CS470 Final Presentation

Option 1

KASSL: Knowledge Distillation Applied in Self Supervised Learning

Team 2. KASSL on the Hill

20160709 Seokwoo Hong

Option 1

Applying Knowledge Distillation to Self-supervised Learning in Image Classification

Background Knowledge distillation (KD) is currently actively applied in supervised learning.

Goal ***With KD, train a small model to perform similarly to large model in self-supervised learning (SSL).***

Approach Transfer feature representation by reducing loss between teacher and student models.

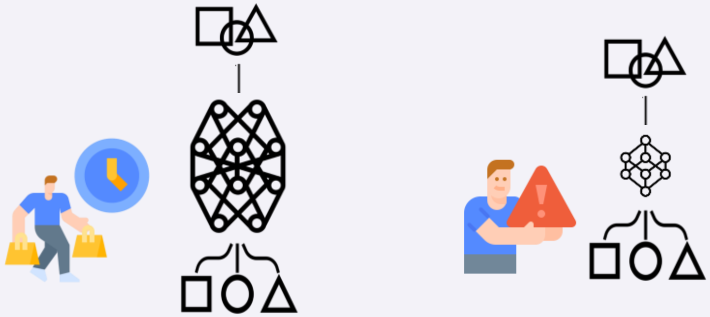
Result Our models (with KD) outperformed baseline in top-1 accuracy:
63.52% ⇒ 77.02% & 81.82%.

result

Meaning

1. **Successfully boosting performance** of small network on image classification task with SSL.
2. **Self-explored problem setting** (KD in SSL) and suggesting working solution.

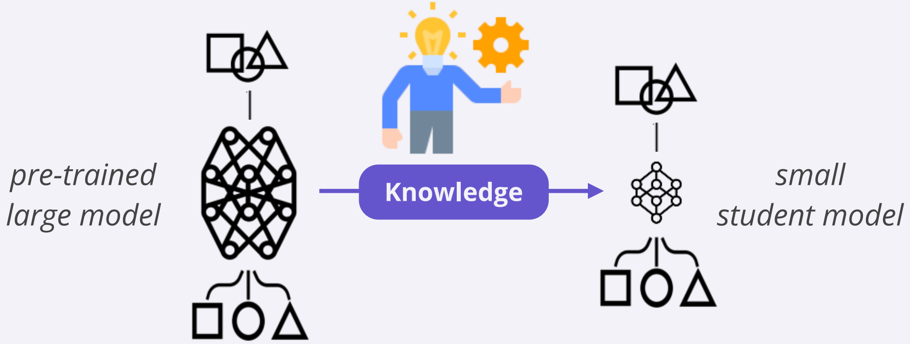
Motivation



Problem

Well-performing models are usually large → Lots of computation resources and time.
Small models are needed to be deployed on end devices but relatively performs worse.

Motivation



Problem

Well-performing models are usually large → Lots of computation resources and time.
Small models are needed to be deployed on end devices but relatively performs worse.

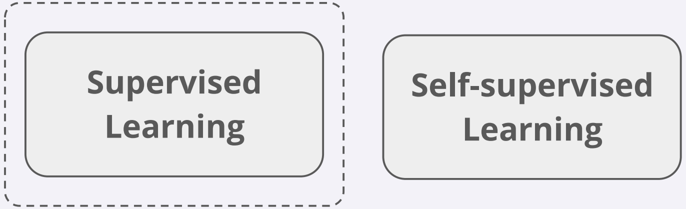
Current Solution

With knowledge distillation (KD), a small model is trained to perform similarly as pre-trained large model.

- A large pre-trained model (teacher) transfers knowledge to a small model (student).

Motivation

Knowledge Distillation Application



Restriction of Solution Area

KD is so far mainly applied in **supervised learning**.

- Soft labels are transferred as the knowledge from a large pre-trained teacher model to a small model.

Motivation

Knowledge Distillation Application

Supervised Learning | **Self-supervised Learning**

Restriction of Solution Area

KD is so far mainly applied in **supervised learning**.

- Soft labels are transferred as the knowledge from a large pre-trained teacher model to a small model.

Beyond the Current Solution Area

Self-supervised learning (SSL) is a rising field in image classification task.

- Small model directly trained with SSL still relatively performs worse without KD.

Objective

With **knowledge distillation**,
train a small model to perform similarly to the large model
which is already pre-trained **in self-supervised learning (SSL)**

Directions of Related Work

DIRECTION 1

Knowledge Distillation

How knowledge distillation is already done ?

DIRECTION 2

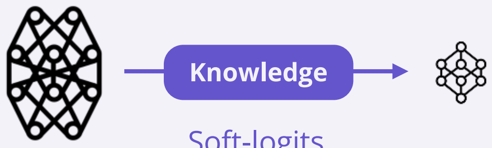
Self-Supervised Learning

How self-supervised learning is done currently?

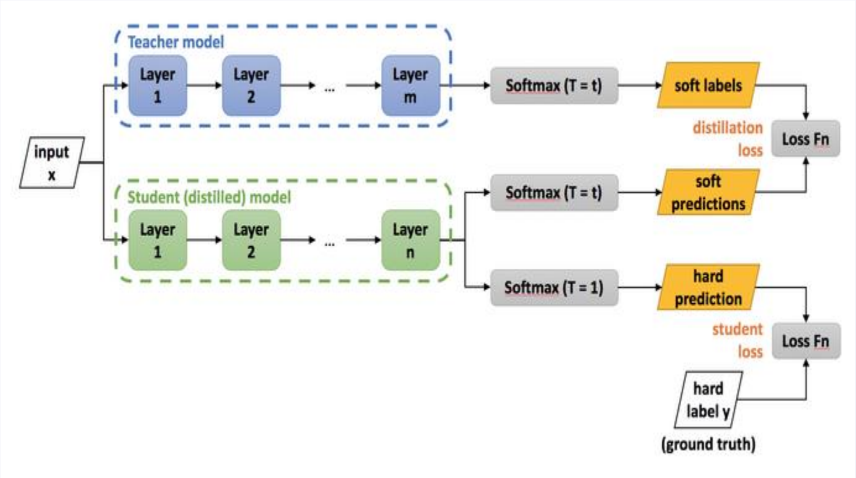
Direction 1. Knowledge Distillation

Knowledge Distillation

The way to distill knowledge depends on kinds of knowledge and how to transfer



Soft-logits
Activation Boundary
Attention Map
Consistency of Feature
...



**Reference: Distilling the knowledge in a neural network*

Direction 2. Self-supervised Learning

What is Self-supervised Learning?

In SSL, the model trains itself to learn one part of the input from another part of the input.

Example of Self-supervised Learning

BYOL: Bootstrap Your Own Latent

- Target: extract target representation / Online: extract target prediction

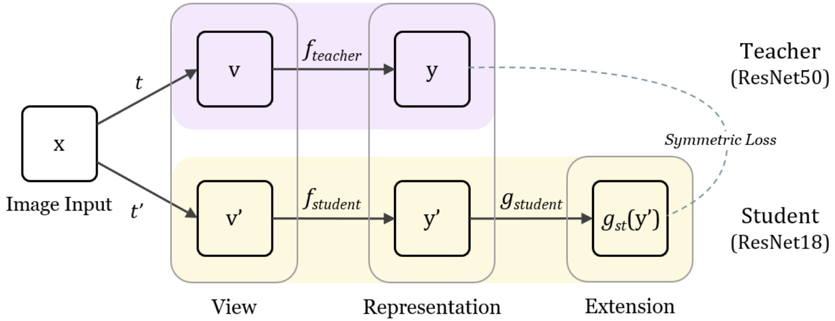
SimCLR: A simple framework for contrastive learning of visual representations

- Data augmentation / contrastive representation learning

MoCo-v2: Improved Baselines with Momentum Contrastive Learning

- Stronger augmentation, MLP projection head

KASSL: *Knowledge distillation Applied to Self-Supervised Learning*

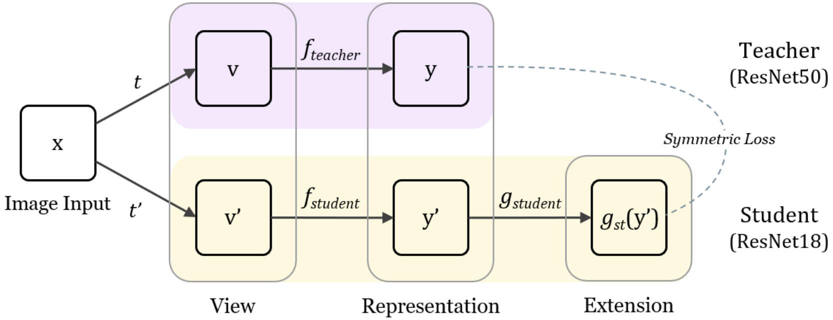


High Level Idea

Transfer knowledge from large model by passing **feature representation**

- Reduce loss between two feature representations made by teacher and student

KASSL: *Knowledge distillation Applied to Self-Supervised Learning*



① $\hat{\theta}_s = argmin_{\theta_s} \sum_i^N \mathcal{L}_{distill}(x_i, \theta_s, \theta_t)$

② $\hat{\theta}_s = -argmin_{\theta_s} \sum_i^N \frac{\sum_j^K y_j g_s(y'_j)}{\sqrt{\sum_j^K y_j^2}\sqrt{\sum_j^K g_s(y'_j)^2}}$

③ $\mathcal{L}_{student,teacher} = \mathcal{L}_{student,teacher} + \widetilde{\mathcal{L}}_{student,teacher}$

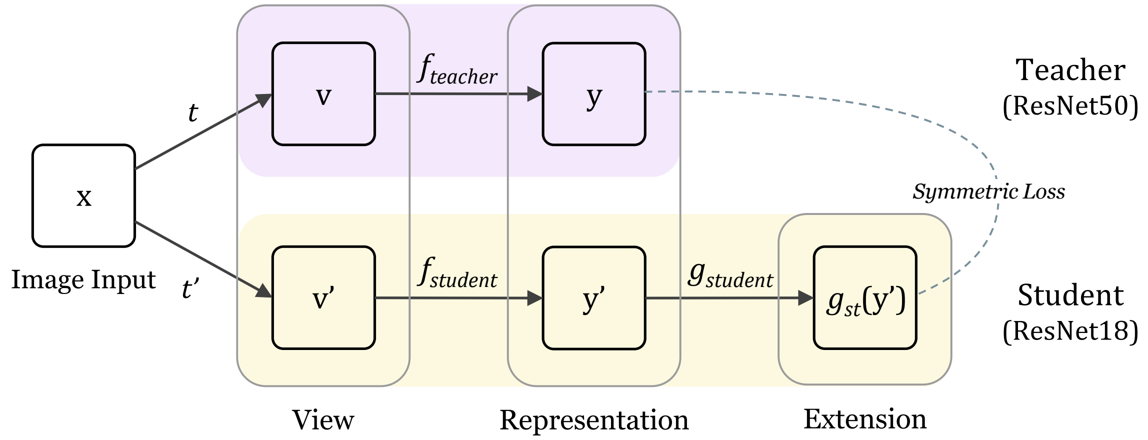
Problem Formulation of KASSL

① Problem formulation : Aim to minimize loss between teacher and student

② Apply negative cosine similarity loss to problem formulation

③ Total loss with Symmetric loss

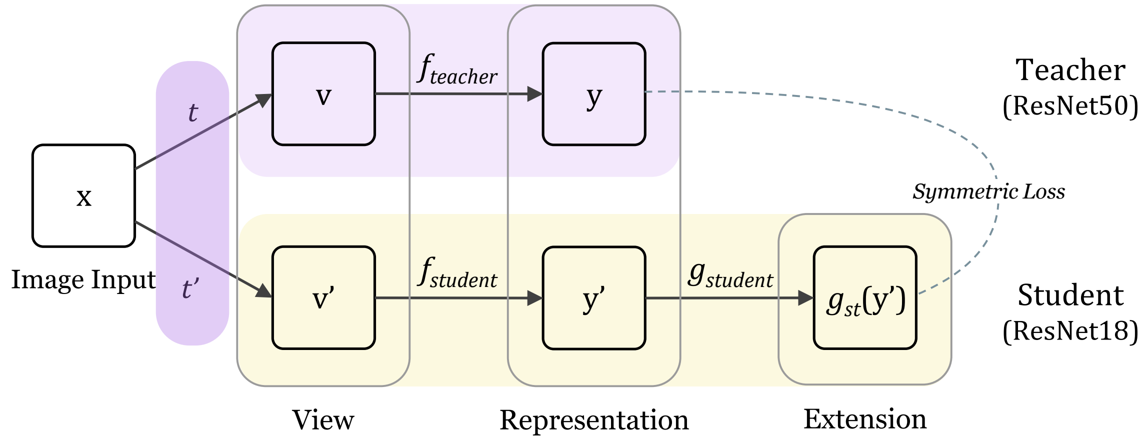
KASSL: *Knowledge distillation Applied to Self-Supervised Learning*



Details

1. Image Augmentation
2. Encoder Extension
3. Use of Symmetric Loss

KASSL: *Knowledge distillation Applied to Self-Supervised Learning*



Details

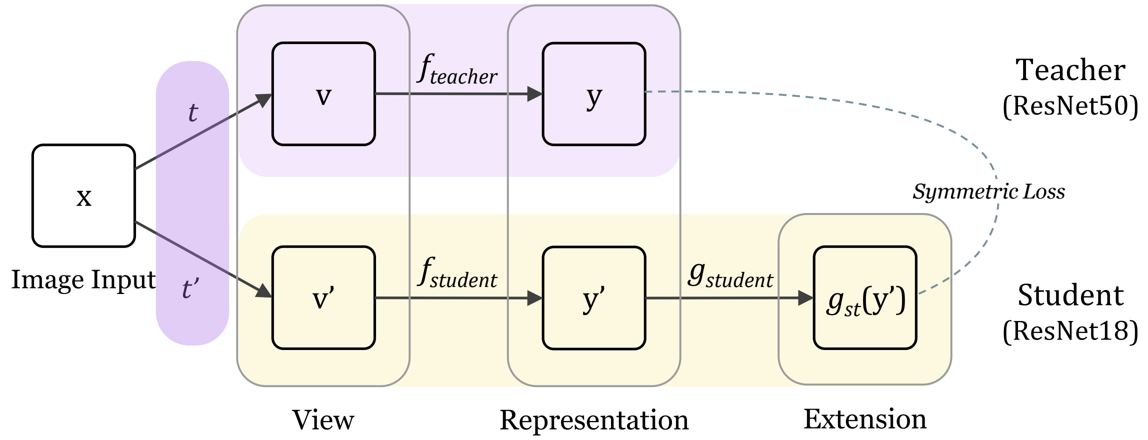
1. **Image Augmentation**
2. Encoder Extension
3. Use of Symmetric Loss

Image Augmentation

Two random image augmentations done similar to SimCLR

- Inspired from the contrastive learning where different image augmentations bring outperforming results.

KASSL: *Knowledge distillation Applied to Self-Supervised Learning*



Details

1. **Image Augmentation**
2. Encoder Extension
3. Use of Symmetric Loss

Augmentation t and t'

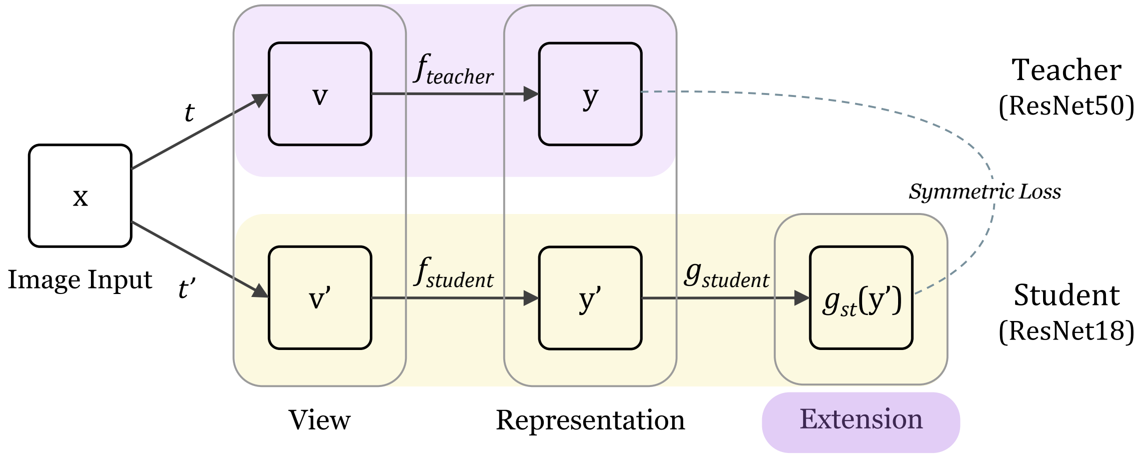
1. Patch selection & resizing
2. flip, color distortion, hue, etc.
3. Gaussian blur, solarization

Image Augmentation

Two random image augmentations done similar to SimCLR

- Inspired from the contrastive learning where different image augmentations bring outperforming results.

KASSL: *Knowledge distillation Applied to Self-Supervised Learning*



Details

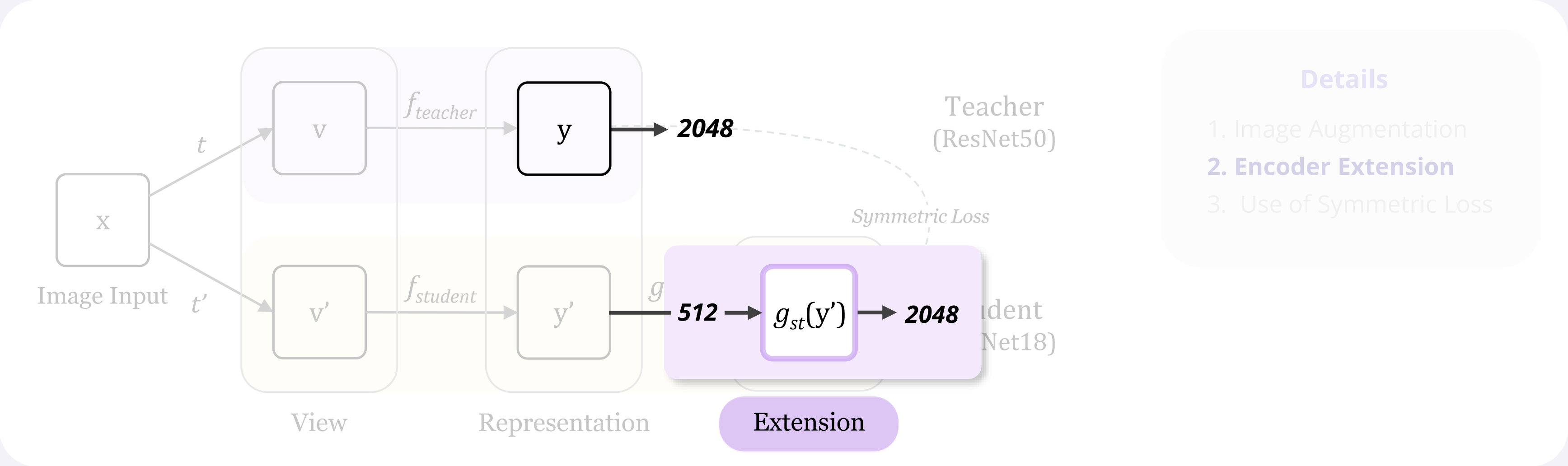
1. Image Augmentation
2. **Encoder Extension**
3. Use of Symmetric Loss

Encoder Extension

Extend student feature vector to fit with the teacher's dimension

- Assuming teacher's representation as the answer, add 4 layers to fit to teacher's dimension.

KASSL: *Knowledge distillation Applied to Self-Supervised Learning*

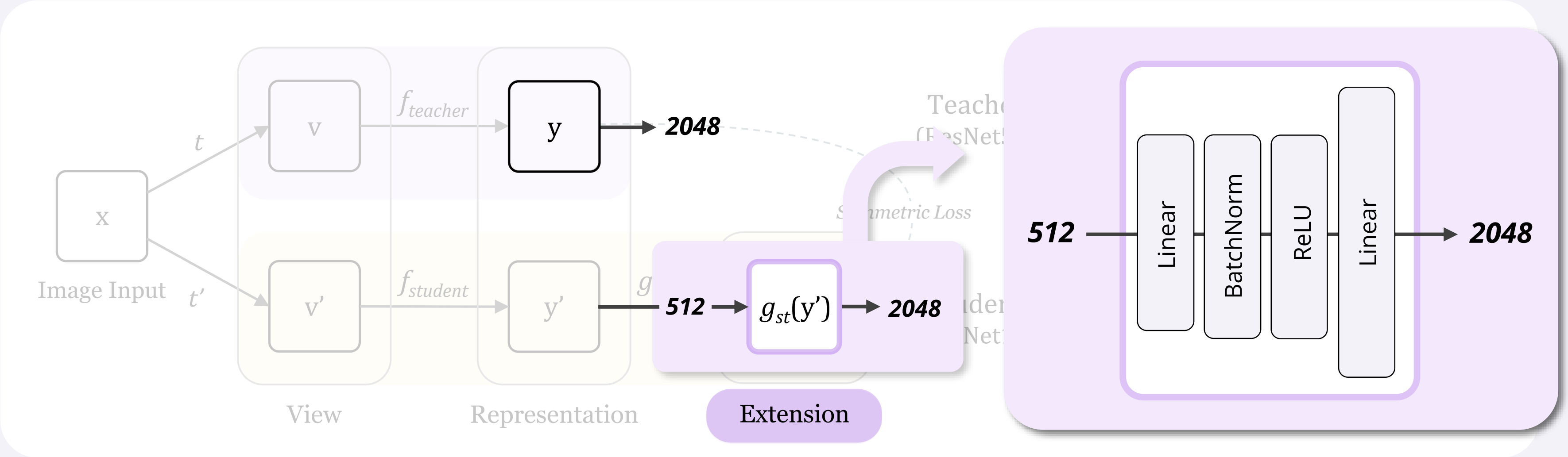


Encoder Extension

Extend student feature vector to fit with the teacher's dimension

- Assuming teacher's representation as the answer, add 4 layers to fit to teacher's dimension.

KASSL: *Knowledge distillation Applied to Self-Supervised Learning*

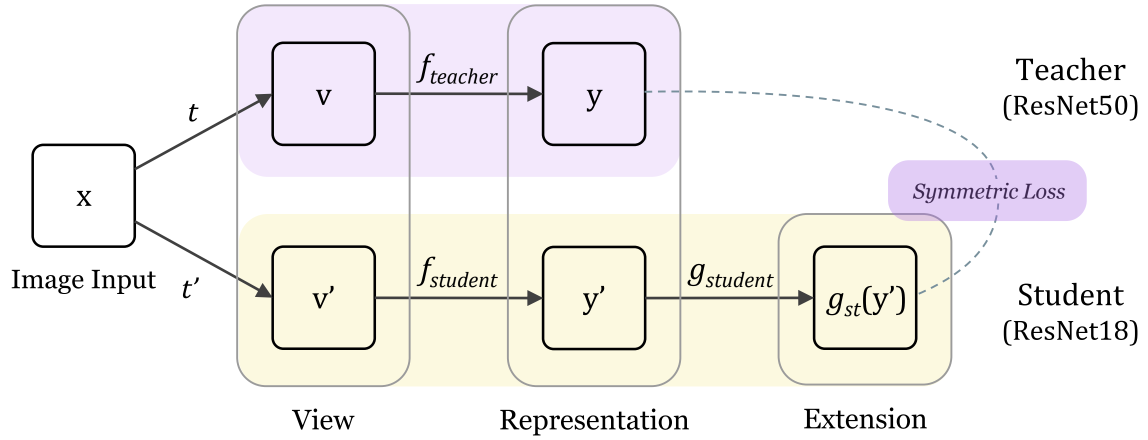


Encoder Extension

Extend student feature vector to fit with the teacher's dimension

- Assuming teacher's representation as the answer, add 4 layers to fit to teacher's dimension.

KASSL: *Knowledge distillation Applied to Self-Supervised Learning*



Details

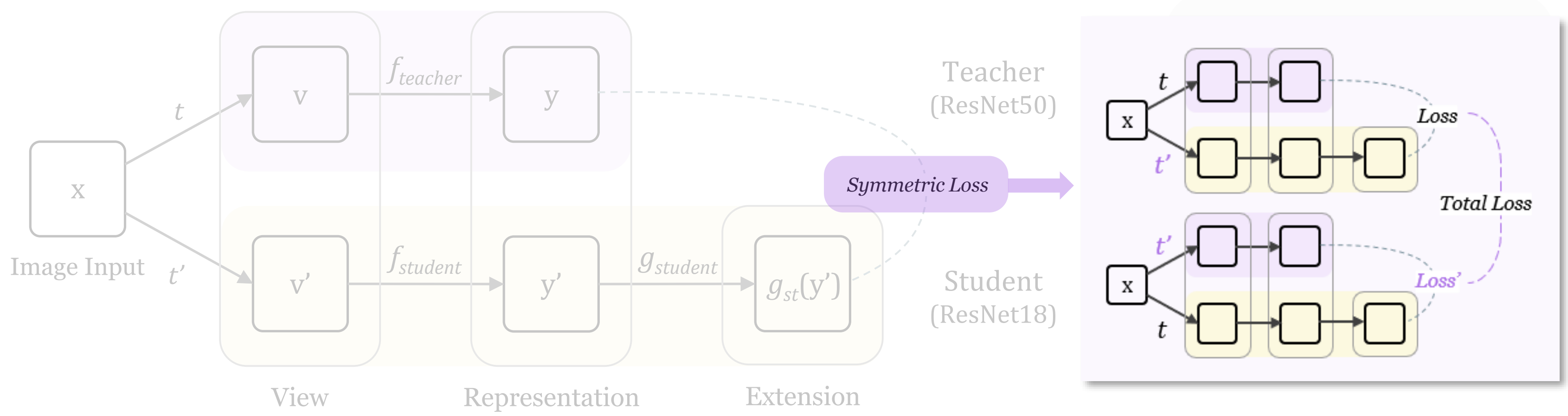
1. Image Augmentation
2. Encoder Extension
3. **Use of Symmetric Loss**

Use of Symmetric Loss

Symmetrize the loss by: **Total Loss = Loss + Loss'**

- Loss' is from reverse augmented image which is feeding t' to the teacher and t to the student.

KASSL: *Knowledge distillation Applied to Self-Supervised Learning*

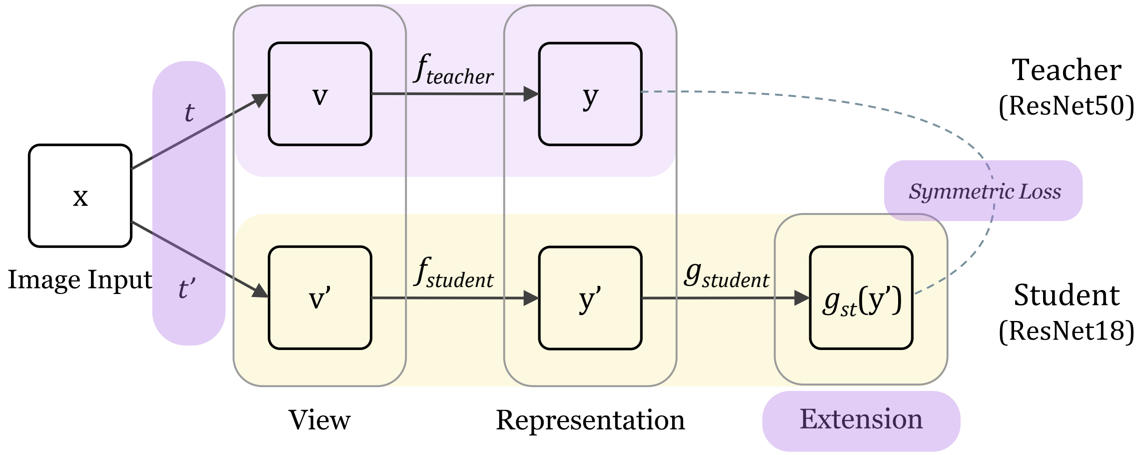


Use of Symmetric Loss

Symmetrize the loss by: **Total Loss = Loss + Loss'**

- Loss' is from reverse augmented image which is feeding t' to the teacher and t to the student.

KASSL: *Knowledge distillation Applied to Self-Supervised Learning*



Details

1. Image Augmentation
2. Encoder Extension
3. Use of Symmetric Loss

With KASSL architecture, we could train a small model to perform similarly to the large model in SSL.

Experiment Plan

Teacher
(ResNet50)

- Pretrained large network trained with existing SSL
- **Learning target** to mimic its performance

Baseline
(ResNet18)

- Baseline small network trained with existing SSL
- **Comparison target** to evaluate our model

Student
(ResNet18)

- Small network trained with **KASSL**
- **Our knowledge distilled result**

Original — KASSL (Original)

Exp 1
Exp 2
Exp 3

Ablation Study (KASSL + Exp 1, 2, 3)

Experiment Plan: Teacher

Teacher (ResNet50)

- Pretrained large network trained with existing SSL
- **Learning target** to mimic its performance

Baseline (ResNet18)

- Baseline small network trained with existing SSL
- **Comparison target** to evaluate our model

Student (ResNet18)

- Small network trained with **KASSL**
- **Our knowledge distilled result**

Original
Exp 1
Exp 2
Exp 3

Pretraining Teacher Network

- **Architecture**: ResNet50
- **Method**: BYOL (default)
  MocoV2 (additional)
- **Dataset**: ImageNet-1k
- Pretrained with 200 epochs
  (we did not train these networks)

Experiment Plan: Baseline

Teacher
(ResNet50)

- Pretrained large network trained with existing SSL
- **Learning target** to mimic its performance

Baseline
(ResNet18)

- Baseline small network trained with existing SSL
- **Comparison target** to evaluate our model

Student
(ResNet18)

- Small network trained with **KASSL**
- **Our knowledge distilled result**

Original
Exp 1
Exp 2
Exp 3

Training Baseline Network

- **Architecture**: ResNet18
- **Method**: BYOL
- **Dataset**: ImageNet100

- 200 epochs
- Hyperparameters based on BYOL

Experiment Plan: Baseline

Teacher
(ResNet50)

- Pretrained large network trained with existing SSL
- **Learning target** to mimic its performance

Baseline
(ResNet18)

- Baseline small network trained with existing SSL
- **Comparison target** to evaluate our model

Student
(ResNet18)

- Small network trained with **KASSL**
- **Our knowledge distilled result**

Original
Exp 1
Exp 2
Exp 3

Training Baseline Network

- **Architecture**: ResNet18
- **Method**: BYOL
- **Dataset**: ImageNet100

- **Subset of ImageNet-1k**
- 10 classes
- 1300 images for training
- 30 images for validation

**Due to the limitation of GPU resource, we chose ImageNet100 instead of 1k*

**Reference: ImageNet-100 from Olga Russakovsky & Fei-Fei, 2008*

Experiment Plan: Student

Teacher
(ResNet50)

- Pretrained large network trained with existing SSL
- **Learning target** to mimic its performance

Baseline
(ResNet18)

- Baseline small network trained with existing SSL
- **Comparison target** to evaluate our model

Student
(ResNet18)

- Small network trained with **KASSL**
- **Our knowledge distilled result**

Original
Exp 1
Exp 2
Exp 3

Distilling Student Network (Original)

- **Architecture**: ResNet18
- **Method**: KASSL (Ours)
- **Dataset**: ImageNet100

- 200 epochs (5 warm up)
- Hyperparameters
  - SGD optimizer with momentum 0.9
  - Learning rate: 0.03
  - Batch size: 64

Experiment Plan: Student

Teacher
(ResNet50)

- Pretrained large network trained with existing SSL
- **Learning target** to mimic its performance

Baseline
(ResNet18)

- Baseline small network trained with existing SSL
- **Comparison target** to evaluate our model

Student
(ResNet18)

- Small network trained with **KASSL**
- **Our knowledge distilled result**

Original
Exp 1
Exp 2
Exp 3

Distilling Student Network (Experiment 1, 2, 3)

- Ablation study done to experiment variations on original KASSL

**Experiments will be explained later.*

- **Architecture**: ResNet18
- **Method**: KASSL + experiments 1, 2, 3
- **Dataset**: ImageNet100

Evaluation Plan

Teacher
(ResNet50)

- Pretrained large network trained with existing SSL
- **Learning target** to mimic its performance

Baseline
(ResNet18)

- Baseline small network trained with existing SSL
- **Comparison target** to evaluate our model

Student
(ResNet18)

- Small network trained with **KASSL**
- **Our knowledge distilled result**

Original
Exp 1
Exp 2
Exp 3

Evaluation Plan

Teacher
(ResNet50)

- Pretrained large network trained with existing SSL
- **Learning target** to mimic its performance

Baseline
(ResNet18)

- Baseline small network trained with existing SSL
- **Comparison target** to evaluate our model

Student
(ResNet18)

- Small network trained with **KASSL**
- **Our knowledge distilled result**

Original
Exp 1
Exp 2
Exp 3

Image Classification

acc1
acc5

acc1
acc5

acc1
acc5

Additional Training to Downstream Task

Evaluation Standard

(1)
Does KASSL mimic teacher target well?

Evaluation Plan

Teacher
(ResNet50)

- Pretrained large network trained with existing SSL
- **Learning target** to mimic its performance

Baseline
(ResNet18)

- Baseline small network trained with existing SSL
- **Comparison target** to evaluate our model

Student
(ResNet18)

- Small network trained with **KASSL**
- **Our knowledge distilled result**

Original
Exp 1
Exp 2
Exp 3

Image Classification

acc1
acc5

acc1
acc5

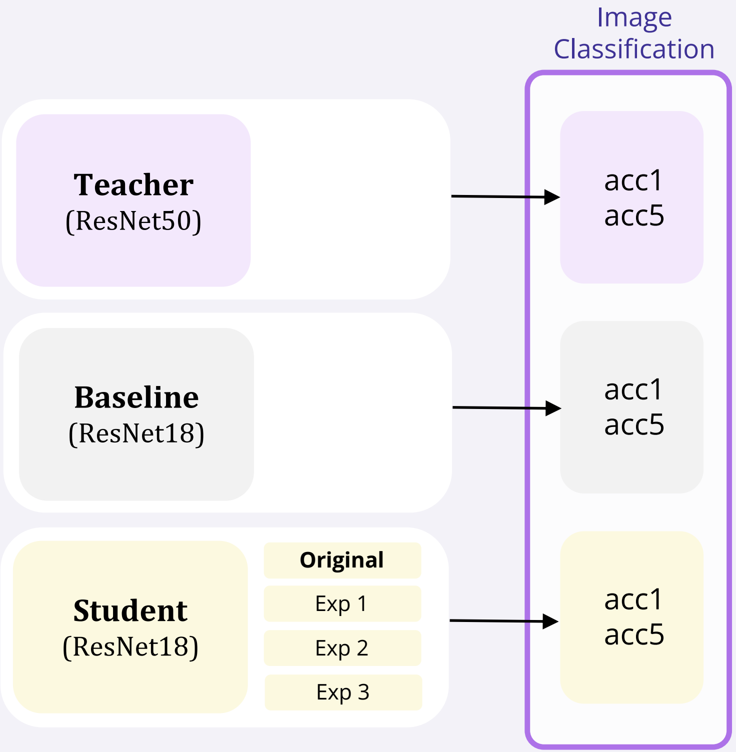
acc1
acc5

Additional Training to Downstream Task

Evaluation Standard

(2)
How does KASSL perform compared to existing method?

Evaluation Result



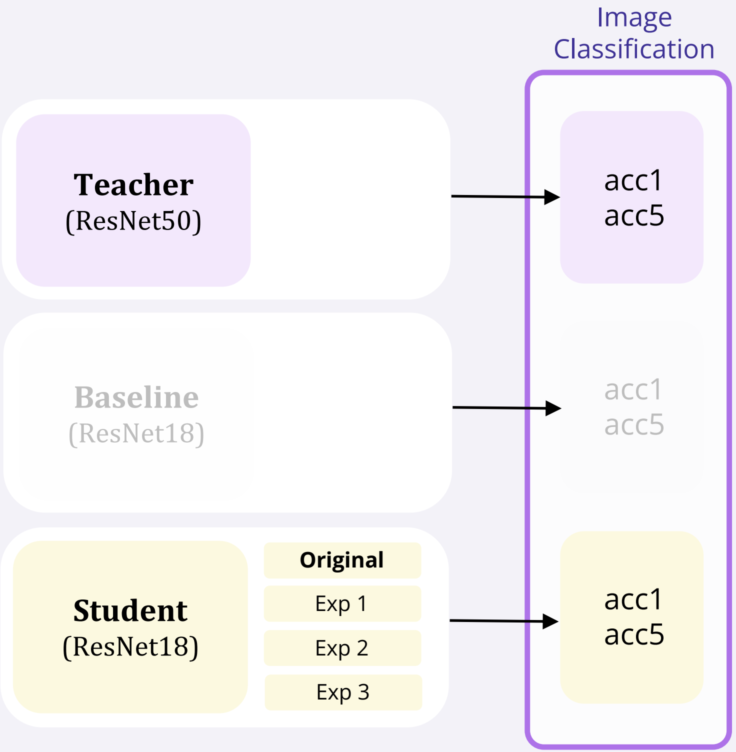
	Teacher	Baseline	Student			
			Original	No Aug	Feature Red	No Aug + Feature Red
Top 1 Acc	81.46	63.52	81.82	81.04	76.18	75.50
Top 5 Acc	96.24	87.32	95.58	95.84	93.10	92.18

Table 1: BYOL Results

	Teacher	Baseline	Student			
			Original	No Aug	Feature Red	No Aug + Feature Red
Top 1 Acc	80.28	-	77.02	77.52	71.56	71.34
Top 5 Acc	95.40	-	94.88	94.66	92.42	92.58

Table 2: MOCO-V2 Results

Evaluation Result



	Teacher	Baseline	Student			
			Original	No Aug	Feature Red	No Aug + Feature Red
Top 1 Acc	81.46	63.52	81.82	81.04	76.18	75.50
Top 5 Acc	96.24	87.32	95.58	95.84	93.10	92.18

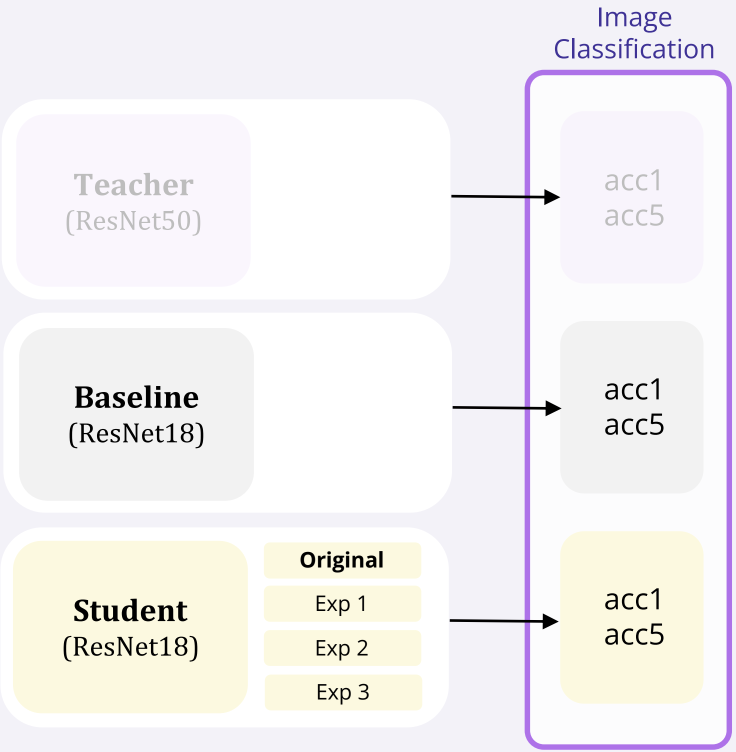
Table 1: BYOL Results

	Teacher	Baseline	Student			
			Original	No Aug	Feature Red	No Aug + Feature Red
Top 1 Acc	80.28	-	77.02	77.52	71.56	71.34
Top 5 Acc	95.40	-	94.88	94.66	92.42	92.58

Table 2: MOCO-V2 Results

➔ ***KASSL overall mimics teacher target well.***

Evaluation Result



	Teacher	Baseline	Student			
			Original	No Aug	Feature Red	No Aug + Feature Red
Top 1 Acc	81.46	63.52	81.82	81.04	76.18	75.50
Top 5 Acc	96.24	87.32	95.58	95.84	93.10	92.18

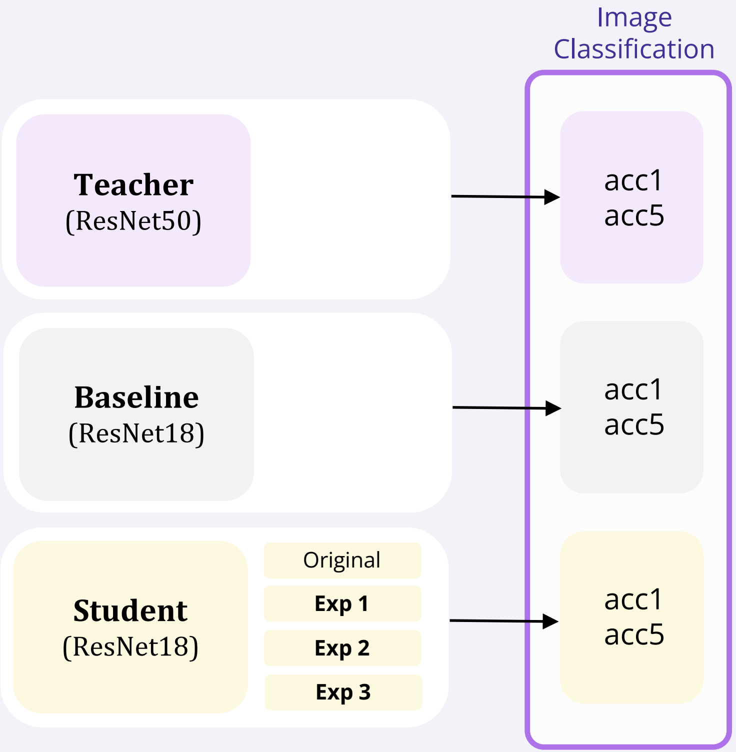
Table 1: BYOL Results

	Teacher	Baseline	Student			
			Original	No Aug	Feature Red	No Aug + Feature Red
Top 1 Acc	80.28	-	77.02	77.52	71.56	71.34
Top 5 Acc	95.40	-	94.88	94.66	92.42	92.58

Table 2: MOCO-V2 Results

➔ ***KASSL outperforms baseline model.***

Evaluation Result



	Teacher	Baseline	Student			
			Original	No Aug	Feature Red	No Aug + Feature Red
Top 1 Acc	81.46	63.52	81.82	81.04	76.18	75.50
Top 5 Acc	96.24	87.32	95.58	95.84	93.10	92.18

Table 1: BYOL Results

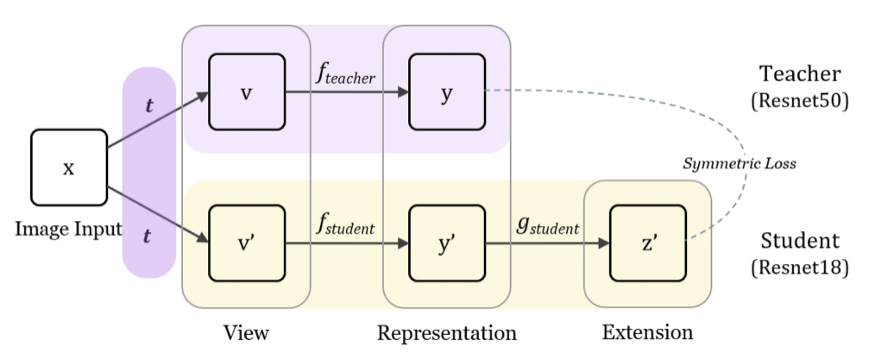
	Teacher	Baseline	Student			
			Original	No Aug	Feature Red	No Aug + Feature Red
Top 1 Acc	80.28	-	77.02	77.52	71.56	71.34
Top 5 Acc	95.40	-	94.88	94.66	92.42	92.58

Table 2: MOCO-V2 Results

Experiment 1 and Result

No Augmentation

- Apply same image into teacher and student network
- Hypothesis: Teacher produces answer for each image



Result

BYOL	Original	No Aug
Top-1	81.82	81.04
Top-5	95.58	95.84

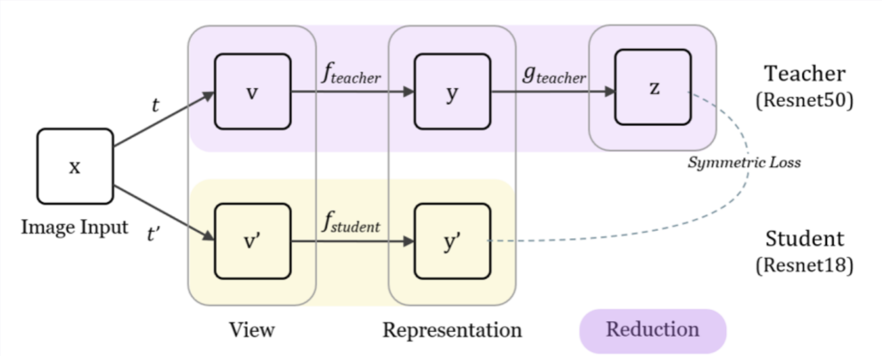
MoCo-v2	Original	No Aug
Top-1	77.02	77.52
Top-5	94.88	94.66

- *Work as well as original method*
- *Possibly work better on larger dataset*
  (e.g. ImageNet-1k)

Experiment 2 and Result

Feature Reduction

- Add projection layer at teacher network
- Hypothesis: Giving generalizable knowledge by reducing the output dimension



Result

BYOL	Original	Feature Red
Top-1	81.82	76.18
Top-5	95.58	93.10

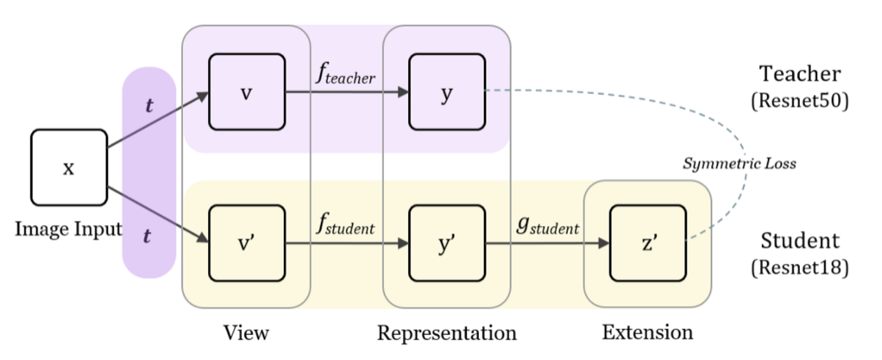
MoCo-v2	Original	Feature Red
Top-1	77.02	71.56
Top-5	94.88	92.42

- *Fail to improve original's performance*
- *Possibly lost information while training*

Experiment 3 and Result

No Augmentation & Feature Reduction

- Combine experiment 1 and 2
- Hypothesis: No Augmentation would complement the negative effect of feature reduction



Result

BYOL	Original	No Aug +F.R
Top-1	81.82	75.50
Top-5	95.58	92.18

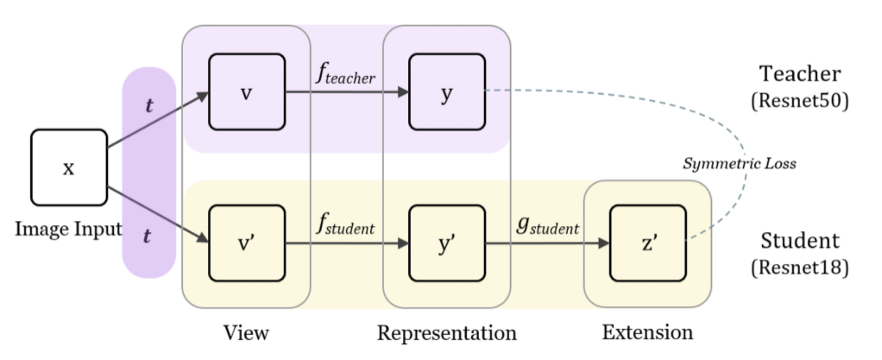
MoCo-v2	Original	No Aug +F.R
Top-1	77.02	71.34
Top-5	94.88	92.58

- *Fail to improve original's performance*
- *Possible that feature reduction has great impact on the model's performance*

Experiment 3 and Result

No Augmentation & Feature Reduction

- Combine experiment 1 and 2
- Hypothesis: No Augmentation would complement the negative effect of feature reduction



Result

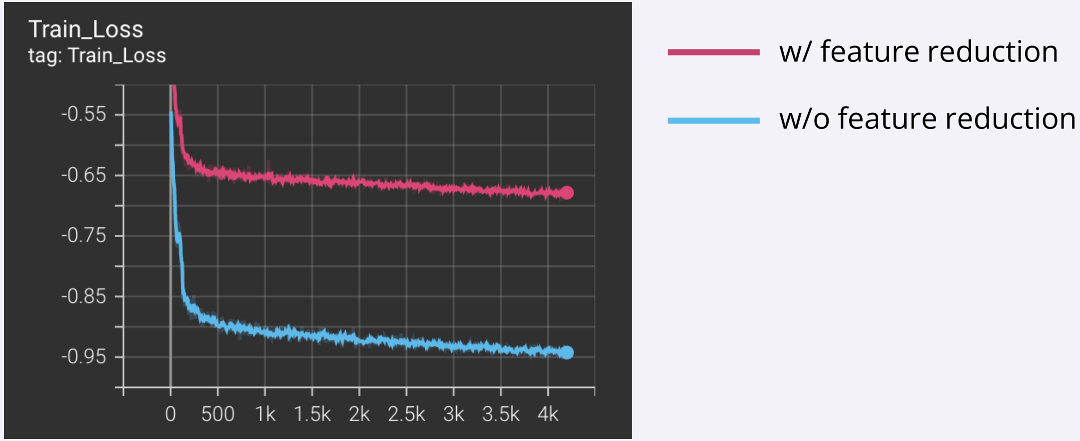
BYOL	Original	No Aug +F.R
Top-1	81.82	75.50
Top-5	95.58	92.18

MoCo-v2	Original	No Aug +F.R
Top-1	77.02	71.34
Top-5	94.88	92.58

- *Fail to improve original's performance*
- *Possible that feature reduction has great impact on the model's performance*

Experiment Result Analysis

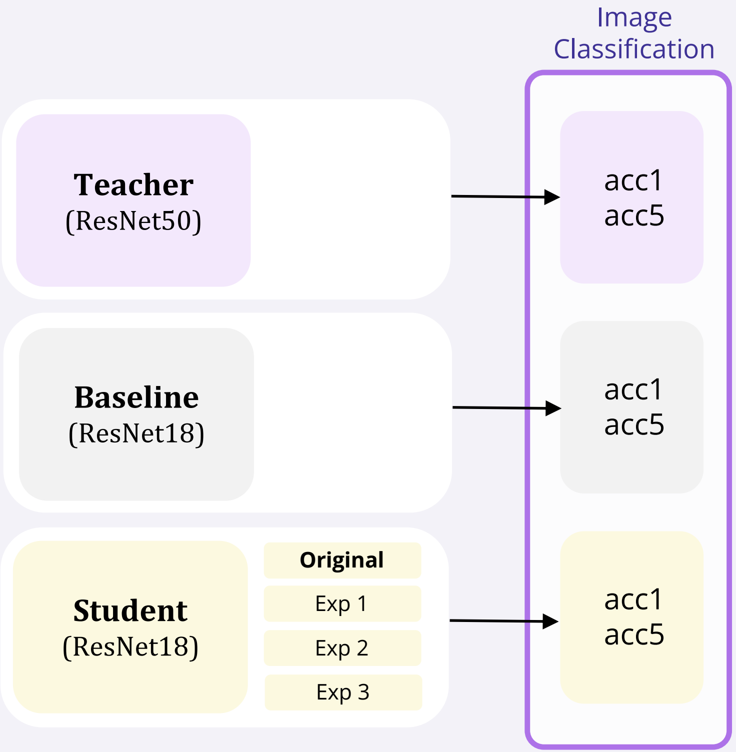
Non-Feature Reduction vs. Feature Reduction



- Training graph w/o feature reduction converges around -0.9
- Training graph w/ feature reduction converges around -0.6

➔ ***Difficulty in learning knowledge when reducing feature representation***

Evaluation Result



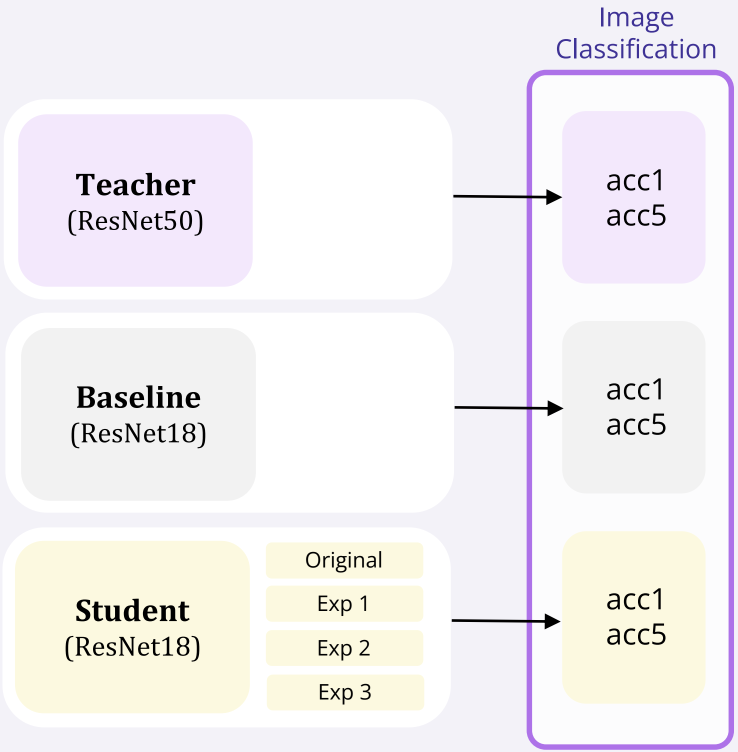
	Teacher	Baseline	Student			
			Original	No Aug	Feature Red	No Aug + Feature Red
Top 1 Acc	81.46	63.52	81.82	81.04	76.18	75.50
Top 5 Acc	96.24	87.32	95.58	95.84	93.10	92.18

Table 1: BYOL Results

	Teacher	Baseline	Student			
			Original	No Aug	Feature Red	No Aug + Feature Red
Top 1 Acc	80.28	-	77.02	77.52	71.56	71.34
Top 5 Acc	95.40	-	94.88	94.66	92.42	92.58

Table 2: MOCO-V2 Results

Evaluation Result



	Teacher	Baseline	Student			
			Original	No Aug	Feature Red	No Aug + Feature Red
Top 1 Acc	81.46	63.52	81.82	81.04	76.18	75.50
Top 5 Acc	96.24	87.32	95.58	95.84	93.10	92.18

Table 1: BYOL Results

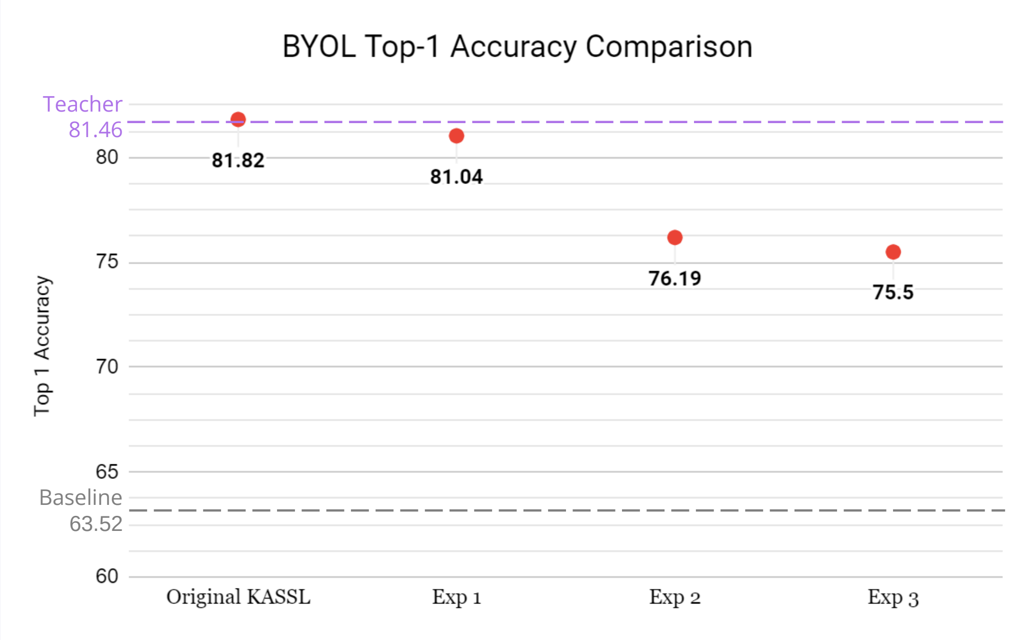
	Teacher	Baseline	Student			
			Original	No Aug	Feature Red	No Aug + Feature Red
Top 1 Acc	80.28	-	77.02	77.52	71.56	71.34
Top 5 Acc	95.40	-	94.88	94.66	92.42	92.58

Table 2: MOCO-V2 Results

Evaluation Result

1. **KASSL VS Teacher**

2. **KASSL VS Baseline**

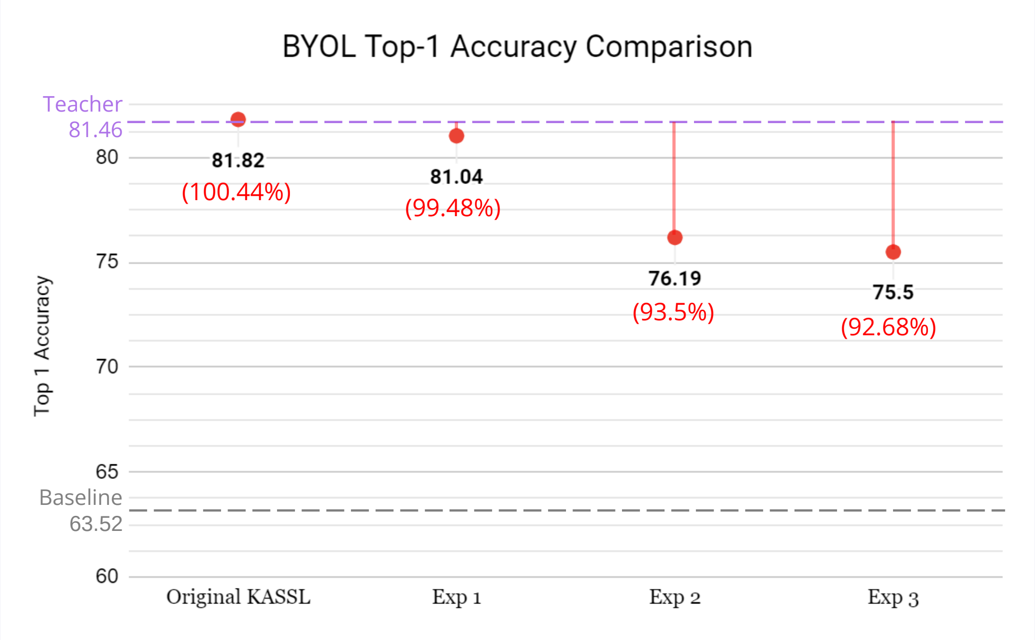


Evaluation Result

1. KASSL VS Teacher

- Mimics teacher's performance by
  - Average: 96.53%
  - Best: 100.44%

2. KASSL VS Baseline



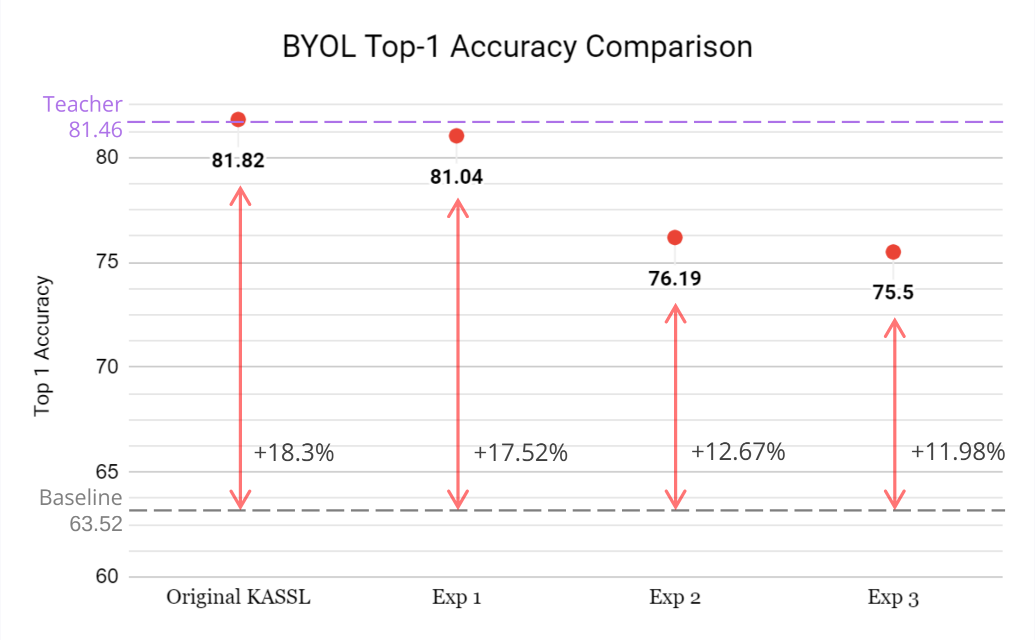
Evaluation Result

1. KASSL VS Teacher

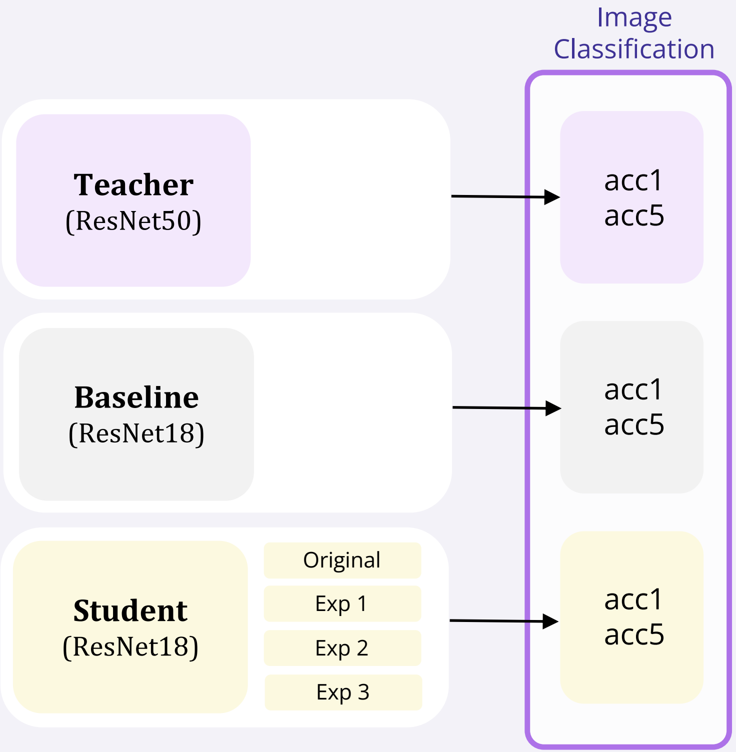
- Mimics teacher's performance by
  - Average: 96.53%
  - Best: 100.44%

2. KASSL VS Baseline

- Outperform existing method by
  - Average: +15.12%
  - Best: +18.3%



Evaluation Result



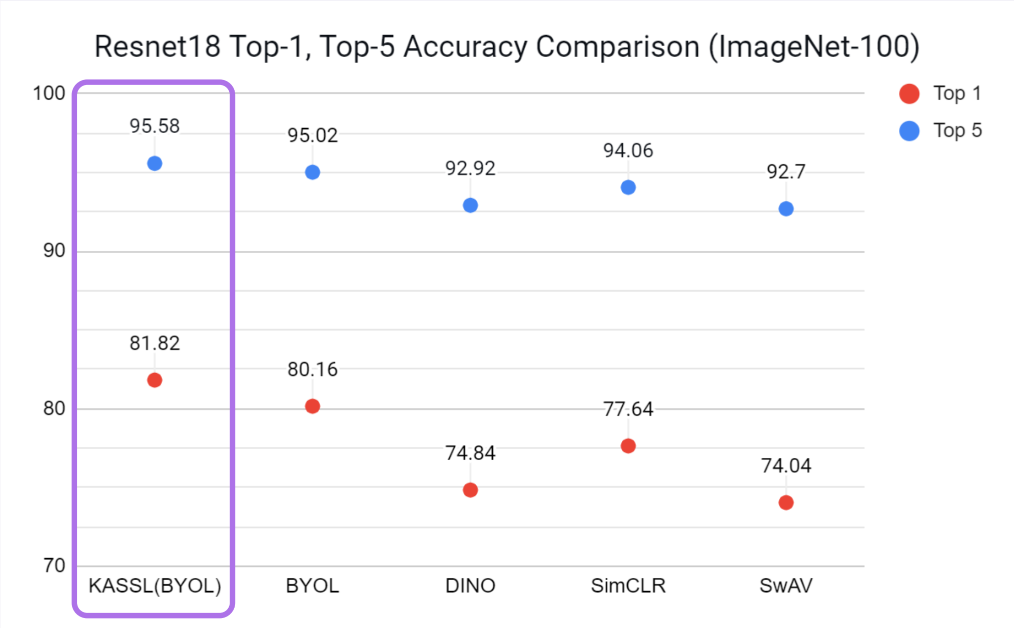
	Teacher	Baseline	Student			
			Original	No Aug	Feature Red	No Aug + Feature Red
Top 1 Acc	81.46	63.52	81.82	81.04	76.18	75.50
Top 5 Acc	96.24	87.32	95.58	95.84	93.10	92.18

Table 1: BYOL Results

	Teacher	Baseline	Student			
			Original	No Aug	Feature Red	No Aug + Feature Red
Top 1 Acc	80.28	-	77.02	77.52	71.56	71.34
Top 5 Acc	95.40	-	94.88	94.66	92.42	92.58

Table 2: MOCO-V2 Results

Evaluation Result



Contribution

New method to train well-performing small networks in SSL domain w/ KD

- Successfully mimic teacher network
- Outperform networks trained by SSL methods
- Independent to how teacher network is trained

New method to train small networks effectively

- Outperform networks trained by SSL methods w/ only half epochs

Limitation and Challenges

Limitation

- Lack of large dataset (e.g. ImageNet-1K)
- Lack of experiment in various student architecture

Challenge

- ~