IBM Research

Alleviating Noisy Data in Image Captioning with Cooperative Distillation

PIERRE DOGNIN, IGOR MELNYK, YOUSSEF MROUEH, INKIT PADHI, MATTIA RIGOTTI, JARRET ROSS, YAIR SCHIFF*

CVPR VizWiz Grand Challenge Workshop
June 14, 2020

Image credit: http://ecollectivedesign.Com/distillation-process-identify-core/

*All authors contributed equally

Motivation: Leverage clean and noisy datasets

Background

✓ Multimodal datasets with noisy labels are ubiquitous, cheap to collect and available abundantly at scale

x Clean multimodal data is expensive to collect and available at a smaller scale

Goal

Leverage noisy and clean datasets to **alleviate shortcomings** of each while **benefiting from** their respective **strengths**:

- Overcome the noise barrier
- Improve the accuracy of the trained models

Setup: Teacher and student models each learning on their respective datasets

- Student model $\mathcal{S}$ learning from noisy dataset S (we used ***Google Conceptual Captioning***[1], GCC, for the student dataset)

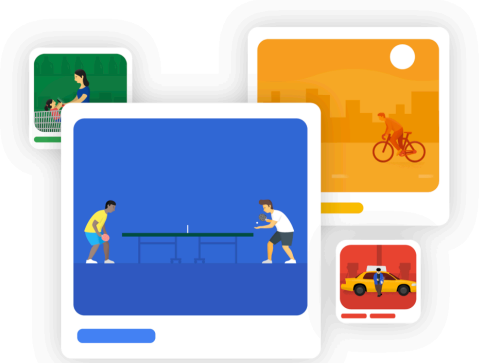


Image credit: https://ai.google.com/research/ConceptualCaptions

- Teacher model $\mathcal{T}$ learning from a clean dataset T (we use ***MS COCO***[2] for the teacher dataset)

COCO
Common Objects in Context

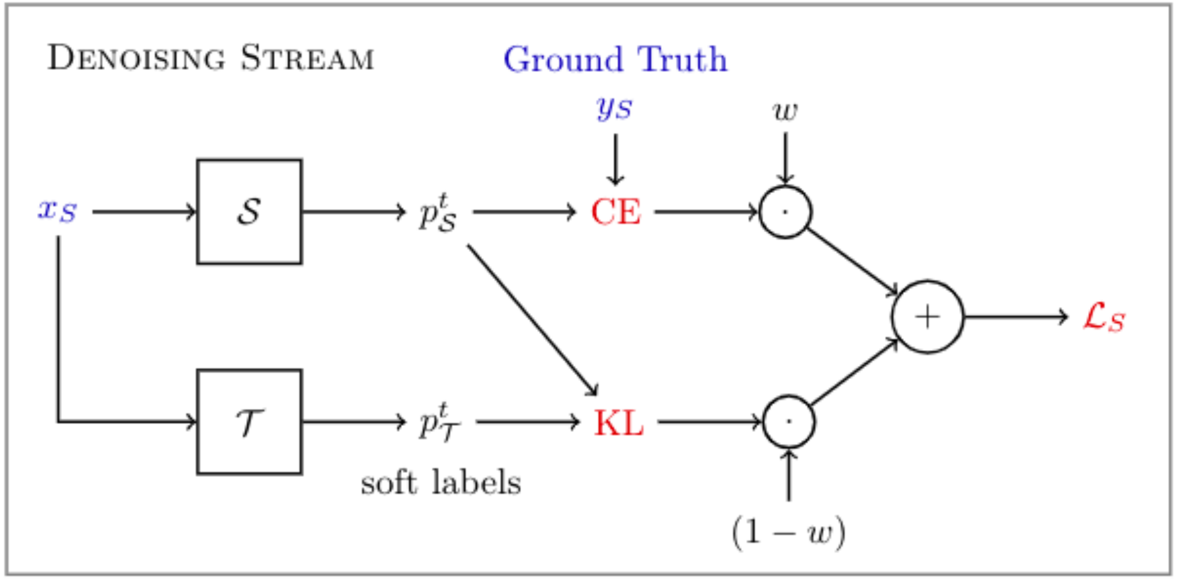
Image credit: http://cocodataset.org/#home

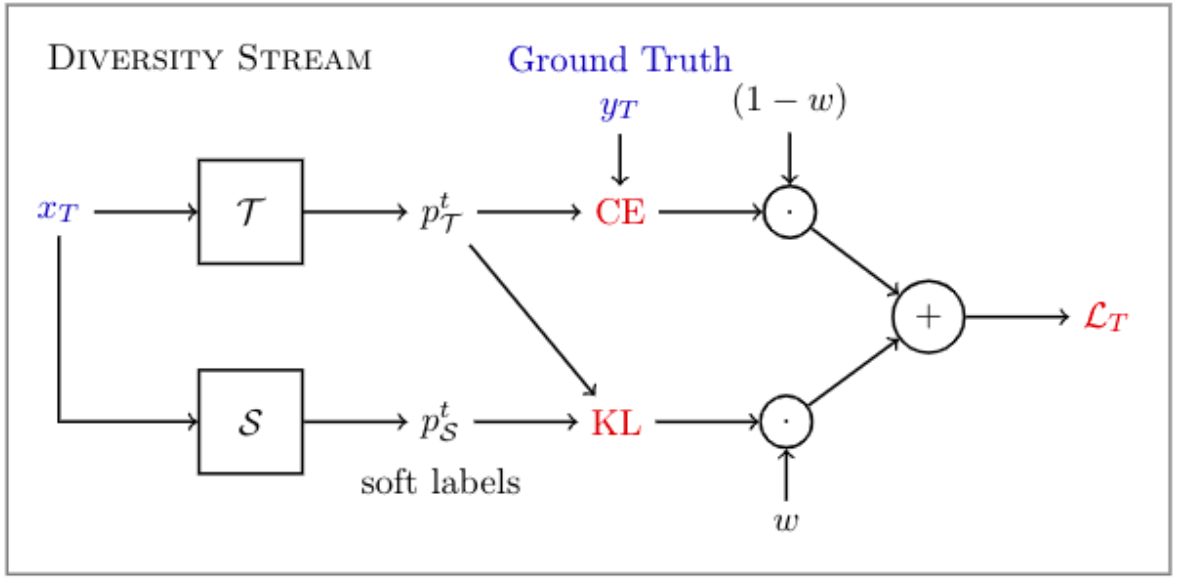
- Both teacher and student are transformer models

1. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of ACL (2018)
2. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Doll ́ar, P., Zitnick, C.L.: Microsoft COCO: common objects in context. EECV (2014)

CoDistill: Alternating between *Denoising* and *Diversity* streams

Student learns, teacher is fixed

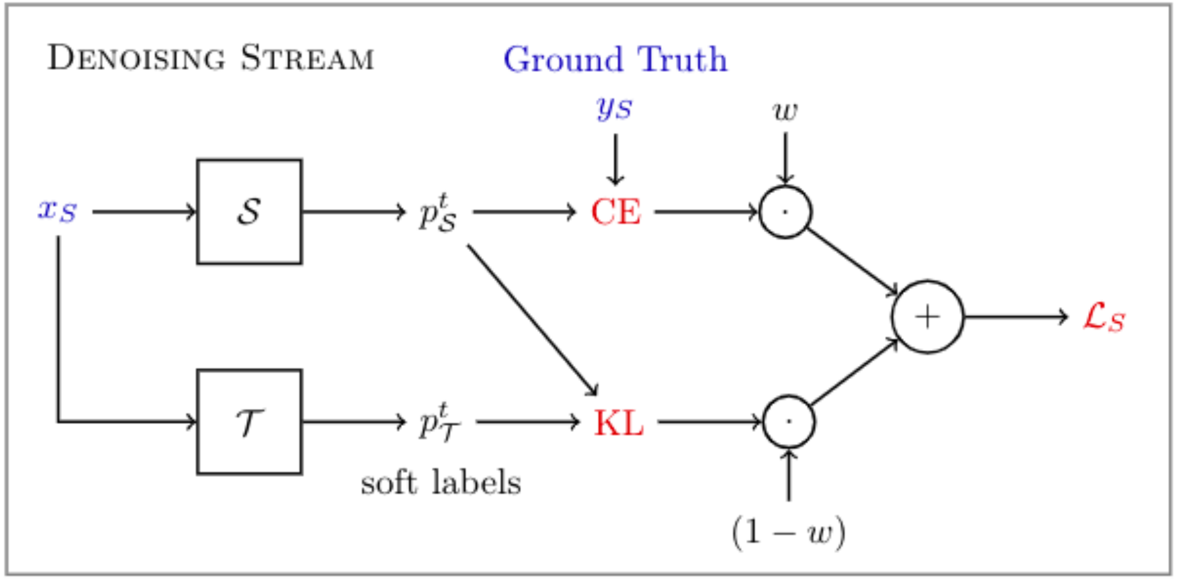


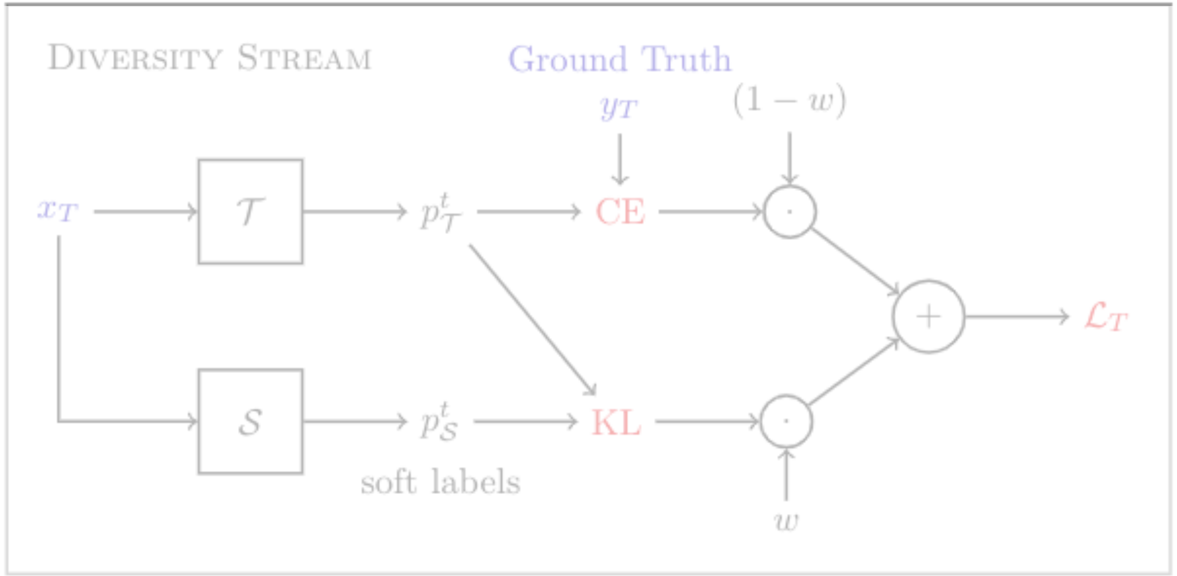


Teacher learns, student is fixed

CoDistill: Alternating between *Denoising* and *Diversity* streams

Student learns, teacher is fixed

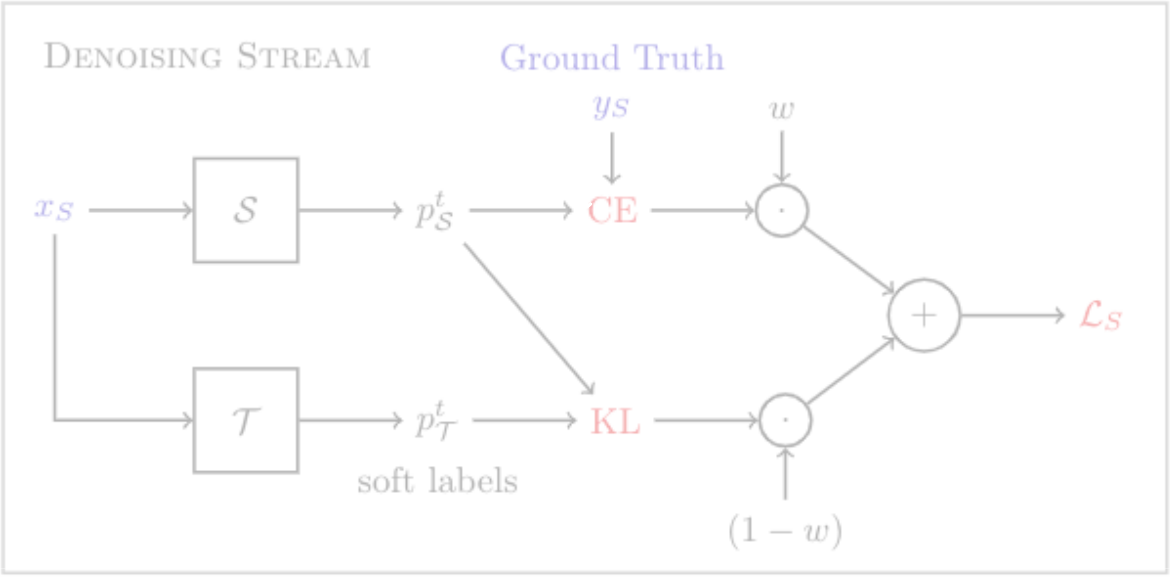


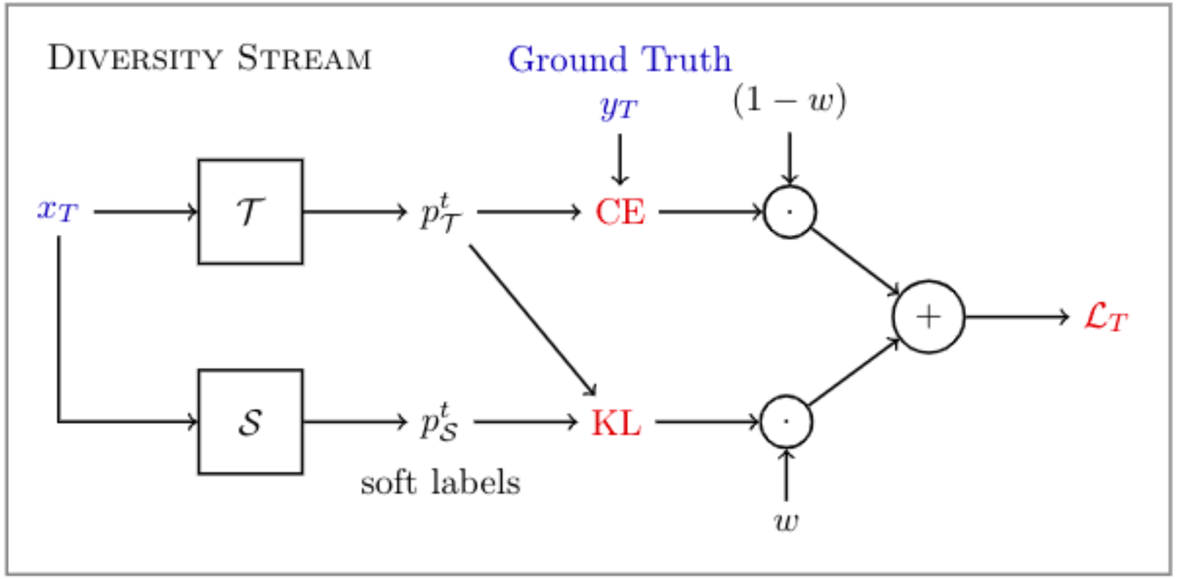


Teacher learns, student is fixed

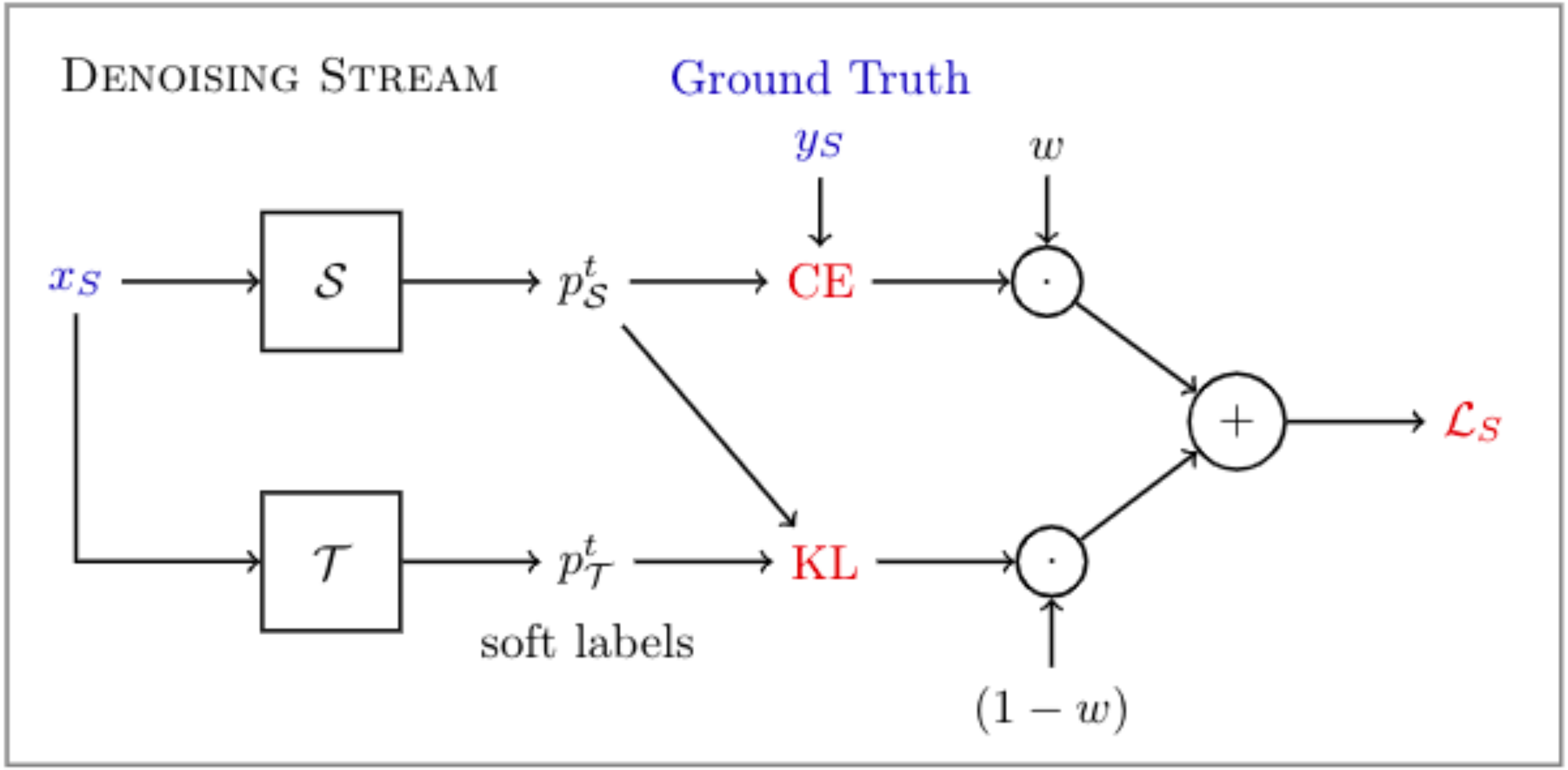
CoDistill: Alternating between *Denoising* and *Diversity* streams

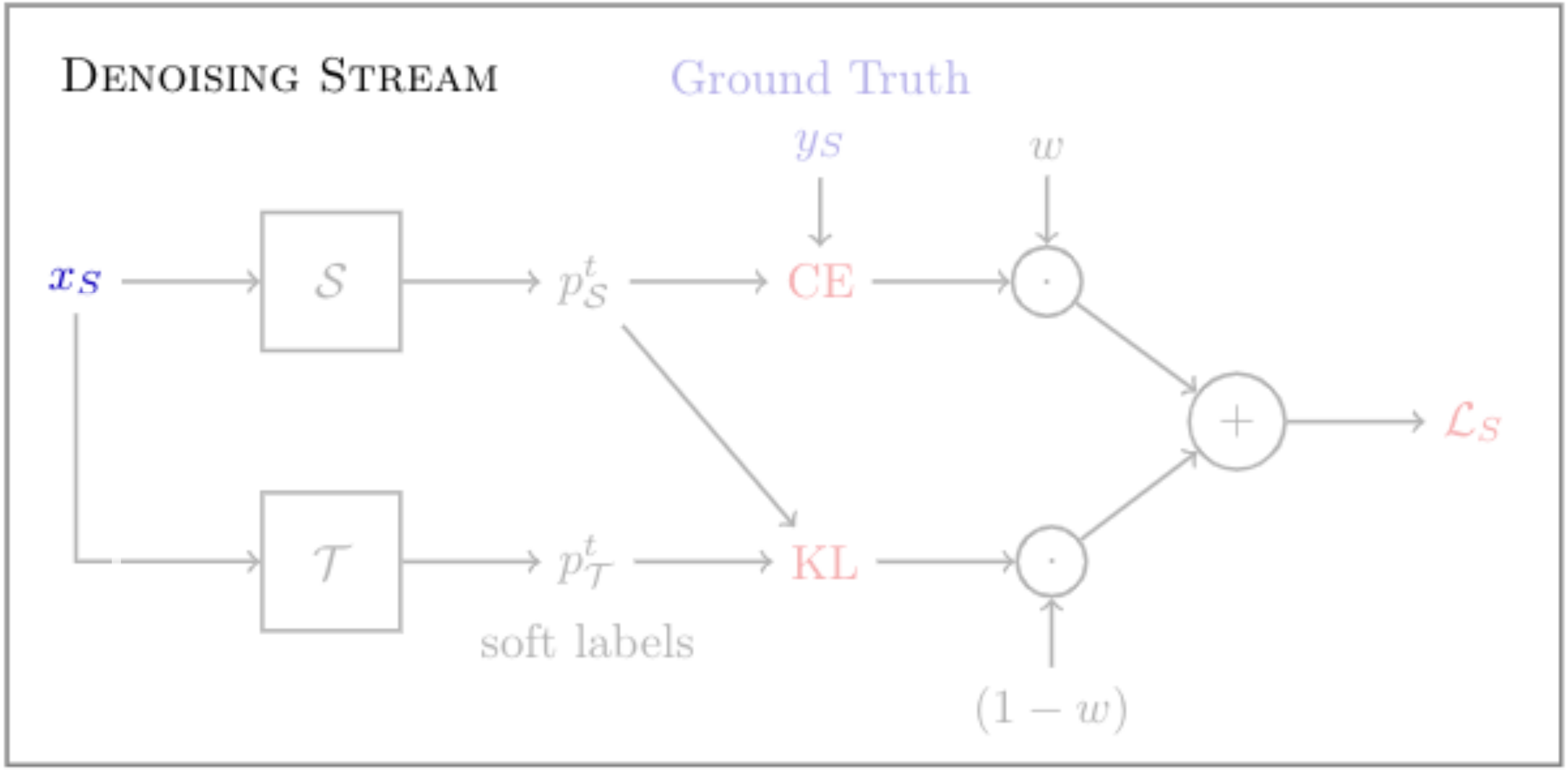
Student learns, teacher is fixed

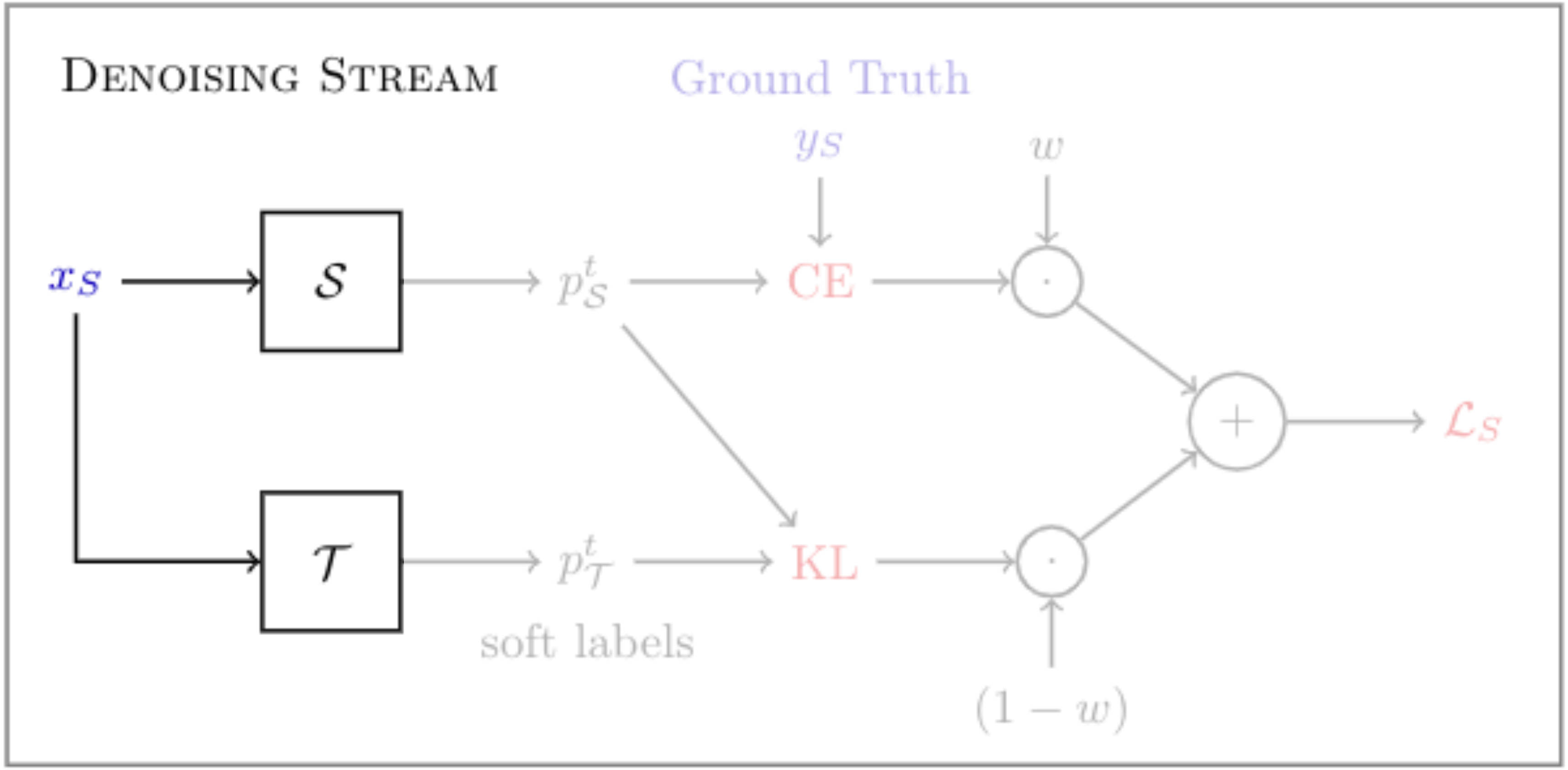




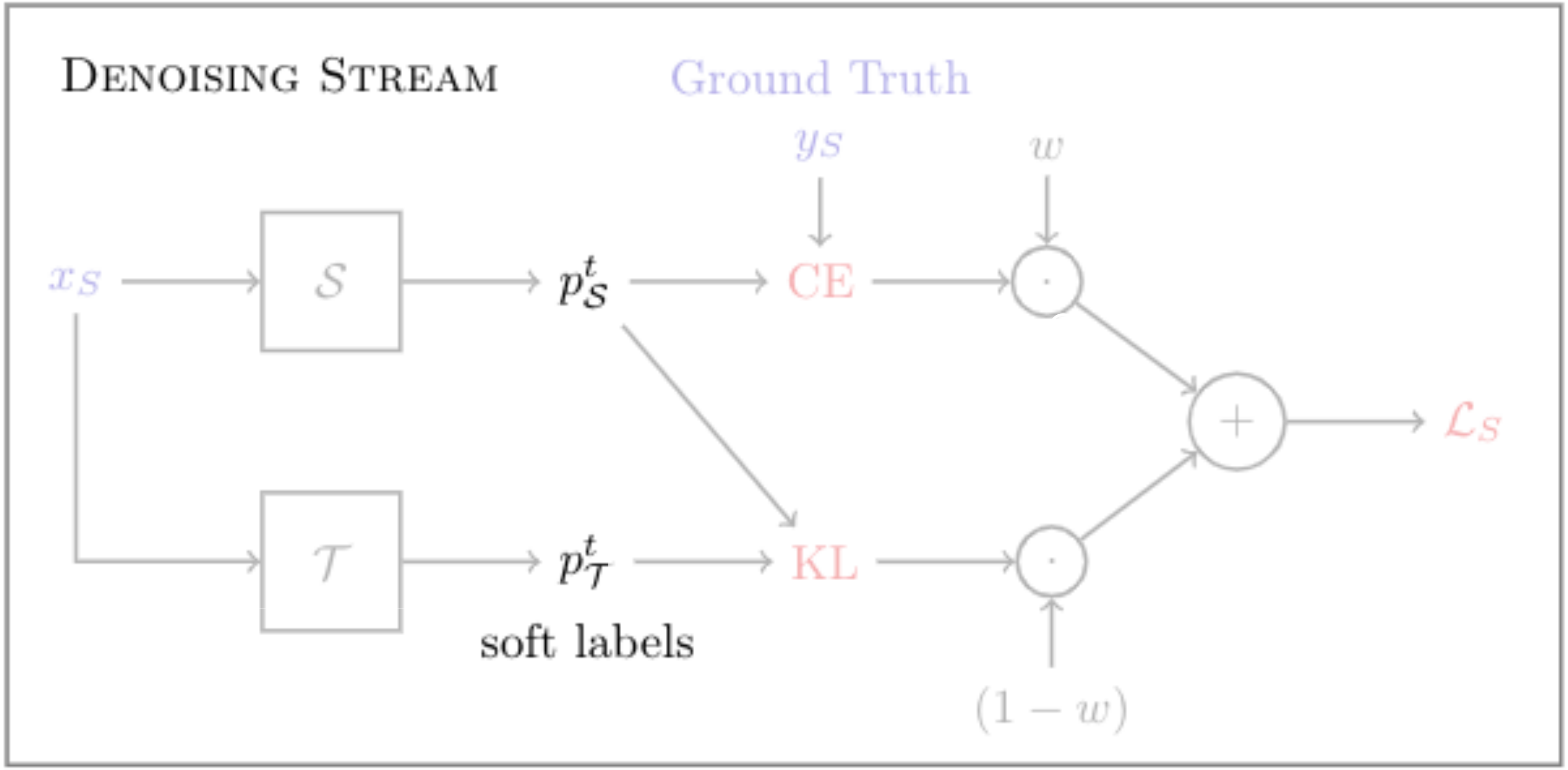
Teacher learns, student is fixed

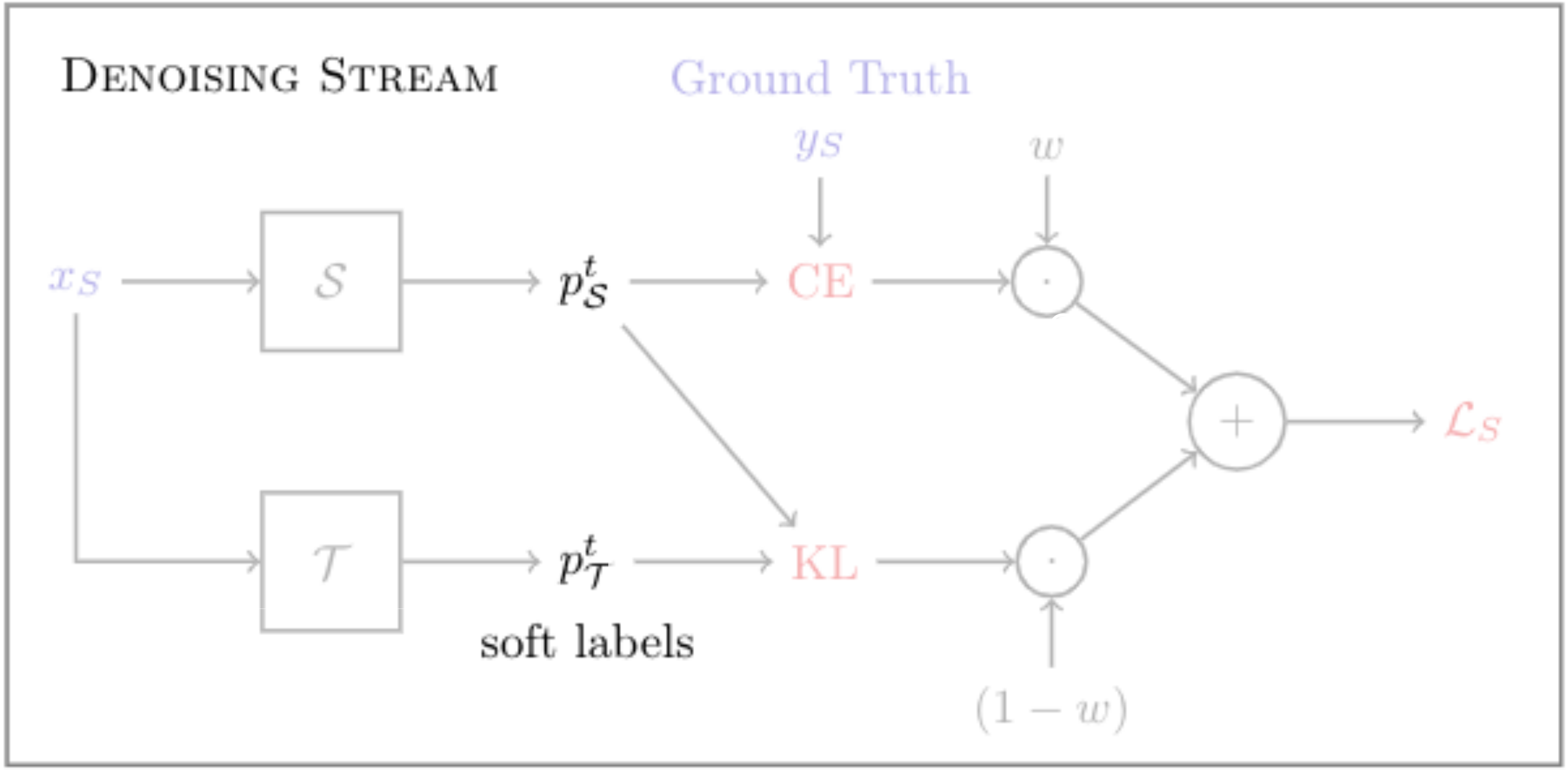

Denoising Stream
Ground Truth
y_S
w
x_S
$\mathcal{S}$
$p^t_{\mathcal{S}}$
CE
$\cdot$
$+$
$\mathcal{L}_S$
$\mathcal{T}$
$p^t_{\mathcal{T}}$
KL
$\cdot$
soft labels
$(1-w)$


DENOISING STREAM
Ground Truth
y_S
w
$\boldsymbol{x_S}$
$\mathcal{S}$
$p^t_{\mathcal{S}}$
CE
$\cdot$
$+$
$\mathcal{L}_S$
$\mathcal{T}$
$p^t_{\mathcal{T}}$
KL
$\cdot$
soft labels
$(1 - w)$

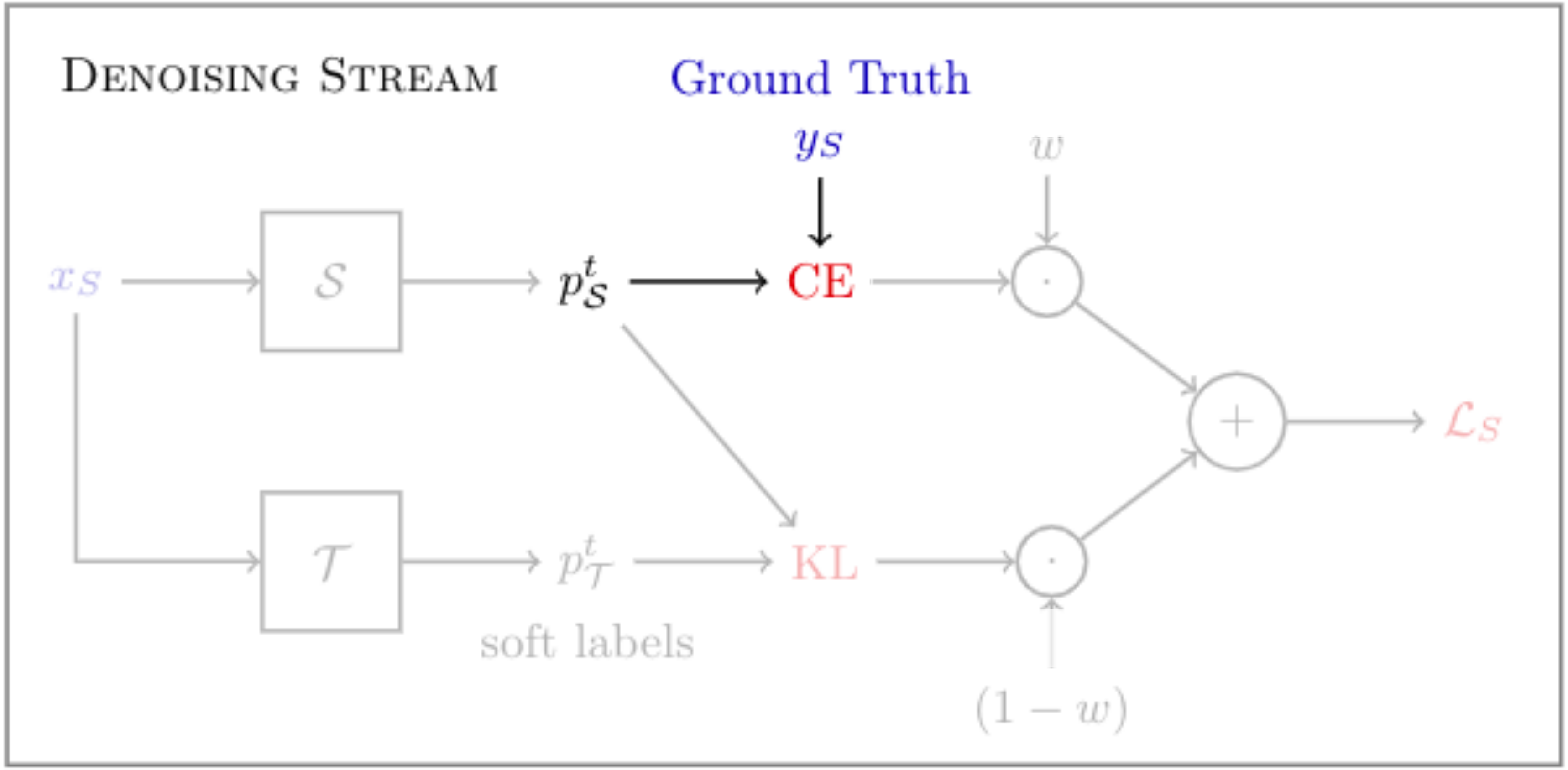


Denoising Stream
Ground Truth
y_S
w
x_S
$\mathcal{S}$
$p_{\mathcal{S}}^t$
CE
$\cdot$
$+$
$\mathcal{L}_S$
$\mathcal{T}$
$p_{\mathcal{T}}^t$
KL
$\cdot$
soft labels
$(1-w)$


DENOISING STREAM
Ground Truth
y_S
w
x_S
$\mathcal{S}$
$p^t_{\mathcal{S}}$
CE
$\cdot$
$+$
$\mathcal{L}_S$
$\mathcal{T}$
$p^t_{\mathcal{T}}$
KL
$\cdot$
soft labels
$(1-w)$



$$w(x_S, y_S)\left(-\sum_{t=1}^{L_S}\langle y_S^t, \log p_{\mathcal{S}}^t(x_S)\rangle\right) + (1 - w(x_S, y_S))\sum_{t=1}^{L_T}\mathrm{KL}(p_{\mathcal{T}}^t, p_{\mathcal{S}}^t)$$

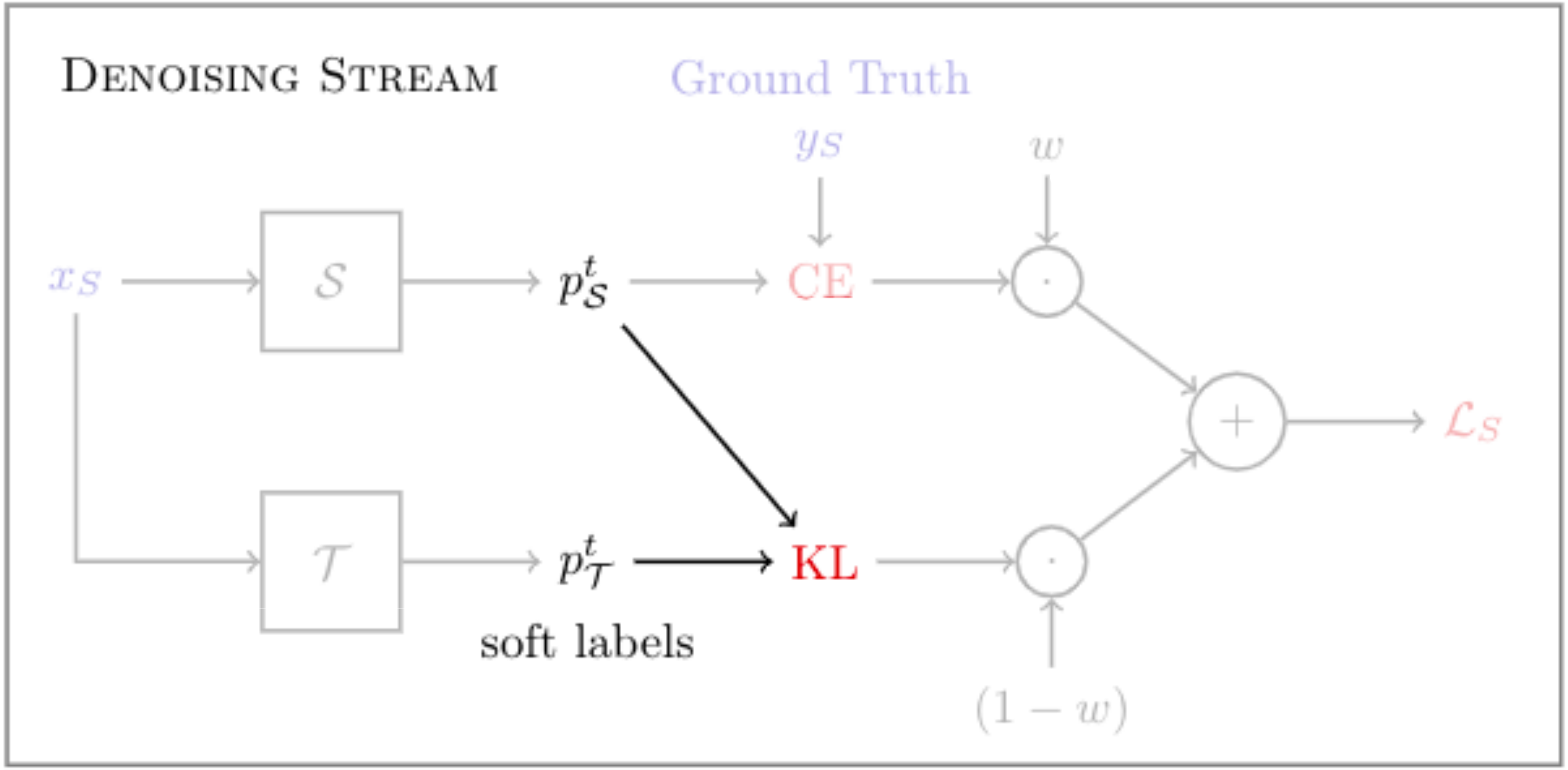


Denoising Stream

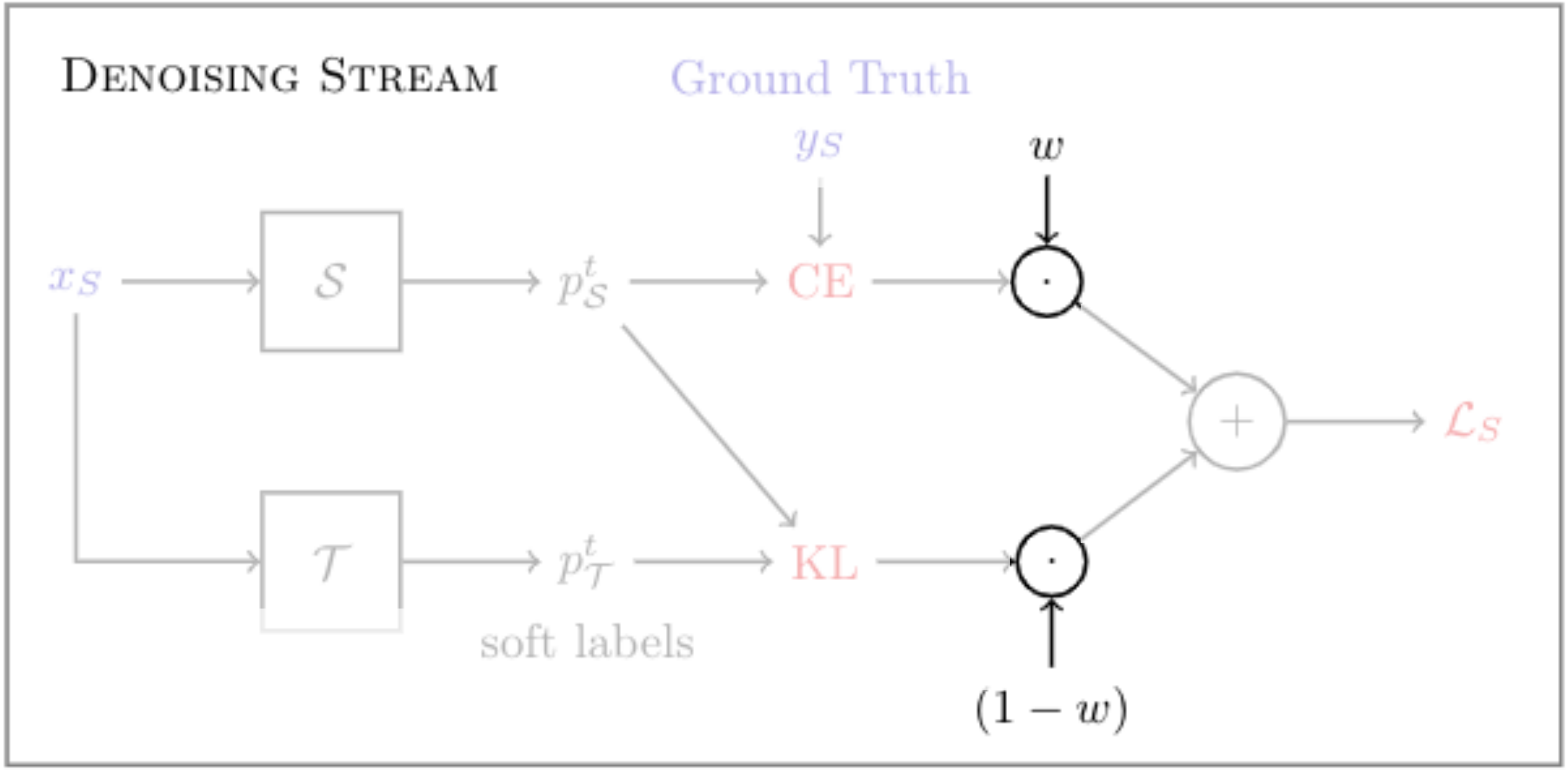
Ground Truth

soft labels

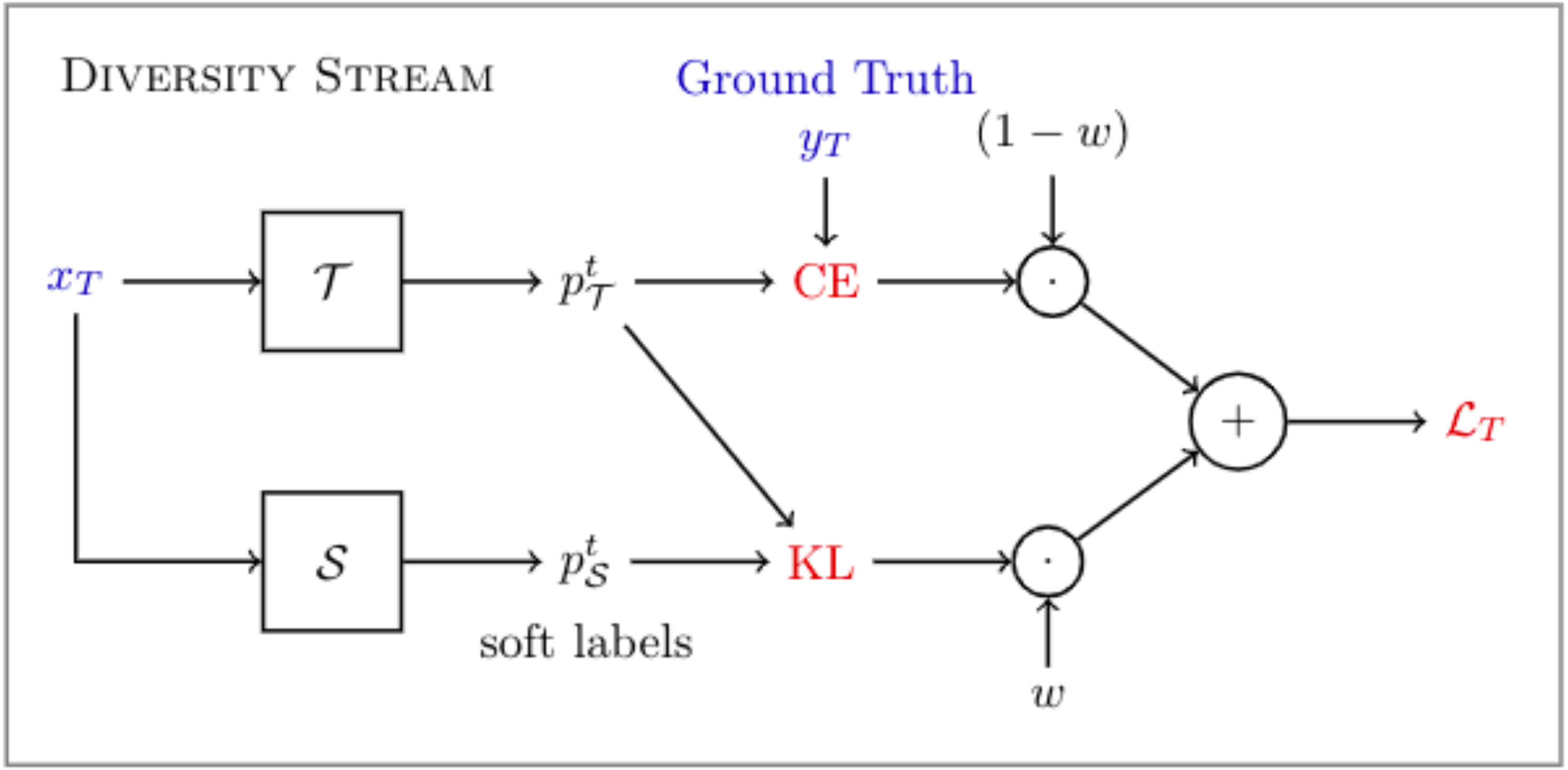
$$w(x_S, y_S)\left(-\sum_{t=1}^{L_S}\langle y^t_S, \log p^t_{\mathcal{S}}(x_S)\rangle\right) + (1 - w(x_S, y_S))\sum_{t=1}^{L_T}\mathrm{KL}(p^t_{\mathcal{T}}, p^t_{\mathcal{S}})$$



$$w(x_S, y_S)\left(-\sum_{t=1}^{L_S}\langle y_S^t, \log p_{\mathcal{S}}^t(x_S)\rangle\right) + (1-w(x_S, y_S))\sum_{t=1}^{L_T}\mathrm{KL}(p_{\mathcal{T}}^t, p_{\mathcal{S}}^t)$$



$$w(x_S, y_S)\left(-\sum_{t=1}^{L_S}\langle y_S^t, \log p_{\mathcal{S}}^t(x_S)\rangle\right) + (1 - w(x_S, y_S))\sum_{t=1}^{L_T}\mathrm{KL}(p_{\mathcal{T}}^t, p_{\mathcal{S}}^t)$$



Diversity Stream

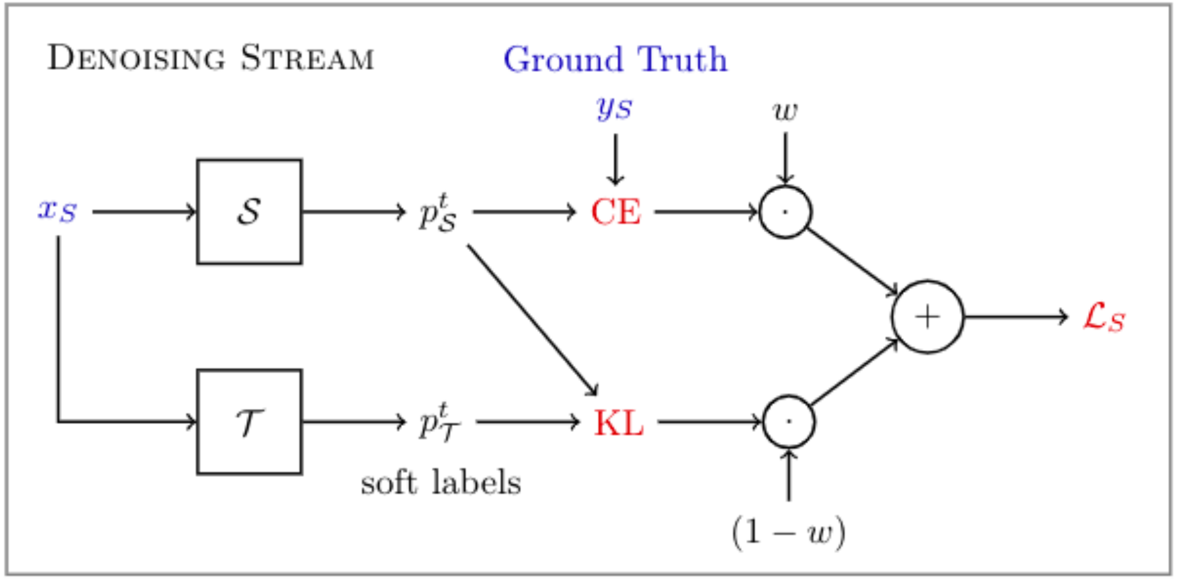
Ground Truth

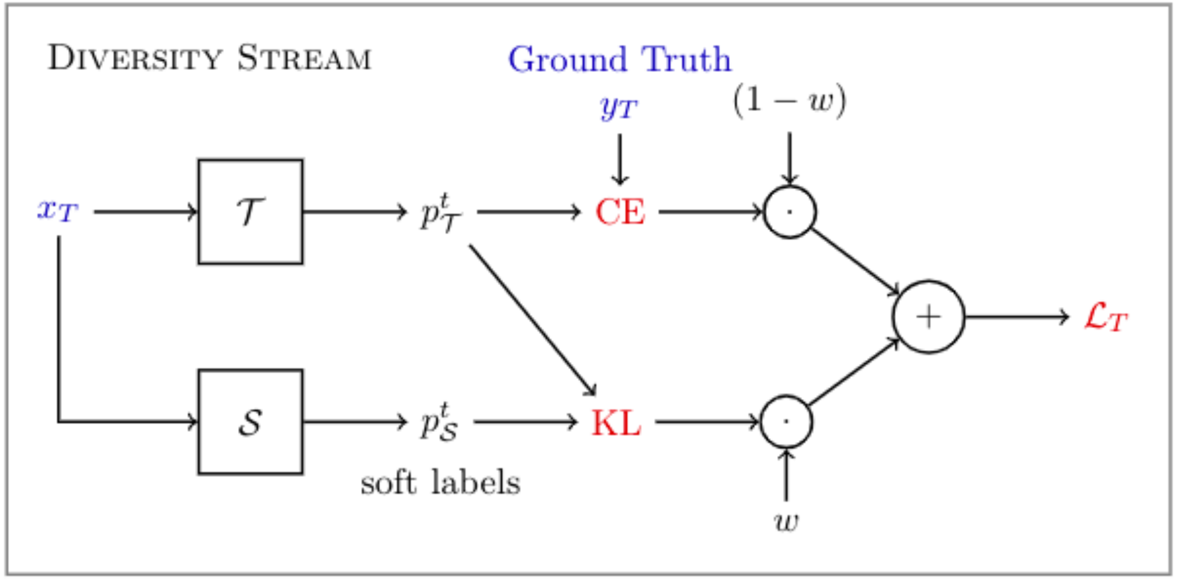
soft labels

$$(1 - w(x_T, y_T))\left(-\sum_{t=1}^{L_T} \langle y_T^t, \log p_{\mathcal{T}}^t(x_T)\rangle\right) + w(x_T, y_T)\sum_{t=1}^{L_S} \mathrm{KL}(p_{\mathcal{S}}^t, p_{\mathcal{T}}^t)$$

CoDistill: Alternating between *Denoising* and *Diversity* streams

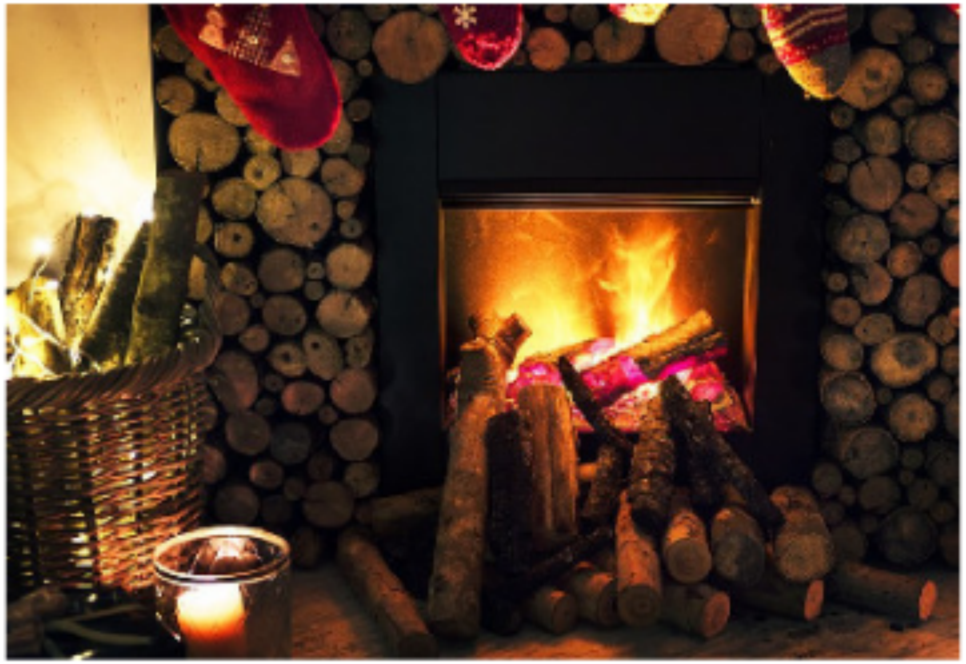
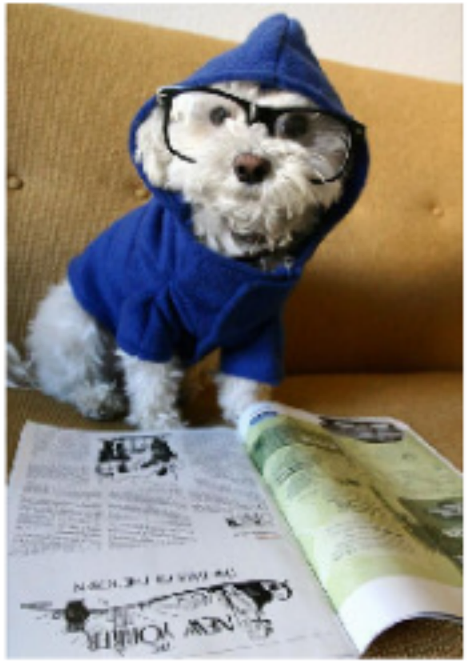
Student learns, teacher is fixed





Teacher learns, student is fixed

Denoising in action

	
GT: things to do before you move into a new house S: a fire in a fireplace surrounded by logs	GT: the most stylish dog on the internet S: a dog wearing a blue hat and glasses

Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of ACL (2018)

IBM Research

Thank you!

Image credit: http://ecollectivedesign.Com/distillation-process-identify-core/