Data-centred Deep Learning Models for Food Fine-grained recognition

Petia Radeva
Full professor
Head of “Artificial Intelligence and Biomedical Applications”
Consolidated Research Group,
Universitat de Barcelona, Spain

- Joint work with **Eduardo Aguilar**. **Bhalaji Nagarajan, Imanol Gonzalez, Jesus Molina, Ricardo Marquez, etc**

UNIVERSITAT DE BARCELONA

Index

1. Why Food Recognition?

2. Self-Supervised Learning for Fine-Grained Recognition

    - Validation of All4One

3. Other Food Recognition works

4. Food Recognition real applications

What are the most popular datasets today?

Dataset	Papers	Benchmarks	Images (K)	Classes	Sizes
Cifar-10	10581	66	60	10	32x32
ImageNet	10046	97	**1400**	**1000**	variable
COCO	7160	78	123	80	
OpenImages			**9000**	600	
MNIST	5911	49	60	10	28x28
Cifar-100	5322	42	60	100	32x32
Cityskapes	2562	37	25	8	
SVHN	2474	11	60	10	32x32
Kitti	2453	120	0,5	11	
CelebA	2408	20	202	10177	178x218
Fashion-MNIST	2150	17	70	10	28x28
CUB-00-2011	2408	37	12	200	
Places	760	4	**2500**	205	
Tiny ImageNet	516	7	31	200	
Places205	468	1	**2500**	205	
Caltech-101	393	6	5	101	300x200
Stanford Cars	392	8	16	196	360x240
Caltech-256	345	4	30	257	

https://www.analyticsvidhya.com/blog/2018/03/comprehensive-collection-deep-learning-datasets/

Large Scale Food Recognition Dataset

Journals & Magazines > IEEE Transactions on Pattern ... > Early Access

Large Scale Visual Food Recognition

Publisher: IEEE Cite This PDF

Weiqing Min ; Zhiling Wang ; Yuxin Liu ; Mengjiang Luo ; Liping Kang ; Xiaoming Wei ; Xia... **All Authors**

Abstract

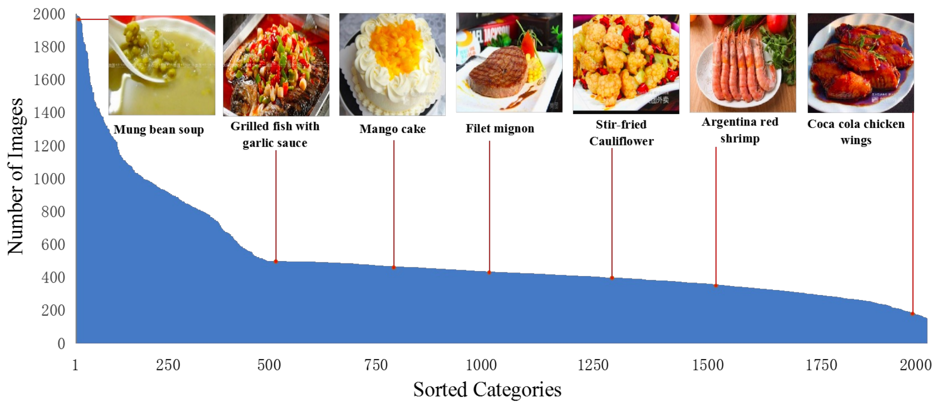
Authors

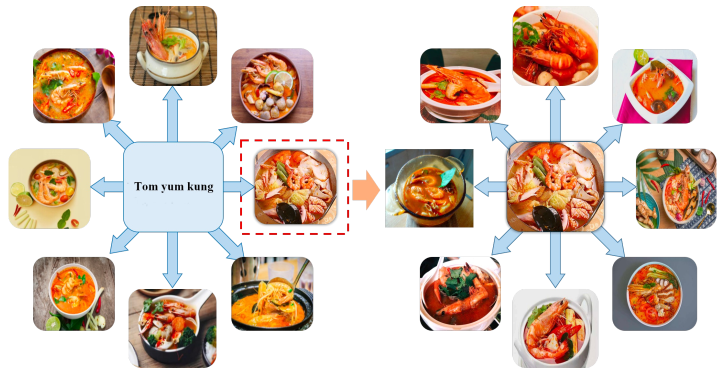
Keywords

Metrics

Abstract:
Food recognition plays an important role in food choice and intake, which is essential to the health and well-being of humans. It is thus of importance to the computer vision community, and can further support many food-oriented vision and multimodal tasks, e.g., food detection and segmentation, cross-modal recipe retrieval and generation.

Meat	Cheese back ribs	Tomahaw	Fried pork in scoop	Sheep roll
Vegetables	Eggplant salad	Fruit salad	Shredded cucumber	Fried eggplant
Bread	Tuna pizza	Beef burger	Seafood pancake	Coconut bread
Snack	Egg tart	Roti prata	Strawberry smoothie	Takoyaki
Fried food	Tonkatsu	Fried chicken	Fried cuttlefish balls	Fried tofu
Seafood	Tempura	Spicy crab	Geoduck sashimi	Cod fish steak
Cereal products	Egg fried rice	Salmon sushi	Pan-fried pork bun	Instant noodles

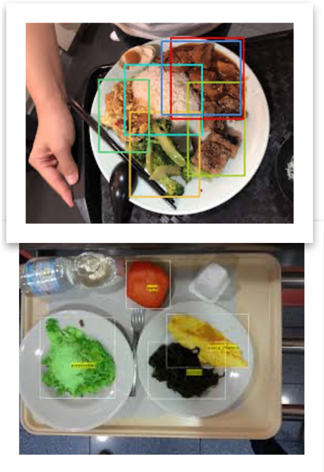


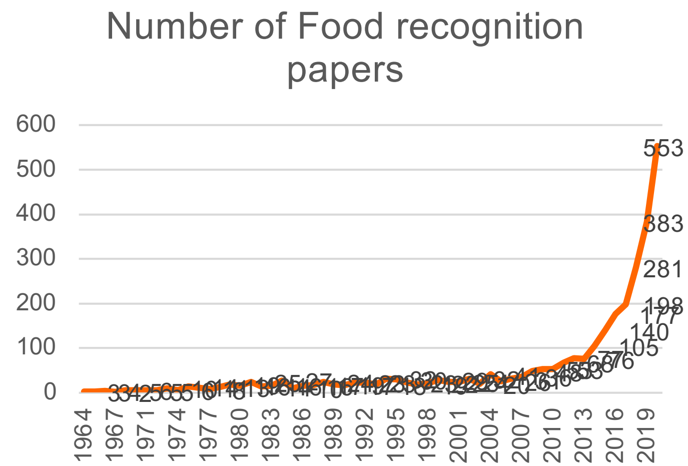


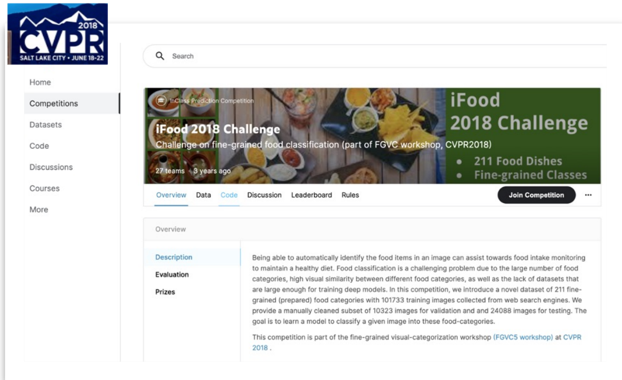
More than 1M images!

Min, Weiqing, et al. "Large scale visual food recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

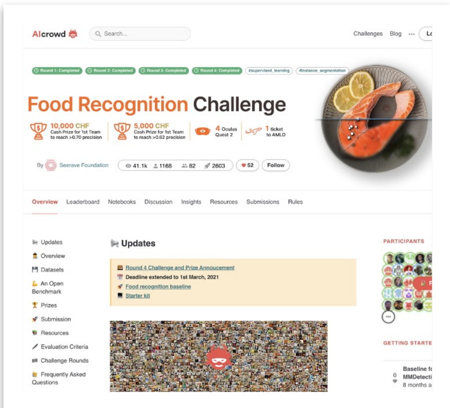
Food recognition popularity







iFood 2011 fine-grained (prepared) food categories with 135733



AICrowd: 26000 annotated segmented images

LargeFineFoodAI: 1,000 fine-grained food categories and over 50,000 images.


Food image analysis

Why is the food recognition a challenge?

Motivation

Food Analysis Problems

Ingredients

- Intra-class variability
- Inter-class similarity

Decreasing in Precision

Intra-class variability example: Apple. Image source: Recipes5k

Inter-class similarity example: Tomato sauce and Curry sauce. Image source: Recipes5k

The food recognition is a Fine-grained recognition problem

Challenges of Food image analysis

Food256: 25.600 images (100 images/class) Classes: 256

Food101 – 101.000 images
(1000 images/class)
Classes: 101

FoodX-251
Classes: 251
140K images

Food1K
Classes: 1000
370K images

Food DB	ImageNet	Future Food DB
150.000 images 231 categories	1.400.000 images 1000 categories	????? images 200.000 categories

Current SoA on Food recognition

- 79% on UECFOOD
- 44% on ChinaFood1000

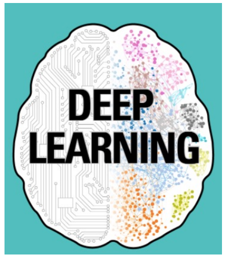
UNIVERSITAT DE BARCELONA

How to leverage from the huge amount of non-annotated data/images?

Self-Supervised Learning allows to leverage big amounts of unlabelled data to make NNs more robust!

Complex Problems Need Powerful Models

What makes this project possible?



Category
noodles and pasta
Dish type
spaghetti carbonara

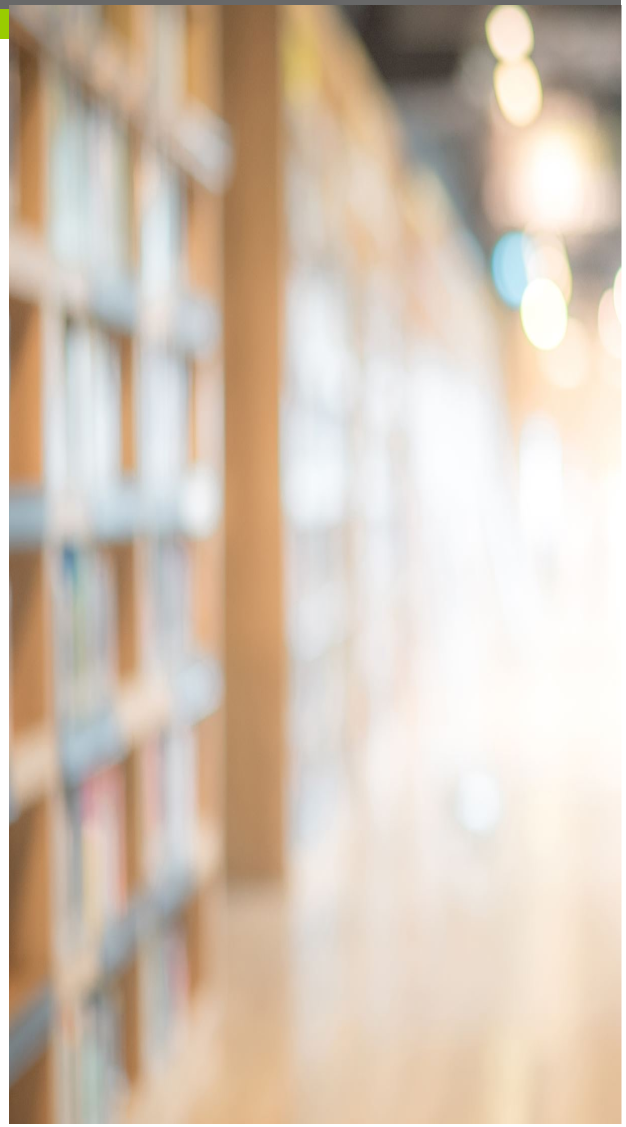
Ingredients
eggs, olive oil, garlic, black pepper, white wine, parmesan cheese, spaghetti, pancetta, grated pecorino

Nutritional information
Energy 710,7kcal
Sugars 3,7 gr
Cholesterol 176,9 mg

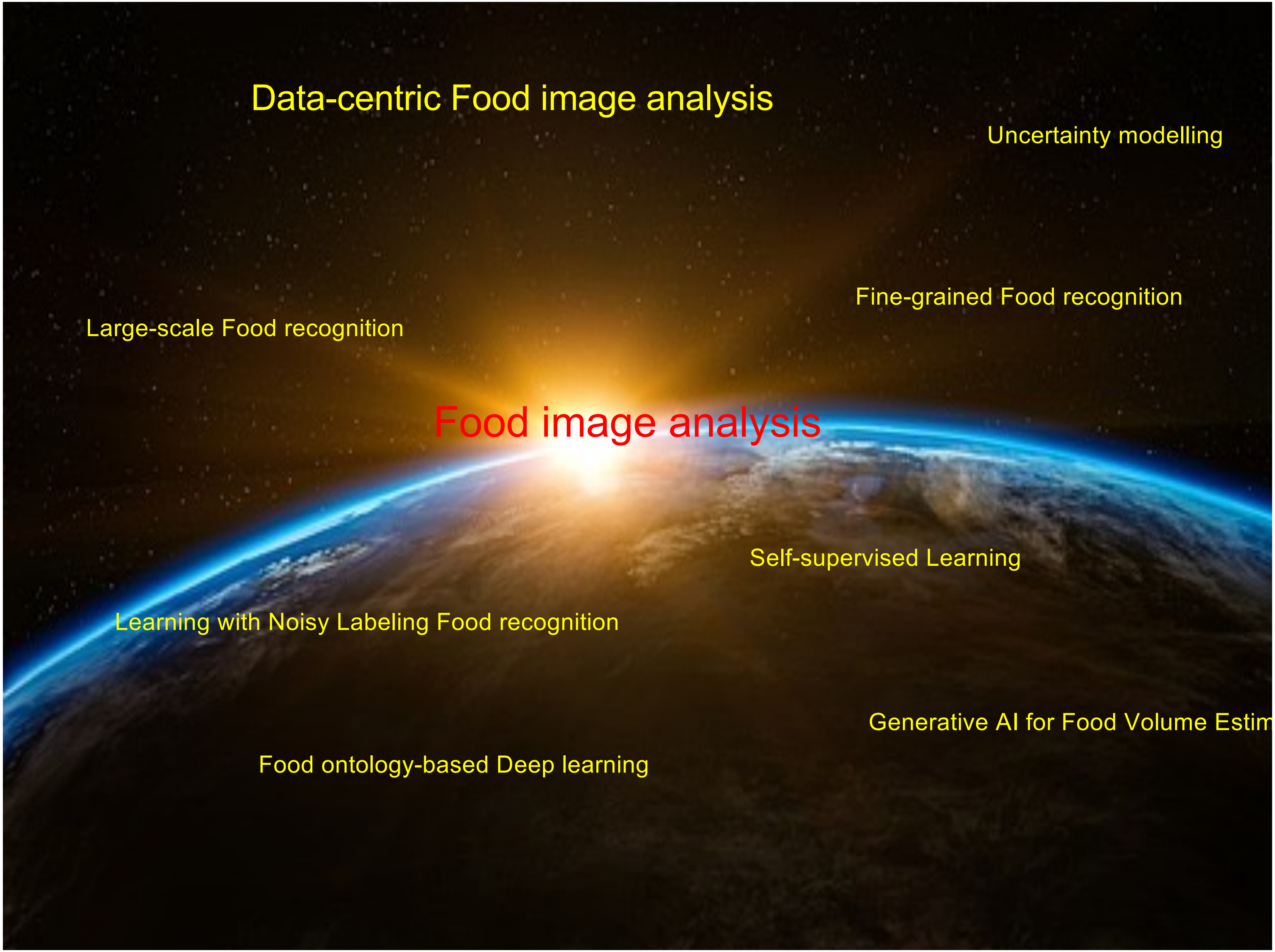
Trends in DL

1. **Transfer Learning**: The use of pre-trained models and transfer learning techniques was becoming more widespread, as they can significantly reduce the amount of data required to train effective models.
   - Multi-task learning
2. **Uncertainty modeling:** refers to the process of quantifying and managing uncertainty or ambiguity in the predictions or decisions made by machine learning models
   - Uncertainty-aware MTL
   - Learning with noisy labeling
3. **Self-Supervised Learning**: methods train models on unlabeled data, which can be particularly useful in cases where labeled data is scarce.
4. **Meta-Learning**: Meta-learning techniques were explored to enable models to learn how to learn, which can lead to faster adaptation to new tasks.
   - Continual learning
5. **Generative AI and Generative Adversarial Networks (GANs)**: GANs were being used for a variety of applications, from image generation to data augmentation and domain adaptation.
   - Uncertainty-aware data augmentation
   - NeRFs
   - Stable diffusion

Trends in DL



- **6. Efficiency and Model Compression**: There was a growing interest in making deep learning models smaller, faster, and more energy-efficient, particularly for edge computing and mobile applications.
    - Scaling by hierarchical DL models
    - Fine-grained recognition
- **7. Multimodal Learning**: Combining information from different sources, such as text and images, to create more comprehensive models for understanding and generating content.
    - Food ontology
- **8. Explainable AI (XAI)**: The need for understanding and interpreting deep learning models became more critical, especially in fields like healthcare.
    - Robust Explainable models
- **9. Federated Learning**: This approach allows for training models across decentralized data sources without sharing raw data, preserving privacy. It gained traction, particularly in applications involving sensitive or proprietary data.
- **10. AI for Healthcare**: Deep learning was increasingly applied to medical image analysis, drug discovery, and patient diagnosis, with a focus on improving healthcare outcomes.
    - Our projects: food data analysis,


Data-centric Food image analysis
Uncertainty modelling
Fine-grained Food recognition
Large-scale Food recognition
Food image analysis
Self-supervised Learning
Learning with Noisy Labeling Food recognition
Generative AI for Food Volume Estim
Food ontology-based Deep learning

Index

1. Why Food Recognition?

2. Self-Supervised Learning for Fine-Grained Recognition

   - Validation of All4One

3. Other Food Recognition works

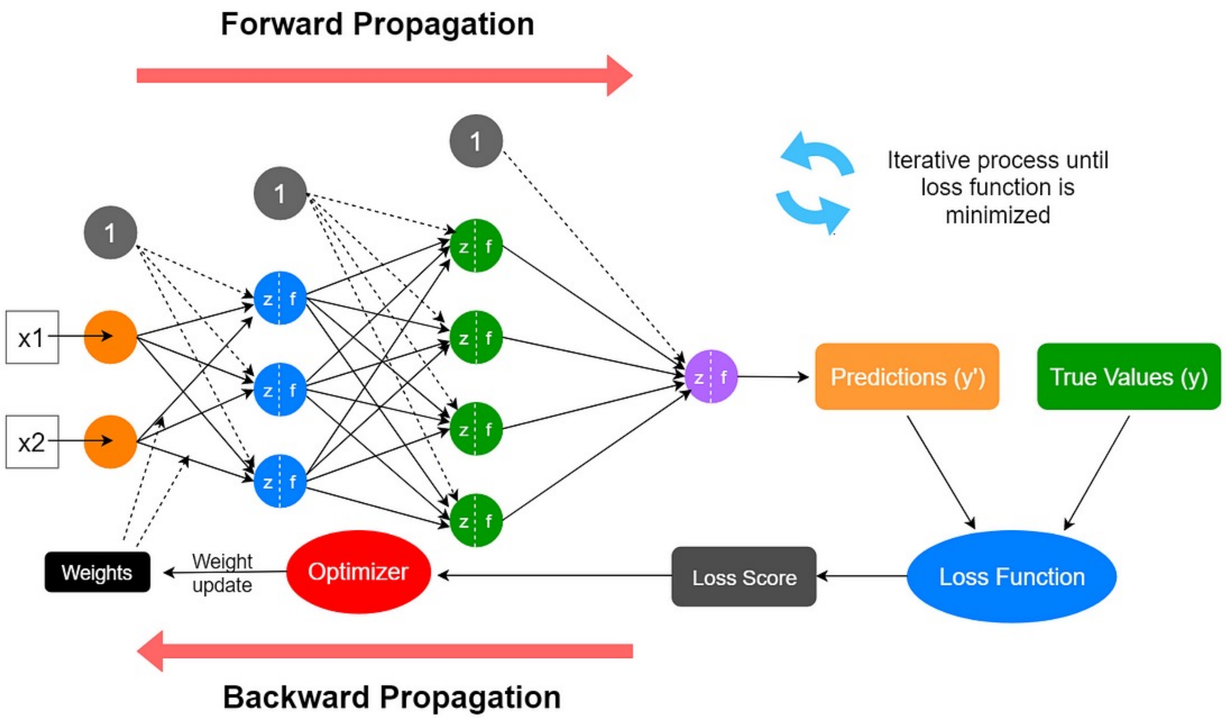
4. Food Recognition real applications

Main pitfall of Deep Learning Models

Neural networks are dragons, but…

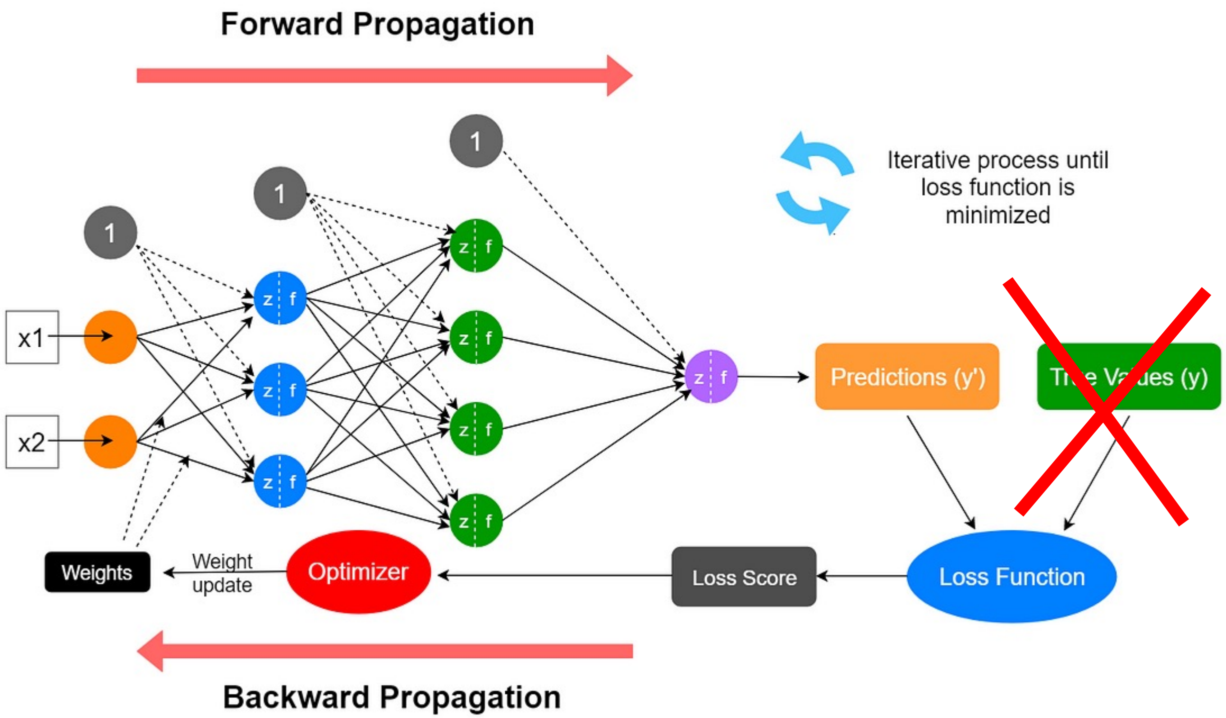
Greedy dragons!

How to make a NN robust?

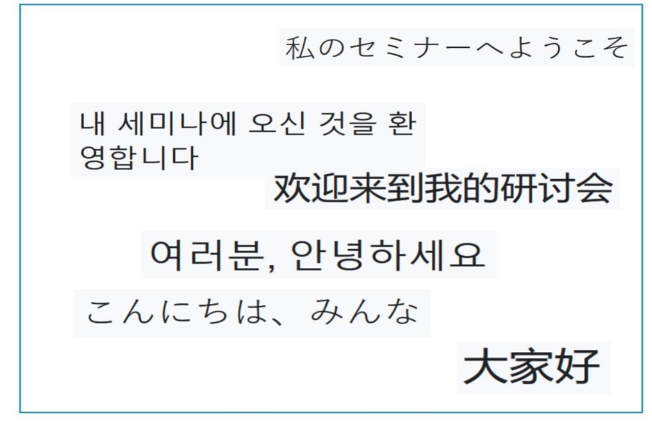


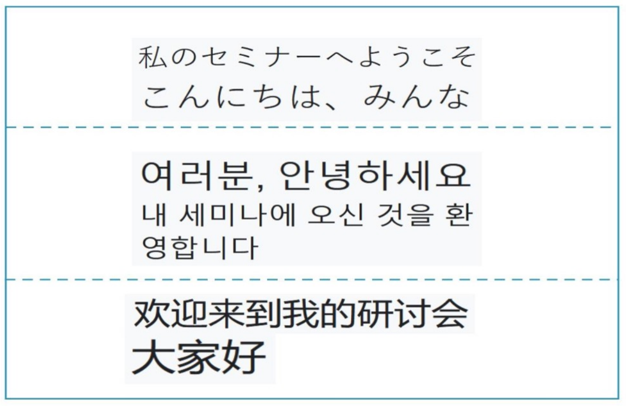
Use data augmentation

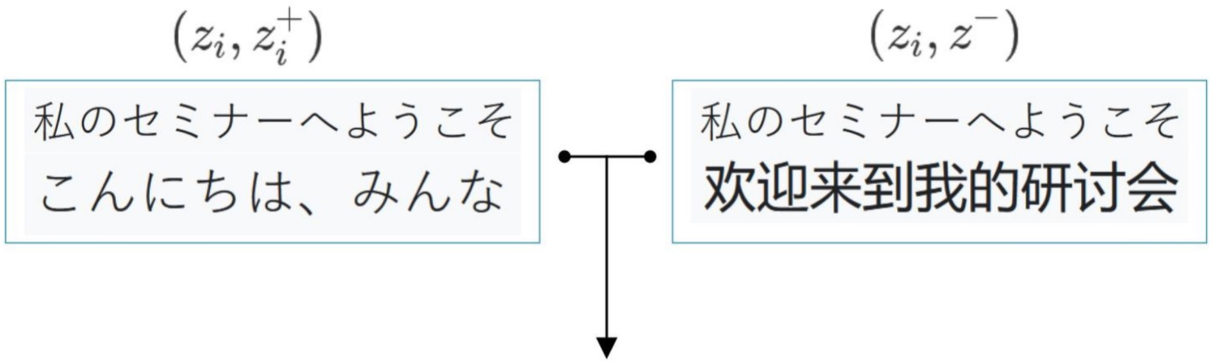
How to make a NN robust with unlabeled data?



What can you do if you have a lot of just data and may be a not-trained model?



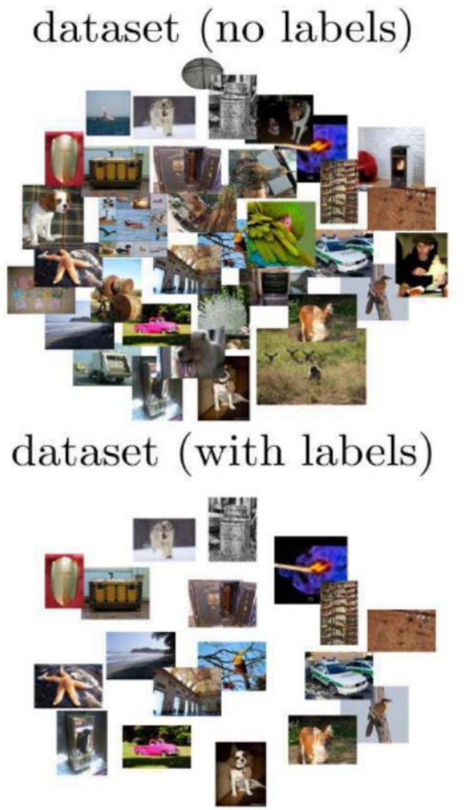




Contrastive loss

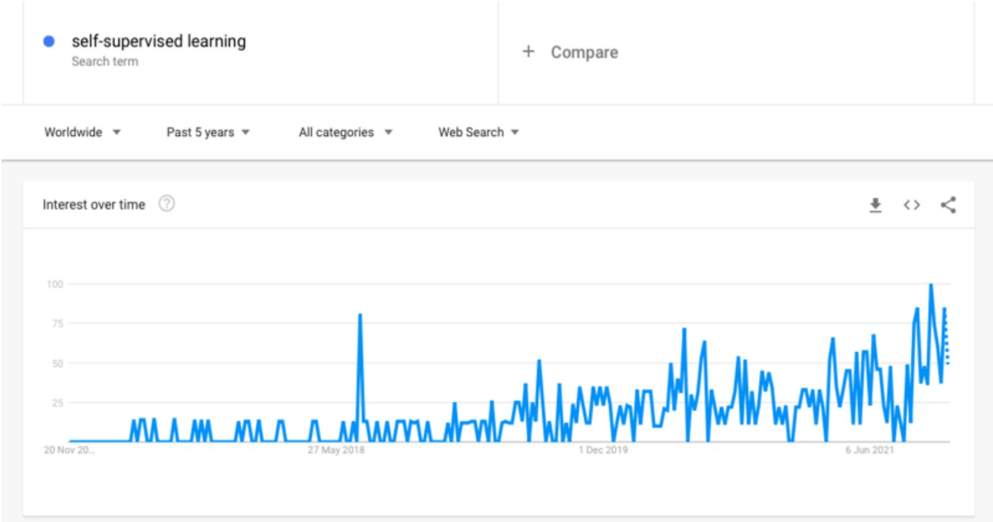
$$\mathcal{L}_i^{\text{InfoNCE}} = -\log \frac{\exp\left(z_i \cdot z_i^+/\tau\right)}{\exp\left(z_i \cdot z_i^+/\tau\right) + \sum_{z^- \in \mathcal{N}_i} \exp\left(z_i \cdot z^-/\tau\right)}$$

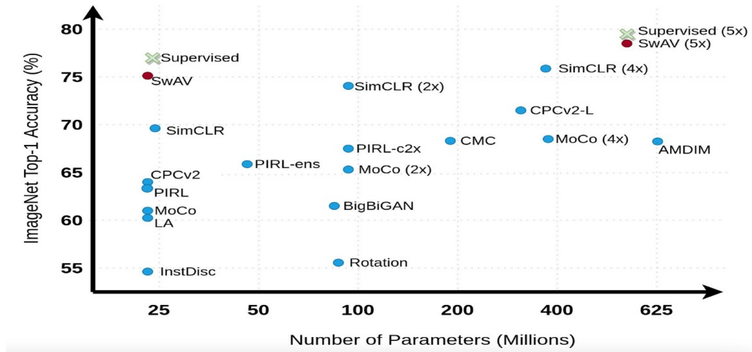
What is self-supervised learning (SSL)?



What would you do if you have thousands of unlabelled images?

Self-Supervised Learning: Benefits & Uses in 2023



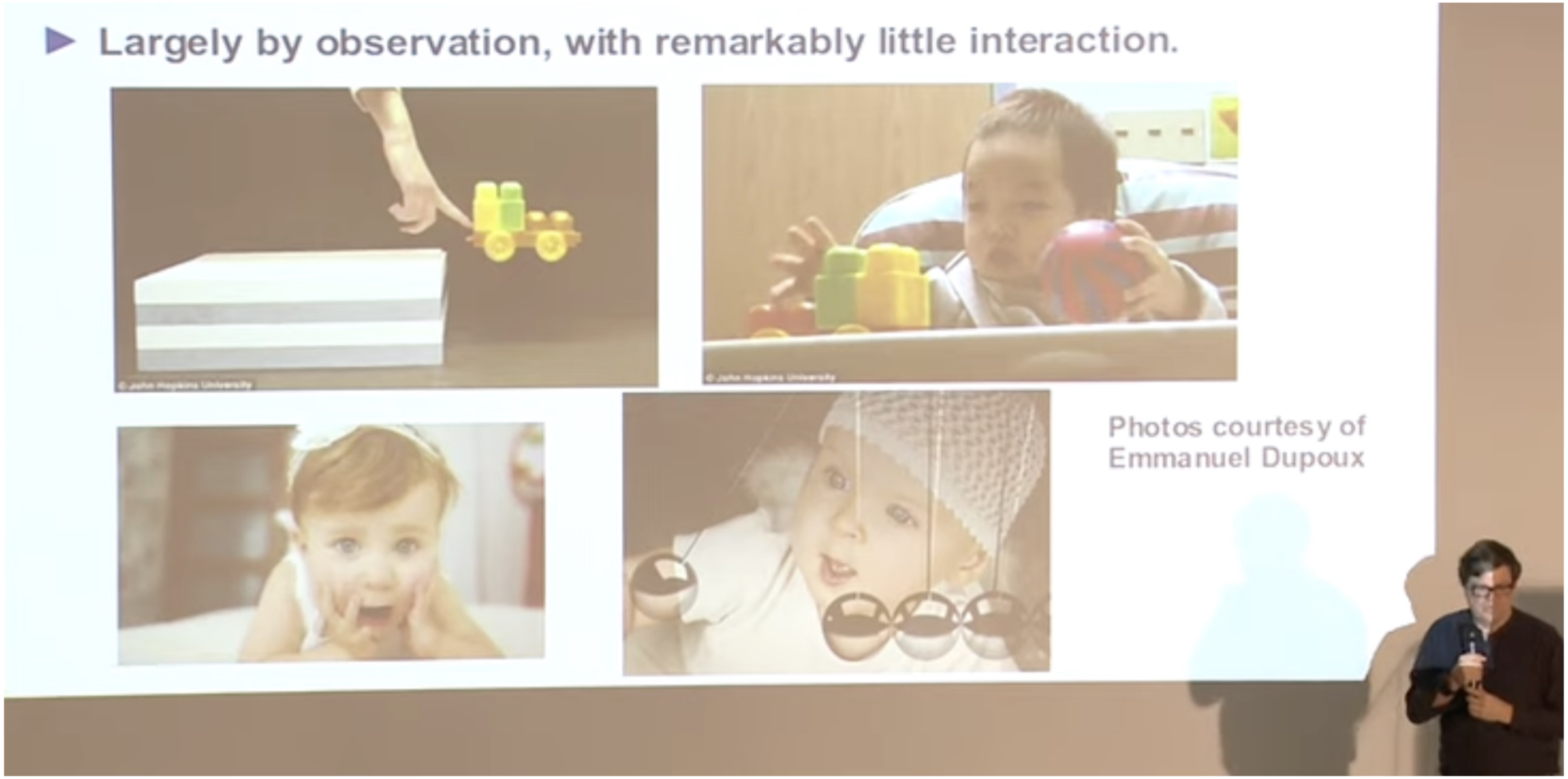


Yann LeCun and Yoshua Bengio: "Self-supervised learning is the key to human-level intelligence"

Yann LeCun, VP and Chief AI Scientist at Facebook, is explaining how self-supervised learning works at PAISS'19

https://research.aimultiple.com/self-supervised-learning/

Babies learn how the world Works by observation



Artifical vs Natural NNs

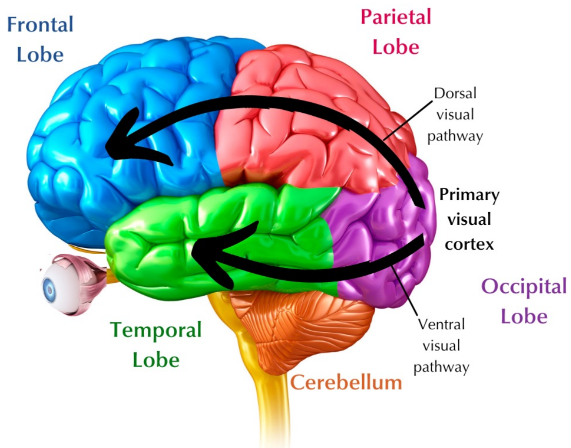
Understand brain through NNs:

- the brain is full of feedback connections, while current models have few such connections, if any.

Next step: use SSL to train highly recurrent networks and see how the activity in NNs compares to real brain activity.

Crucial step: match the activity of NNs in SSL models to the activity of individual biological neurons.

"No doubt that 90% of what the brain does is self-supervised learning," Blake Richards, a computational neuroscientist at McGill University and Mila, the Quebec Artificial Intelligence Institute.



Hypothesis: the visual systems of humans and other primates are the best studied of all animal sensory systems,

- neuroscientists struggle to explain why they include two separate pathways:
  - the **ventral visual stream**, which is responsible for recognizing objects and faces, and
  - the **dorsal visual stream**, which processes movement (the "what" and "where" pathways, respectively).

CVI NOW

SimCLR by the Google AI team

Introduces projectors: a learnable nonlinear transformation between the representation and the contrastive loss

Positive sampling: Given a batch of N samples, the pretext task P generates two augmented views x^a_i and x^+_i for each sample x_i of the batch.

Negative sampling: the rest of the images x^-_i on the same batch to form the negative pairs (x^a_i, x^-_i).

Batch sizes of 8196 are used.

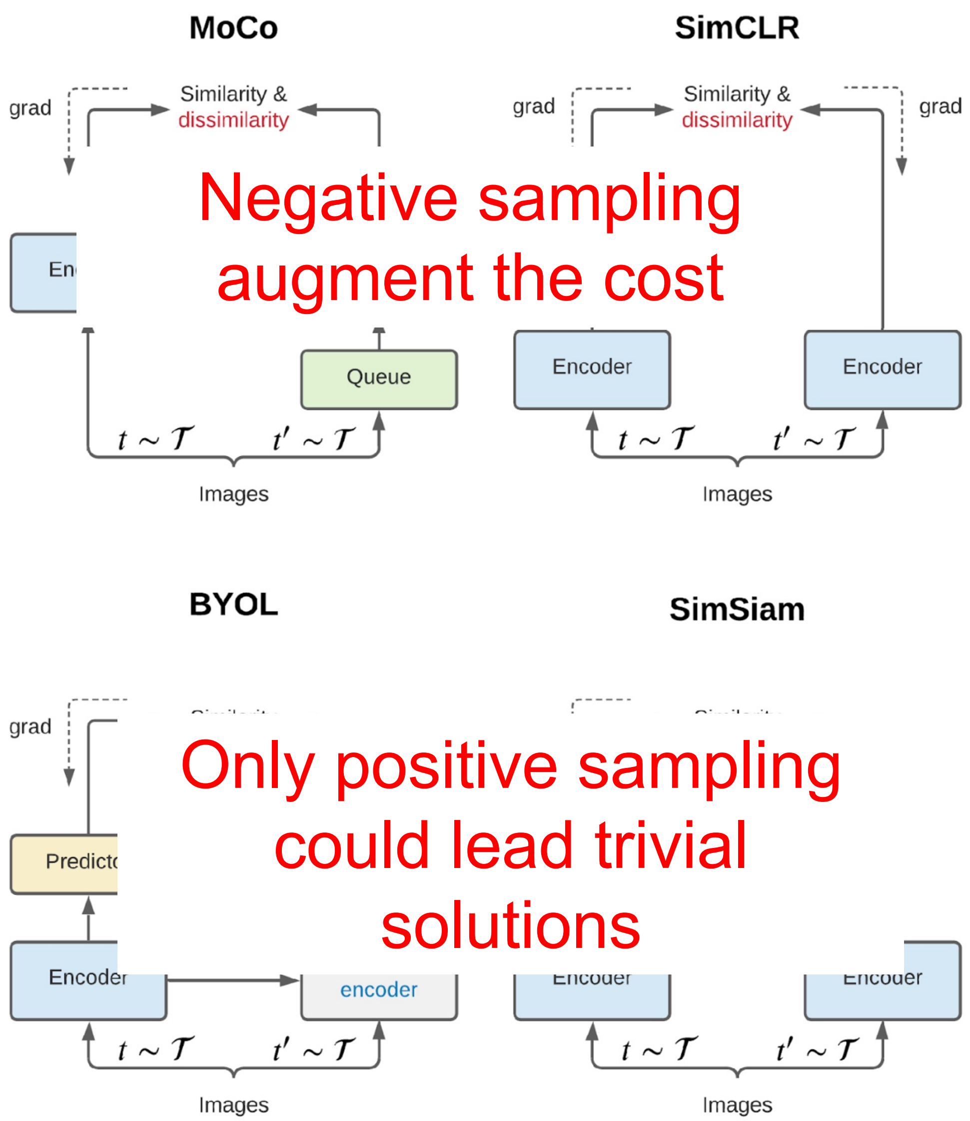
Loss function:

$$L_i^{SimCLR} = -log(\frac{exp(z_i^a \cdot z_i^+/\tau)}{\sum_{k=1}^{N} exp(z_i^a \cdot z_k^-/\tau)})$$

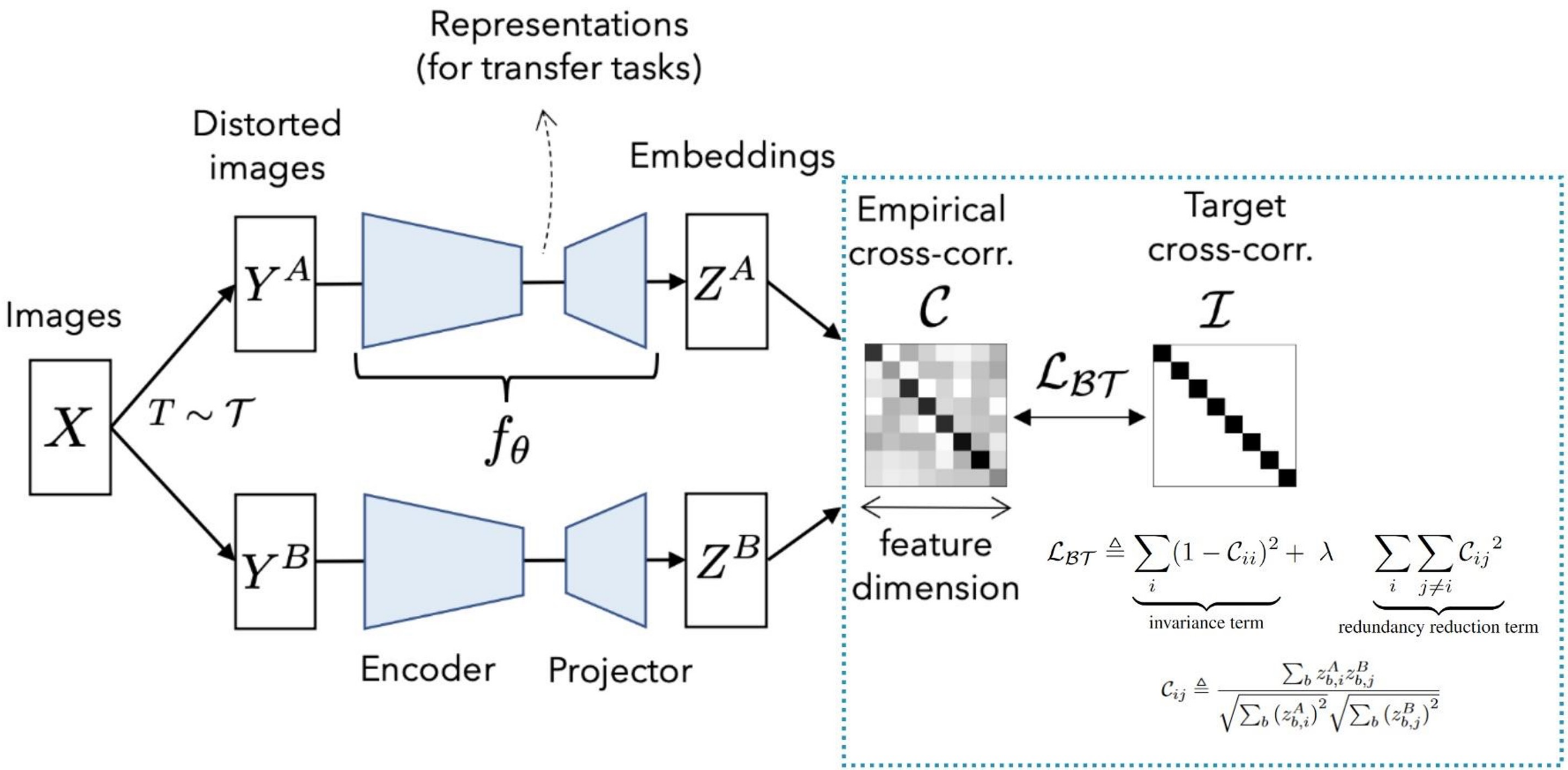
Ting Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". 37th ICMLL 2020, pp. 1597–1607.

https://analyticsindiamag.com/what-is-contrastive-self-supervised-learning/

State-of-the-art Contrastive SSL models



SSL Framework: Barlow Twins

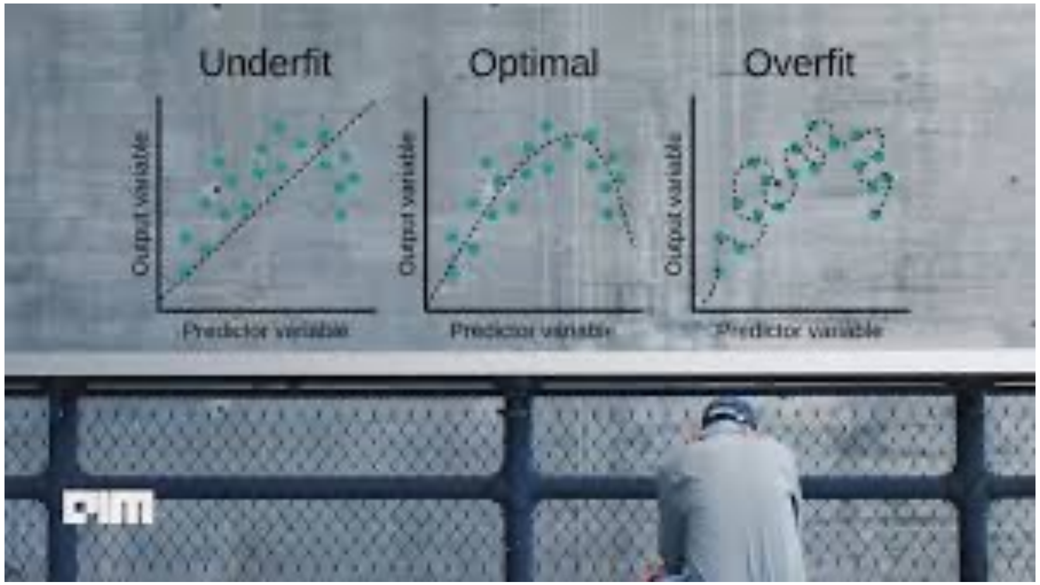


J. Zbontar et al. “Barlow Twins: Self-Supervised Learning via Redundancy Reduction”. 38th ICML. 2021, pp. 12310–12320

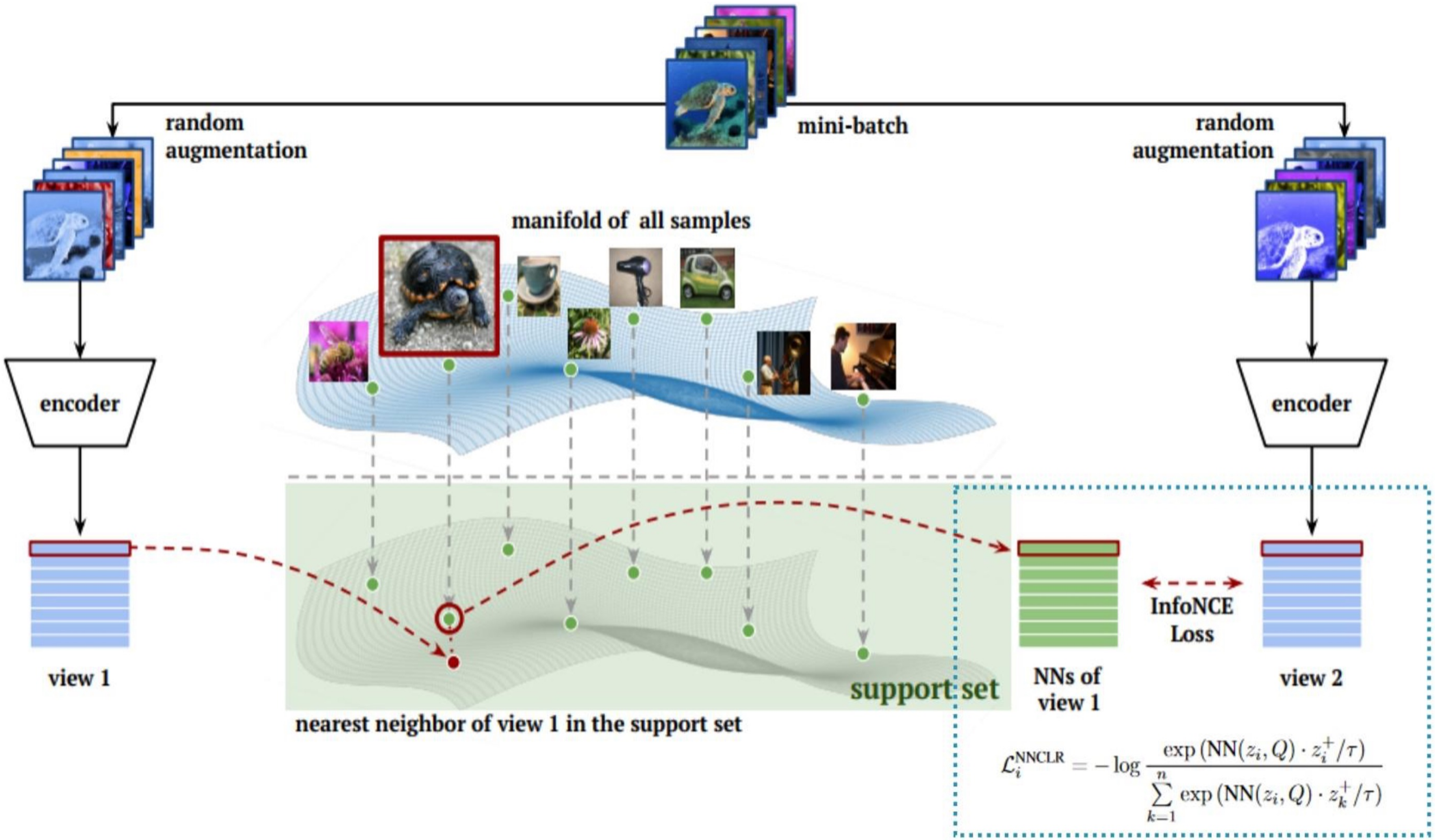
Self-supervised Learning

- Still SSL learn too fast!

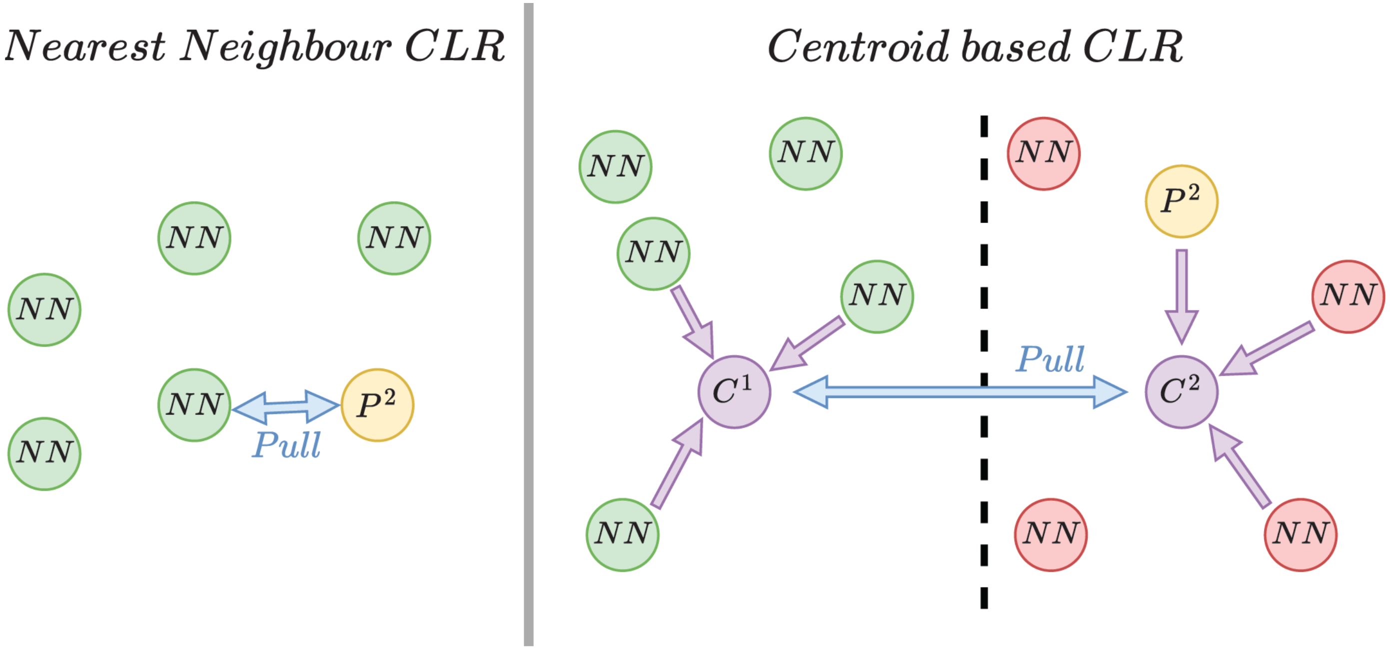
- Overfitting the domain!



SSL Framework: NNCLR



Centroid based CLR



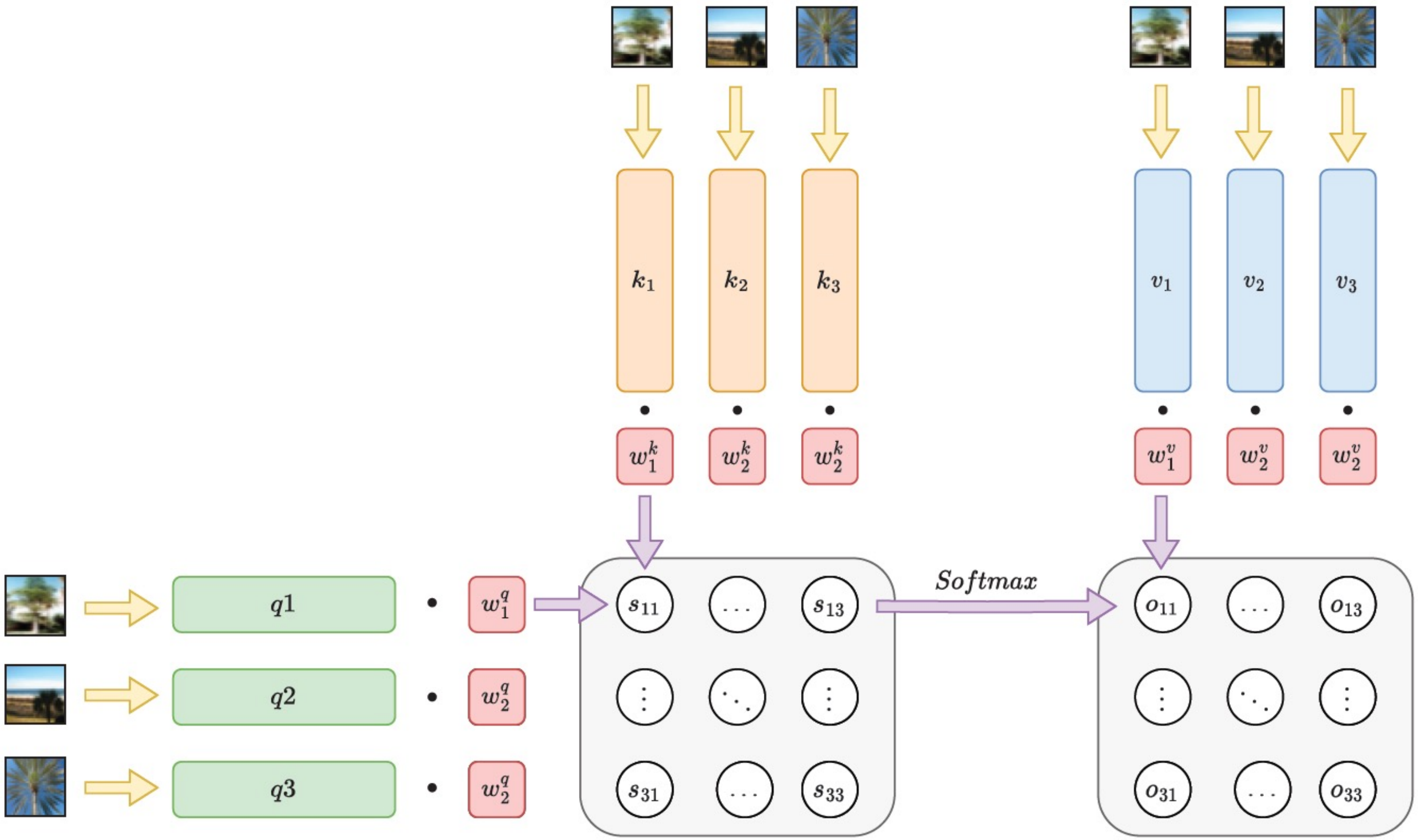
Common neighbour contrastive approaches only contrast the first neighbour.

We create representations that contain contextual information from the k NNs
- Contrast it in a single objective computation:

$$L_i^{centroid} = -log\left(\frac{exp(c_i^1 \cdot c_i^2/\tau)}{\sum_{n=1}^{n} exp(c_i^1 \cdot c_n^2/\tau)}\right)$$

Estepa, I. G., Sarasúa, I., Nagarajan, B., & Radeva, P. (2023). All4One: Symbiotic Neighbour Contrastive Learning via Self-Attention and Redundancy Reduction. ICCV'2023, *arXiv:2303.09417*.

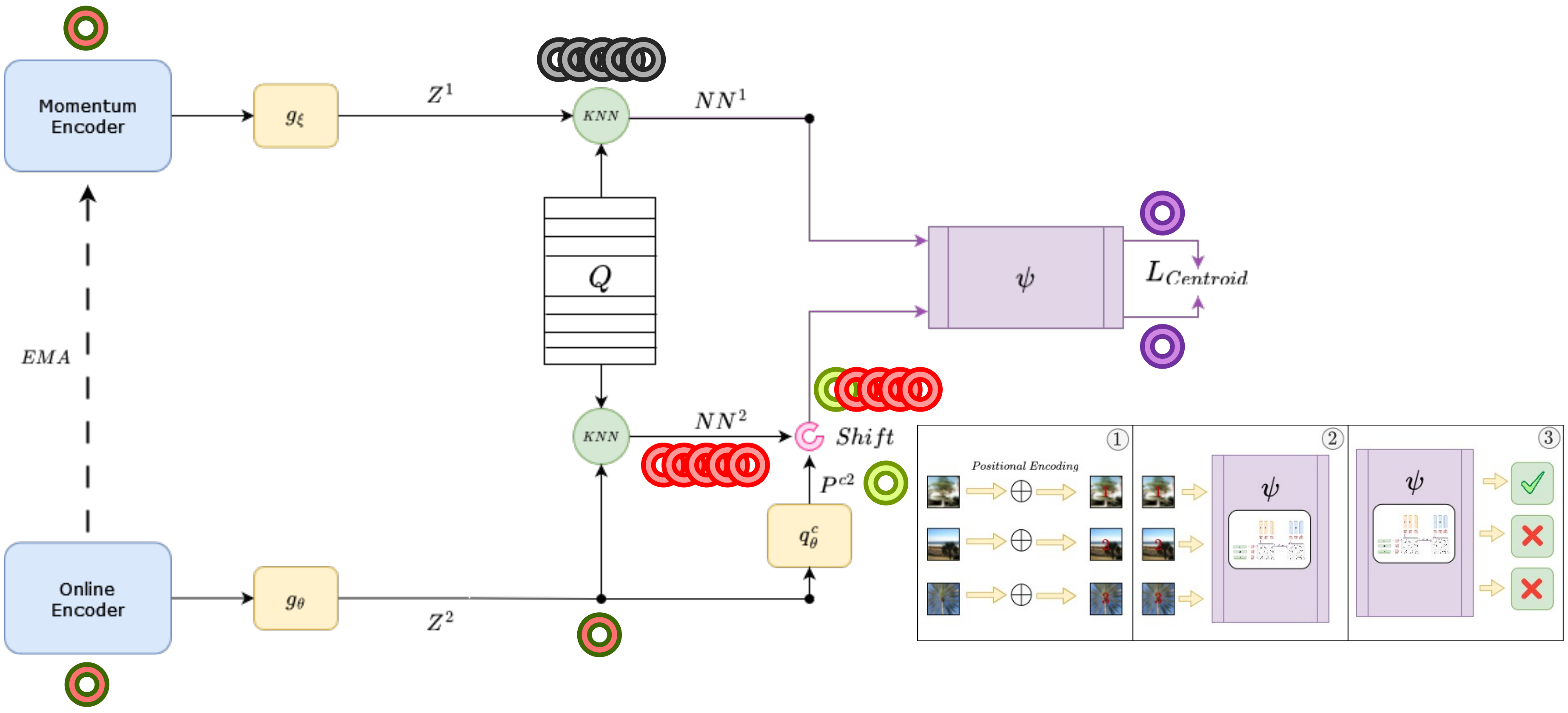
Self-Attention in order to obtain "context-aware" representations



$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{n}})\mathbf{V}$$

Estepa, I. G., Sarasúa, I., Nagarajan, B., & Radeva, P. (2023). All4One: Symbiotic Neighbour Contrastive Learning via Self-Attention and Redundancy Reduction. ICCV'2023, *arXiv:2303.09417*.

All for One: Centroid contrast



Estepa, I. G., Sarasúa, I., Nagarajan, B., & Radeva, P. (2023). All4One: Symbiotic Neighbour Contrastive Learning via Self-Attention and Redundancy Reduction. ICCV'2023, *arXiv:2303.09417*.

Index

1. Why Food Recognition?

2. Self-Supervised Learning for Fine-Grained Recognition

   - Validation of All4One

3. Other Food Recognition works

4. Food Recognition real applications

Dataset and Evaluation Metrics

Dataset

Food-101

Evaluation Metrics

SSL Model

Top 1 and top 5 accuracy using a k-NN classifier

Classifier

Overall accuracy
Variance
Entropy
Mutual Information

Quantitative Results: CIFAR

CIFAR-10

Method	Backbone	Epochs	Acc@1 (Online)	Acc@5 (Online)
BYOL	ResNet18	1000	92.58	99.79
DeepCluster V2	ResNet18	1000	88.85	99.58
DINO	ResNet18	1000	89.52	99.71
MoCo V2+	ResNet18	1000	92.94	99.79
MoCo V3	ResNet18	1000	93.10	99.80
ReSSL	ResNet18	1000	90.63	99.62
SimCLR	ResNet18	1000	90.74	99.75
Simsiam	ResNet18	1000	90.51	99.72
SwAV	ResNet18	1000	89.17	99.68
VIbCReg	ResNet18	1000	91.18	99.74
VICReg	ResNet18	1000	92.07	99.74
W-MSE	ResNet18	1000	88.67	99.68
Barlow Twins	ResNet18	1000	92.10	99.73
NNCLR	ResNet18	1000	91.88	99.78
All4One (Ours)	ResNet18	1000	**93.24**	**99.88**

CIFAR-100

Method	Backbone	Epochs	Acc@1 (Online)	Acc@5 (Online)	k-NN Acc@1 (Online)
BYOL	ResNet18	1000	70.46	91.96	-
DeepCluster V2	ResNet18	1000	63.61	88.09	-
DINO	ResNet18	1000	66.76	90.34	-
MoCo V2+	ResNet18	1000	69.89	91.65	-
MoCo V3	ResNet18	1000	68.83	90.57	-
ReSSL	ResNet18	1000	65.92	89.73	-
SimCLR	ResNet18	1000	65.78	89.04	-
Simsiam	ResNet18	1000	66.04	89.62	-
SwAV	ResNet18	1000	64.88	88.78	-
VIbCReg	ResNet18	1000	67.37	90.07	-
VICReg	ResNet18	1000	68.54	90.83	-
W-MSE	ResNet18	1000	61.33	87.26	-
NNCLR	ResNet18	1000	69.62	91.52	-
NNCLR*	ResNet18	1000	69.17	91.70	62.16
Barlow Twins	ResNet18	1000	70.90	91.91	-
Barlow Twins*	ResNet18	1000	71.21	92.46	63.11
MSF*	ResNet18	1000	67.84	91.64	63.36
All4One (Ours)	ResNet18	1000	**72.17**	**93.35**	**64.84**

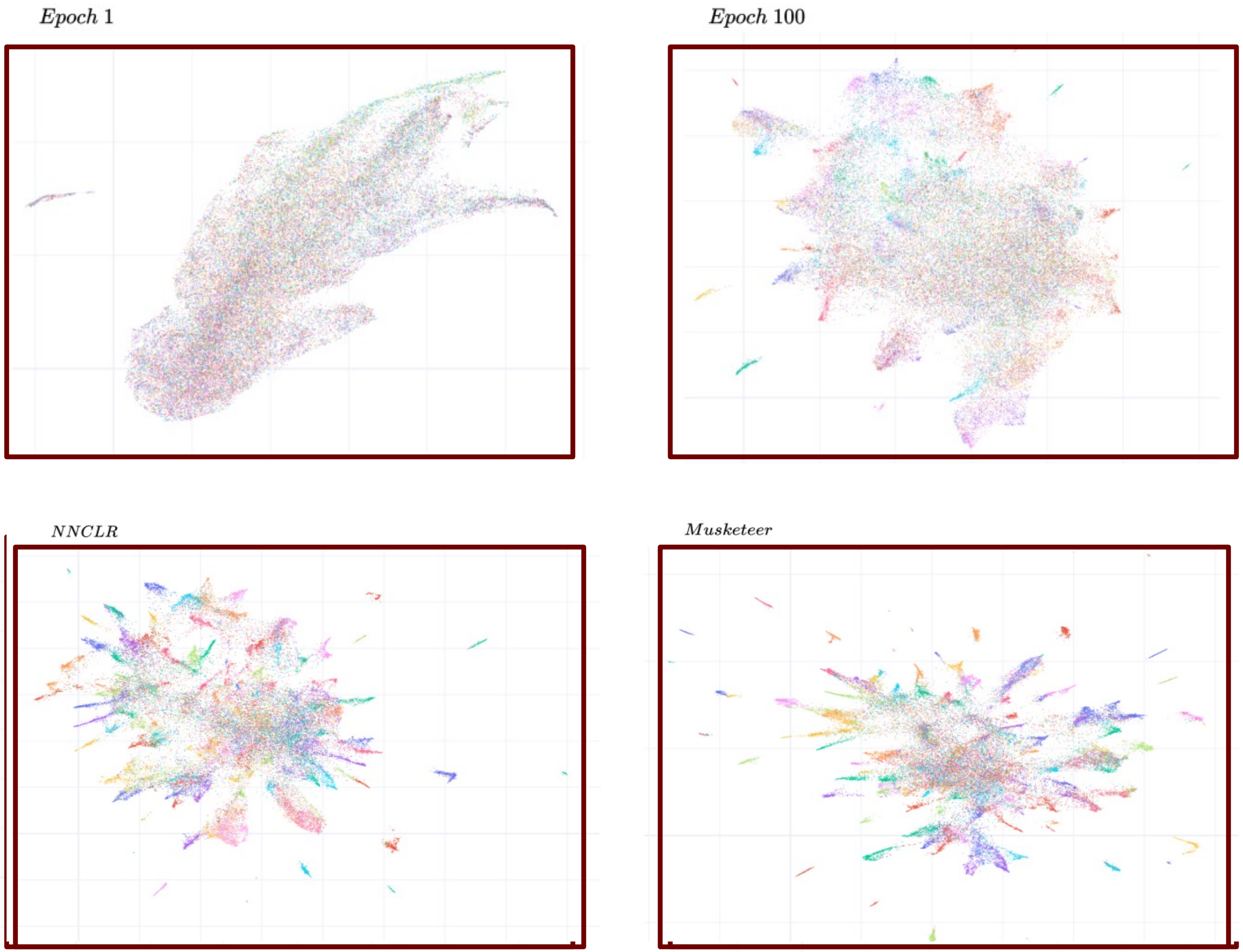
Estepa, I. G., Sarasúa, I., Nagarajan, B., & Radeva, P. (2023). All4One: Symbiotic Neighbour Contrastive Learning via Self-Attention and Redundancy Reduction. ICCV'2023, *arXiv:2303.09417*.

Quantitative Results: ImageNet-100

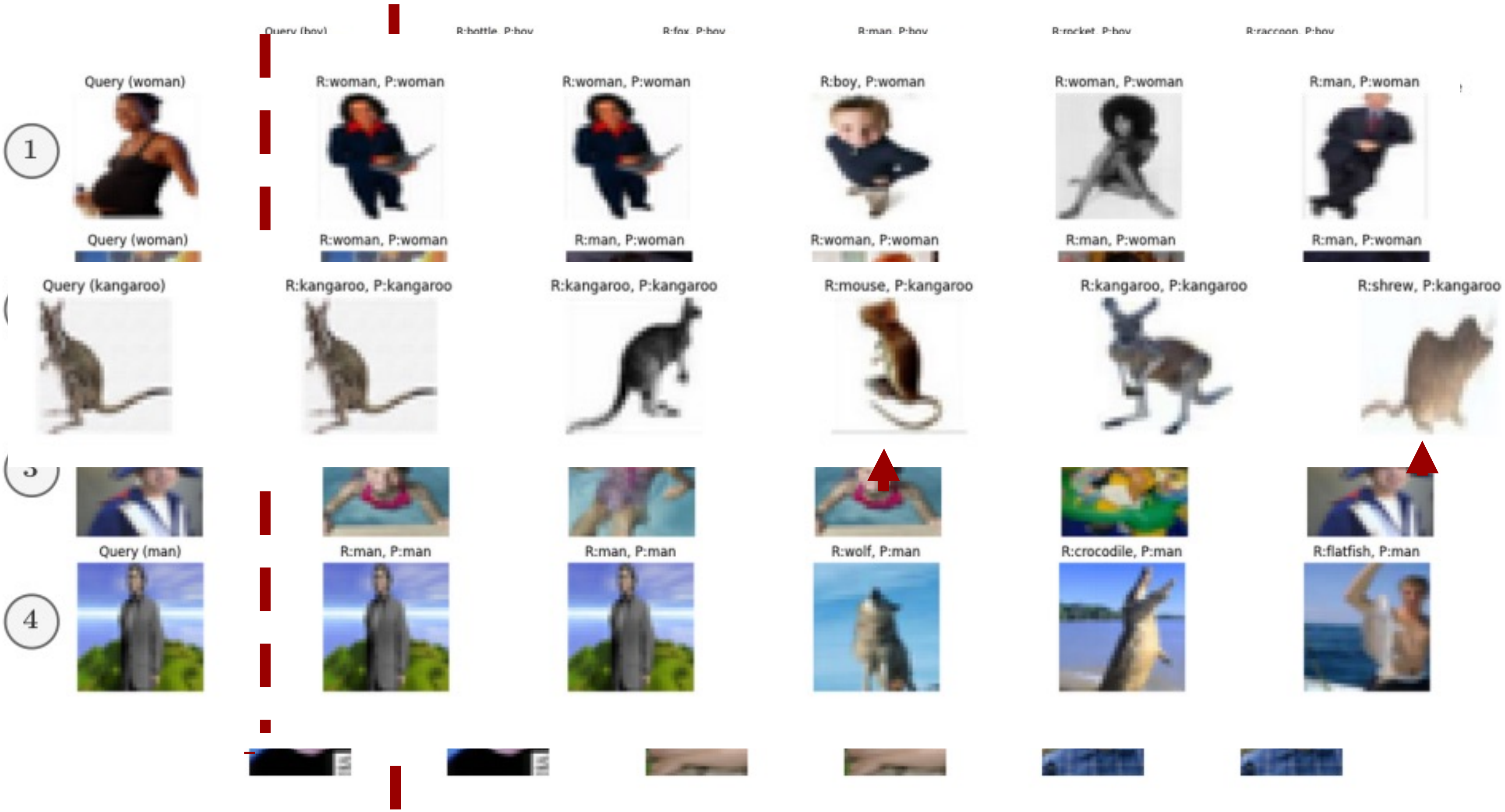
Method	Backbone	Epochs	Acc@1 (online)	Acc@5 (online)
BYOL ++	ResNet18	400	80.16	95.02
DeepCluster V2	ResNet18	400	75.36	93.22
DINO	ResNet18	400	74.84	92.92
MoCo V2+ ++	ResNet18	400	78.20	95.50
MoCo V3 ++	ResNet18	400	80.36	95.18
ReSSL	ResNet18	400	76.92	94.20
SimCLR ++	ResNet18	400	77.64	94.06
Simsiam	ResNet18	400	74.54	93.16
SwAV	ResNet18	400	74.04	92.70
VIbCReg	ResNet18	400	79.86	94.98
VICReg ++	ResNet18	400	79.22	95.06
W-MSE	ResNet18	400	67.60	90.94
Barlow Twins ++	ResNet18	400	80.38	95.28
NNCLR ++	ResNet18	400	79.80	95.28
All4One **(Ours)**	ResNet18	400	**81.93**	**96.23**

Estepa, I. G., Sarasúa, I., Nagarajan, B., & Radeva, P. (2023). All4One: Symbiotic Neighbour Contrastive Learning via Self-Attention and Redundancy Reduction. ICCV'2023, *arXiv:2303.09417*.

Qualitative Analysis: UMAP



Qualitative Analysis: NN Retrieval



Estepa, I. G., Sarasúa, I., Nagarajan, B., & Radeva, P. (2023). All4One: Symbiotic Neighbour Contrastive Learning via Self-Attention and Redundancy Reduction. ICCV'2023, *arXiv:2303.09417*.

Index

1. Why Food Recognition?

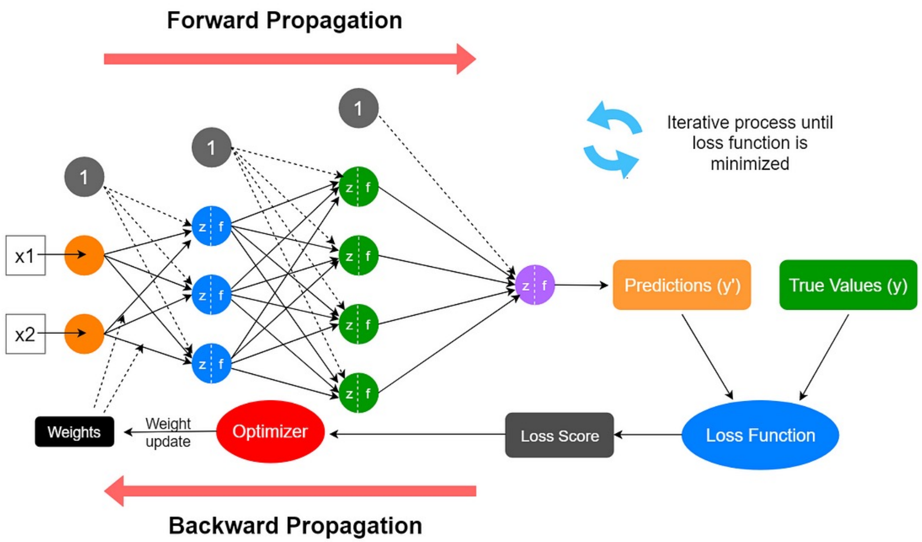
2. Self-Supervised Learning for Fine-Grained Recognition

   - Validation of All4One

3. Other Food Recognition works

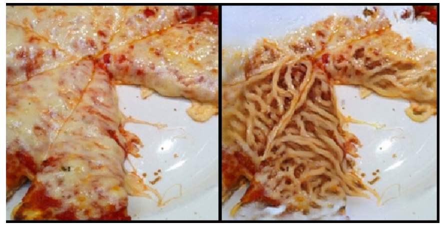
4. Food Recognition real applications

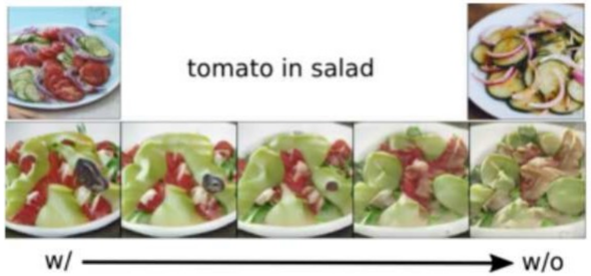
How to make a NN robust?

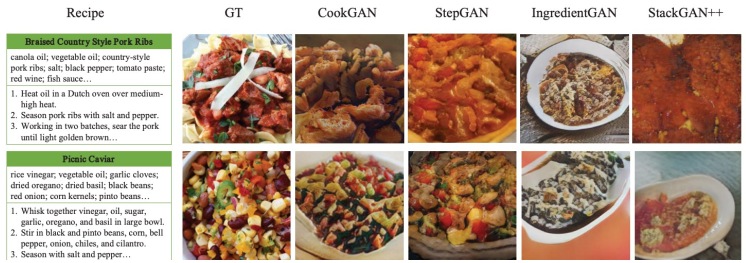


Use data augmentation
Is it enough?

Synthetic image generation



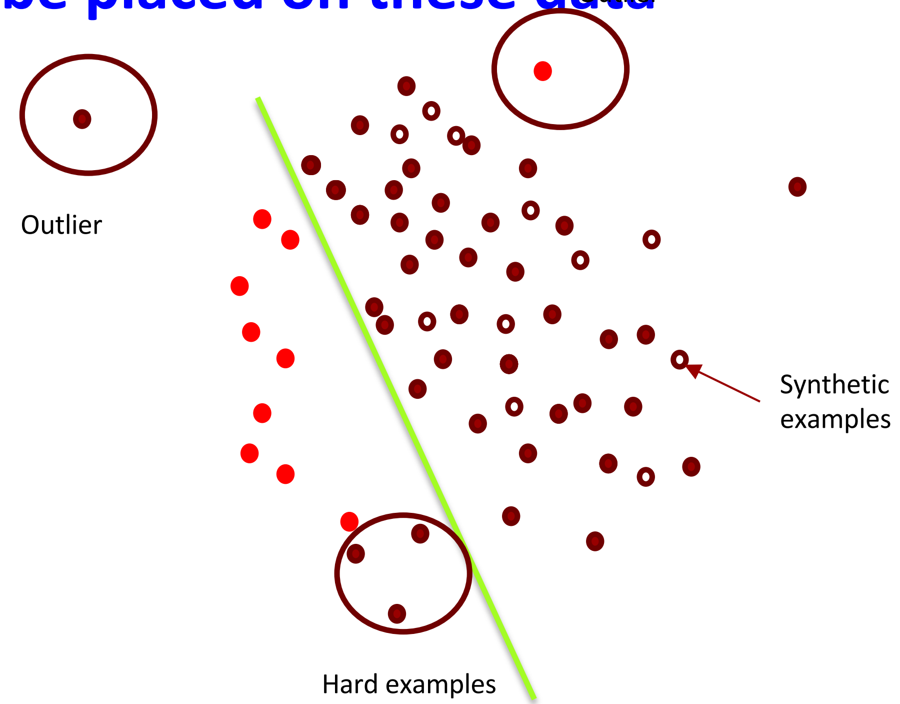




Using synthetic data but how?

- Some images could be more beneficial for the classification than others

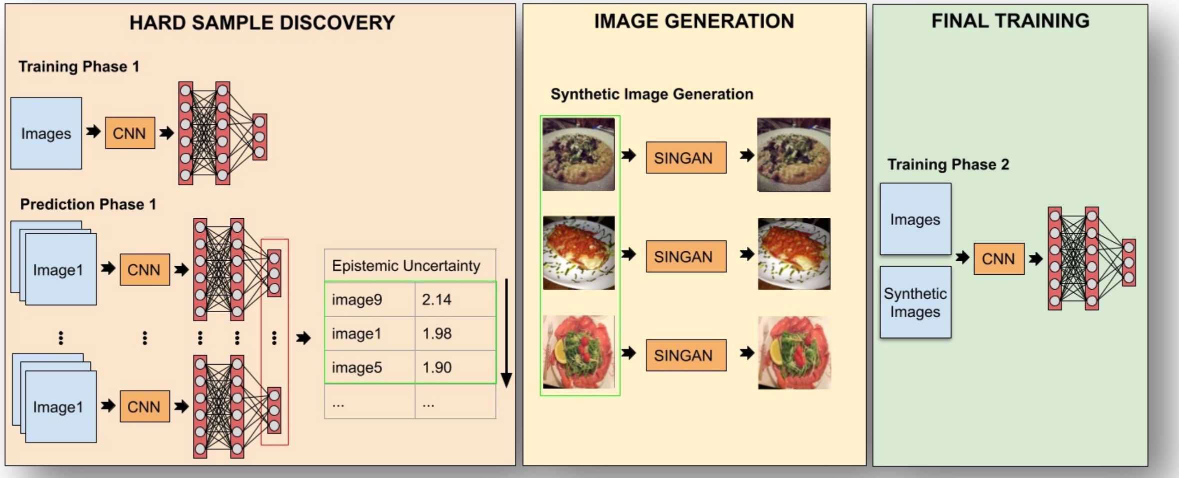
More emphasis should be placed on these data



- **Hypothesis:** Estimating the uncertainty can help us decide the most appropriate samples and classes to perform additional data augmentation methods.

E. Aguilar, and P. Radeva. "Class-conditional Data Augmentation Applied to Image Classification." International Conference on Computer Analysis of Images and Patterns (CAIP),2019.

Use Uncertainty for Data Augmentation

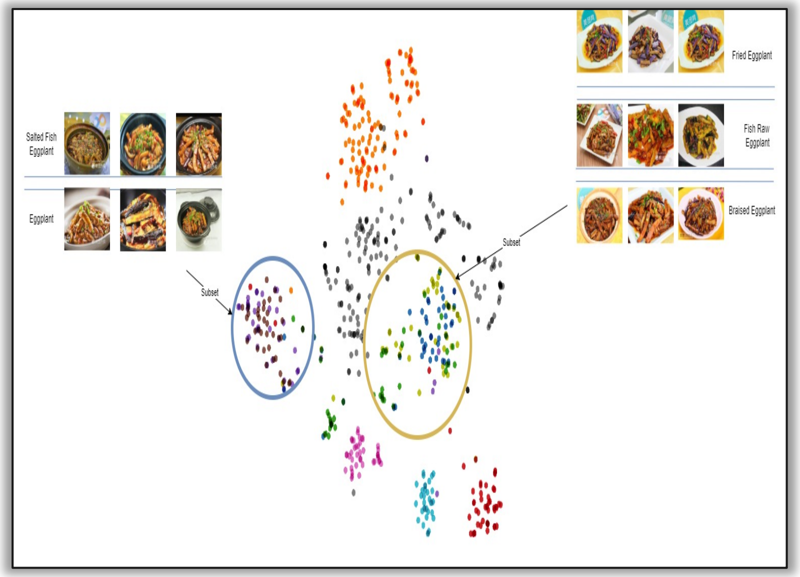


Use the data augmentation applied class-conditionally to improve the results in terms of accuracy and also to reduce the overall epistemic uncertainty.

During the prediction phase, the same image is fed to the CNN several times to calculate the epistemic uncertainty given by the model for that image

E. Aguilar, and P. Radeva. "Class-conditional Data Augmentation Applied to Image Classification." International Conference on Computer Analysis of Images and Patterns (CAIP),2019.

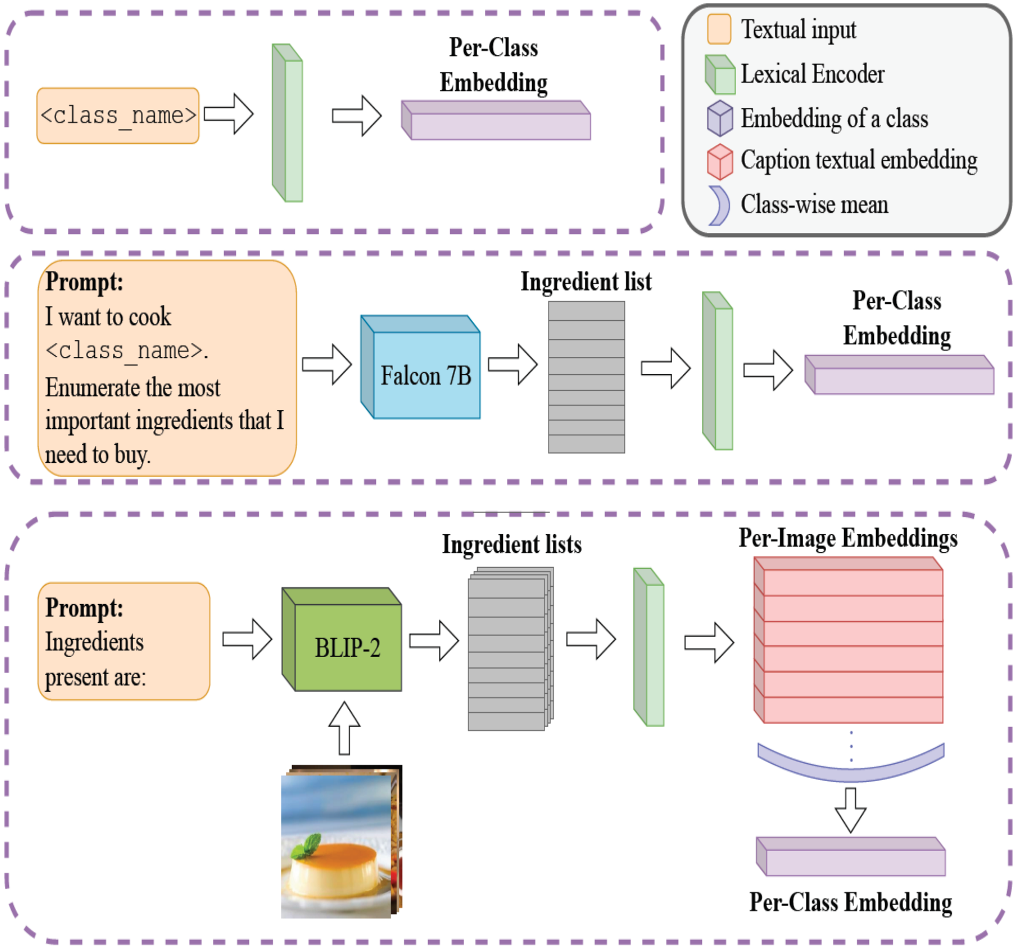
Fine-grained recognition



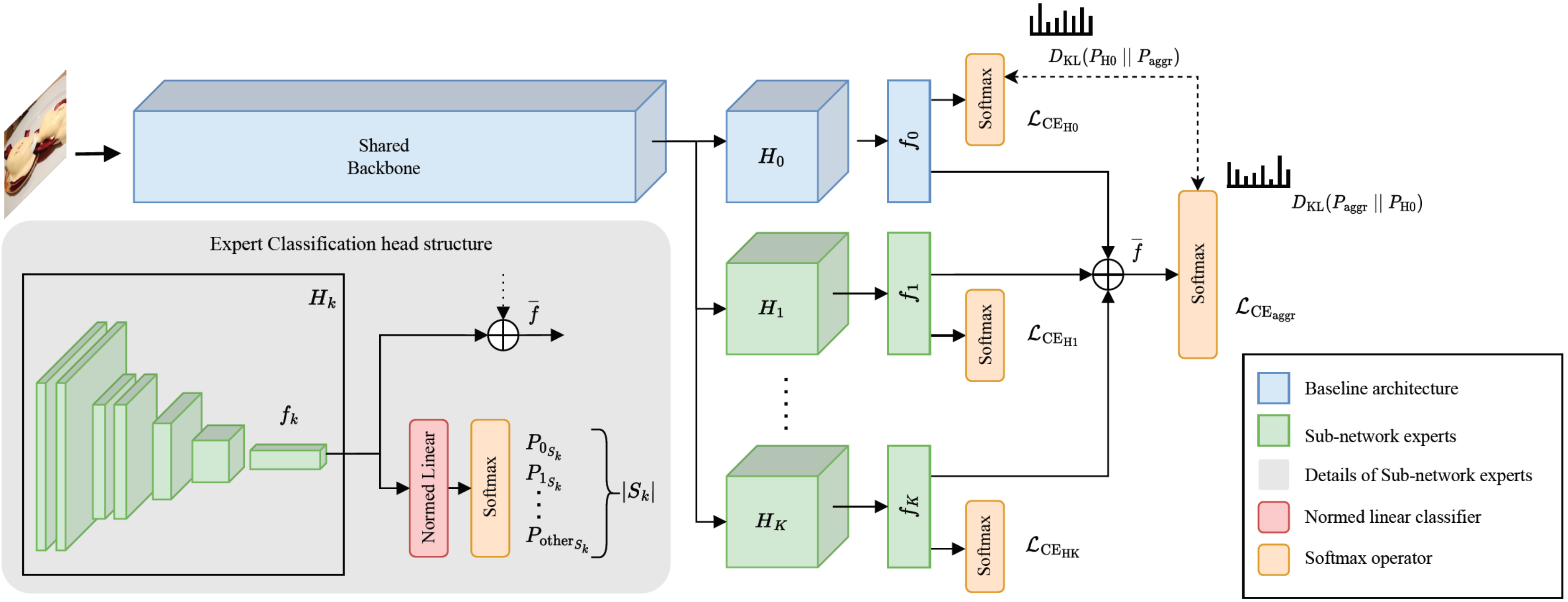
Ródenas, J., Nagarajan, B., Bolaños, M., & Radeva, P. (2022, October). Learning Multi-Subset of Classes for Fine-Grained Food Recognition. In *Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management* (pp. 17-26).

Obtaining lexical embeddings

- Use the **class labels** from the dataset directly, pre-process them and **pass** them to the **lexical encoder**.
- **Leverage LLMs** to obtain a **list of ingredients** for a given class name.
  - The class name is used as a prompt to **generate the list of ingredients**.
  - A text composed of the class name and comma-separated ingredients is used as the textual representation of each class, which is **the input to the lexical encoder**.
- Feed with the **food images** and a **textual prompt** asking for the visible ingredients in the image.
  - A querying transformer is used to **query the input (image)** using the prompt.
  - This enables the model to produce **a list of ingr**edients, in this case only the visible ones, which is used as a textual representation of each image.
  - The **captions** are produced per image, obtaining several **lists of ingredients for each class**, which are **further processed** through the lexical encoder.
- The textual representations are processed by a **neural lexical encoder** which transforms the input sentence into a fixed-length embedding in the lexical feature space.
- A **clustering** is applied to detect similar classes.



Dining on Details: LLM-Guided Expert Networks for Fine-Grained Food Recognition



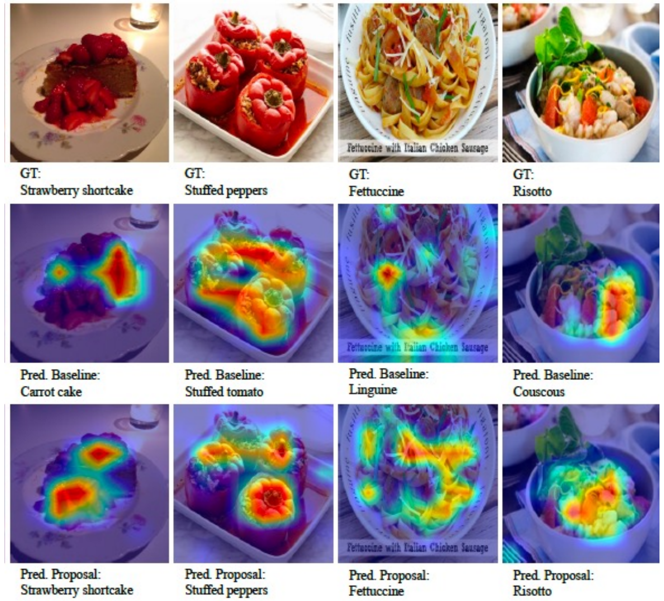
- Trained through an end-to-end **multi-task learning** process, this method **enhances performance** in the fine-grained food recognition task, showing exceptional prowess with highly similar classes

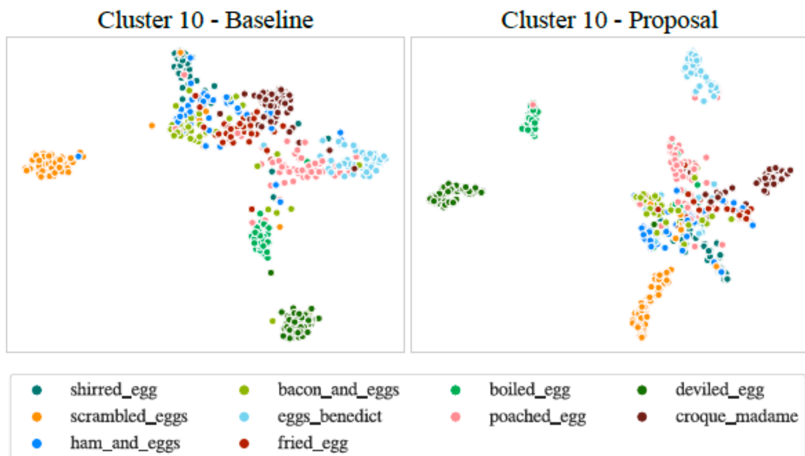
Validation

Method	Test Accuracy (%)
Grafit [61] (ICCV'21)	93.7
EffNet-B7 [59] (ICML'19)†	93.0
PMG [8] (CVPR'21)†§	87.5
FGFR [53] (Madima'22)§	93.8
DoD + SwinV2-B§	**94.9**

Comparison of DoD with SoTA methods in Food-10

† =bigger image size. °◊ =subset-based method



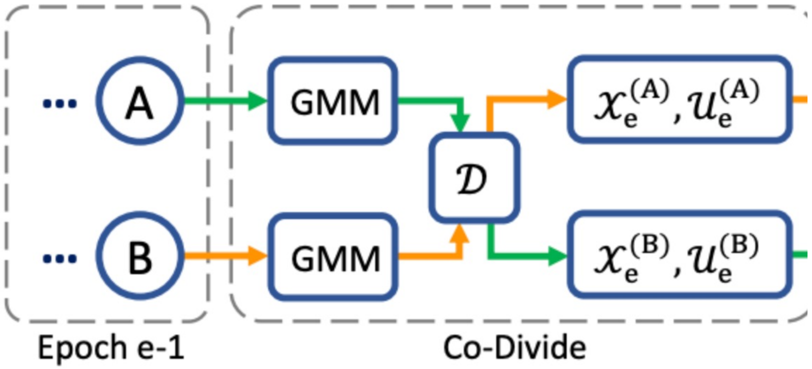


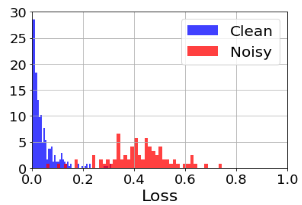
UMAP of the baseline and DoD

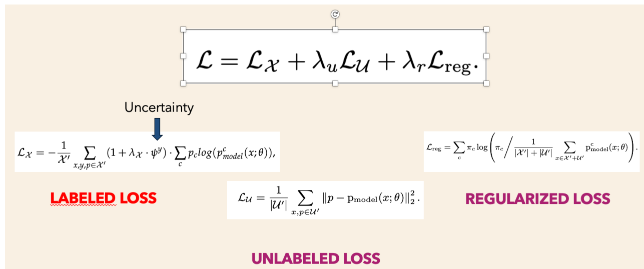
Learning with Noisy Labeling

Did you know that:

- 3.4% average error rate across all datasets,
- 6% for ImageNet
- MNIST digits dataset contains 15 (human-validated) label errors in the test set.







Tatjer, Albert, et al. "CCLM: Class-Conditional Label Noise Modelling." Iberian Conference on Pattern Recognition and Image Analysis. Cham: Springer Nature Switzerland, 2023. Best paper award.

Towards Volume Estimation: MomentsNeRF





Index

1. Why Food Recognition?

2. Self-Supervised Learning for Fine-Grained Recognition

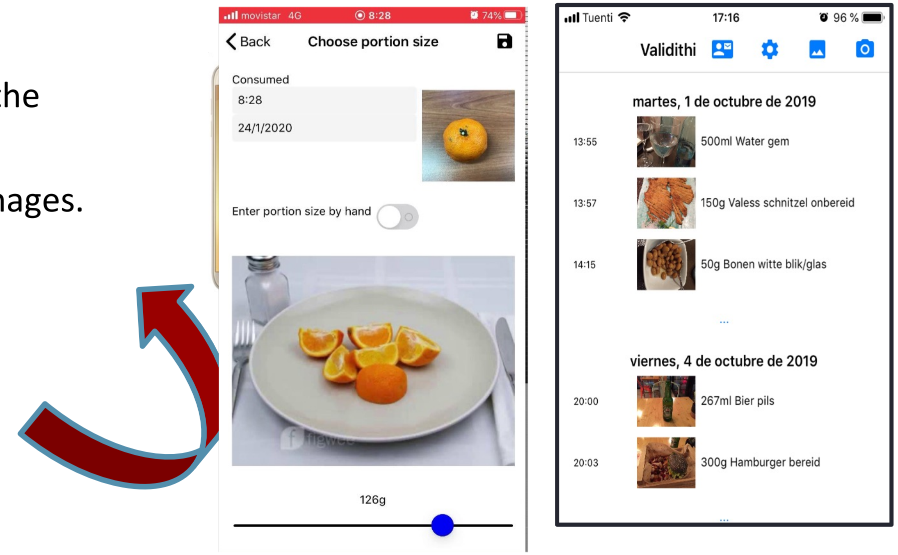
    - Validation of All4One

3. Other Food Recognition works

4. Food Recognition real applications

SUCCESS STORY: FOOD intake monitoring of kidney transplant patients

LogMeal is a HealthApp and API in the cloud that is able to automatically recognize and analyze food from images.



Validithi (EIT Health, 2019/20)

- **Automatic food diary** construction (UB).
- **Accurate, objective and continuous** food intake monitoring (UB).
- Semi-automatic **volume estimation** (Nestle).
- **Meal planner** and **health recommendations (**Nestle**)**.

umcg UNIVERSITAT DE BARCELONA Nestlé CVC

SUCCESS STORY Greenhabit: a serious game to promote change behaviour

Ted Talk of Chantal Linders: “Manage the Monster in Your Head”

Greenhabit (EIT Digital, 2020/21)

Conclusions

- Food recognition is a perfect test domain for powerful Machine/Deep Learning models

- **Food** Image Analysis is **highly underexplored problem** that could convert in an important **benchmark** for CV algorithms.

- ALI4One **combines its neighbour contrast objective with a feature redundancy reduction** objective, being beneficial in its overall performance.
  - **consistently outperforms SoTA** instance discrimination frameworks on popular image classification benchmarking datasets, namely, CIFAR-10, CIFAR-100 and ImageNet-100.

- Multiple other CV problems are highly relevant for food image analysis: uncertainty modelling, multi-task learning, vision-language models, fine-grained recognition, multi-scale classification, etc.

- Multiple **real applications** and **professional opportunities**

Our small group

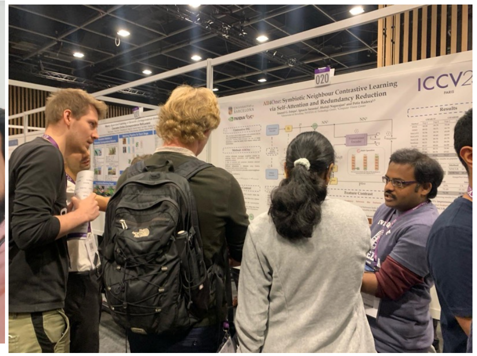














Thank you!
petia.ivanova@ub.edu
UNIVERSITAT DE BARCELONA