华东师范大学

Knowledge Distillation (2015-2019)

Yuang Liu
AIDA, ECNU
frankliu624@gmail.com

2019/10/6

Knowledge Distillation (2015-2019)

- Introduction & Motivation
- Survey of Knowledge Distillation
- Applications of Knowledge Distillation
- Knowledge Distillation and GAN
- Beyond Knowledge Distillation

Introduction

Model compression and acceleration

- Parameter pruning and sharing
- Low-rank factorization and sparsity
- Transferred/compact convolutional filters
- Knowledge distillation

Buciluă et al. "Model compression." SIGKDD 2006

Do Deep Nets Really Need to be Deep?

- Training Shallow Nets to Mimic Deep Nets
  - Mimic Learning via Regressing Logit with L2 Loss
  - Speeding-up Mimic Learning by Introducing a Linear Layer
- The Capacity and Representational Power of Shallow Models

Ba, Jimmy, and Rich Caruana. "Do deep nets really need to be deep?." NIPS 2014

Knowledge Distillation

Knowledge Distillation aims to compress and improve the model by transfering knowledge from deep nets to a small network.

Survey of KD

- three basic work
- kinds of knowledge
- multi-teacher and Self-KD

Distilling the knowledge in a neural network

The pioneer of KD.

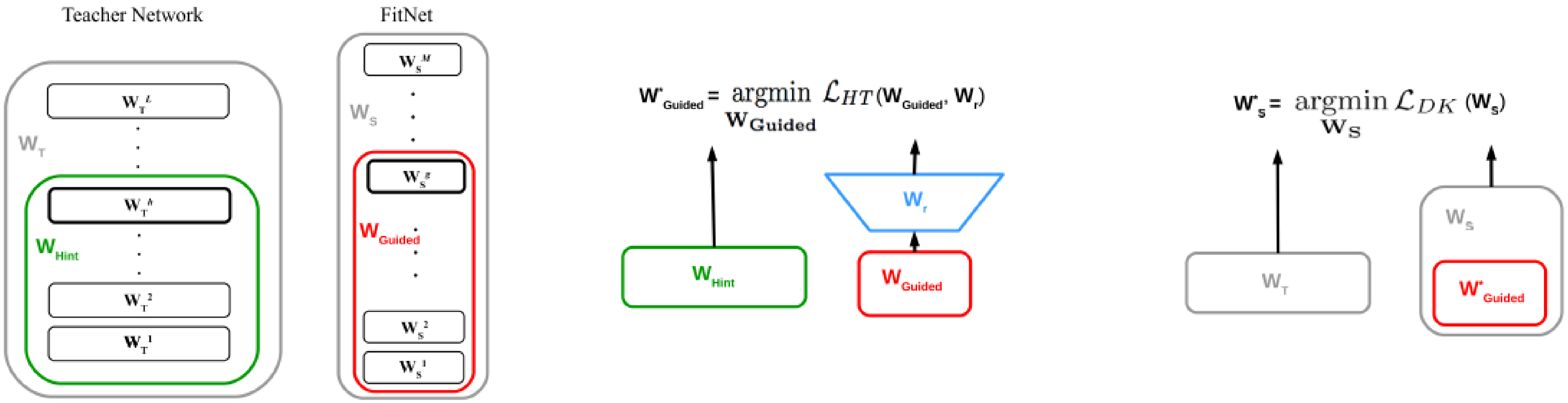
- Distilling the knowledge in an ensemble of models into a single model.
- Achieve model compression and performance improvement.
- Transfer knowledge from teacher to student through soft-targets.

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

Hinton et al. "Distilling the knowledge in a neural network." arXiv:1503.02531

FitNets: Hints for Thin Deep Nets

The first one considering intermediate layer.



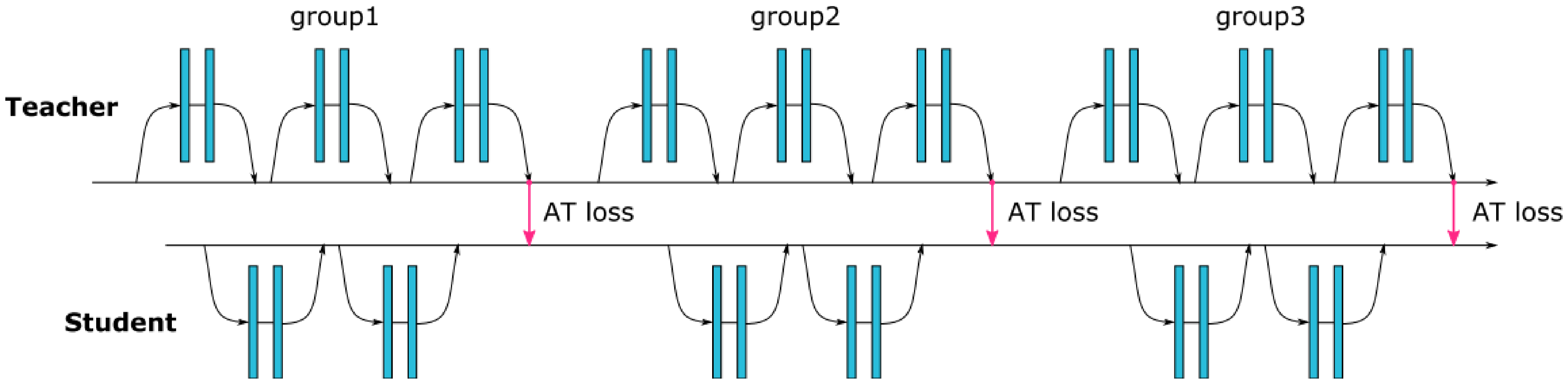
(a) Teacher and Student Networks (b) Hints Training (c) Knowledge Distillation

$$\mathcal{L}_{HT}(\mathbf{W_{Guided}}, \mathbf{W_r}) = \frac{1}{2}||u_h(\mathbf{x}; \mathbf{W_{Hint}}) - r(v_g(\mathbf{x}; \mathbf{W_{Guided}}); \mathbf{W_r})||^2$$

Romero, Adriana et al. "Fitnets: Hints for thin deep nets." arXiv:1412.6550

Attention Transfer

- Activation-based attention transfer



$$\mathcal{L}_{AT} = \mathcal{L}(\mathbf{W}_S, x) + \frac{\beta}{2} \sum_{j \in \mathcal{I}} \left\| \frac{Q_S^j}{\|Q_S^j\|_2} - \frac{Q_T^j}{\|Q_T^j\|_2} \right\|_p$$

- Gradient-based attention transfer

Attention Map

Zagoruyko et al. "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer." ICLR 2017

Privileged Information

- Vapnik, Vladimir, and Rauf Izmailov. "Learning using privileged information: similarity control and knowledge transfer." MLR 2015
- Lopez-Paz, David, et al. "Unifying distillation and privileged information." *arXiv 2015*

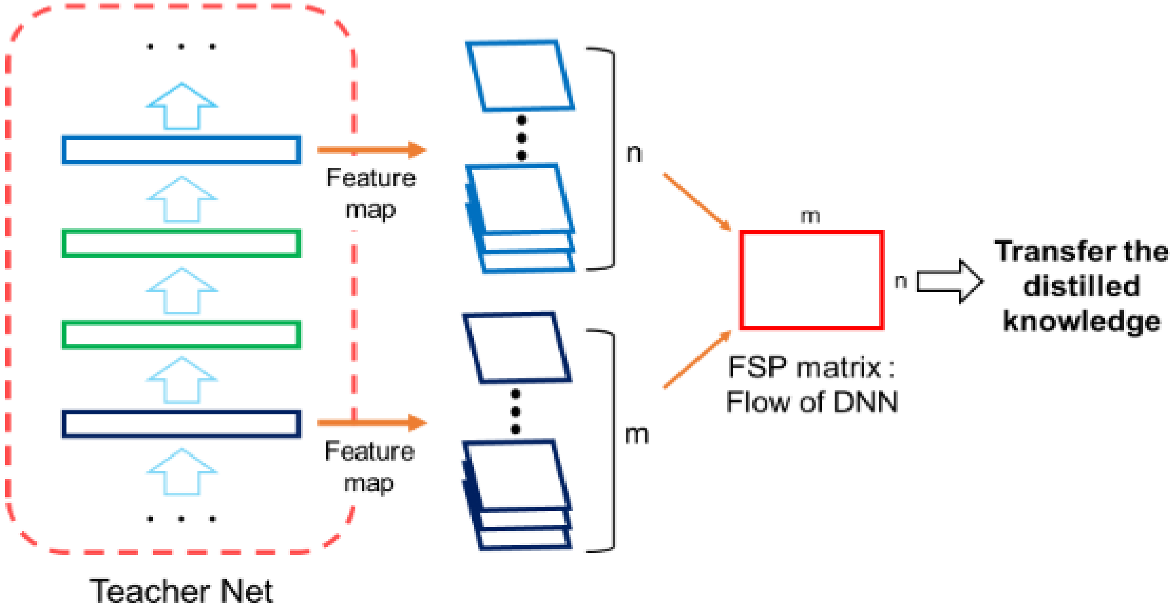
- Privileged Information

  teacher knows but student not.

Survey of KD

- three basic work
- **kinds of knowledge**
- multi-teacher and Self-KD

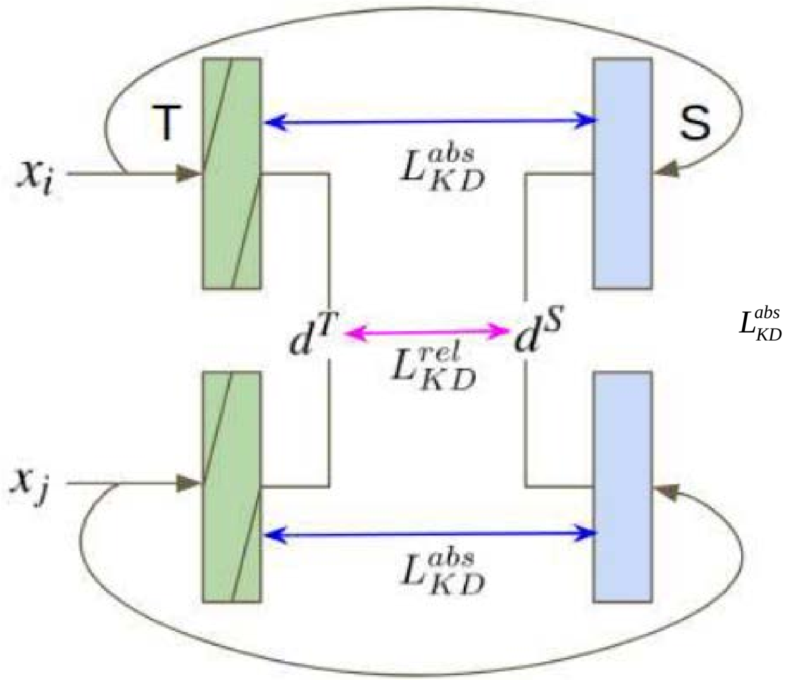
FSP Matrix

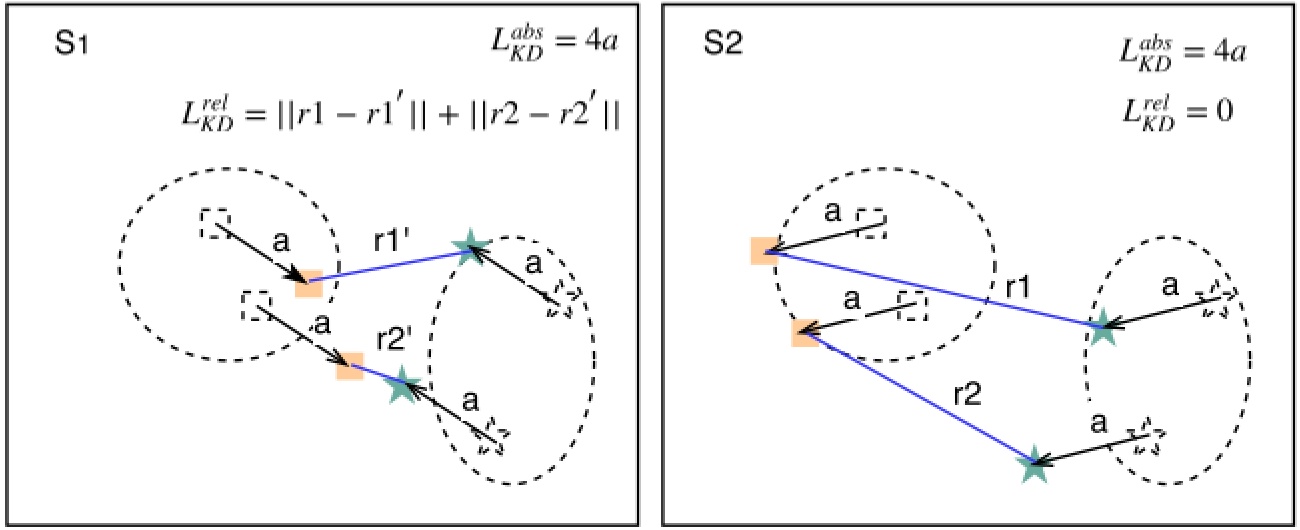


- the flow of solving a problem(FSP) can be defined as the relationship between features from two layers.
- FSP loss for hint, not use label

Yim, Junho et al. "A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning." CVPR 2017

Metrics Learning

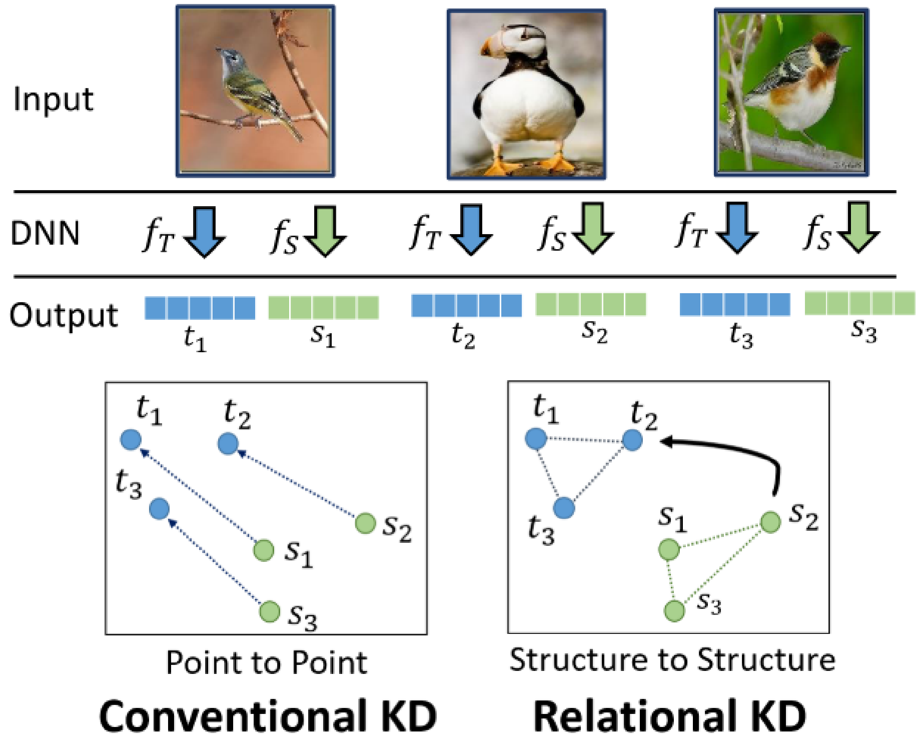


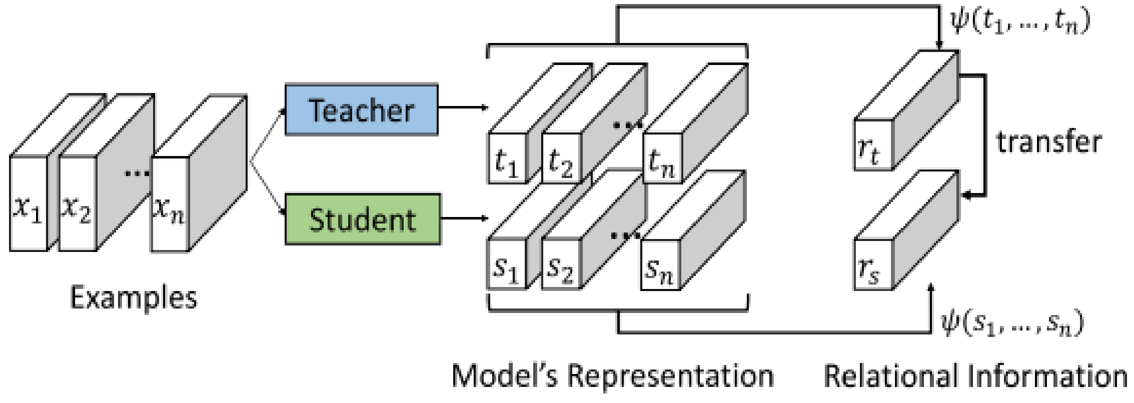


- $L_{KD}^{abs} = \left\| F^S(x_i) - F^T(x_i) \right\|$
- $L_{KD}^{rel} = \left| d^S - d^T \right|, \quad \begin{aligned} d^S &= \left\| F^S(x_i) - F^S(x_j) \right\| \\ d^T &= \left\| F^T(x_i) - F^T(x_j) \right\| \end{aligned}$

Yu, Lu et al. "Learning Metrics from Teachers: Compact Networks for Image Embedding." CVPR 2019

Relational Knowledge Distillation(RKD)



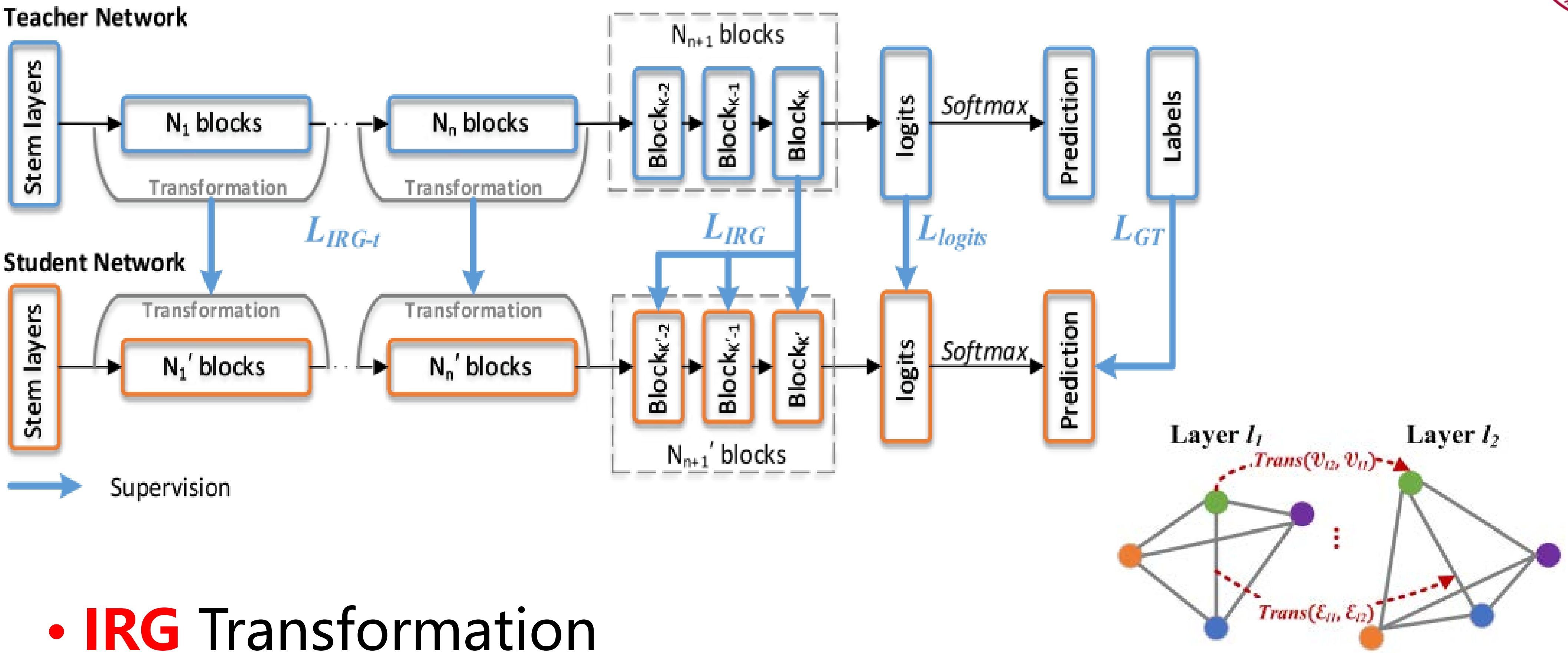


Relational Knowledge Distillation

- Relation between logits
- could work with AT or FitNet

Park, Wonpyo et al. "Relational Knowledge Distillation." CVPR 2019

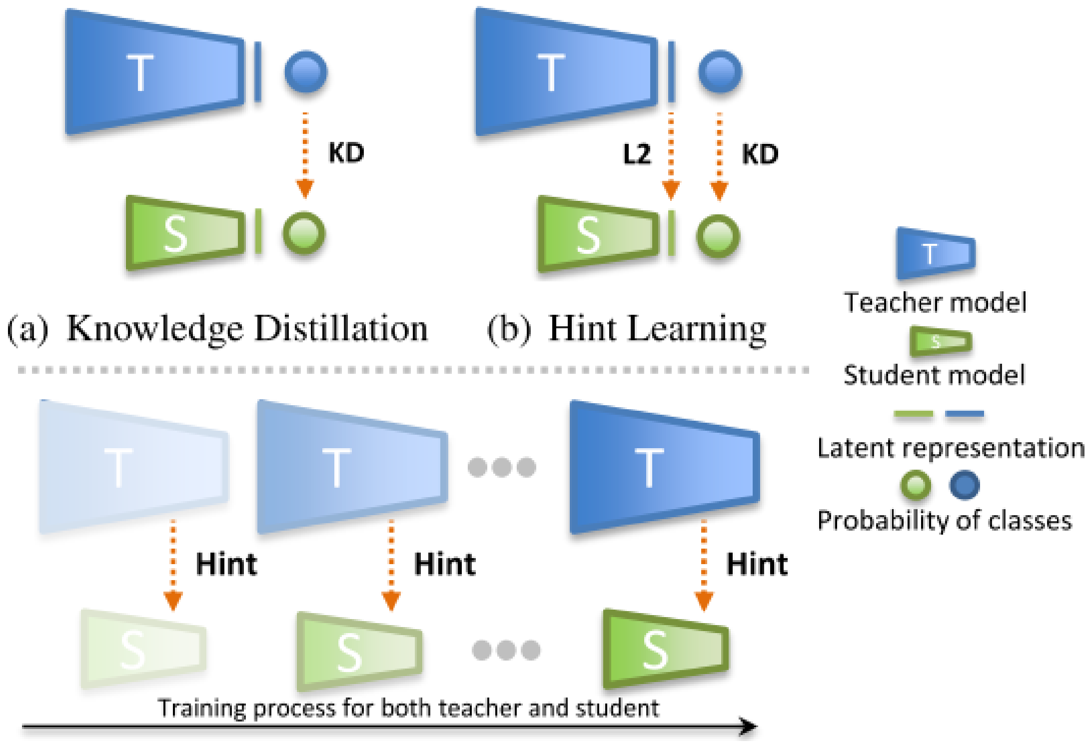
Instance Relationship Graph



- **IRG** Transformation
- Multi-Type Knowledge Loss: $L_{IRG-t}, L_{IRG}, L_{\text{logits}}, L_{GT}$

Liu, Yufan et al. "Knowledge Distillation via Instance Relationship Graph." CVPR 2019

Route Constrained Optimization



Algorithm 1 Route Constrained Optimization

Require: anchor points set from pre-trained teacher network: $C_1, C_2, ..., C_n$, student network with parameter W_i

$i = 1$

Randomly initialize W_i

while $i \leq n$ **do**

  Initialize teacher network with C_i anchor, get W_{C_i}

  if $i > 1$ **then**

    Initialize W_i with W_{i-1}

  end if

  update the W_i by optimizing $L_{KD}(W_i, W_{C_i})$

  $i = i + 1$

end while

get W_n as the final weights of student.

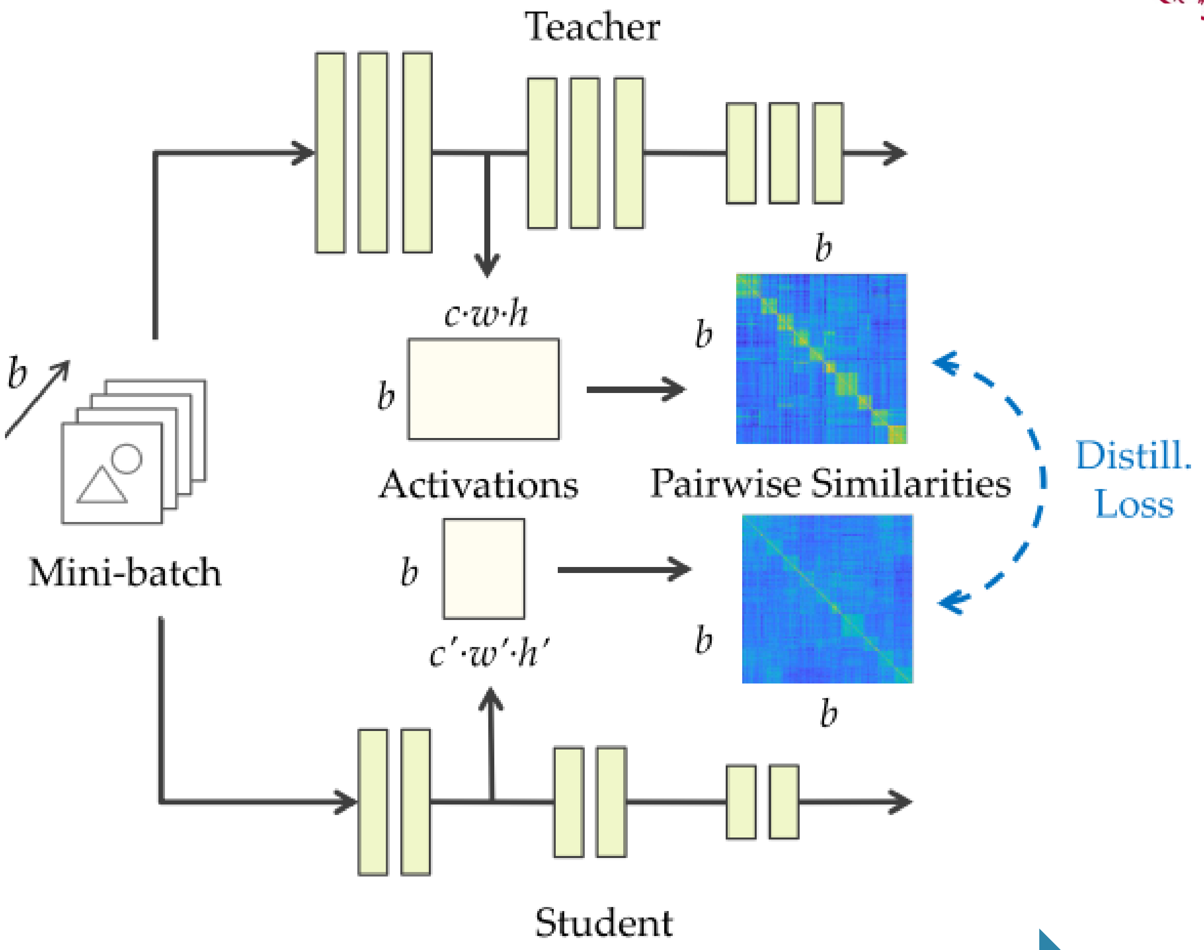
- **just multi hint**

$$Loss_i = H(y, \phi_s(x; W_s)) + \lambda H(\phi_s(x; W_s), \phi_t(x; W_{C_i}))$$

Jin, Xiao et al. "Knowledge Distillation via Route Constrained Optimization." ICCV 2019

Similarity-Preserving

- Activations
- Pair-wise Similarities
- a new hint by **similarity maps** between instance



Tung, Frederick, and Mori Greg. "Similarity-Preserving Knowledge Distillation." ICCV 2019

Survey of KD

- three basic work
- kinds of knowledge
- **multi-teacher and Self-KD**

Noisy Teachers

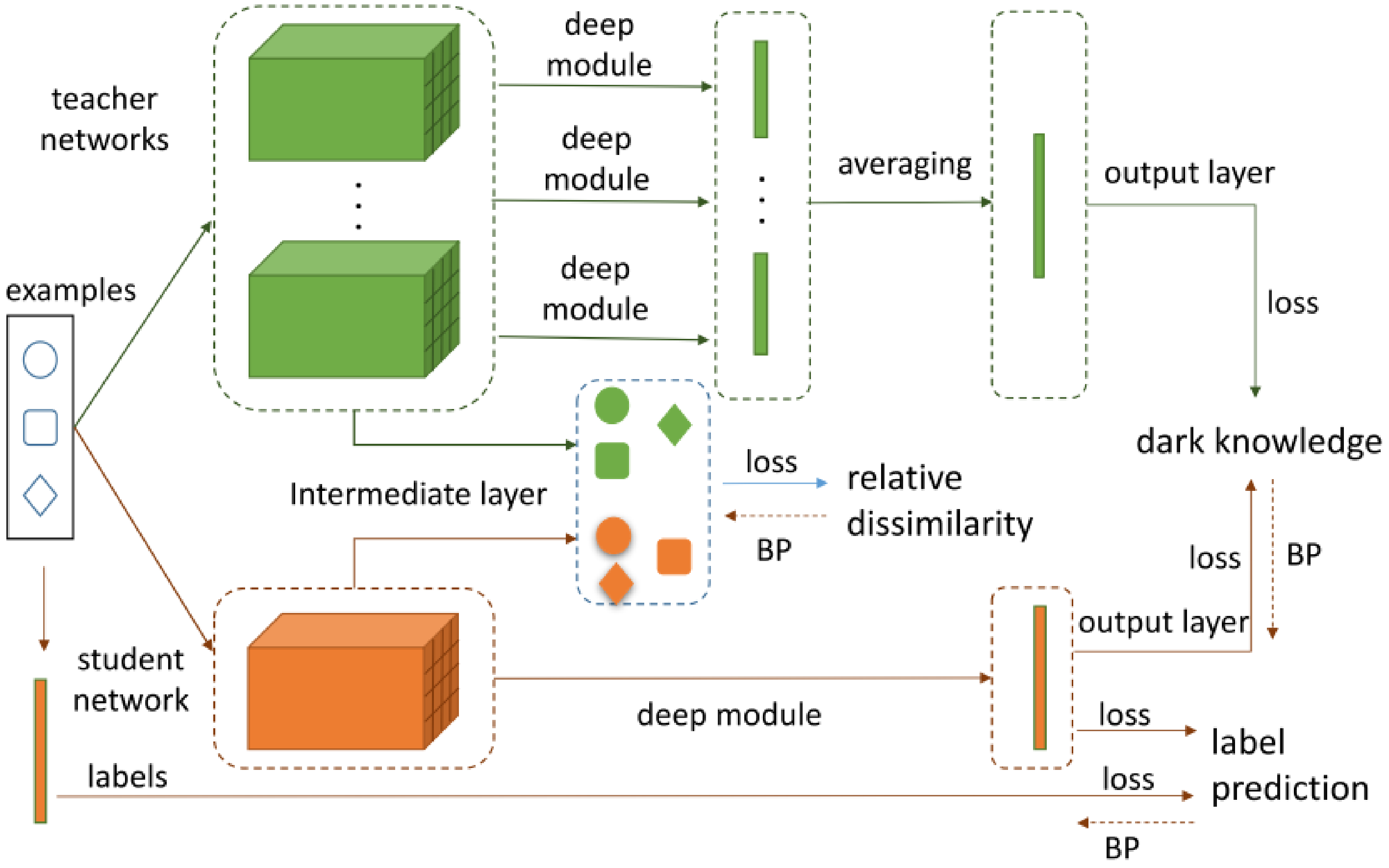
- noise-based regularization
- add noise to logits from teacher
- estimate multi-teacher
- different Noise Level, different performance

Algorithm 1 Training Student with Logit Perturbation

Input: Training Data $\mathcal{D}=(x,z)$, probability α, std σ
Initialization: θ_0 = Model Parameters of student model
for each mini-batch $\mathcal{D}_t = \{x_t, z_t\}$ **do**
  Generate ξ, $\xi \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$
  Select samples from x_t with probability α
  Perturb corresponding logit values in z_t using Eqn (2)
  Calculate L2 loss using Eqn (3)
  Update model parameters θ_t using Eqn (4)
end for
Output: θ = Trained Student Model Parameters

Sau et al. “Deep Model Compression: Distilling Knowledge from Noisy Teachers.” arXiv:1610.09650v2

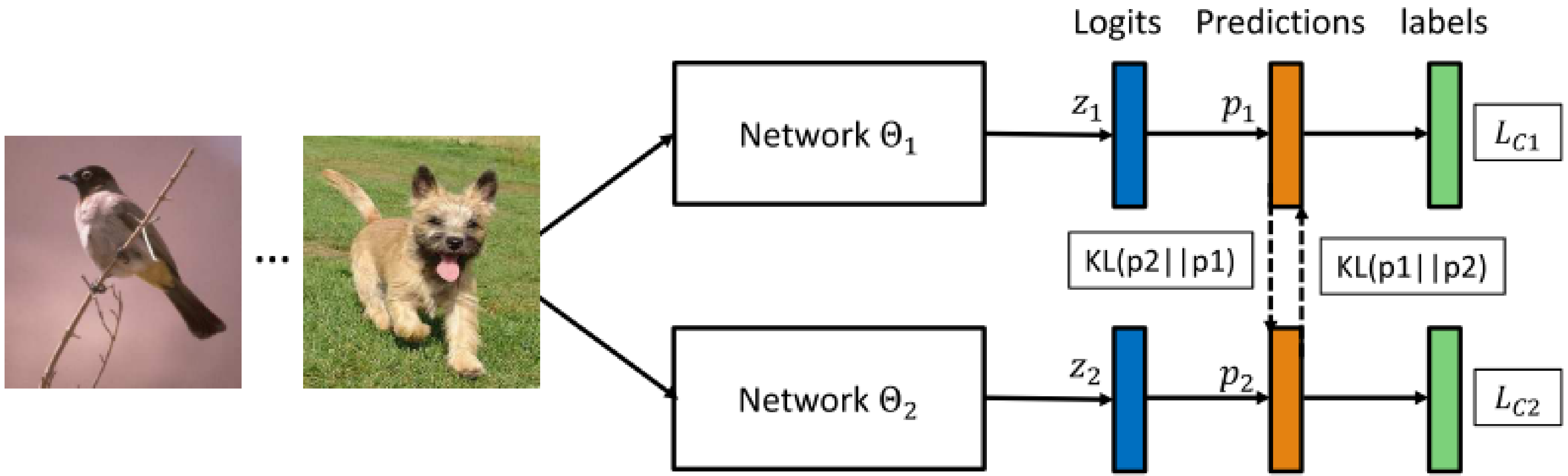
Learning from Multiple Teacher Networks



- Average strategy, not work independently
- Relative dissimilarity by triplet selection

You, Shan et al. "Learning from Multiple Teacher Networks." KDD 2017

Deep Mutual Learning(DML)

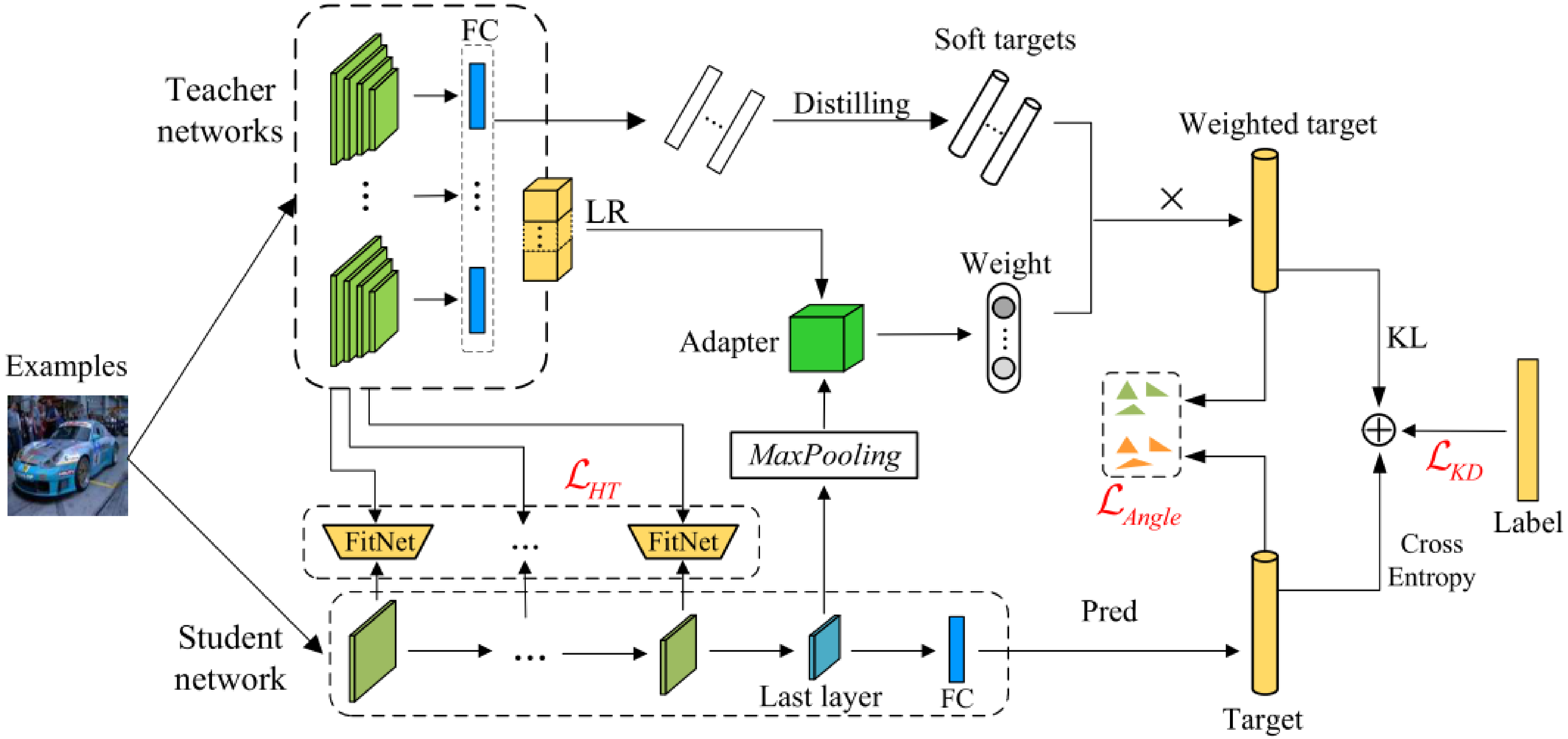


- only *softmax* logits and KL loss
- learn from each other

$$L_{\Theta_k} = L_{C_k} + \frac{1}{K-1} \sum_{l=1, l \neq k}^{K} D_{KL}(\boldsymbol{p}_l \| \boldsymbol{p}_k)$$

Zhang, Ying et al. "Deep Mutual Learning." CVPR 2018

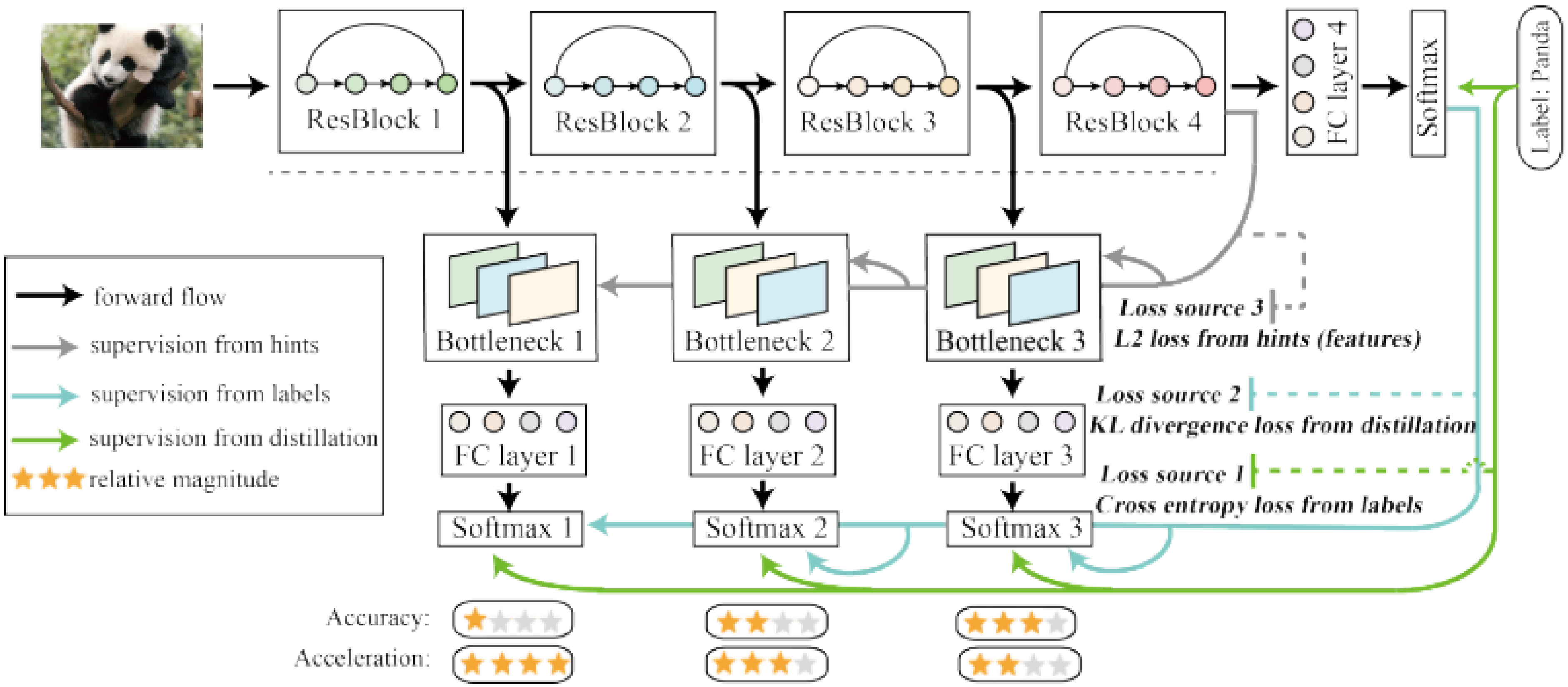
Adaptive Multi-Teacher Multi-Level Knowledge Distillation (Ours)



- Instance-level adaptive weight for teachers
- Multi-Group Hint strategy

Liu, Yuang, Zhang Wei, and Wang jun. "Adaptive Multi-Teacher Multi-Level Knowledge Distillation." 2019.09

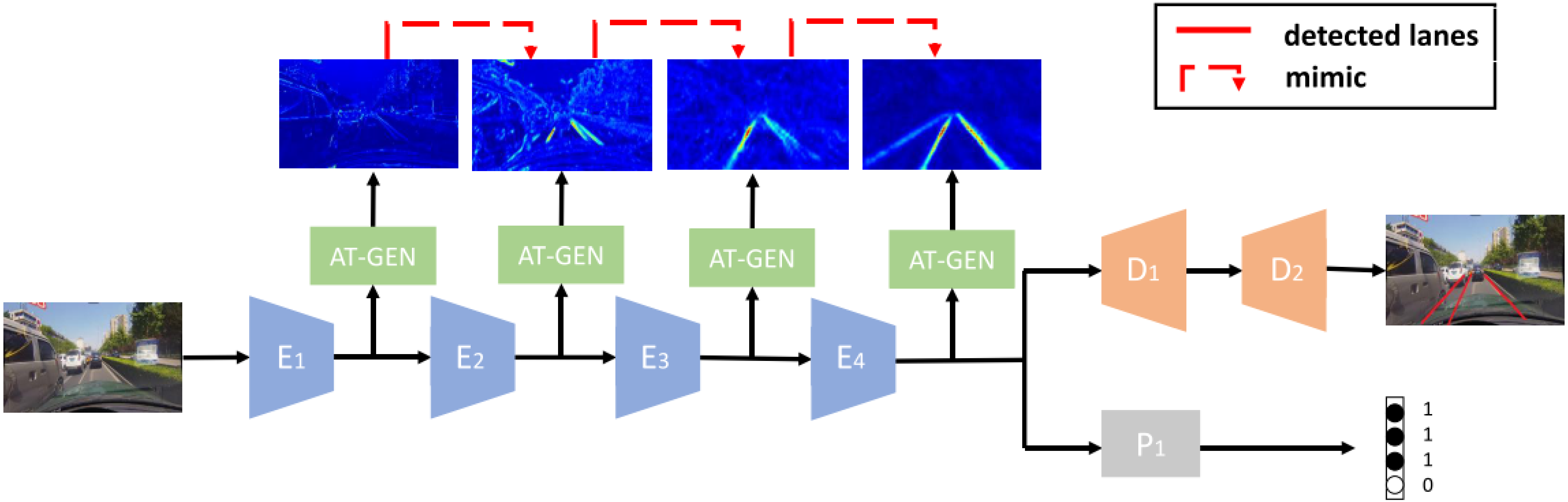
Be Your Own Teacher



- 3 loss sources： hints， KL， labels
- the deepest guides all the front layers

Zhang, Linfeng et al. "Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation." ICCV 2019

Self Attention Distillation



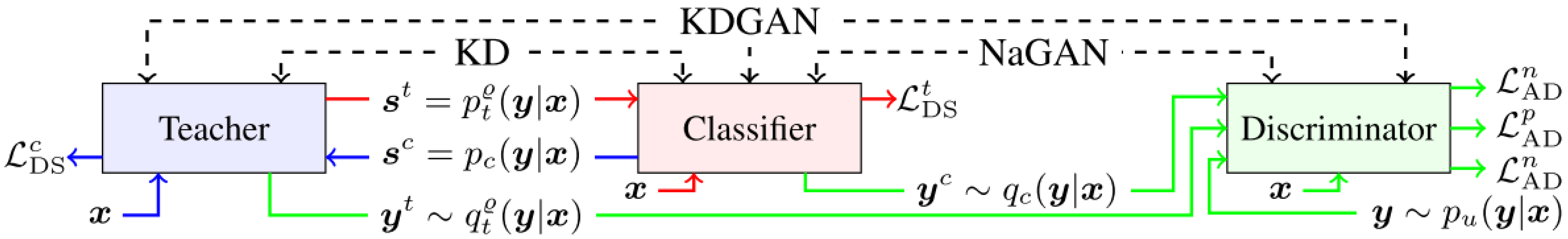
- Attention map from AT-GEN
- deeper layer hints thinner layer

Hou, Yuenan et al. "Learning Lightweight Lane Detection CNNs by Self Attention Distillation." ICCV 2019

Applications of KD

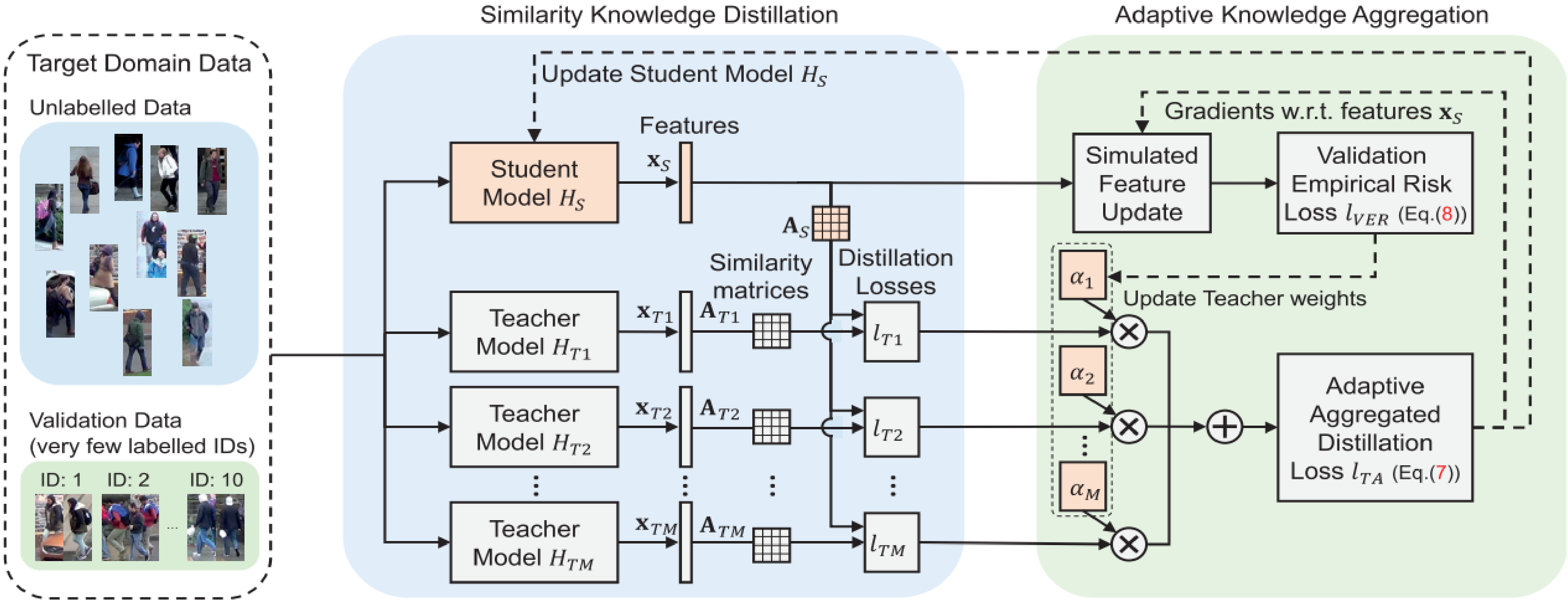
for GAN

- Mainly to compress and improve the generator,
- but limited in classification task.



Wang, Xiaojie et al. "KDGAN: knowledge distillation with generative adversarial networks." NIPS 2018

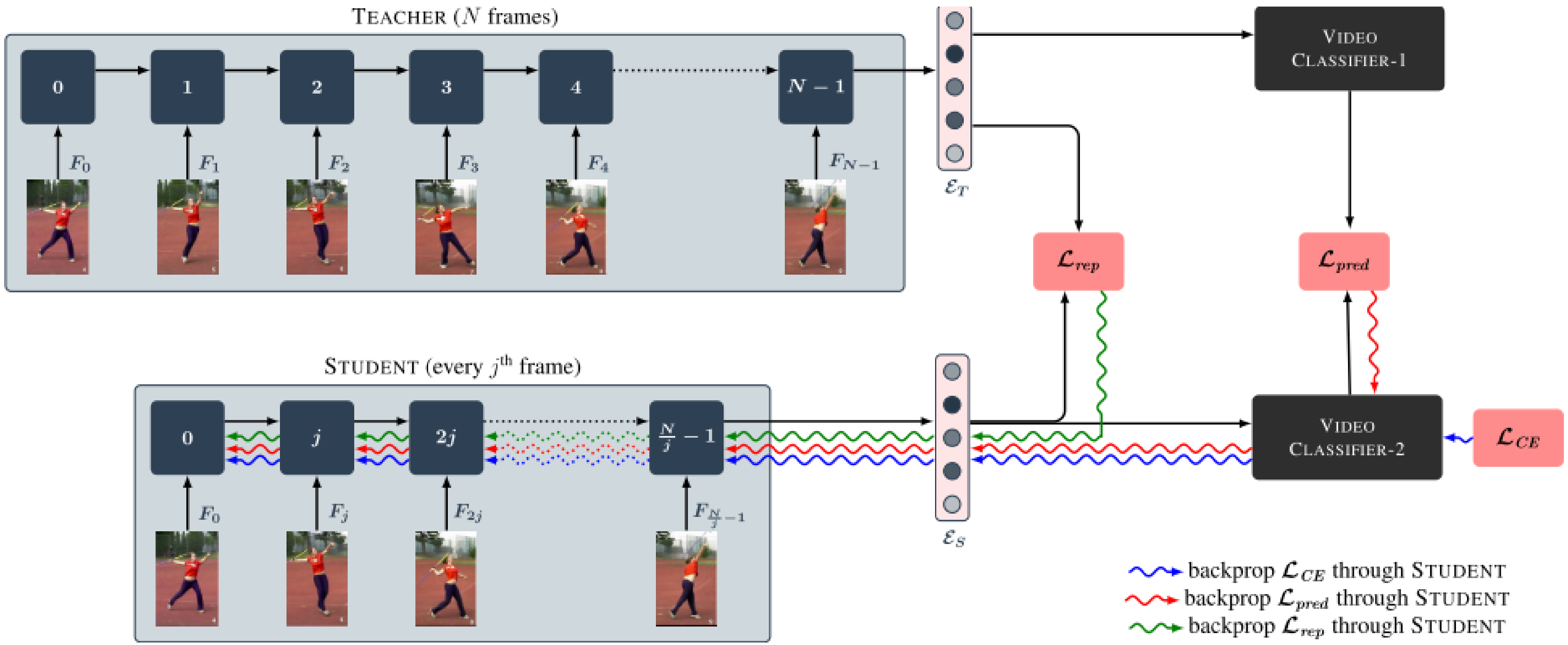
for Person Re-identification



- Similarity Knowledge Distillation
- Multiple Teachers, Adaptive Knowledge Aggregation

Wu, Ancong et al. "Distilled Person Re-identification: Towars a More Scalable System. “ CVPR 2019

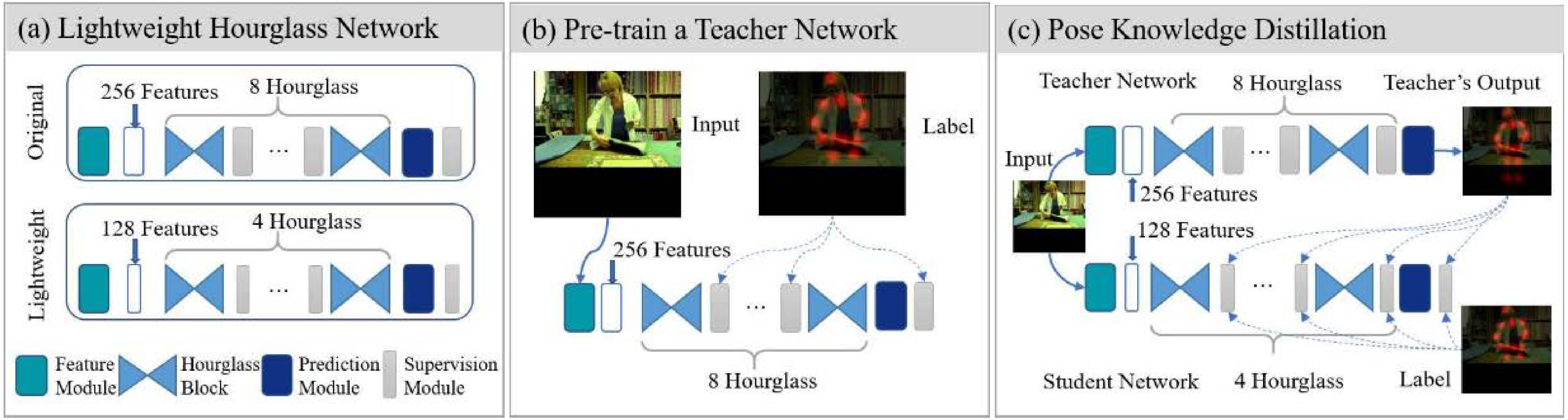
for Video Classification



- few frame, small student
- original KD with class label, no hint

Bhardwaj, Shweta et al. "Efficient Video Classification Using Fewer Frames." CVPR 2019

for Pose Estimation



- confidence map as knowledge
- no hint

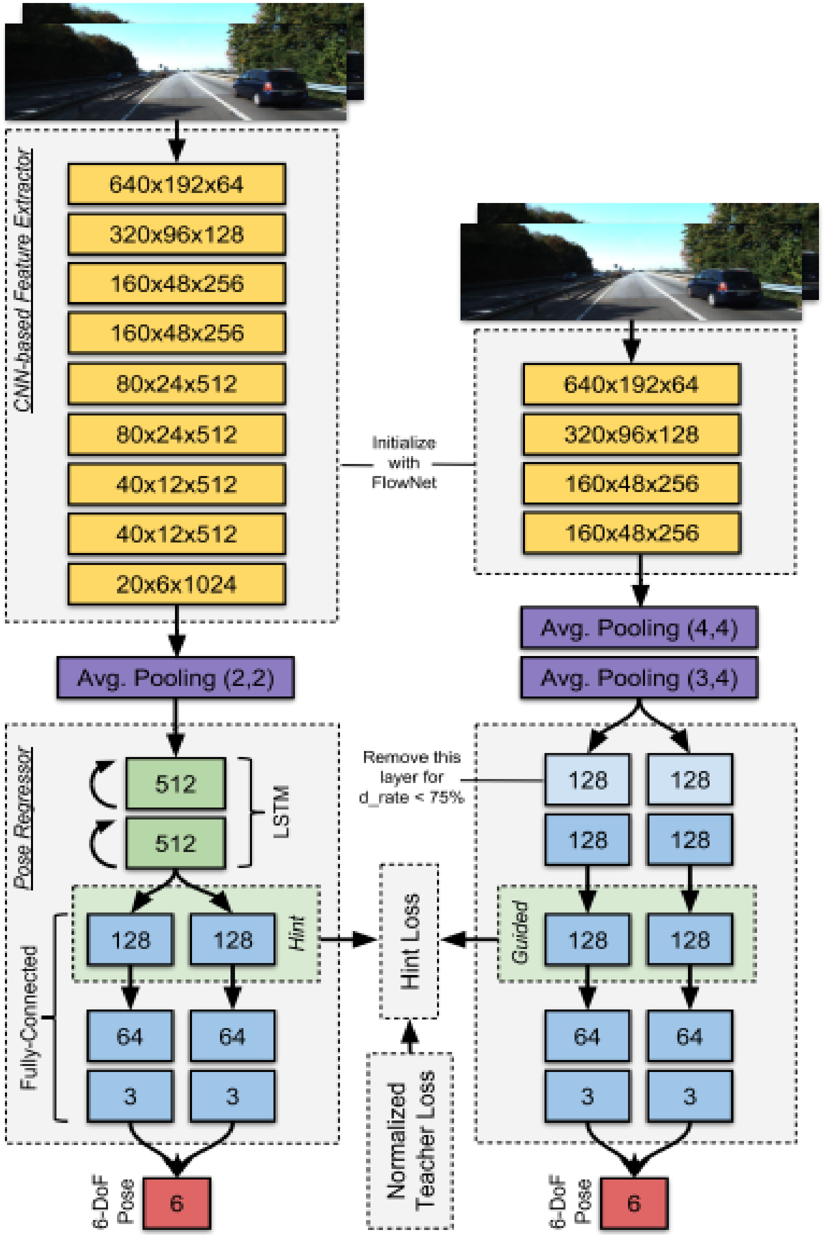
Zhang, Feng et al. "Fast Human Pose Estimation." CVPR2019.

for Pose Regressor

Loss:

- attentive imitation loss(AIL)
- attentive hint training(AHT)

only hint



Saputra et al. "Distilling knowledge from a deep pose regressor network." arXiv:1908.00858 (2019).

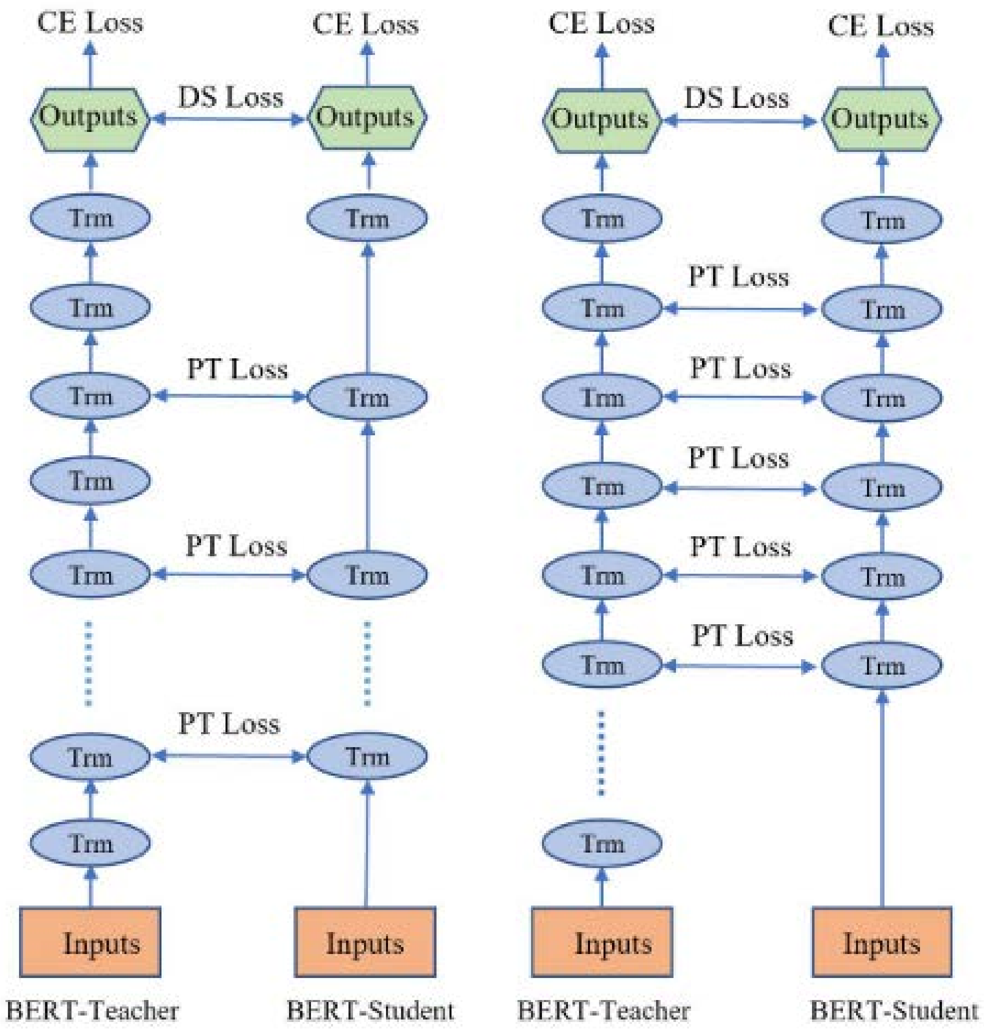
for BERT(NLP)

PT Loss:

- PKD-Last
- PKD-Skip

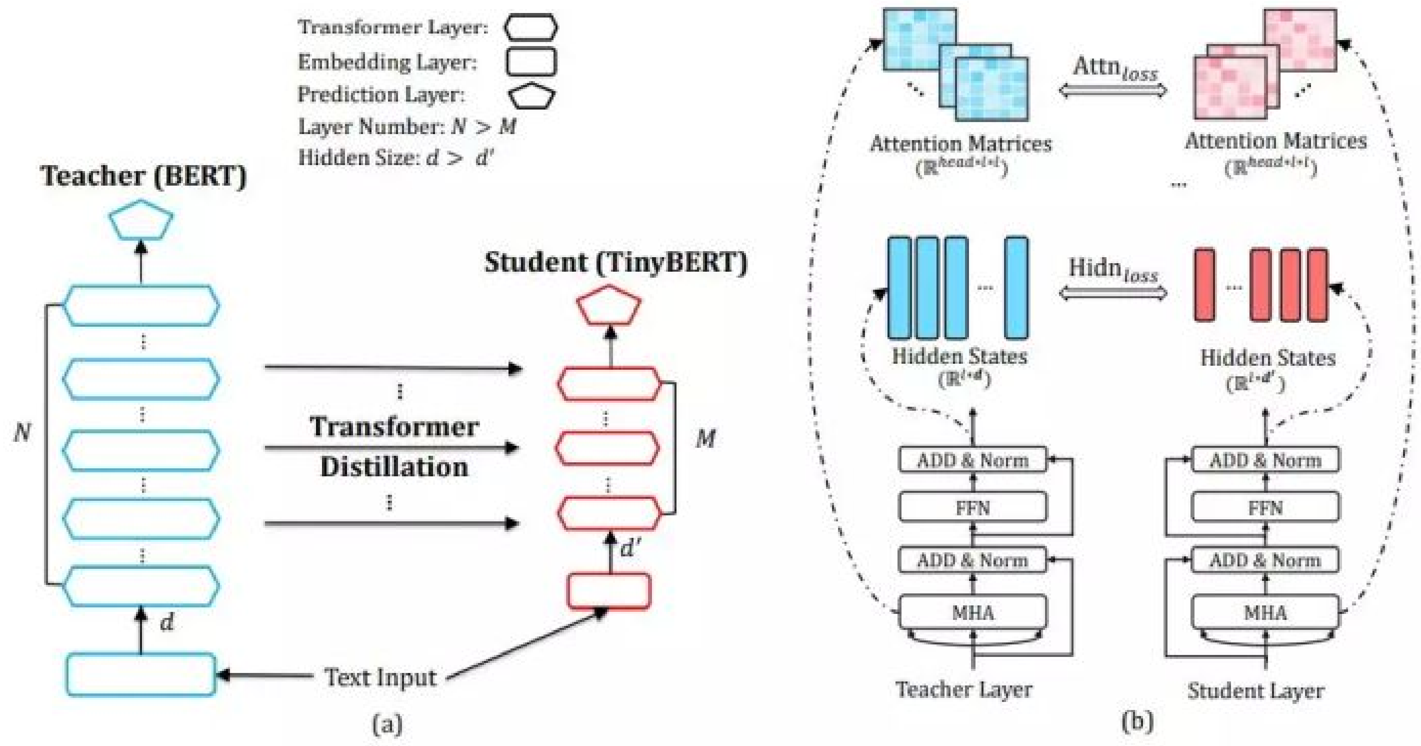
CE loss + hint(PT)

no new idea.



Sun, Siqi et al. "Patient Knowledge Distillation for BERT Model Compression." arXiv:1908.09355

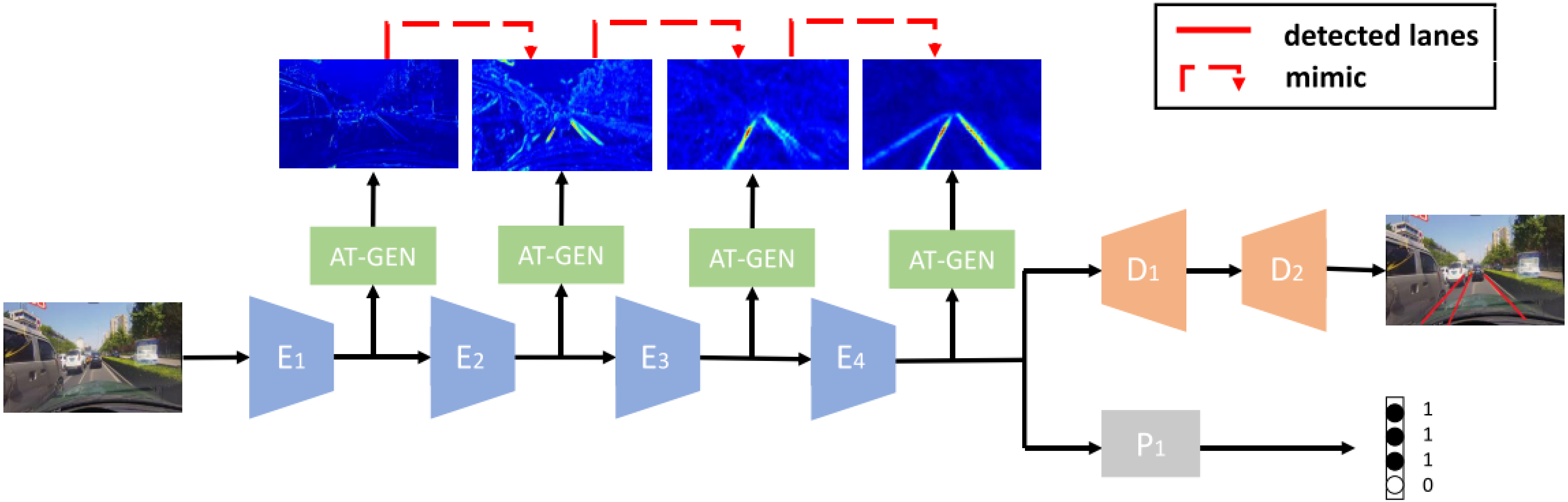
for BERT(NLP) – TinyBERT



- Transformer(hint)
- 2-step : Pre-training and fine-tuning

Jiao, Xiaoqi et al. "TinyBERT: Distilling BERT for Natural Language Understanding. *arXiv: 1909.10351*

for Lane Detection



- Self Attention Distillation
- no additional labels

Hou, Yuenan et al. “Learning Lightweight Lane Detection CNNs by Self Attention Distillation.” ICCV 2019

for Semantic Segmentation

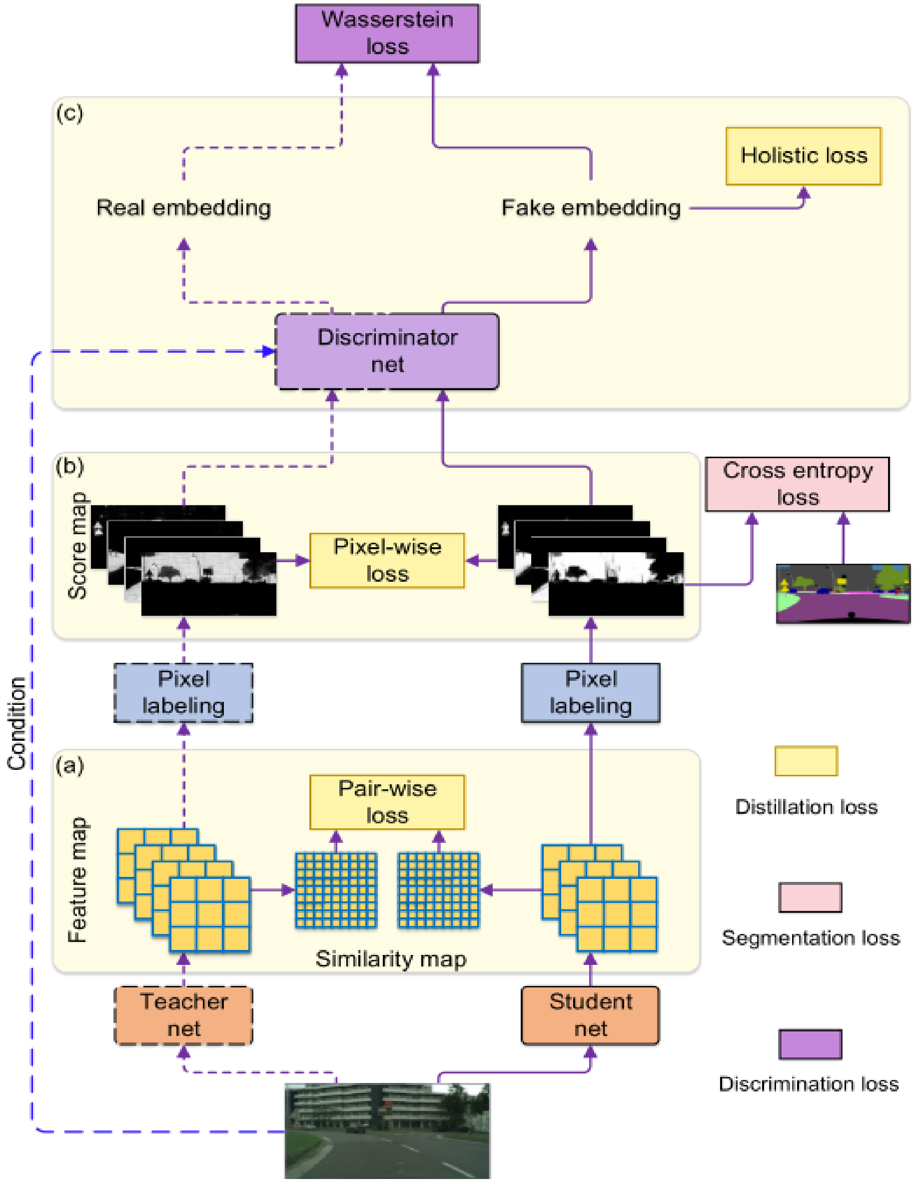
Loss:

- Pixel-wise loss
- Pair-wise loss
- Discrimination loss

knowledge: Similarity map

hint

Compare

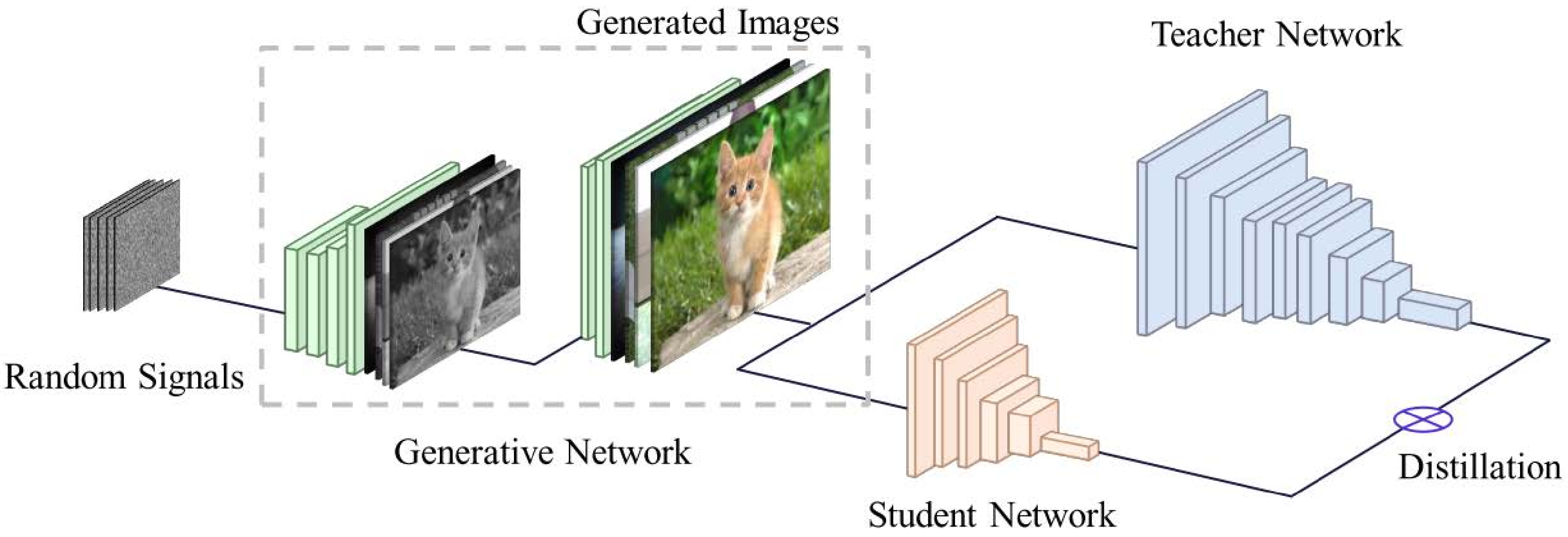


Liu, Yifan et al. "Structured Knowledge Distillation for Semantic Segmentation." CVPR 2019

KD and GAN

- GAN generates data for KD
- GAN for KD to adversarial
- KD for GAN to compress G

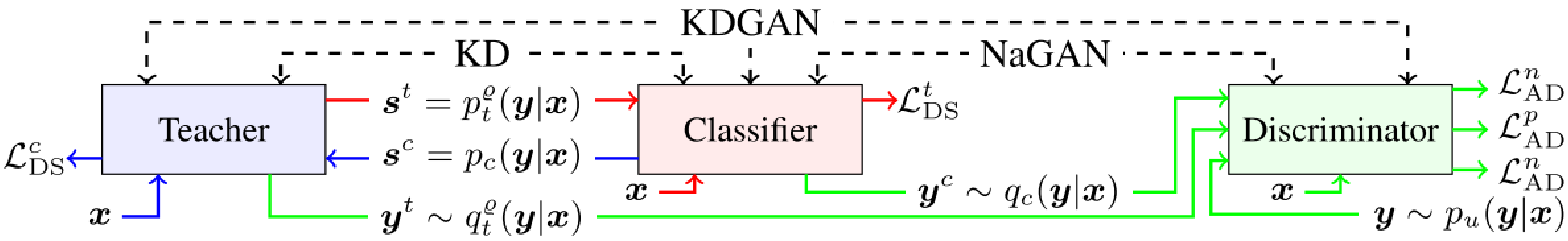
DAFL: Data-Free Learning of Student Networks



- 3 losses for GAN: one-hot, activation, info-entropy
- original KD, no hint
- Teacher as Discriminator
- *but a little poor*

Chen, Hanting et al. "DAFL: Data-Free Learning of Student Networks." ICCV 2019

KDGAN

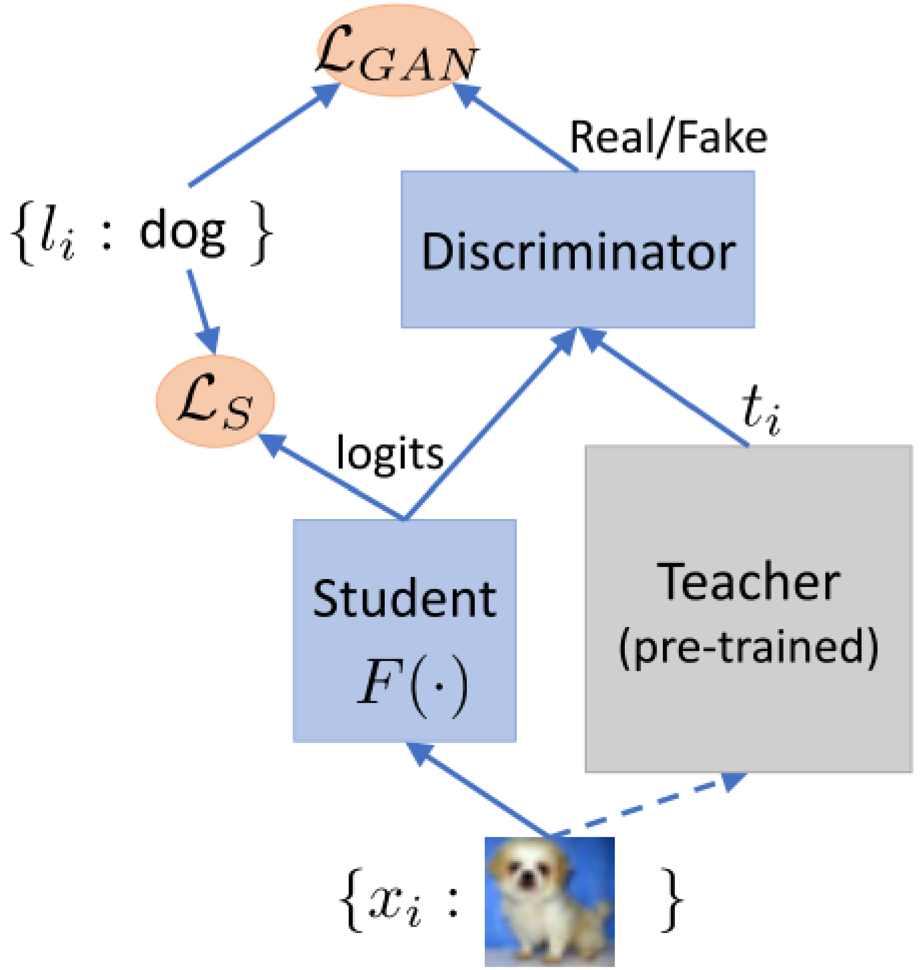


- Mainly to compress the Generator(Classifier)
- Simultaneously train teacher and classifier
- Through logits, no hint

Wang, Xiaojie. “KDGAN:Knowledge Distillation with Generative Adversarial Networks.” NIPS 2018

KD with cGAN

- Add a Discriminator
- through logits



Xu, Zheng et al. "Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks." arXiv:1709.00513.

Beyond KD

- label smoothing principle
- challenge
- news
- idea

When Does Label Smoothing Help?

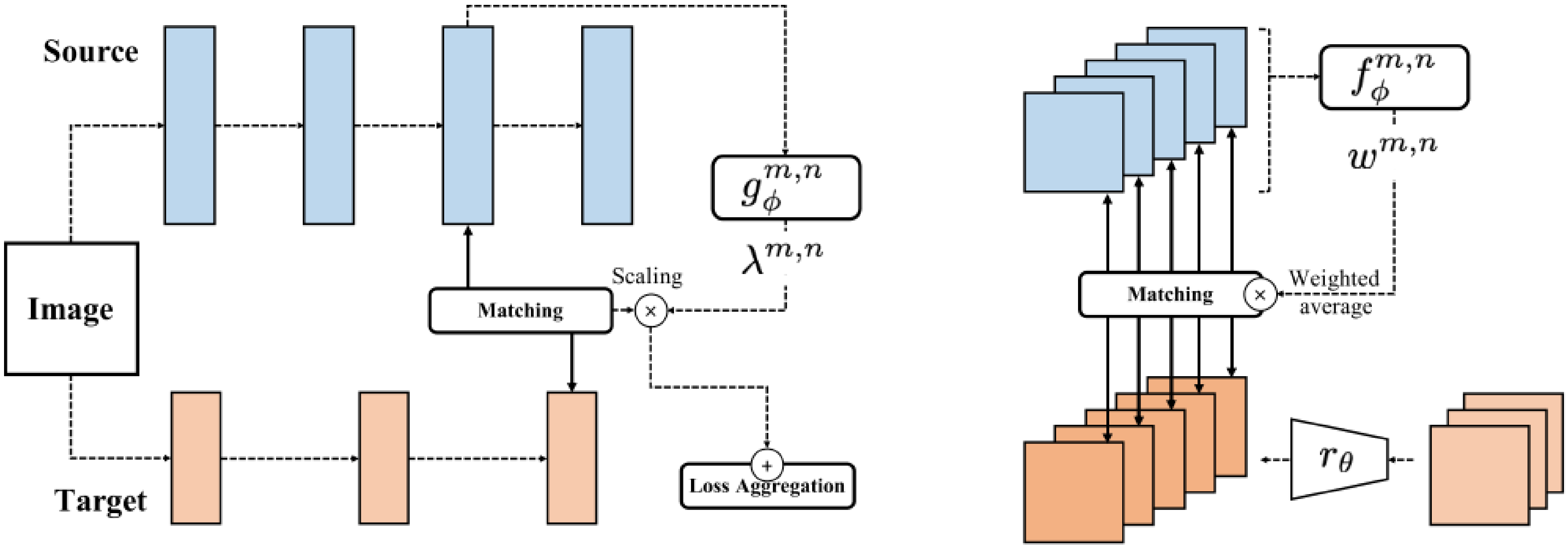
- poorly understood
- prevent from over-confident
- teacher is trained with label smoothing， wrong
- label smoothing can hurt distillation
- reduces mutual information

Müller, Rafael, Kornblith, and Hinton. "When Does Label Smoothing Help?" NIPS 2019

Challenge

1. combination strategy of Multi-teacher
   - logits: our adaptive, average, random, max entropy
   - intermediate info: our group hint(multi-level), triplet selection(You et al. 2017), self-KD(Hou et al. 2019, Zhang et al. 2019), other?
2. usable knowledge for GAN or vision task
   - feature maps
   - relational info，e.g. similarity
   - no labels or logits

new: Learning What and Where to Transfer



(a) Where to transfer (b) What to transfer

- based on Meta-learning
- What: which features and how much knowledge from each feature
- Where: which pairs of layers should be matched for knowledge transfer

Jang, Yunhun et al. "Learning What and Where to Transfer." ICML 2019

Idea for KD

1. Meta-KD

   prediction、layer、weight、gradient、attention map

2. KD-GAN for Image synthesis task
3. finding new knowledge form
4. KD for segment, detection, Identification, NLP, etc.

References

1. Buciluǎ, Cristian et al. "Model compression." *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006.
2. Ba, Jimmy, and Rich Caruana. "Do deep nets really need to be deep?." NIPS 2014.
3. Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531* (2015).
4. Romero, Adriana et al. "Fitnets: Hints for thin deep nets." *arXiv preprint arXiv:1412.6550* (2014).
5. Zagoruyko, Sergey, and Nikos Komodakis. "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer." *arXiv preprint arXiv:1612.03928* (2016).
6. Vapnik, Vladimir, and Rauf Izmailov. "Learning using privileged information: similarity control and knowledge transfer." *Journal of machine learning research* 16.2023-2049 (2015): 2.
7. Lopez-Paz, David et al. "Unifying distillation and privileged information." *arXiv preprint arXiv:1511.03643* (2015).
8. Wang, Xiaojie et al. "KDGAN: knowledge distillation with generative adversarial networks." NIPS 2018.
9. Wu, Ancong et al. "Distilled Person Re-identification: Towars a More Scalable System. " CVPR 2019.
10. Bhardwaj, Shweta et al. "Efficient Video Classification Using Fewer Frames." CVPR 2019.
11. Zhang, Feng et al. "Fast Human Pose Estimation." CVPR2019.
12. Saputra et al. "Distilling knowledge from a deep pose regressor network." *arXiv preprint arXiv:1908.00858* (2019).

References

13. Xu, Zheng et al. "Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks." *arXiv preprint arXiv:1709.00513* (2017).
14. Sun, Siqi et al. "Patient Knowledge Distillation for BERT Model Compression." *arXiv preprint arXiv:1908.09355.*
15. Jiao, Xiaoqi et al. "TinyBERT: Distilling BERT for Natural Language Understanding." *arXiv preprint arXiv: 1909.10351.*
16. Hou, Yuenan et al. "Learning Lightweight Lane Detection CNNs by Self Attention Distillation." ICCV 2019.
17. Liu, Yifan et al. "Structured Knowledge Distillation for Semantic Segmentation." CVPR 2019.
18. Yim, Junho et al. "A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning." CVPR 2017.
19. Yu, Lu et al. "Learning Metrics from Teachers:Compact Networks for Image Embedding." CVPR 2019.
20. Park, Wonpyo et al. "Relational Knowledge Distillation." CVPR 2019.
21. Liu, Yufan et al. "Knowledge Distillation via Instance Relationship Graph." CVPR 2019.
22. Jin, Xiao et al. "Knowledge Distillation via Route Constrained Optimization." ICCV 2019.
23. Tung, Frederick, and Mori Greg. "Similarity-Preserving Knowledge Distillation." ICCV 2019.
24. Chen, Hanting et al. "DAFL: Data-Free Learning of Student Networks." ICCV 2019.
25. Wang, Xiaojie. "KDGAN:Knowledge Distillation with Generative Adversarial Networks." NIPS 2018.

References

26. Sau, Bharat Bhusan et al. "Deep Model Compression: Distilling Knowledge from Noisy Teachers." *arXiv preprint arXiv:1610.09650v2*
27. You, Shan et al. "Learning from Multiple Teacher Networks." KDD 2017.
28. Zhang, Ying et al. "Deep Mutual Learning." CVPR 2018.
29. Zhang, Linfeng et al. "Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation." ICCV 2019.
30. Müller, Rafael, Kornblith, and Hinton. "When Does Label Smoothing Help?" NIPS 2019.
31. Cheng, Yu et al. "A Survey of Model Compression and Acceleration for Deep Neural Networks." IEEE 2017.
32. Jang, Yunhun et al. "Learning What and Where to Transfer." ICML 2019.
33. **Liu, Yuang**, Zhang Wei, and Wang jun. "Adaptive Multi-Teacher Multi-Level Knowledge Distillation." 2019.09(ours)

华东师范大学

Thanks for your listening!

Yuang Liu
AIDA, ECNU
frankliu624@gmail.com