# Knowledge Distillation with the Reused Teacher Classifier

**Supervisor:** Diego Coello de Portugal

**Team Members:**
Wut Hmone Hnin Hlaing
Steffen Roeber
Apurv Kumar
Abin Baby
Brhanu Atsbaha

# Agenda

- Knowledge Distillation
- Base paper approach
- Dataset
- Experiment Results
- Discussion
- Conclusion

# Knowledge distillation

- Process of transferring the knowledge from a large model to smaller model
- Achieve similar performance while reducing computation
- Object detection, semantic segmentation, training of transformers

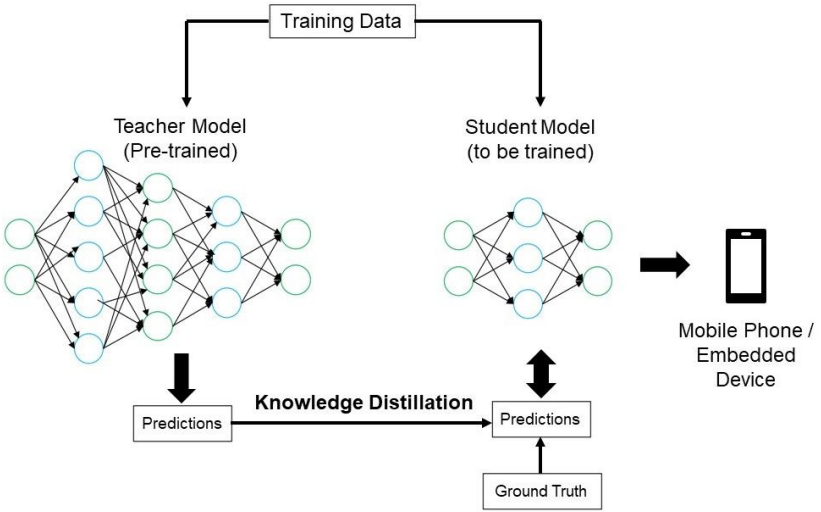
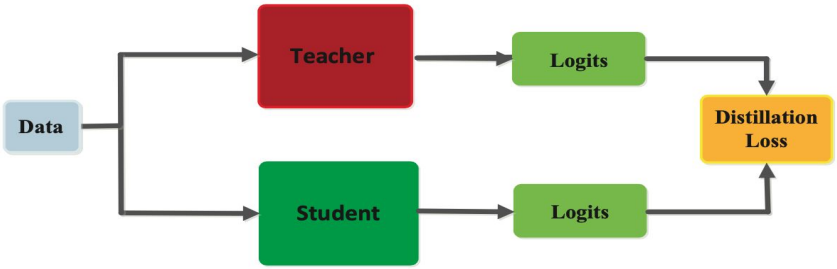
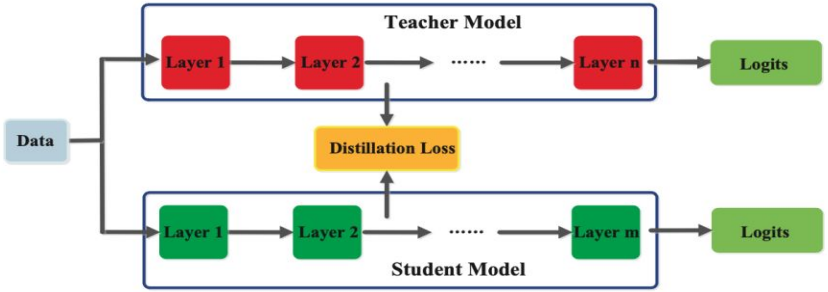



Image src: https://www.analyticsvidhya.com/blog/2022/01/knowledge-distillation-theory-and-end-to-end-case-study/

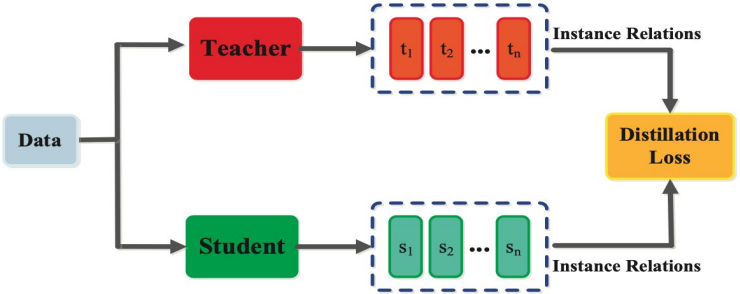
# Types of Knowledge Distillation




Response-based knowledge




Feature-based knowledge

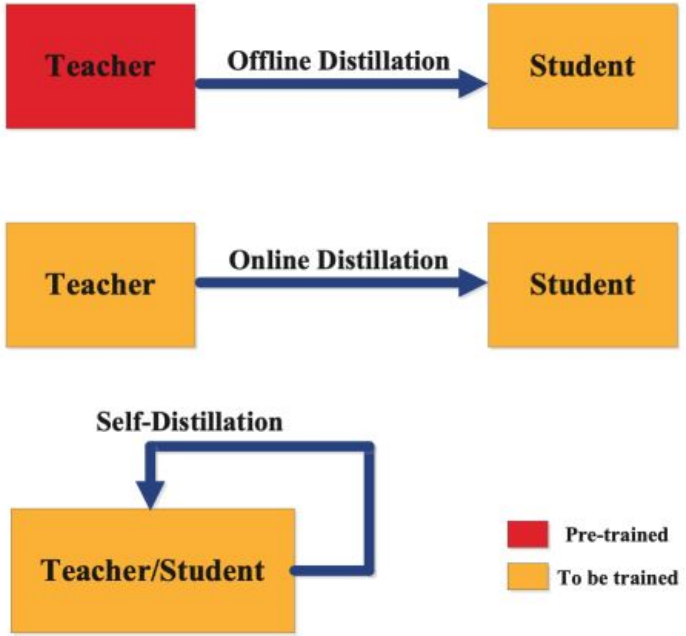



Relation-based knowledge

Source of images: https://neptune.ai/blog/knowledge-distillation

# Training types

- Offline distillation
- Online distillation
- Self-distillation

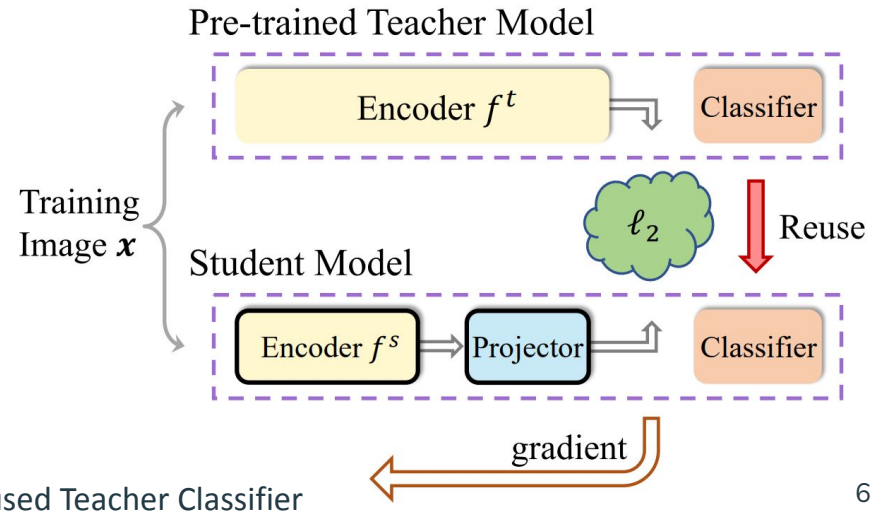



Img src: https://neptune.ai/blog/knowledge-distillation

# Base Paper approach (SimKD)

- Feature-based offline distillation
- Align deep features before the classification layer
- Uses a convolution projector to align feature dimensions, which renders the technique applicable to various teacher and student architectures.
- Re-using the pre-trained classifier from the teacher

$$\mathcal{L}_{\text{SimKD}} = \|\boldsymbol{f}^t - \mathcal{P}(\boldsymbol{f}^s)\|_2^2$$




Source: Chen et al (2022) Knowledge Distillation with the Reused Teacher Classifier

# Base Paper approach (SimKD) continued...

**Teacher model architectures**: WRN-40-2 ResNet-32x4 ResNet-110x2, etc.

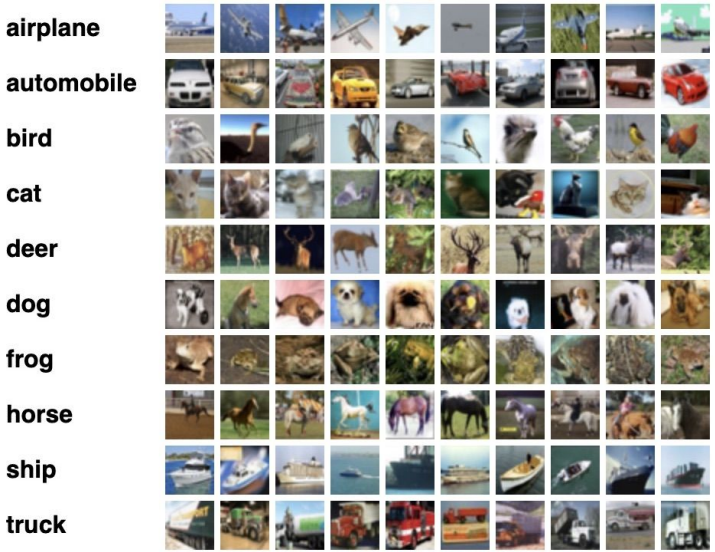
**Student model architectures**: WRN-40-1, ResNet-8x4, ResNet-110, ShuffleNetV2, MobileNetV2 etc.

| Student | WRN-40-1 | ResNet-8x4 | ResNet-110 | ResNet-116 | VGG-8 | ResNet-8x4 | ShuffleNetV2 |
|---|---|---|---|---|---|---|---|
| | 71.92 ± 0.17 | 73.09 ± 0.30 | 74.37 ± 0.17 | 74.46 ± 0.09 | 70.46 ± 0.29 | 73.09 ± 0.30 | 72.60 ± 0.12 |
| SimKD | **75.56 ± 0.27** | **78.08 ± 0.15** | **77.82 ± 0.15** | **77.90 ± 0.11** | **75.76 ± 0.12** | **76.75 ± 0.23** | **78.39 ± 0.27** |
| Teacher | WRN-40-2 | ResNet-32x4 | ResNet-110x2 | ResNet-110x2 | ResNet-32x4 | WRN-40-2 | ResNet-32x4 |
| | 76.31 | 79.42 | 78.18 | 78.18 | 79.42 | 76.31 | 79.42 |

Source: Chen et al. (2022) Knowledge Distillation with the Reused Teacher Classifier

# Dataset

- Cifar-100 (consisting of 60,000 32x32 colour images with a total of 100 target classes)

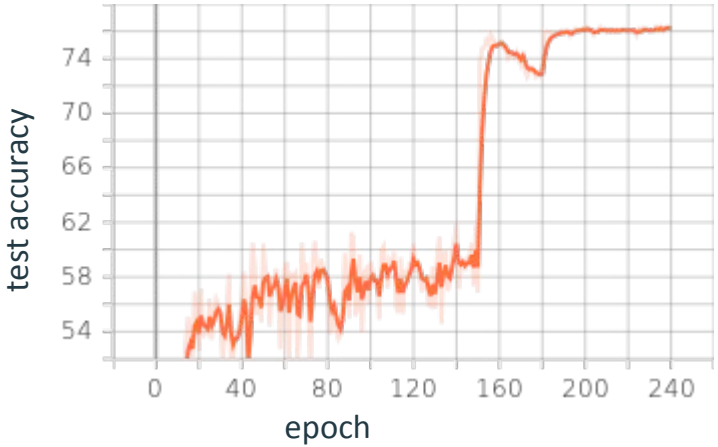



Source: https://paperswithcode.com/dataset/cifar-100

# Experiment: Reproduce Baseline

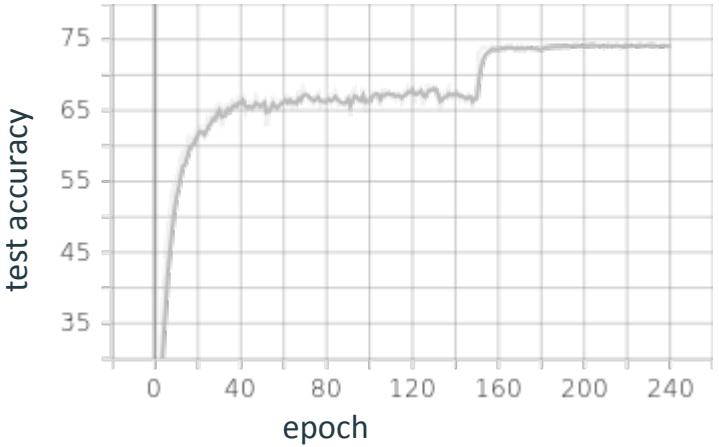
- Teacher: WRN-40-2, Student: **ResNet-8x4**
- Paper Baseline: Reproduce results (SimKD)

Teacher Test Acc: **76.3**




**Test accuracy during training**

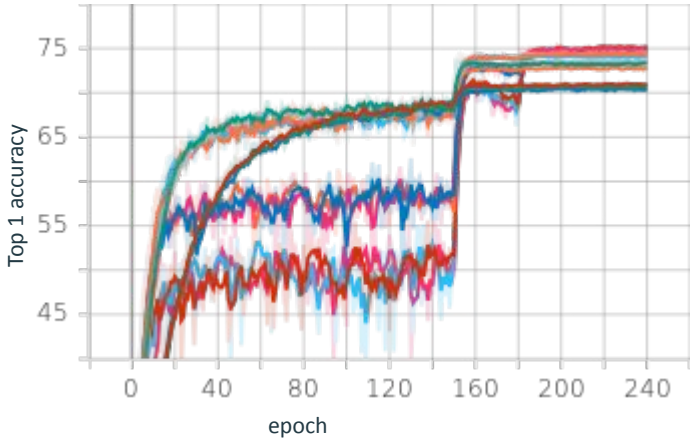
Student Test Acc: **76.02**



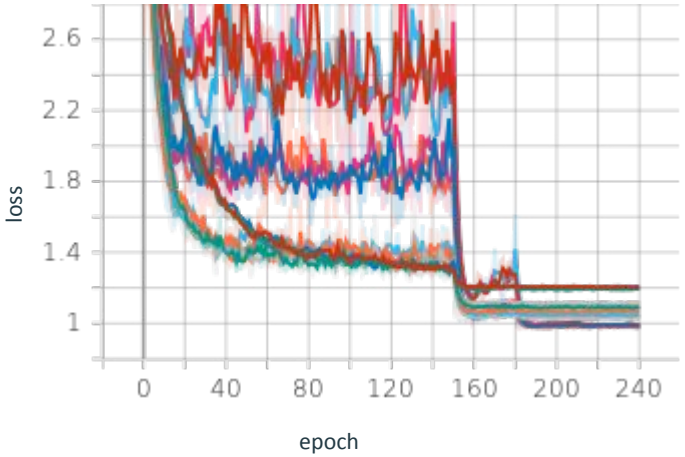

**Test accuracy during testing**

# Cosine Similarity - Detail

- Loss : **cosine similarity,** learning rate: **(0.001, 0.005, 0.01, 0.05, 0.1)**
- Epoch: **240**, weight delay epoch: **[150, 180, 210]**, Trial - **3 trial for each**
- Teacher - **WRN-40-2**, Student: **ResNet-8x4**




**Test accuracy during training**




**Test loss during training**

# Cosine Similarity - Detail

- Loss : **cosine similarity,** learning rate: **(0.001,0.005, 0.01, 0.05, 0.1)**
- Epoch: **240**, weight delay epoch: **[150, 180, 210]**
- Teacher - **WRN-40-2**, Student: **ResNet-8x4**

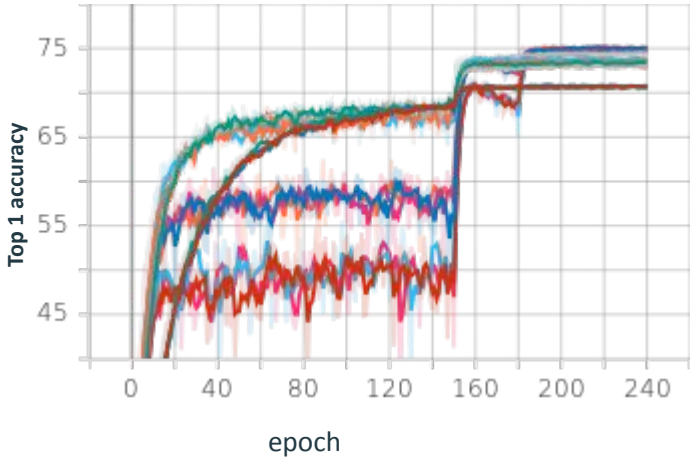

**Learning rate vs test top 1 accuracy**

Teacher Test Acc: **76.3**

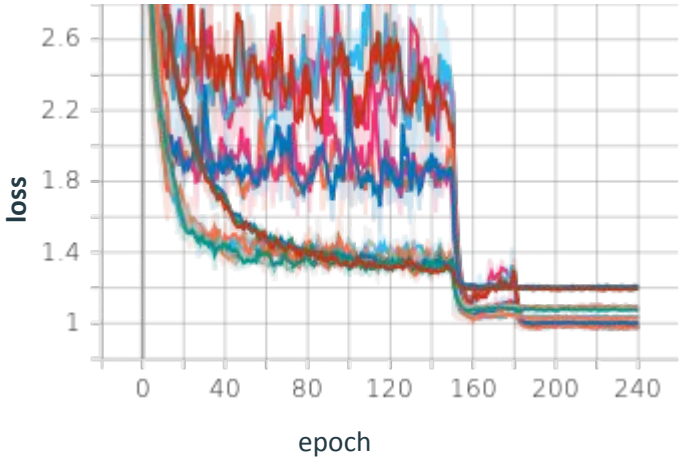
| Learning rate | Student Test Acc |
|---|---|
| 0.001 | 70.89 ± 0.21 |
| 0.005 | 73.48 ± 0.44 |
| 0.01 | 74.49 ± 0.17 |
| 0.05 | 74.95 ± 0.30 |
| **0.1** | **75.05 ± 0.34** |

# KL Divergence - Detail

- Loss : **KL divergence,** learning rate: **(0.001, 0.005, 0.01, 0.05, 0.1)**
- Epoch: 240, weight decay epoch: **[150, 180, 210]**, Trial - **3 trial for each**
- Teacher - **WRN-40-2**, Student: **ResNet-8x4**




**Test accuracy during training**




**Test loss during training**

# KL Divergence - Detail

- Loss : **KL divergence,** learning rate: (0.001,0.005, 0.01, 0.05, 0.1)
- Epoch: 240, weight delay epoch: [150, 180, 210]
- Teacher - **WRN-40-2**, Student: **ResNet-8x4**

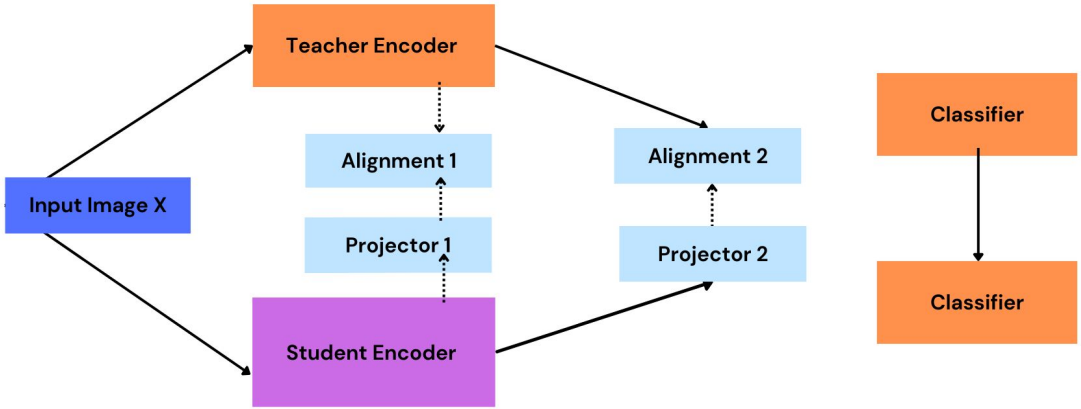

**Learning rate vs test top 1 accuracy**

Teacher Test Acc: **76.3**

| Learning rate | Student Test Acc |
|---|---|
| 0.001 | 70.95 ± 0.07 |
| 0.005 | 73.38 ± 0.32 |
| 0.01 | 74.17 ± 0.4 |
| **0.05** | **75.27 ± 0.05** |
| 0.1 | 75.03 ± 0.08 |

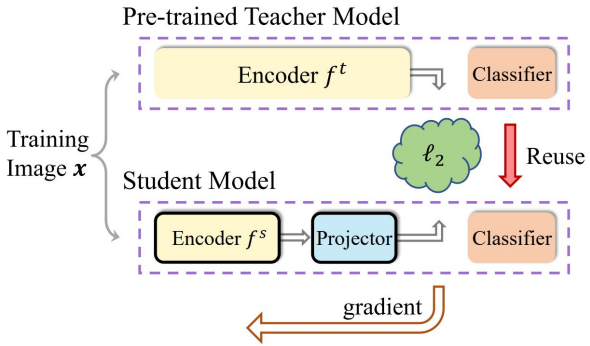
# Experiments: Multiple projector (bias)

Teacher - **ResNet32x4**




2/3 model layers | 1/3 remaining model layers

Student - **MobileNetV2x2** - **30 %** of Teacher parameters



# Experiments: Multiple projector (bias)

**Learning Rates and Training Strategy:**

- Three different learning rates were experimented with—**0.05**, **0.01**, and **0.001**.
- For each learning rate, **three separate** runs were conducted. However, complete results (i.e., all three runs) were only available for the **0.01** learning rate, which also yielded the best performance.

**Feature Alignment for multiple bias Projectors:**

- The feature alignment is conducted through projectors, which process the intermediate outputs of both models.
- Feature alignment happens **twice during training: once after 2/3 of the model's layers** and **again at the end**, following the strategy outlined in the original reference paper.

# Experiments: Multiple projector (bias)

**Projector Weighting:**

The projectors are weighted differently in terms of their contribution to the overall loss:

- The **first projector** (applied after 2/3 of the models) contributes **1/3** to the total loss.
- The **final projector** (applied at the end of the models) contributes **2/3** to the total loss.

# Experiments - Multiple Projectors (bias) - L2 loss

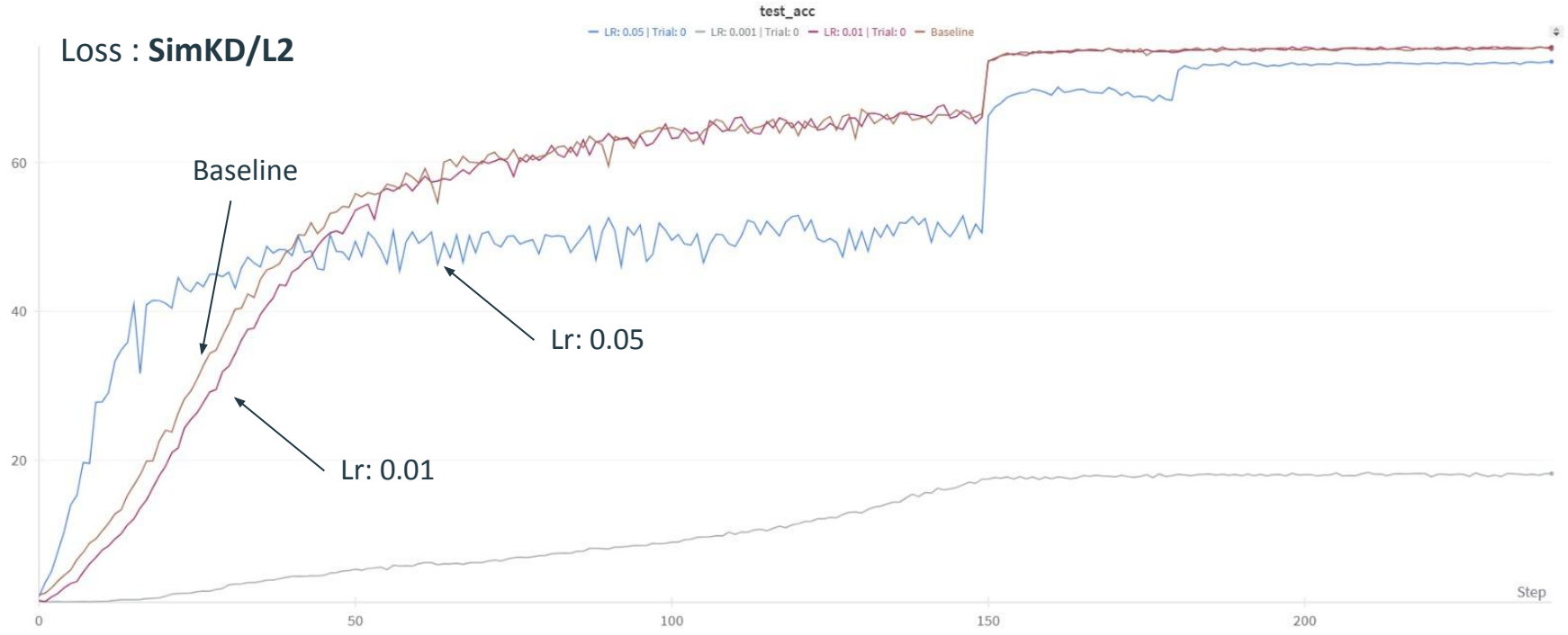
- Loss : **SimKD ,** learning rate: **(0.01, 0.05, 0.001)**
- Epoch: 240, weight decay epoch: **[150, 180, 210]**, Trial - **3 trial for each**
- Teacher - **resnet32x4**, Student: **MobileNetV2_1_0**

Teacher Test Acc: **79.42**

| **Learning rate** | **Student Test Acc** |
|---|---|
| **0.01** | **75.52 ± 0.18** |
| 0.05 | 73.99 ± 0.39 |
| 0.001 | 25.04 ± 0.51 |

**SimKD Acc:** 75.43 ± 0.26

# Experiments - Multiple Projectors (bias) - L2 loss



# Experiments: Multiple Projectors - KL-Div loss

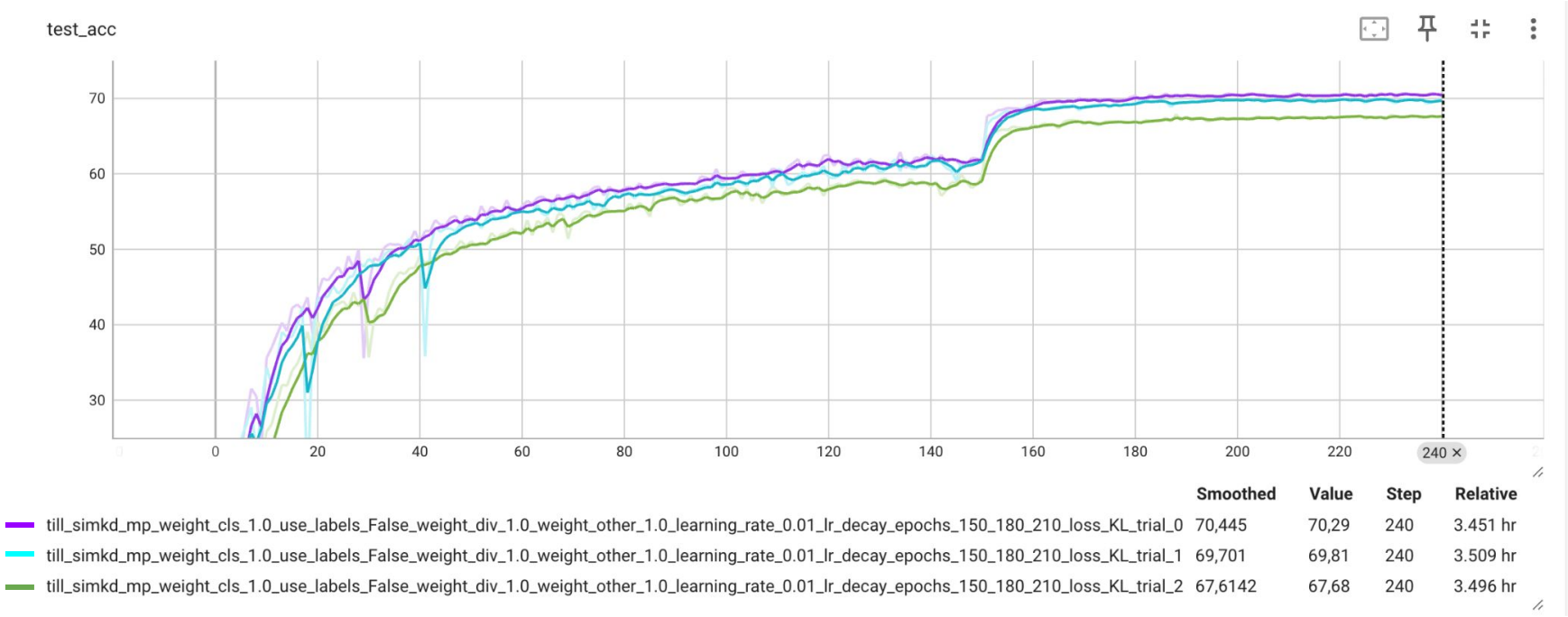
- Loss : **KL-div,** learning rate: **0.01**
- Epoch: 240, weight decay epoch: **[150, 180, 210]**, Trial - **3 trial for each**
- Teacher - **resnet32x4**, Student: **MobileNetV2_1_0**
- After observing the previous results, Lr 0.01 was chosen to conduct further computationally expensive experiments

Teacher Test Acc: **79.42**

| Learning rate | Student Test Acc |
|---|---|
| 0.01 | 66.88 ± 3.78 |

**SimKD Acc:** 75.43 ± 0.26

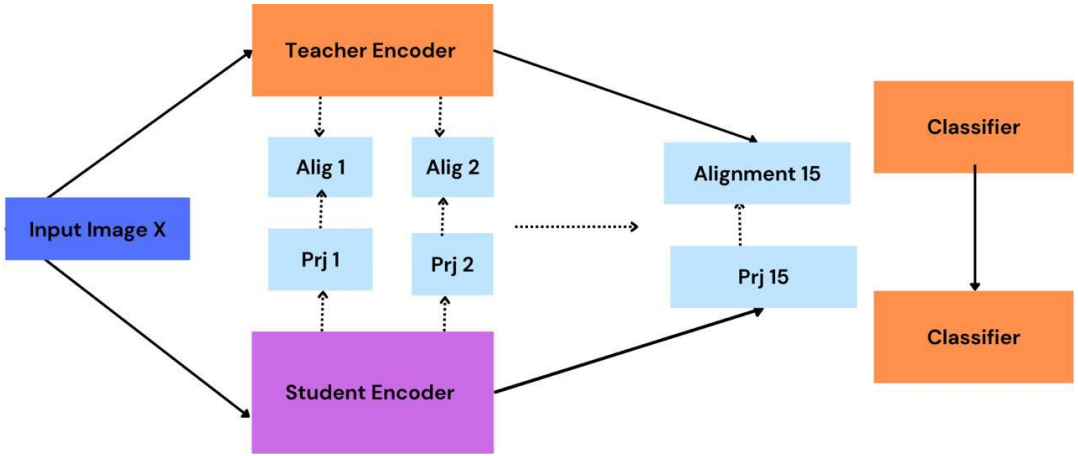
# Experiments - Multiple Projectors (bias) - KL-Div loss




| | Smoothed | Value | Step | Relative |
|---|---|---|---|---|
| till_simkd_mp_weight_cls_1.0_use_labels_False_weight_div_1.0_weight_other_1.0_learning_rate_0.01_lr_decay_epochs_150_180_210_loss_KL_trial_0 | 70,445 | 70,29 | 240 | 3.451 hr |
| till_simkd_mp_weight_cls_1.0_use_labels_False_weight_div_1.0_weight_other_1.0_learning_rate_0.01_lr_decay_epochs_150_180_210_loss_KL_trial_1 | 69,701 | 69,81 | 240 | 3.509 hr |
| till_simkd_mp_weight_cls_1.0_use_labels_False_weight_div_1.0_weight_other_1.0_learning_rate_0.01_lr_decay_epochs_150_180_210_loss_KL_trial_2 | 67,6142 | 67,68 | 240 | 3.496 hr |

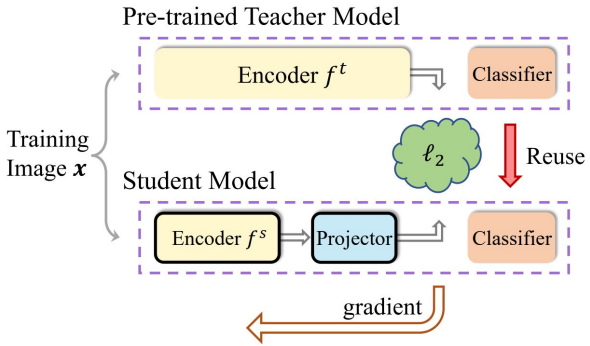
# Experiments: All layer projector (unbiased)

Teacher - **ResNet32x4**




All layers of student models

Student - **MobileNetV2x2**



# Experiments: All-layer projector (unbiased)

**Learning Rates and Training Strategy:**

- Three different learning rates were experimented with—**0.05**, **0.01**, and **0.001**.
- For each learning rate, **three separate** runs were conducted. However, complete results (i.e., all three runs) were only available for the **0.01** learning rate, which also yielded the best performance.

**Feature Alignment and unbiased (all layers) Projectors:**

- The feature alignment is conducted through projectors, which process the intermediate outputs of both models.
- Feature alignment happens at **all of the student model's layers,** following the strategy outlined in the original reference paper.
- Projectors are added after every layer, ensuring continuous and detailed alignment throughout the training process.
- This approach refines feature representations progressively, enhancing knowledge transfer and improving overall model performance.

# Experiments - All layers projector (unbiased) - L2 loss

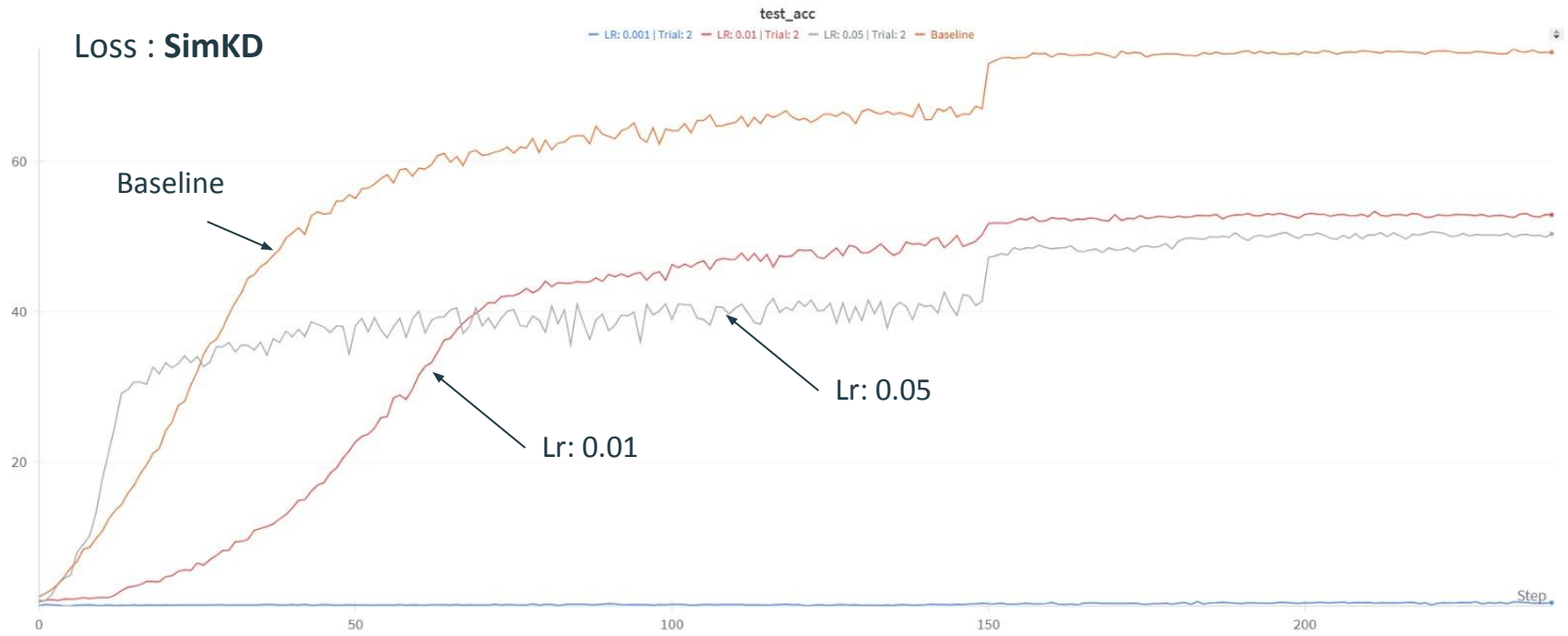
- Loss : **SimKD/L2 ,** learning rate: **(0.01, 0.05, 0.001)**
- Epoch: 240,  weight decay epoch: **[150, 180, 210]**, Trial - **3 trial for each**
- Teacher - **resnet32x4**, Student: **MobileNetV2x2**

Teacher Test Acc: **79.42**

| **Learning rate** | **Student Test Acc** |
|---|---|
| **0.01** | **53.23 ± 0.38** |
| 0.05 | 50.81 ± 0.26 |
| 0.001 | 1.45 ± 0.21 |

**SimKD Acc:** 75.43 ± 0.26

# Experiments: All-layer projector (unbiased) - L2 loss



# Experiments - All layers projector (unbiased)- KL Div

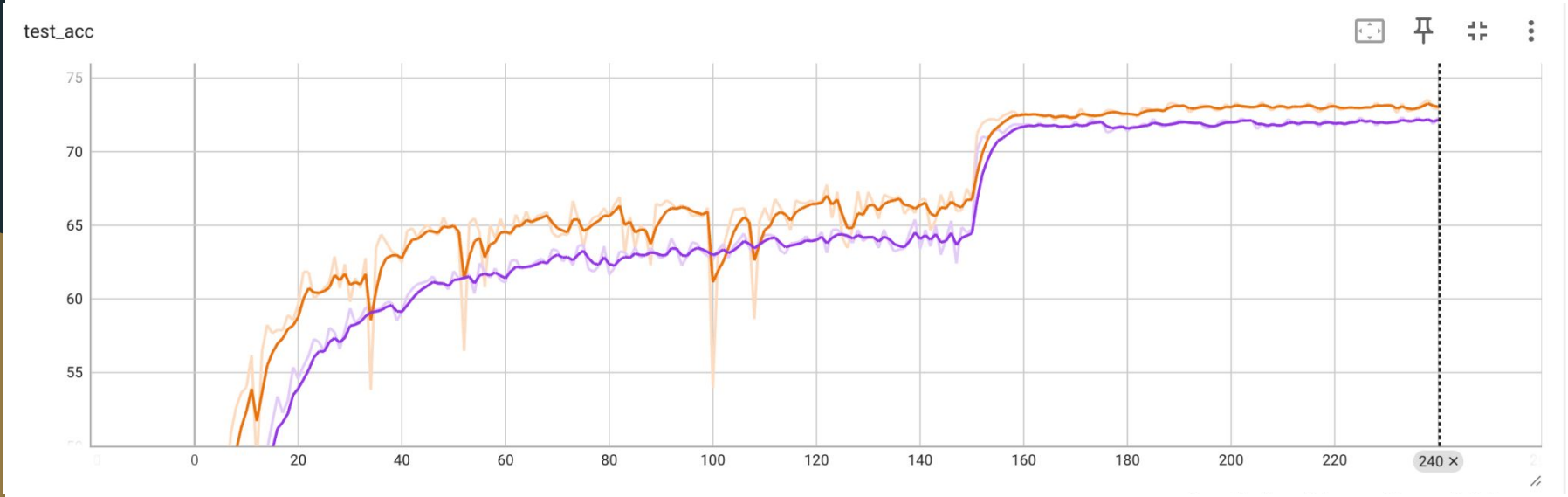
- Loss : **KL-div,** learning rate: **0.01**
- Epoch: 240, weight decay epoch: **[150, 180, 210]**, Trial - **2 trial for each**
- Teacher - **resnet32x4**, Student: **MobileNetV2x2**

Teacher Test Acc: **79.42**

| **Learning rate** | **Student Test Acc** |
|---|---|
| 0.01 | 72.6 ± 0.41 |

**SimKD Acc:** 75.43 ± 0.26

# Experiments: All layers projector (unbiased)- KL Div



# Experiments: Using Labels - L2 loss

- Loss : **SimKD/L2 ,** learning rate: **(0.01, 0.05, 0.001)**
- Epoch: 240, weight decay epoch: **[150, 180, 210]**, Trial - **3 trial for each**
- Teacher - **resnet32x4**, Student: **MobileNetV2x2**
- We only let the student learn, if the teacher made a correct prediction.
- The number of removed images and labels varies from epoch to epoch and is around 10(±3)/50000 images

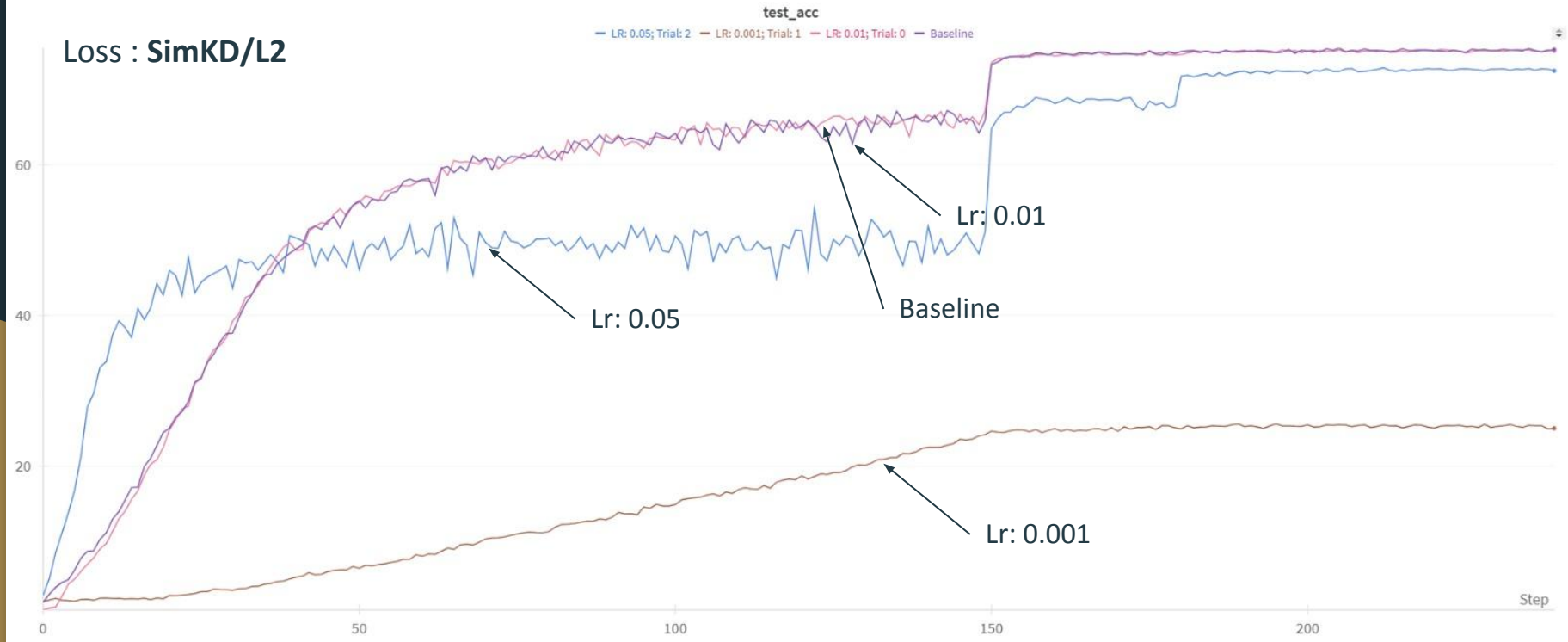
Teacher Test Acc: **79.42**

| Learning rate | Student Test Acc |
|---|---|
| **0.01** | **75.41 ± 0.48** |
| 0.05 | 73.16 ± 0.7 |
| 0.001 | 18.36 ± nan |

**SimKD Acc:** 75.43 ± 0.26

# Experiments: Using labels - L2 loss

Loss : **SimKD/L2**



# Experiments: Using Labels - KL-Div loss

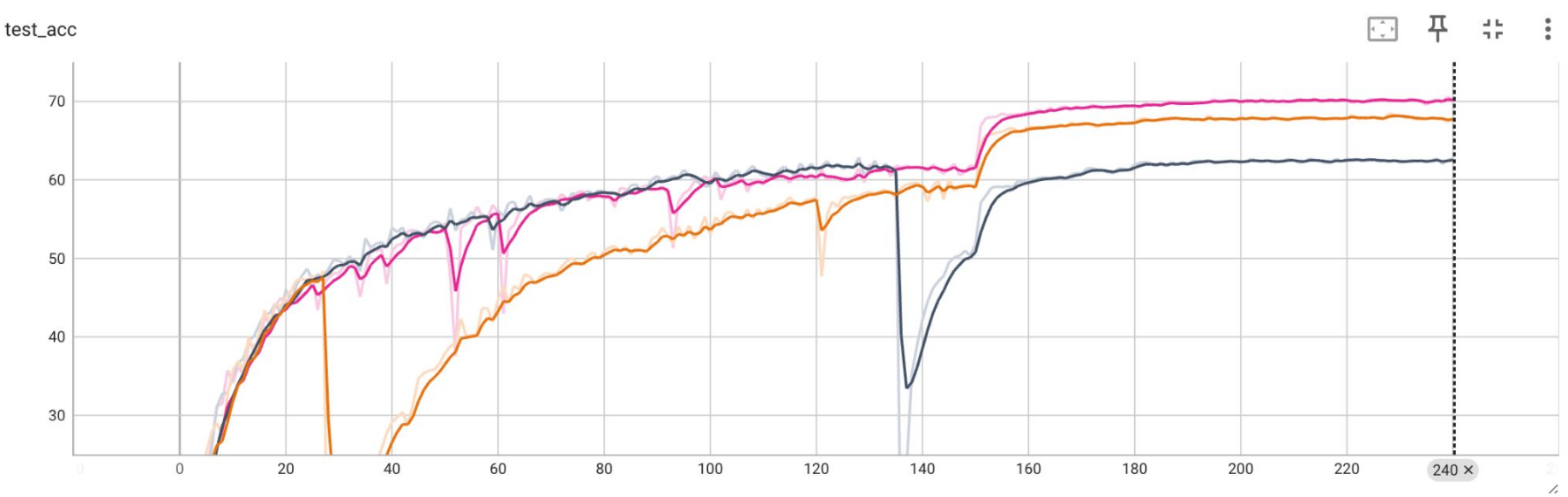
- Loss : **KL-Div,** learning rate: **0.01**
- Epoch: 240,  weight decay epoch: **[150, 180, 210]**, Trial - **3 trial for each**
- Teacher - **resnet32x4**, Student: **MobileNetV2x2**
- We only let the student learn, if the teacher made a correct prediction.
- The number of removed images and labels varies from epoch to epoch and is on average 11(±2)/50000 images

Teacher Test Acc: **79.42**

| Learning rate | Student Test Acc |
|---|---|
| 0.01 | 66.87 ± 3.78 |

**SimKD Acc:** 75.43  ± 0.26

# Experiments: Using labels - KL-Div loss



# Discussion - Results

| Teacher Arc | Student Arc | Projector type | Distillation Loss | Learning rate (best) | Test acc (best) | Improvement from Baseline |
|---|---|---|---|---|---|---|
| **resnet32x4** | **MobileNetV2x2** | **Multiple projector (bias)** | **L2 loss** | **0.01** | **75.52 ± 0.18** | **75.43 (yes)** |
| resnet32x4 | MobileNetV2x2 | Multiple projector (bias) | KL divergence | 0.01 | 66.88 ± 3.78 | 75.43 (no) |
| resnet32x4 | MobileNetV2x2 | All layers projector (unbiased) | L2 loss | 0.01 | 53.23 ± 0.38 | 75.43 (no) |
| resnet32x4 | MobileNetV2x2 | All layers projector (unbiased) | KL divergence | 0.01 | 72.6 ± 0.41 | 75.43 (no) |
| **resnet32x4** | **MobileNetV2x2** | **Using ground truth labels** | **L2 loss** | **0.01** | **75.41 ± 0.48** | **75.43 (on-par)** |
| resnet32x4 | MobileNetV2x2 | Using ground truth labels | KL divergence | 0.01 | 66.87 ± 3.78 | 75.43 (no) |

# Discussion

## Using multiple projectors seems to help,

- Fine-grained knowledge transfer at every layer.
- Better gradient flow for stable learning.
- Captures detailed teacher knowledge.
- Reduces overfitting to noisy features.

## but at what cost?

- More training times and resources
  a. Without additional projector - training computed in less than 2 hours
  b. With additional projector - 2x longer training time
- However, the improvement is not significant.

# Conclusion

- We conducted a series of experiments using various distillation losses and different projector methods to evaluate their effectiveness.
- We experimented with omitting examples were the teacher makes wrong prediction to stabilise the learning process, but the additional computation also requires extra time
- Additionally, incorporating extra projectors or applying projectors to all layers appeared to enhance performance in certain scenarios. However, this approach also introduced a significant increase in training time, which may not be ideal for all applications.
- To provide a comprehensive overview of our findings, we plan to compile a detailed summary report of these experiments, including insights into the trade-offs observed and recommendations for future work.

# Q&A

# Base paper details

- Reusing the teacher classifier largely overlooked in literature, hypothesis transfer learning (HTL) aims to utilize the learned source domain classifier to help the training of the target, only a small amount of labeled target dataset and no source dataset are accessible.
- Vanilla KD losses: (CE loss) + (T^2*KL loss)
- Data: CIFAR-100, ImageNet, standard data augmentation and normalize all images by channel means and standard deviations

# Base paper details

SGD optimizer with 0.9 Nesterov momentum for all datasets. For CIFAR-100, the total training epoch is set to 240 and the learning rate is divided by 10 at 150th, 180th and 210th epochs. The initial learning rate is set to 0.01 for MobileNet/ShuffleNetseries architectures and 0.05 for other architectures. The mini-batch size is set to 64 and the weight decay is set to $5 \times 10{-}4$ . For ImageNet, the initial learning rate is set to

0.1 and then divided by 10 at 30th, 60th, 90th of the total 120 training epochs. The mini-batch size is set to 256 and the weight decay is set to $1 \times 10{-}4$ . All results are reported

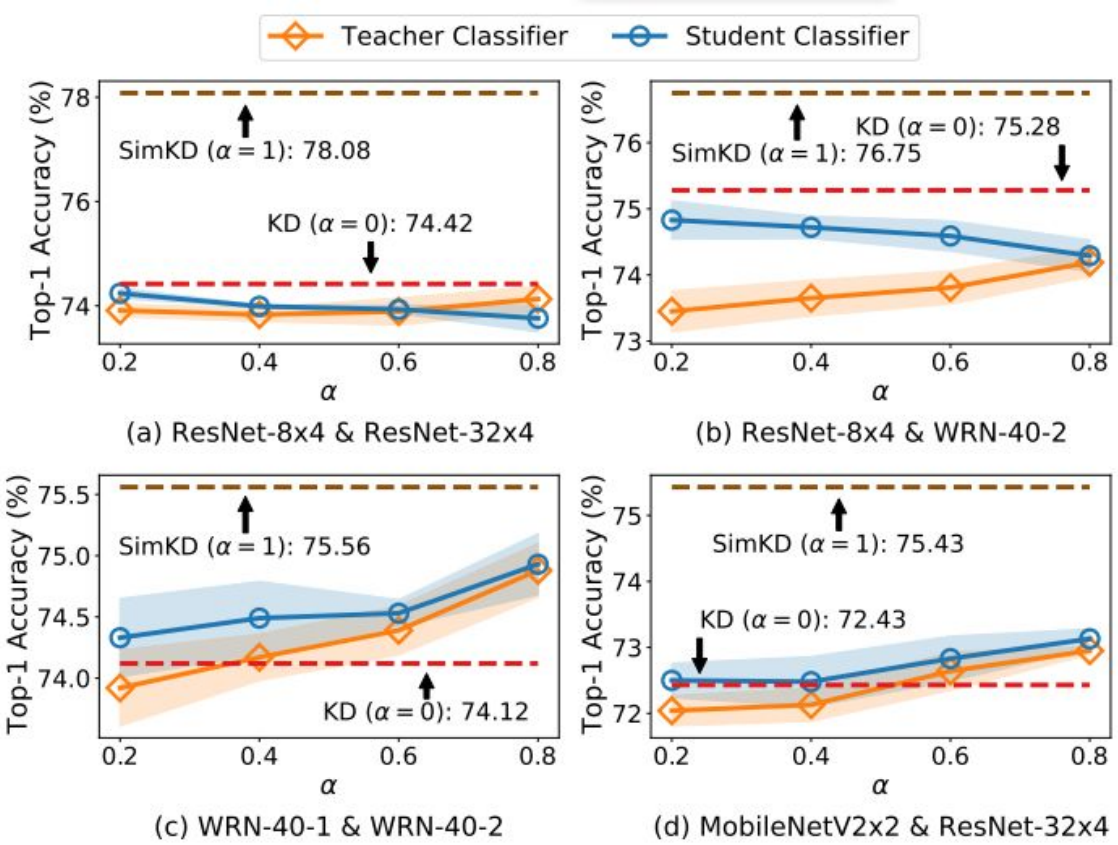
in means (standard deviations) over 4 trials, except for the results on ImageNet are reported in a single trial. The temperature T in the KD loss is set to 4 throughout this paper

| Student | WRN-40-1<br>71.92 ± 0.17 | ResNet-8x4<br>73.09 ± 0.30 | ResNet-110<br>74.37 ± 0.17 | ResNet-116<br>74.46 ± 0.09 | VGG-8<br>70.46 ± 0.29 | ResNet-8x4<br>73.09 ± 0.30 | ShuffleNetV2<br>72.60 ± 0.12 |
|---|---|---|---|---|---|---|---|
| KD [24] | 74.12 ± 0.29 | 74.42 ± 0.05 | 76.25 ± 0.34 | 76.14 ± 0.32 | 72.73 ± 0.15 | 75.28 ± 0.18 | 75.60 ± 0.21 |
| FitNet [40] | 74.17 ± 0.22 | 74.32 ± 0.08 | 76.08 ± 0.13 | 76.20 ± 0.17 | 72.91 ± 0.18 | 75.02 ± 0.31 | 75.82 ± 0.22 |
| AT [55] | 74.67 ± 0.18 | 75.07 ± 0.03 | 76.67 ± 0.28 | 76.84 ± 0.25 | 71.90 ± 0.13 | 75.74 ± 0.09 | 75.41 ± 0.10 |
| SP [48] | 73.90 ± 0.17 | 74.29 ± 0.07 | 76.43 ± 0.39 | 75.99 ± 0.26 | 73.12 ± 0.10 | 74.84 ± 0.08 | 75.77 ± 0.08 |
| VID [1] | 74.59 ± 0.17 | 74.55 ± 0.10 | 76.17 ± 0.22 | 76.53 ± 0.24 | 73.19 ± 0.23 | 75.56 ± 0.13 | 75.22 ± 0.07 |
| CRD [46] | 74.80 ± 0.33 | 75.59 ± 0.07 | 76.86 ± 0.09 | 76.83 ± 0.13 | 73.54 ± 0.19 | 75.78 ± 0.27 | 77.04 ± 0.61 |
| SRRL [52] | 74.64 ± 0.14 | 75.39 ± 0.34 | 76.75 ± 0.14 | 77.19 ± 0.09 | 73.23 ± 0.16 | 76.12 ± 0.18 | 76.19 ± 0.35 |
| SemCKD [6] | 74.41 ± 0.16 | 76.23 ± 0.04 | 76.62 ± 0.14 | 76.69 ± 0.48 | 75.27 ± 0.13 | 75.85 ± 0.16 | 77.62 ± 0.32 |
| SimKD | **75.56 ± 0.27** | **78.08 ± 0.15** | **77.82 ± 0.15** | **77.90 ± 0.11** | **75.76 ± 0.12** | **76.75 ± 0.23** | **78.39 ± 0.27** |
| Teacher | WRN-40-2<br>76.31 | ResNet-32x4<br>79.42 | ResNet-110x2<br>78.18 | ResNet-110x2<br>78.18 | ResNet-32x4<br>79.42 | WRN-40-2<br>76.31 | ResNet-32x4<br>79.42 |

Table 1. Top-1 test accuracy (%) of various knowledge distillation approaches on CIFAR-100.

| Student | ShuffleNetV1<br>71.36 ± 0.25 | WRN-16-2<br>73.51 ± 0.32 | ShuffleNetV2<br>72.60 ± 0.12 | MobileNetV2<br>65.43 ± 0.29 | MobileNetV2x2<br>69.06 ± 0.10 | WRN-40-2<br>76.35 ± 0.18 | ShuffleNetV2x1.5<br>74.15 ± 0.22 |
|---|---|---|---|---|---|---|---|
| KD [24] | 74.30 ± 0.16 | 74.90 ± 0.29 | 76.05 ± 0.34 | 69.07 ± 0.47 | 72.43 ± 0.32 | 77.70 ± 0.13 | 76.82 ± 0.23 |
| FitNet [40] | 74.52 ± 0.03 | 74.70 ± 0.35 | 76.02 ± 0.21 | 68.64 ± 0.27 | 73.09 ± 0.46 | 77.69 ± 0.23 | 77.12 ± 0.24 |
| AT [55] | 75.55 ± 0.19 | 75.38 ± 0.18 | 76.84 ± 0.19 | 68.62 ± 0.31 | 73.08 ± 0.14 | 78.45 ± 0.24 | 77.51 ± 0.31 |
| SP [48] | 74.69 ± 0.32 | 75.16 ± 0.32 | 76.60 ± 0.22 | 68.73 ± 0.17 | 72.99 ± 0.27 | 78.34 ± 0.08 | 77.18 ± 0.19 |
| VID [1] | 74.76 ± 0.22 | 74.85 ± 0.35 | 76.44 ± 0.32 | 68.91 ± 0.33 | 72.70 ± 0.22 | 77.96 ± 0.33 | 77.11 ± 0.35 |
| CRD [46] | 75.34 ± 0.24 | 75.65 ± 0.08 | 76.67 ± 0.27 | 70.28 ± 0.24 | 73.67 ± 0.26 | 78.15 ± 0.14 | 77.66 ± 0.22 |
| SRRL [52] | 75.18 ± 0.39 | 75.46 ± 0.13 | 76.71 ± 0.27 | 69.34 ± 0.16 | 73.48 ± 0.36 | 78.39 ± 0.19 | 77.55 ± 0.26 |
| SemCKD [6] | 76.31 ± 0.20 | 75.65 ± 0.23 | 77.67 ± 0.30 | 69.88 ± 0.30 | 73.98 ± 0.32 | 78.74 ± 0.17 | 79.13 ± 0.41 |
| SimKD | **77.18 ± 0.26** | **77.17 ± 0.32** | **78.25 ± 0.24** | **70.71 ± 0.41** | **75.43 ± 0.26** | **79.29 ± 0.11** | **79.54 ± 0.26** |
| Teacher | ResNet-32x4<br>79.42 | ResNet-32x4<br>79.42 | ResNet-110x2<br>78.18 | WRN-40-2<br>76.31 | ResNet-32x4<br>79.42 | ResNet-32x4<br>79.42 | ResNet-32x4<br>79.42 |

Table 2. Top-1 test accuracy (%) of various knowledge distillation approaches on CIFAR-100.



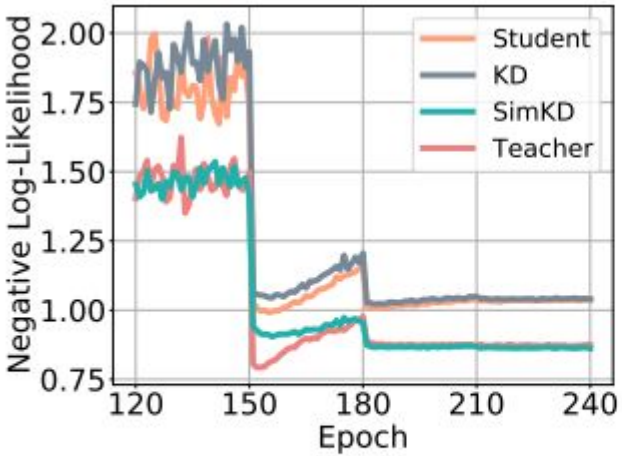

train the student feature encoder with its associated classifier jointly and then report the test accuracies of student models by using their own classifiers or the reused teacher classifiers.

| | Student | KD [24] | AT [55] | SP [48] | VID [1] | CRD [46] | SRRL [52] | SemCKD [6] | SimKD | Teacher |
|---|---|---|---|---|---|---|---|---|---|---|
| 1/4 Epoch | 49.34 | 52.75 | 52.85 | 53.57 | 53.22 | 55.44 | 55.14 | 53.14 | **61.73** | 54.50 |
| 1/2 Epoch | 64.98 | 66.69 | 66.69 | 66.36 | 66.64 | 67.25 | 67.36 | 66.89 | **69.26** | 70.55 |
| Full Epoch | 70.58 | 71.29 | 71.18 | 71.08 | 71.11 | 71.25 | 71.46 | 71.41 | **71.66** | 76.26 |

Table 3. Top-1 test accuracy (%) comparison on ImageNet for different training epochs. We adopt ResNet-18 as the student model.

| Student | Sequential | SimKD | Teacher |
|---|---|---|---|
| WRN-40-1 | $74.48 \pm 0.04$ | $75.56 \pm 0.27$ | WRN-40-2 |
| ResNet-8x4 | $51.97 \pm 0.19$ | $78.08 \pm 0.15$ | ResNet-32x4 |
| ResNet-110 | $77.63 \pm 0.05$ | $77.82 \pm 0.15$ | ResNet-110x2 |
| ResNet-116 | $77.75 \pm 0.03$ | $77.90 \pm 0.11$ | ResNet-110x2 |
| VGG-8 | $35.72 \pm 1.33$ | $75.76 \pm 0.12$ | ResNet-32x4 |
| ResNet-8x4 | $45.03 \pm 0.44$ | $76.75 \pm 0.23$ | WRN-40-2 |
| ShuffleNetV2 | $21.56 \pm 0.31$ | $78.39 \pm 0.27$ | ResNet-32x4 |

Table 4. Training a new student classifier from scratch.




| | Accuracy (%) |
|---|---|
| Student | 73.09 ± 0.30 |
| KD [24] | 74.42 ± 0.05 |
| SimKD | **78.08 ± 0.15** |
| SimKD+ | **78.47 ± 0.08** |
| SimKD++ | **78.88 ± 0.05** |
| Teacher | 79.42 |

Figure 5. Comparison of the top-1 test accuracy (%) and negative log-likelihood (Student: ResNet-8x4, Teacher: ResNet-32x4).

$$\text{Pruning Ratio} = 1 - \frac{\sharp\text{param}_{\text{se}} + \sharp\text{param}_{\text{proj}} + \Delta}{\sharp\text{param}_{\text{t}}}$$

$$\Delta = \sharp\text{param}_{\text{tc}} - \sharp\text{param}_{\text{sc}},$$

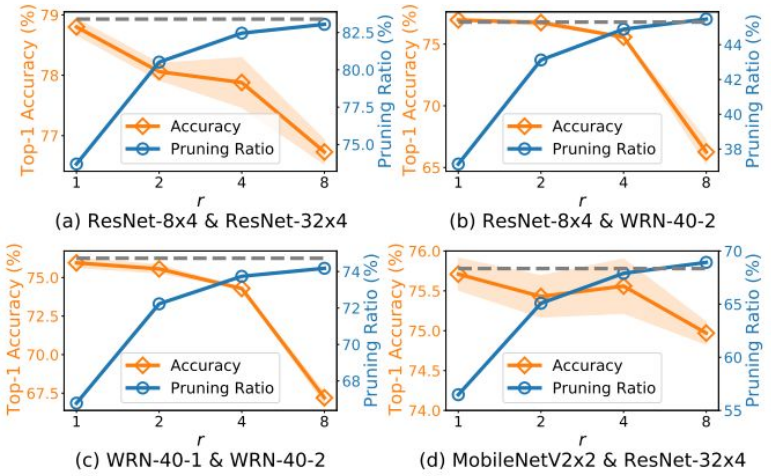



Figure 6. Trade-off between test accuracy and pruning ratio. The pruning ratio of the vanilla KD is drawn with the gray dashed line.

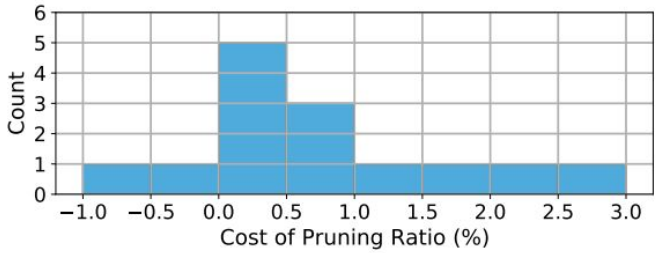



Figure 7. A histogram of the pruning ratio cost.

| Method | ① | ② |
|---|---|---|
| Student | $72.60 \pm 0.12$ | $72.60 \pm 0.12$ |
| AVEG | $75.94 \pm 0.20$ | $76.33 \pm 0.14$ |
| AEKD [14] | $75.99 \pm 0.18$ | $76.17 \pm 0.43$ |
| AEKD-F [14] | $77.24 \pm 0.32$ | $77.08 \pm 0.28$ |
| SimKD$_v$ | $\mathbf{77.43 \pm 0.21}$ | $\mathbf{77.60 \pm 0.23}$ |
| SimKD | $\mathbf{78.59 \pm 0.31}$ | $\mathbf{78.59 \pm 0.05}$ |

Table 6. Results of the multi-teacher KD. We adopt ShuffleNetV2 as the student model and train it under two groups of pre-trained teacher models: ① includes three ResNet-32x4. ② includes two ResNet-32x4 and one ResNet-110x2.

| Projector | Test loss ($\ell_2$) | Accuracy (%) |
|---|---|---|
| 1x1Conv | $0.345 \pm 0.001$ | $75.15 \pm 0.27$ |
| 1x1Conv-1x1Conv | $0.343 \pm 0.001$ | $75.71 \pm 0.33$ |
| 1x1Conv-3x3Conv (DW)-1x1Conv | $0.306 \pm 0.001$ | $77.76 \pm 0.12$ |
| 1x1Conv-3x3Conv-1x1Conv | $\mathbf{0.301 \pm 0.001}$ | $\mathbf{78.08 \pm 0.15}$ |

Table 5. Comparison of projectors. "1x1/3x3Conv" denotes a convolutional layer with 1x1/3x3 kernel size. "DW" denotes depthwise separable convolutions. Standard batch normalization and ReLU activation are used after each layer.

| Method | Require data? | WRN-40-1 | WRN-16-2 |
|---|---|---|---|
| Student | Yes | $71.92 \pm 0.17$ | $73.51 \pm 0.32$ |
| ZSKT [34] | No | $33.60 \pm 3.88$ | $45.03 \pm 1.73$ |
| DAFL [9] | No | $45.32 \pm 1.46$ | $45.94 \pm 1.66$ |
| CMI [16] | No | $64.80 \pm 0.35$ | $65.11 \pm 0.43$ |
| CMI+SimKD | No | $\mathbf{66.78 \pm 0.29}$ | $\mathbf{67.31 \pm 0.89}$ |

Table 7. Results of the data-free KD. We adopt WRN-40-2 as the teacher model with two different student models.

| Input dimension | Operator | Output dimension |
|---|---|---|
| $H \times W \times C_s$ | 1x1 Conv | $H \times W \times C_t/r$ |
| $H \times W \times C_t/r$ | 3x3 Conv | $H \times W \times C_t/r$ |
| $H \times W \times C_t/r$ | 1x1 Conv | $H \times W \times C_t$ |

Table S.1. Projector structure. "1x1/3x3Conv" denotes a convolutional layer with 1x1/3x3 kernel size. Standard batch normalization and ReLU activation are used after each convolutional layer. $r$ is the reduction ratio.

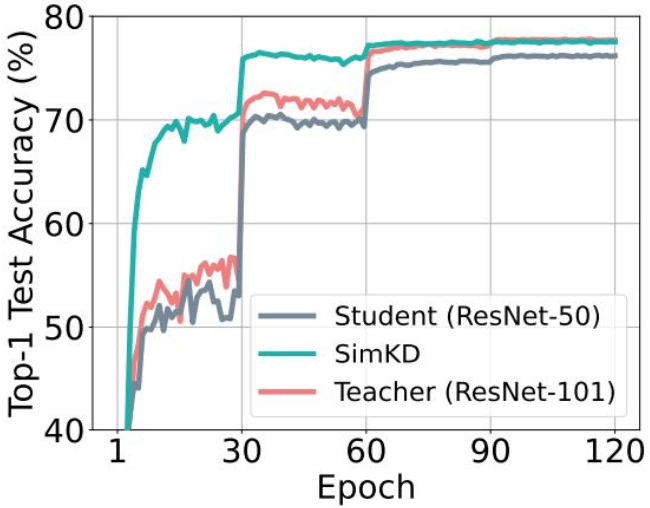



Figure S.1. The test accuracy (%) of ResNet-50 & ResNet-101 ImageNet. Our SimKD achieves faster model convergence.

| | Test Accuracy | Pruning Ratio |
|---|---|---|
| Student | 73.09 ± 0.30 | 83.40% |
| SimKD | **78.08 ± 0.15** | **80.52%** |
| SimKD+ | **78.47 ± 0.08** | 19.21% |
| SimKD++ | **78.88 ± 0.05** | 15.97% |
| Teacher | 79.42 | 0% |

Table S.10. Comparison of reusing different teacher layers.

(a) Vanilla KD [24]. (b) Our SimKD.

Figure S.2. Visualizations of all test images from CIFAR-100 with t-SNE [49]. Features extracted by the teacher and student models are depicted with magenta and cyan colors, respectively, and they are almost indistinguishable in our SimKD. Best viewed in color.

# Additional Experiment result details

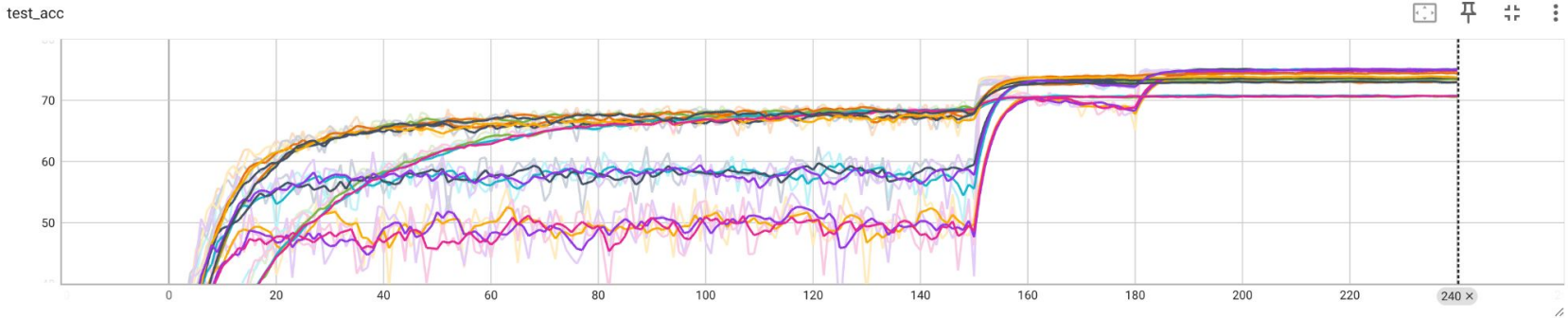
## General Training Parameters

```
{
    "batch_size": 64,
    "epochs": 240,
    "gpu_id": 0,
    "learning_rate": 0.05, #0.001, 0.005,0.01, 0.1
    "lr_decay_epochs": [150,180,210],
    "lr_decay_rate": 0.1,
    "weight_decay": 0.0005,
    "optimizer": "SGD",
    "momentum": 0.9,
    "dataset": "cifar100",
    "model_student": "resnet8x4",
    "path_t": "resnet38x2_best.pth",
    "trial": "0",
    "kd_T": 4,
    "distill": "kd", # cosine similarity
    "cls": 1.0, # scalar for classification
    "div": 1.0, # scalar divergence loss
    "beta": 0.0, # scalar for other loss
    "factor": 2,
    "soft": 1.0,
    "hint_layer": 1,
    "feat_dim": 128,
    "mode": "exact",
    "nce_k": 16384,
    "nce_t": 0.07,
    "nce_m": 0.5,
    "deterministic": false,
    "skip_validation": false,
    "model_teacher": "resnet38x2" # WRN-40-2 }
```

# Individual **KL Divergence** Loss Experiment

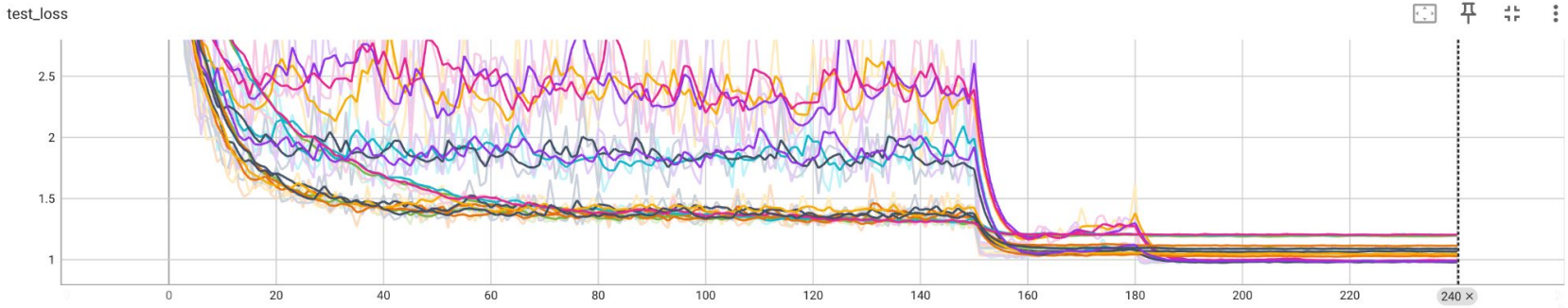
| | | | | Metrics | | | Parameters | |
|---|---|---|---|---|---|---|---|---|
| Run Name | Created | Duration | epoch | test_acc | test_loss | distill | learning_rate |
| S_resnet8x4_T_resnet3... | 3 days ago | 1.4h | 222 | 73.25 | 1.0825066... | simkd_kl | 0.005 |
| S_resnet8x4_T_resnet3... | 3 days ago | 1.4h | 198 | 73.01 | 1.1012697... | simkd_kl | 0.005 |
| S_resnet8x4_T_resnet3... | 3 days ago | 1.4h | 221 | 73.59 | 1.0822051... | simkd_kl | 0.005 |
| S_resnet8x4_T_resnet3... | 3 days ago | 1.4h | 238 | 74.95 | 0.9862127... | simkd_kl | 0.1 |
| S_resnet8x4_T_resnet3... | 3 days ago | 1.4h | 197 | 74.33 | 1.0089675... | simkd_kl | 0.1 |
| S_resnet8x4_T_resnet3... | 3 days ago | 1.4h | 227 | 75.41 | 0.9707022... | simkd_kl | 0.1 |
| S_resnet8x4_T_resnet3... | 3 days ago | 1.5h | 188 | 70.93 | 1.2004726... | simkd_kl | 0.001 |
| S_resnet8x4_T_resnet3... | 3 days ago | 1.4h | 187 | 75.25 | 0.9777719... | simkd_kl | 0.05 |
| S_resnet8x4_T_resnet3... | 3 days ago | 1.4h | 186 | 74.22 | 1.0404529... | simkd_kl | 0.01 |
| S_resnet8x4_T_resnet3... | 3 days ago | 1.5h | 195 | 70.64 | 1.2068628... | simkd_kl | 0.001 |
| S_resnet8x4_T_resnet3... | 3 days ago | 1.4h | 237 | 75.13 | 0.9864139... | simkd_kl | 0.05 |
| S_resnet8x4_T_resnet3... | 3 days ago | 1.4h | 236 | 74.19 | 1.0287499... | simkd_kl | 0.01 |
| S_resnet8x4_T_resnet3... | 3 days ago | 1.5h | 211 | 71.17 | 1.1811063... | simkd_kl | 0.001 |
| S_resnet8x4_T_resnet3... | 3 days ago | 1.4h | 210 | 74.93 | 0.9920148... | simkd_kl | 0.05 |
| S_resnet8x4_T_resnet3... | 4 days ago | 1.4h | 185 | 74.14 | 1.0328612... | simkd_kl | 0.01 |

## **KL Divergence** - Test **Top 1 Acc** Detail




| Run | Smoothed | Value | Step | Relative |
|---|---|---|---|---|
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.001_lr_decay_epochs_150_180_210_trial_0 | 70.6805 | 70.65 | 240 | 1.361 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.001_lr_decay_epochs_150_180_210_trial_1 | 70.5783 | 70.56 | 240 | 1.368 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.001_lr_decay_epochs_150_180_210_trial_2 | 70.7196 | 70.67 | 240 | 1.368 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.005_lr_decay_epochs_150_180_210_trial_0 | 73.4456 | 73.52 | 240 | 1.358 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.005_lr_decay_epochs_150_180_210_trial_1 | 72.9924 | 73.08 | 240 | 1.363 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.005_lr_decay_epochs_150_180_210_trial_2 | 72.9537 | 73.05 | 240 | 1.369 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.01_lr_decay_epochs_150_180_210_trial_0 | 73.6004 | 73.42 | 240 | 1.32 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.01_lr_decay_epochs_150_180_210_trial_1 | 73.7406 | 73.55 | 240 | 1.311 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.01_lr_decay_epochs_150_180_210_trial_2 | 74.4201 | 74.37 | 240 | 1.314 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.05_lr_decay_epochs_150_180_210_trial_0 | 75.0717 | 75.12 | 240 | 1.306 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.05_lr_decay_epochs_150_180_210_trial_1 | 75.0597 | 75.18 | 240 | 1.301 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.05_lr_decay_epochs_150_180_210_trial_2 | 75.0451 | 75.17 | 240 | 1.312 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.1_lr_decay_epochs_150_180_210_trial_0 | 74.8906 | 74.87 | 240 | 1.32 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.1_lr_decay_epochs_150_180_210_trial_1 | 74.8166 | 74.92 | 240 | 1.319 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.1_lr_decay_epochs_150_180_210_trial_2 | 74.8374 | 74.84 | 240 | 1.312 hr |

## **KL Divergence** - Test **Loss** Detail

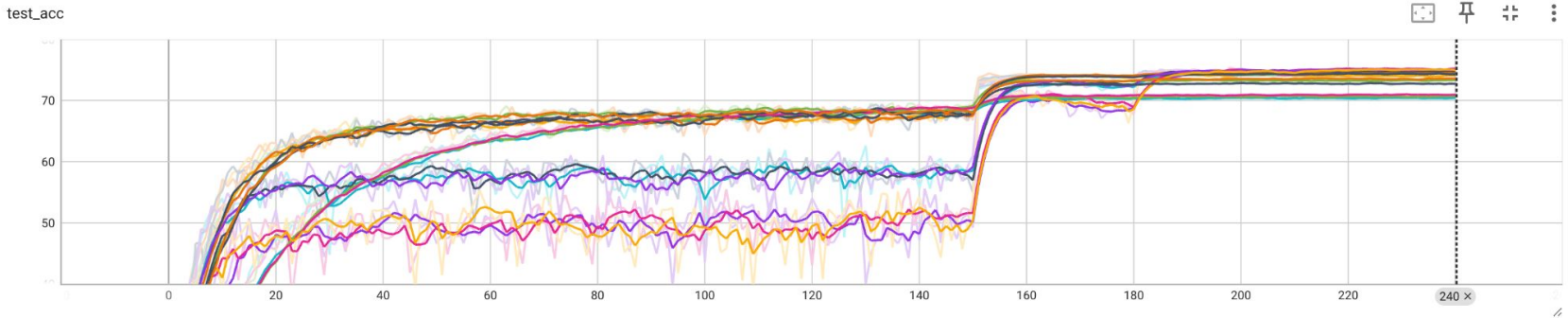



| Run | Smoothed | Value | Step | Relative |
|---|---|---|---|---|
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.001_lr_decay_epochs_150_180_210_trial_0 | 1.2049 | 1.2058 | 240 | 1.361 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.001_lr_decay_epochs_150_180_210_trial_1 | 1.1971 | 1.2004 | 240 | 1.368 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.001_lr_decay_epochs_150_180_210_trial_2 | 1.203 | 1.2068 | 240 | 1.368 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.005_lr_decay_epochs_150_180_210_trial_0 | 1.0867 | 1.0779 | 240 | 1.358 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.005_lr_decay_epochs_150_180_210_trial_1 | 1.1124 | 1.1097 | 240 | 1.363 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.005_lr_decay_epochs_150_180_210_trial_2 | 1.0907 | 1.0963 | 240 | 1.369 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.01_lr_decay_epochs_150_180_210_trial_0 | 1.0666 | 1.0677 | 240 | 1.32 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.01_lr_decay_epochs_150_180_210_trial_1 | 1.0489 | 1.0543 | 240 | 1.311 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.01_lr_decay_epochs_150_180_210_trial_2 | 1.0289 | 1.0253 | 240 | 1.314 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.05_lr_decay_epochs_150_180_210_trial_0 | 0.9902 | 0.9907 | 240 | 1.306 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.05_lr_decay_epochs_150_180_210_trial_1 | 0.9879 | 0.9897 | 240 | 1.301 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.05_lr_decay_epochs_150_180_210_trial_2 | 0.9787 | 0.9803 | 240 | 1.312 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.1_lr_decay_epochs_150_180_210_trial_0 | 0.9926 | 0.9915 | 240 | 1.32 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.1_lr_decay_epochs_150_180_210_trial_1 | 0.983 | 0.9822 | 240 | 1.319 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_kd_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.1_lr_decay_epochs_150_180_210_trial_2 | 0.9847 | 0.986 | 240 | 1.312 hr |

# Individual **Cosine Similarity** Loss Experiment

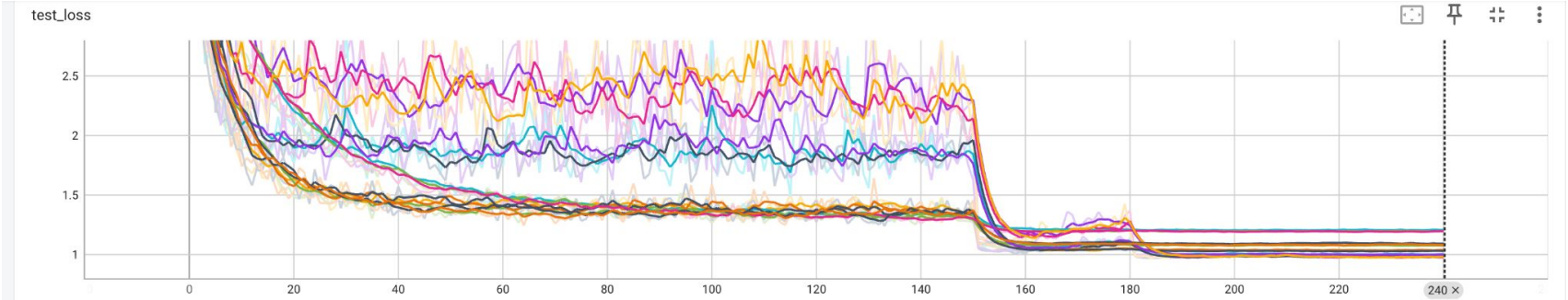
| | | Run Name | Created | Duration | Metrics | | | Parameters | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | epoch | test_acc | test_loss | distill | learning_rate |
| | | Group: similarity 15 | - | | 204.9333... | 73.77333... | 1.055152... | - | 0.0332 |
| | | S_resnet8x4_T_resnet3... | 3 days ago | 1.5h | 223 | 73 | 1.0896979... | similarity | 0.005 |
| | | S_resnet8x4_T_resnet3... | 3 days ago | 1.5h | 235 | 73.87 | 1.0732818... | similarity | 0.005 |
| | | S_resnet8x4_T_resnet3... | 4 days ago | 1.5h | 157 | 73.57 | 1.0652806... | similarity | 0.005 |
| | | S_resnet8x4_T_resnet3... | 4 days ago | 1.5h | 217 | 74.68 | 0.9820224... | similarity | 0.1 |
| | | S_resnet8x4_T_resnet3... | 4 days ago | 1.5h | 238 | 75.14 | 0.9756460... | similarity | 0.1 |
| | | S_resnet8x4_T_resnet3... | 4 days ago | 1.5h | 198 | 75.35 | 0.9888904... | similarity | 0.1 |
| | | S_resnet8x4_T_resnet3... | 4 days ago | 1.5h | 172 | 70.66 | 1.2075932... | similarity | 0.001 |
| | | S_resnet8x4_T_resnet3... | 4 days ago | 1.5h | 185 | 75.03 | 0.9726548... | similarity | 0.05 |
| | | S_resnet8x4_T_resnet3... | 4 days ago | 1.5h | 220 | 74.61 | 1.0382163... | similarity | 0.01 |
| | | S_resnet8x4_T_resnet3... | 4 days ago | 1.5h | 206 | 70.92 | 1.1905091... | similarity | 0.001 |
| | | S_resnet8x4_T_resnet3... | 4 days ago | 1.5h | 200 | 75.2 | 0.9997150... | similarity | 0.05 |
| | | S_resnet8x4_T_resnet3... | 4 days ago | 1.5h | 169 | 74.3 | 1.0278089... | similarity | 0.01 |
| | | S_resnet8x4_T_resnet3... | 5 days ago | 1.5h | 202 | 71.09 | 1.1890707... | similarity | 0.001 |
| | | S_resnet8x4_T_resnet3... | 5 days ago | 1.5h | 232 | 74.61 | 1.0013551... | similarity | 0.05 |
| | | S_resnet8x4_T_resnet3... | 5 days ago | 1.5h | 220 | 74.57 | 1.0255390... | similarity | 0.01 |

## **Cosine Similarity** - Test **Top 1 Acc** Detail



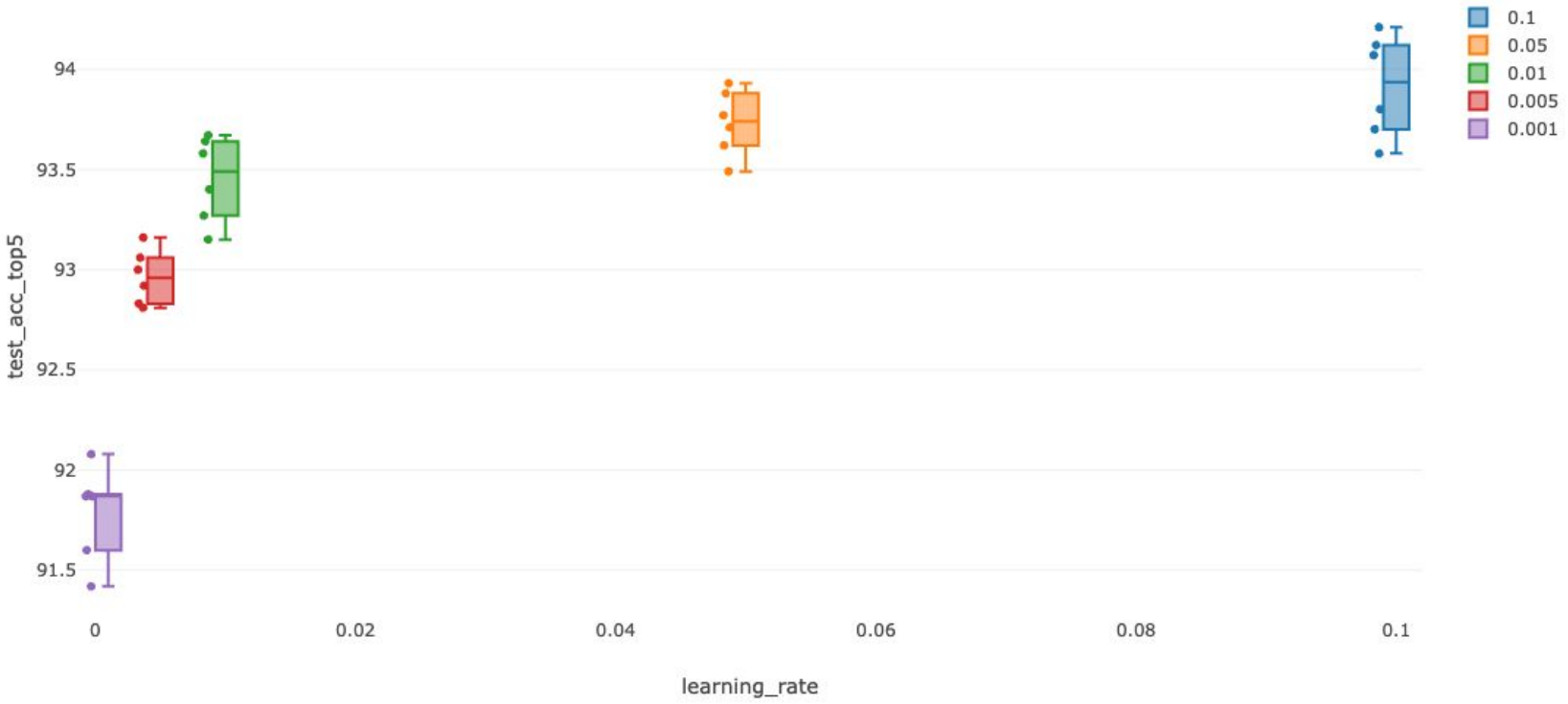

| Run | Smoothed | Value | Step | Relative |
|---|---|---|---|---|
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.001_lr_decay_epochs_150_180_210_trial_0 | 70.9387 | 70.98 | 240 | 1.533 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.001_lr_decay_epochs_150_180_210_trial_1 | 70.6886 | 70.68 | 240 | 1.533 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.001_lr_decay_epochs_150_180_210_trial_2 | 70.4089 | 70.55 | 240 | 1.54 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.005_lr_decay_epochs_150_180_210_trial_0 | 73.3299 | 73.39 | 240 | 1.504 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.005_lr_decay_epochs_150_180_210_trial_1 | 73.5541 | 73.5 | 240 | 1.5 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.005_lr_decay_epochs_150_180_210_trial_2 | 72.7204 | 72.69 | 240 | 1.51 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.01_lr_decay_epochs_150_180_210_trial_0 | 74.2818 | 74.27 | 240 | 1.496 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.01_lr_decay_epochs_150_180_210_trial_1 | 74.0419 | 73.98 | 240 | 1.459 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.01_lr_decay_epochs_150_180_210_trial_2 | 74.3783 | 74.19 | 240 | 1.491 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.05_lr_decay_epochs_150_180_210_trial_0 | 74.4699 | 74.56 | 240 | 1.472 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.05_lr_decay_epochs_150_180_210_trial_1 | 75.0159 | 75.14 | 240 | 1.463 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.05_lr_decay_epochs_150_180_210_trial_2 | 74.7277 | 74.76 | 240 | 1.467 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.1_lr_decay_epochs_150_180_210_trial_0 | 75.1825 | 75.33 | 240 | 1.464 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.1_lr_decay_epochs_150_180_210_trial_1 | 74.9614 | 74.79 | 240 | 1.486 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.1_lr_decay_epochs_150_180_210_trial_2 | 74.544 | 74.67 | 240 | 1.477 hr |

## **Cosine Similarity** - Test **Loss** Detail

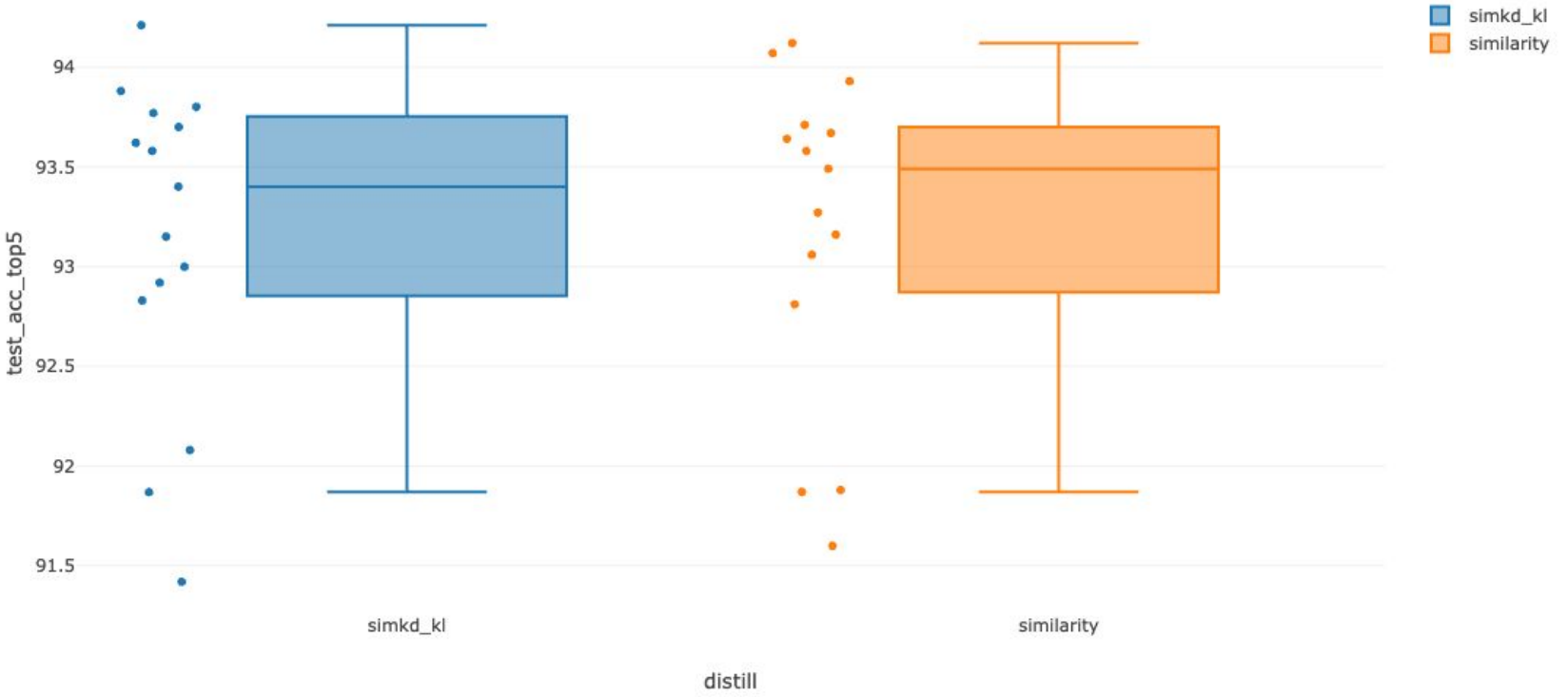



| Run | Smoothed | Value | Step | Relative |
|---|---|---|---|---|
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.001_lr_decay_epochs_150_180_210_trial_0 | 1.1965 | 1.1987 | 240 | 1.533 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.001_lr_decay_epochs_150_180_210_trial_1 | 1.196 | 1.195 | 240 | 1.533 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.001_lr_decay_epochs_150_180_210_trial_2 | 1.2061 | 1.2007 | 240 | 1.54 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.005_lr_decay_epochs_150_180_210_trial_0 | 1.0759 | 1.0736 | 240 | 1.504 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.005_lr_decay_epochs_150_180_210_trial_1 | 1.0818 | 1.084 | 240 | 1.5 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.005_lr_decay_epochs_150_180_210_trial_2 | 1.092 | 1.092 | 240 | 1.51 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.01_lr_decay_epochs_150_180_210_trial_0 | 1.0299 | 1.0275 | 240 | 1.496 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.01_lr_decay_epochs_150_180_210_trial_1 | 1.0344 | 1.035 | 240 | 1.459 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.01_lr_decay_epochs_150_180_210_trial_2 | 1.0378 | 1.0396 | 240 | 1.491 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.05_lr_decay_epochs_150_180_210_trial_0 | 1.0031 | 1.0057 | 240 | 1.472 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.05_lr_decay_epochs_150_180_210_trial_1 | 1.0005 | 1.0006 | 240 | 1.463 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.05_lr_decay_epochs_150_180_210_trial_2 | 0.9787 | 0.9786 | 240 | 1.467 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.1_lr_decay_epochs_150_180_210_trial_0 | 0.977 | 0.975 | 240 | 1.464 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.1_lr_decay_epochs_150_180_210_trial_1 | 0.9816 | 0.9854 | 240 | 1.486 hr |
| S_resnet8x4_T_resnet38x2_dataset_cifar100_distill_similarity_weight_cls_1.0_weight_div_1.0_weight_other_0.0_learning_rate_0.1_lr_decay_epochs_150_180_210_trial_2 | 0.9833 | 0.9809 | 240 | 1.477 hr |

Test **Top 5** Accuracy vs Learning rate
( KL Divergence, Cosine Similarity)



Test **Top 5** Accuracy vs Distillation (KL Divergence, Cosine Similarity)



# END