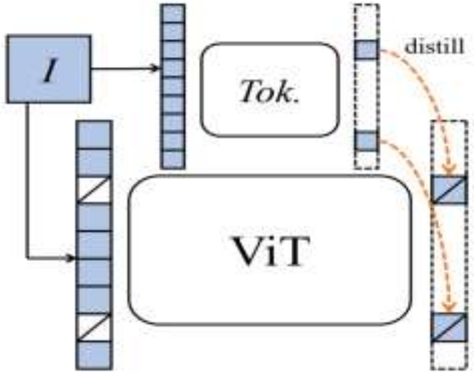
Mind Your Augmentation:
The Key to Decoupling Dense Self-supervised Learning

Congpei Qiu[1*], Tong Zhang[2*], Yanhao Wu[1], Wei Ke[1‡], Mathieu Salzmann[2], Sabine Süsstrunk[2]

[1]School of Software Engineering, Xi’an Jiaotong University, China
[2]School of Computer and Communication Sciences, EPFL, Switzerland

Background



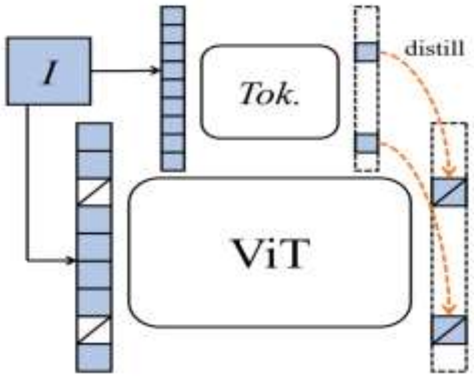
iBOT [1]

- Self-supervised Learning based on Masked Image Modelling progresses significantly in processing dense-level vision information

➞ DINO v2 [2] applies the MIM objective of IBOT as dense-level supervision

[1] Zhou, Jinghao, et al. "ibot: Image bert pre-training with online tokenizer." arXiv preprint arXiv:2111.07832 (2021).
[2] Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision." arXiv preprint arXiv:2304.07193 (2023).

Background

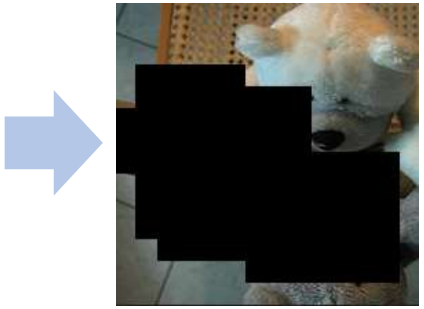


iBOT [1]

- Self-supervised Learning based on Masked Image Modelling progresses significantly in processing dense-level vision information

→ DINO v2 [2] applies the MIM objective of IBOT as dense-level supervision

Blockwise Mask

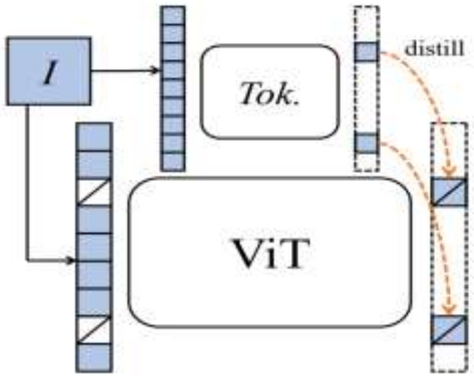


- High mask ratio leads to the loss of key semantics in multi-object images

  → Limited learning efficiency for accessing dense-level patterns

[1] Zhou, Jinghao, et al. "ibot: Image bert pre-training with online tokenizer." arXiv preprint arXiv:2111.07832 (2021).
[2] Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision." arXiv preprint arXiv:2304.07193 (2023).

Background

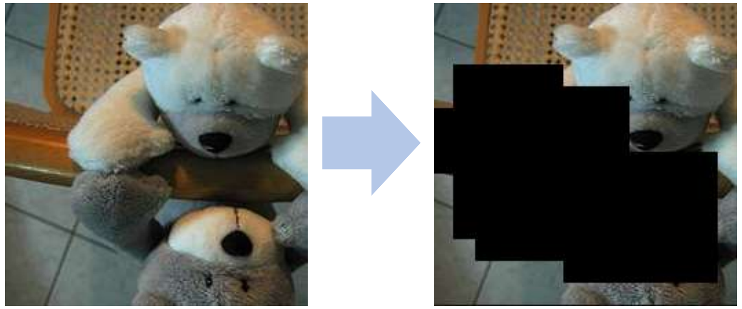


iBOT [1]

- Self-supervised Learning based on Masked Image Modelling progresses significantly in processing dense-level vision information

→ DINO v2 [2] applies the MIM objective of IBOT as dense-level supervision

Blockwise Mask

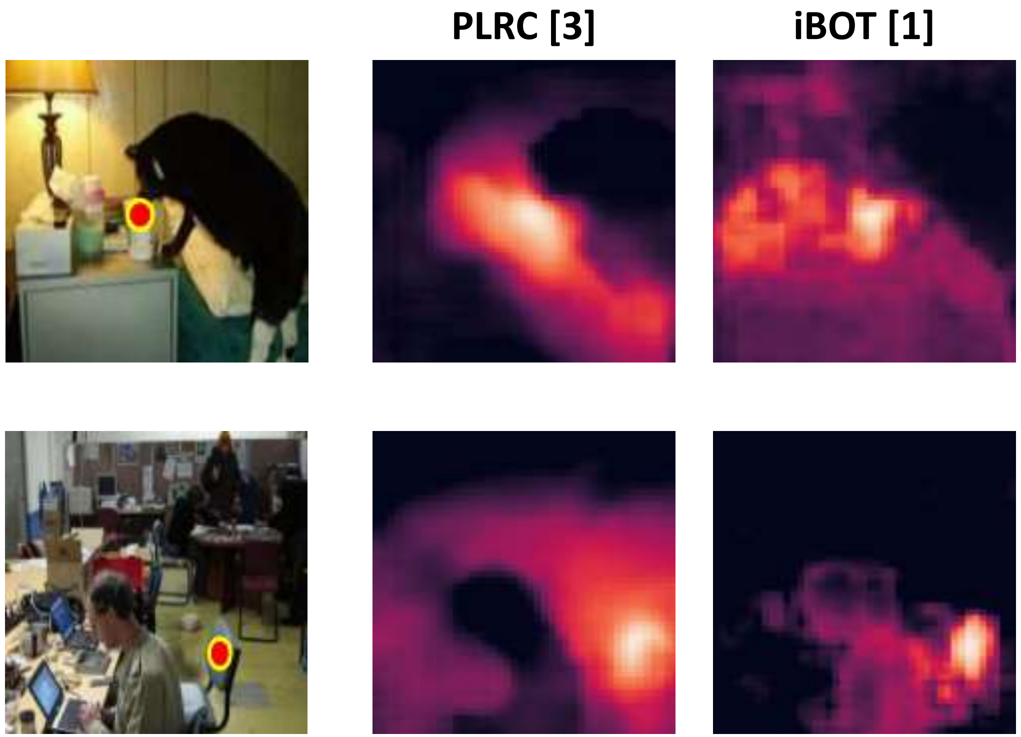


➤ Strong augmentations lead to coupling shortcut in Dense SSL

[1] Zhou, Jinghao, et al. "ibot: Image bert pre-training with online tokenizer." arXiv preprint arXiv:2111.07832 (2021).
[2] Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision." arXiv preprint arXiv:2304.07193 (2023).

Coupling Issue in Dense SSL

Visualization examples of pre-trained models taking coupling shortcut :

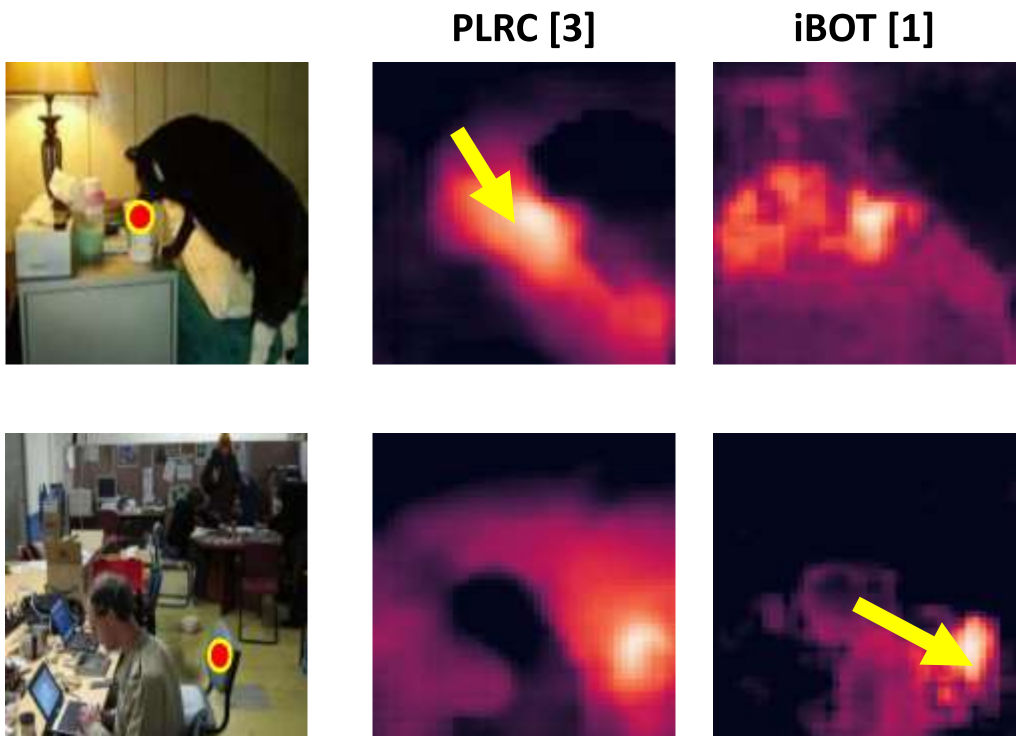


*** We show the visualization examples following [2], the query point-level feature is marked by the red dot**

[3] Bai, Yutong, et al. "Point-level region contrast for object detection pre-training." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

Coupling Issue in Dense SSL

Visualization examples of pre-trained models taking coupling shortcut :

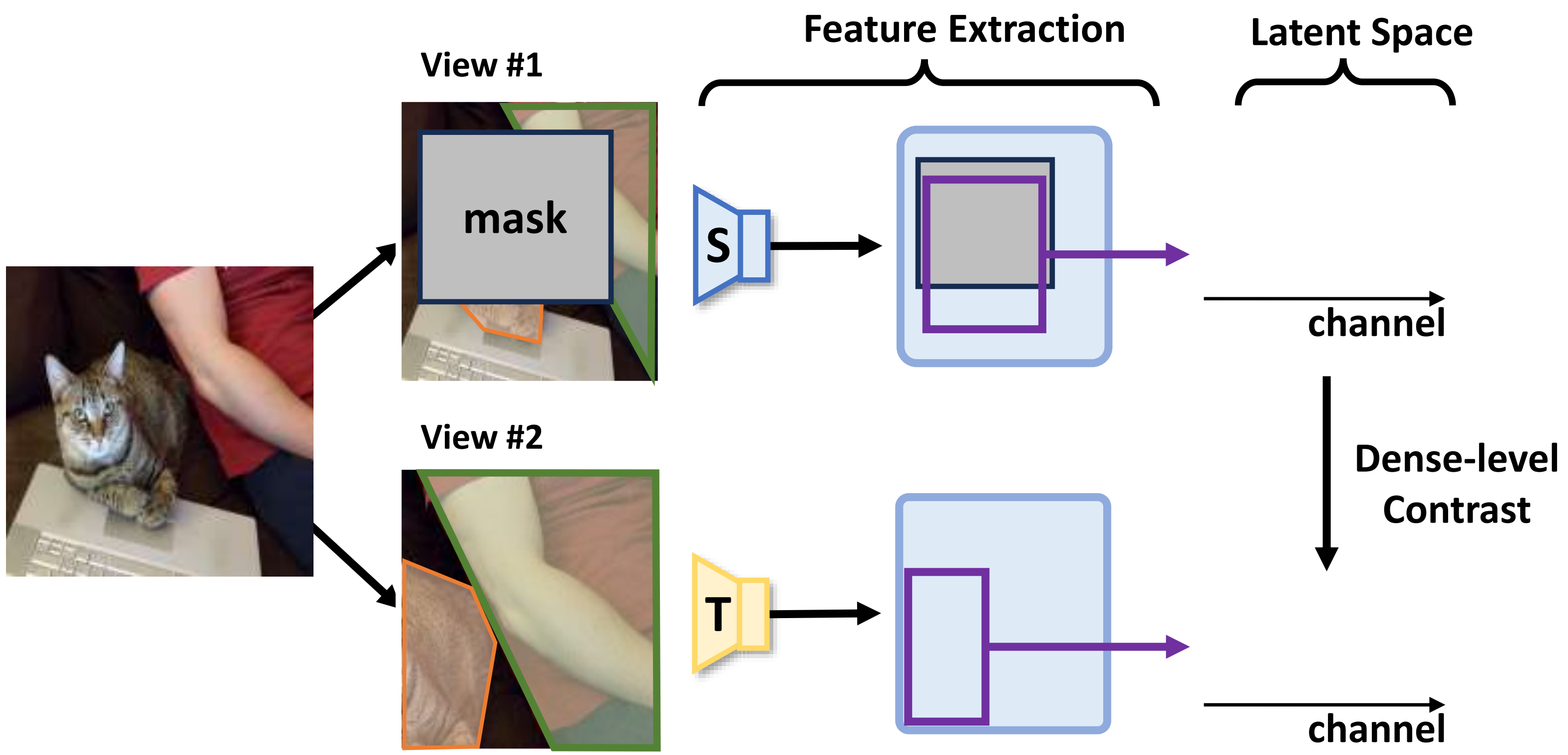


*** We show the visualization examples following [2], the query point-level feature is marked by the red dot**

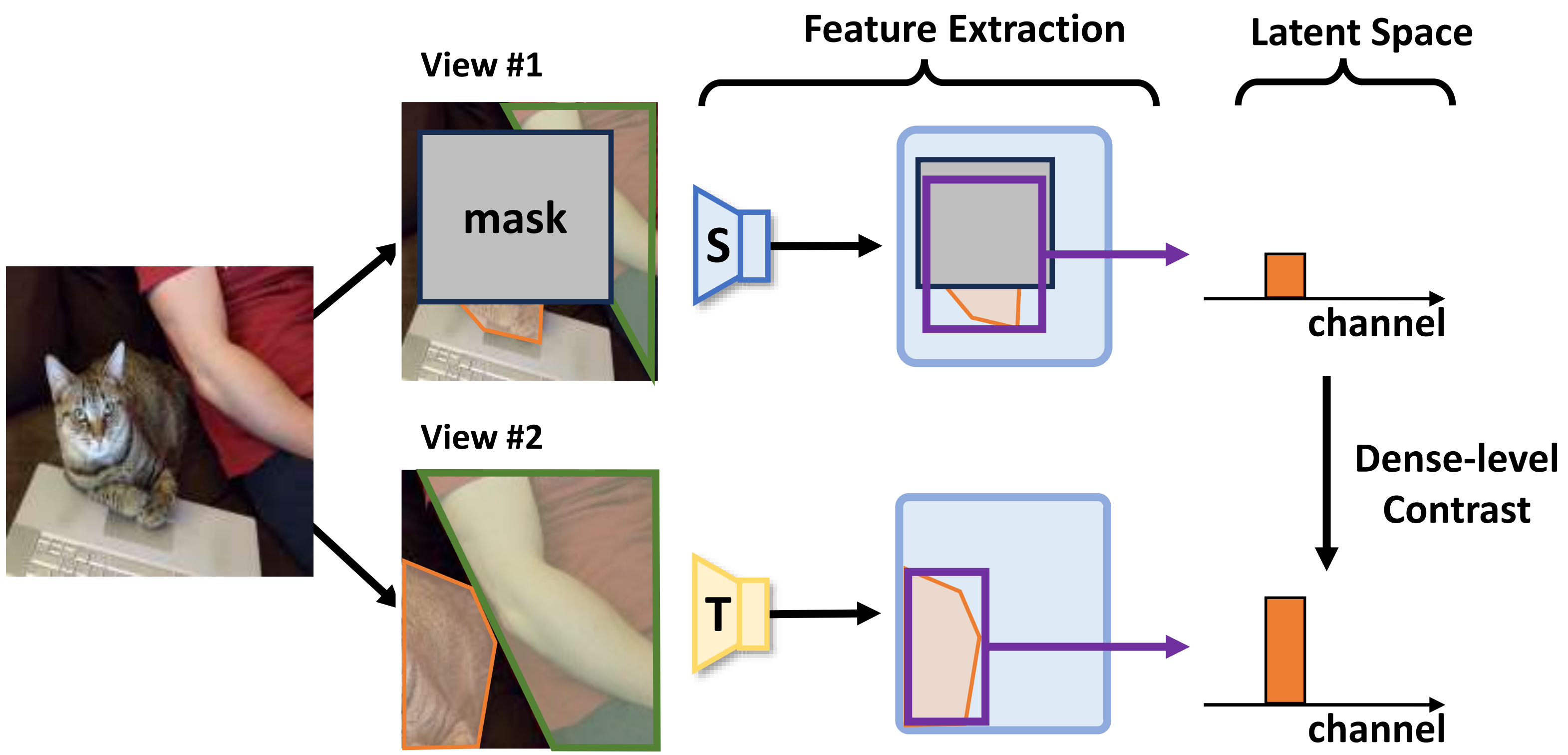
- Dense-level representations are heavily entangled with their surroundings

[3] Bai, Yutong, et al. "Point-level region contrast for object detection pre-training." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

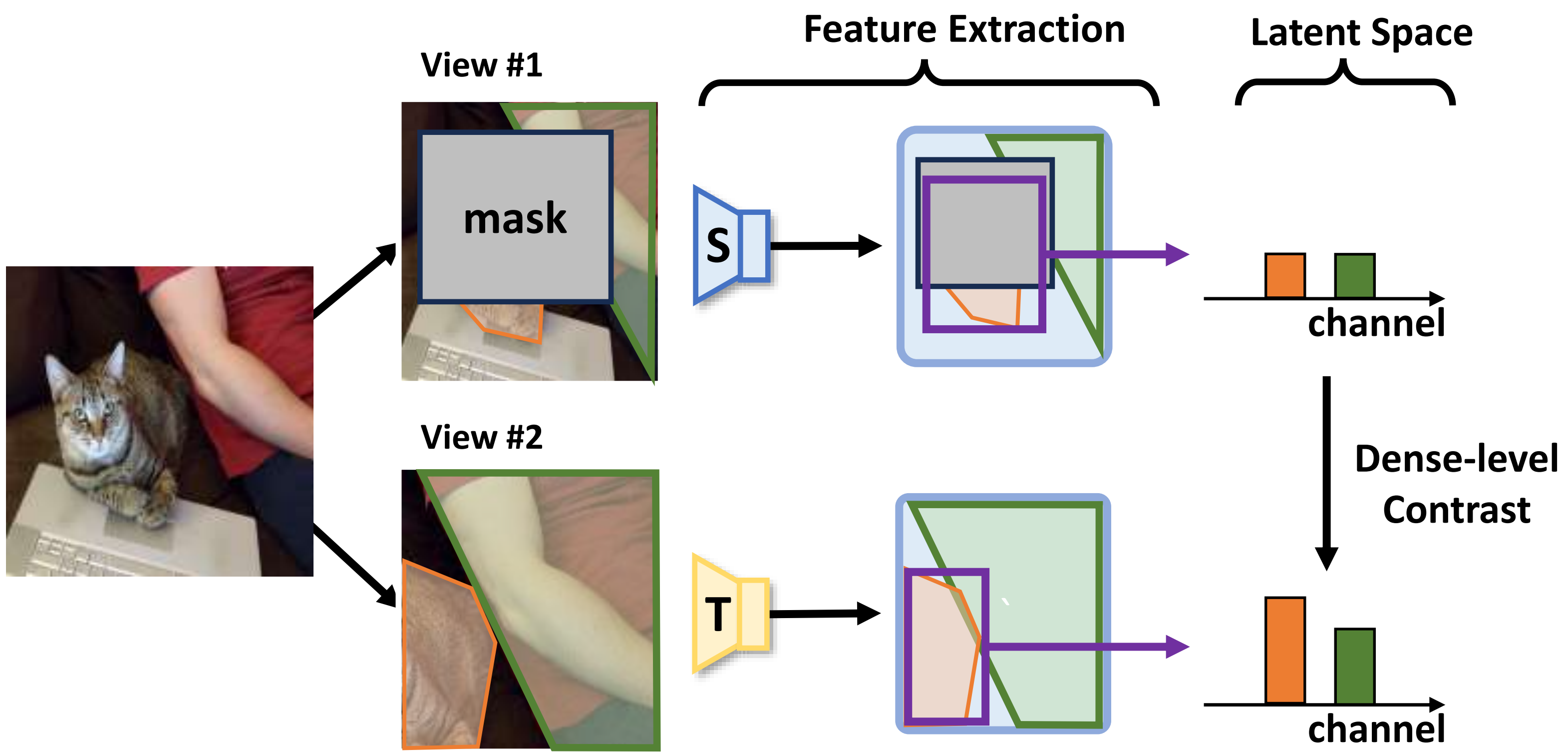
Coupling Issue in Dense SSL



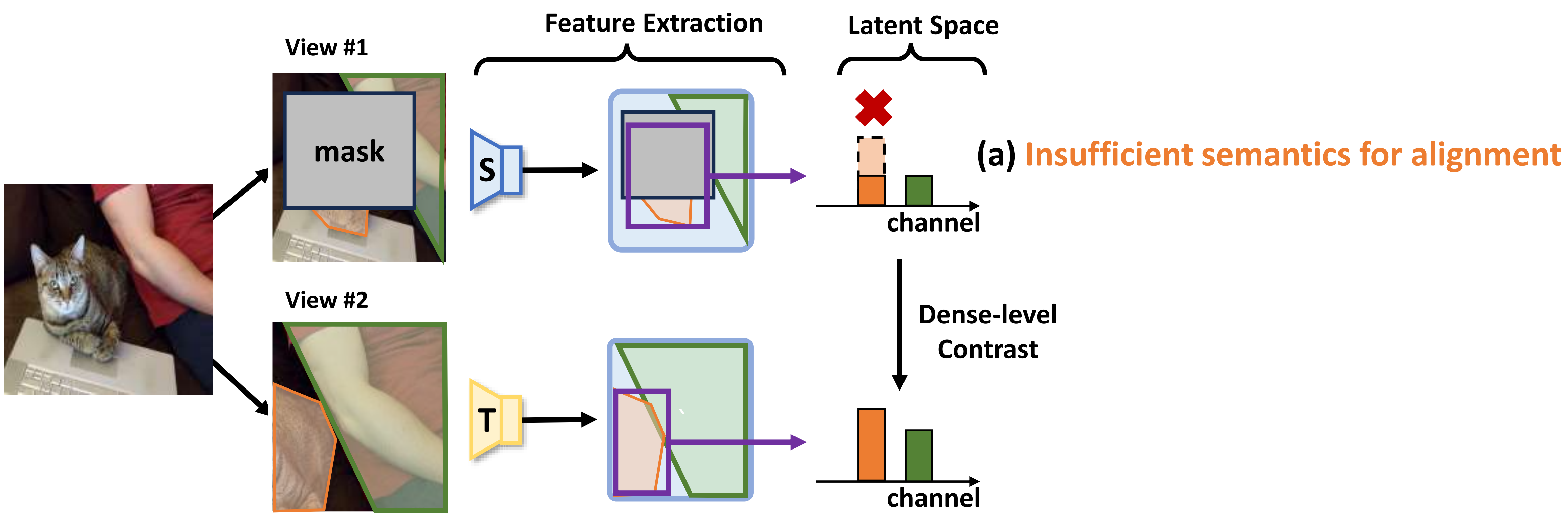
Coupling Issue in Dense SSL



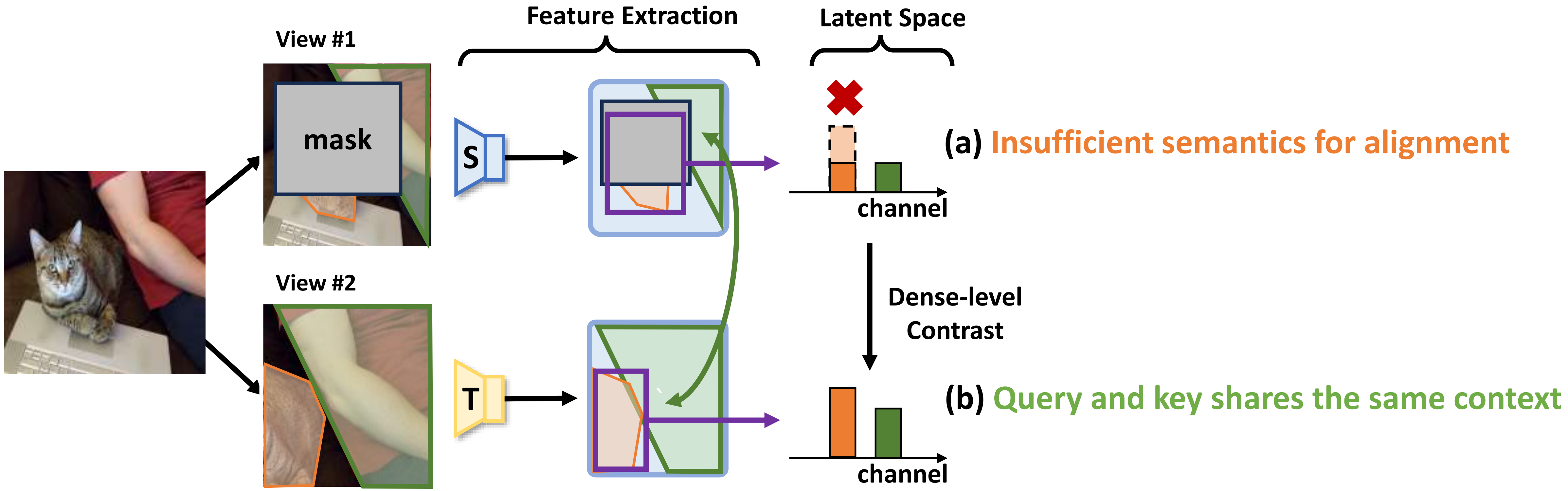
Coupling Issue in Dense SSL



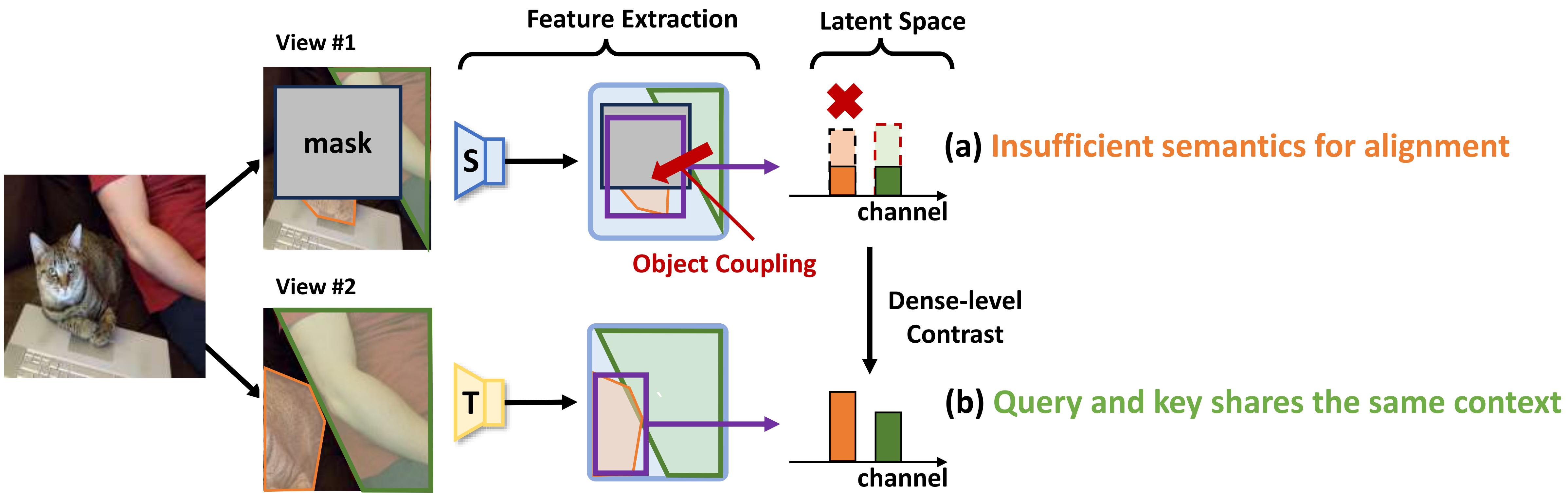
Coupling Issue in Dense SSL



Coupling Issue in Dense SSL

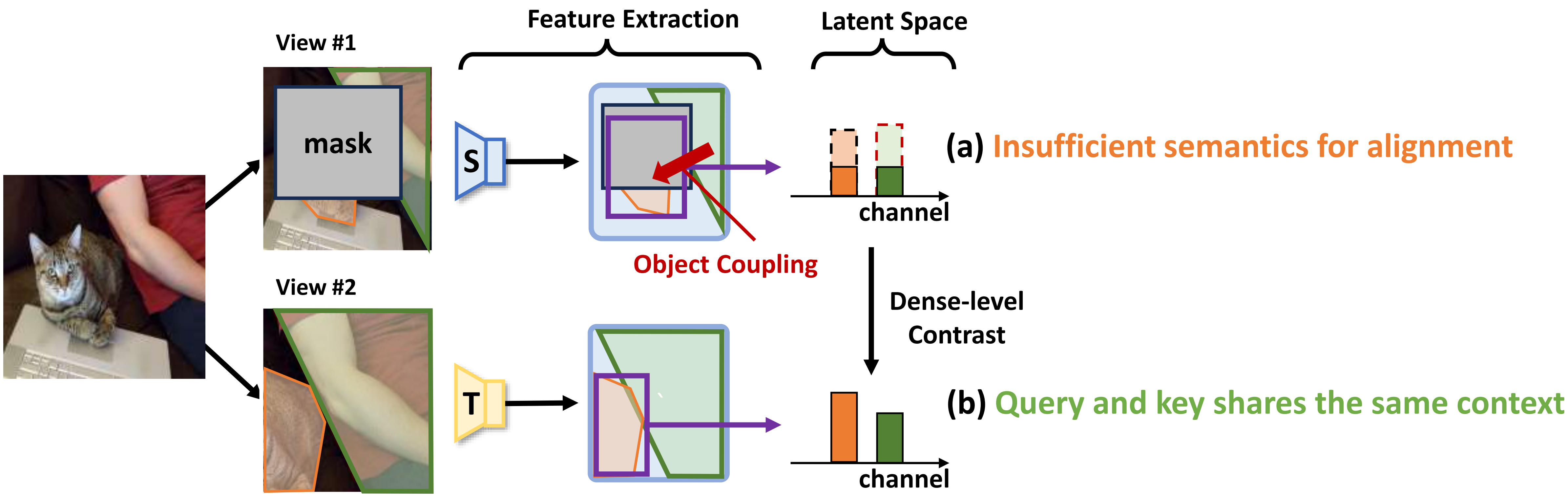


Coupling Issue in Dense SSL



(c) Shared information leaked from the context becomes shortcut for alignment

Coupling Issue in Dense SSL

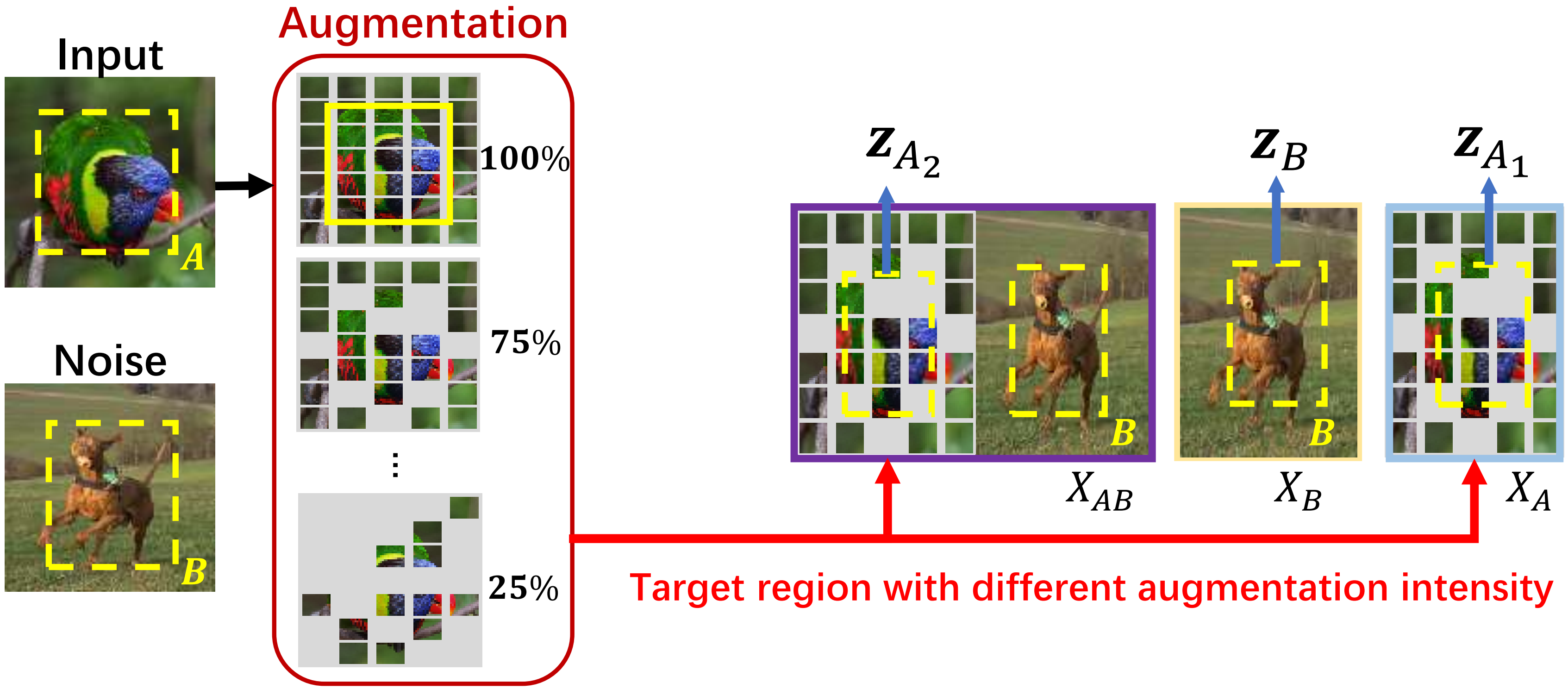


(c) Shared information leaked from the context becomes shortcut for alignment

Coupling

Empirical Study of Coupling

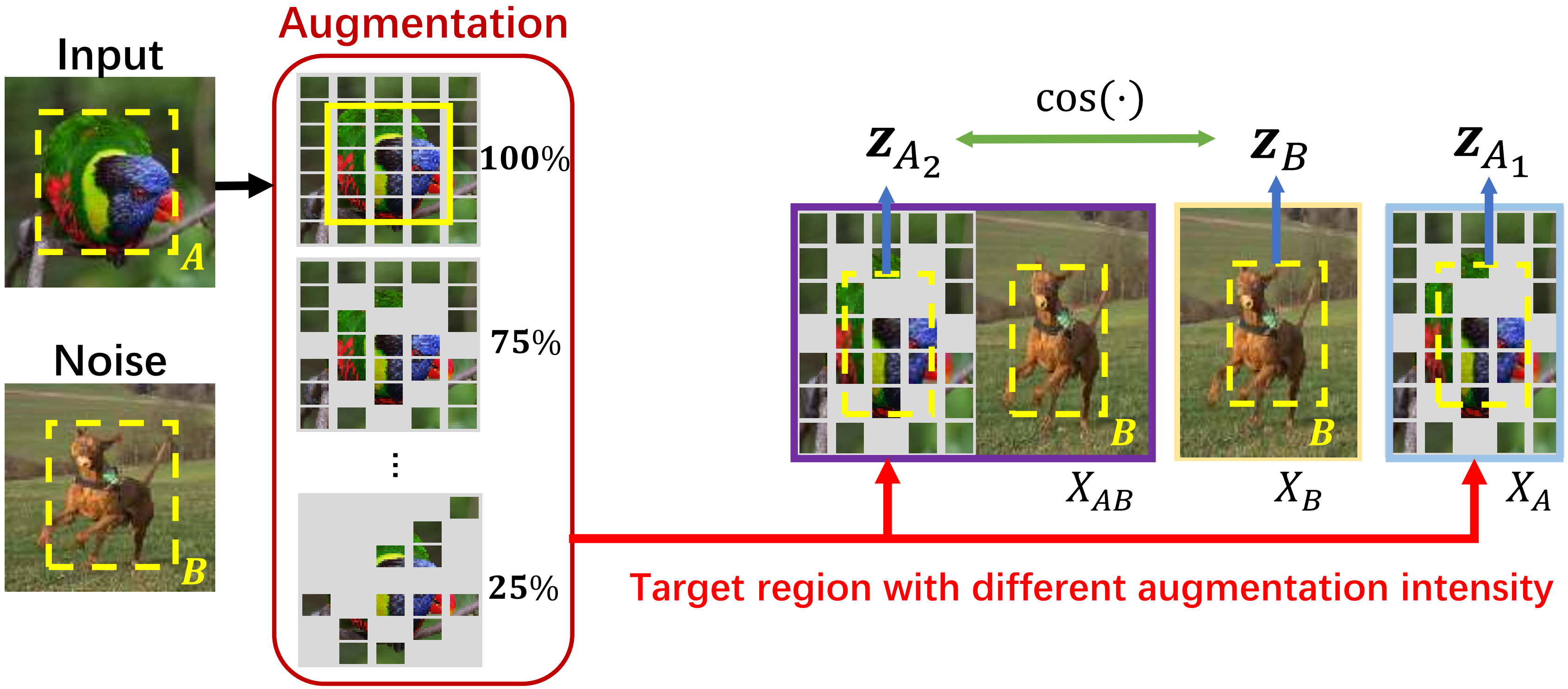
- A pipeline for quantitatively measuring coupling in pre-trained models



$$\boldsymbol{Coupling\ Rate}\ (CR) = \frac{\max\left(\frac{\pi}{2} - \theta\left(\boldsymbol{z}_{A_2}, \boldsymbol{z}_B\right), \epsilon\right)}{\max\left(\frac{\pi}{2} - \theta\left(\boldsymbol{z}_{A_1}, \boldsymbol{z}_B\right), \epsilon\right)}$$

Empirical Study of Coupling

- A pipeline for quantitatively measuring coupling in pre-trained models

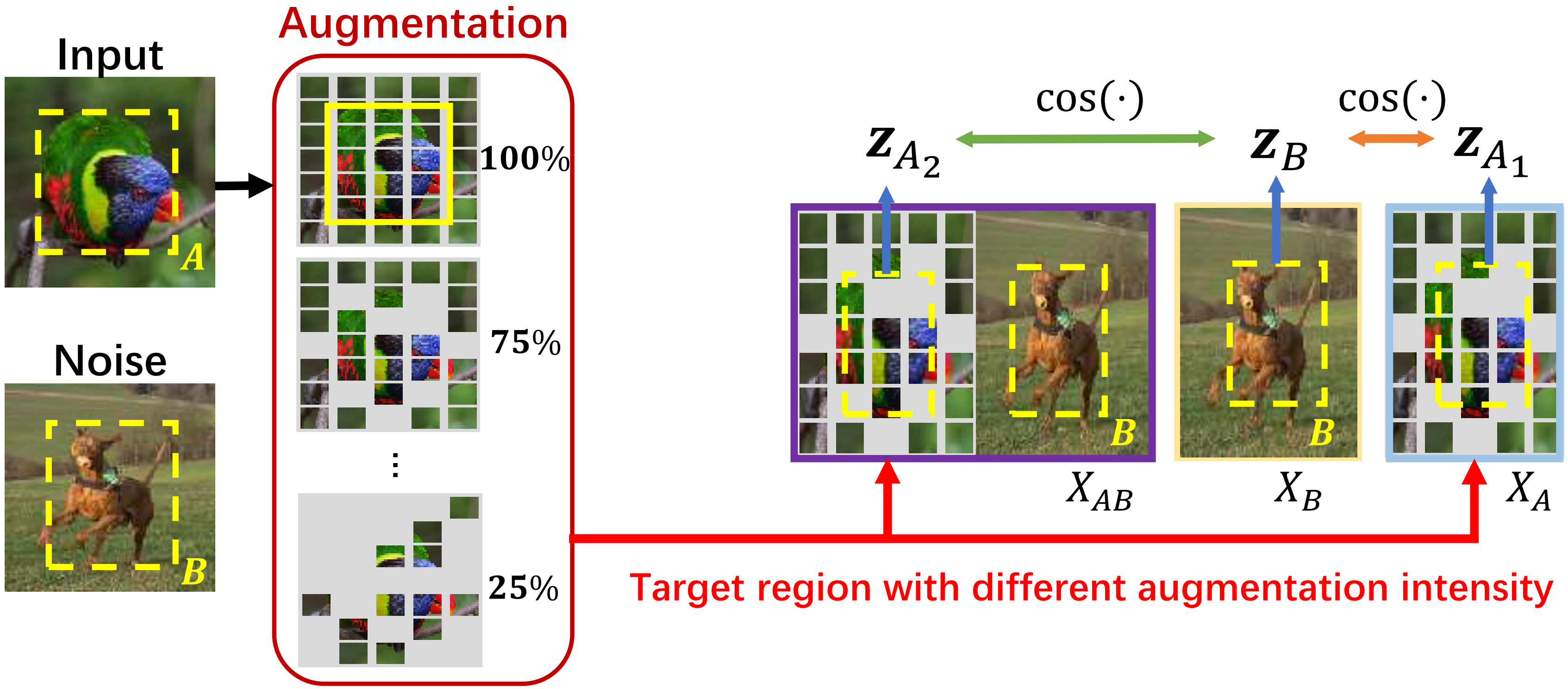


$$\boldsymbol{Coupling\ Rate}\ (CR) = \frac{\max\left(\frac{\pi}{2} - \theta\left(z_{A_2}, z_B\right), \epsilon\right)}{\max\left(\frac{\pi}{2} - \theta\left(z_{A_1}, z_B\right), \epsilon\right)}$$

$\longrightarrow$ Biased similarity when information leaks from B

Empirical Study of Coupling

- A pipeline for quantitatively measuring coupling in pre-trained models



$$\textit{Coupling Rate}\ (CR) = \frac{\max\left(\frac{\pi}{2} - \theta\left(z_{A_2}, z_B\right), \epsilon\right)}{\max\left(\frac{\pi}{2} - \theta\left(z_{A_1}, z_B\right), \epsilon\right)}$$

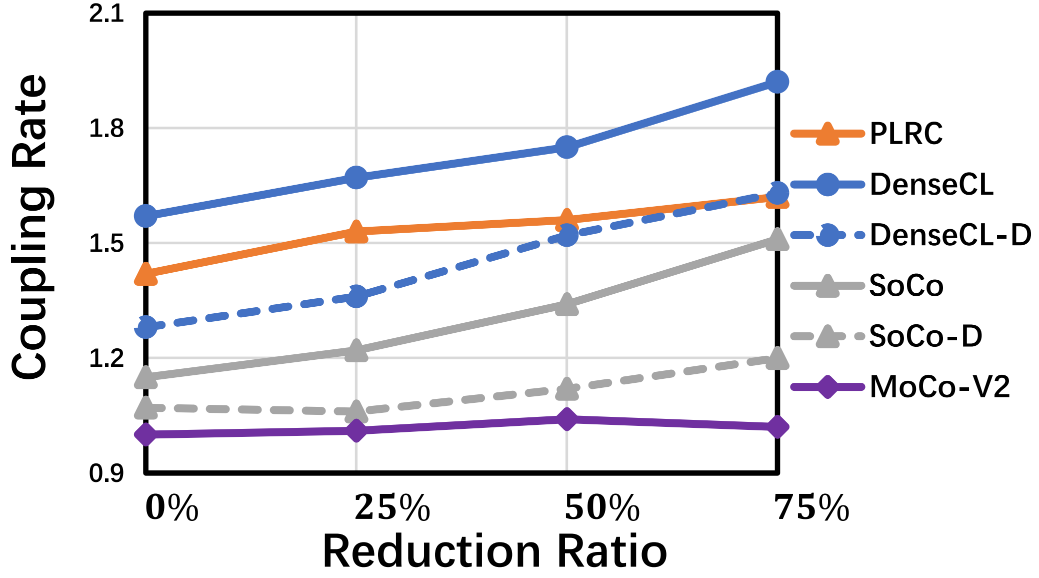
→ Biased similarity when information leaks from B

→ The true correlation between two objects

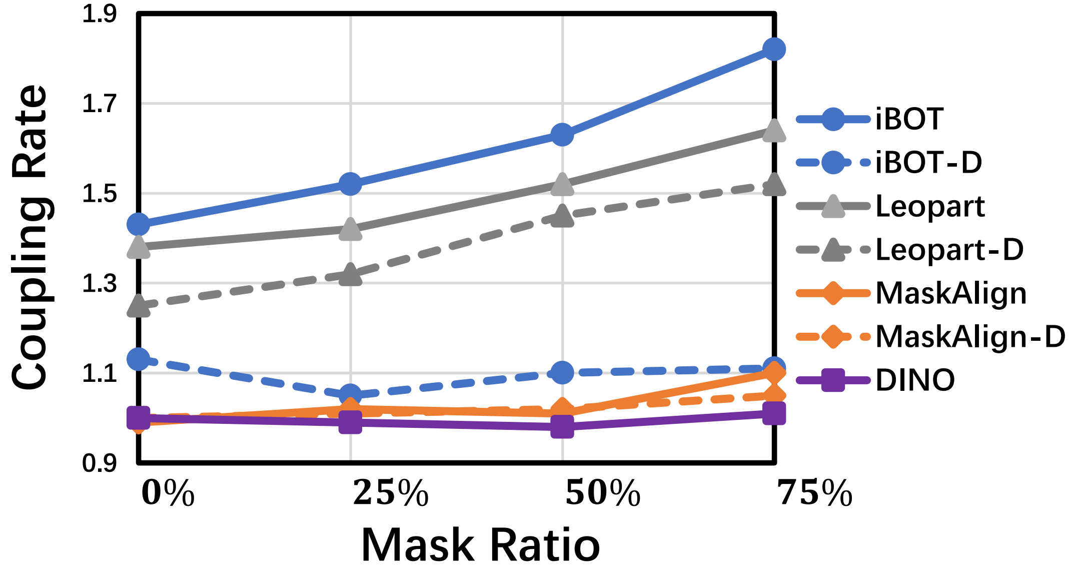
Empirical Study of Coupling

- A pipeline for quantitatively measuring coupling in pre-trained models

(a) CR value on CNNs



(b) CR value on ViTs



*** '-D' denotes models pre-trained with the de-coupling branch**

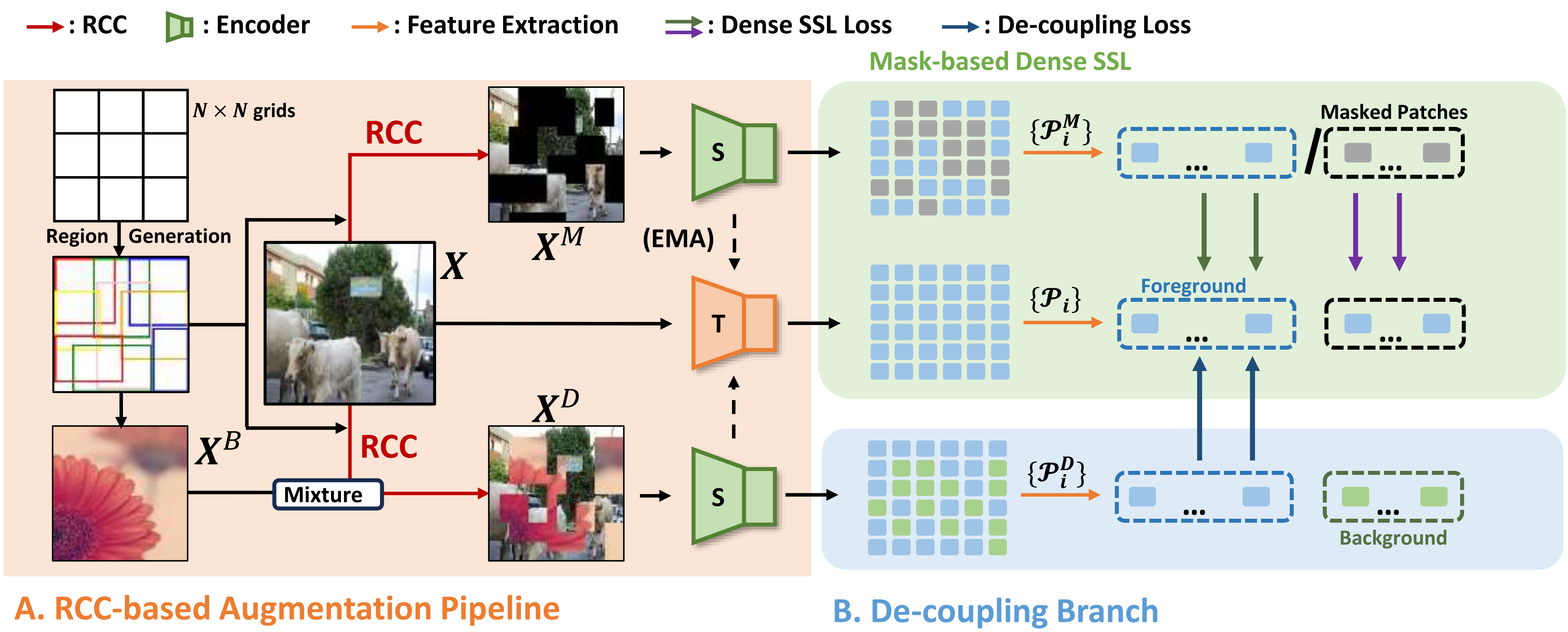
$$\boldsymbol{Coupling\ Rate}\ (CR) = \frac{\max\left(\frac{\pi}{2} - \theta\left(z_{A_2}, z_B\right), \epsilon\right)}{\max\left(\frac{\pi}{2} - \theta\left(z_{A_1}, z_B\right), \epsilon\right)}$$

→ Biased similarity when information leaks from B

→ The true correlation between two objects

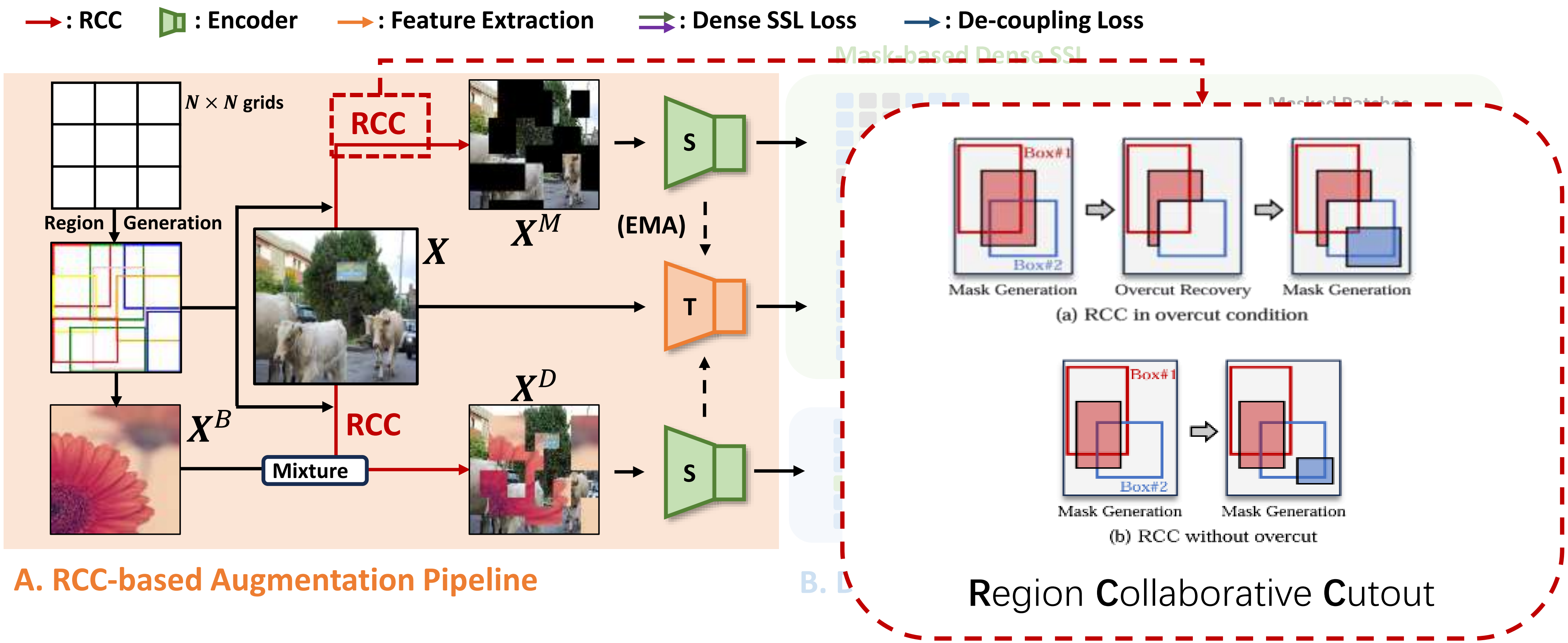
De-coupling Dense-level SSL

- A generalizable de-coupling strategy for dense-level SSL



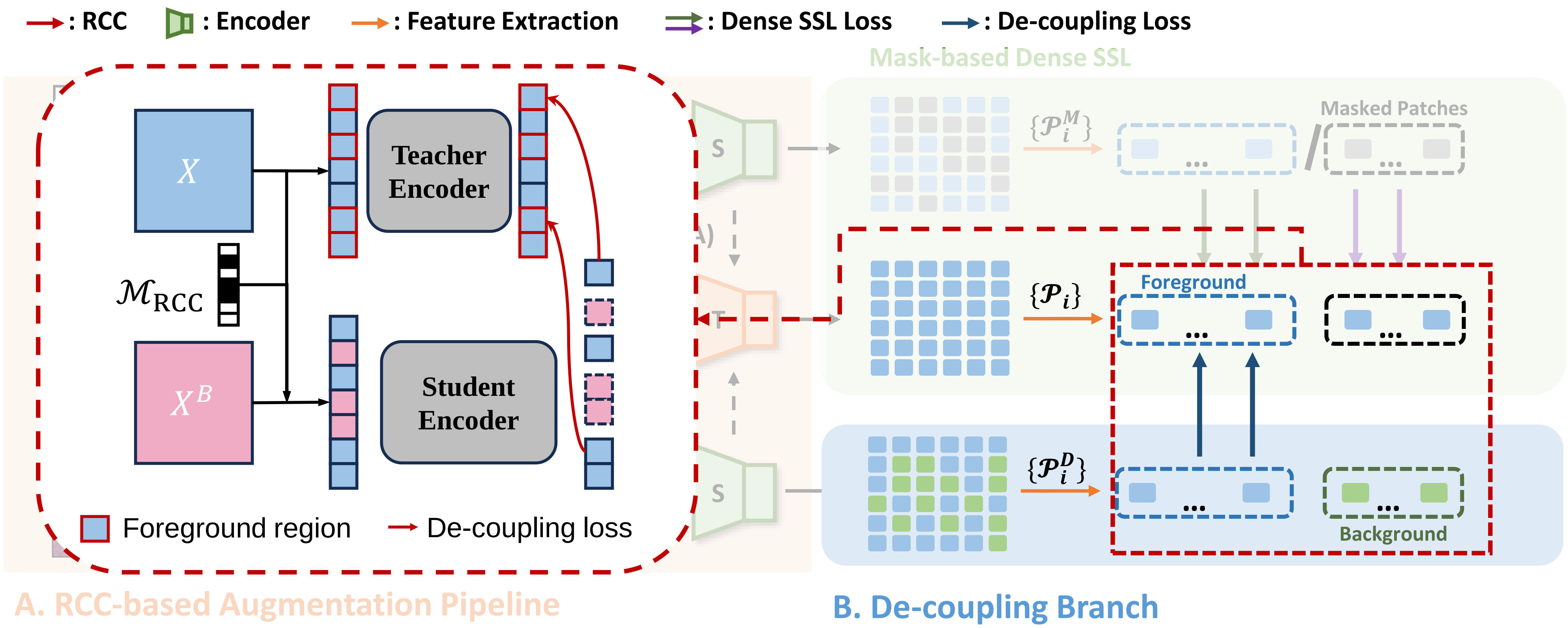
De-coupling Dense-level SSL

- A generalizable de-coupling strategy for dense-level SSL



De-coupling Dense-level SSL

- A generalizable de-coupling strategy for dense-level SSL



Experiments

(a) CNN-based models (ResNet50)

Method	VOC Det.			COCO Det.			COCO ISeg.		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
MoCo† v2	54.6	81.0	60.4	37.8	57.4	41.0	32.9	54.1	35.2
ReSim†	56.6	81.7	63.5	38.3	57.8	41.4	33.5	54.4	35.6
DenseCL	56.7	81.7	63.0	38.5	58.1	41.5	33.6	54.8	35.7
DenseCL-D	<u>57.2</u>	<u>82.2</u>	63.7	39.3	58.7	42.6	34.2	55.7	36.5
PLRC	57.1	82.1	<u>63.8</u>	39.8	<u>59.6</u>	<u>43.7</u>	**35.9**	56.9	**38.6**
SoCo†	56.8	81.7	63.5	38.5	57.9	41.5	33.4	54.6	35.4
SoCo-D	**57.8**	**82.5**	**64.4**	**40.3**	**60.1**	**44.0**	<u>35.1</u>	**56.9**	<u>37.6</u>

(b) ViT-based models (ViT-S)

Method	COCO Det.			COCO ISeg.			ADE Seg.
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	mIoU
iBOT	42.3	61.2	45.6	37.0	58.3	39.4	39.9
iBOT-D	**45.1**	**64.3**	**48.7**	**39.1**	**61.2**	**41.7**	**41.6**
MaskAlign	45.6	65.2	49.7	39.6	62.0	42.4	43.7
MaskAlign-D	**46.7**	**66.4**	**50.5**	**40.5**	**63.2**	**43.5**	**44.3**

(c) Affinity visualization

Experiments

(a) CNN-based models (ResNet50)

Method	VOC Det.			COCO Det.			COCO ISeg.		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
MoCo† v2	54.6	81.0	60.4	37.8	57.4	41.0	32.9	54.1	35.2
ReSim†	56.6	81.7	63.5	38.3	57.8	41.4	33.5	54.4	35.6
DenseCL	56.7	81.7	63.0	38.5	58.1	41.5	33.6	54.8	35.7
DenseCL-D	57.2	82.2	63.7	39.3	58.7	42.6	34.2	55.7	36.5
PLRC	57.1	82.1	63.8	39.8	59.6	43.7	**35.9**	56.9	**38.6**
SoCo†	56.8	81.7	63.5	38.5	57.9	41.5	33.4	54.6	35.4
SoCo-D	**57.8**	**82.5**	**64.4**	**40.3**	**60.1**	**44.0**	35.1	**56.9**	37.6

(b) ViT-based models (ViT-S)

Method	COCO Det.			COCO ISeg.			ADE Seg.
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	mIoU
iBOT	42.3	61.2	45.6	37.0	58.3	39.4	39.9
iBOT-D	**45.1**	**64.3**	**48.7**	**39.1**	**61.2**	**41.7**	**41.6**
MaskAlign	45.6	65.2	49.7	39.6	62.0	42.4	43.7
MaskAlign-D	**46.7**	**66.4**	**50.5**	**40.5**	**63.2**	**43.5**	**44.3**

(c) Affinity visualization

Models with the de-coupling strategy

- learns dense semantics more efficiently and achieves better dense prediction performance

Experiments

(a) CNN-based models (ResNet50)

Method	VOC Det.			COCO Det.			COCO ISeg.		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
MoCo† v2	54.6	81.0	60.4	37.8	57.4	41.0	32.9	54.1	35.2
ReSim†	56.6	81.7	63.5	38.3	57.8	41.4	33.5	54.4	35.6
DenseCL	56.7	81.7	63.0	38.5	58.1	41.5	33.6	54.8	35.7
DenseCL-D	57.2	82.2	63.7	39.3	58.7	42.6	34.2	55.7	36.5
PLRC	57.1	82.1	63.8	39.8	59.6	43.7	**35.9**	56.9	**38.6**
SoCo†	56.8	81.7	63.5	38.5	57.9	41.5	33.4	54.6	35.4
SoCo-D	**57.8**	**82.5**	**64.4**	**40.3**	**60.1**	**44.0**	35.1	**56.9**	37.6

(b) ViT-based models (ViT-S)

Method	COCO Det.			COCO ISeg.			ADE Seg.
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	mIoU
iBOT	42.3	61.2	45.6	37.0	58.3	39.4	39.9
iBOT-D	**45.1**	**64.3**	**48.7**	**39.1**	**61.2**	**41.7**	**41.6**
MaskAlign	45.6	65.2	49.7	39.6	62.0	42.4	43.7
MaskAlign-D	**46.7**	**66.4**	**50.5**	**40.5**	**63.2**	**43.5**	**44.3**

(c) Affinity visualization

Models with the de-coupling strategy

- learns dense semantics more efficiently and achieves better dense prediction performance
- acquires dense-level representations with higher consistency with the object regions

Thank You !

https://openreview.net/forum?id=WQYHbr36Fo

Q&A: qiucongpei@gmail.com