Understanding Negative Samples in Instance Discriminative Self-supervised representation Learning

Kento Nozawa[1,2]

Issei Sato[1]

1 THE UNIVERSITY OF TOKYO

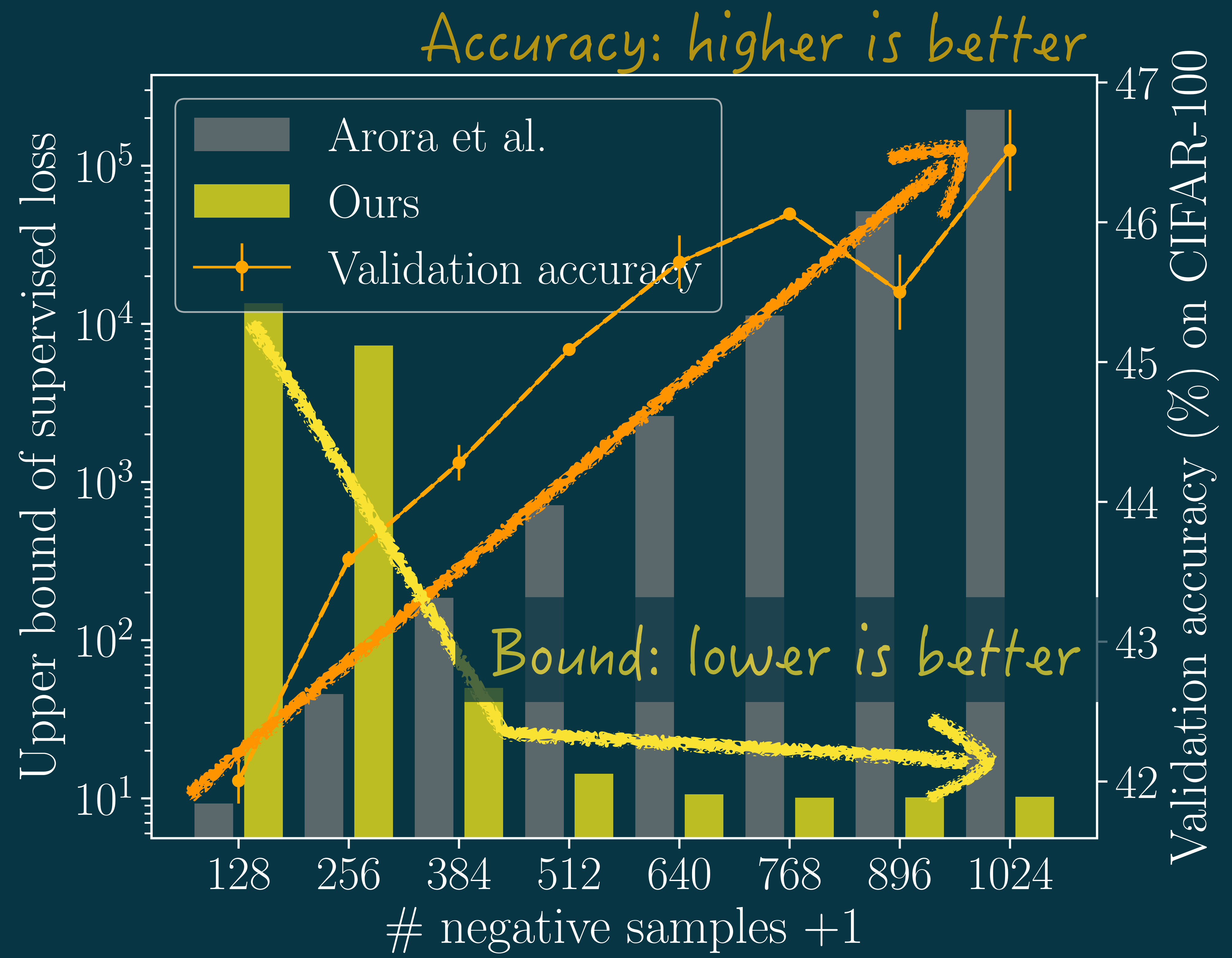
2 AIP

Paper: https://arxiv.org/abs/2102.06866
Code: https://github.com/nzw0301/Understanding-Negative-Samples

Short summary of this talk

- We point out the inconsistency between self-supervised learning's common practice and an existing theoretical analysis.
  - Practice: Large # negative samples don't hurt classification performance.
  - Theory: they hurt classification performance.

- We propose an novel analysis using Coupon collector's problem.



Instance discriminative self-supervised representation learning

Goal: Learn generic feature encoder $\mathbf{f}$, for example deep neural nets, for a downstream task, such as classification.

Feature representations help a linear classifier to attain classification accuracy comparable to a supervised method from scratch.

Overview of Instance discriminative self-supervised representation learning

Draw $K + 1$ samples from an unlabeled dataset.

- $\mathbf{x}$: anchor sample.
- $\mathbf{x}^-$: negative sample. It can be a set of samples $\{\mathbf{x}_k^-\}_{k=1}^K$.

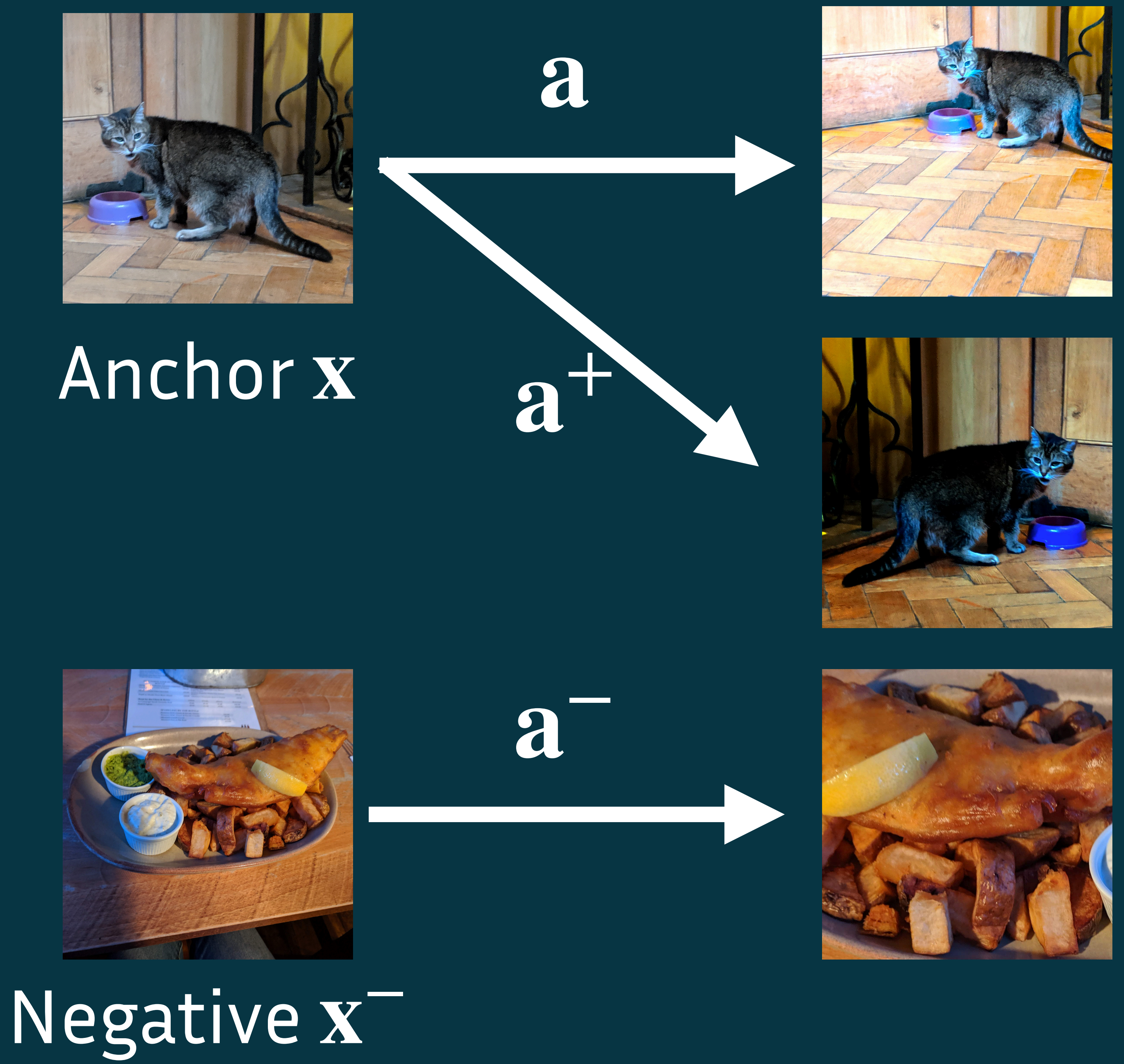
Anchor $\mathbf{x}$

Negative $\mathbf{x}^-$

Overview of Instance discriminative self-supervised representation learning

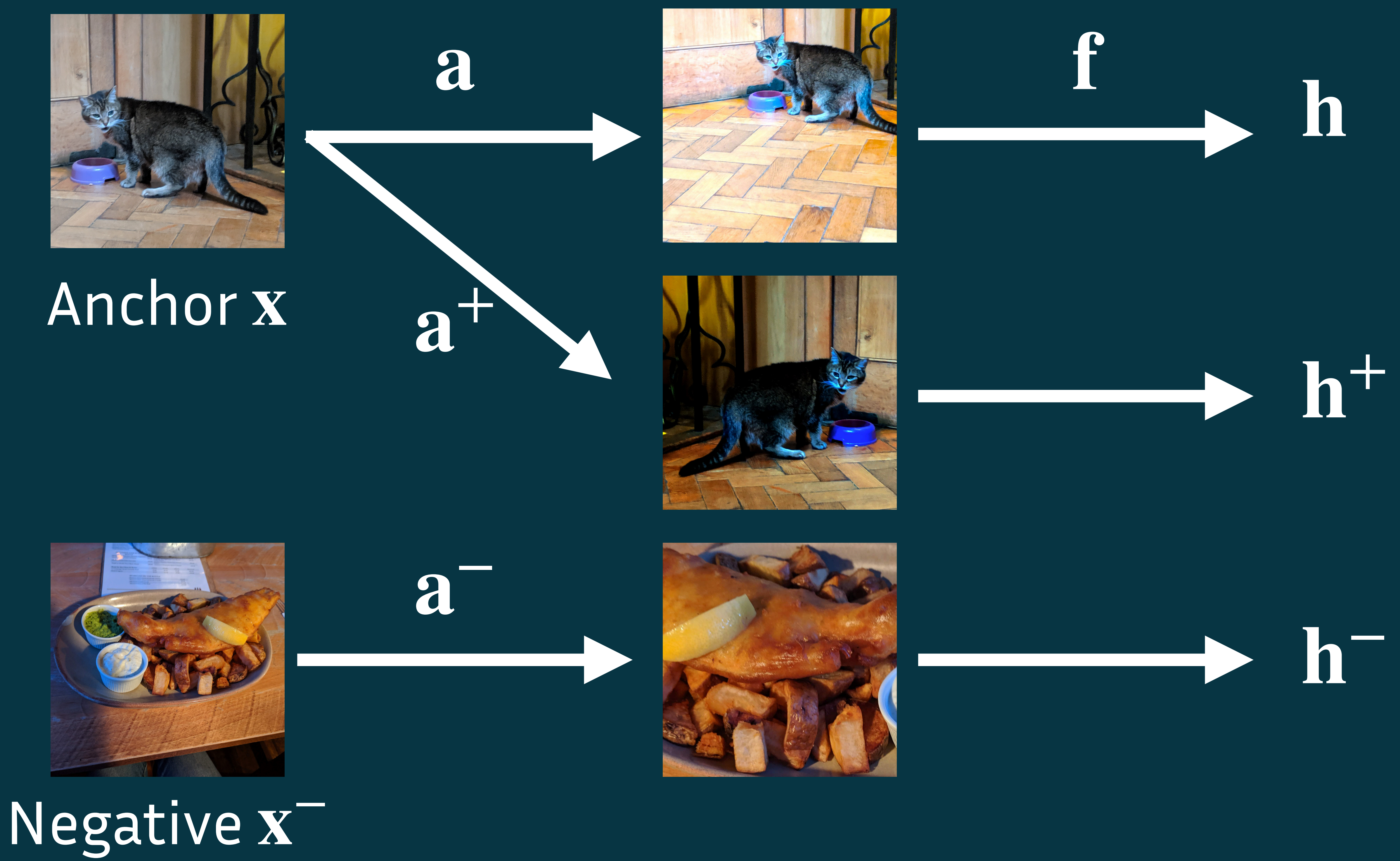
Apply data augmentation to the samples:

- For the anchor sample $\mathbf{x}$, we draw and apply two data augmentations $\mathbf{a}, \mathbf{a}^{+}$.
- For negative sample $\mathbf{x}^{-}$, we draw and apply single data augmentation $\mathbf{a}^{-}$.



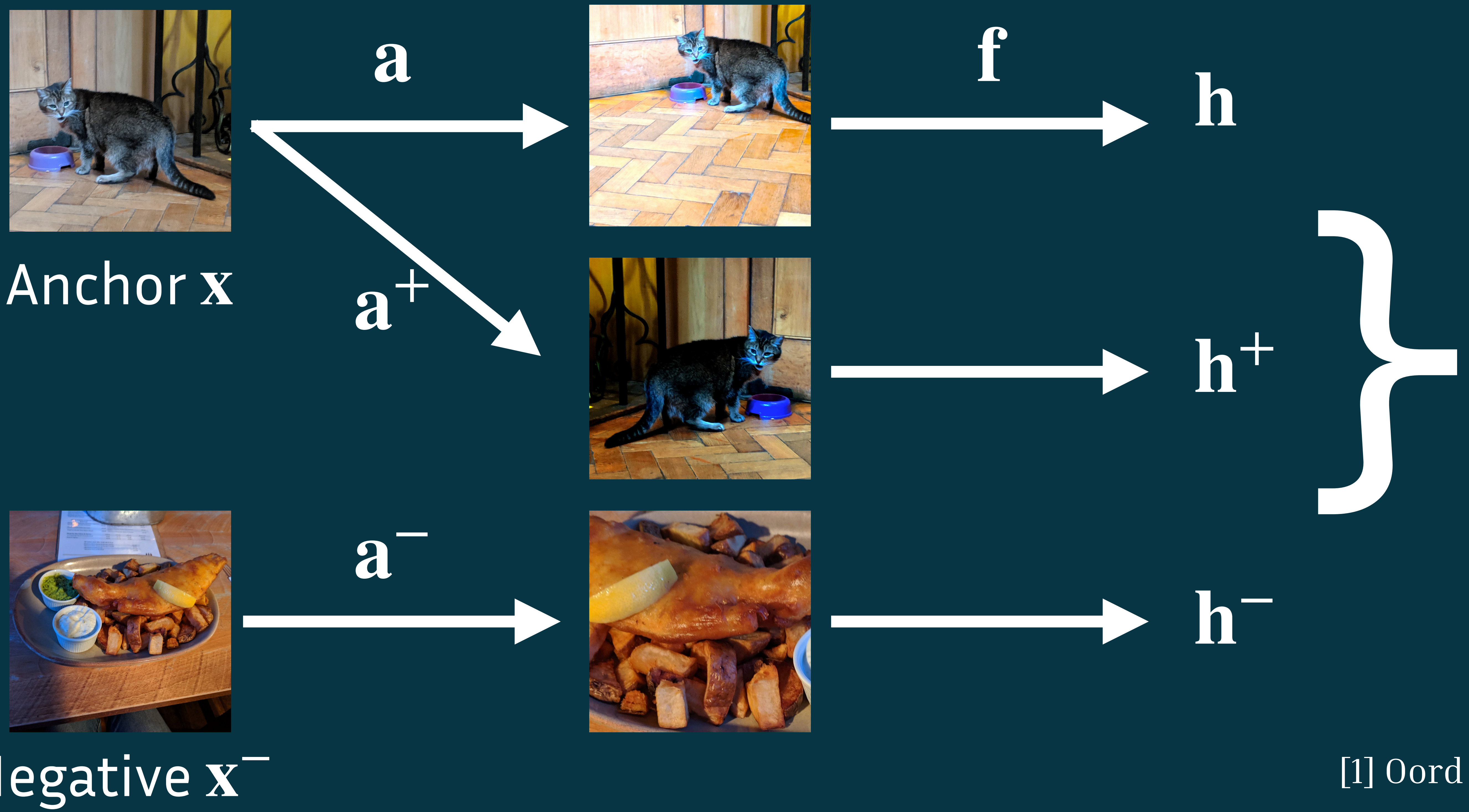
Overview of Instance discriminative self-supervised representation learning

Feature encoder $\mathbf{f}$ maps augmented samples to feature vectors $\mathbf{h}, \mathbf{h}^+, \mathbf{h}^-$.



Overview of Instance discriminative self-supervised representation learning

- Minimize a contrastive loss given feature representations.
  - sim(. · .): a similarity function, such as cosine similarity.
- Learned $\hat{\mathbf{f}}$ works as a feature extractor for a downstream task.



Contrastive loss function, ex. InfoNCE [1]:

$$-\ln \frac{\exp[\mathrm{sim}(\mathbf{h}, \mathbf{h}^+)]}{\exp[\mathrm{sim}(\mathbf{h}, \mathbf{h}^+)] + \exp[\mathrm{sim}(\mathbf{h}, \mathbf{h}^-)]}$$

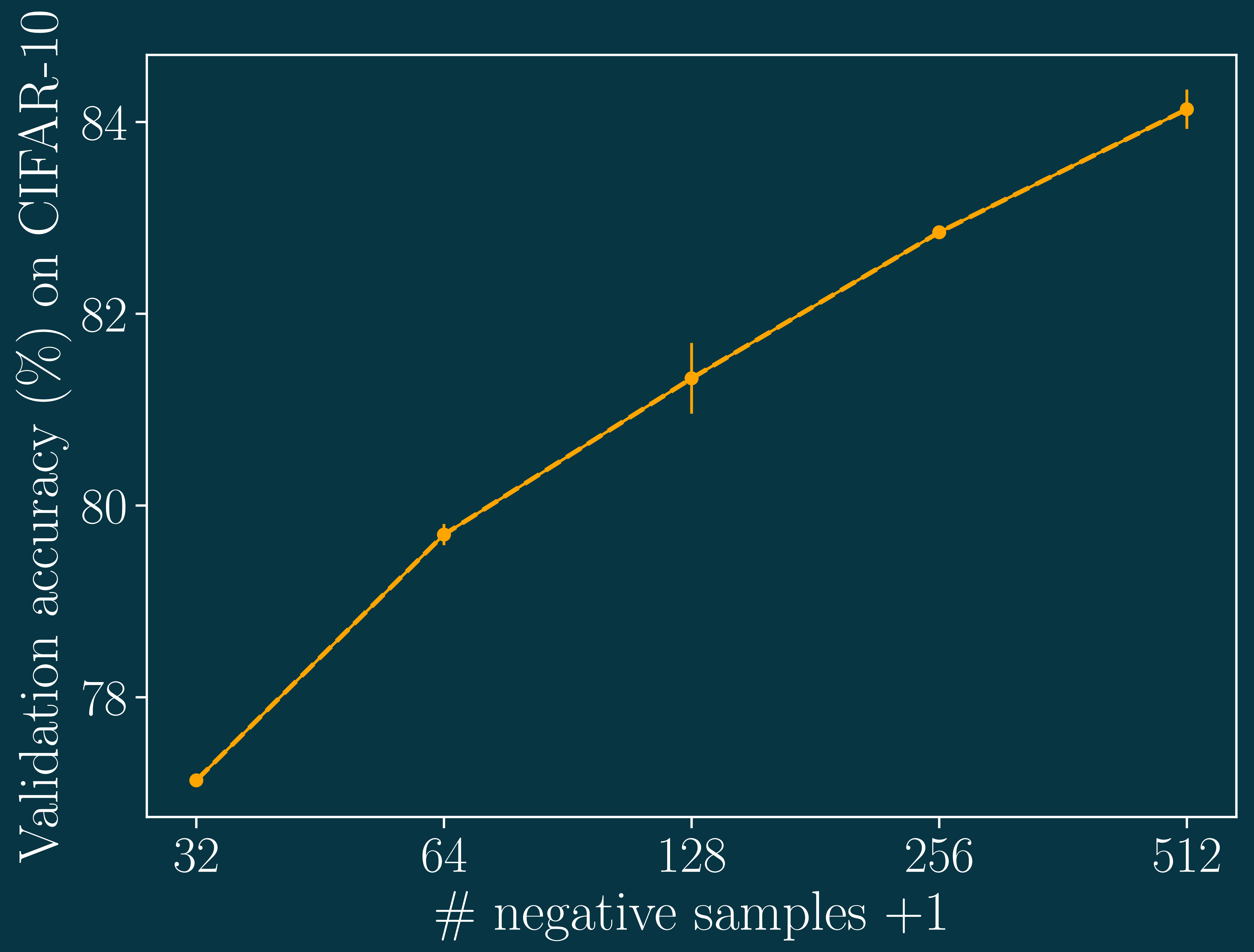
[1] Oord et al. Representation Learning with Contrastive Predictive Coding, *arXiv*, 2018.

Common technique: use large # negative samples K

By increasing # negative samples, learned $\widehat{\mathbf{f}}$ yields informative features for linear classifier in practice.

For ImageNet,

- MoCo [2]: $K = 65\,536$.
- SimCLR [3]: $K = 8\,190$ or even more.



[2] He et al. Momentum Contrast for Unsupervised Visual Representation Learning, In *CVPR*, 2020.

[3] Chen et al. A Simple Framework for Contrastive Learning of Visual Representations, In *ICML*, 2020.

A theory of contrastive representation learning

Informal bound [4] modified for self-supervised learning:

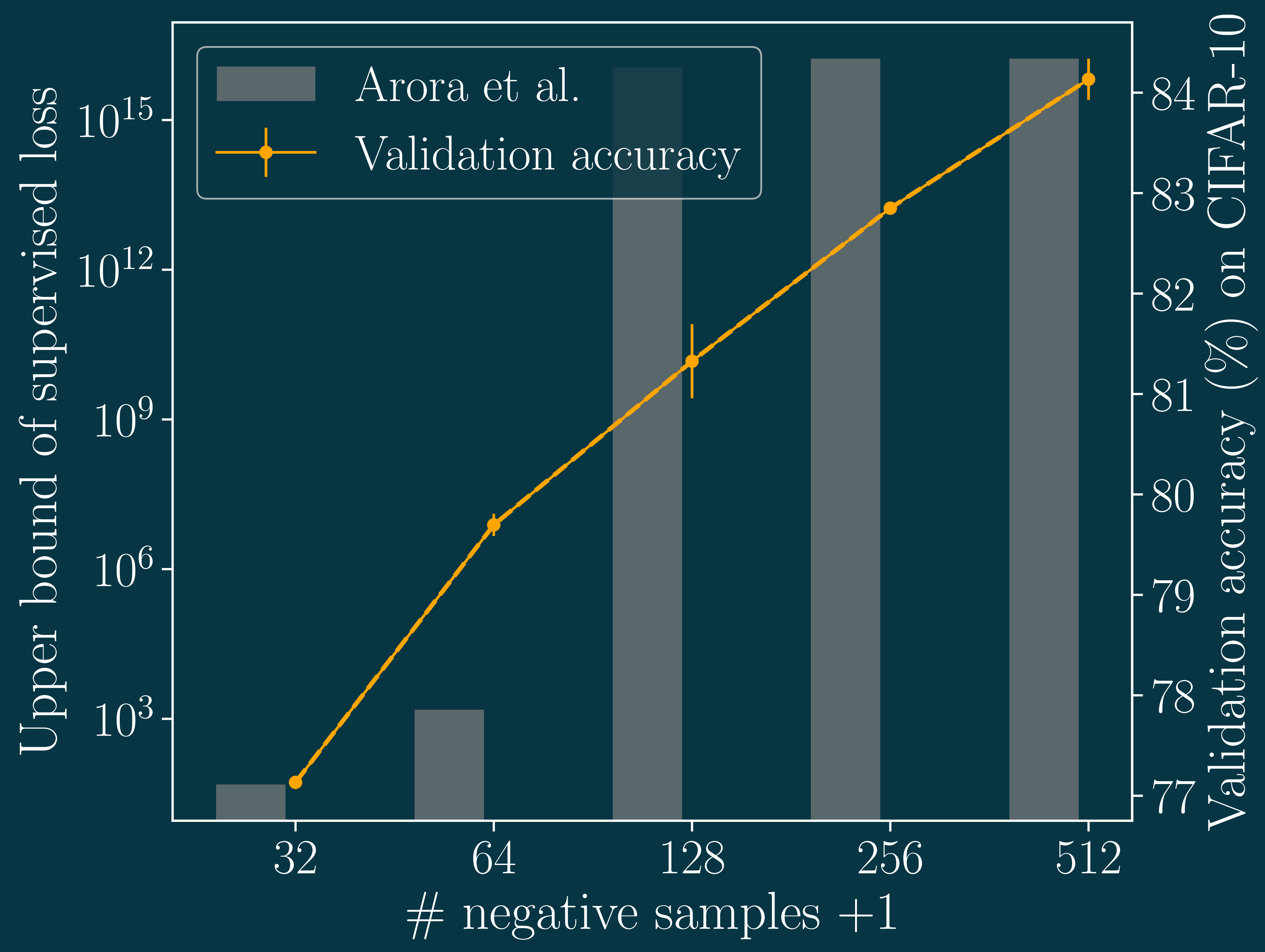
$$L_{\text{cont}}(\mathbf{f}) \geq (1-\tau_K)(L_{\text{sup}}(\mathbf{f}) + L_{\text{sub}}(\mathbf{f})) + \tau_K \underbrace{\ln(\text{Col}+1)}_{\text{Collison term}} + d(\mathbf{f}).$$

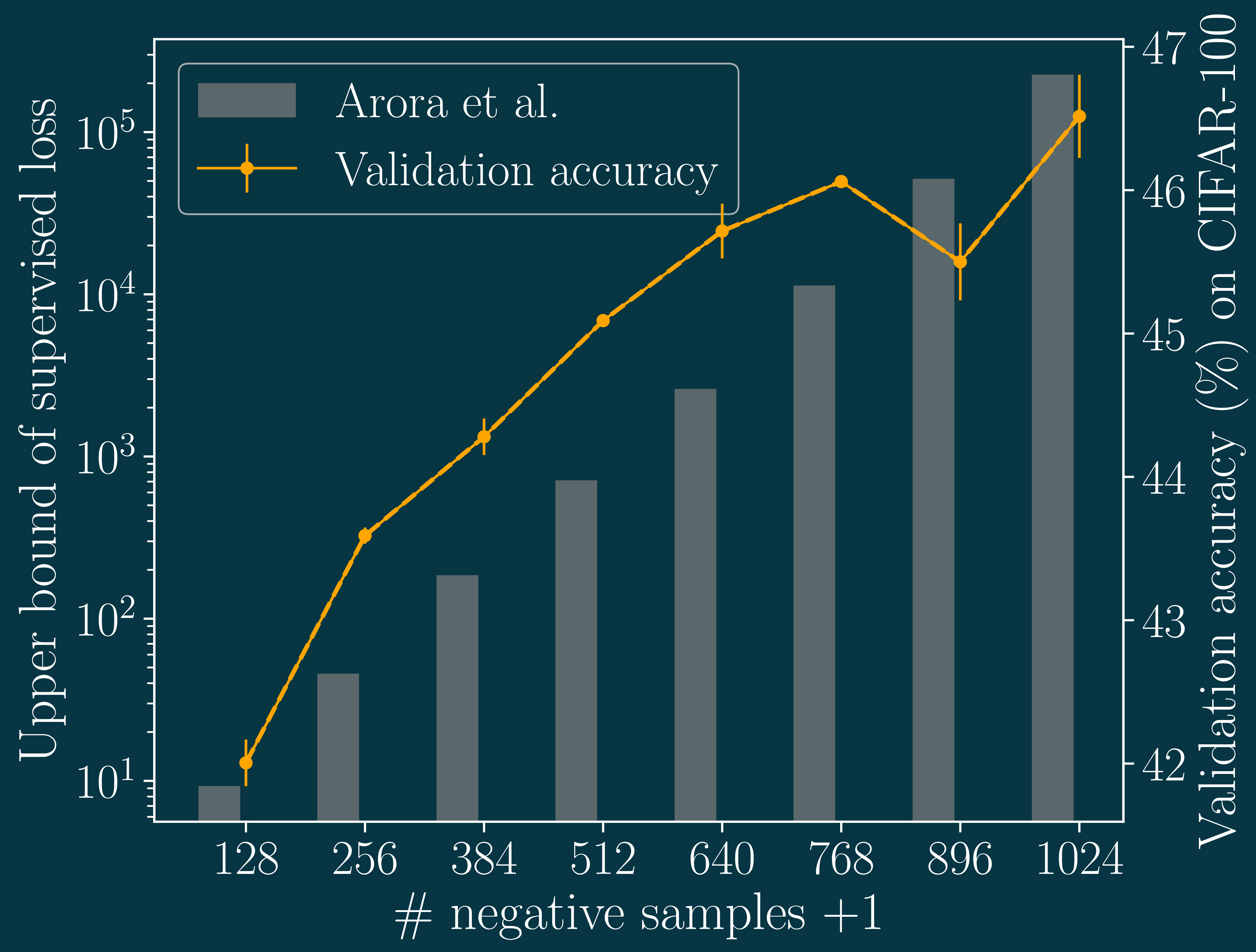
- τ_K: Collision probability that anchor's label appears in negatives' one.
- $L_{\text{sup}}(\mathbf{f})$: Supervised loss with $\mathbf{f}$.
- $L_{\text{sub}}(\mathbf{f})$: Supervised loss over subset of labels with $\mathbf{f}$.
- Col: the number of duplicated negative labels with the anchor's label.
- $d(\mathbf{f})$: a function of $\mathbf{f}$, but almost constant term in practice.

[4] Arora et al. A Theoretical Analysis of Contrastive Unsupervised Representation Learning, In *ICML*, 2019.

The bound of L_{sup} explodes with large K

- The bound on CIFAR-10, where # classes is 10 with $K = 31$:
  - About 96 % samples contribute the collision term not related to the supervised loss due to τ_K.
- Plots rearranged upper bound: $L_{\mathrm{sup}}(\mathbf{f}) \leq (1 - \tau_K)^{-1} \left[L_{\mathrm{cont}}(\mathbf{f}) - \tau_K \ln(\mathrm{Col} + 1) - d(\mathbf{f}) \right] - L_{\mathrm{sub}}(\mathbf{f})$:





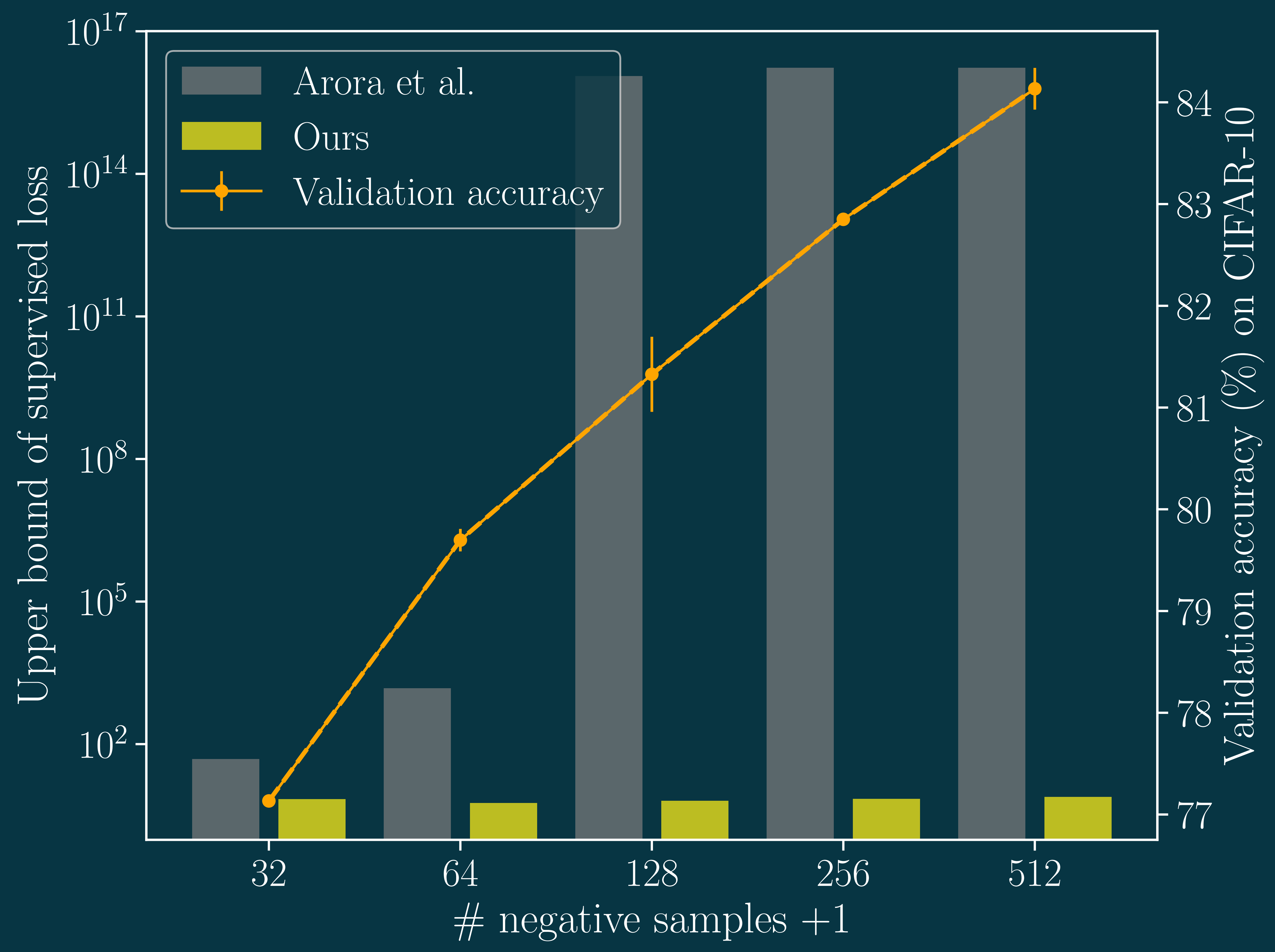
Contributions: novel lower bound of contrastive loss

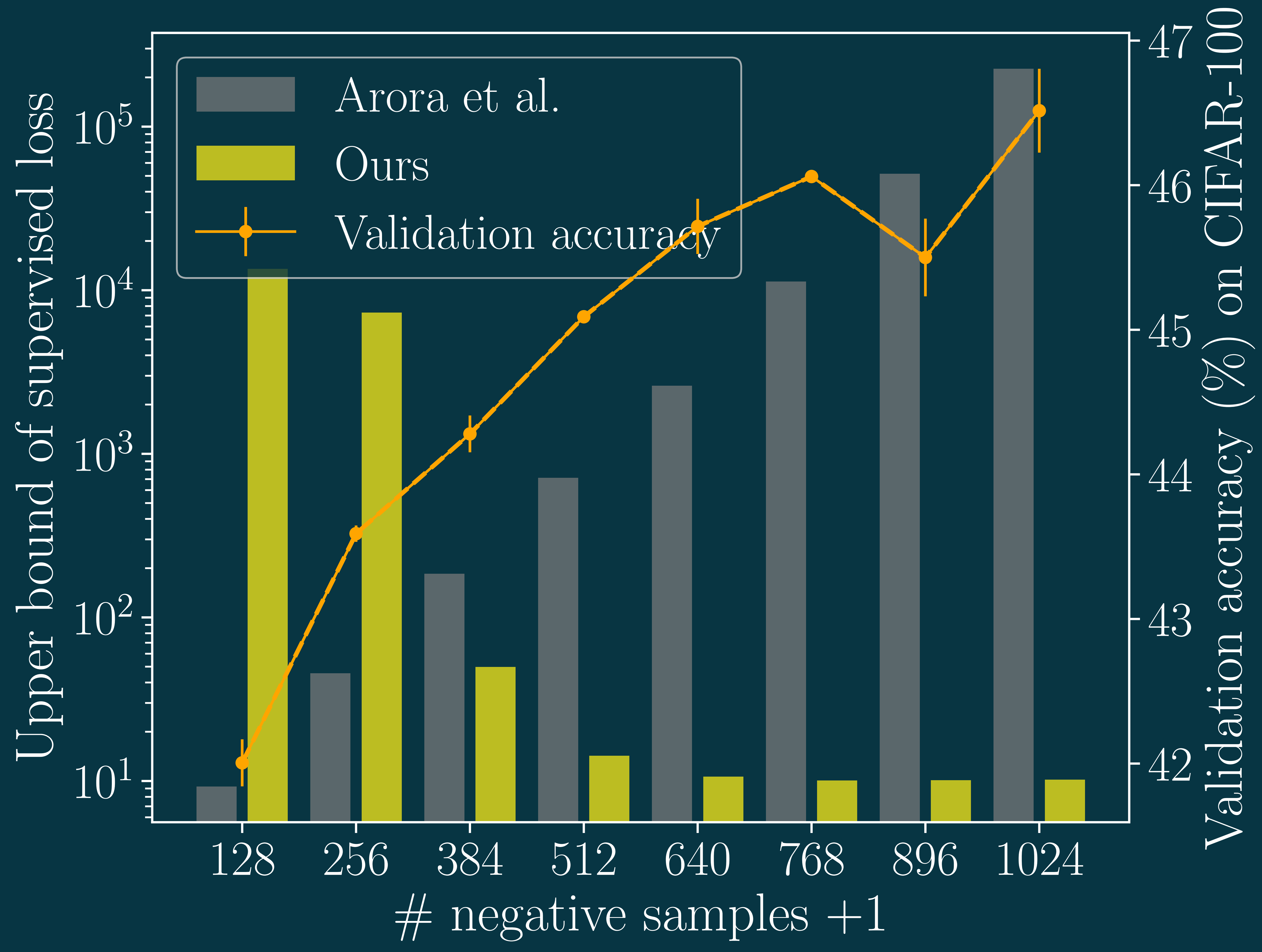
Informal proposed bound:

$$L_{\mathrm{cont}}(\mathbf{f}) \geq \frac{1}{2}\left\{ \upsilon_{K+1} L_{\mathrm{sup}}(\mathbf{f}) + (1-\upsilon_{K+1}) L_{\mathrm{sub}}(\mathbf{f}) + \ln(\mathrm{Col}+1) \right\} + d(\mathbf{f}).$$

- Key idea: replace collision probability τ with Coupon collector's problem's probability υ_{K+1} that $K+1$ samples' labels include the all supervised labels.
- Additional insight: the expected $K+1$ to draw all supervised labels from ImageNet-1K is about 7700.

Our bound doesn’t explode





Conclusion

- We pointed out the inconsistency between self-supervised learning's common practice and the existing bound.
  - Practice: Large K doesn't hurt classification performance.
  - Theory: large K hurts classification performance.

- We proposed the new bound using Coupon collector's problem.

- Additional results:
  - Upper bound of the collision term.
  - Optimality when $\upsilon = 0$ with too small K.
  - Experiments on a NLP dataset.