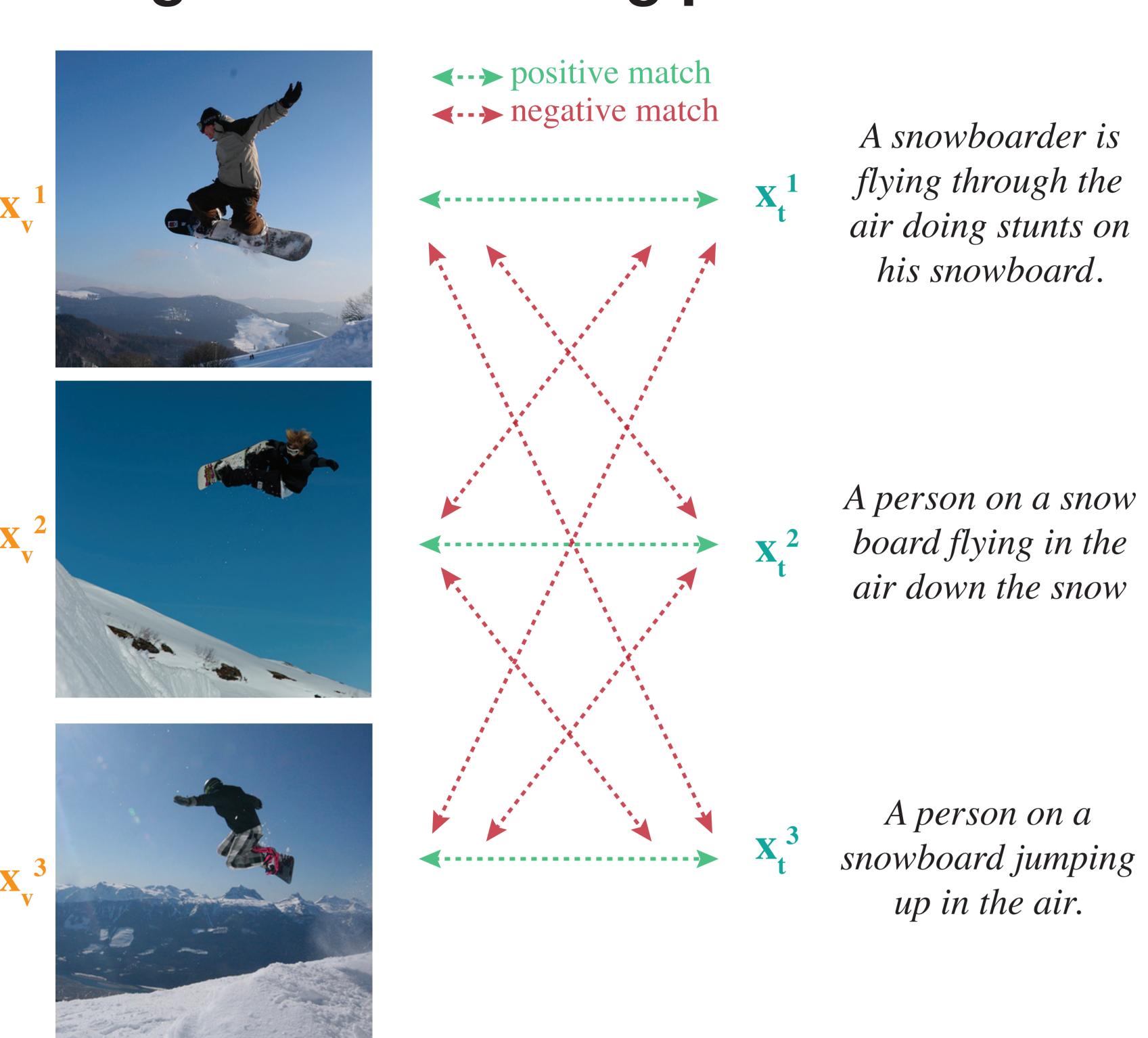
Improved Probabilistic Image-Text Representations

Sanghyuk Chun



Motivation: Inherent ambiguity in image-text matching problem



There are many false negatives (FNs) in dataset, and it naturally leads to ambiguous supervision to the model

Image

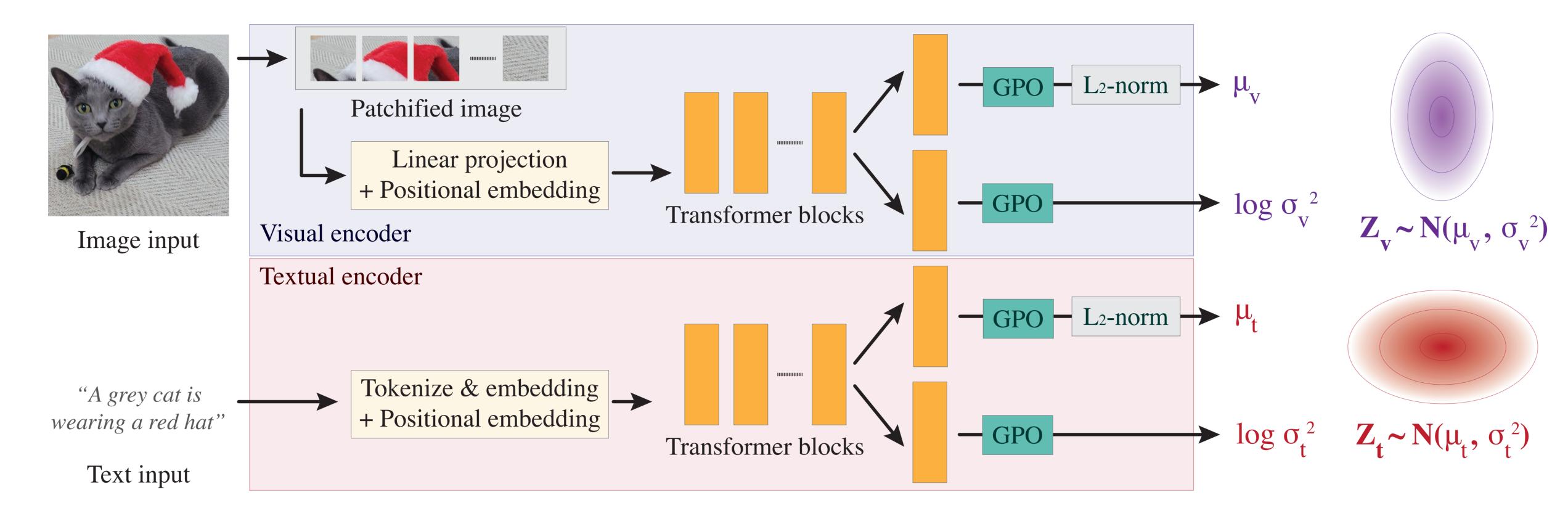
Deterministic embedding space

Match

Caption

Approximating the textual embeddings as the same embedding point \mathbf{z}_{i} . The match between z_{i}^{1} and z_{i}^{\prime} become "uncertain" → Visual/textual embeddings will be located closely as they are semantically almost same. Image-text match (ITM) supervision becomes uncertain (either matched or not), but a deterministic embedding space cannot capture uncertainty. → The existing deterministic approaches enforces to the textual embeddings to be different by making \mathbf{z}_{i}^{1} differs from others using HNM, although they are semantically the same. Probabilistic embedding space If two distributions have a certain match, an embedding has a low uncertainty value. If the match between \mathbb{Z}_{i}^{1} and \mathbb{Z}_{i}^{\prime} become "uncertain", the embedding uncertainty (i.e., variance becomes larger) → A probabilistic embedding space captures uncertainty of noisy ITM supervision. If a match is uncertain, then the variance of embeddings becomes larger (Z_t^1, Z_t^2, Z_t^3) , if a match is certain, then the variance becomes smaller (\mathbf{Z}_{\cdot}^{4})

PCME++ An improved probabilistic VL model



How can we train a proper probabilistic embedding space? In other words, how to measure a distance between two distributions? Here, we expect three properties for the probabilistic embeddings:

- 1) There exists a proper probabilistic distance between two distributions
- 2) If the match between two distributions is certain, then the distributions should have small variances
- 3) If the match between two distributions is uncertain, then the variances should be large.

Popular distances such as Wasserstien dist, KL divergence cannot satisfy (2).

Monte-Carlo approximiation of Monte-Carlo approximiation of matching probability (Chun et al. 2021) $p_{ heta}(m|x_{lpha},x_{eta})pprox rac{1}{J^2}\sum s(-a\|z_{lpha}-z_{eta}\|_2+b)$

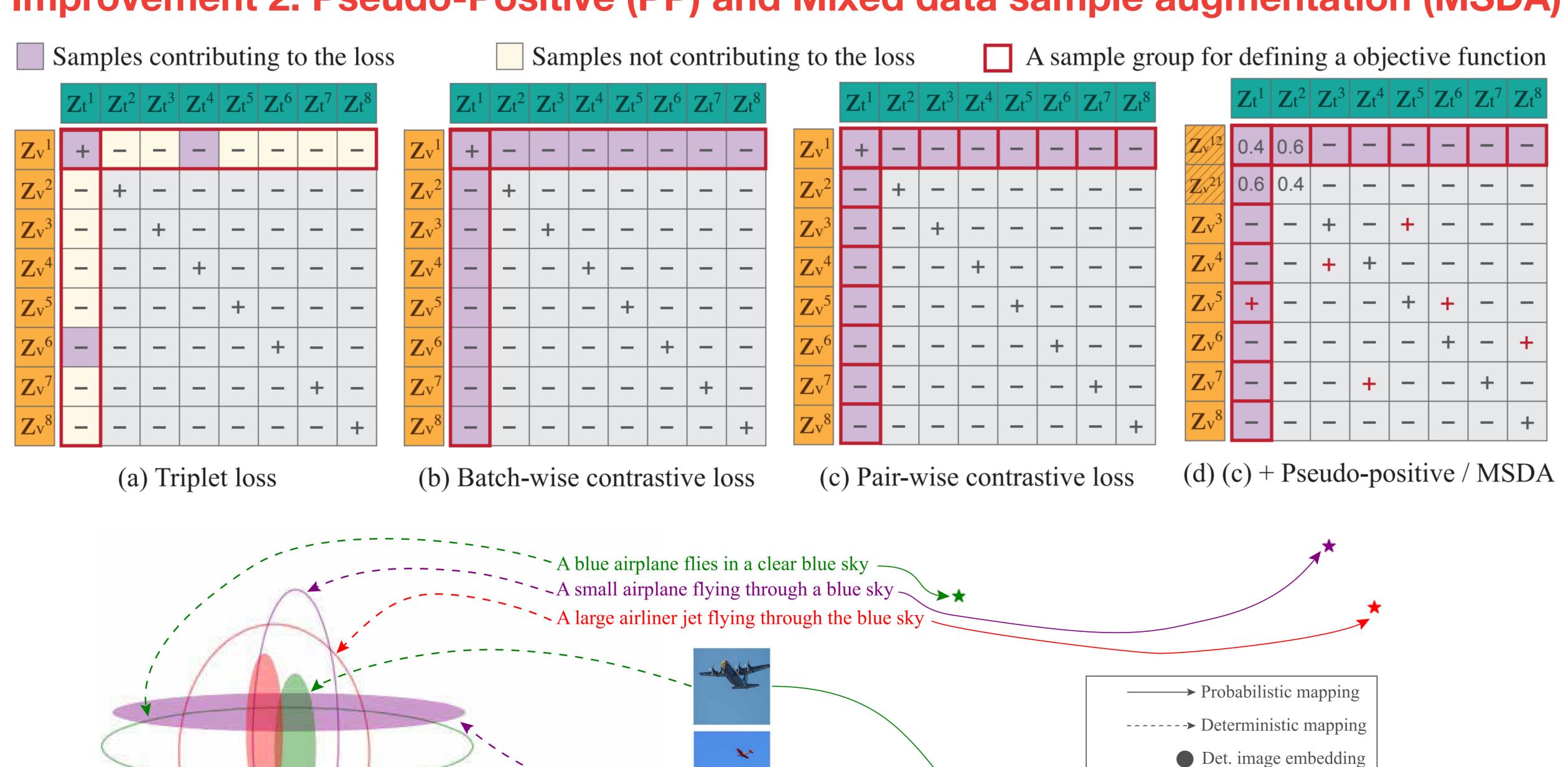
It satisfies all conditions, but the MC sampling itself has flaws in computation and accuracy

Improvement 1: Closed-form Sampled Distance (CSD)

$$d(\mathbf{Z}_{v}, \mathbf{Z}_{t}) = \mathbb{E}_{\mathbf{Z}_{v}, \mathbf{Z}_{t}} \|\mathbf{Z}_{v} - \mathbf{Z}_{t}\|_{2}^{2} = \|\mu_{v} - \mu_{t}\|_{2}^{2} + \|\sigma_{v}^{2} + \sigma_{t}^{2}\|_{1}$$

 $\mathcal{L}_{\text{match}} = -m_{vt} \log \operatorname{sigmoid}(-a \cdot d(\mathbf{Z}_v, \mathbf{Z}_t) + b) - (1 - m_{vt}) \log \operatorname{sigmoid}(a \cdot d(\mathbf{Z}_v, \mathbf{Z}_t) - b)$

Improvement 2: Pseudo-Positive (PP) and Mixed data sample augmentation (MSDA)



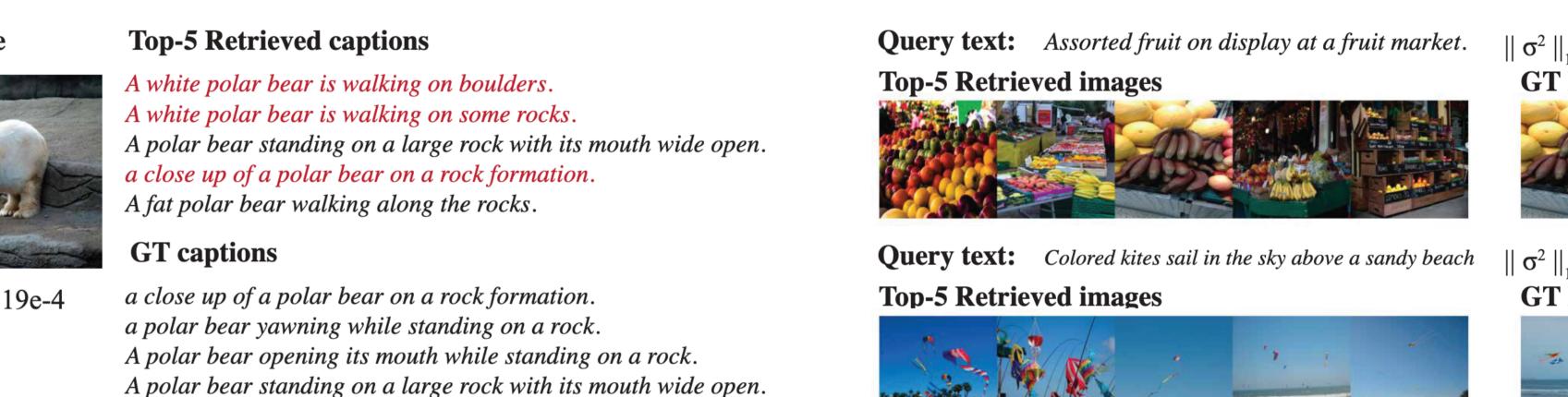
Experimental results

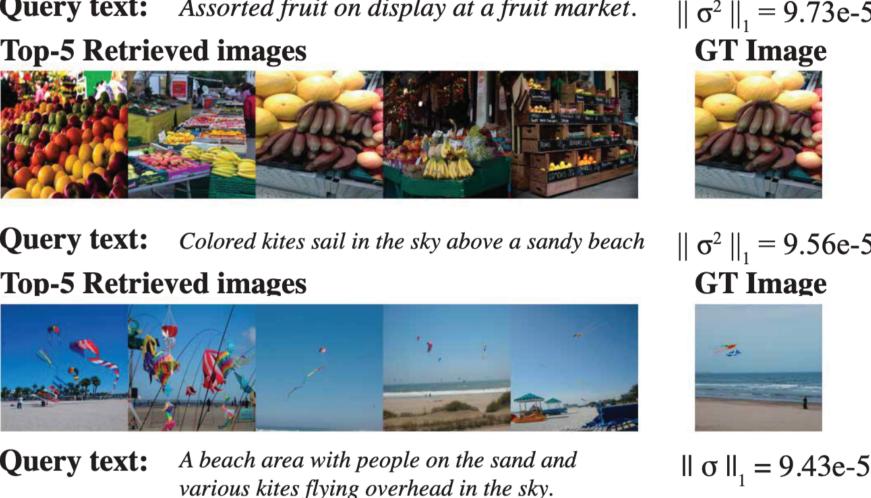
COCO Caption results

COCO Noisy correspondence results

3.6.1.1	D 10		-						ECCV	V Caption	on	COC	CO
Method	Prob?	mAP@R	R-P	R@I	5K R@1	RSUM	Noise	Method	mAP@R	R-P	R@1	5K R@1	RSUM
CLIP ZS^{\dagger}	×	26.8	36.9	67.1	40.3	471.9		VSE∞	37.0	46.3	79 7	51.8	518.6
$VSE\infty$	X	<u>40.0</u>	<u>49.5</u>	83.1	55.2	<u>536.5</u>	20%						$\frac{310.8}{212.8}$
P2RM	X	39.0	48.7	82.0	51.7	530.2							
DAA	X	39.2	49.0	82.0	52.9	530.9							514.6
InfoNCE	X	39.0	48.7	81.7	53.0	532.6		PCME	<u>37.6</u>	47.6	79.2	48.7	520.7
PCME	/	39.1	48.9	81.4	53.0	532.0		PCME++ (ours)	37.7	47.6	80.0	<u>50.4</u>	524.6
PCME++ (μ only)	X	39.5	49.1	82.7	55.2	536.2		NCR [†]	35.9	46.0	78.0	48.8	518.6
	/							DECL^\dagger	_	_	_	49.4	518.2
PCME++ (SWA)		40.2	49.8	82.9	55.2	537.3			18.0	28.5	13.7		394.1
CLIP ZS [†]	×	28.0	37.8	72.2	46.4	491.6	50%						499.5
$VSE\infty$	X	20.2	31.5	46.2	22.7	424.3							
	X												508.0
	/							PCME++ (ours)	35.7	45.8	76.3	<u>45.5</u>	<u>511.0</u>
PCME++	~	42.1	50.8	88.8	64.3	554.7		NCR [†]	34.0	44.3	75.1	<u>45.5</u>	508.5
	$VSE\infty$ $P2RM$ DAA $InfoNCE$ $PCME$ $PCME++ (\mu \text{ only})$ $PCME++$ $PCME++ (SWA)$ $CLIP ZS^{\dagger}$ $VSE\infty$ $InfoNCE$ $PCME$	$\begin{array}{cccccccccccccccccccccccccccccccccccc$	Method Prob? mAP@R CLIP ZS† \times 26.8 VSE∞ \times 40.0 P2RM \times 39.0 DAA \times 39.2 InfoNCE \times 39.0 PCME \checkmark 39.1 PCME++ (μ only) \times 39.5 PCME++ \checkmark 40.1 PCME++ (SWA) \checkmark 40.2 CLIP ZS† \times 28.0 VSE∞ \times 20.2 InfoNCE \times 35.6 PCME \checkmark 41.2	Method Prob? mAP@R R-P CLIP ZS† \times 26.8 36.9 VSE∞ \times 40.0 49.5 P2RM \times 39.0 48.7 DAA \times 39.2 49.0 InfoNCE \times 39.0 48.7 PCME \checkmark 39.1 48.9 PCME++ (μ only) \times 39.5 49.1 PCME++ \checkmark 40.1 49.7 PCME++ (SWA) \checkmark 40.2 49.8 CLIP ZS† \times 28.0 37.8 VSE∞ \times 20.2 31.5 InfoNCE \times 35.6 45.8 PCME \checkmark 41.2 50.3	CLIP ZS^{\dagger} X 26.8 36.9 67.1 VSE ∞ X 40.0 49.5 83.1 P2RM X 39.0 48.7 82.0 DAA X 39.2 49.0 82.0 InfoNCE X 39.0 48.7 81.7 PCME Y 39.1 48.9 81.4 PCME++ (μ only) X 39.5 49.1 82.7 PCME++ Y 40.1 49.7 83.1 PCME++ (SWA) Y 40.2 49.8 82.9 CLIP ZS † X 28.0 37.8 72.2 VSE ∞ X 20.2 31.5 46.2 InfoNCE X 35.6 45.8 75.6 PCME Y 41.2 50.3 86.0	Method Prob? mAP@R R-P R@1 5K R@1 CLIP ZS† χ 26.8 36.9 67.1 40.3 VSE∞ χ 40.0 49.5 83.1 55.2 P2RM χ 39.0 48.7 82.0 51.7 DAA χ 39.2 49.0 82.0 52.9 InfoNCE χ 39.0 48.7 81.7 53.0 PCME γ 39.1 48.9 81.4 53.0 PCME++ (μ only) χ 39.5 49.1 82.7 55.2 PCME++ γ 40.1 49.7 83.1 55.1 PCME++ (SWA) γ 40.2 49.8 82.9 55.2 CLIP ZS† χ 28.0 37.8 72.2 46.4 VSE∞ χ 20.2 31.5 46.2 22.7 InfoNCE χ 35.6 45.8 75.6 45.9 PCME ψ 41.2	Method Prob? mAP@R R-P R@1 5K R@1 RSUM CLIP ZS† \times 26.8 36.9 67.1 40.3 471.9 VSE∞ \times 40.0 49.5 83.1 55.2 536.5 P2RM \times 39.0 48.7 82.0 51.7 530.2 DAA \times 39.2 49.0 82.0 52.9 530.9 InfoNCE \times 39.0 48.7 81.7 53.0 532.6 PCME \checkmark 39.1 48.9 81.4 53.0 532.0 PCME++ (μ only) \times 39.5 49.1 82.7 55.2 536.2 PCME++ \checkmark 40.1 49.7 83.1 55.1 537.0 PCME++ (SWA) \checkmark 40.2 49.8 82.9 55.2 537.3 CLIP ZS† \times 28.0 37.8 72.2 46.4 491.6 VSE∞ \times 20.2 31.5	Method Prob? mAP@R R-P R@1 5K R@1 RSUM Noise CLIP ZS† \times 26.8 36.9 67.1 40.3 471.9 VSE∞ \times 40.0 49.5 83.1 55.2 536.5 P2RM \times 39.0 48.7 82.0 51.7 530.2 DAA \times 39.2 49.0 82.0 52.9 530.9 InfoNCE \times 39.0 48.7 81.7 53.0 532.6 PCME \checkmark 39.1 48.9 81.4 53.0 532.0 PCME++ (μ only) \times 39.5 49.1 82.7 55.2 536.2 PCME++ \checkmark 40.1 49.7 83.1 55.1 537.0 PCME++ (SWA) \checkmark 40.2 49.8 82.9 55.2 537.3 CLIP ZS† \times 28.0 37.8 72.2 46.4 491.6 VSE∞ \times 20.2	Method Prob? mAP@R R-P R@1 5K R@1 RSUM Noise Method CLIP ZS† \times 26.8 36.9 67.1 40.3 471.9 VSE ∞ \times 40.0 49.5 83.1 55.2 536.5 DAA DAA \times 39.0 48.7 82.0 51.7 530.2 530.9 DAA InfoNCE PCME \times 39.0 48.7 81.7 53.0 532.6 PCME PCME \times 39.0 48.7 81.7 53.0 532.6 PCME PCME \times 39.1 48.9 81.4 53.0 532.6 PCME PCME++ (ours) NCR† PCME++ (ours) NCR† DECL* NCR† DECL† PCME++ (ours) NCR† DECL† VSE ∞ NCR† DECL† VSE ∞ InfoNCE YSE ∞ InfoNCE PCME PCME	Method Prob? mAP@R R-P R@1 5K R@1 RSUM Noise Method mAP@R CLIP ZS [†] X 26.8 36.9 67.1 40.3 471.9 VSE∞ 37.0 VSE∞ X 40.0 49.5 83.1 55.2 536.5 DAA 6.7 P2RM X 39.0 48.7 82.0 51.7 530.2 DAA 6.7 InfoNCE X 39.0 48.7 81.7 53.0 532.6 PCME 9CME 37.6 PCME Y 39.1 48.9 81.4 53.0 532.0 PCME++ (ours) 37.7 PCME++ Y 40.1 49.7 83.1 55.1 537.0 PCME++ (ours) NCR [†] 35.9 PCME++ (SWA) Y 40.2 49.8 82.9 55.2 537.3 VSE∞ NCR [†] 18.0 CLIP ZS [†] X 28.0 37.8 72.2 46.4 491.6 YSE∞ <td>Method Prob? mAP@R R-P R@1 5K R@1 RSUM Noise Method mAP@R R-P CLIP ZS† X 26.8 36.9 67.1 40.3 471.9 VSE∞ 37.0 46.3 VSE∞ X 40.0 49.5 83.1 55.2 536.5 DAA 6.7 12.5 PRM X 39.0 48.7 82.0 51.7 530.2 DAA 6.7 12.5 DAA X 39.0 48.7 81.7 53.0 532.6 PCME 35.9 46.3 PCME X 39.0 48.7 81.7 53.0 532.6 PCME 37.6 47.6 PCME++ (µ only) X 39.5 49.1 82.7 55.2 536.2 PCME++ (ours) 37.7 47.6 PCME++ (SWA) Y 40.1 49.7 83.1 55.1 537.0 DECL† - - - CLIP ZS† X 28.0</td> <td>Method Prob? mAP@R R-P R@1 SK R@1 RSUM Noise Method mAP@R R-P R@1 CLIP ZS† X 26.8 36.9 67.1 40.3 471.9 VSE∞ 37.0 46.3 79.7 VSE∞ X 40.0 49.5 83.1 55.2 536.5 DAA 6.7 12.5 18.5 P2RM X 39.0 48.7 82.0 51.7 530.2 DAA 6.7 12.5 18.5 DAA 39.0 48.7 81.7 53.0 532.6 PCME 35.9 46.3 76.1 PCME 39.0 48.7 81.7 53.0 532.6 PCME 37.6 47.6 79.2 PCME+ (μ only) X 39.5 49.1 82.7 55.2 537.0 NCR† 35.9 46.0 78.0 PCME++ (SWA) Y 40.1 49.7 83.1 55.1 537.0 55.2 537.3 YS</td> <td>Method Prob? mAP@R R-P R@1 5K R@1 RSUM Noise Method mAP@R R-P R@1 5K R@1 CLIP ZS† X 26.8 36.9 67.1 40.3 471.9 VSE∞ X 40.0 49.5 83.1 55.2 536.5 P2RM X 39.0 48.7 82.0 51.7 530.2 DAA X 39.0 48.7 81.7 53.0 52.9 530.9 InfoNCE X 39.0 48.7 81.7 53.0 532.6 PCME Y 39.1 48.9 81.4 53.0 532.6 PCME ++ (μ only) X 39.5 49.1 82.7 55.2 536.2 PCME++ (SWA) Y 40.1 49.7 83.1 55.1 537.0 PCME++ (SWA) Y 40.2 49.8 82.9 55.2 537.3 CLIP Zs† X 28.0 37.8 72.2 46</td>	Method Prob? mAP@R R-P R@1 5K R@1 RSUM Noise Method mAP@R R-P CLIP ZS† X 26.8 36.9 67.1 40.3 471.9 VSE∞ 37.0 46.3 VSE∞ X 40.0 49.5 83.1 55.2 536.5 DAA 6.7 12.5 PRM X 39.0 48.7 82.0 51.7 530.2 DAA 6.7 12.5 DAA X 39.0 48.7 81.7 53.0 532.6 PCME 35.9 46.3 PCME X 39.0 48.7 81.7 53.0 532.6 PCME 37.6 47.6 PCME++ (µ only) X 39.5 49.1 82.7 55.2 536.2 PCME++ (ours) 37.7 47.6 PCME++ (SWA) Y 40.1 49.7 83.1 55.1 537.0 DECL† - - - CLIP ZS† X 28.0	Method Prob? mAP@R R-P R@1 SK R@1 RSUM Noise Method mAP@R R-P R@1 CLIP ZS† X 26.8 36.9 67.1 40.3 471.9 VSE∞ 37.0 46.3 79.7 VSE∞ X 40.0 49.5 83.1 55.2 536.5 DAA 6.7 12.5 18.5 P2RM X 39.0 48.7 82.0 51.7 530.2 DAA 6.7 12.5 18.5 DAA 39.0 48.7 81.7 53.0 532.6 PCME 35.9 46.3 76.1 PCME 39.0 48.7 81.7 53.0 532.6 PCME 37.6 47.6 79.2 PCME+ (μ only) X 39.5 49.1 82.7 55.2 537.0 NCR† 35.9 46.0 78.0 PCME++ (SWA) Y 40.1 49.7 83.1 55.1 537.0 55.2 537.3 YS	Method Prob? mAP@R R-P R@1 5K R@1 RSUM Noise Method mAP@R R-P R@1 5K R@1 CLIP ZS† X 26.8 36.9 67.1 40.3 471.9 VSE∞ X 40.0 49.5 83.1 55.2 536.5 P2RM X 39.0 48.7 82.0 51.7 530.2 DAA X 39.0 48.7 81.7 53.0 52.9 530.9 InfoNCE X 39.0 48.7 81.7 53.0 532.6 PCME Y 39.1 48.9 81.4 53.0 532.6 PCME ++ (μ only) X 39.5 49.1 82.7 55.2 536.2 PCME++ (SWA) Y 40.1 49.7 83.1 55.1 537.0 PCME++ (SWA) Y 40.2 49.8 82.9 55.2 537.3 CLIP Zs† X 28.0 37.8 72.2 46

Observation 1: When scaling-up backbone, det. methods are overfitted, but prob. methods are not. Observation 2: PCME++ successfully handles NC although it is not designed for tackling NC.





Large-scale pre-training with PCME++

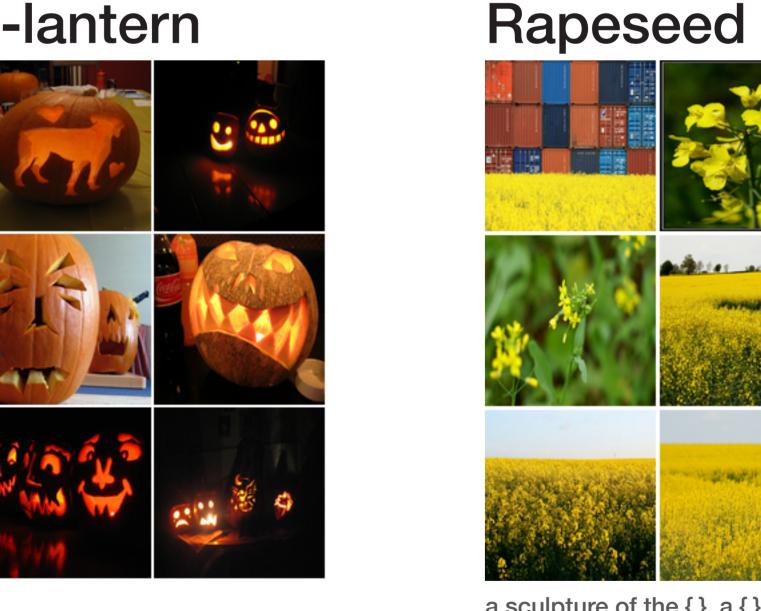
A large polar bear stands on rock with an open mouth

Model	Prompts	Top-1 ImageNet Zero-shot acc	100 -	
CLIP	"A photo of { · }" All 80 prompts	31.85 35.50	ation - 09	
PCME++	"A photo of { · }" All 80 prompts	30.43 34.22	Population - 09	
	Top-K certain prompts Best top-K for each class	34.22 41.82	20 -	des la colonia de la colonia d
_				0 20

ViT-B/16 trained on CC3M,12M,RedCaps



a close-up photo of a { }, a photo of the clean { }. art of the { }. a cropped photo of the { }. a close-up photo of the { }. a photo of a clean { }. a close-up photo of the {



graffiti of a { }. a rendition of the { }. a photo of the small { }. a photo of one { }. a photo of the dirty

Check out more details in...



Det. text embedding



a cropped photo of a { }

Check out my other related works!

PCME (CVPR'21)

ECCV Caption (ECCV'22)



