101 Xie et al. (2017)	CNN	84.29	88.79
SWIN-S3 Liu et al. (2021b)	Trafo	83.93	71.13
TWINS-pcpvt Chu et al. (2021)	Trafo	83.14	60.99
VOLO-D2 Yuan et al. (2021)	Trafo	85.19	58.68
ResMLP-36 Touvron et al. (2021)	MLP	79.77	44.69
DLA102 Yu et al. (2018)	CNN	78.03	33.27
SWSL-ResNext50 Xie et al. (2021)	CNN	82.18	25.03
ViT-P16 Dosovitskiy et al. (2021)	Trafo	81.40	22.05
gMLP-S16 Liu et al. (2021a)	MLP	79.64	19.42
COAT-lite Xu et al. (2021)	Trafo	79.09	11.01
MixNet Tan and Chen (2019)	MLP	78.98	7.33
RegNety-006 Radosavovic et al. (2020)	CNN	75.25	6.06
MixNet Tan and Chen (2019)	MLP	76.00	4.13
XCiT-nano-12-P8 El-Nouby et al. (2021)	Trafo	73.92	3.05

parameter search as a random search over the parameter grid. Ultimately, we found a parameter setting using a high momentum of 0.9999 in combination with a high interpolation frequency (every other iteration) and a learning rate of 0.01 with weight score 0.7 to work best on average. Unlike simple KL Distillation based transfer, a fixed hyperparameter combination now results in both a positive transfer delta on average, and a significantly increased number of teachers from which each student can learn from (c.f. Fig. 4a)

For our final proposed *KL+DP transfer* approach, we again conducted a similar parameter search. However, unlike *XE-KL+MCL transfer*, the *KL+DP* approach **does not introduce additional hyperparameters** compared to the standard KL distillation based setup. Consequently, we utilize a grid of $lr \in \{1\text{e-}3, 1\text{e-}4\}$, $\lambda \in \{0.5, 0.75, 0.9, 1\}$ and $T \in \{0.1, 1, 10\}$. Note that while we ablated the use of an external cross-entropy classification loss, we found the best performance to consistently come for $\lambda = 1$ - by turning of the auxiliary classification objective. This provides strong evidence that an external measures for training stability are no longer required. Finally, across all remaining experiments, we utilize a learning rate of 1e-4 and a temperature of 1. While more in-depth parameter searches could likely provide a parameter combination that would improve the average success rate, we believe that results achieved in its current setting to offer sufficient *proof-of-concept*.

A.4 Model Lists: Large-scale Studies on Stratified ImageNet Subsets

Table 5 presents a comprehensive summary of the pretrained teacher and student models employed in our evaluation of various transfer techniques on the 10% subset of the ImageNet dataset (§5.1). These models were carefully chosen to encompass diverse architecture families, demonstrate varying performance levels, and exhibit a range of model sizes. This selection allows us to thoroughly examine the efficacy of knowledge transfer methods in different scenarios and settings. Note that for the exploration of complementary context (§3) we leveraged an even broader set of 466 teacher-student pairs comprising of 301 individual pretrained models randomly drawn from the `timm` Wightman (2019) library.

A.5 Models Evaluated on Full ImageNet

Table 6 showcases the detailed specifications of the student and teacher models employed in our full-scale ImageNet experiments (refer to Section 5.1). In the context of knowledge transfer from multiple teacher models (§4.3), we utilized the same set of teacher models in combination with a subset of student models.

Table 6: Selection of student an teacher models used for the experiments on full ImageNet. The student models were selected to contain multiple architecture types and cover a wide range of model sizes and performance levels.

Student Models	Type	Acc.	# Param.
XCiT-large-24-p16 El-Nouby et al. (2021)	Trafo	82.89	189.10
PiT-B Heo et al. (2021)	Trafo	82.44	73.76
PiT-XS Heo et al. (2021)	Trafo	78.19	11.00
Gluon-SeNet154 He et al. (2018)	CNN	81.23	115.09
ConvNext Liu et al. (2022)	CNN	84.57	50.22
ResNetV2-50-dist He et al. (2016b)	CNN	82.80	25.55
Mixer-B16-miil Tolstikhin et al. (2021)	MLP	82.30	59.88
ResMLP-24-dist Touvron et al. (2021)	MLP	80.76	30.02

Teacher Models	Type	Acc.	# Param.
SWSL-ResNext101 Xie et al. (2017)	CNN	84.29	88.79
VOLO-D2 Yuan et al. (2021)	Trafo	85.19	58.68
ResMLP-36 Touvron et al. (2021)	MLP	79.77	44.69
CoaT-lite-mini Xu et al. (2021)	Trafo	79.09	11.01

B Extended Experimental Results

In this section, we present additional experimental results of our experiments conducted in Section 5.

B.1 Additional experiments on variants of distillation-based knowledge transfer

In the following subsection, we present supplementary experiments conducted to enhance the performance of knowledge transfer variants for knowledge transfer among pretrained models.

Using a cross-entropy plus distillation transfer objective. As an alternative to the KL divergence used in Equation (1) we additionally investigated the potential of using a cross-entropy loss between the soft targets of the teacher and the student model, similar to Hinton et al. (2015). However, our results showed no advantage in using a cross-entropy loss over KL divergence. In fact, we observed an average transfer delta that was 1.2 percentage points lower when using cross-entropy loss compared to KL divergence on a set of 60 teacher-student pairs. We also explored the use of a warmup epoch where only the student model's linear layers are trained using KL divergence loss, but found no improvement in transfer performance.

Restricting the set of classes for computing the distillation-based transfer loss. In our supplementary experiments, we investigate the impact of limiting the distillation loss to focus only on the top-10 or top-100 most probable classes. This approach aimed to address the challenge posed by the large number of classes in the ImageNet dataset, specifically the potential bias towards matching the long tails of the soft target distributions. To evaluate this hypothesis, we compared the KL divergence between full soft targets and subsets of soft targets. By selecting the top-10 and top-100 most probable classes based on the teacher's predictions, we observed that some teacher-student pairs exhibited higher divergence over all classes compared to the selected subsets. This indicated the influence of classes with low prediction probabilities on the KL divergence. Motivated by these findings, we further examined the impact of considering only the top-10 or top-100 classes on the transfer performance. Across six teacher-student pairs, using the top-10 divergence resulted in an average increase of 0.20 percentage points in transfer delta. Moreover, we observed that the magnitude of improvements aligned with the differences between the top-10 and total KL divergence. Our findings suggest that limiting the divergence to selected classes can be advantageous when dealing with a large number of classes, although the magnitude of improvements remains limited.

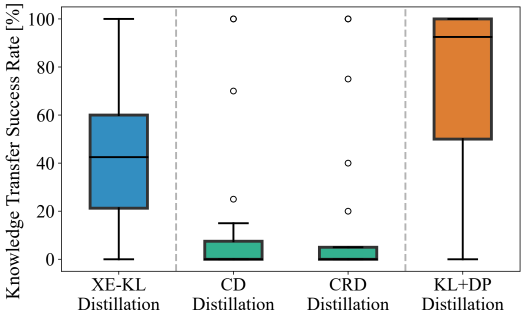


Figure 6: Share of teacher increasing student performance (success rate) for contrastive distillation (green) vs classification-guided distillation (blue) and continual learning based KL+DP (orange).

Contrastive distillation for knowledge transfer between arbitrary models To understand how well contrastive distillation techniques are suited for knowledge transfer between arbitrary

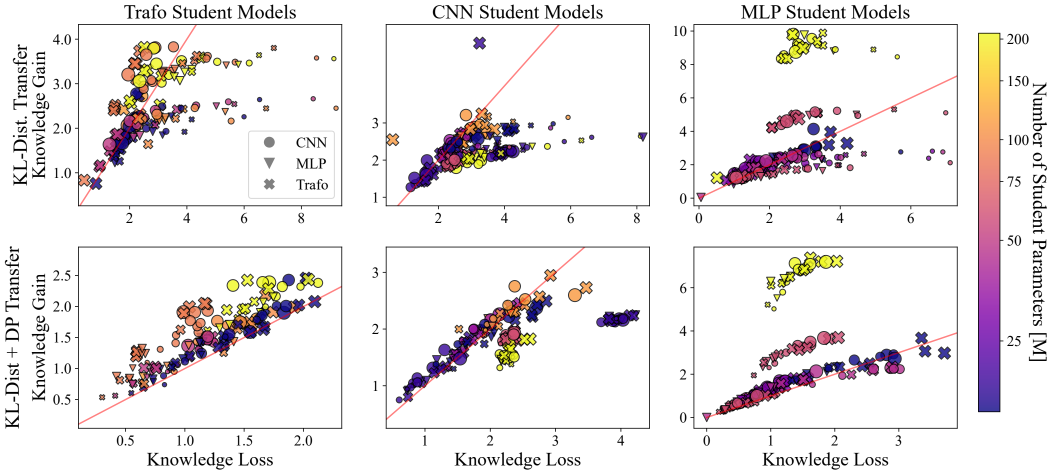


Figure 7: Share of transferred knowledge (knowledge gain) visualized against the share of knowledge lost for vanilla KL distillation and our proposed KL+DP distillation approach. Student models are grouped by their respective architecture type. Each marker represents one teacher-student pair. The color of the markers represents the size of the student, while marker shapes determine the teacher architecture. The marker size visualizes the teacher's performance. Results showcase a clear benefit of KL+DP, moving most points to areas of positive knowledge transfer (above red diagonal).

pretrained models, we measure the average transfer success rate for both CD and CRD distillation transfer (§A.2), with results shown in Fig. 6. We leverage the same experimental setup on 10% ImageNet as for the other transfer approaches (see §5.1). The experimental results clearly show the contrastive distillation approaches to be unable to improve the student model for most teacher models. On closer examination of the results we can see that the contrastive distillation approaches result in similar levels of knowledge transfer from the teacher to the student, but appear to also incur much stronger overall overwriting, causing the student to lose large portions of its previous knowledge. While very suitable for distillation to untrained students, this behaviour is unfortunately not well suited for knowledge transfer between already trained expert models.

B.2 Extended results on knowledge transfer between pretrained models

For our knowledge transfer success rate experiments conducted in Section 5.1, we provide an extended and more detailed version for Figure 4a in Figure 7. Using a scatterplot, we relate the share of knowledge transferred to the student model (knowledge gain) versus the share of the students pretrained knowledge that is overwritten during the transfer process (knowledge loss). Each student model is denoted by a respective color choice associated with its parameter count. Symbol sizes and colors denote both family and performance of the respective teacher models. The red line denotes an equal trade-off between knowledge gain and loss, with upper-diagonal entries indicating a positive knowledge transfer. Comparing the results of vanilla *KL-Dist. transfer* and the continual learning based *Kl+DP transfer*, we see that a vast majority of points are pushed up the diagonal, allowing for transfer even from weaker models (small symbols, heavily scattered towards the lower diagonal area in the normal knowledge distillation approach). This behaviour also highlights that normal knowledge distillation approaches generally overwrite knowledge instead of augmenting, and is reflected in our correlation studies in Figure 4a.

Overall, these results simply extend the insights provided in the main part of this work from a more detailed point of view, highlighting that a continual learning treatment of the knowledge transfer problem can significantly raise the transfer success rate. However, we note that this more finegrained perspective *provides better support on the detrimental aspect of stronger visual inductive biases for general knowledge transfer*, as we found CNN students to generally perform worst, even when leveraging *KL+DP transfer*.

Table 7: Knowledge Transfer results on full ImageNet, from four teacher to eight selected student models. The tables include the individual transfer deltas of all teacher-student pairs.

Teachers → ↓ **Students**	$\Delta_{\text{transf.}}$ KL-Dist.				$\Delta_{\text{transf.}}$ KL-Dist.+DP				$\Delta_{\text{transf.}}$ KL-Dist.+DP (unsup.)			
	SWSL-ResNext101	Volo-D2	ResMLP36	CoaT-lite-mini	SWSL-ResNext101	Volo-D2	ResMLP36	CoaT-lite-mini	SWSL-ResNext101	Volo-D2	ResMLP36	CoaT-lite-mini
XCiT-P16	0.95	1.40	-0.55	-0.88	0.93	0.90	0.36	0.42	1.29	1.45	0.57	0.52
PiT-B	1.16	1.42	-0.24	-0.74	0.74	0.86	0.31	0.31	1.35	1.59	0.43	0.45
PiT-XS	0.54	0.43	0.14	-0.71	0.55	0.44	0.37	0.23	0.51	0.53	0.29	0.08
SeNet154	0.38	-0.07	-0.20	-0.32	0.38	0.02	0.22	0.48	0.42	0.13	0.09	0.38
ConvNext	0.23	0.41	-1.10	-1.57	0.49	0.44	0.26	0.12	0.44	0.38	0.22	0.09
ResNetV2	0.34	0.11	-0.23	-0.56	0.32	0. 17	0.28	0.13	0.34	0.18	0.29	0.11
Mixer-B16	0.32	0.22	-0.64	-1.07	0.31	0.22	0.11	-0.05	0.35	0.26	0.07	-0.02
ResMLP-24	0.58	0.43	-0.16	-0.26	0.57	0.45	0.10	0.20	0.57	0.36	0.10	0.13

Table 8: The table below shows the results of knowledge transfer with our proposed KL-Dist. + DP transfer approach on the full ImageNet. It includes two metrics that describe the changes in the positive and negative prediction flips, and extends the information provided in Table 1. For each student, we report the mean and standard deviation over all teacher models, which can be found in Table 6.

Students	Type	Acc.	# Param.	$\Delta_{\text{transf.}}$	$\Delta_{\rho^{pos}}$	$\Delta_{\rho^{neg}}$
XCiT-P16 (El-Nouby et al., 2021)	Trafo	82.89	189.10	**0.65** (±0.26)	-0.74 (±0.25)	-0.08 (±0.04)
PiT-B (Heo et al., 2021)	Trafo	82.44	73.76	**0.55** (±0.25)	-0.64 (±0.28)	-0.08 (±0.05)
PiT-XS (Heo et al., 2021)	Trafo	78.19	10.62	**0.40** (±0.12)	-0.45 (±0.13)	-0.05 (±0.04)
SeNet154 (He et al., 2018)	CNN	81.23	115.09	**0.27** (±0.17)	-0.35 (±0.14)	-0.07 (±0.04)
ConvNext (Liu et al., 2022)	CNN	84.57	50.22	**0.33** (±0.14)	-0.37 (±0.09)	-0.04 (±0.06)
ResNetV2 (He et al., 2016b)	CNN	82.80	25.55	**0.23** (±0.08)	-0.35 (±0.09)	-0.13 (±0.03)
Mixer-B16 (Tolstikhin et al., 2021)	MLP	82.30	59.88	**0.15** (±0.13)	-0.16 (±0.08)	-0.01 (±0.07)
ResMLP-24 (Touvron et al., 2021)	MLP	80.76	30.02	**0.33** (±0.19)	-0.30 (±0.16)	+0.03 (±0.04)

The following table shows the individual transfer deltas of the teacher-student pairs from Table 1 and Table 3.

To further support our analysis in Section 5.1, we have provide additional results regarding the change in the share of positive and negative prediction flips during knowledge transfer. Positive prediction flips ρ^{pos} refer to cases where the teacher was correct, but the student was incorrect. In contrast, negative prediction flips ρ^{neg} refer to cases where the teacher was incorrect, but the student was correct. To measure this change, we defined two new metrics, pos-flips delta $\Delta_{\rho^{pos}}$ and neg-flips delta $\Delta_{\rho^{neg}}$, similar to the transfer delta. We present the mean and standard deviation for both metrics for all student models using our *KL+DP transfer* approach in Table 8, extending the results from Table 1.

Our goal with knowledge transfer is to transfer complementary knowledge, i.e., the positive prediction flips. This means that the number of samples where the teacher is correct but the student is incorrect should decrease as much as possible. However, we must simultaneously preserve the student's previous knowledge. As a result, the number of samples where the student is correct and the teacher is incorrect (negative prediction flips) should not decrease.

The experimental results conclusively demonstrate the effectiveness of our approach in reducing the share of positive prediction flips for all student models. This underlines the capability of our approach to transfer complementary knowledge between models. Moreover, the minor changes in the negative prediction flips provide compelling evidence of the approach's ability to preserve the student's previous knowledge.

B.3 Extended results on the impact of different student model properties on knowledge transfer

In this section, we provide a closer assessment of the impact of the student model properties on the knowledge transfer behaviour, as measured through the transfer delta. In particular, we look at performance, size (measured by the number of parameters) and model family. For this assessment, we selected for each model property pairs or triplets of students with similar values for two of the three properties to isolate each single variable as well as possible. While an exact intervention can not be made by simply leveraging pretrained models, this setup does provide more controlled insights, which we visualize in Figure 8 for experiments conducted on 10% ImageNet using the *KL+DP transfer* approach.

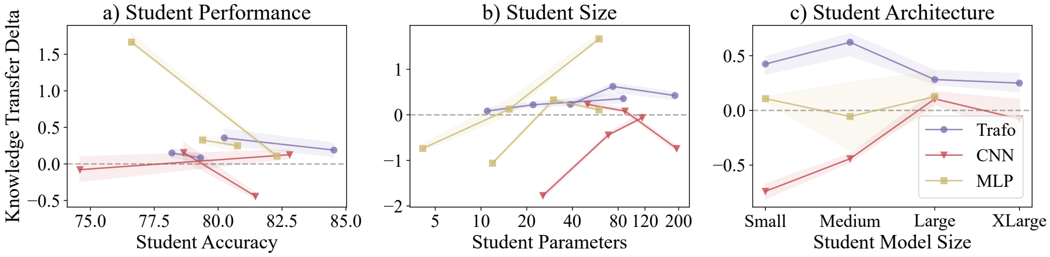


Figure 8: Evaluation of the impact of the student model properties a) performance, b) size (measured by the number of parameters) and c) architecture type on the knowledge transfer delta. Each marker represents a selected student model distilled with 20 different teacher models. We group students into pairs or triplets based on the remaining model properties by connecting the respective markers.

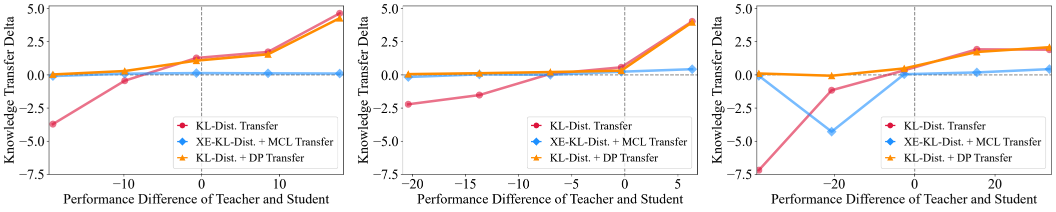
Note that each marker represents *one evaluated student model with all 20 teacher models*. We connect the pairs or triples of students that can be compared, with the color of the lines and markers representing the model family of the student.

Our results replicate insights noted in the main part of this work, particularly Figure 5 (right). We find that even when controlling for other factors such as initial accuracy, the overall student capacity appears strongly correlated with the ability to receive new knowledge without overwriting previous. This is a particularly pronounced behavior in models with strong visual inductive bias such as CNNs. The rightmost subfigure showcases that when looking at the average behavior of a model family (divided into different model sizes), that scale can offer emergent transfer capabilities in CNNs - while not available before - for any type of specific architecture - increased sizes can allow for notably improved transferability.

B.4 Extended results on additional datasets

To substantiate our results on ImageNet we additionally conduct experiments on the CUB200 Wah et al. (2011), Caltech256 Griffin et al. (2007), and Stanford-Cars Krause et al. (2013) datasets.

For each datasets we combine the nine student and four teacher models as shown in Table 6 resulting in a total of 36 teacher-student combination. We fine-tune the classification layer of the student and teacher models using dataset-specific data before initiating knowledge transfer. We employ the dataset's training data as the transfer set.



(a) CUB200

(b) Caltech256

(c) Stanford Cars

Figure 9: Knowledge transfer delta based on teacher-student performance difference for three additional datasets: a) CUB200, b) Caltech256, and c) Stanford Cars. We compare simple *KL-Dist.* transfer with *XE-KL-Dist.+MCL* transfer and *KL-Dist.+DP* Transfer. The teacher-student pairs are categorized into bins determined by equipartitions of their respective performance differences. To mitigate the influence of outliers, we report the mean transfer delta of the top 25% within each bin and approach.

Across all datasets, we consistently observe the *KL-Dist.+DP transfer* approach to not only enable the transfer of knowledge from less proficient teachers without compromising student performance but to also demonstrate the capacity to transfer substantial knowledge portions in cases where the teacher outperforms the student significantly, aligning with the effectiveness of the straightforward *KL-Dist. transfer*. These results are in line with our observations on ImageNet (c.f. Figure 4b) and underline the strengths of *KL+DP transfer*.

B.5 Extended results on knowledge transfer under domain shifts

We further explore knowledge transfer in the setting of a domain shift between the teacher and student model. For this purpose we fine tune the teacher model on the domainnet infograph dataset Peng et al. (2019) before conducting knowledge transfer. The transfer process is executed on the 10% subset of ImageNet. Our comprehensive assessment encompasses a cohort of 9 distinct student models and 4 teacher models (see Table 6).

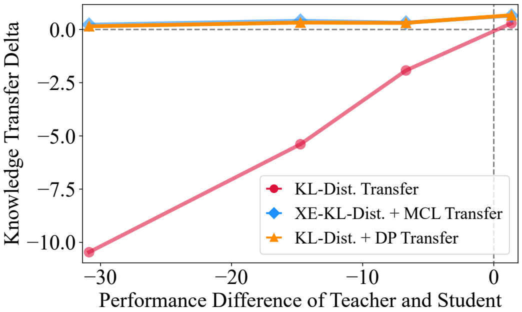


Figure 10: Inter-domain knowledge transfer delta analysis for *KL-Dist.* and *KL-Dist.+DP* transfer. We investigate the transfer delta resulting from knowledge transfer from a teacher model trained on DomainNet Infograph to an ImageNet-pretrained student model.

Notably, our findings underscore the efficacy of the *KL-Dist.+DP transfer* approach, which facilitates the transfer of knowledge from the Infograph-trained teacher to the student model on the ImageNet domain, thereby improving the student's performance. In stark contrast, the conventional *KL-Dist. transfer* demonstrates a substantial decrease in student accuracy, particularly when using a less proficient teacher.

B.6 Extended results on the transfer from multiple teachers

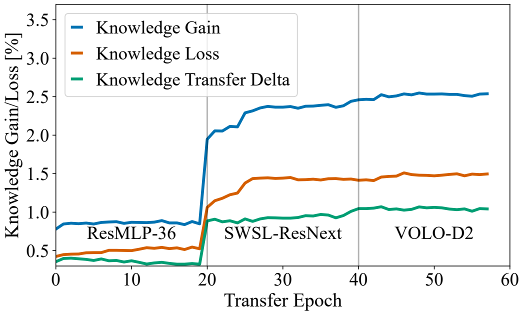


Figure 11: Knowledge transfer (knowledge gain) and loss of the student previous knowledge (knowledge loss) during the sequential training of PiT-B Heo et al. (2021) with three different teacher models sorted by ascending performance.

Finally, we present additional insights on the sequential knowledge transfer from multiple teacher models to a single pretrained student model. For all multi-teacher knowledge transfer experiments we select three student models (XCiT-P16, Twins, PiT-B) and three teacher models (SWSL-ResNext101, VOLO-D2, ResMLP-36) from Tab. 6. Appendix B.6 visualizes the knowledge transfer (knowledge gain), the share of the student's pretrain knowledge being lost (knowledge loss) and the overall transfer delta over the transfer epochs for the PiT-B Heo et al. (2021) student model presented in §5.2. As noted there, we distill the student with three different teacher models (see Table 6). For this particular visualization, we order teachers by ascending performance, but find positive continual transfer also to be achievable from other sequences. For each teacher, we allocate a fixed transfer budget of 20 epochs. As noted already in Table 2, the figure visually highlights that positive transfer deltas can be gained going from one teacher to the subsequent one (stronger transfer delta compared to the strongest single student, $\Delta_{\text{dist}} = 1.04$), but with returns diminishing. We can attribute this to the increased rate of forgetting - while knowledge gain is steadily rising, continuously moving the student from its initial pretraining weights induces increasingly stronger knowledge loss, even when leveraging *Kl+DP transfer*.

For further insights, we compare the results of our multi-teacher experiments using *KL-Dist.+DP transfer* to vanilla *KL-Dist. transfer* (Tab. 9). The results clearly show that sequential *KL-Dist.*

Table 9: *Knowledge transfer from multiple teachers* into a pretrained student using sequential, and soup-based vanilla KL-Dist. transfer (c.f. §4.3). We compare with transfer deltas obtained from the single teacher knowledge transfer.

Students	Type	Acc.	# Param.	$\Delta_{\text{transf.}}$ Single Teacher			$\Delta_{\text{transf.}}$ Multiple Teachers	
				Mean	Min	Max	Sequ.	Soup
XCiT-P16 El-Nouby et al. (2021)	Transf.	82.89	189.1	0.47	-0.85	**1.31**	0.48	0.89
Twins Chu et al. (2021)	Transf.	83.68	99.27	-0.04	-1.04	**0.63**	0.01	0.43
PiT-B Heo et al. (2021)	Transf.	82.44	73.76	0.69	-0.24	**1.39**	0.80	1.19

transfer cannot achieve larger gains as the best teacher alone but results in performance gains in the range of the average transfer delta across the three teachers. This again shows that rather than transferring only the complementary knowledge vanilla *KL-Dist. transfer* overwrites the student's previous knowledge with the knowledge of the teacher model. Thus when sequentially transferring knowledge from multiple teachers improvements from the previous transfer are lost during transfer from the subsequent teacher. Note that the vanilla *KL-Dist. transfer* approach cannot be directly applied to transfer knowledge from multiple teacher models in parallel, hence we omit this baseline.