

# Revisiting Label Smoothing & Knowledge Distillation Compatibility: What was Missing?

*International Conference on Machine Learning (ICML) 2022*

**Keshigeyan Chandrasegaran**



**Ngoc-Trung Tran \***



**Yunqing Zhao \***



**Ngai-Man Cheung**

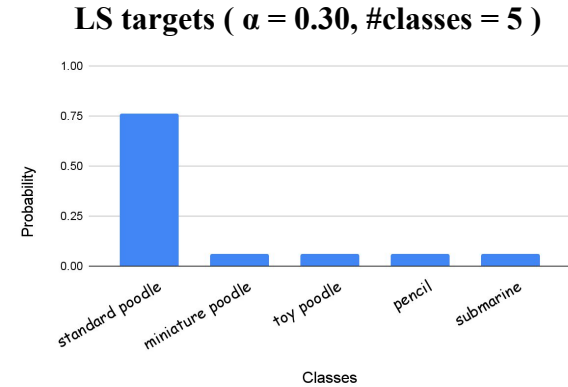
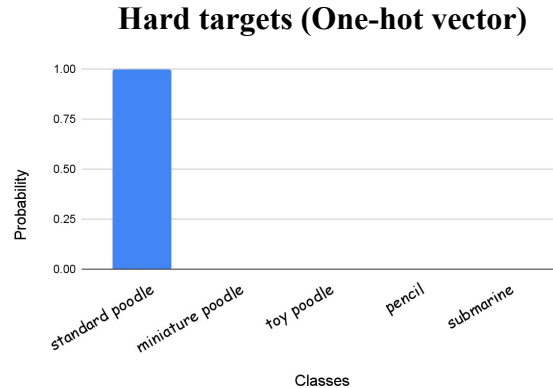


# Label Smoothing (LS)

Label Smoothing (LS) (Szegedy et al., 2016) was originally formulated as a regularization strategy to alleviate models' overconfidence.



Standard poodle



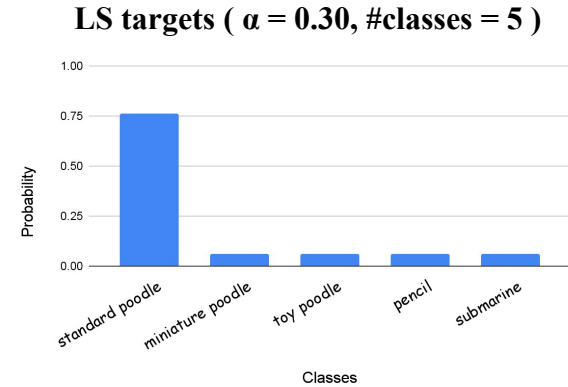
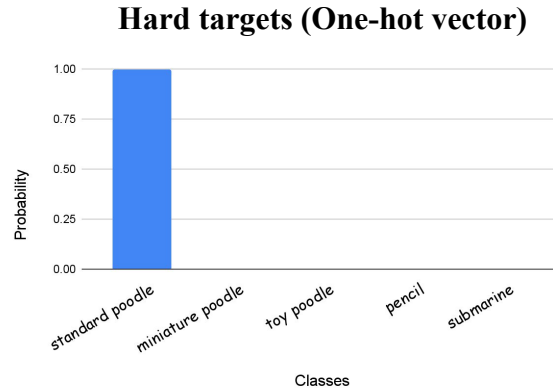
LS replaces original hard target distribution by a **mixture of original hard target distribution and the uniform distribution** (characterized by a mixture parameter  $\alpha$ ).

# Label Smoothing (LS)

Label Smoothing (LS) (Szegedy et al., 2016) was originally formulated as a regularization strategy to alleviate models' overconfidence.



Standard poodle



Practically used in many tasks including **image classification** (He et al., 2019), **NLP** (Vaswani et al., 2017) and **speech recognition** (Chiu et al., 2018).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. CVPR

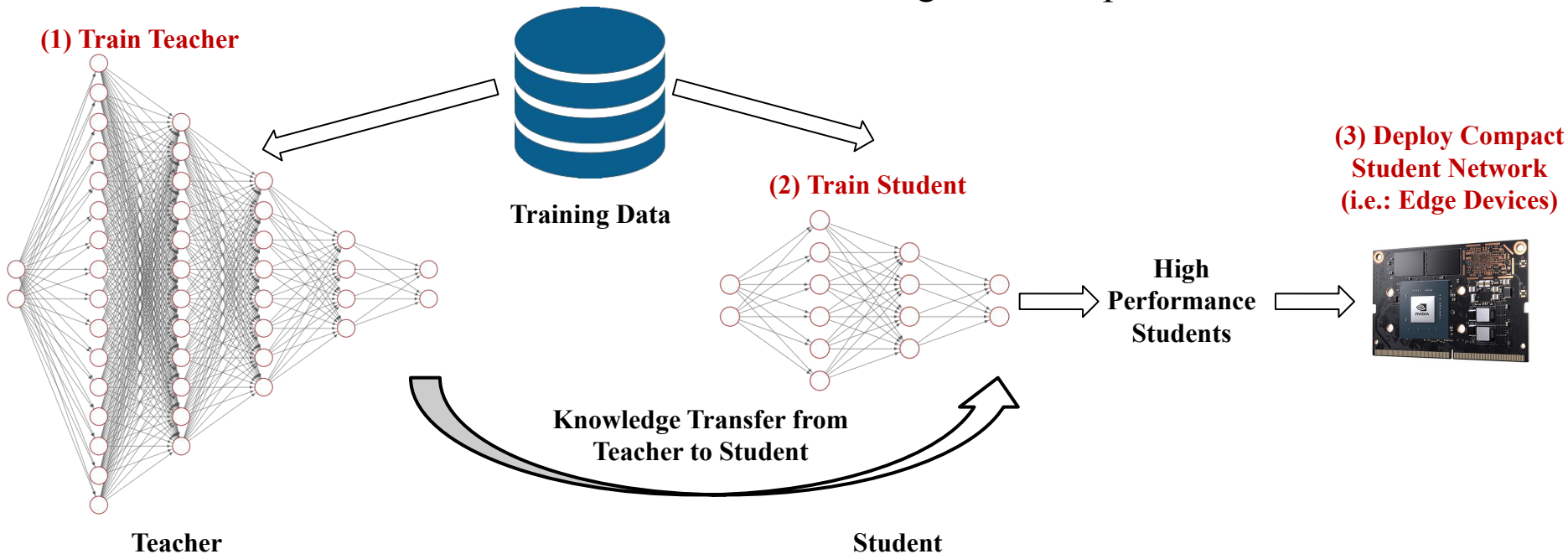
He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., & Li, M. (2019). Bag of tricks for image classification with convolutional neural networks. CVPR

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Chiu, C. C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., ... & Bacchiani, M. (2018, April). State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE ICASSP*

# Knowledge Distillation (KD)

Knowledge Distillation (KD) (Hinton et al., 2015) uses a larger capacity teacher model/ ensemble of teacher models to transfer knowledge to a compact student model.

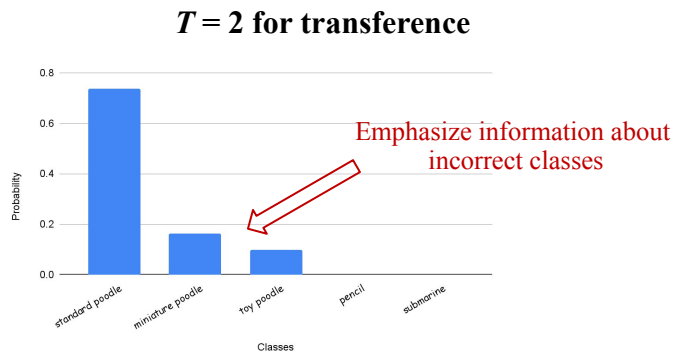
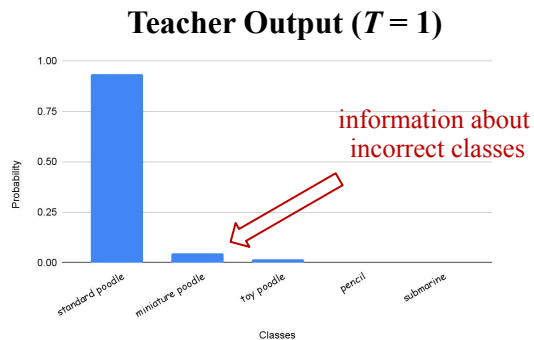


Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. In NIPS Deep Learning and Representation Learning Workshop, 2015.

# Knowledge Distillation (KD)



Standard poodle



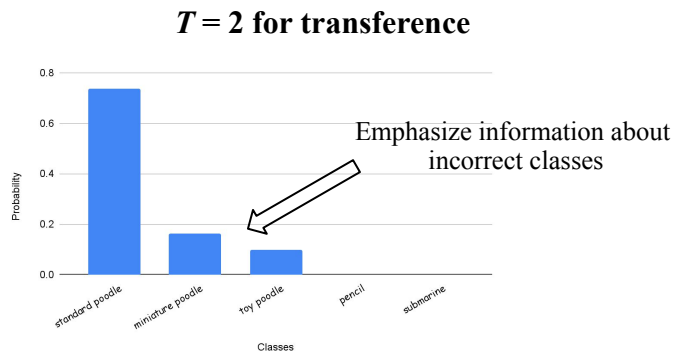
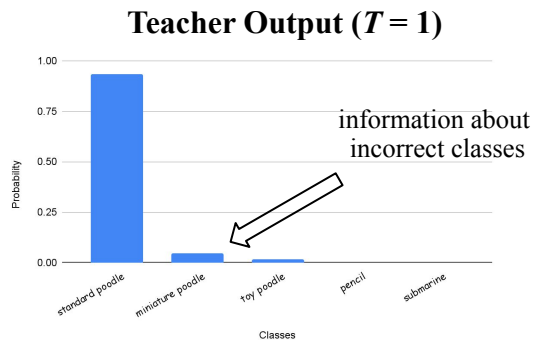
Increased  $T$  for better transference in many problems. Empirically identified.

In KD, a temperature  $T$  is used to facilitate the **transference**: an **increased  $T$**  may produce more suitable soft targets that have more **emphasis on the probabilities of incorrect classes**

# Knowledge Distillation (KD)



Standard poodle



Increased  $T$  for better transference in many problems. Empirically identified.

KD methods have been widely used in **visual recognition** (Peng et al., 2019), **NLP** (Hu et al., 2018), **speech recognition** (Perez et al., 2020), **self-supervised learning** (Fang et al., 2020) and **neural architecture search** (Wang et al., 2021).

Z. Peng, Z. Li, J. Zhang, Y. Li, G. -J. Qi and J. Tang, "Few-Shot Image Recognition With Knowledge Transfer," ICCV 2019

Hu, M., Peng, Y., Wei, F., Huang, Z., Li, D., Yang, N., & Zhou, M. (2018, January). Attention-Guided Answer Distillation for Machine Reading Comprehension. In *EMNLP*.

Perez, A., Sanguineti, V., Morerio, P., & Murino, V. (2020). Audio-visual model distillation using acoustic images. *WACV*

Fang, Z., Wang, J., Wang, L., Zhang, L., Yang, Y., & Liu, Z. (2020, September). SEED: Self-supervised Distillation For Visual Representation. In *ICLR*.

Wang, D., Gong, C., Li, M., Liu, Q., & Chandra, V. (2021, July). AlphaNet: improved training of supernet with alpha-divergence. In *ICML*. PMLR.

# LS and KD Compatibility

**Does LS in a teacher network suppress the effectiveness of KD?**

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help?. *Advances in neural information processing systems*, 32.

Shen, Z., Liu, Z., Xu, D., Chen, Z., Cheng, K. T., & Savvides, M. (2021). Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study. In *ICLR*

# LS and KD Compatibility

## Does LS in a teacher network suppress the effectiveness of KD?

“If a teacher network is trained with label smoothing, knowledge distillation into a student network is much less effective.”

“Label smoothing can hurt distillation”

[ Müller et al., 2019 ]

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help?. *Advances in neural information processing systems*, 32.

Shen, Z., Liu, Z., Xu, D., Chen, Z., Cheng, K. T., & Savvides, M. (2021). Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study. In *ICLR*



# LS and KD Compatibility

## Does LS in a teacher network suppress the effectiveness of KD?

“If a teacher network is trained with label smoothing, knowledge distillation into a student network is much less effective.”

“Label smoothing can hurt distillation”

[ Müller et al., 2019 ]

“Label smoothing will not impair the predictive performance of students.”

“Label smoothing is compatible with knowledge distillation”

[ Shen et al., 2021 ]

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help?. *Advances in neural information processing systems*, 32.

Shen, Z., Liu, Z., Xu, D., Chen, Z., Cheng, K. T., & Savvides, M. (2021). Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study. In *ICLR*

# LS and KD Compatibility : Research Gap

	Information Erasure (Incompatibility)	Distance enlargement (compatibility)	Conclusion
Müller et al. 2019	LS erases relative information in the logits		LS-trained teacher can hurt KD
Shen et al. 2021	With LS, some relative information in the logits is still retained	LS enlarges the distance between semantically similar classes	Benefits outweigh disadvantages. LS is compatible with KD.

Studied in isolation, both these contradictory arguments are convincing and supported empirically, although the later does not address the contradictory findings / results of Müller et al. (2019)

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help?. *Advances in neural information processing systems*, 32.

Shen, Z., Liu, Z., Xu, D., Chen, Z., Cheng, K. T., & Savvides, M. (2021). Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study. In *ICLR*

# LS and KD Compatibility : Research Gap

	Information Erasure (Incompatibility)	Distance enlargement (compatibility)	Conclusion
Müller et al. 2019	LS erases relative information in the logits		LS-trained teacher can hurt KD
Shen et al. 2021	With LS, some relative	LS enlarges the distance	Benefits outweigh

**Should you smooth a teacher network?  
THIS REMAINS UNCLEAR!**

studied in isolation, both these contradictory arguments are convincing and supported empirically, although the later does not address the contradictory findings / results of Müller et al. (2019)

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help?. *Advances in neural information processing systems*, 32.

Shen, Z., Liu, Z., Xu, D., Chen, Z., Cheng, K. T., & Savvides, M. (2021). Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study. In *ICLR*

# Revisiting LS and KD Compatibility: Our Contributions

**Does LS in a teacher network suppress the effectiveness of KD?**

# Revisiting LS and KD Compatibility: Our Contributions

## Does LS in a teacher network suppress the effectiveness of KD?

- Our contributions are the **discovery, analysis and validation of systematic diffusion** as the missing concept which is instrumental in **resolving these contradictory findings**.

# Revisiting LS and KD Compatibility: Our Contributions

## Does LS in a teacher network suppress the effectiveness of KD?

- Our contributions are the **discovery, analysis and validation of systematic diffusion** as the missing concept which is instrumental in **resolving these contradictory findings**.
- We conduct large-scale experiments including **image classification, neural machine translation** and **compact student distillation** tasks spanning across multiple datasets and teacher-student architectures to **qualitatively / quantitatively show Systematic Diffusion**.

# Revisiting LS and KD Compatibility: Our Contributions

## Does LS in a teacher network suppress the effectiveness of KD?

- Our contributions are the **discovery, analysis and validation of systematic diffusion** as the missing concept which is instrumental in **resolving these contradictory findings**.
- We conduct large-scale experiments including **image classification, neural machine translation** and **compact student distillation** tasks spanning across multiple datasets and teacher-student architectures to **qualitatively / quantitatively show Systematic Diffusion**.
- **As a rule of thumb**, we suggest practitioners to **use an LS-trained teacher with a low-temperature transfer** (i.e.,  $T = 1$ ) to render high performance students.

# Revisiting LS and KD Compatibility: Systematic Diffusion in Student

- We discover that in the presence of an LS-trained teacher, KD at higher  $T$  **systematically diffuses** penultimate layer representations learnt by the student **towards semantically similar classes**.



# Revisiting LS and KD Compatibility: Systematic Diffusion in Student

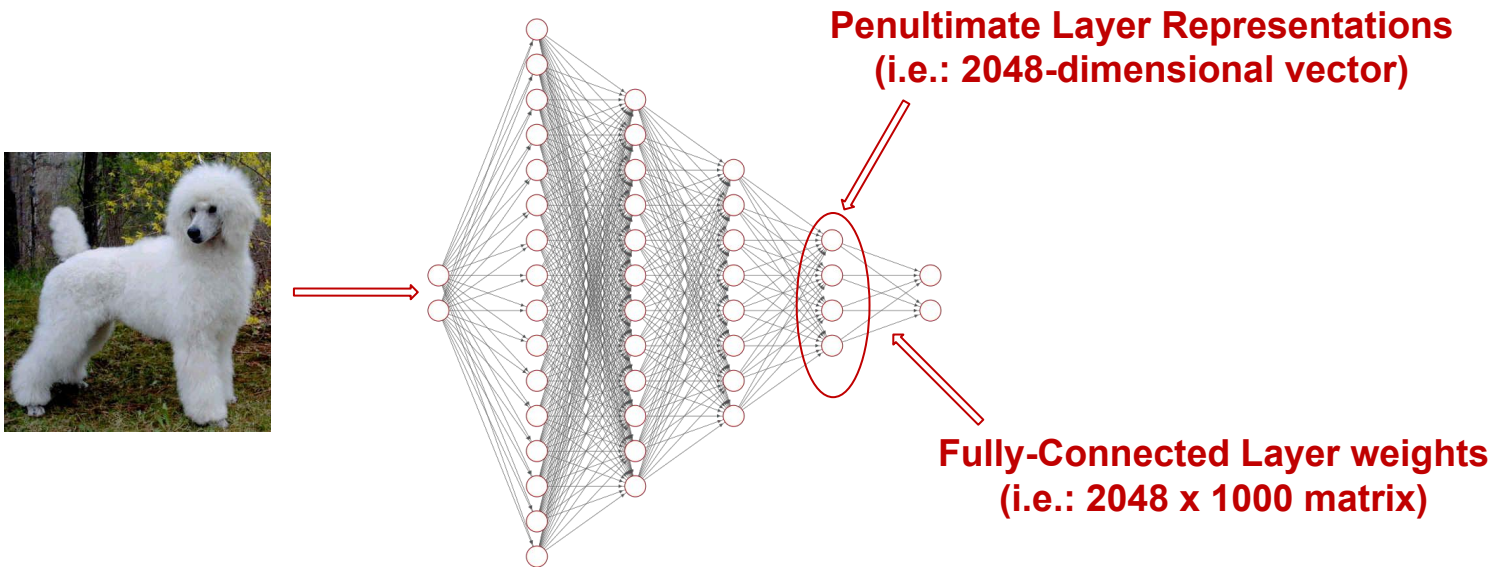
- We discover that in the presence of an LS-trained teacher, KD at higher  $T$  **systematically diffuses** penultimate layer representations learnt by the student **towards semantically similar classes**.
- This systematic diffusion is critical as it directly **curtails the distance enlargement benefits between semantically similar classes** when distilling from an LS-trained teacher

# Revisiting LS and KD Compatibility: Systematic Diffusion in Student

- We discover that in the presence of an LS-trained teacher, KD at higher  $T$  **systematically diffuses** penultimate layer representations learnt by the student **towards semantically similar classes**.
- This systematic diffusion is critical as it directly **curtails the distance enlargement benefits between semantically similar classes** when distilling from an LS-trained teacher
- Therefore, in the presence of an LS-trained teacher, **KD at increased temperatures is rendered ineffective**.

# Penultimate Layer Visualization to demonstrate Systematic Diffusion

We use linear projections of the Penultimate Layer Representations (Müller et al., 2019) to **qualitatively** demonstrate Systematic Diffusion.



# Systematic Diffusion using three-class analysis



Standard poodle

**Target class**



Miniature poodle

**Semantically similar class**



Submarine

**Semantically dissimilar class**

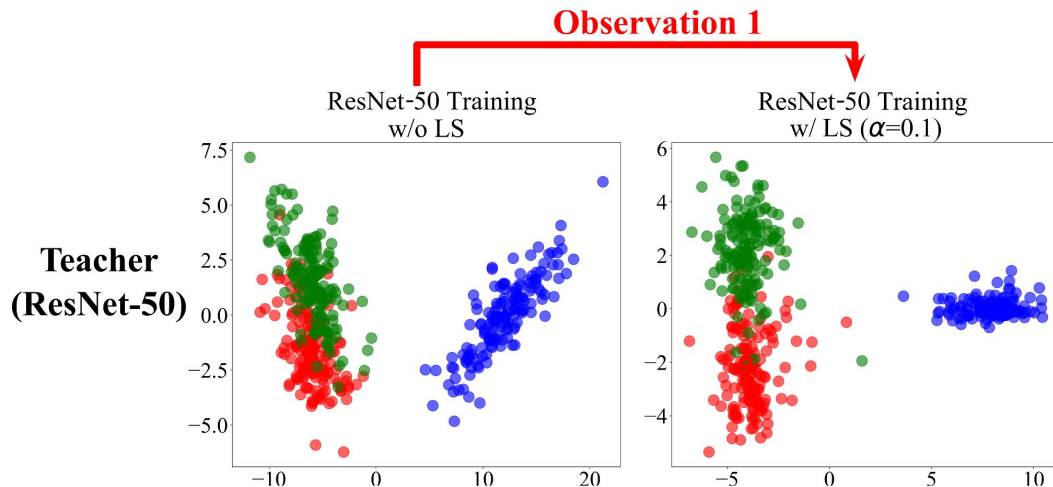
We follow experiment setups of Müller et al., 2019 and Shen et al., 2021

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help?. *Advances in neural information processing systems*, 32.

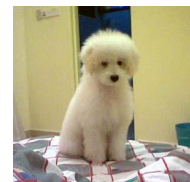
Shen, Z., Liu, Z., Xu, D., Chen, Z., Cheng, K. T., & Savvides, M. (2021). Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study. In *ICLR*

# Results : Penultimate Layer Visualization

● standard\_poodle    ● miniature\_poodle    ● submarine



Standard poodle



Miniature poodle



Submarine

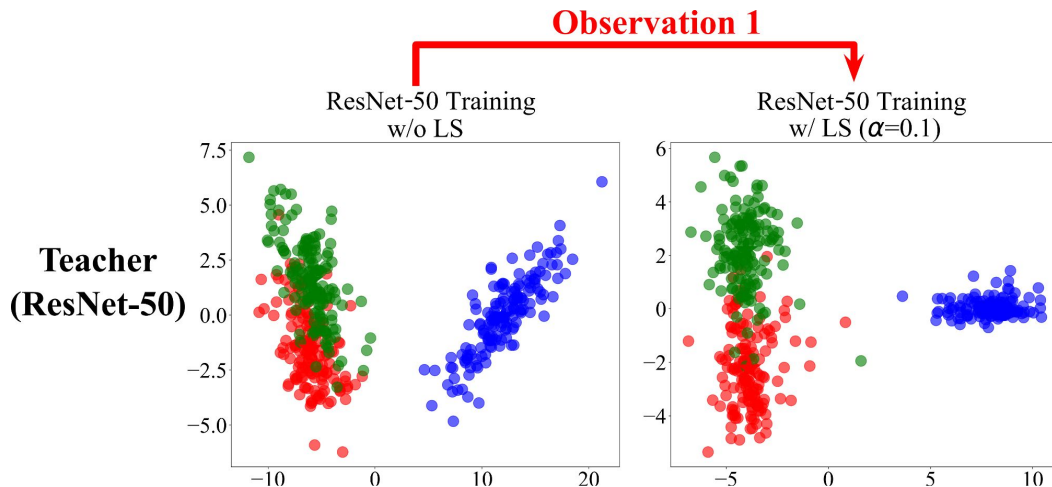
**Observation 1:** The use of LS on the teacher leads to **tighter clusters which shows information erasure in logits**. Information about resemblances to instances of different classes is essential for KD (Müller et al., 2019) → **LS and KD Incompatibility**

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help?. *Advances in neural information processing systems*, 32.

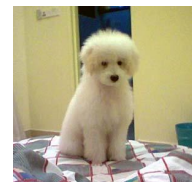
Shen, Z., Liu, Z., Xu, D., Chen, Z., Cheng, K. T., & Savvides, M. (2021). Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study. In *ICLR*

# Results : Penultimate Layer Visualization

● standard\_poodle      ● miniature\_poodle      ● submarine



Standard poodle



Miniature poodle



Submarine

**Observation 1:** Increase in central cluster distance between semantically similar classes (standard poodle, miniature poodle) can be observed with the use of LS (Shen et al., 2021) → **LS and KD Compatibility**

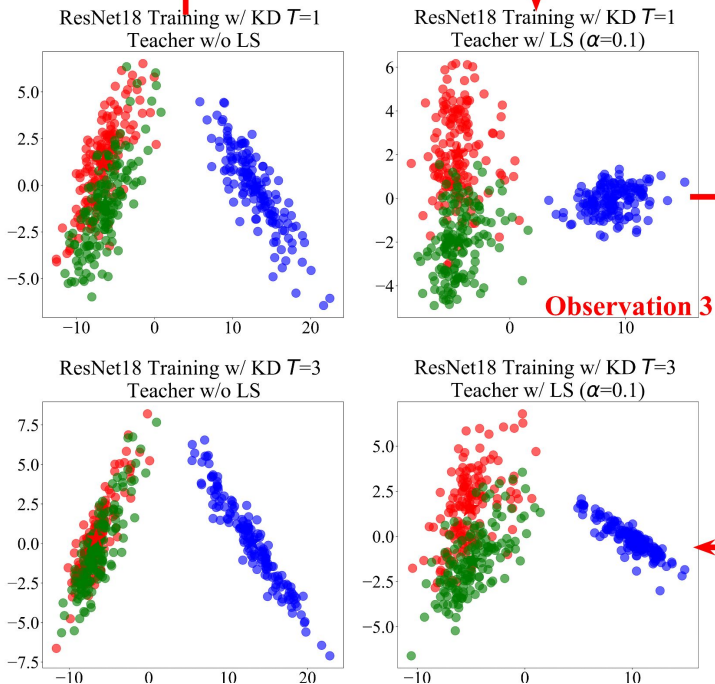
Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help?. *Advances in neural information processing systems*, 32.

Shen, Z., Liu, Z., Xu, D., Chen, Z., Cheng, K. T., & Savvides, M. (2021). Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study. In *ICLR*

# Results : Penultimate Layer Visualization

● standard\_poodle    ● miniature\_poodle    ● submarine

Observation 2



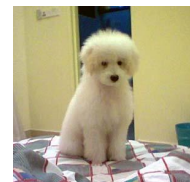
**Observation 2:** We visualize the student's representations.

Both information erasure in logits' and increase in central distance between semantically similar classes can be observed in the student.

This confirms the transfer of this drawback / benefit from the teacher to the student.



Standard poodle



Miniature poodle

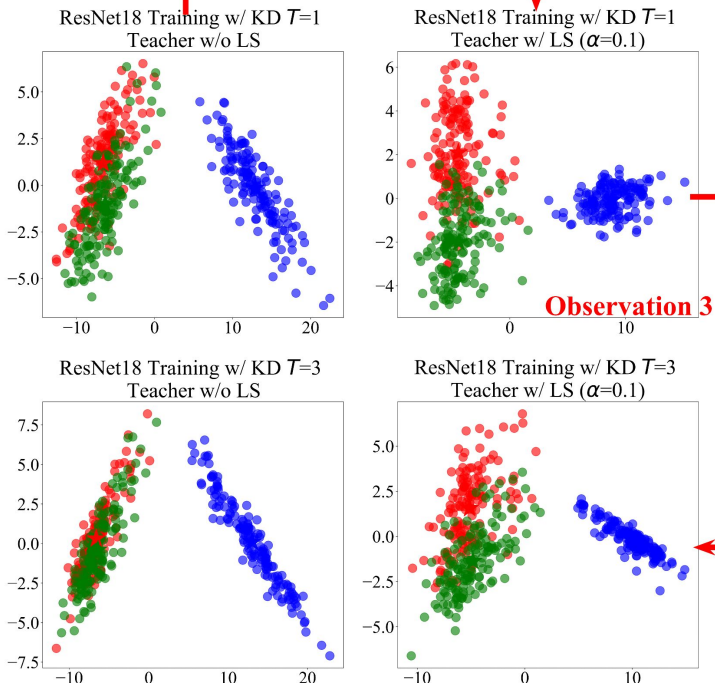


Submarine

# Results : Penultimate Layer Visualization

● standard\_poodle    ● miniature\_poodle    ● submarine

Observation 2



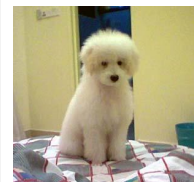
**Observation 3 (Systematic Diffusion):**  
KD of an increased  $T$  causes **systematic diffusion** of representations between semantically similar classes (**standard poodle, miniature poodle**).

This **curtails the central distance enlargement benefits** between semantically similar classes due to the use of an LS-trained teacher.

Systematic Diffusion  $\rightarrow$  **LS and KD Incompatibility**



Standard poodle



Miniature poodle



Submarine



# Diffusion Index ( $\eta$ ) to Quantify Systematic Diffusion

The principal idea of this metric is to **quantify the distance change between clusters in the student** when distilled from an LS-trained teacher at higher  $T$ .

The design of the metric is to **quantify and verify that the diffusion is systematic**: i.e., quantify Observation 3

# Diffusion Index ( $\eta$ ) to Quantify Systematic Diffusion

**Target class**  
(i.e.: Standard poodle)

**Normalized Distance between the centroid of target class  $\pi$  and class  $k$  when distilled at  $T_2$**

**Normalized Distance between the centroid of target class  $\pi$  and class  $k$  when distilled at  $T_1$**

**Distillation Temperatures**

**Set of Classes**  
 $S_1$ : semantically similar classes  
 $S_2$ : semantically dissimilar classes

$$\eta(T_1, T_2; \pi, S) = \frac{1}{|S|} \sum_{k \in S} \frac{d(\mathbf{c}_\pi(T_2), \mathbf{c}_k(T_2)) - d(\mathbf{c}_\pi(T_1), \mathbf{c}_k(T_1))}{d(\mathbf{c}_\pi(T_1), \mathbf{c}_k(T_1))}$$

# Diffusion Index ( $\eta$ ) to Quantify Systematic Diffusion

Target class  
(i.e.: Standard poodle)

Normalized Distance between the  
centroid of target class  $\pi$  and class  $k$   
when distilled at  $T_2$

Normalized Distance between the  
centroid of target class  $\pi$  and class  $k$   
when distilled at  $T_1$

Distillation  
Temperatures

Set of Classes  
 $S_1$ : semantically similar classes  
 $S_2$ : semantically dissimilar classes

$$\eta(T_1, T_2; \pi, S) = \frac{1}{|S|} \sum_{k \in S} \frac{d(\mathbf{c}_\pi(T_2), \mathbf{c}_k(T_2)) - d(\mathbf{c}_\pi(T_1), \mathbf{c}_k(T_1))}{d(\mathbf{c}_\pi(T_1), \mathbf{c}_k(T_1))}$$

Given  $T_1 < T_2$  Systematic Diffusion if  
 $\eta(T_1, T_2; \pi, S_1) < 0$  &  $\eta(T_1, T_2; \pi, S_2) > 0$

# Experiments

<b>Task</b>	<b>Datasets</b>	<b>Architectures</b>
Image Classification	ImageNet-1K	ResNet-18, ResNet-50
Neural Machine Translation	En – De (IWSLT) En – Ru (IWSLT)	Transformers
Fine-grained Image Classification	CUB200-2011	ResNet-18, ResNet-50, ConvNeXt-T
Compact Student Distillation	ImageNet-1K CUB200-2011	EfficientNet-B0 MobileNetV2

# Experiments

Task	Datasets	Architectures
<b>Image Classification</b>	<b>ImageNet-1K</b>	<b>ResNet-18, ResNet-50</b>
Neural Machine Translation	En – De (IWSLT) En – Ru (IWSLT)	Transformers
Fine-grained Image Classification	CUB200-2011	ResNet-18, ResNet-50, ConvNeXt-T
Compact Student Distillation	ImageNet-1K CUB200-2011	EfficientNet-B0 MobileNetV2

# Results (ImageNet-1K) : KD using LS-trained teacher

We show Top1/ Top5 Accuracies

A. ImageNet-1K : ResNet-50 to ResNet-18, ResNet-50 KD

	$T \backslash \alpha$	$\alpha = 0.0$	$\alpha = 0.1$
Teacher : ResNet-50	-	76.130 / 92.862	76.196 / 93.078
Student : ResNet-18	$T = 1$	71.547 / 90.297	<b>71.616 / 90.233</b>
	$T = 2$	71.349 / 90.359	68.428 / 89.139
	$T = 3$	69.570 / 89.657	66.570 / 88.631
	$T = 64$	66.230 / 88.730	65.472 / 89.564

# Results (ImageNet-1K) : KD using LS-trained teacher

We show Top1/ Top5 Accuracies

A. ImageNet-1K : ResNet-50 to ResNet-18, ResNet-50 KD

	$T \backslash \alpha$	$\alpha = 0.0$	$\alpha = 0.1$
Teacher : ResNet-50	-	76.130 / 92.862	76.196 / 93.078
Student : ResNet-18	$T = 1$	71.547 / 90.297	<b>71.616 / 90.233</b>
	$T = 2$	71.349 / 90.359	68.428 / 89.139
	$T = 3$	69.570 / 89.657	66.570 / 88.631
	$T = 64$	66.230 / 88.730	65.472 / 89.564

In the presence of an LS-trained teacher, at higher  $T$ , KD is rendered ineffective due to Systematic Diffusion in student.

# Results (ImageNet-1K) : KD using LS-trained teacher

We show Top1/ Top5 Accuracies

A. ImageNet-1K : ResNet-50 to ResNet-18, ResNet-50 KD

	$T \backslash \alpha$	$\alpha = 0.0$	$\alpha = 0.1$
Teacher : ResNet-50	-	76.130 / 92.862	76.196 / 93.078
Student : ResNet-18	$T = 1$	71.547 / 90.297	<b>71.616 / 90.233</b>
	$T = 2$	71.349 / 90.359	68.428 / 89.139
	$T = 3$	69.570 / 89.657	66.570 / 88.631
	$T = 64$	66.230 / 88.730	65.472 / 89.564

Rapid degrade in Student performance (Systematic Diffusion) with increased  $T$  in the presence of LS-trained teacher compared to baseline



# Results (ImageNet-1K) : KD using LS-trained teacher

We show Top1/ Top5 Accuracies

A. ImageNet-1K : ResNet-50 to ResNet-18, ResNet-50 KD

	$T \backslash \alpha$	$\alpha = 0.0$	$\alpha = 0.1$
Teacher : ResNet-50	-	76.130 / 92.862	76.196 / 93.078
Student : ResNet-18	$T = 1$	71.547 / 90.297	<b>71.616 / 90.233</b>
	$T = 2$	71.349 / 90.359	68.428 / 89.139
	$T = 3$	69.570 / 89.657	66.570 / 88.631
	$T = 64$	66.230 / 88.730	65.472 / 89.564

LS-trained teacher with a low-temperature transfer (i.e.,  $T = 1$ ) obtains the best ResNet-18 student

# Results (ImageNet-1K): $\eta$ measurements showing Systematic Diffusion

$S_1$  and  $S_2$  selected using standard, pre-defined ImageNet knowledge graph  
(WordNet, Fellbaum, 1998)

Set 1 : ResNet-18 student

Target class	$Train : S_1$	$Train : S_2$	$Val : S_1$	$Val : S_2$
Chesapeake Bay retriever	-0.392	0.162	-1.082	0.269
curly-coated retriever	-0.578	0.179	-2.024	0.383
flat-coated retriever	-1.729	0.380	-3.320	0.655
golden retriever	-0.880	0.228	-2.594	0.555
Labrador retriever	-2.758	0.501	-4.618	0.840

Set 2 : ResNet-18 student

Target class	$Train : S_1$	$Train : S_2$	$Val : S_1$	$Val : S_2$
thunder snake	-2.316	0.376	-3.584	0.511
ringneck snake	-0.463	0.058	-0.757	0.094
hognose snake	-1.528	0.258	-4.067	0.631
water snake	-2.028	0.326	-3.053	0.478
king snake	-2.474	0.521	-4.577	0.840

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press. ISBN: 978-0-262-06197-1

<https://observablehq.com/@mbostock/imagenet-hierarchy>

# Results (ImageNet-1K): $\eta$ measurements showing Systematic Diffusion

$S_1$  and  $S_2$  selected using standard, pre-defined ImageNet knowledge graph

Set 1 : ResNet-18 student

Target class	$Train : S_1$	$Train : S_2$	$Val : S_1$	$Val : S_2$
Chesapeake Bay retriever	-0.392	0.162	-1.082	0.269
curly-coated retriever	-0.578	0.179	-2.024	0.383
flat-coated retriever	-1.729	0.380	-3.320	0.655
golden retriever	-0.880	0.228	-2.594	0.555
Labrador retriever	-2.758	0.501	-4.618	0.840

Set 2 : ResNet-18 student

Target class	$Train : S_1$	$Train : S_2$	$Val : S_1$	$Val : S_2$
thunder snake	-2.316	0.376	-3.584	0.511
ringneck snake	-0.463	0.058	-0.757	0.094
hognose snake	-1.528	0.258	-4.067	0.631
water snake	-2.028	0.326	-3.053	0.478
king snake	-2.474	0.521	-4.577	0.840

$$\eta(T_1 = 1, T_2 = 3; \pi, S_1) < 0$$

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press. ISBN: 978-0-262-06197-1

<https://observablehq.com/@mbostock/imagenet-hierarchy>

# Results (ImageNet-1K): $\eta$ measurements showing Systematic Diffusion

$S_1$  and  $S_2$  selected using standard, pre-defined ImageNet knowledge graph

Set 1 : ResNet-18 student

Target class	$Train : S_1$	$Train : S_2$	$Val : S_1$	$Val : S_2$
Chesapeake Bay retriever	-0.392	0.162	-1.082	0.269
curly-coated retriever	-0.578	0.179	-2.024	0.383
flat-coated retriever	-1.729	0.380	-3.320	0.655
golden retriever	-0.880	0.228	-2.594	0.555
Labrador retriever	-2.758	0.501	-4.618	0.840

Set 2 : ResNet-18 student

Target class	$Train : S_1$	$Train : S_2$	$Val : S_1$	$Val : S_2$
thunder snake	-2.316	0.376	-3.584	0.511
ringneck snake	-0.463	0.058	-0.757	0.094
hognose snake	-1.528	0.258	-4.067	0.631
water snake	-2.028	0.326	-3.053	0.478
king snake	-2.474	0.521	-4.577	0.840

$$\eta(T_1 = 1, T_2 = 3; \pi, S_2) > 0$$

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press. ISBN: 978-0-262-06197-1

<https://observablehq.com/@mbostock/imagenet-hierarchy>

# Revisiting LS and KD Compatibility : Systematic Diffusion is Critical

	Information Erasure (Incompatibility)	Distance enlargement (compatibility)	<b>Systematic Diffusion (Incompatibility)</b>	<b>Conclusion</b>	
Müller et al. 2019	LS erases relative information in the logits			LS-trained teacher can hurt KD	
Shen et al. 2021	With LS, some relative information in the logits is still retained	LS enlarges the distance between semantically similar classes		Benefits outweigh disadvantages. LS is compatible with KD.	
Our work	<b>Lower <math>T</math> (i.e.: <math>T=1</math>)</b>	We agree with Shen et al., 2021 in information erasure	We validate the inheritance of distance enlargement in the student (Not shown in prior work)	<b>With KD of lower <math>T</math> (i.e.: <math>T=1</math>), there is lower degree of systematic diffusion. This doesn't curtail the distance enlargement benefit.</b>	<b>At lower levels of systematic diffusion in student, LS is compatible with KD</b>
	<b>Increase of <math>T</math></b>	The loss of logits relative information cannot be recovered with an increased $T$	We agree with Shen et al., 2021 observation, but the distance enlargement is curtailed at an increased $T$ .	<b>With KD of increased <math>T</math>, there is systematic diffusion of penultimate representations towards semantically similar classes, curtailing the distance enlargement benefits.</b>	<b>At higher levels of systematic diffusion in student, LS and KD are not compatible.</b>

# Revisiting LS and KD Compatibility : Key Takeaways for Practitioners

**Systematic Diffusion** can be qualitatively observed using **Penultimate Layer Visualization** and quantitatively measured using our **proposed  $\eta$** .

**As rule of thumb**, we suggest **using an LS-trained teacher with a low-temperature transfer** (i.e.,  $T = 1$ ) to render high performance students.

Code / Pre-trained models



Thank you