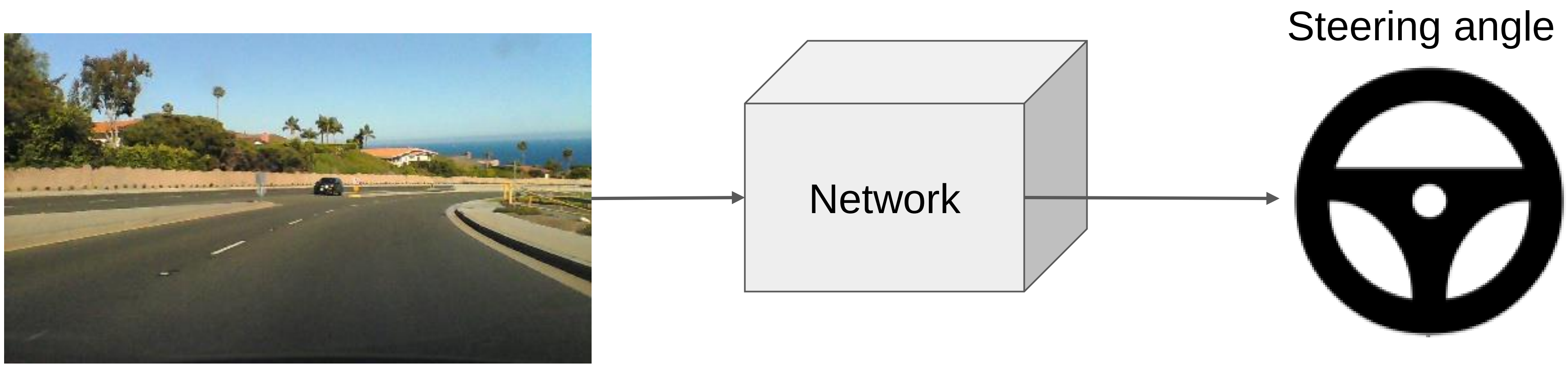
Small-shot Multi-modal Distillation for Vision-based Autonomous Steering

Yu Shen, Luyu Yang, Xijun Wang,
Ming C. Lin

Paper ID:1818

Target Task

- Learn to steer in end-to-end autonomous driving
- Perception and control



Motivation

Multi-modal distillation in autonomous driving require ***paired*** data with different modalities.

However, sometimes we only have a few auxiliary modality data

- expensive expert-labeled data
- sensing data from a low-frequency sensor
- online inferred data with high computational complexity

How to solve such a small-shot multi-modal distillation problem?

Contributions

A novel framework to distill knowledge from multi-modality model to single-modality model

- small-shot auxiliary modality distillation network (*AMD-S-Net*)

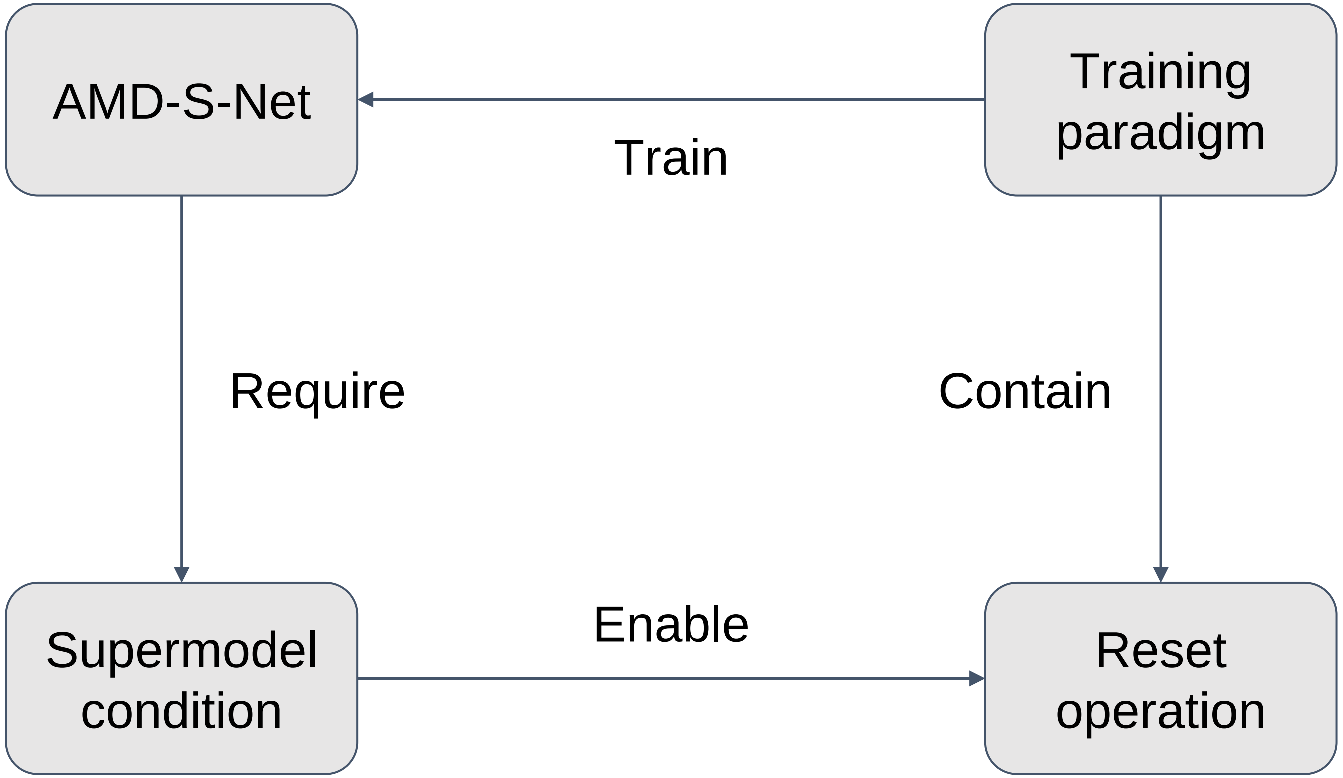
Which is trained with our training paradigm and must satisfy a specific *supermodel condition*.

AMD-S-Net also contains a specific framework design to fully distill the information

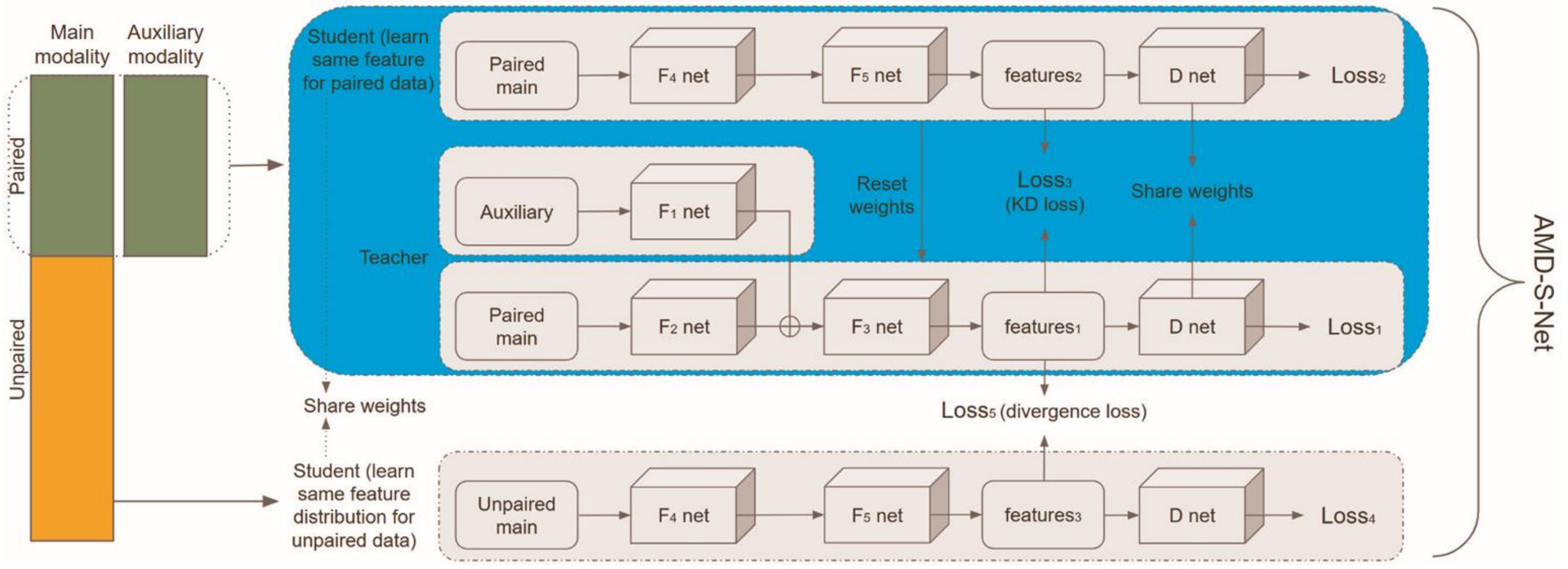
- consistency supervision for the pairwise data
- distribution divergence supervision for the unpaired data.

A novel knowledge distillation *training paradigm* that enables teachers to explore and provide student's local loss landscape information in a higher dimension to students, boosting performance.

Relation of Key Concepts

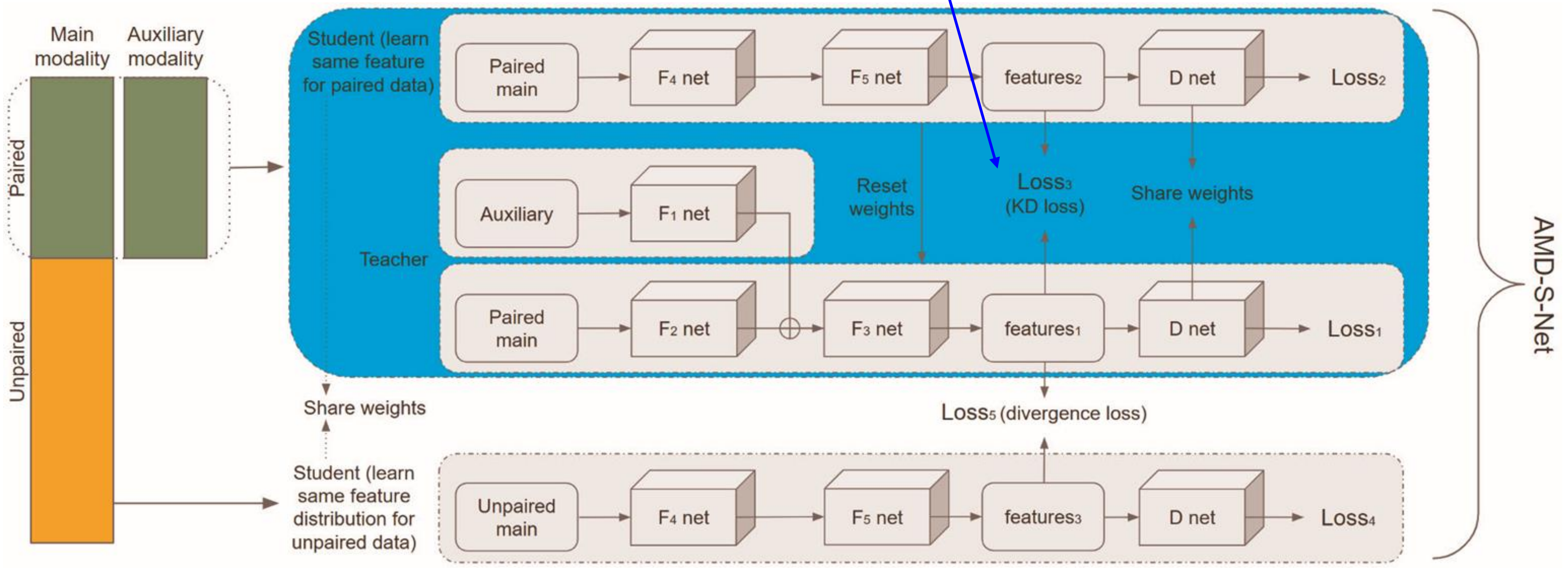


Framework (AMD-S-Net)

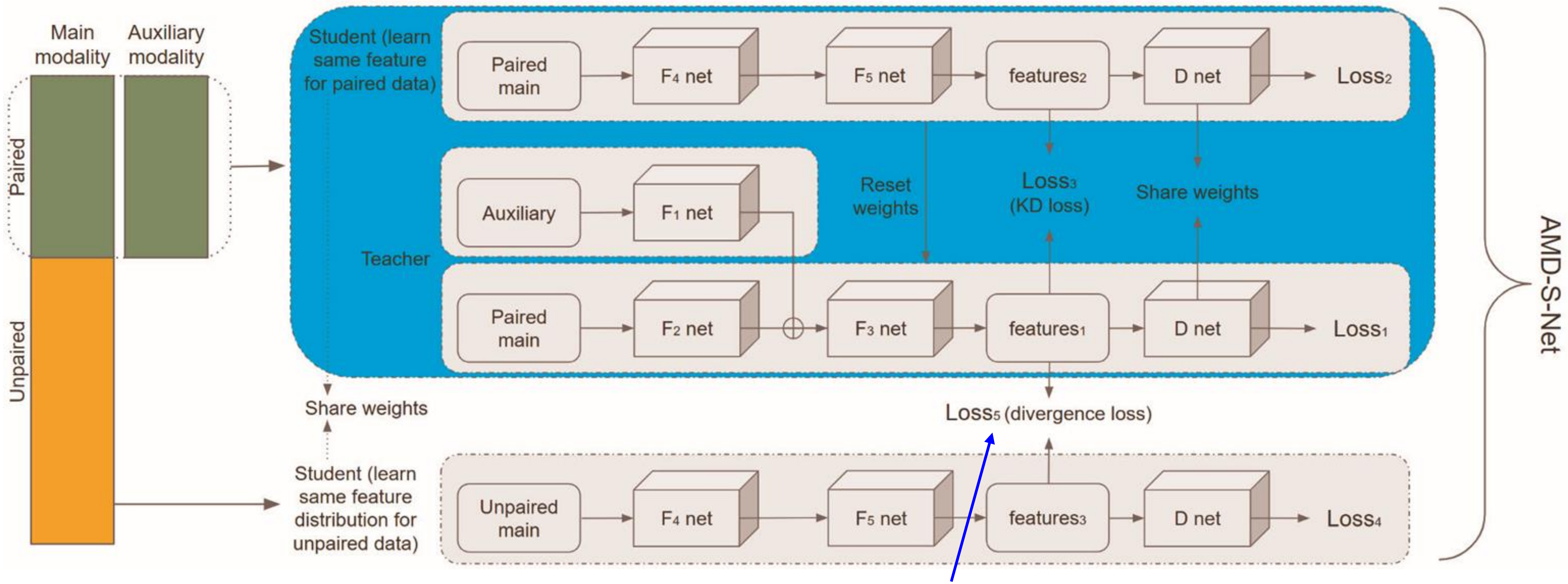


Framework (AMD-S-Net)

consistency supervision for the pairwise data



Framework (AMD-S-Net)

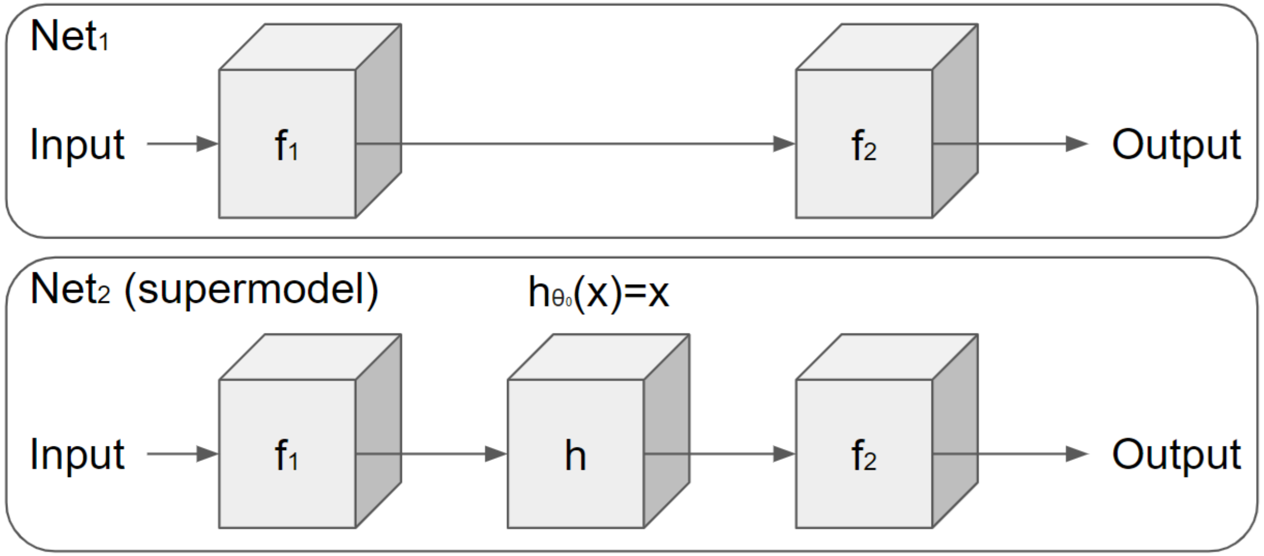


distribution divergence supervision for the unpaired data

Supermodel Condition

Definition 1. *Given a model $M^{(A)}_{\theta_A}(I_A)$ (weights θ_A and input I_A), and a model $M^{(B)}_{\theta_B}(I_B)$ (weights θ_B and input I_B), if for any θ_A, there is a θ_B, such that $M^{(A)}_{\theta_A}(I_A) = M^{(B)}_{\theta_B}(I_B)$ for any arbitrary valid input data I_A and its superset I_B. We call model M_B as a "supermodel" of M_A.*

Example:



Reset Operation

Definition 2. *Given a model $M^{(A)}_{\theta_A}(I_A)$ (weights θ_A and input I_A), and its supermodel $M^{(B)}_{\theta_B}(I_B)$ (weights θ_B and input I_B), we define "reset B with A" to be the process of constructing a new θ_B that meet $M^{(A)}_{\theta_A}(I_A) = M^{(B)}_{\theta_B}(I_B)$ for given θ_A and any arbitrary valid input data I_A and its superset I_B.*

Example:

Suppose B is a supermodel of A (e.g.. B=A+A'). reset B with A is constructing such

$$\theta_B = [\theta_A, 0]$$

, where θ_A is the weights of A and 0 is the weights of A'.

Without VS With Our Training Paradigm

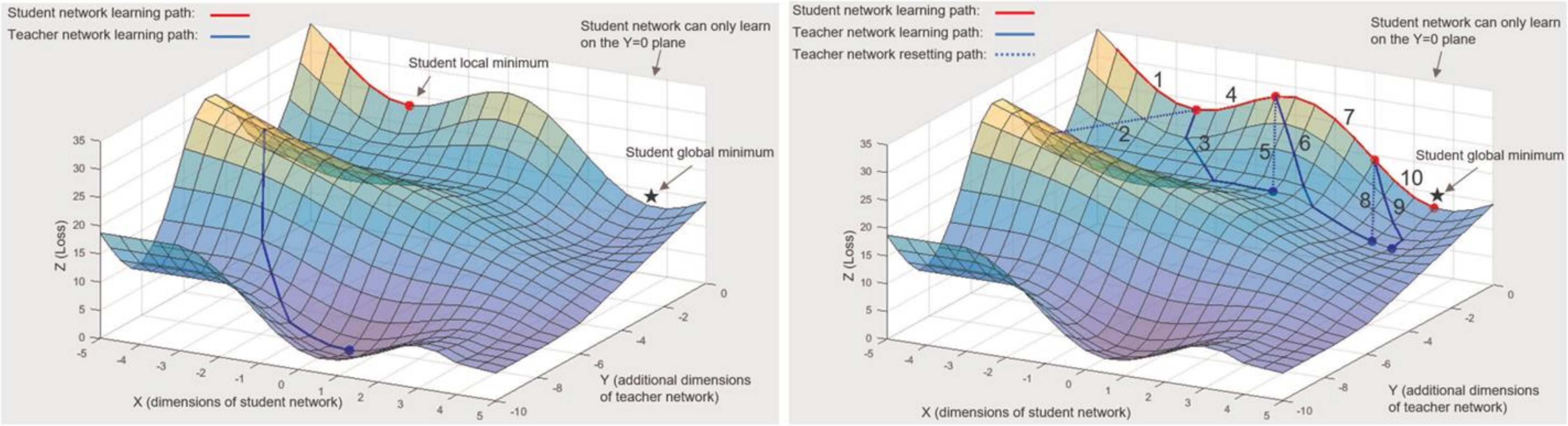
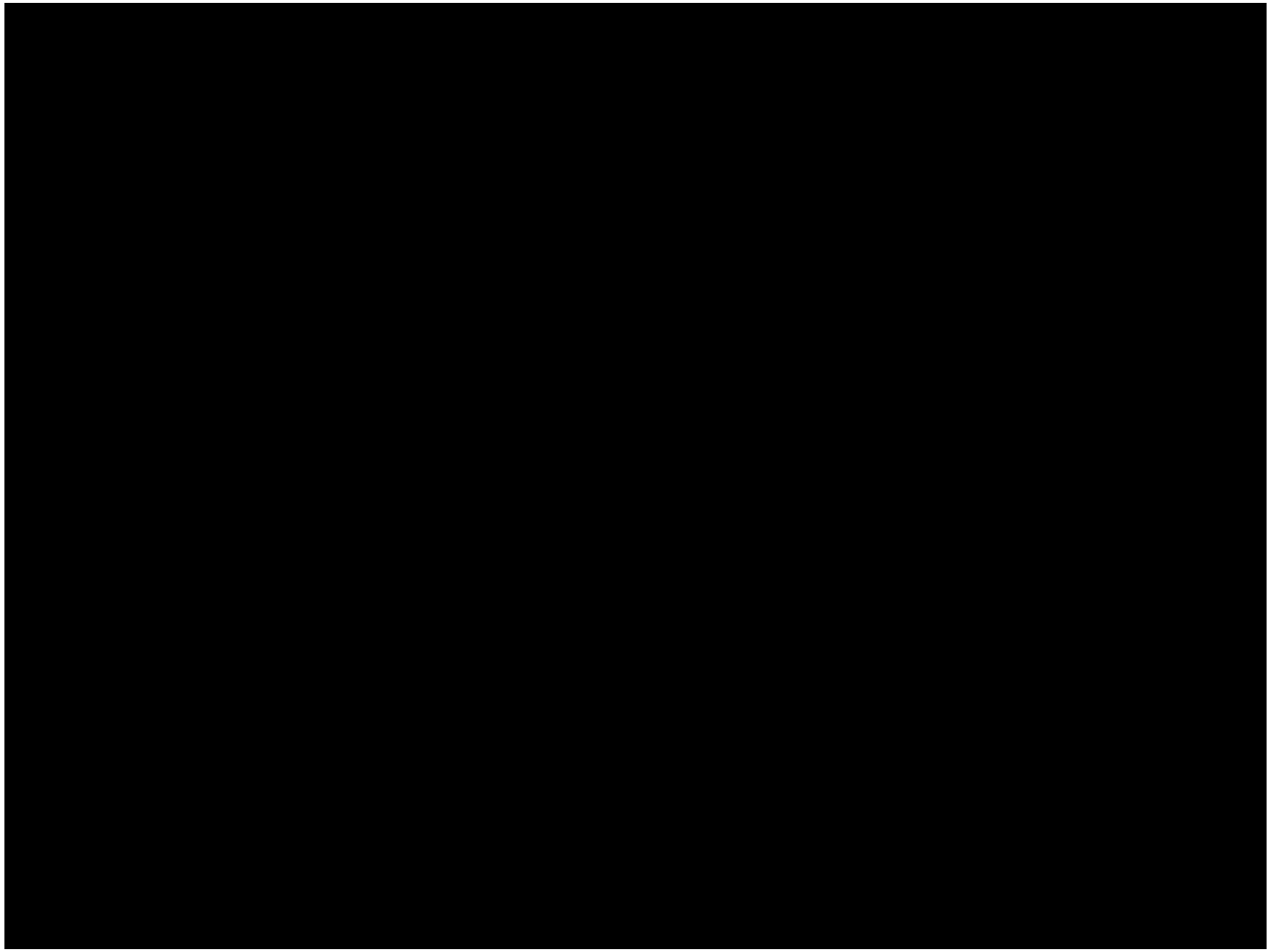


Fig. 2. **Training Path Comparison on Loss Landscape.** Given the teacher network is a *supermodel* of the student network, the student parameter space (along X axis with Y=0) is a subspace of the teacher parameter space (XY plane). LEFT: Without our training paradigm, the teacher is not aware of the student states, the training path and the final state of the teacher can be far away from the student space, i.e. the landscape may be totally different, thus providing limited guidance and lead to the student getting stuck in a local minimum. RIGHT: In our method, the teacher is reset to the student states at the beginning of each round, and does optimization with additional dimensions but within a certain range of the student space, teaching the student with local landscape information and potential direction to a better solution. The number 1∼10 is the step order of these processes, see details in Sec. III-C.

Our Training Paradigm



Comparison (AMD-S-Net)

	Accuracy (%) on different angle threshold τ (degree)						
Method	$\tau = 1.5$	$\tau = 3.0$	$\tau = 7.5$	$\tau = 15$	$\tau = 30$	$\tau = 75$	**Mean**
Oracle (100% auxiliary modality data)	42.7	68.0	88.0	94.4	96.6	98.6	81.4
one stream (RGB only)	27.3	49.0	77.4	90.2	95.4	98.1	72.9
two streams (shared regressor)	25.9	47.2	77.7	88.4	93.6	97.8	71.8
Modified Xiao et al. [1]	40.8	64.1	84.7	92.7	95.8	98.2	79.4
Modified DMCL [26]	39.1	67.5	88.3	93.9	96.7	98.2	80.6
Ours (AMD-S-Net)	**52.6**	**72.7**	**91.3**	**95.0**	**97.0**	**98.3**	**84.5**

TABLE I

Performance comparison for AMD-S-Net under the small amount of auxiliary modality data setting (20%). OUR METHOD OUTPERFORMS OTHER METHODS BY UP TO **12.7%** MEAN ACCURACY IMPROVEMENT.

Comparison (Our Training Paradigm)

Method	Accuracy on different threshold τ (%)							
	$\tau = 1.5$	$\tau = 3.0$	$\tau = 7.5$	$\tau = 15$	$\tau = 30$	$\tau = 75$	**Mean**	**Improvement**
train vanilla								
Teacher (img+seg)	40.8	64.1	84.7	92.7	95.8	98.2	79.4	
Student (img)	27.3	49.0	77.4	90.2	95.4	98.1	72.9	
existing distillation methods								
kd [4]	23.4	41.2	68.9	83.7	92.1	97.2	67.7	
hint [11]	28.3	47.6	77.8	89.2	95.0	98.4	72.7	
similarity [13]	20.6	38.9	66.7	81.5	92.6	98.0	66.4	
correlation [15]	21.7	39.5	70.0	86.8	94.6	98.2	68.5	
rkd [16]	26.2	46.5	74.8	87.9	94.1	97.8	71.2	
pkt [9]	30.3	51.0	78.2	88.4	94.4	98.2	73.4	
abound [10]	24.8	45.2	74.9	87.3	93.7	97.7	70.6	
factor [8]	26.8	47.8	76.9	88.8	94.7	98.0	72.2	
fsp [6]	27.1	47.7	74.4	87.9	94.4	97.8	71.6	
attention [7]	27.1	47.0	73.1	84.9	92.8	98.3	70.5	
existing distillation methods with our training paradigm								
kd [4]	30.4	53.7	78.5	88.3	94.8	97.8	73.9	**6.2**
hint [11]	52.7	71.2	88.8	93.6	95.5	97.1	83.1	**10.4**
similarity [13]	52.6	72.7	91.3	95.0	97.0	98.3	84.5	**18.1**
correlation [15]	21.7	39.7	71.2	87.0	94.4	98.2	68.7	**0.2**
rkd [16]	32.4	53.8	79.5	89.3	94.7	97.9	74.6	**3.4**
pkt [9]	54.2	72.5	90.0	94.8	96.7	98.3	84.4	**11**
abound [10]	24.9	45.3	75.1	87.1	93.5	97.7	70.6	0
factor [8]	54.3	72.3	90.1	94.8	96.7	98.3	84.4	**12.2**
fsp [6]	27.5	48.4	75.0	87.5	94.3	97.8	71.8	**0.2**
attention [7]	46.2	68.1	86.8	93.4	96.6	98.2	81.5	**11**

TABLE X

Comparison with knowledge distillation methods on Audi dataset (100% RGB image + 20% segmentation) with Nvidia PilotNet [30]. First section in the table shows the performance of teacher and student network trained directly. Second section shows the performance of student with different knowledge distillation methods (train student from start, using the pretrained teacher model in the previous section). Third section shows the performance of student after using our technique based on other methods (take the teacher and student network in the second section of this table as init model, and retrain the model with our method). By comparing between the second and third section, we can see our method increase the performance of most existing methods with up to 18.1%.

Comparison (Other Tasks)

Model	DS↑	RC↑	IP↓	CP↓	CV↓	CL↓	RLI↓	SSI↓
RGB	21.0	60.5	0.49	**0.01**	0.15	0.08	**0.14**	0.04
RGB+PC	11.2	52.9	**0.37**	0.02	0.22	0.01	0.38	**0.02**
Ours	**22.0**	**63.1**	0.45	0.02	**0.05**	**0.00**	0.20	0.03

TABLE III

PERFORMANCE COMPARISON ON LONG ROUTES WAY POINTS PREDICTION BETWEEN BASE (100% RGB), MULTI-MODALITY (28% RGB + 28% POINT CLOUD), AND OUR METHOD (100% RGB + 28% POINT CLOUD). DS: AVG. DRIVING SCORE, RC: AVG. ROUTE COMPLETION, IP: AVG. INFRACTION PENALTY, CP: COLLISIONS WITH PEDESTRIANS, CV: COLLISIONS WITH VEHICLES, CL: COLLISIONS WITH LAYOUT, RLI: RED LIGHTS INFRACTIONS, SSI: STOP SIGN INFRACTIONS.

ICRA
LONDON · 2023

Comparison (Other Tasks)

	Accuracy (%) on different modalities (ID:1∼6)						
Method	1	2	3	4	5	6	mean
Other KD	84.92	**62.98**	68.75	61.10	70.35	43.17	65.2
Ours	**87.42**	62.29	**70.86**	**66.34**	**71.97**	**49.49**	**68.1**

TABLE IV

Performance comparison on handwritten classification task. OUR METHOD OUTPERFORMS OTHER KD METHODS WITH 2.9% ON AVERAGE.

Comparison (Different Backbones)

		Accuracy (%) on various angle threshold τ (degree)				
Backbone	Method	$\tau = 1.5$	$\tau = 3.0$	$\tau = 7.5$	$\tau = 15$	**mAcc**
PilotNet	SIM	20.6	38.9	66.7	81.5	66.4
PilotNet	SIM+ours	**52.6**	**72.7**	**91.3**	**95.0**	**84.5**
ResNet34	SIM	30.1	54.4	85.5	**94.1**	76.6
ResNet34	SIM+ours	**37.2**	**60.2**	**85.7**	93.3	**78.6**
ShuffleV2	SIM	39.9	61.3	81.4	89.8	77.7
ShuffleV2	SIM+ours	**47.0**	**71.2**	**90.1**	**94.9**	**83.0**
MobileNetV2	SIM	31.1	51.4	78.2	89.4	73.9
MobileNetV2	SIM+ours	**52.9**	**71.8**	**89.7**	**94.6**	**84.0**
WRN	SIM	22.8	42.9	76.9	92.2	71.7
WRN	SIM+ours	**37.7**	**64.7**	**89.8**	**94.6**	**80.3**

TABLE VI

PERFORMANCE COMPARISON ON DIFFERENT BACKBONES. OUR METHOD OUTPERFORMS SIM [13] ON PILOTNET [30], RESNET34 [43], SHUFFLEV2 [44], MOBILENETV2 [45], AND WRN [46] WITH UP TO 18.1% ACCURACY IMPROVEMENT.

Comparison (Robustness)

	Clean	Blur				Noise		
	Clean	Defocus	Glass	Motion	Zoom	Gauss	Shot	Impulse
RGB only	73.1	72.7	71.8	69.8	72.3	67.9	66.9	67.0
$20\%\mathcal{I}^A$	74.8	74.3	73.1	73.2	74.2	69.2	68.3	68.6
$100\%\mathcal{I}^A$	77.1	75.5	75.2	73.1	76.3	71.4	70.1	70.3

	Clean	Weather				Digital			**mAcc**
	Clean	Snow	Frost	Fog	Bright	Contrast	Pixel	JPEG	**mAcc**
RGB only	73.1	62.8	56.5	54.2	64.2	39.9	73.3	70.7	**65**
$20\%\mathcal{I}^A$	74.8	68.1	65.4	63.8	67.6	65.4	74.8	71.8	**69.8**
$100\%\mathcal{I}^A$	77.1	63.8	58.7	56.4	65.8	62.0	77.2	75.3	**69.4**

TABLE VIII

AVERAGE ACCURACY(%) OF OUR METHOD ON CLEAN AND PERTURBED DATA (GENERATED WITH IMAGENET-C EFFECTS [48]). THE LAST COLUMN "MEAN" IS THE MEAN ACCURACY ON ALL PERTURBED DATA (BLUR, NOISE, WEATHER AND DIGITAL). WE SHOW THAT BOTH BASIC AND SMALL-SHOT AUXILIARY MODALITY LEARNING CAN GET HIGHER ACCURACY THAN THE BASE METHOD (ABOUT 4.7% IN AVERAGE), I.E., HIGHER ROBUSTNESS.

Conclusion

A novel framework to distill knowledge from multi- to single- modality model

small-shot auxiliary modality distillation network (*AMD-S-Net*)

- Among the first that only use a small amount of auxiliary modality data for training
- A specific architecture design to fully distill the information
  - consistency supervision for the pairwise data
  - distribution divergence supervision for the unpaired data.

Performance improvement

- Up to 12% compared to other AML methods
- Up to 18% compared to other knowledge distillation methods

Thank you!