Visual representations that transfer

Diane Larlus

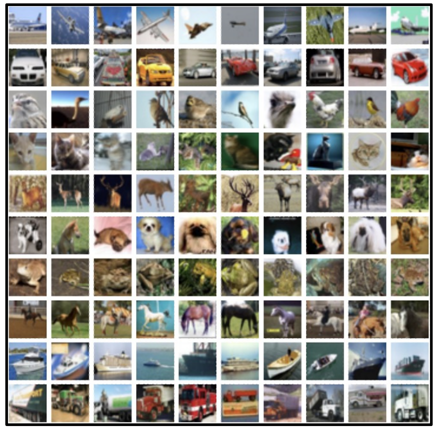
Principal Scientist at NAVER LABS Europe

Self-Supervised Learning, Theory and Practice Workshop – NeurIPS 2023

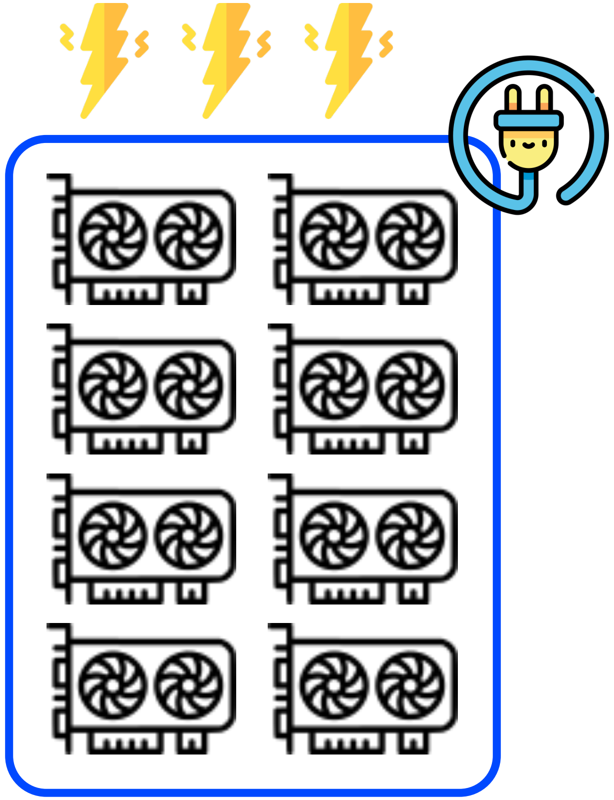
December 16th, 2023

NAVER LABS

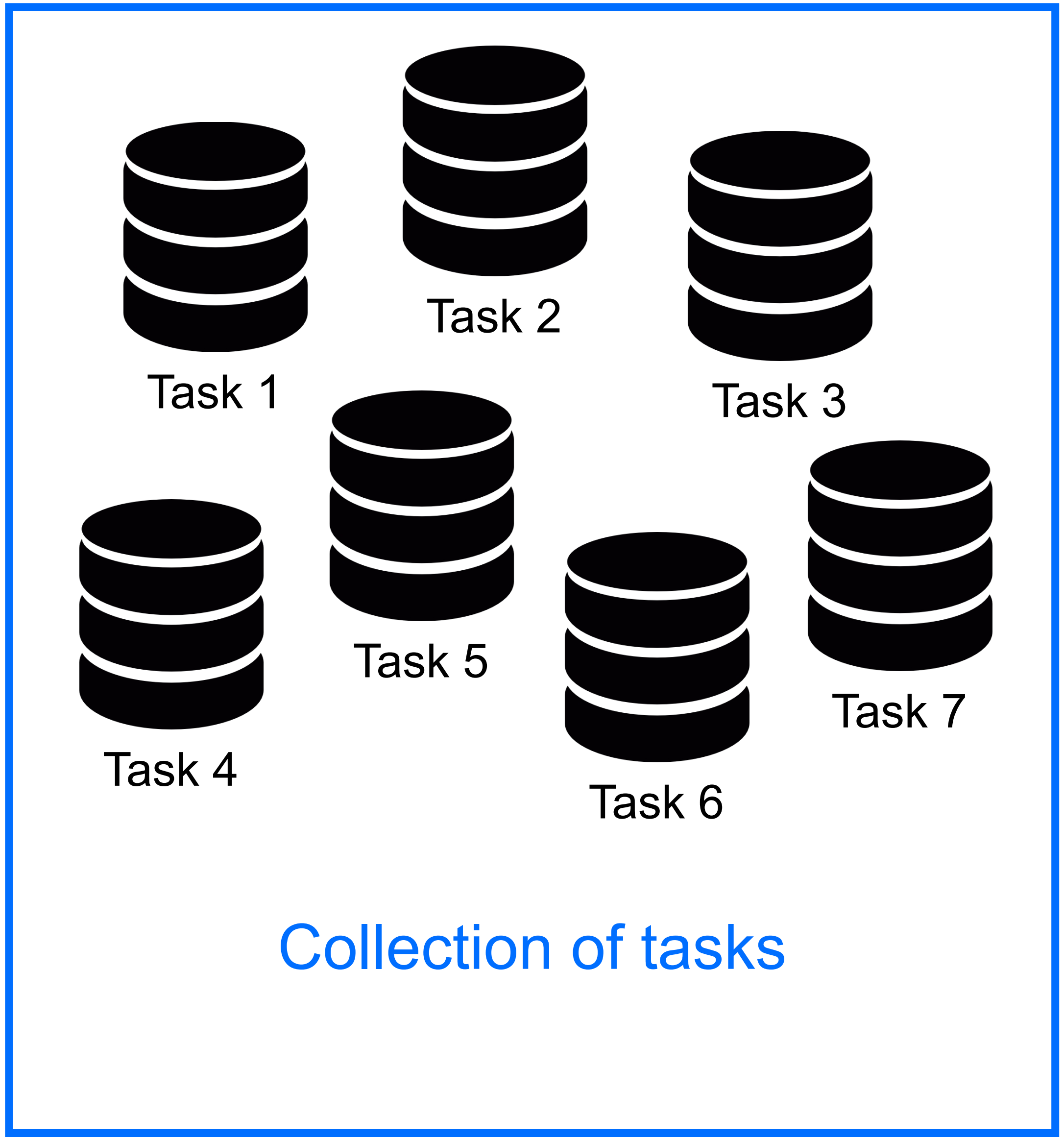
- A large image collection with labels
- A powerful neural architecture
- Lots of compute

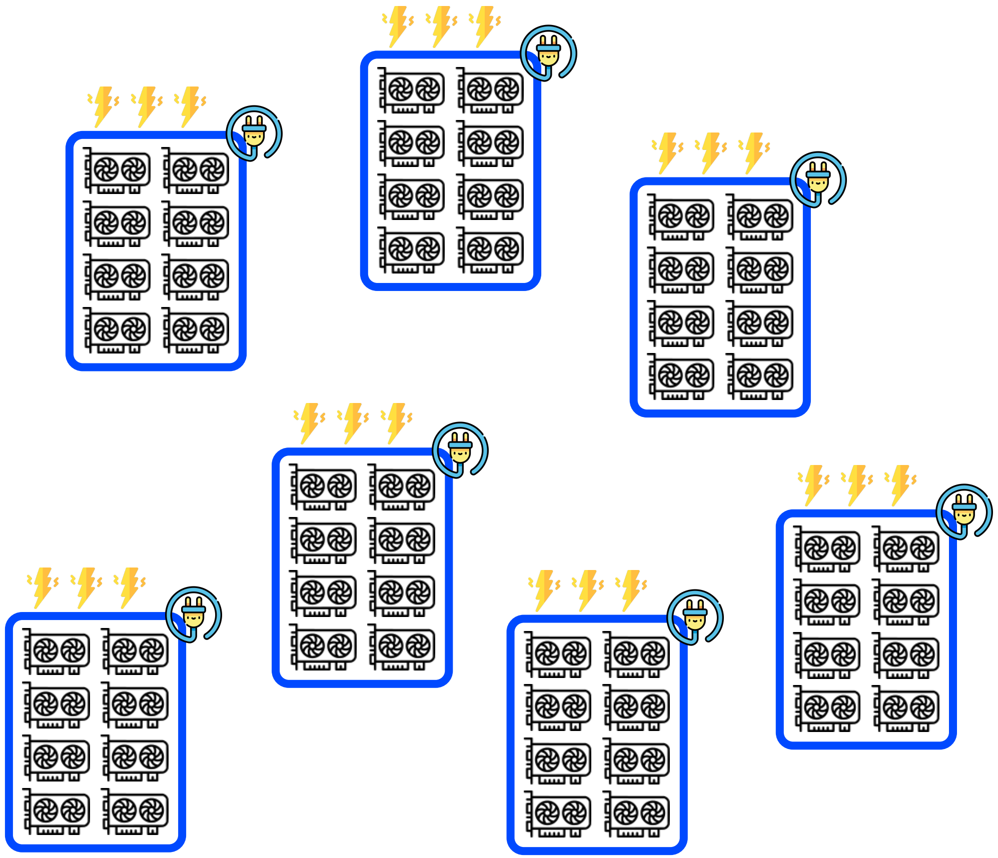


IMAGENET

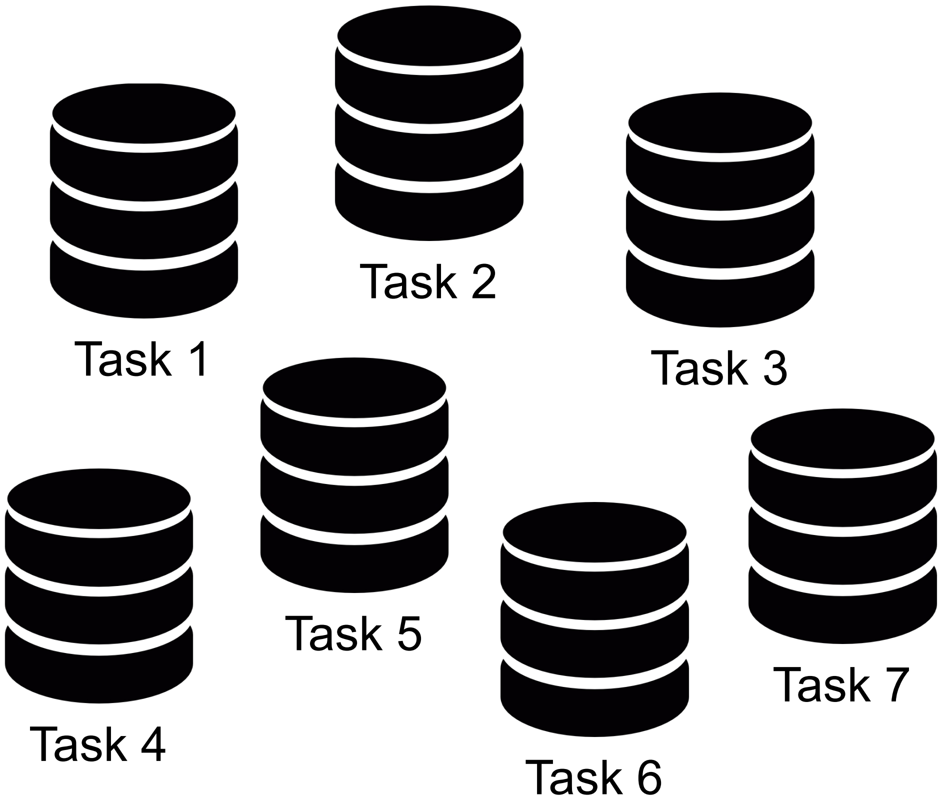


Multiple tasks



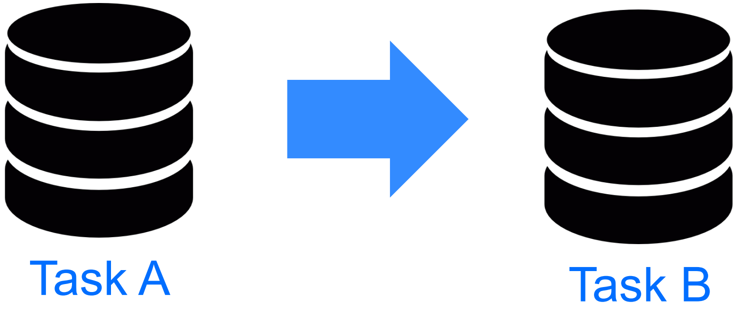


The case of two tasks

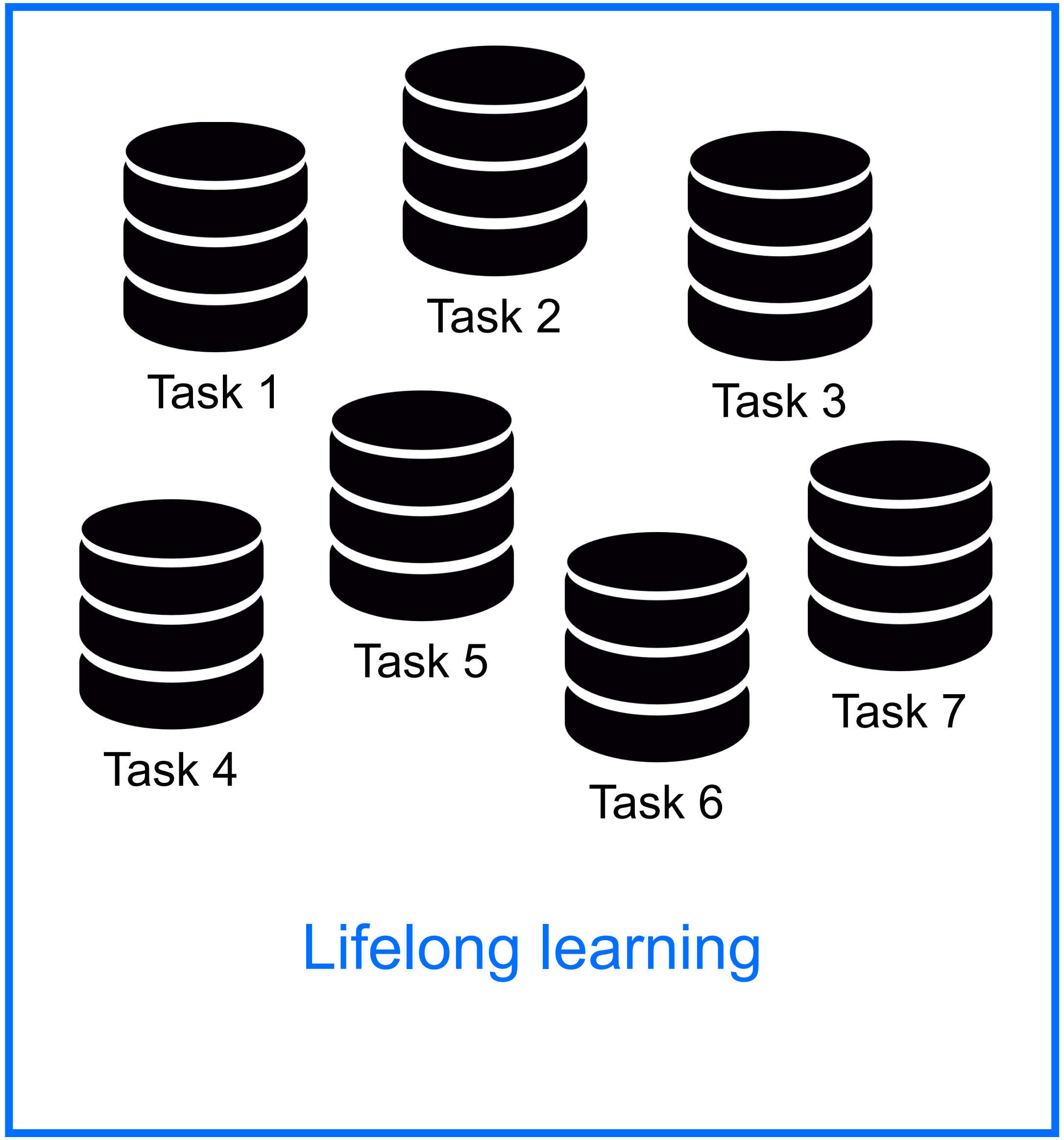


Lifelong learning

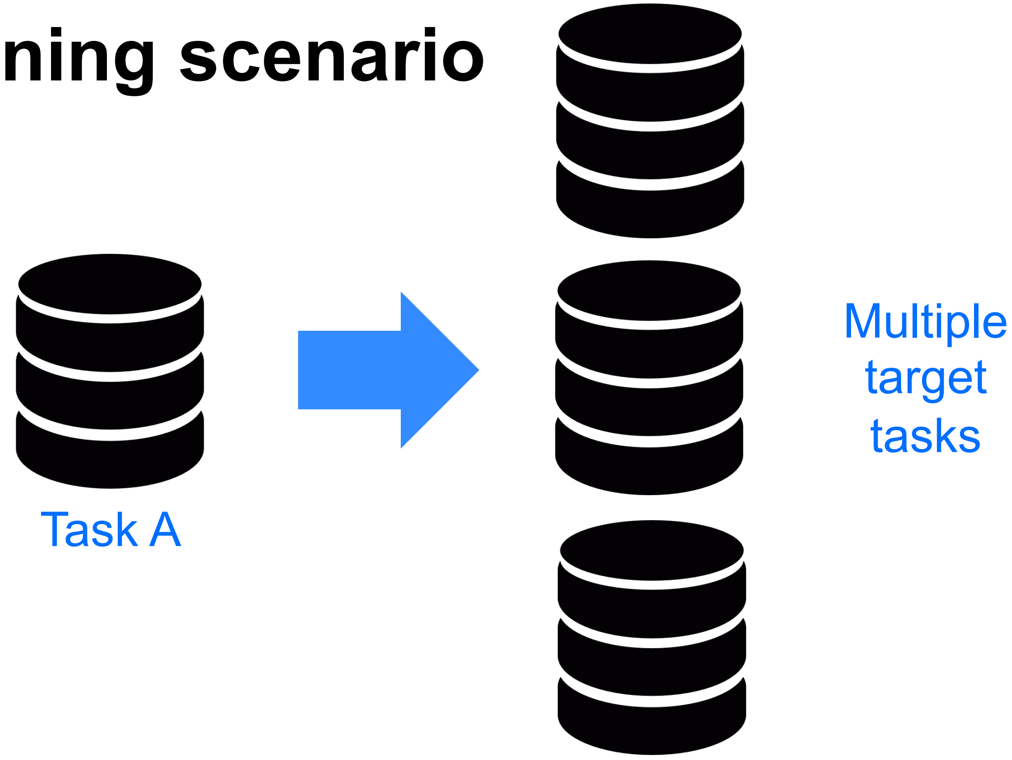
Is Task A useful for Task B?



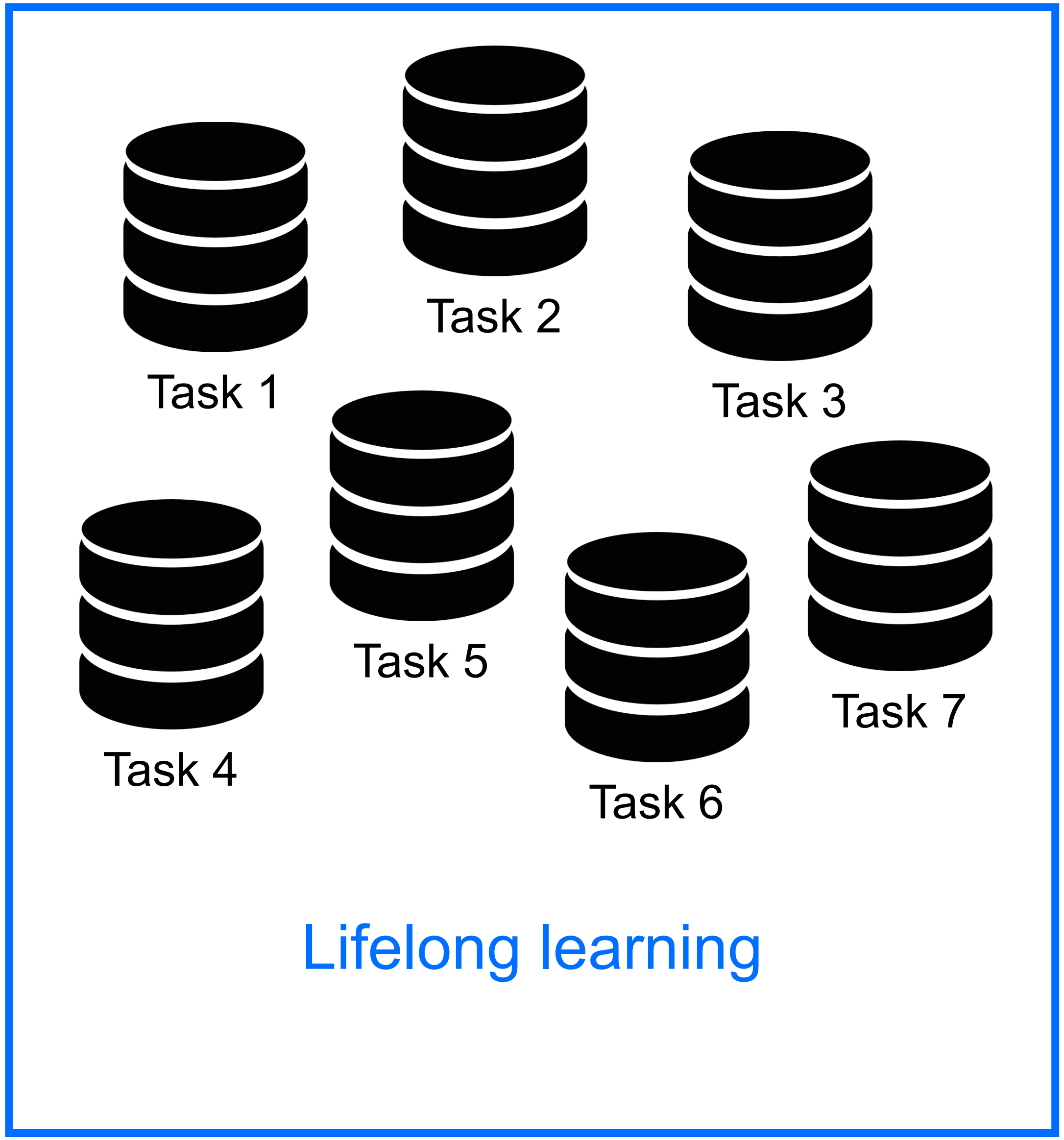
- How should we train on Task A?
- How should we adapt on Task B?



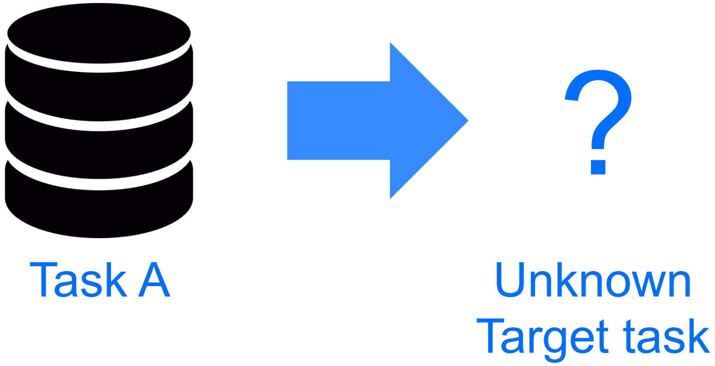
Pretraining scenario



- How should we train on Task A?

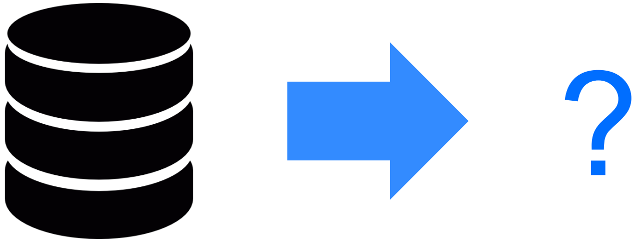


Pretraining scenario



- How should we train on Task A?

Transferable visual representations



Learning transferable visual representations

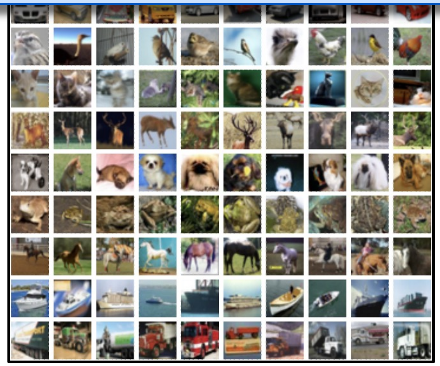
Proxy task

Reducing annotation cost

Fully-Supervised Classification

Images + labels

IMAGENET



Model

Visual representations

Target tasks

Image Classification

Object Detection

Instance Segmentation

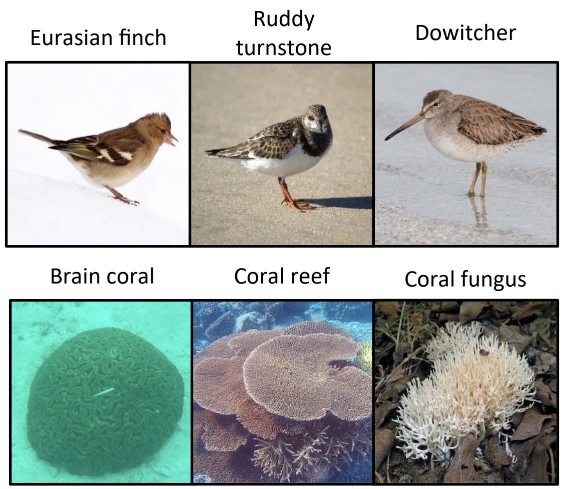
Image Retrieval

Reducing annotation cost

Fully-Supervised

fine-grained annotations

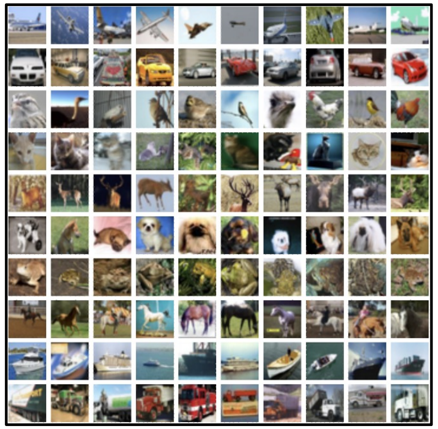
expert knowledge



Self-supervised

annotation-free images

no annotation required



Proxy task

Fully-supervised classification or
Self-supervised approaches, etc.

Model

Visual representations

How well does the produced visual representation transfer?

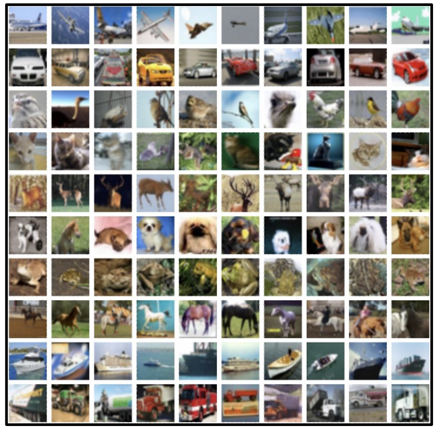
Target tasks

??

Proxy task

Fully-supervised classification or
Self-supervised approaches, etc.

IMAGENET



Model

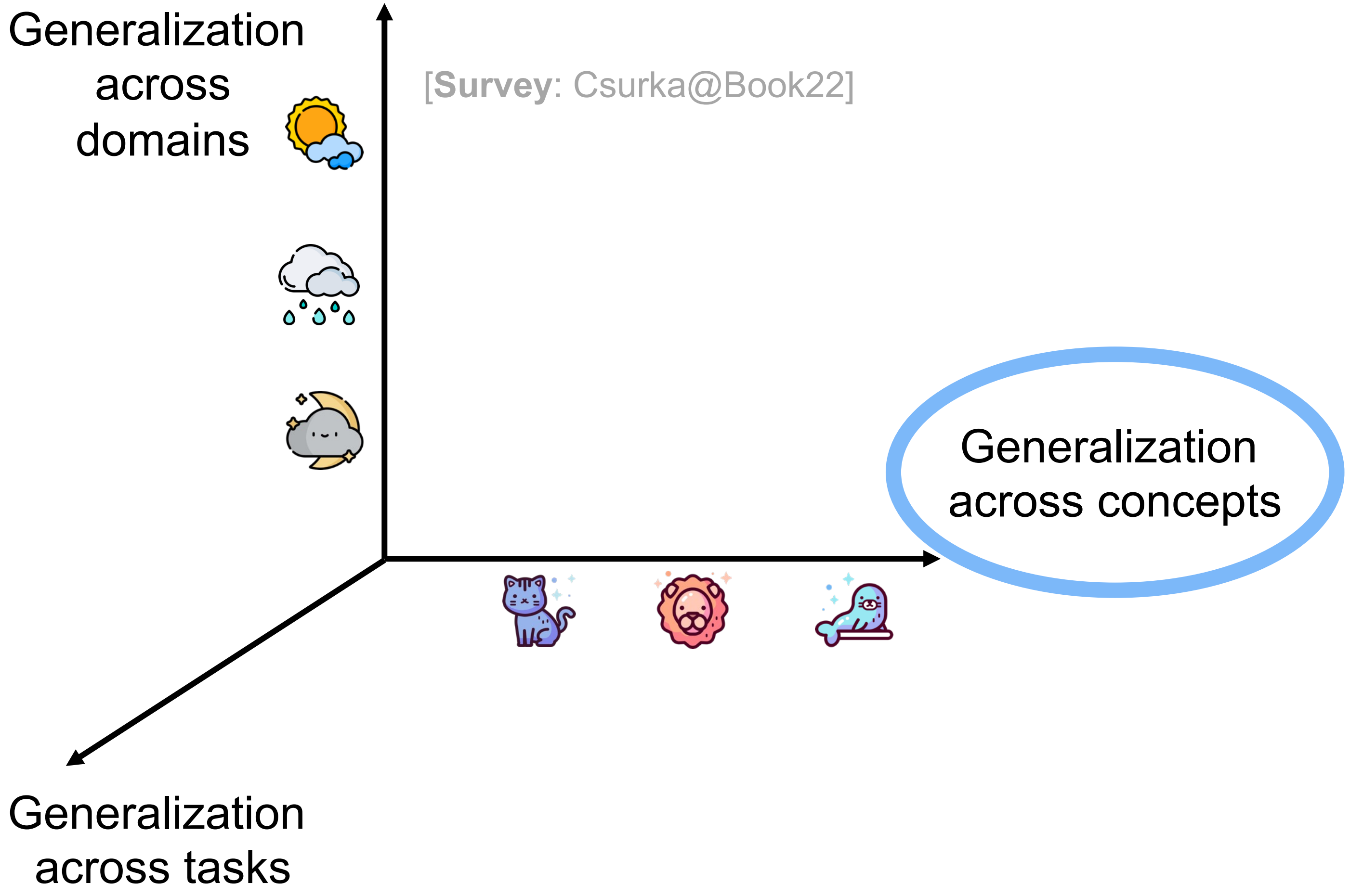
Visual representations

Measure performance on (many) other datasets

Evaluation of visual representations

Proxy task

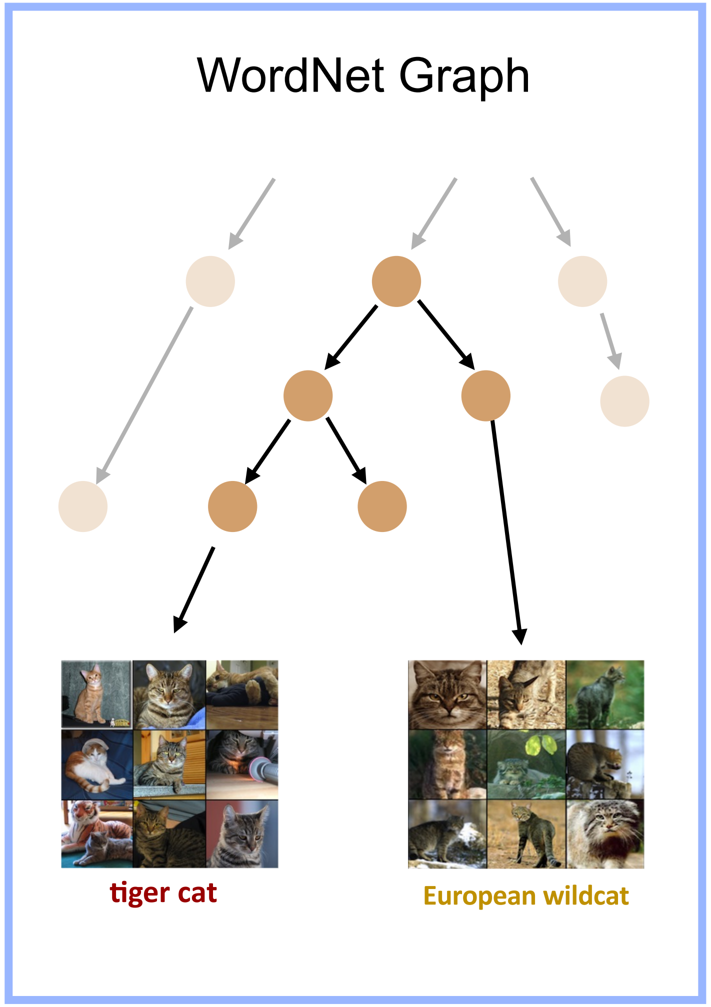
Target task



When training a model on a set of ***seen*** *concepts,*
how well does it generalize to ***new, unseen*** *set of concepts ?*

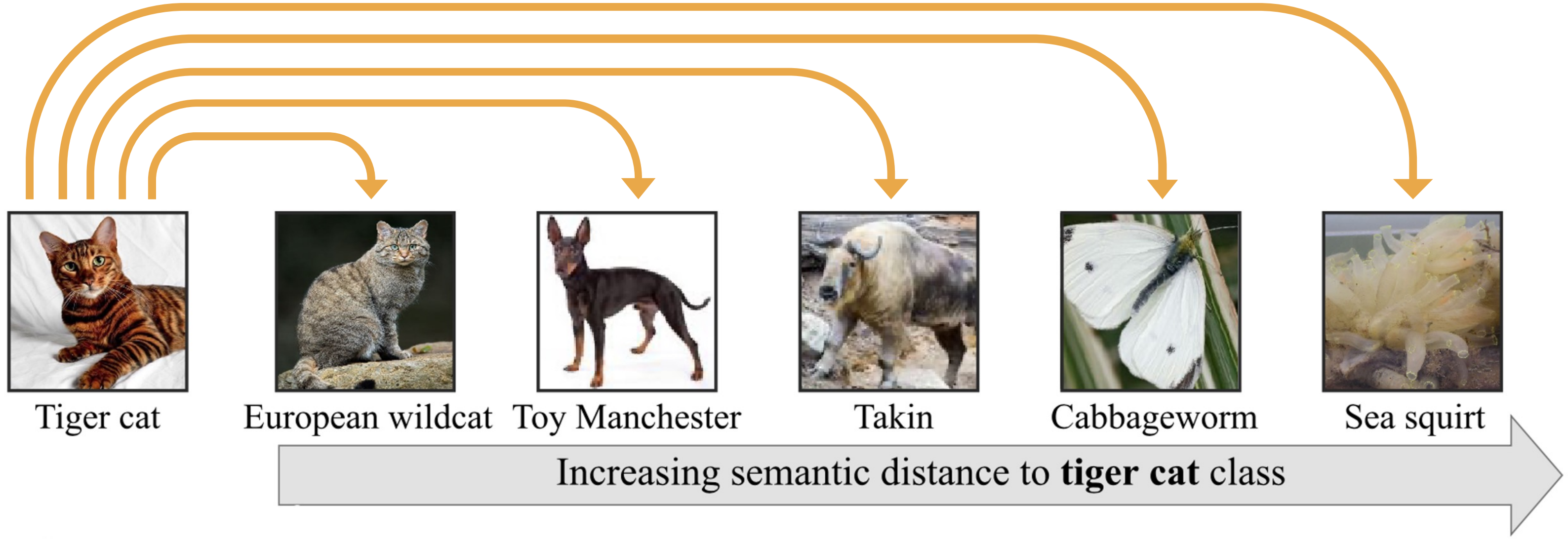
Seen: Cats
Unseen: Caticorns

Measure the semantic distance between concepts



[**Lin**: Lin@ICML1998]

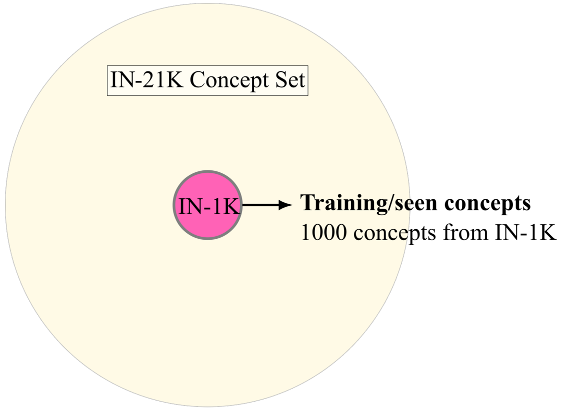
Measure the semantic distance between concepts

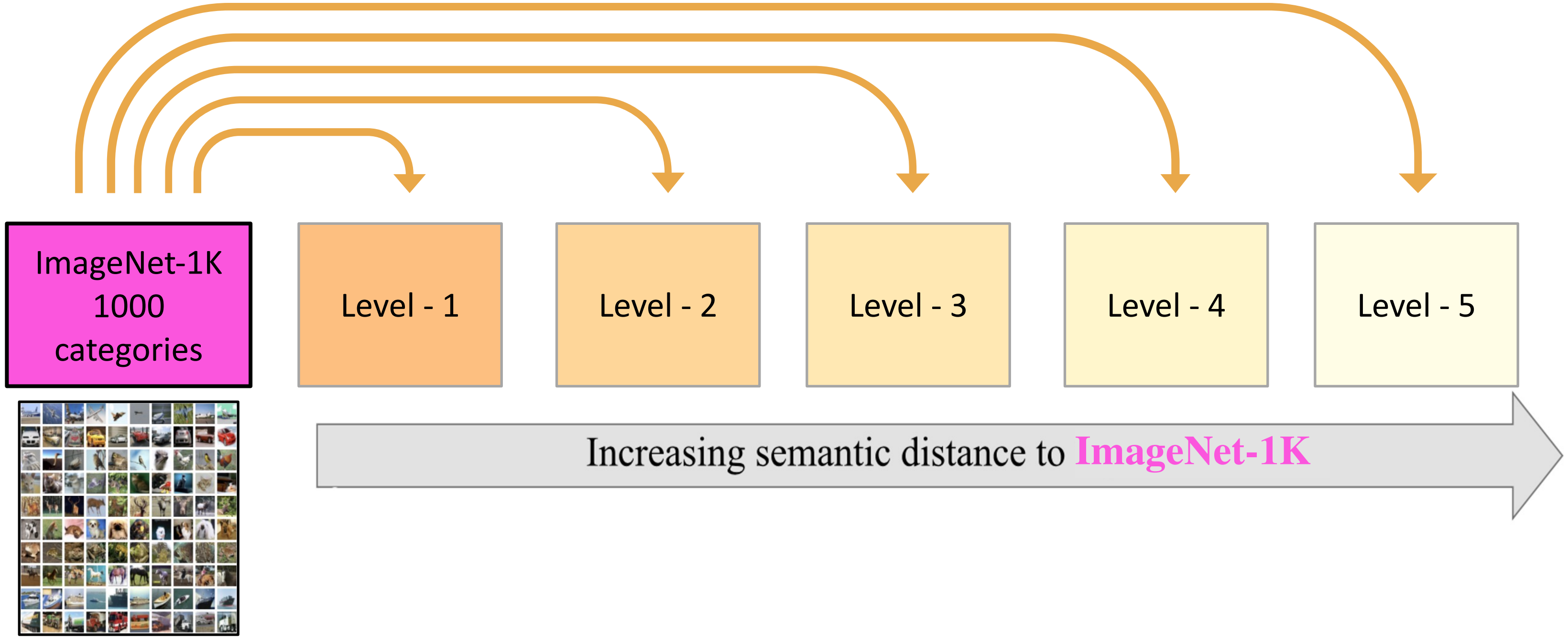


[**Lin**: Lin@ICML1998]

Measure the semantic distance between sets of concepts

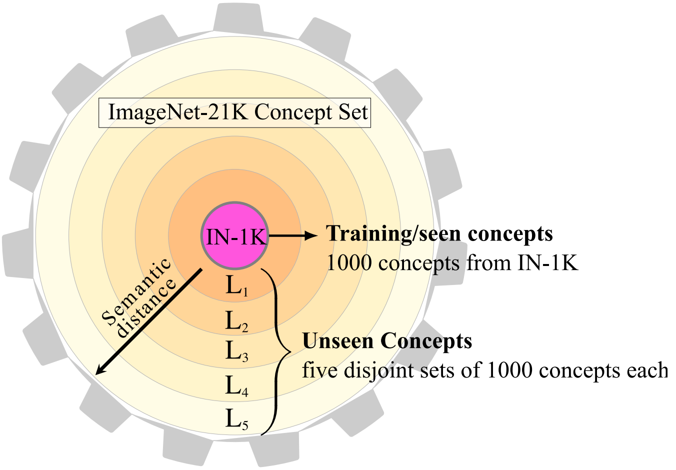
[**ImageNet**: Deng@CVPR2009]



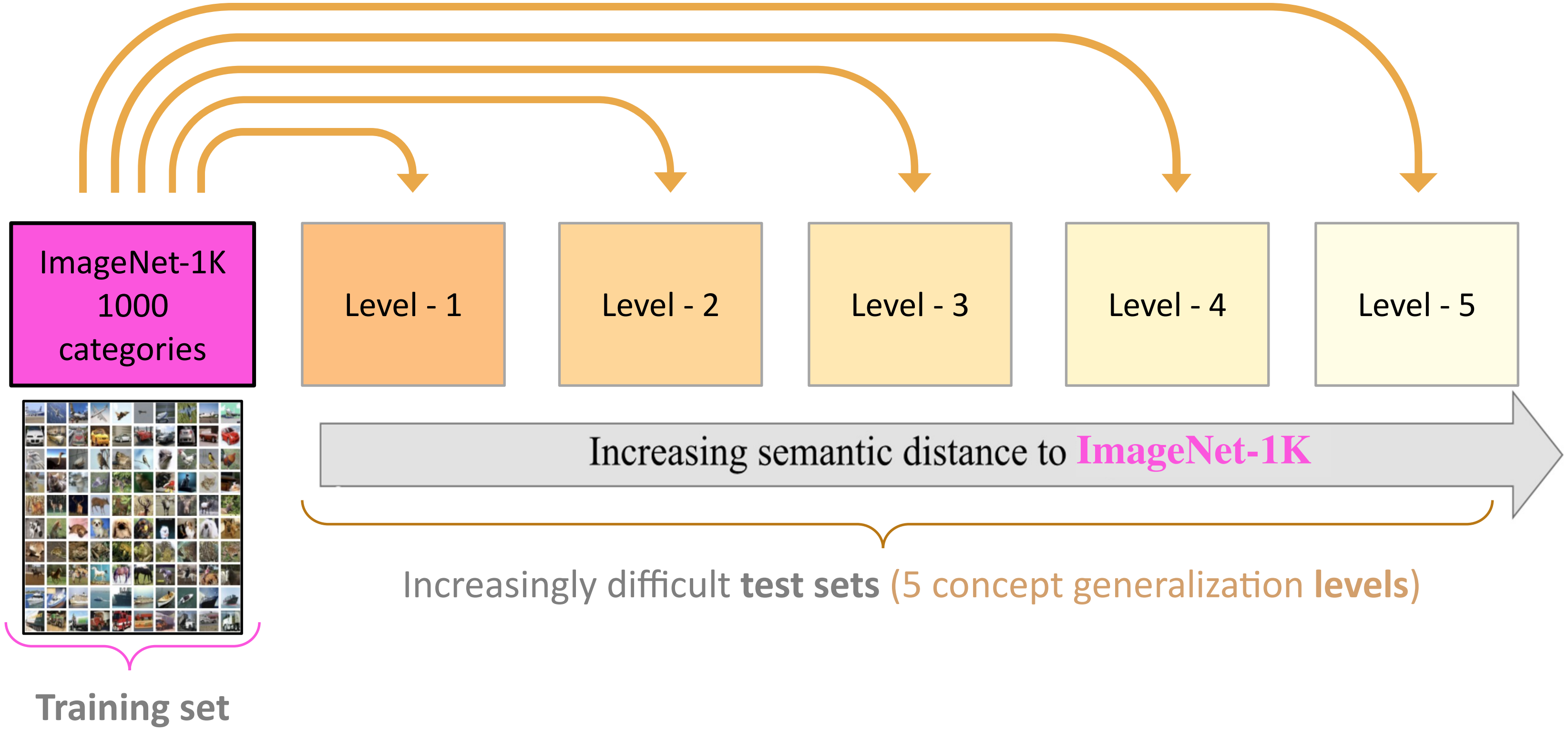


Measure the semantic distance between sets of concepts

[ImageNet: Deng@CVPR2009]



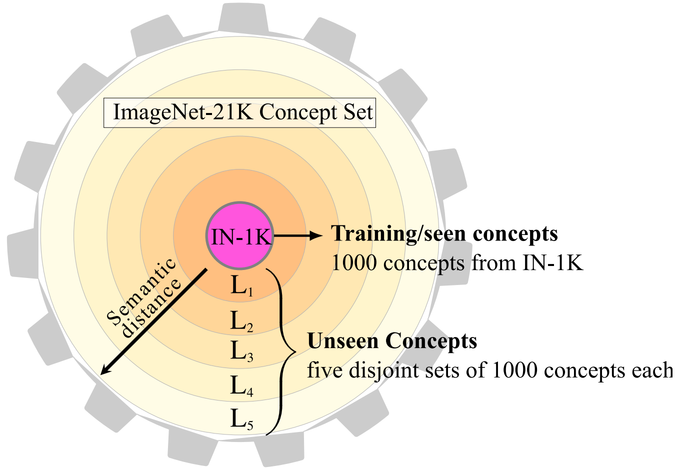
Proposed CoG benchmark



Observations

- It is harder to generalize to semantically distant concepts
- Recent **self-supervised** approaches generalize better
- Label-based augmentations hurt concept generalization

Reference
Concept generalization in visual representation learning
Mert Bulent Sariyildiz, Yannis Kalantidis, Diane Larlus, Karteek Alahari
ICCV 2021



Proposed **CoG** benchmark

ResNet50	**Baseline model** from the torchvision package (25.5M)
Architecture: *Models with different backbone*	
a-T2T-ViT-t-14	Visual transformer (21.5M)
a-DeiT-S	Visual transformer (22M)
a-DeiT-S-distilled	Distilled *a*-DeiT-S (22M)
a-Inception-v3	CNN with inception modules (27.2M)
a-NAT-M4	Neural architecture search model (7.6M)
a-EfficientNet-B1	Neural architecture search model (7.8M)
a-DeiT-B-distilled	Bigger version of *a*-DeiT-S-distilled (87.6M)
a-ResNet152	Bigger version of ResNet50 (60.2M)
a-VGG19	Simple CNN architecture (143.5M)
Self-supervision: *ResNet50 models trained in this framework*	
s-SimCLR-v2	Online instance discrimination (ID)
s-MoCo-v2	ID with momentum encoder and memory bank
s-SwAV	Online clustering
s-BYOL	Negative-free ID with momentum encoder
s-MoCHi	ID with negative pair mining
s-InfoMin	ID with careful positive pair selection
s-OBoW	Online bag-of-visual-words prediction
s-CompReSS	Distilled from SimCLR-v1 (with ResNet50x4)
Regularization: *ResNet50 models with additional regularization*	
r-MixUp	Label-associated data augmentation
r-Manifold-MixUp	Label-associated data augmentation
r-CutMix	Label-associated data augmentation
r-ReLabel	Trained on a "multi-label" version of IN-1K
r-Adv-Robust	Adversarially robust model
r-MEAL-v2	Distilled ResNet50
Use of web data: *ResNet50 models using additional data*	
d-MoPro	Trained on WebVision-V1 ($\sim 2\times$)
d-Semi-Sup	Pretrained on YFCC-100M ($\sim 100\times$), then fine-tuned on IN-1K
d-Semi-Weakly-Sup	Pretrained on IG-1B ($\sim 1000\times$), then fine-tuned on IN-1K
d-CLIP	Trained on WebImageText ($\sim 400\times$)

Observations

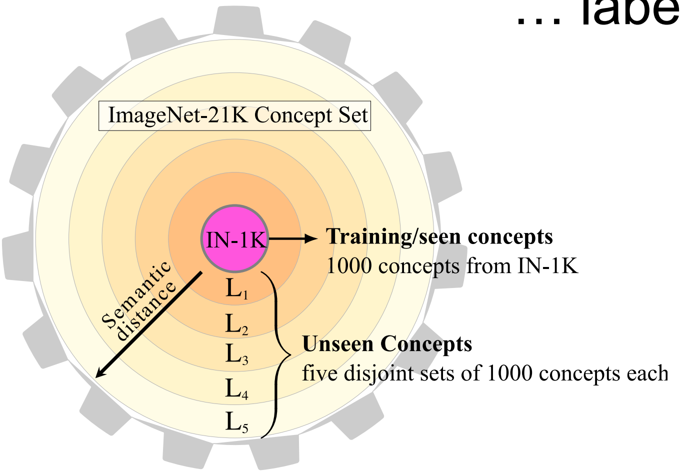
- Recent **self-supervised** approaches generalize better

Yes, but ..

… a good model should shine **both** on

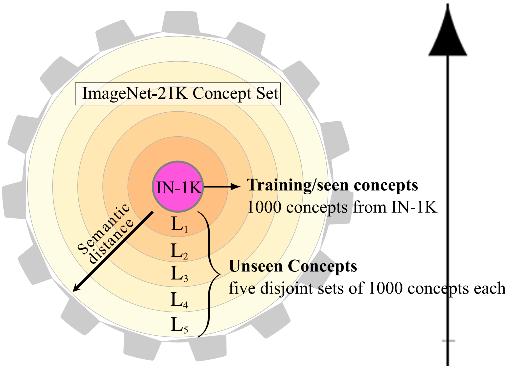
- **Training** task
- **Transfer** tasks

… labels shouldn't hurt



Proposed **CoG** benchmark

Performance **trade-off** between the **training** task and **transfer**



CoG
+
8 datasets

Transfer

… a good model should shine **both** on

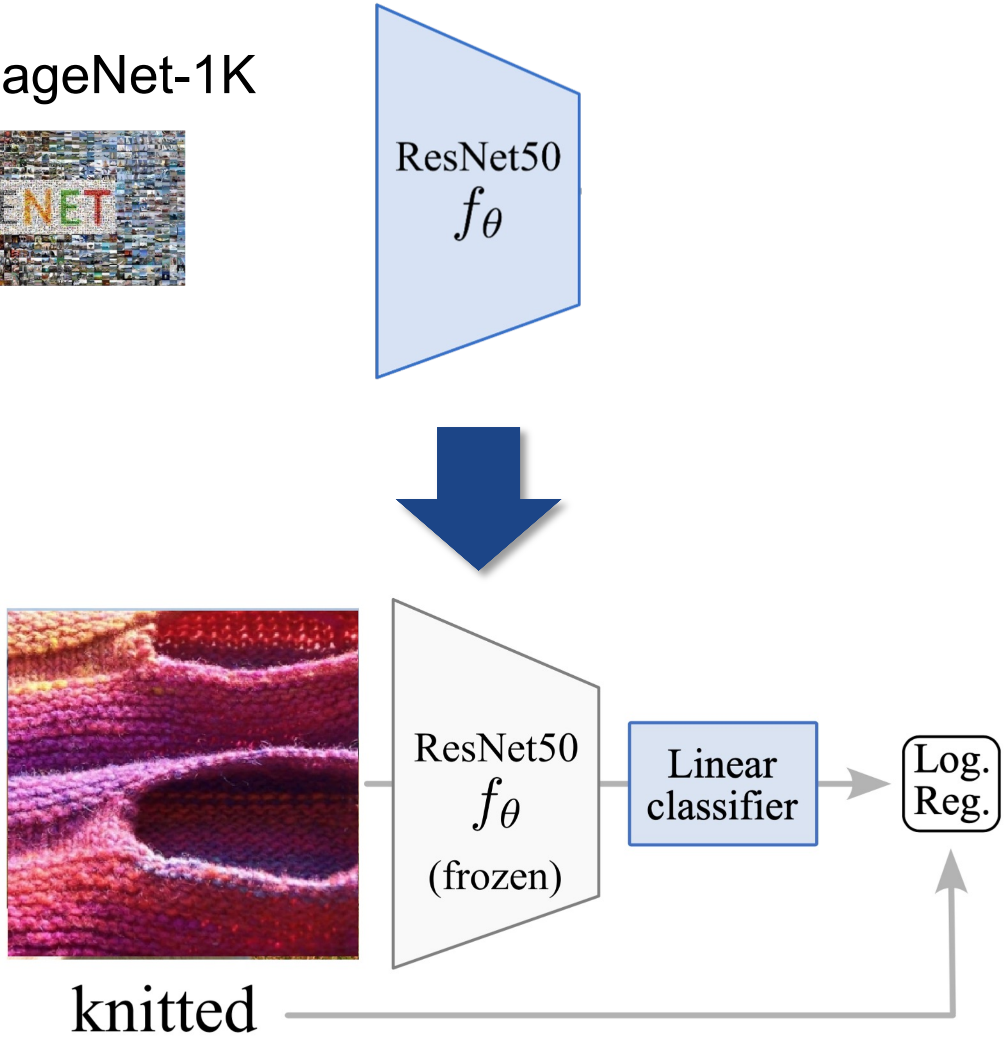
- **Training** task
- **Transfer** tasks

… labels shouldn't hurt

Training

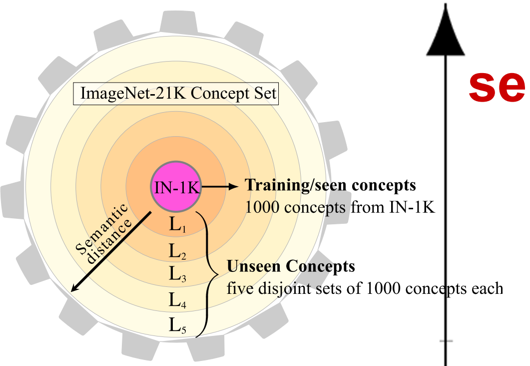
Performance **trade-off** between the **training** task and **transfer**

Train on ImageNet-1K

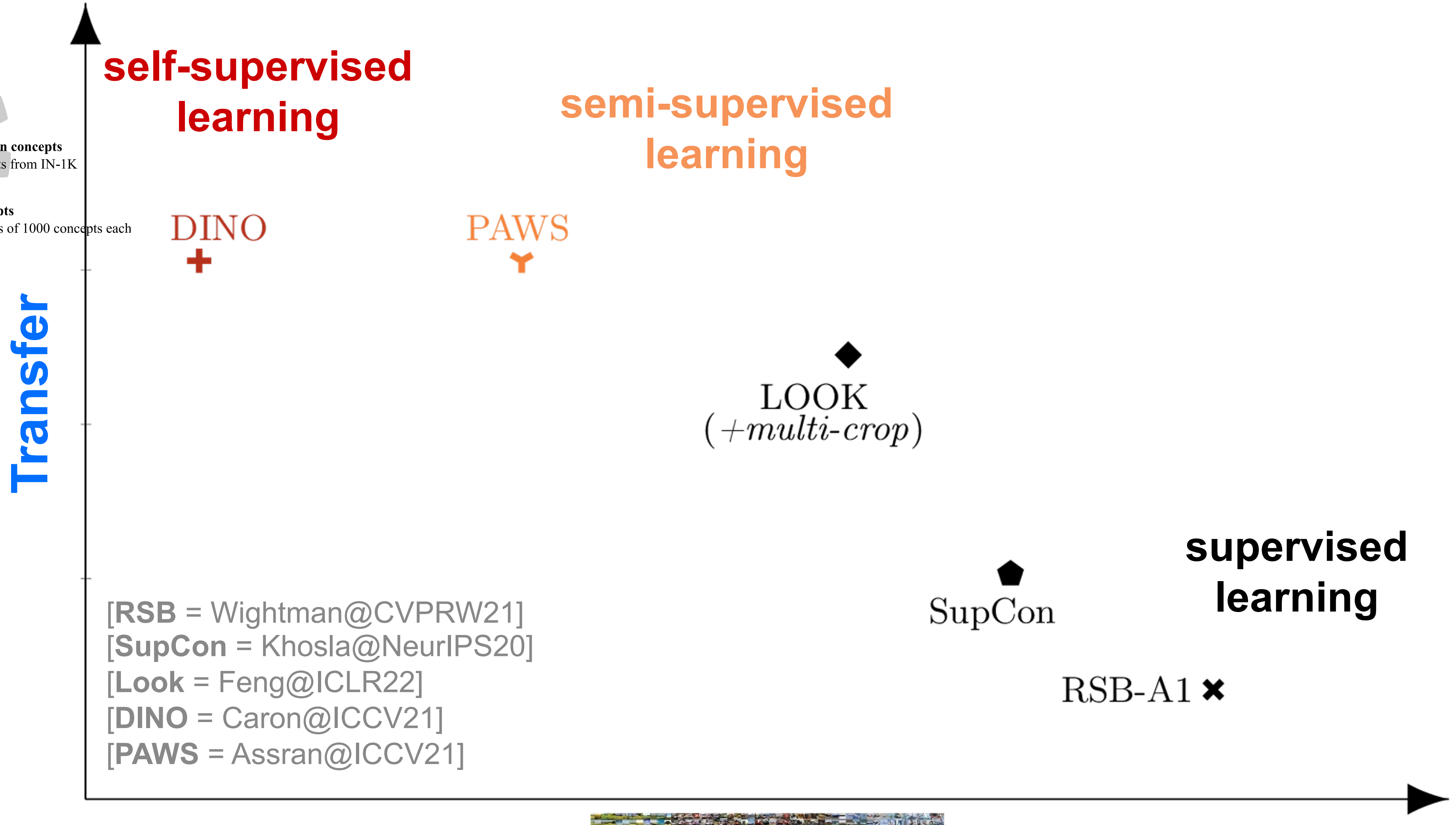


For the **Training** task + every **Transfer** task

Performance **trade-off** between the training task and transfer

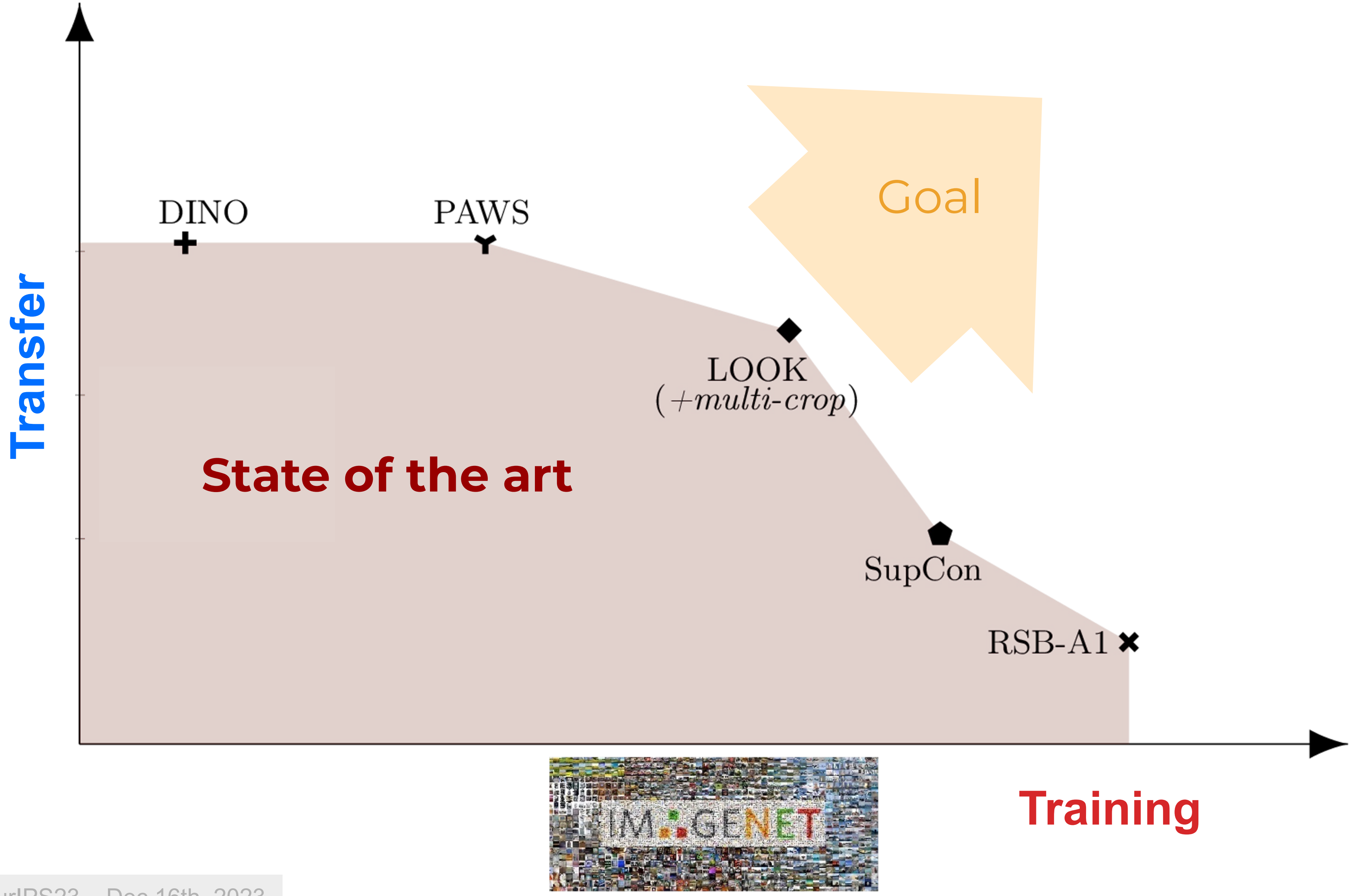


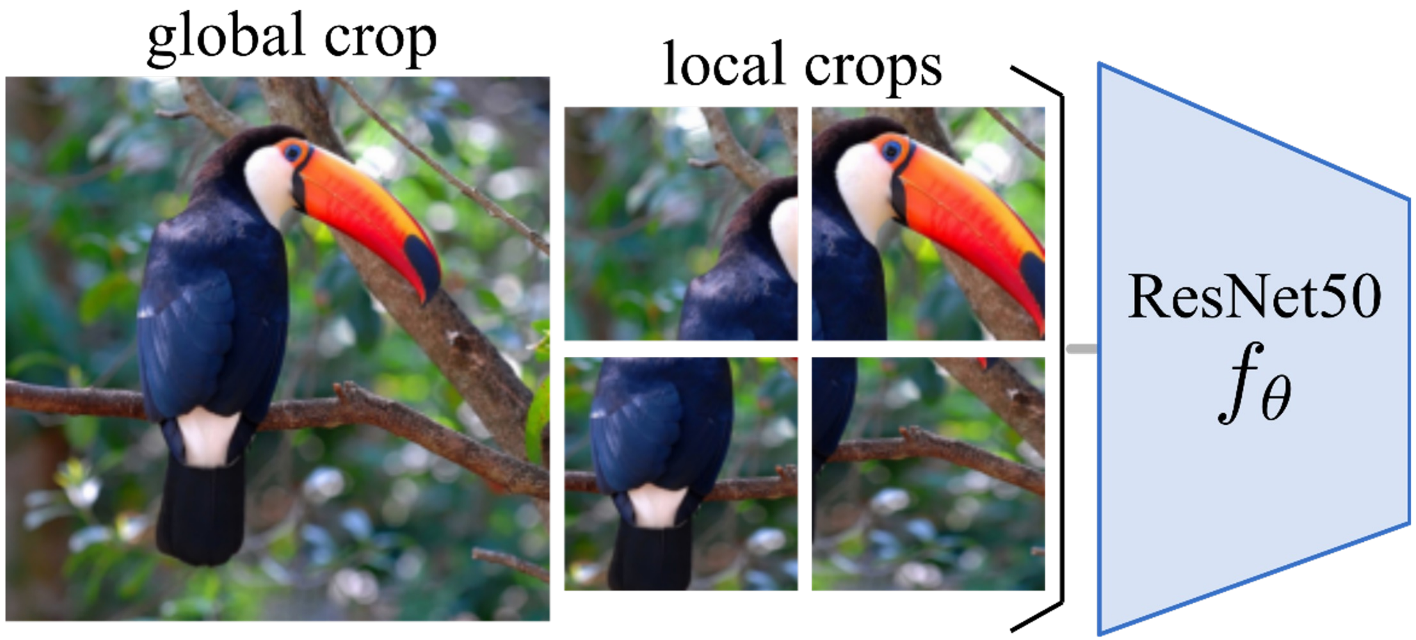
CoG
+
8 datasets



Training

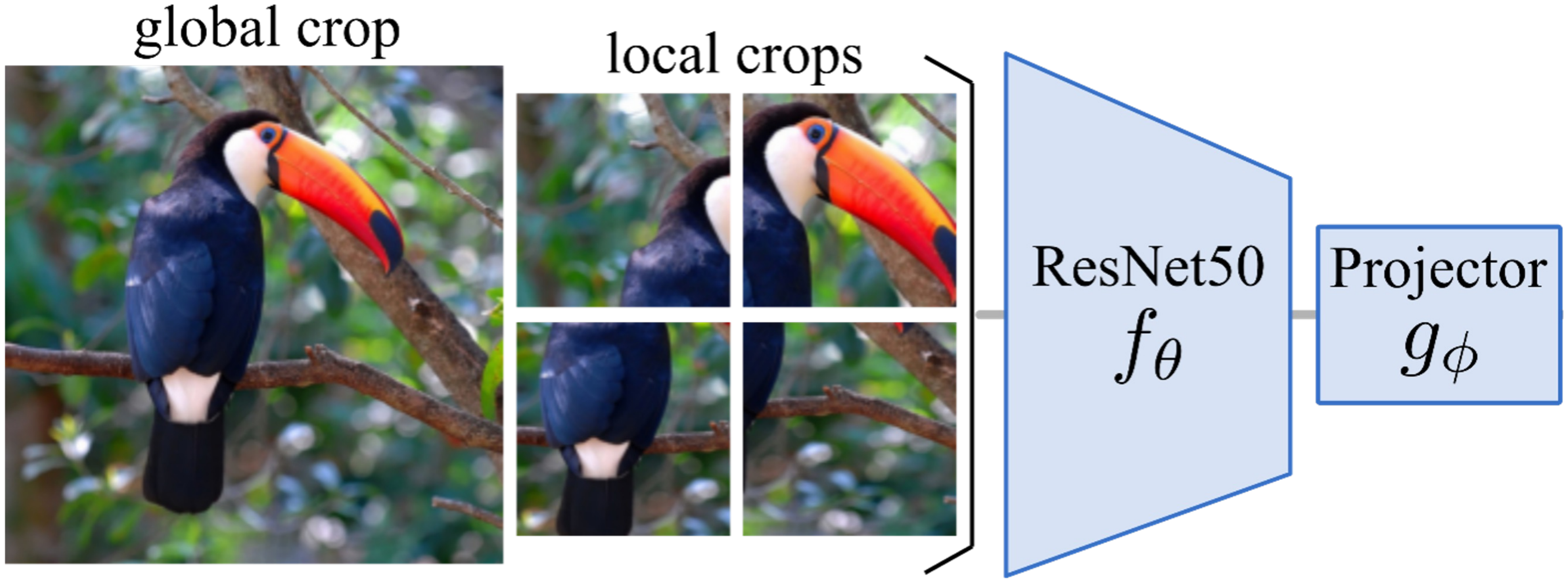
Increasing results both on the training task and transfer





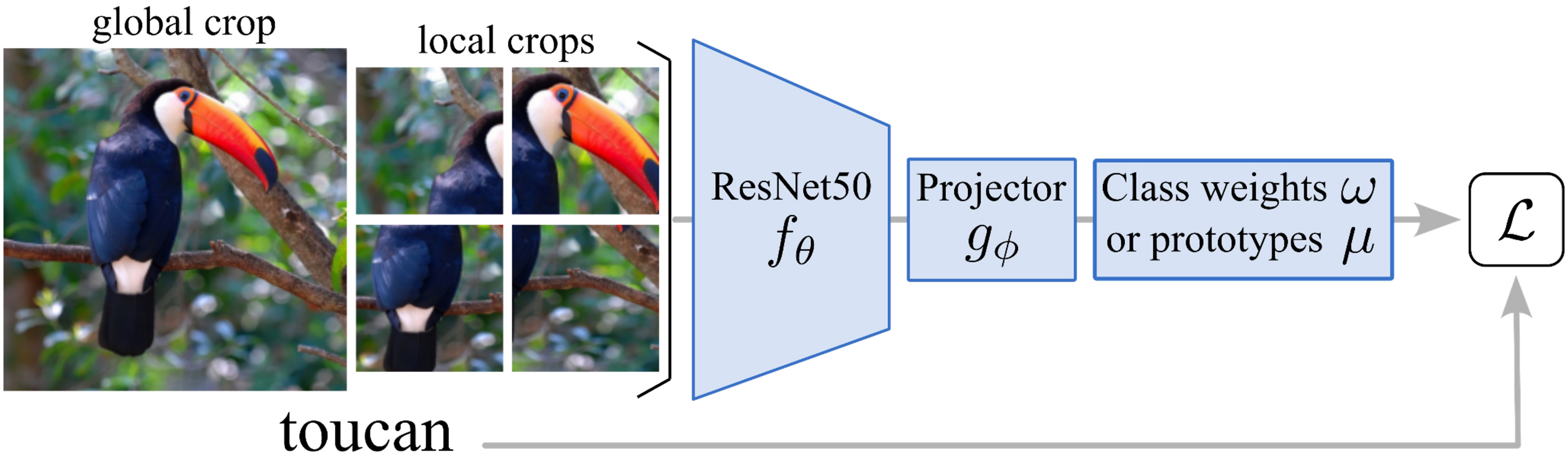
1. Multi-crop data augmentation

[**SWAV** = Caron@NeurIPS20]
[**DINO** = Caron@ICCV21]



1. Multi-crop data augmentation
2. Expendable projector head

[**SimCLR** = Chen@ICML20]
[Wang@CVPR22]

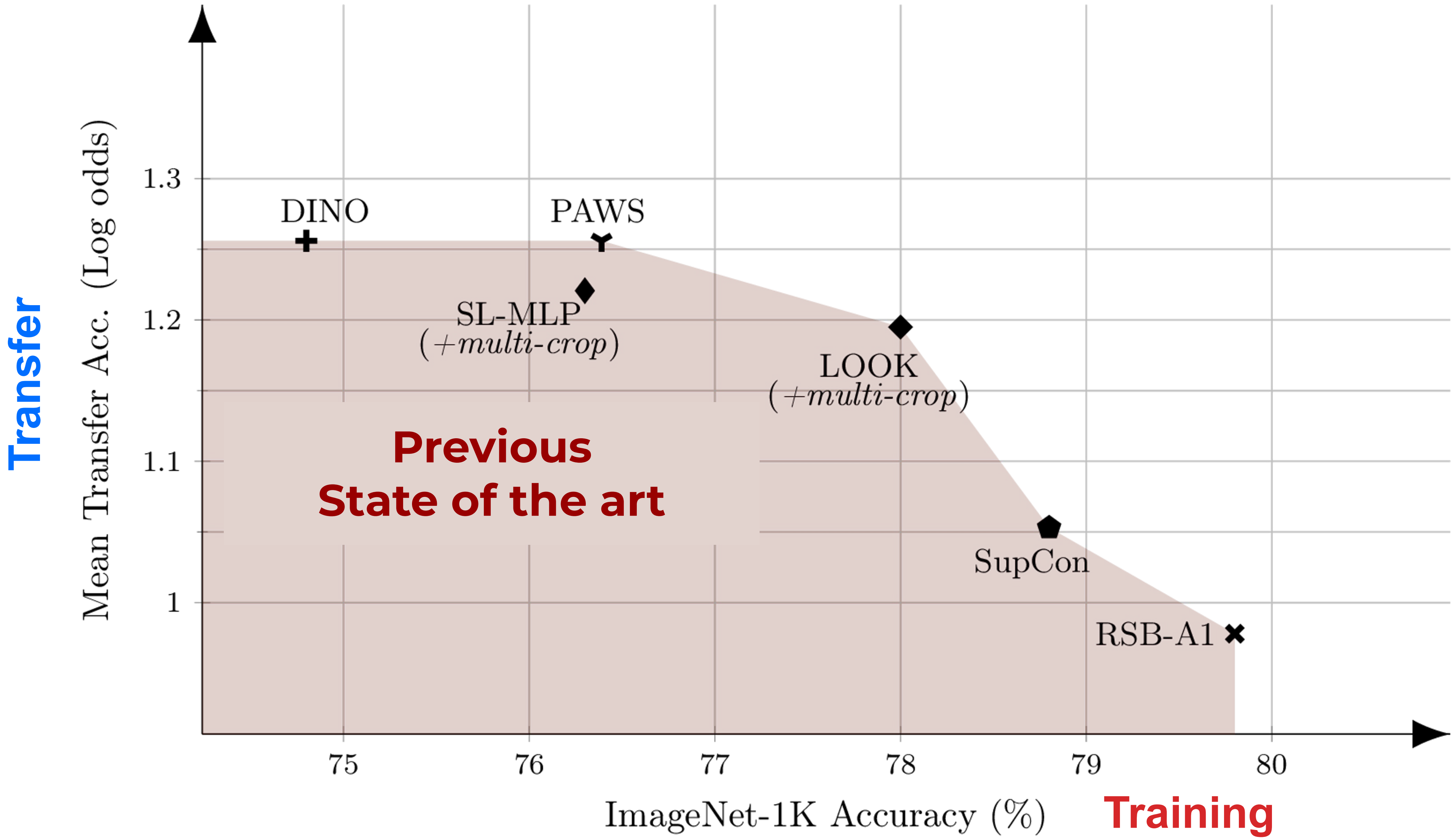


1. Multi-crop data augmentation
2. Expendable projector head
3. (*optional*) Replace class weights with class prototypes

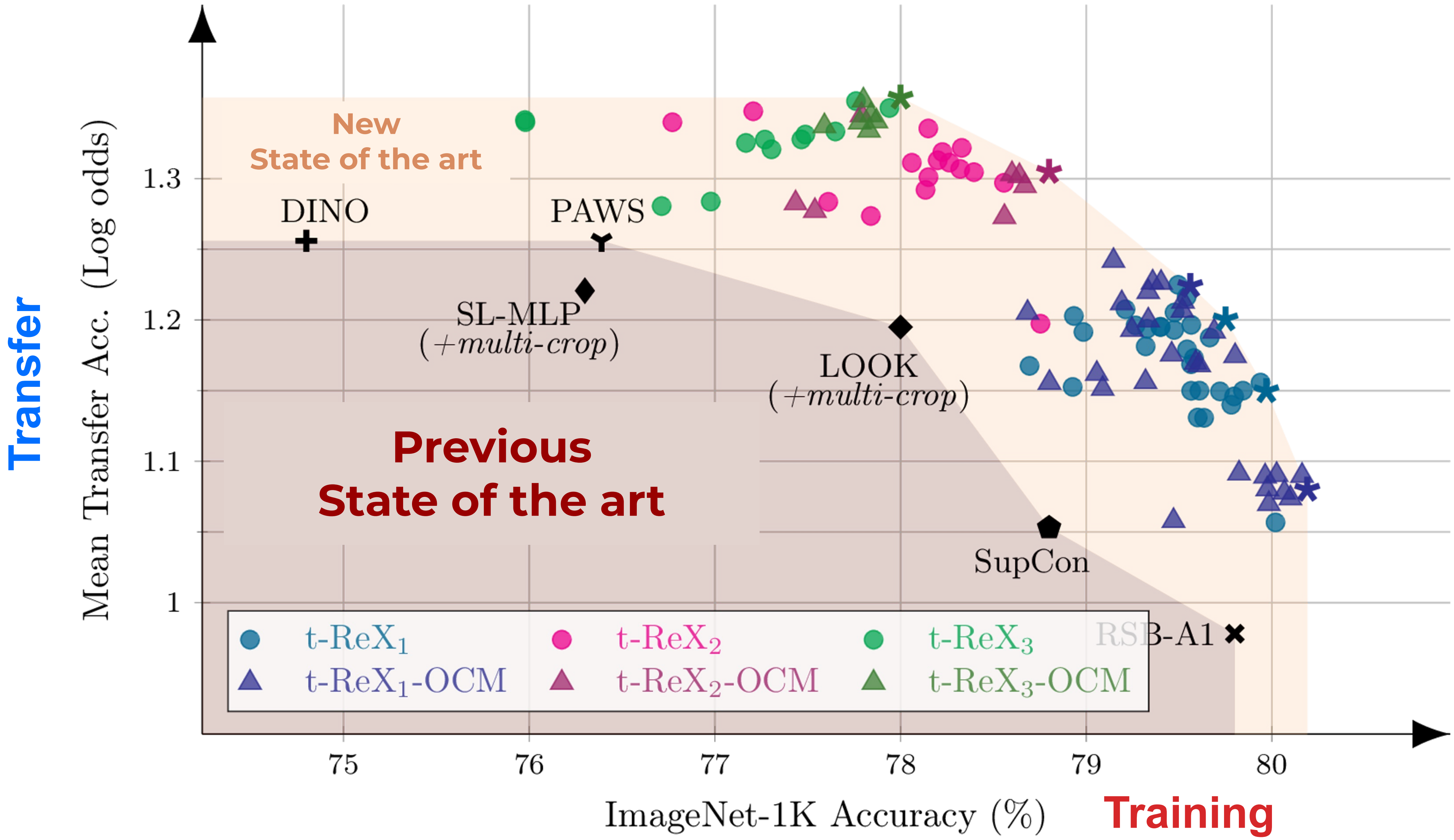
Nearest Class Means (NCM)

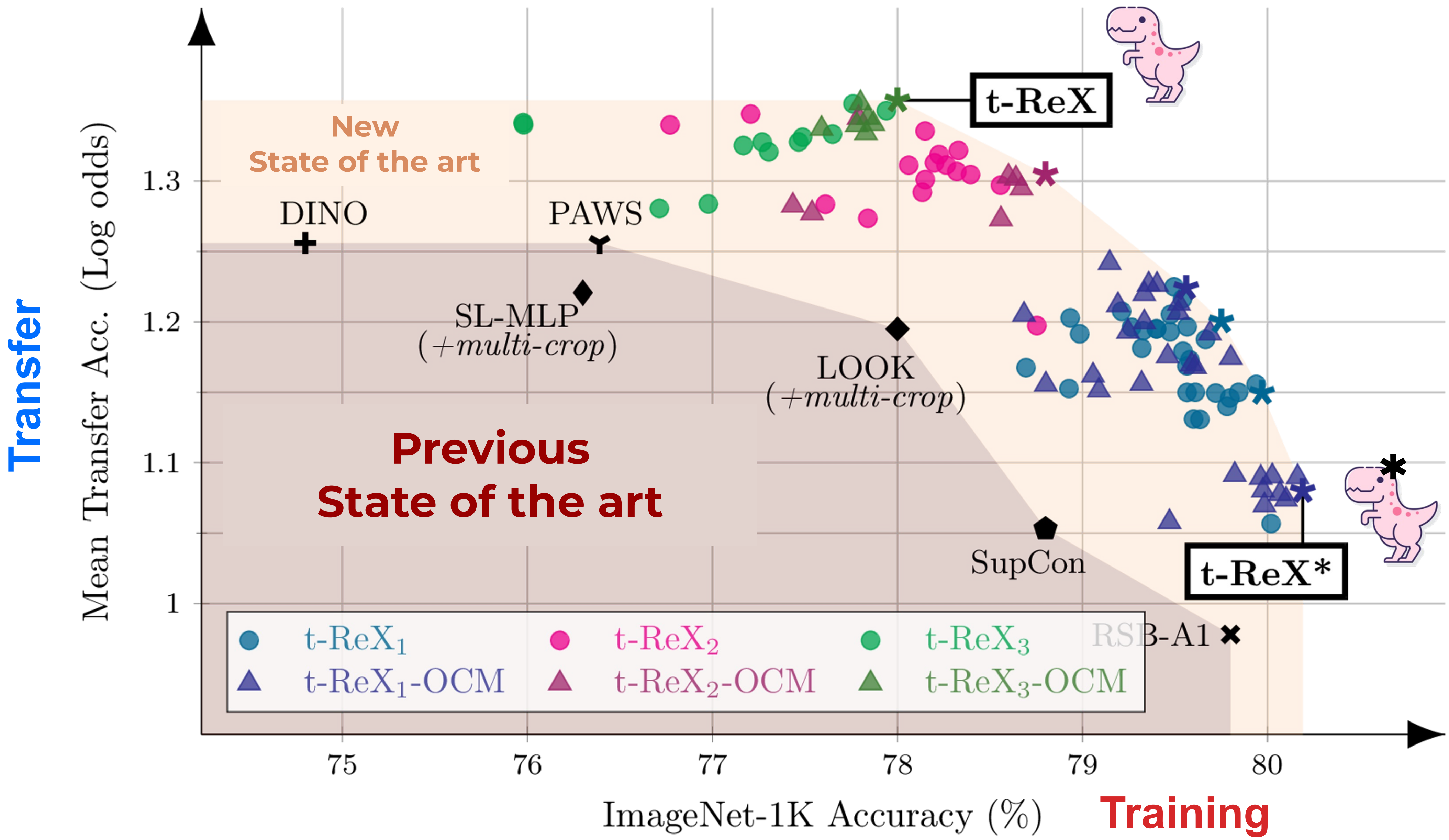
[**NCM** = Mensink@ECCV12]
[**DeepNCM** = Guerriero@W-ICLR18]

Comparison with the SOTA



Comparison with the SOTA




Transfer
Mean Transfer Acc. (Log odds)
1.3
1.2
1.1
1
New
State of the art
DINO
PAWS
SL-MLP
(+multi-crop)
LOOK
(+multi-crop)
Previous
State of the art
SupCon
RSB-A1
t-ReX
t-ReX*
t-ReX1
t-ReX2
t-ReX3
t-ReX1-OCM
t-ReX2-OCM
t-ReX3-OCM
75
76
77
78
79
80
ImageNet-1K Accuracy (%)
Training

Take home message

T-ReX

T-Rex is **state of the art** for **Transfer** “despite” being supervised

- Multi-crop data augmentation helps
- Expendable projector controls **Training** / **Transfer** trade-off

Reference
No Reason for No Supervision: Improved Generalization in Supervised Models
Mert Bulent Sariyildiz, Yannis Kalantidis, Karteek Alahari, Diane Larlus
ICLR 2023

Proxy task

Reducing annotation cost

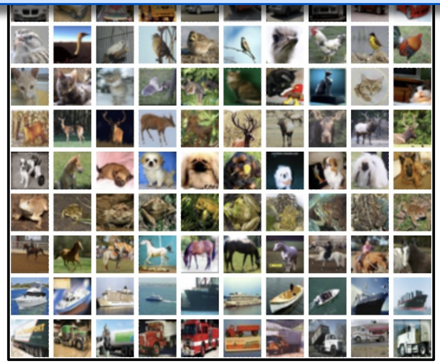
Fully-Supervised Classification

Images + labels

Model

Visual representations

IMAGENET



Target tasks

Image Classification

Object Detection

Instance Segmentation

Image Retrieval

Reducing annotation cost

Fully-Supervised
fine-grained annotations

Eurasian finch | Ruddy turnstone | Dowitcher

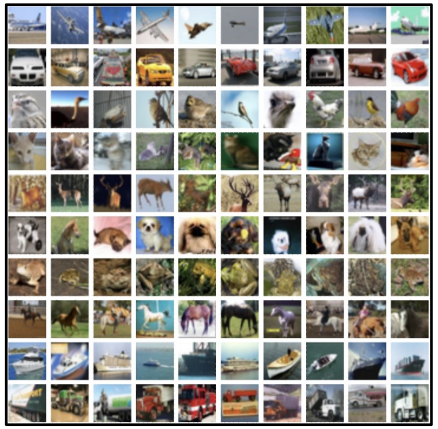
Brain coral | Coral reef | Coral fungus

Caption-supervised
side information

a statue of a man stands in front of an old red bus.
a big and red bus with many displays for people to watch.
a red double decker bus parked next to a statue.
the double decker bus is beside a statue near restaurant tables.
a view of a bus sitting in front a small wooden statue.

a busy street with cars and trucks down it
an intersection with a view that looks towards a small downtown area.
cars parked on the side of the street and traveling down the road
an intersection with a stop light on a city street.
a street filled with lots of traffic under a traffic light.

Self-supervised
annotation-free images



Weak annotations

Reducing annotation cost

Fully-Supervised
fine-grained annotations

Eurasian finch | Ruddy turnstone | Dowitcher

Brain coral | Coral reef | Coral fungus

Caption-supervised
side information

a statue of a man stands in front of an old red bus.
a big and red bus with many displays for people to watch.
a red double decker bus parked next to a statue.
the double decker bus is beside a statue near restaurant tables.
a view of a bus sitting in front a small wooden statue.

a busy street with cars and trucks down it
an intersection with a view that looks towards a small downtown area.
cars parked on the side of the street and traveling down the road
an intersection with a stop light on a city street.
a street filled with lots of traffic under a traffic light.

Self-supervised
annotation-free images

Learning transferable visual representations

Input: Image

Caption

"Little girl holding red umbrella"

Mask a token

"Little girl holding red [MASK]"

Visual representation
(learnt from scratch)

Textual representation

Multi-modal network =
Auxiliary modules

[MASK] = Umbrella

[**ICMLM** = Sariyildiz@ECCV20]

[**VirTex** = Desai@CVPR21]

Learning transferable visual representations

Weak annotations

Reducing annotation cost

Dataset scale

[**ICMLM** = Sariyildiz@ECCV20]

[**VirTex** = Desai@CVPR21]

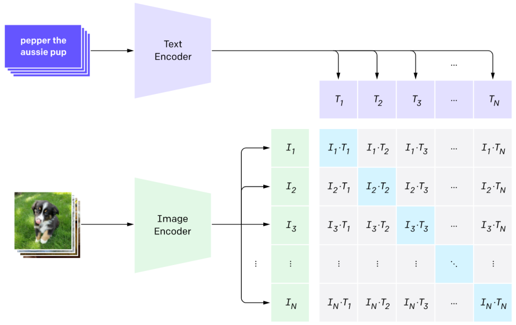
Caption-supervised

side information

smaller sets

a statue of a man stands in front of an old red bus.
a big and red bus with many displays for people to watch.
a red double decker bus parked next to a statue.
the double decker bus is beside a statue near restaurant tables.
a view of a bus sitting in front a small wooden statue.

[**CLIP** = Radford@ICLM21]



Unfiltered

Image + Text

large scale

Text-to-image generation

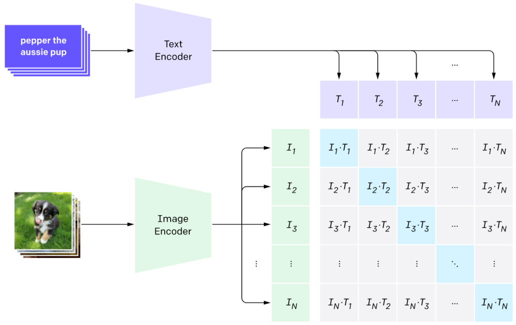
[**DALL-E** = Ramesh@ICML21]

[**DALL-E2** = Saharia@NeurIPS21]

[**DALL-E3** = Betker@Website23]

[**Stable diffusion** = Rombach@CVPR22]

[**CLIP** = Radford@ICLM21]



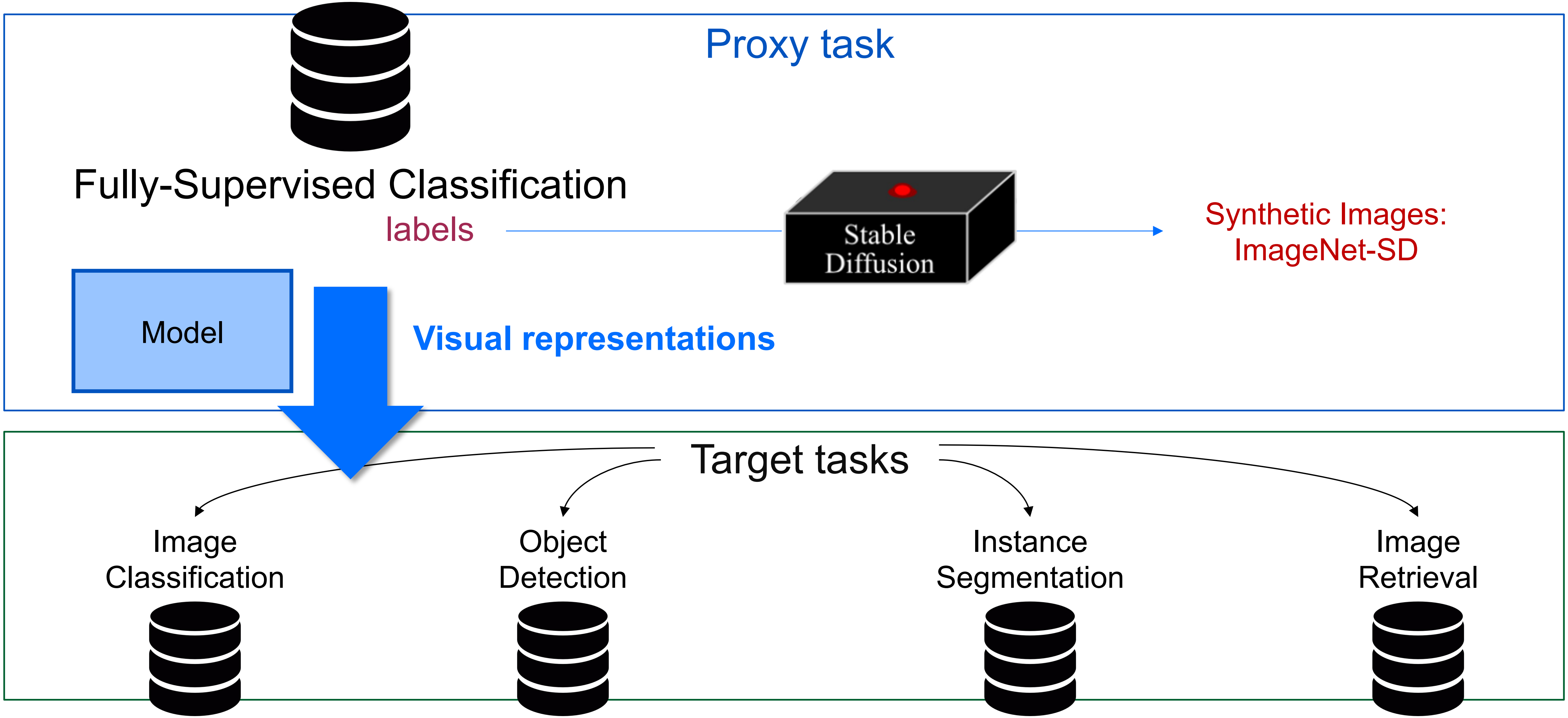
Unfiltered
Image + Text
large scale

Stable Diffusion

cat

[**Stable Diffusion** = Rombach@CVPR22]

Do we still need actual images
to pretrain visual representations?


Proxy task
Fully-Supervised Classification
labels
Stable
Diffusion
Synthetic Images:
ImageNet-SD
Model
Visual representations
Target tasks
Image
Classification
Object
Detection
Instance
Segmentation
Image
Retrieval

Generating images – the naïve solution

prompt = class name

Synthetic Image

papillon

lorikeet

pirate ship

Semantic errors

Lack of diversity

Domain issues

"papillon" class in ImageNet

"pirate ship" class in ImageNet

Generating images – improving the prompt

prompt = class name

prompt = class name, hypernym*

prompt = class name, description*

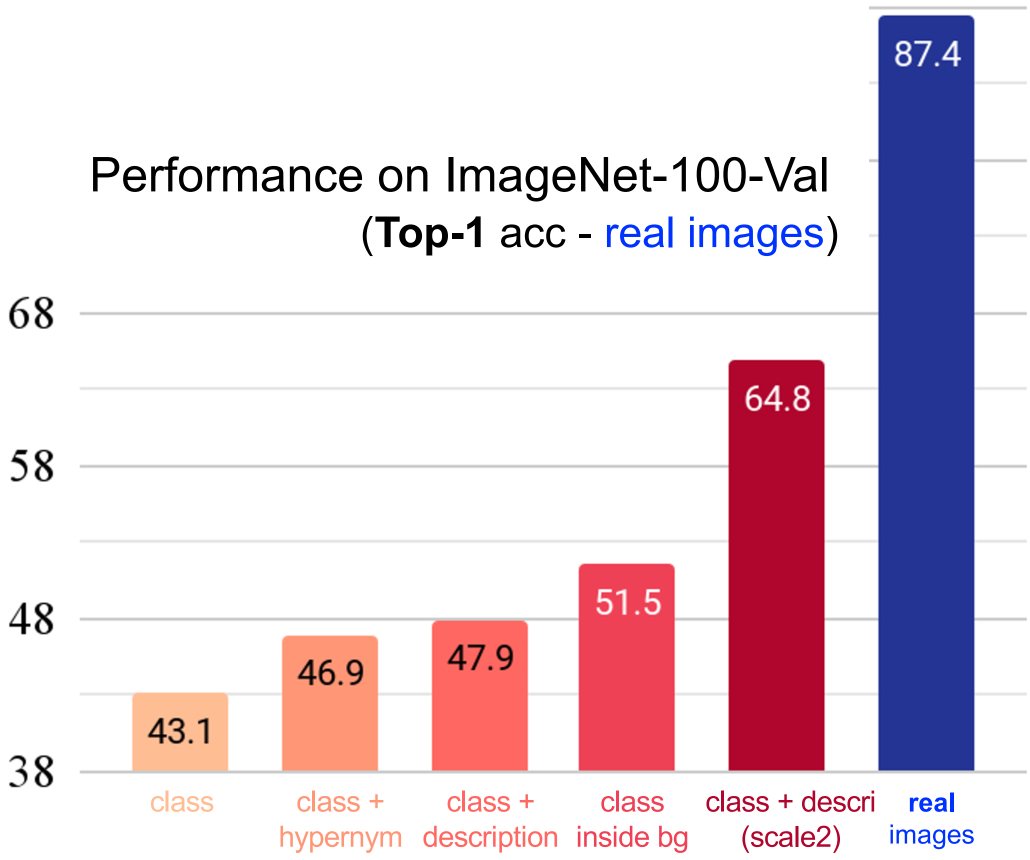
prompt = class name, hypernym inside background**

prompt = class name, description (+ reduce guidance scale)

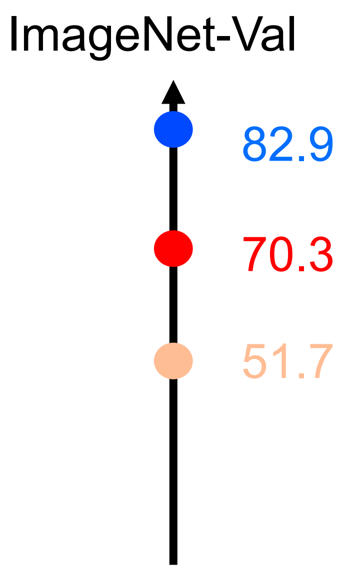
* from **Wordnet** lexical database

** from **Places 365** dataset

How well does each model perform when classiying real images?



How well does each model perform when classiying real images?



Top-5 acc

Model trained on ImageNet-1k

ImageNet-SD (better prompt)

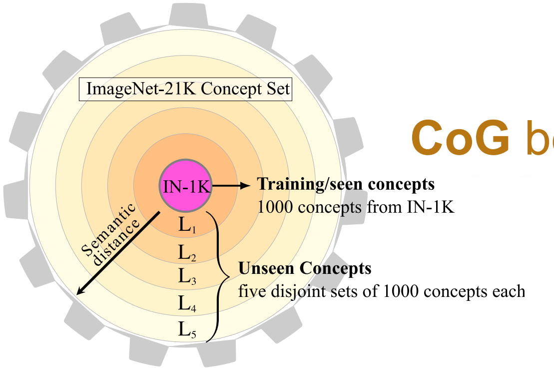
ImageNet-SD (naïve) name

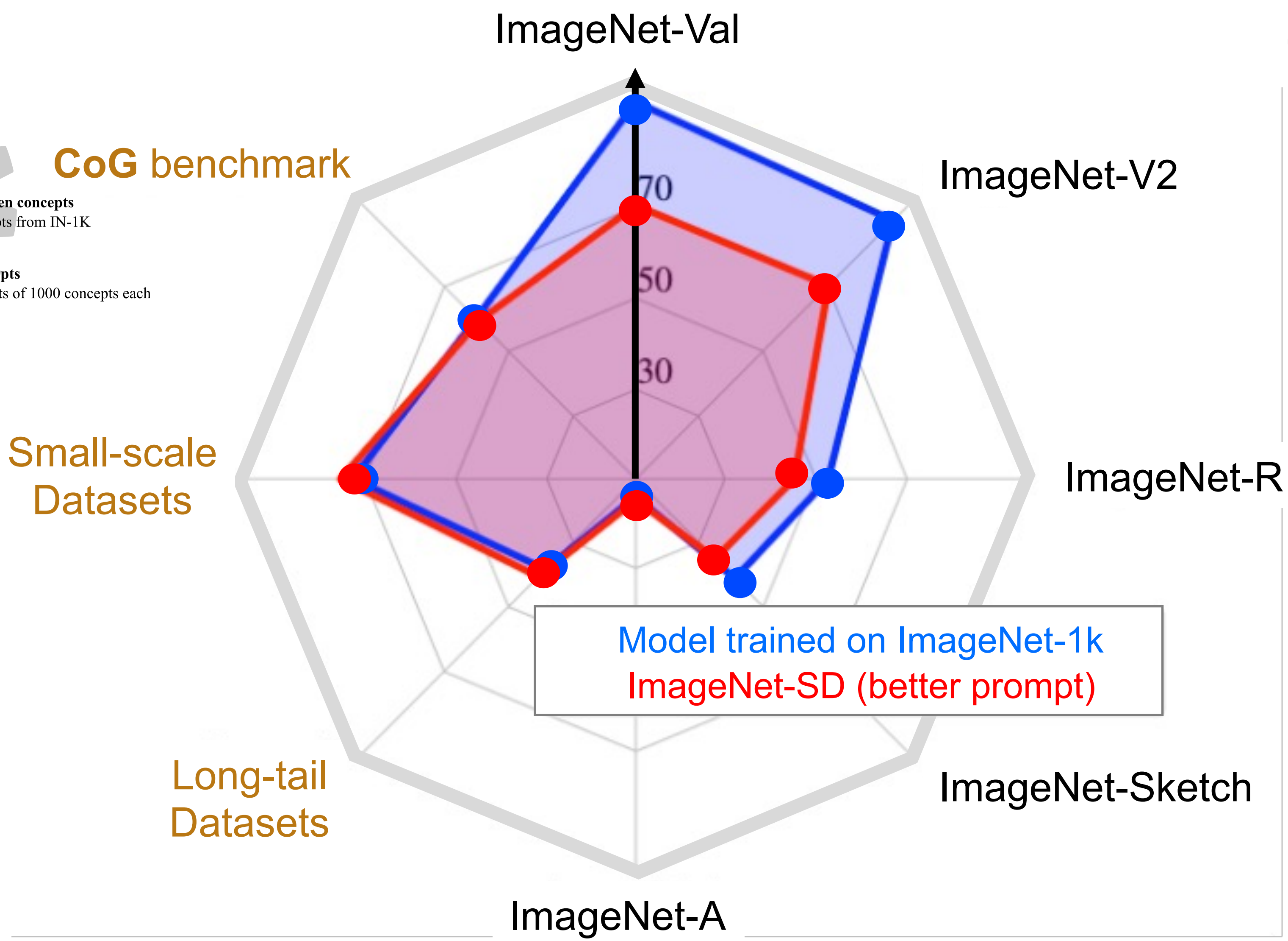
prompt = class name

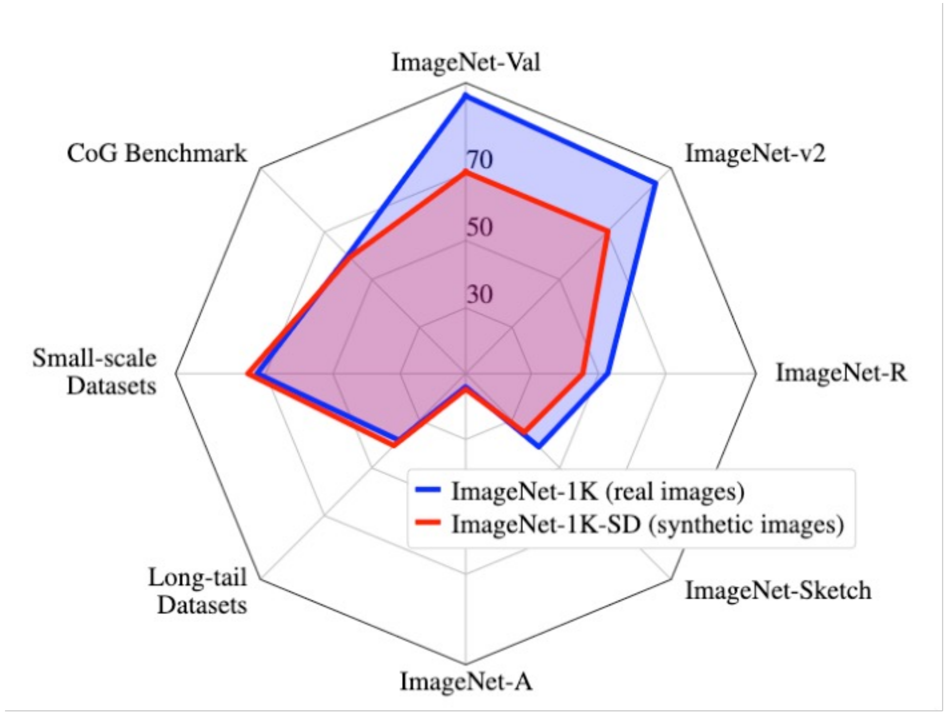
prompt = class name, description (+ reduce guidance scale)

Training with synthetic images – evaluation

For ImageNet variants:
Top-5 acc







Do we still need actual images to pretrain visual representations?

- Promising results on the ImageNet variants
- Strong transfer results

Reference
Fake it till you make it: Learning transferable representations from synthetic ImageNet clones
Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, Yannis Kalantidis
CVPR 2023

Proxy task

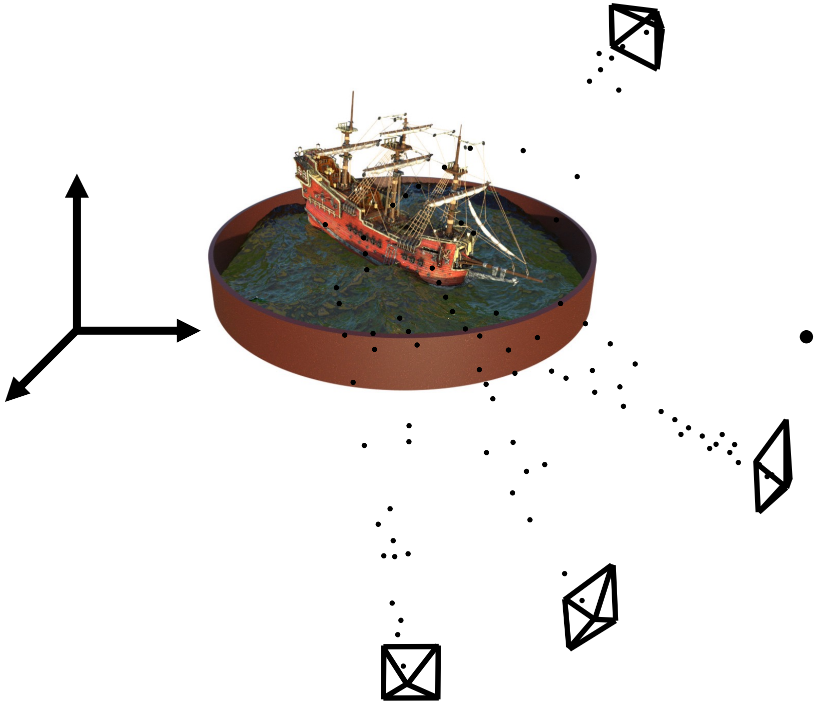
Fully-supervised classification or
Self-supervised approaches, etc.

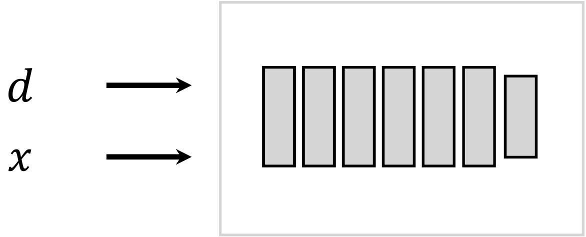
Model

Visual representations

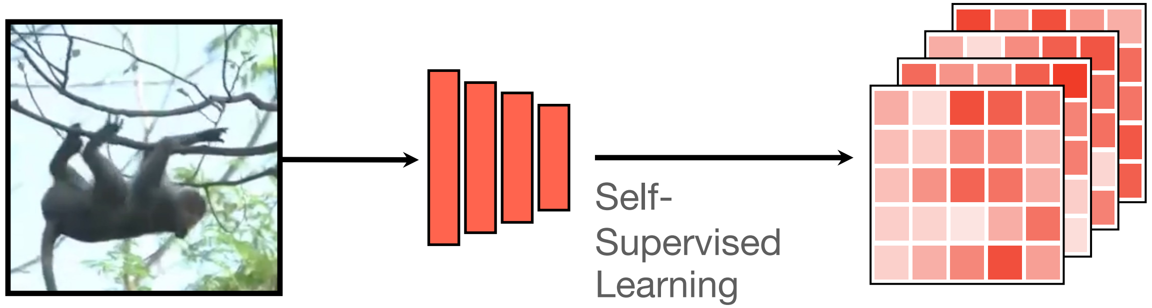
Target tasks

What about 3D tasks?





[**NERF** = Mildenhall et al. ECCV20]

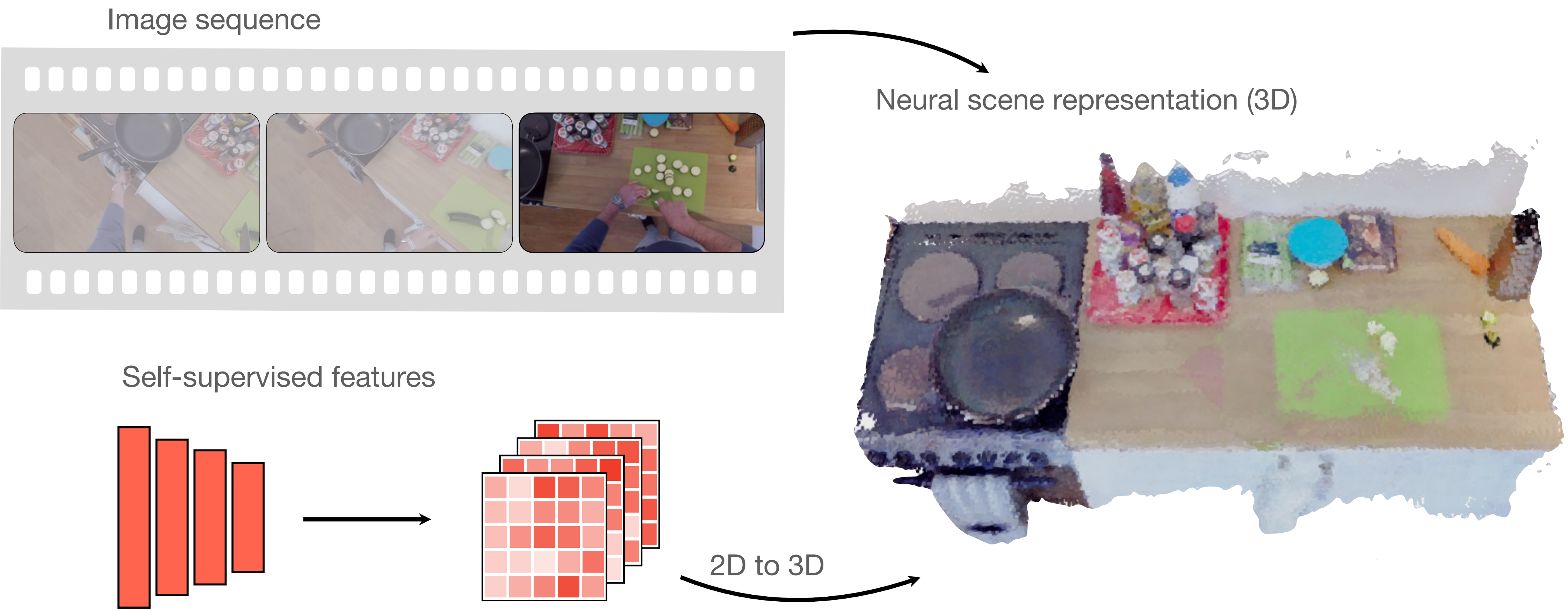


[**DINO**: Caron @ ICCV21]

Proposed
Neural Feature Fusion Fields

N3F

Fusing Image-Level and 3D Scene Representations



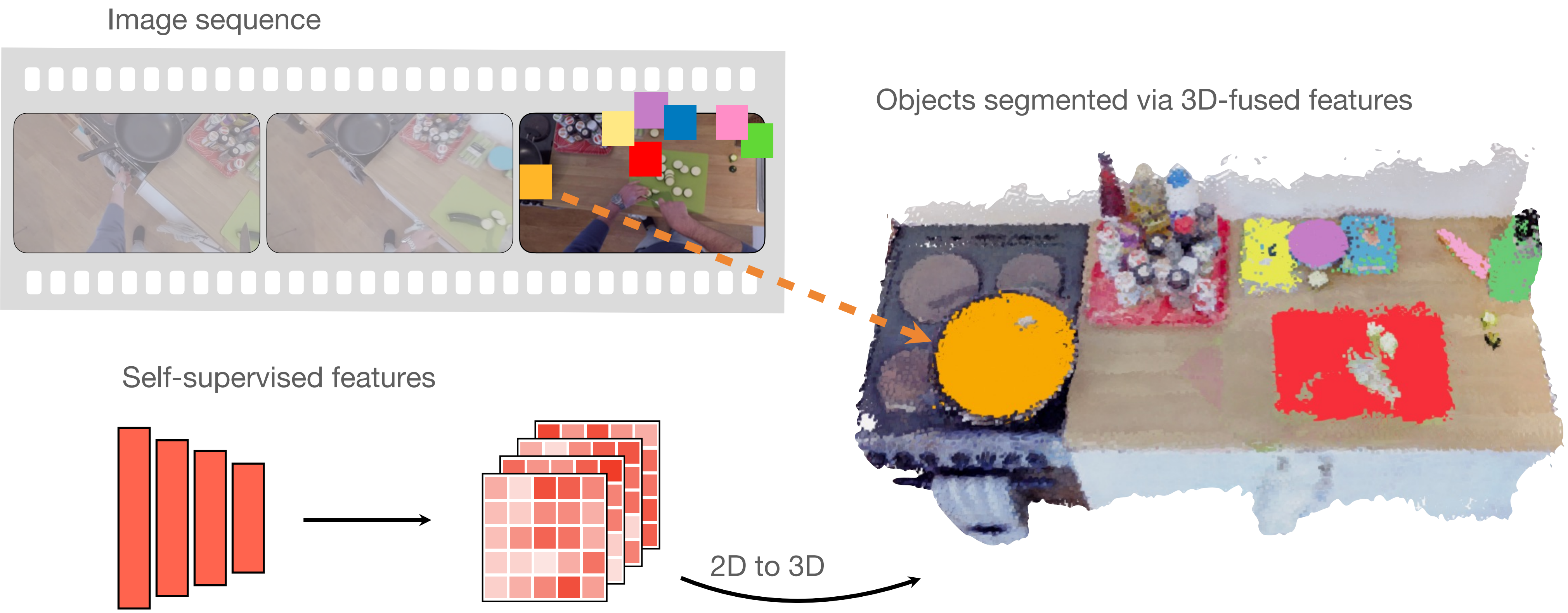
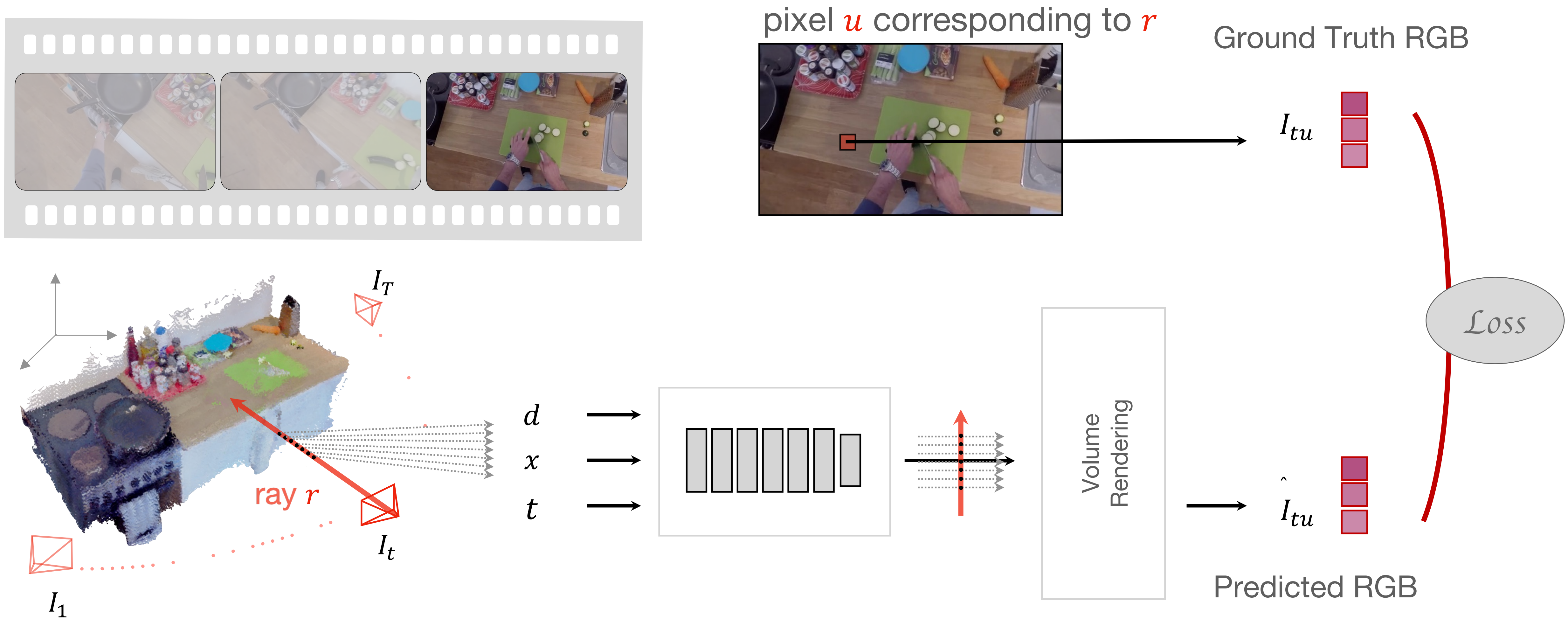
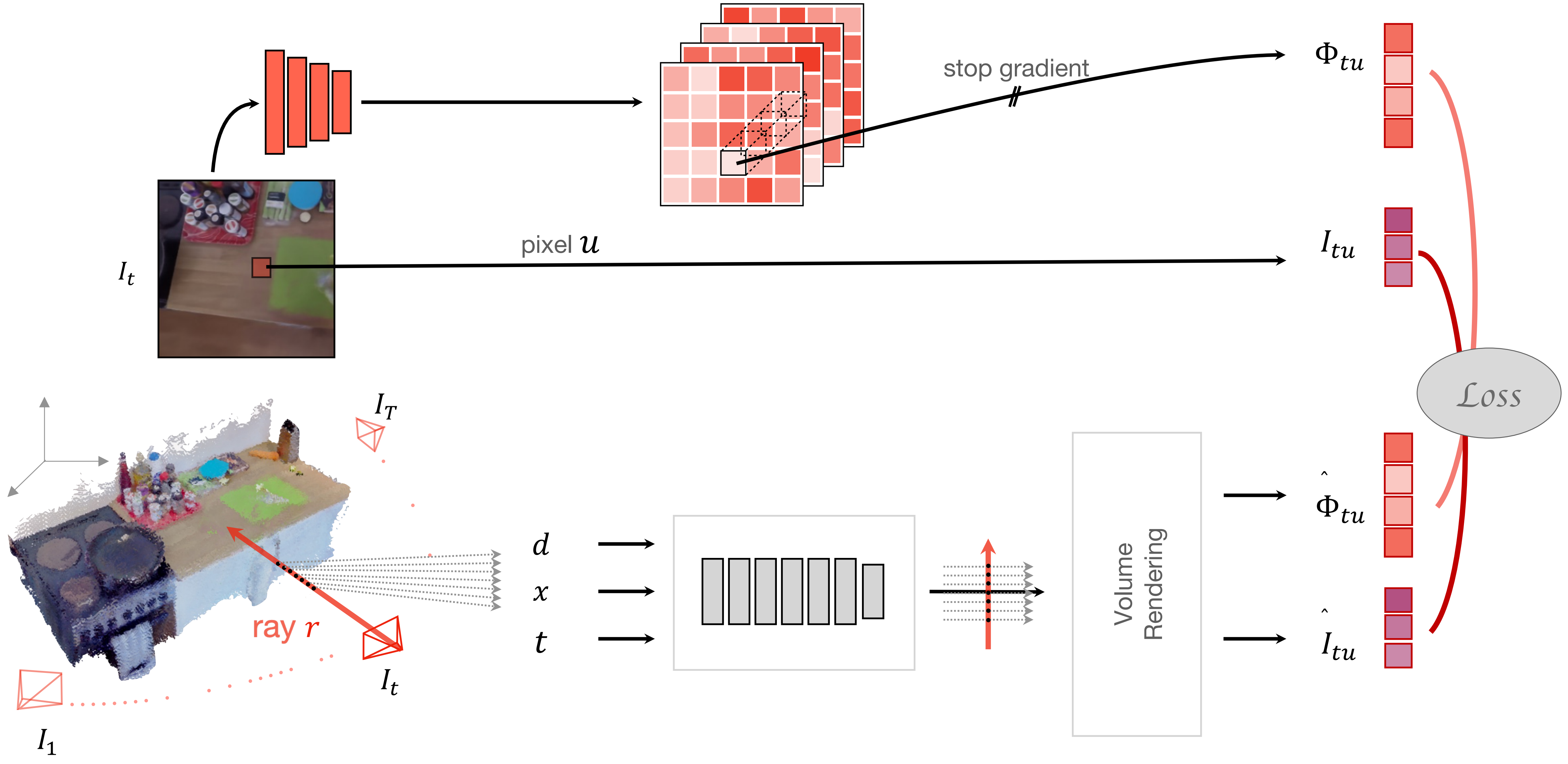


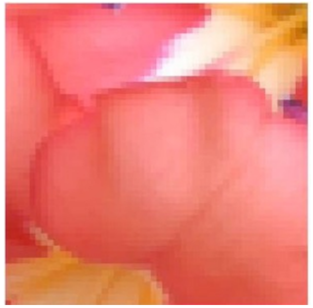
Image sequence
Objects segmented via 3D-fused features
Self-supervised features
2D to 3D


pixel u corresponding to r
Ground Truth RGB
I_{tu}
I_T
d
x
t
ray r
I_t
I_1
Volume Rendering
$\hat{I}_{tu}$
Loss
Predicted RGB

Neural Feature Fusion Fields (N3F)



Query



Full radiance field	Foreground	Background
		

Concurrent work: Kobayashi et al. Decomposing NeRF for Editing via Feature Field Distillation. NeurIPS22.

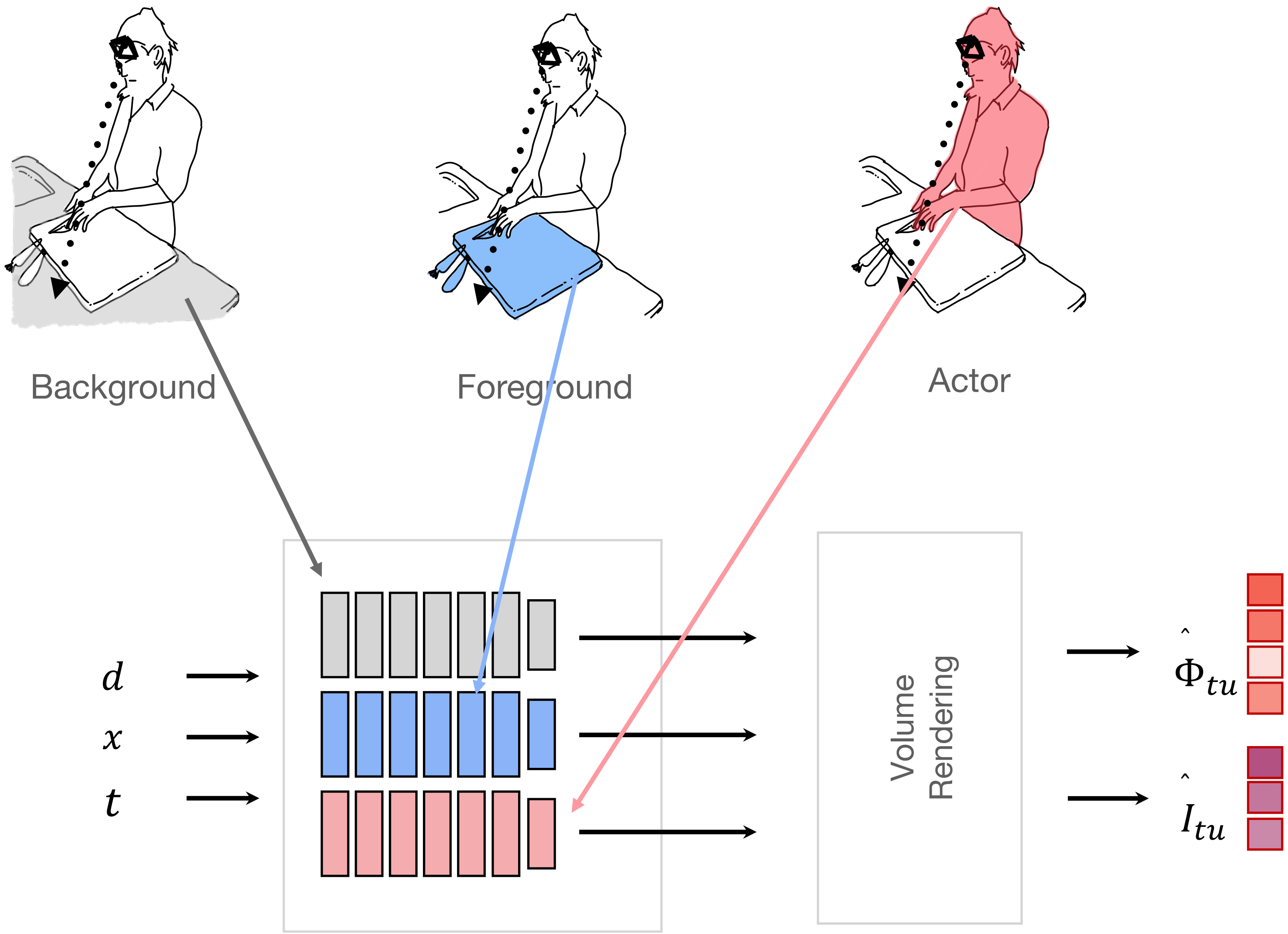
Applying **N3F** to Dynamic Scenes

- Objects moved frequently
- Actor is heavily occluding the scene

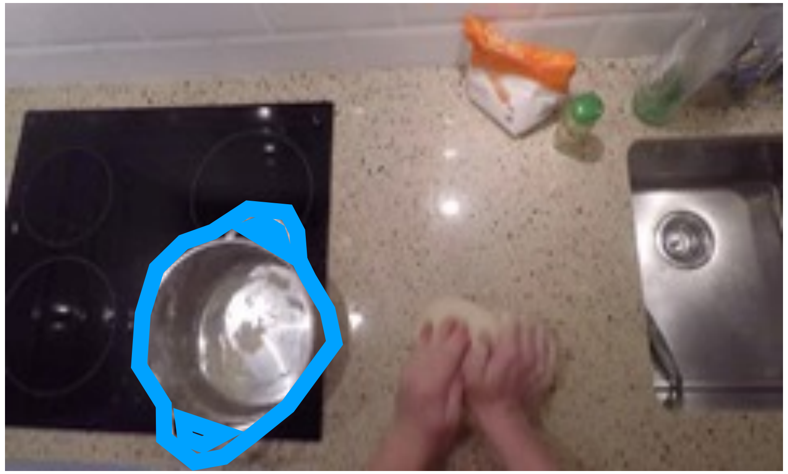
[**EPIC-KITCHENS** = Damen@IJCV21]

[**NeuralDiff** = Tschernezki@3DV21]

Applying **N3F** to Dynamic Scenes



[**NeuralDiff** = Tschernezki@3DV21]



Query

With object

Without object (edited)

Distillation of 2D self-supervised features into 3D scenes

Works for static scenes as well as for complex egocentric scenes

Potential applications: object retrieval, scene editing, language guided manipulation [F3RM: Shen@CoRL23]

Reference
Neural Feature Fusion Fields (N3F): 3D Distillation of Self-Supervised 2D Image Representations
Vadim Tschernezki, Iro Laina, Diane Larlus, Andrea Vedaldi
3DV 2022

Thanks!

Joint work with ..

Concept generalization in visual representation learning
Mert Bülent Sariyildiz, Yannis Kalantidis, Diane Larlus, Karteek Alahari
International Conference in Computer Vision (**ICCV**) 2021

No Reason for No Supervision: Improved Generalization in Supervised Models
Mert Bülent Sariyildiz, Yannis Kalantidis, Karteek Alahari, Diane Larlus
International Conference in Representation Learning (**ICLR**) 2023

Fake it till you make it: Learning transferable representations from synthetic ImageNet clones
Mert Bülent Sariyildiz, Yannis Kalantidis, Diane Larlus, Karteek Alahari
Conference in Computer Vision and Pattern Recognition (**CVPR**) 2023

Neural Feature Fusion Fields (N3F): 3D Distillation of Self-Supervised 2D Image Representations
Vadim Tschernezki, Iro Laina, Diane Larlus, Andrea Vedaldi
International Conference on 3D Vision (**3DV**) 2022


Bülent Sariyildiz


Karteek Alahari


Yannis Kalantidis


Vadim Tschernezki


Iro Laina


Andrea Vedaldi

Credit icons: https://www.flaticon.com/free-icons

End