Supplemental Material for Generalized Cross-Task Knowledge Distillation via Relationship Matching

Han-Jia Ye, Su Lu, De-Chuan Zhan

In the supplemental material, we provide

- Detailed derivations to interpret the comparison matching (*c.f.* Section 4.2.2 in the main paper)
- Detailed definition of harmonic mean for Generalized Knowledge Distillation (GKD) evaluation (*c.f.* Section 5.1 in the main paper)
- More experiments and analyses of REFILLED (*c.f.* Section 5 in the main paper)

1 Interpretation of Comparison Matching

The main idea of comparison matching in section 4.2.1 in the main paper is to align the similarity predictions over tuples between teacher and student. Recall that given a tuple $(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{P}}, \mathbf{x}_{i1}^{\mathcal{N}}, \ldots, \mathbf{x}_{iK}^{\mathcal{N}})$ with one target neighbor $\mathbf{x}_i^{\mathcal{P}}$ and K impostors $\{\mathbf{x}_{i1}^{\mathcal{N}}, \ldots, \mathbf{x}_{iK}^{\mathcal{N}}\}$, we optimize the embedding through

$$\min_{\phi} \sum_i \mathbf{KL}\left(p_i(\phi_T) \parallel p_i(\phi)\right) \ . \tag{1}$$

Here $p_i(\phi_T)$ and $p_i(\phi)$ are the validness probability based on teacher's and student's embedding ϕ_T and ϕ, respectively.

We provide the concrete derivation for Eq. 6 in the main paper to explain the effect of comparison matching and show how the teacher's estimation of the tuples influences the embedding optimization. We first illustrate the idea in a triplet form — an anchor $\mathbf{x}_i$ with one positive neighbor $\mathbf{x}_i^{\mathcal{P}}$ and *one negative impostor* $\mathbf{x}_i^{\mathcal{N}}$, which is easy to explain. Then we extend the analysis to the general case.

When there is only one impostor in the tuple, we can simplify the term $p_i(\phi)$ in Eq. 3 in the main paper as

$$\sigma\left(\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{N}}\right) - \mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{P}}\right)\right) = \sigma\left(\mathbf{Diff}_{\mathbf{x}_i}\right) \ . \tag{2}$$

where $\sigma(x) = 1/(1+\exp(-x))$ is the logistic function squashing the input into $[0, 1]$. We define $\mathbf{Diff}_{\mathbf{x}_i} = \mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{N}}\right) - \mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{P}}\right)$, $\rho_i = 1 - p_i(\phi_T)$, and logistic loss $\iota(x) = \ln(1 + \exp(-x))$. In the vanilla case, we can obtain the embedding via minimizing the triplets with the logistic loss, which pushes impostor $\mathbf{x}_i^{\mathcal{N}}$ away and pulls the target neighbor $\mathbf{x}_i^{\mathcal{P}}$ close.

We can reformulate Eq. 4 in the main paper as

$$\begin{aligned}
&\mathbf{KL}\left(p_i(\phi_T) \parallel p_i(\phi)\right) \qquad\qquad (3)\\
= \ & p_i(\phi_T) \ln \frac{p_i(\phi_T)}{p_i(\phi)} + (1 - p_i(\phi_T)) \ln \frac{1 - p_i(\phi_T)}{1 - p_i(\phi)} \\
= \ & \underbrace{p_i(\phi_T) \ln p_i(\phi_T)}_{\text{constant}} - p_i(\phi_T) \ln p_i(\phi) \\
& + \ \underbrace{(1 - p_i(\phi_T)) \ln (1 - p_i(\phi_T))}_{\text{constant}} \\
& - \ (1 - p_i(\phi_T)) \ln (1 - p_i(\phi))
\end{aligned}$$

Then we have

$$\begin{aligned}
&\mathbf{KL}\left(p_i(\phi_T) \parallel p_i(\phi)\right) \\
\cong \ & - p_i(\phi_T) \ln p_i(\phi) \ - \ \ln(1 - p_i(\phi)) \\
& + \ p_i\left(\phi_T\right) \ln\left(1 - p_i\left(\phi\right)\right) \\
= \ & - p_i\left(\phi_T\right)\left(-\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{P}}\right) - \ln \Delta\right) + \ln \Delta \\
& + \ \mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{N}}\right) + p_i\left(\phi_T\right)\left(-\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{N}}\right) - \ln \Delta\right)
\end{aligned}$$

The notation $\cong$ neglects the constant term in the equation. We set $\Delta \triangleq \exp\left(-\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{P}}\right)\right) + \exp\left(-\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{N}}\right)\right)$. Thus,

$$\begin{aligned}
&\mathbf{KL}\left(p_i(\phi_T) \parallel p_i(\phi)\right) \\
= \ & p_i(\phi_T)\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{P}}\right) + \left(1 - p_i\left(\phi_T\right)\right)\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{N}}\right) + \ln \Delta \\
= \ & p_i\left(\phi_T\right)\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{P}}\right) + \left(1 - p_i\left(\phi_T\right)\right)\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{N}}\right) \\
& + \ \ln\left(\exp\left(-\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{P}}\right)\right) + \exp\left(-\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{N}}\right)\right)\right) \\
= \ & \left(p_i(\phi_T) - 1\right)\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{P}}\right) + \left(1 - p_i\left(\phi_T\right)\right)\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{N}}\right) \\
& + \ \ln\left(1 + \exp\left(- \left(\mathbf{Diff}_{\mathbf{x}_i}\right)\right)\right) \\
= \ & \rho_i\left(\mathbf{Diff}_{\mathbf{x}_i}\right) + \iota\left(\mathbf{Diff}_{\mathbf{x}_i}\right) \ .
\end{aligned}$$

At last, minimizing the KL-divergence equals minimizing two losses. The first part is a rectification term and the second part is the vanilla logistic loss. Therefore, the comparison matching weakens the effect of the student's embedding updates from the label supervision by incorporating the

- *H.-J. Ye, L. Han, and D.-C. Zhan are with State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China.*
E-mail: {yehj, lus, zhandc}@lamda.nju.edu.cn

teacher's knowledge. More discussions are in Section 4.4.1 in the main paper.

In the following, we extend the previous analysis to the general case when $K > 1$. Based on Eq. 1, we have

$$\begin{aligned} &\mathbf{KL}\left(p_i(\phi_T) \,\|\, p_i(\phi)\right) \\ \cong\ &\underbrace{\mathbf{KL}\left(p_i(\phi_T) - \mathbf{e}_0 \,\|\, p_i(\phi)\right)}_{\text{rectification term}} + \underbrace{\mathbf{KL}\left(\mathbf{e}_0 \,\|\, p_i(\phi)\right)}_{\text{contrastive loss}} . \end{aligned} \quad (4)$$

We set $\mathbf{e}_0 = [1, 0, \ldots, 0]$ as an all-zero value vector except for the first element, whose size is the same as $p_i(\phi)$. Particularly, we have

$$\begin{aligned} &\mathbf{KL}\left(\mathbf{e}_0 \,\|\, p_i(\phi)\right) \\ \cong\ & -\ln \mathbf{s}_\tau\left(\left[-\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{P}}\right), \ldots, -\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_{iK}^{\mathcal{N}}\right)\right]\right) \\ =\ & -\ln \frac{\exp\left(-\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{P}}\right)\right)}{\exp\left(-\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{P}}\right)\right) + \sum_{k=1}^{K} \exp\left(-\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_{ik}^{\mathcal{N}}\right)\right)} \\ =\ & \ln\left(1 + \sum_{k=1}^{K} \exp\left(\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{P}}\right) - \mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_{ik}^{\mathcal{N}}\right)\right)\right) . \end{aligned} \quad (5)$$

By minimizing the KL-divergence between the student's similarity prediction $p_i(\phi)$ with $\mathbf{e}_0$, we get a multi-impostor extension of the logistic loss. With a bit abuse of notation, we also denote the loss in Eq. 5 as $\iota\left(\mathbf{Diff}_{\mathbf{x}_i}\right)$. Eq. 5 promotes the probability between the anchor and the similar term to be the largest one — the similarity between the anchor and the target neighbor becomes larger than those between the anchor and impostors. The comparison matching adds another rectification term, *i.e.*, the first term in Eq. 4. If the teacher's prediction is close to the similarity supervision indicated by the label, then $p_i(\phi_T)$ should be close to $\mathbf{e}_0$ — the first term approaches to zero and Eq. 4 degenerates to the vanilla loss.

On the contrary, if the teacher's comparison over a tuple is not consistent with their relationship indicated by the class labels, *e.g.*, those impostors should not be too distant from the anchor and the anchor should have relative close distance between target neighbor and impostors. In this case, $p_i(\phi_T) - \mathbf{e}_0$ becomes a vector with all negative elements except for the first one. We can analyze the effect of the rectification term by decoupling the influence of the first and other elements in $p_i(\phi_T) - \mathbf{e}_0$. We define

$$p_i^{\mathcal{P}}(\phi) = \frac{\exp\left(-\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{P}}\right)\right)}{\exp\left(-\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{P}}\right)\right) + \sum_{k=1}^{K} \exp\left(-\mathbf{D}_\phi\left(\mathbf{x}_i, \mathbf{x}_{ik}^{\mathcal{N}}\right)\right)} ,$$

then we can derive $(p_i^{\mathcal{P}}(\phi) - 1) \cdot \iota\left(\mathbf{Diff}_{\mathbf{x}_i}\right)$ from the first element of the rectification term, which weakens the force of the contrastive loss. In addition, the other elements in the rectification term make the impostors in the tuples not far away. Thus, the rectification term mitigates the supervision from the labels and injects the teacher's similarity estimation of the tuple.

2 Harmonic Mean for GKD

There are two sets of classes, *i.e.*, the old classes $\mathcal{C}'$ used to train the teacher, and the current classes $\mathcal{C}$ in the student's task, in the cross-task distillation scenario. Given a model discerning the joint classes in $\mathcal{C}' \cup \mathcal{C}$, we collect the same number of instances from $\mathcal{C}$ and $\mathcal{C}'$ to construct the test set.

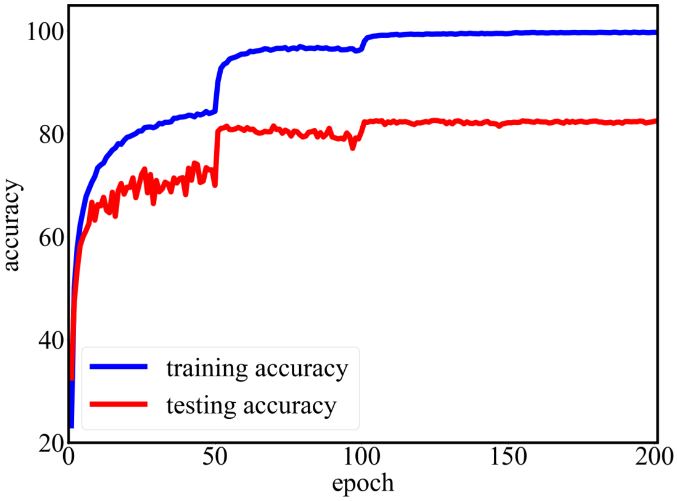


Figure 1: Convergence curve of ReFilled on CIFAR.

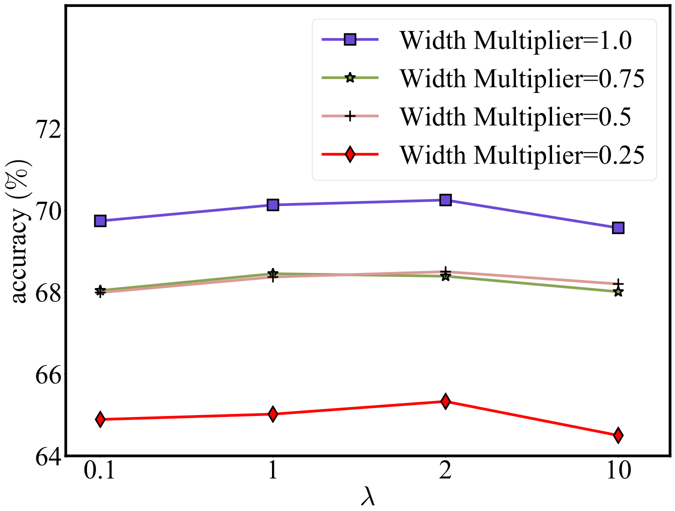


Figure 2: ReFilled's accuracy on CUB dataset width different λ values. Class overlap ratio is set to 50%. Four student networks are considered, *i.e.*, MobileNet-{1.0, 0.75, 0.5, 0.25}.

For an instance $\mathbf{x} \in \mathcal{C}' \cup \mathcal{C}$, we may measure the performance of the model by averaging the classification accuracy over $\mathcal{C}' \cup \mathcal{C}$ for all instances in the test set. We abbreviate this process as $\mathcal{C}' \cup \mathcal{C} \rightarrow \mathcal{C}' \cup \mathcal{C}$.

Directly computing the average classification accuracy will be biased towards the current classes, since the student model has no access to the old class instances. Similar to [1], [2], we consider the harmonic mean accuracy as a more balanced measure, which is a joint measure over two kinds of mean accuracy. First, we measure the joint classification accuracy $\mathrm{Acc}_{\mathcal{C}'}$ for instances from the class set $\mathcal{C}'$, *i.e.*, $\mathcal{C}' \rightarrow \mathcal{C}' \cup \mathcal{C}$. Then we compute $\mathrm{Acc}_{\mathcal{C}}$ as the mean accuracy based on $\mathcal{C} \rightarrow \mathcal{C}' \cup \mathcal{C}$. The harmonic mean accuracy is

$$\frac{2\mathrm{Acc}_{\mathcal{C}'}\mathrm{Acc}_{\mathcal{C}}}{\mathrm{Acc}_{\mathcal{C}'} + \mathrm{Acc}_{\mathcal{C}}} .$$

A model has a high harmonic mean value only it has a relative higher $\mathrm{Acc}_{\mathcal{C}'}$ and $\mathrm{Acc}_{\mathcal{C}}$. In other words, the harmonic mean measures the balanced classification ability over both old and current classes.

Overlap Ratio = 0%				
Width Multiplier	1	0.75	0.5	0.25
Student	71.25	67.56	66.85	64.48
RKD	72.24	68.42	66.85	65.74
AML	72.86	68.79	68.59	66.83
ReFilled$^{\text{EMB}}$	71.88	68.02	67.40	65.12
ReFilled^{-}	74.07	70.93	70.62	67.58
ReFilled	75.13	71.67	71.06	68.22

Overlap Ratio = 25%				
Width Multiplier	1	0.75	0.5	0.25
Student	71.30	71.08	68.56	65.71
RKD	72.07	71.70	68.56	66.43
AML	72.35	72.05	70.37	67.20
ReFilled$^{\text{EMB}}$	72.04	71.36	68.90	66.29
ReFilled^{-}	74.14	73.09	72.33	68.96
ReFilled	75.09	73.92	72.99	70.04

Overlap Ratio = 50%				
Width Multiplier	1	0.75	0.5	0.25
Student	68.20	66.11	65.23	62.26
RKD	68.72	66.82	65.58	62.79
AML	67.94	67.34	66.29	63.64
ReFilled$^{\text{EMB}}$	68.79	66.56	65.68	62.94
ReFilled^{-}	69.14	67.43	67.18	64.62
ReFilled	70.25	68.39	68.50	65.33

Overlap Ratio = 75%				
Width Multiplier	1	0.75	0.5	0.25
Student	65.53	66.73	64.10	60.81
RKD	65.89	67.28	64.66	61.35
AML	66.32	66.92	65.03	62.09
ReFilled$^{\text{EMB}}$	66.47	67.20	64.56	61.37
ReFilled^{-}	67.01	67.93	66.10	62.35
ReFilled	67.28	68.35	66.72	63.03

Overlap Ratio = 100%				
Width Multiplier	1	0.75	0.5	0.25
Student	67.76	67.98	64.91	62.17
RKD	67.23	68.25	65.73	62.04
AML	67.06	68.35	66.27	62.69
ReFilled$^{\text{EMB}}$	68.13	68.34	65.72	63.05
ReFilled^{-}	68.96	69.07	68.53	63.10
ReFilled	68.77	69.10	68.44	63.33

Table 1: Mean accuracy on GKD tasks upon CUB. MobileNets with different width multipliers in $\{1, 0.75, 0.5, 0.25\}$ are used as students. The teacher is MobileNet with width multiplier 1. Different overlap ratios $\{0\%, 25\%, 50\%, 75\%, 100\%\}$ between teacher's and student's class set are considered.

3 More Experiment Results

Full results for GKD. We list the full results of GKD evaluations on CUB and CIFAR-100 in Table 1 and Table 2 respectively. These results are in correspondence with those in Figure 5(a) - Figure 5(h). From these two tables, we can see that ReFilled outperforms other comparison methods

Overlap Ratio = 0%				
(depth, width)	(40, 2)	(16, 2)	(40, 1)	(16, 1)
Student	81.02	78.94	78.98	73.70
RKD	81.46	79.23	78.80	73.45
AML	79.99	79.11	78.99	73.68
ReFilled$^{\text{EMB}}$	81.35	79.43	79.34	74.02
ReFilled^{-}	81.79	79.74	79.60	74.02
ReFilled	82.60	80.70	80.18	74.42

Overlap Ratio = 20%				
(depth, width)	(40, 2)	(16, 2)	(40, 1)	(16, 1)
Student	80.34	76.12	75.43	70.84
RKD	80.56	76.77	75.82	71.05
AML	80.05	76.95	75.37	70.79
ReFilled$^{\text{EMB}}$	79.89	76.45	75.99	71.35
ReFilled^{-}	80.90	77.32	76.95	71.53
ReFilled	81.40	77.82	77.24	72.28

Overlap Ratio = 40%				
(depth, width)	(40, 2)	(16, 2)	(40, 1)	(16, 1)
Student	78.86	75.67	74.98	69.36
RKD	79.33	75.58	74.69	69.07
AML	79.98	75.84	74.28	68.87
ReFilled$^{\text{EMB}}$	79.27	76.03	75.38	69.78
ReFilled^{-}	79.84	76.16	75.30	69.58
ReFilled	80.40	76.26	75.62	70.08

Overlap Ratio = 60%				
(depth, width)	(40, 2)	(16, 2)	(40, 1)	(16, 1)
Student	78.90	76.37	75.14	68.48
RKD	78.69	76.20	75.50	68.23
AML	79.32	76.45	75.23	68.64
ReFilled$^{\text{EMB}}$	79.33	76.90	75.67	69.78
ReFilled^{-}	80.02	76.66	75.79	68.72
ReFilled	80.66	76.66	76.52	69.50

Overlap Ratio = 80%				
(depth, width)	(40, 2)	(16, 2)	(40, 1)	(16, 1)
Student	80.50	77.43	76.96	72.16
RKD	81.21	77.65	77.34	72.05
AML	81.06	77.20	77.06	72.35
ReFilled$^{\text{EMB}}$	80.70	77.79	77.54	72.46
ReFilled^{-}	81.92	78.24	78.33	73.54
ReFilled	82.56	78.76	79.28	73.92

Overlap Ratio = 100%				
(depth, width)	(40, 2)	(16, 2)	(40, 1)	(16, 1)
Student	80.66	77.94	76.35	71.56
RKD	80.52	78.03	76.82	72.04
AML	80.73	78.24	77.15	71.79
ReFilled$^{\text{EMB}}$	81.02	78.32	76.67	72.08
ReFilled^{-}	81.74	78.74	77.77	73.06
ReFilled	81.58	78.60	78.04	73.52

Table 2: Mean accuracy on GKD tasks upon CIFAR-100. Wide ResNets with different depth and width parameters are used as students, and teacher is Wide ResNet with depth 40 and width 2. Different overlap ratios in $\{0\%, 20\%, 40\%, 60\%, 80\%, 100\%\}$ are considered. ReFilled outperforms other comparison methods and baselines.

(depth, width)	(16, 4)	(28, 2)	(16, 2)
Student	77.28	75.12	72.68
KD [3]	78.31	76.57	73.53
FitNet [4]	78.15	76.06	73.7
AT [5]	77.93	76.20	73.44
Jacobian [6]	77.82	76.3	73.29
Overhaul [7]	79.11	78.02	75.92
REFILLED	**79.95**	**79.04**	**76.33**

Table 3: The average classification accuracy of standard KD on CIFAR-100. The teacher is trained with Wide ResNet with depth 28 and width 4, which gets 78.91% test accuracy. The student is learned with Wide ResNet with different configurations of (depth, width).

and baseline methods when class overlap ratio ranges in $[0\%, 100\%]$. REFILLED with adaptive weights improves better than the degenerated version REFILLED$^-$ when the overlap ratio is low.

Comparisons with more methods on standard KD. We compare with more methods on standard KD. We follow the setups in [7] on CIFAR-100. We set the teacher as a Wide ResNet with depth 28 and width 4. The test accuracy of the teacher is 78.91%. Different architectures of the Wide ResNets student are investigated, namely, (16,4), (28,2), and (16-2). Table 3 shows the classification accuracy. REFILLED improves the student and gets the best results in all cases.

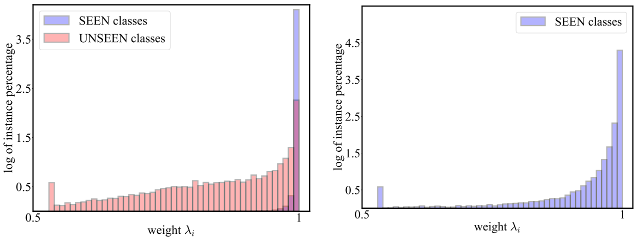
Number of negative instances K. In REFILLED, we formulate the similarity relationship between an instance $\mathbf{x}_i$ and other instances into a tuple, *i.e.*, $(\mathbf{x}_i, \mathbf{x}_i^{\mathcal{P}}, \mathbf{x}_{i1}^{\mathcal{N}}, \ldots, \mathbf{x}_{iK}^{\mathcal{K}})$. Here K is the number of negative instances in a tuple. In experiments, we set K to the minimum number of available negative instances of each anchor point in a mini-batch. We discuss the influence of K on model performance.

We set $K = \min(K', \hat{K})$ where K' is the number of semi-hard negatives for a pair of anchor and neighbor instances. $\hat{K}$ is a manually tuned hyper-parameter. If $\hat{K}$ is set to 1, tuples sampled in each mini-batch degenerate to triplets, which means we keep only one impostor ($K = 1$) for each comparison tuple. On the other hand, if $\hat{K}$ is set to a very large value, K always equals K'. In this case, there are a lot of negative instances in a tuple. Table 4 shows embedding quality on CUB dataset when overlap ratio is 50% and 0%. We set both teacher and student as MobileNet-1.0. We report the accuracy of NCM classifier based on student's embeddings after the first stage distillation. A larger K' value means we use more negative instances in a tuple. The results indicate that using more impostors in a tuple facilitates the distillation of the embedding.

Convergence curve. We plot convergence curve of REFILLED's training process on CIFAR-100 in Figure 1. We set the overlap ratio to 60%. The student is WRN-(40,2). We find that REFILLED successfully converges to a stable point. Convergence curves under other experiment configurations are similar to the plotted one.

$\hat{K}$	1 (triplet)	5	∞ (tuple)
overlap ratio = 0%	30.54	31.42	**32.69**
overlap ratio = 50%	29.61	29.23	**30.27**

Table 4: Quality of student's embedding trained with different $\hat{K}$ on CUB dataset. A larger $\hat{K}$ means there are more impostors in a tuple. Both teacher and student are MobileNet-1.0. Embeddings trained with tuples outperforms those trained with triplets.



(a) Overlap ratio = 60%. (b) Overlap ratio = 100%.

Figure 3: The histogram of λ_i over all target task instances with a learned REFILLED model. We set $\lambda = 1$ and get $\lambda_i \in (0, 1]$. The classes the teacher trained on are denoted as "seen", otherwise denoted as "unseen". (a) When the class overlap ratio is 60%, the weights of instances from seen classes are observably higher than those of unseen classes, and it is easy to recognize unseen classes. (b) When the ratio is 100%, almost all instances have weights close to 1.

Influence of the hyper-parameter λ. We study the influence of hyper-parameter λ. Figure 2 shows REFILLED's testing accuracy on CUB with different λ values, and the class overlap ratio is set to 50%. Various architectures of the student network are considered. The results indicate that REFILLED achieves the highest performance when $\lambda = 2$, and the influence of λ on the model's performance is limited.

The effect of the adaptive weight λ_i. The instance-specific weights λ_i in Eq. 9 in the main paper is a key component for GKD. Since there simultaneously exist cross-task instances (denoted as unseen) and same-task instances (denoted as seen) in the student's training set, and the teacher may predict with different confidences to them. For example, the teacher will have higher confidence over those seen instances where the teacher is trained from, and lower confidence otherwise.

We investigate the distribution of λ_i in two GKD scenarios on CIFAR-100 where the class overlap is 60% and 100%. The distribution (histogram) of λ_i based on a learned REFILLED model in GKD is shown in Figure 3a. We set λ to 1 so that we have $\lambda_i \in (0, 1]$. We find weights for instances belonging to the seen classes are observably larger than those for instances belonging to unseen classes. This means our re-weight strategy can successfully recognize seen/unseen classes and put larger weights on those familiar instances automatically. We compute the AUC when we use λ_i to differentiate seen and unseen classes in the GKD scenario, whose value is 0.959. Thus, λ_i extracts useful information from the teacher's supervision adaptively.

We also plot the results on standard KD in Figure 3b. In standard KD, all instances come from seen classes, and the learned REFILLED model uses larger weights for most instances. There are still a small number of instances that have weights $\lambda_i \approx 0.5$, which may be noisy ones that the teacher cannot provide confident predictions. The teacher's supervision will be weakened with a relatively smaller λ_i.

References

[1] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning - the good, the bad and the ugly," in *CVPR*, 2017, pp. 3077–3086.

[2] H.-J. Ye, H. Hu, and D.-C. Zhan, "Learning adaptive classifiers synthesis for generalized few-shot learning," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1930–1953, 2021.

[3] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.

[4] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *ICLR*, 2015.

[5] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *ICLR*, 2017.

[6] S. Srinivas and F. Fleuret, "Knowledge transfer with jacobian matching," in *ICML*, 2018, pp. 4730–4738.

[7] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *ICCV*, 2019, pp. 1921–1930.