Few-Shot Learning

Jeong-Woon Kang
Master's Course
School of Information Convergence Engineering
Pusan National University
dnsdudwk1027@naver.com

2022.07.14

Contents

- Few-shot learning

Few-shot learning

Meta-Learning : Learning-to-learn, 학습을 잘 하는 방법을 학습
Few-Shot Learning : 적은 데이터만을 가지고 좋은 성능 도출

Few-shot learning은 meta learning의 한 종류

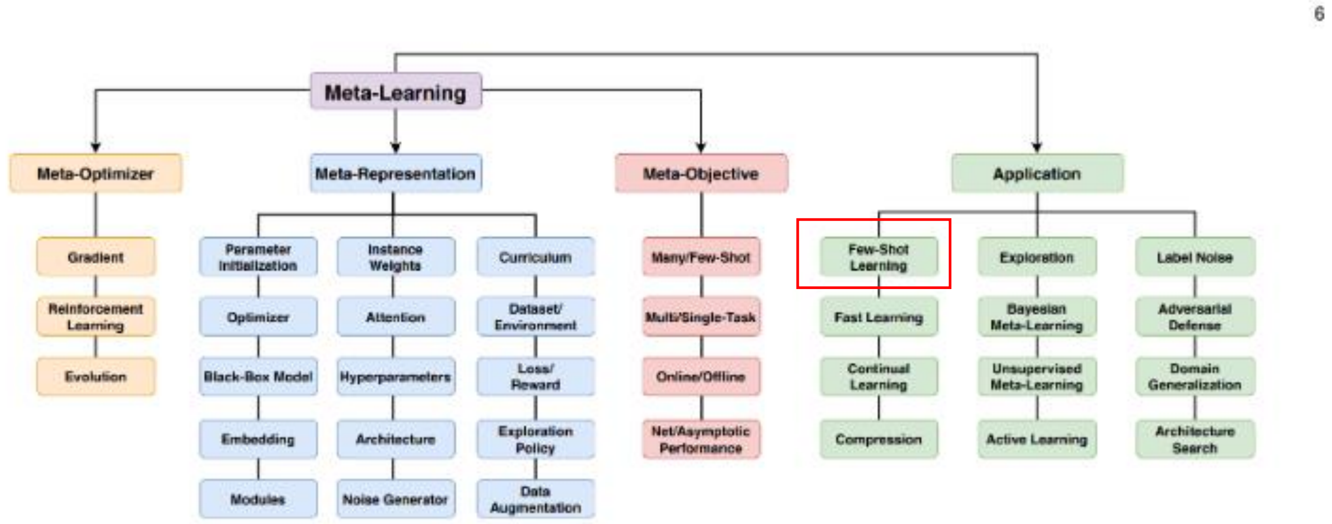


Fig. 1. Overview of the meta-learning landscape including algorithm design (meta-optimizer, meta-representation, meta-objective), and applications.

[wangshusen/DeepLearning (github.com)]

[Few-Shot Learning (1/3): Basic Concepts – YouTube]

[Few-Shot Learning (2/3): Siamese Networks – YouTube]

[Few-Shot Learning (3/3): Pretraining + Fine-tuning – YouTube]

[One-shot learning of object categories | IEEE Journals & Magazine | IEEE Xplore]

[Introduction to Few-Shot Learning & Meta Learning (velog.io)]

Support Set

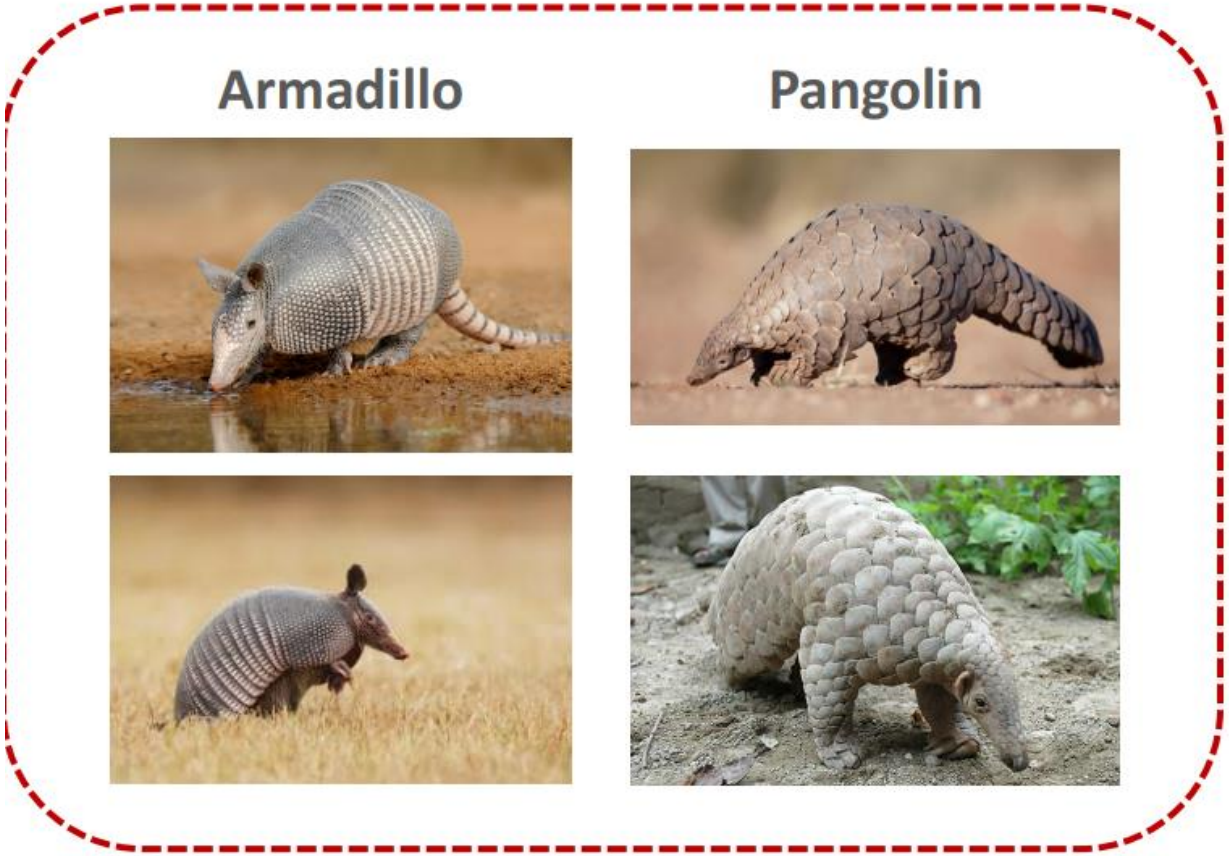
Armadillo

Pangolin

쿼리: 데이터베이스에 대한 특정 정보 요청

여기에서는 평가 데이터로 이해

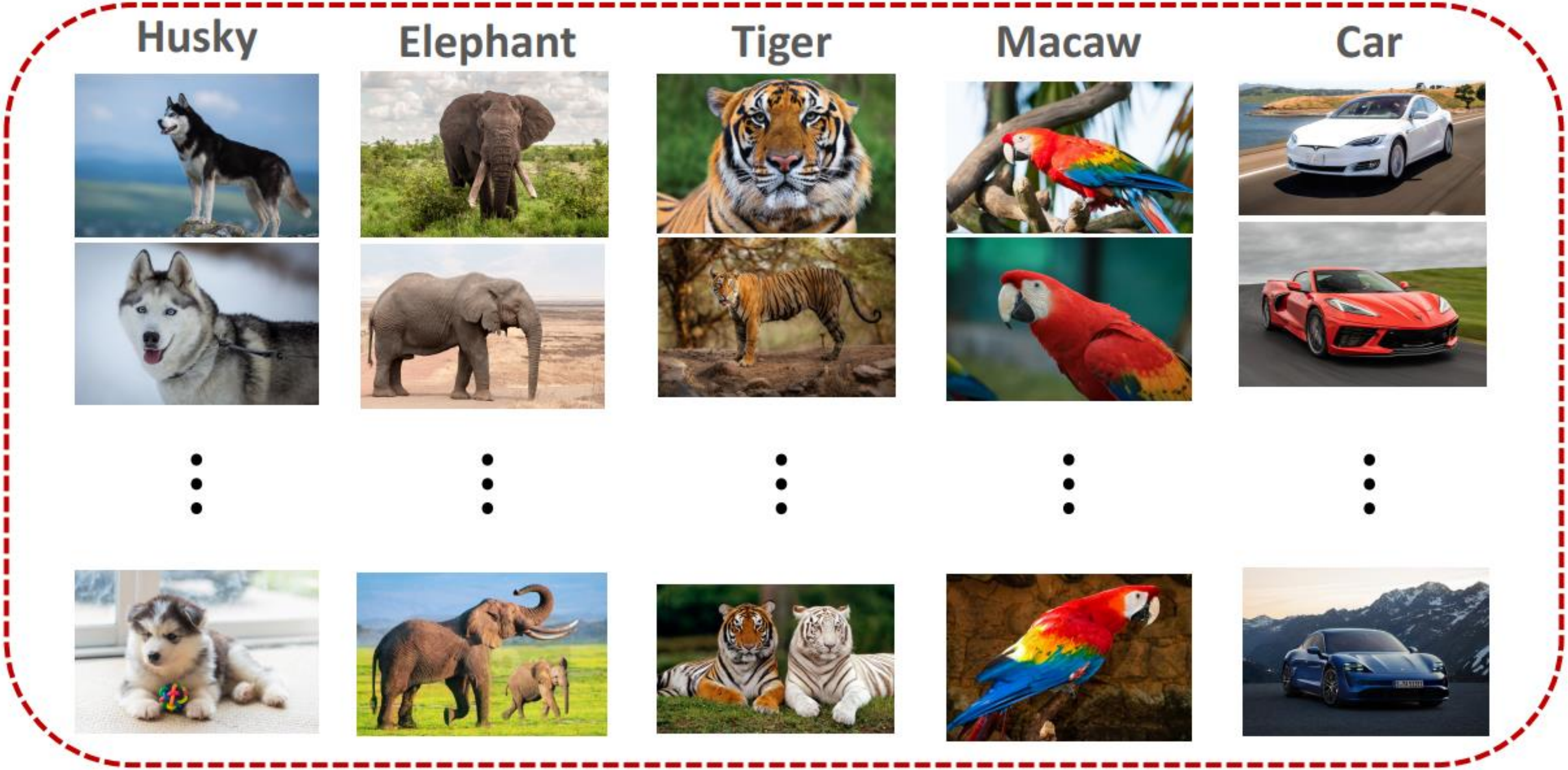
Support Set



Query

Armadillo or Pangolin?

Training Set








Few-Shot Learning

Query:

Support Set:

Fox

Squirrel

Rabbit

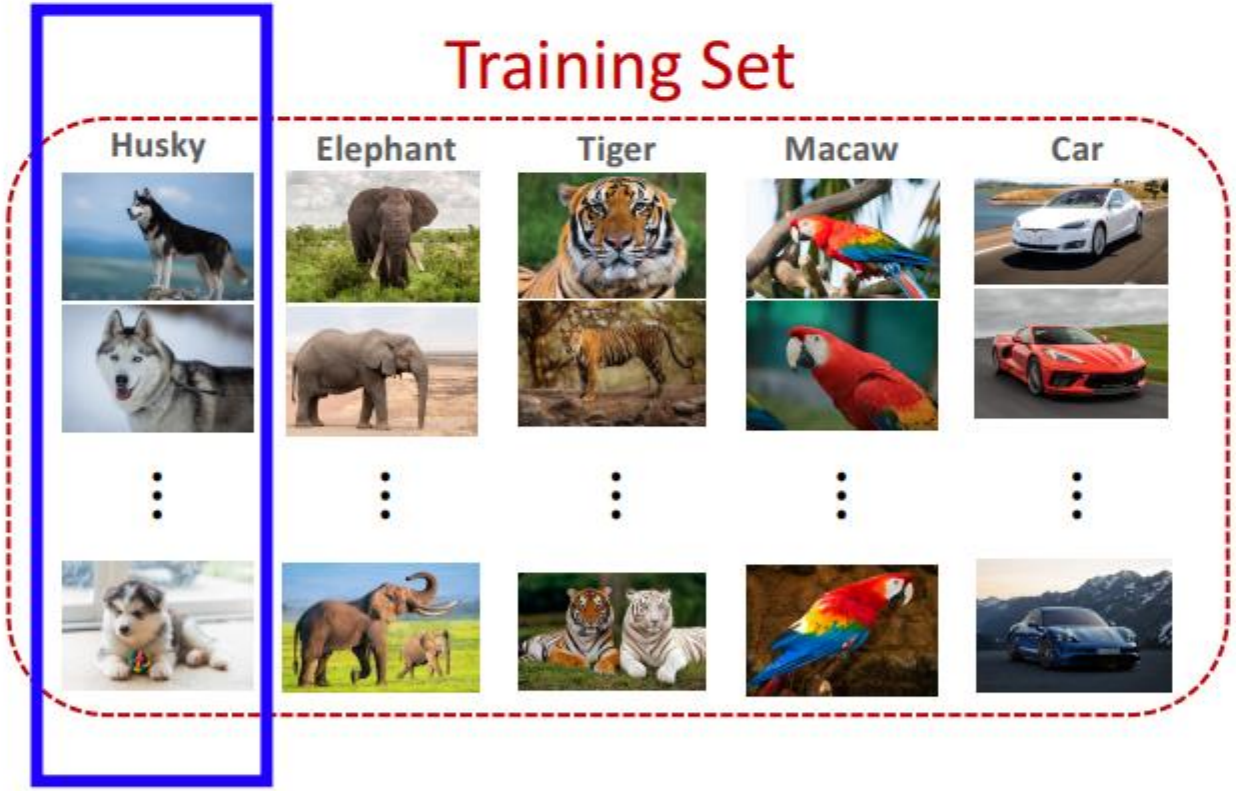
Hamster

Otter

Beaver

Supervised Learning vs. Few-Shot Learning

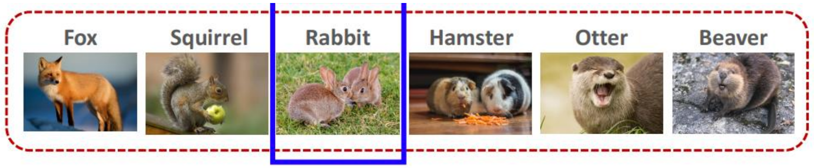
- Traditional supervised learning:
  - Test samples are never seen before.
  - Test samples are from known classes.



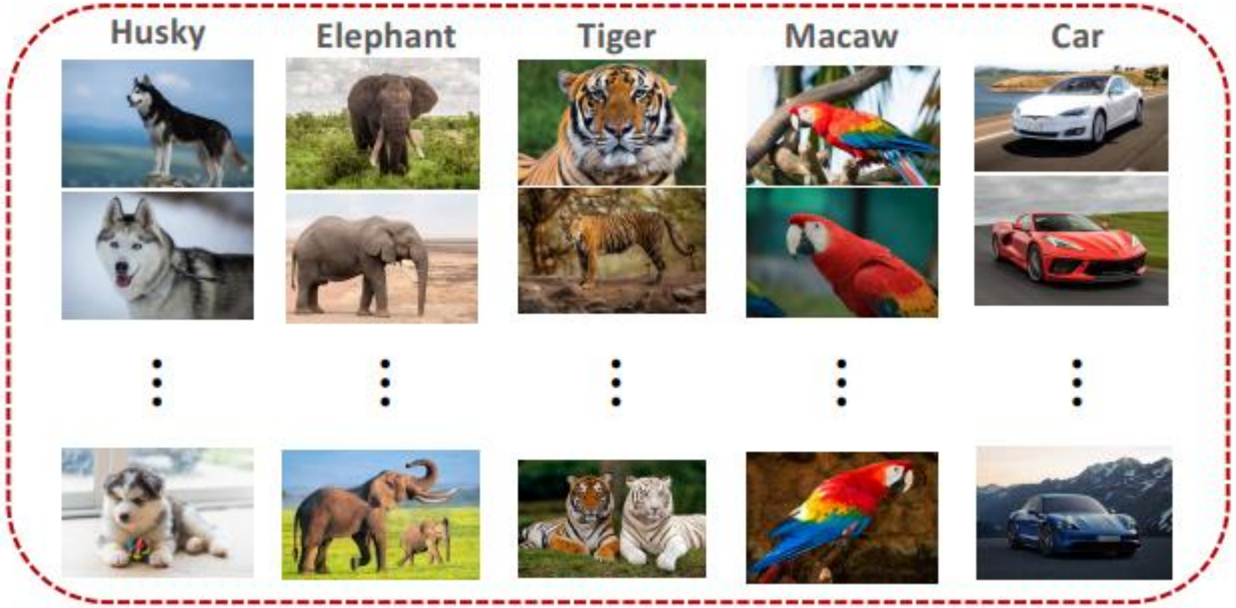
Supervised Learning vs. Few-Shot Learning

- Few-shot learning:
  - Query samples are never seen before.
  - Query samples are from unknown classes.

Support Set:



Training Set

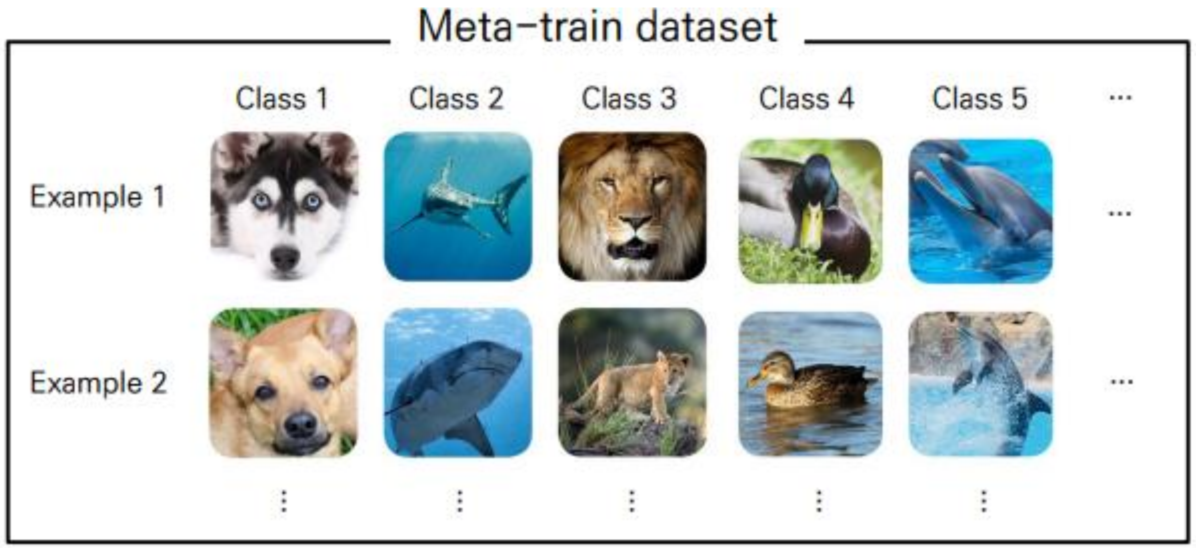


Query Sample

Training Strategy

❖ Episode training

$$\theta = \underset{\theta}{\operatorname{argmax}}\ E_{L\sim T}\left[E_{S\sim L,B\sim L}\left[\sum_{(x,y)\in B} logP_\theta(y|x,S)\right]\right]$$



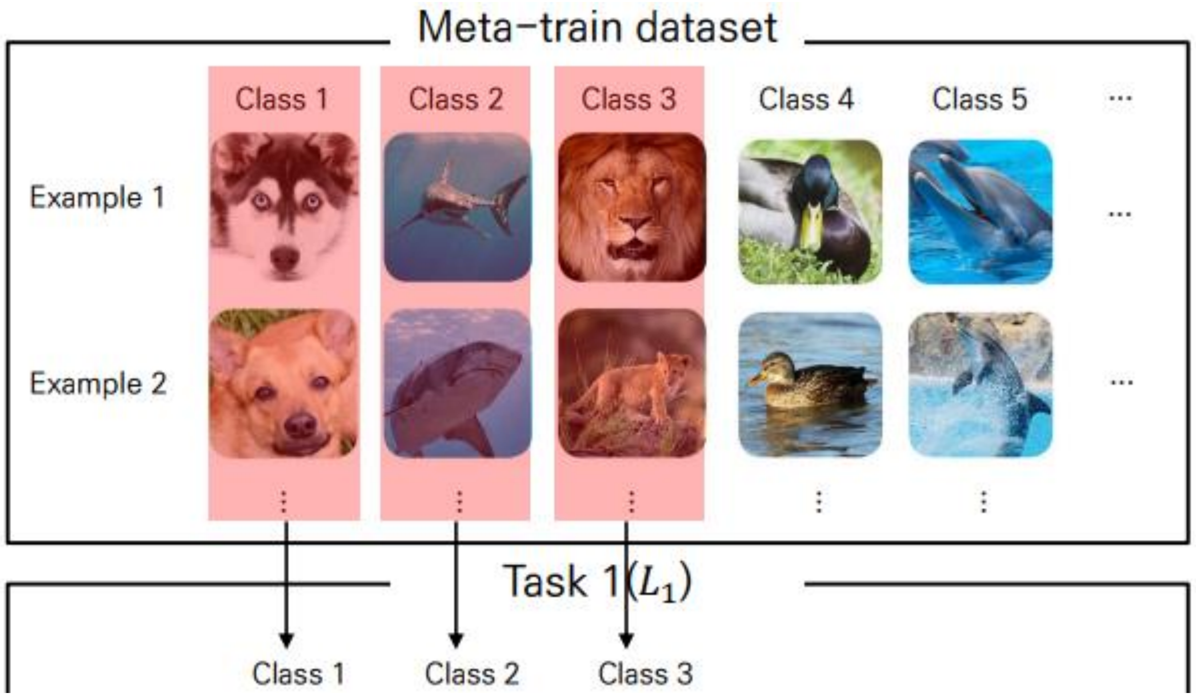
T(task): 가능한 label set L에 대한 분포

[Matching Networks for One Shot Learning (neurips.cc)]

❖ Episode training

$$\theta = \underset{\theta}{\operatorname{argmax}}\ E_{L\sim T}\left[E_{S\sim L,B\sim L}\left[\sum_{(x,y)\in B} logP_\theta(y|x,S)\right]\right]$$

- Task sampling($E_{L\sim T}$)
  - 모든 class 중 학습시킬 N개 class 샘플링



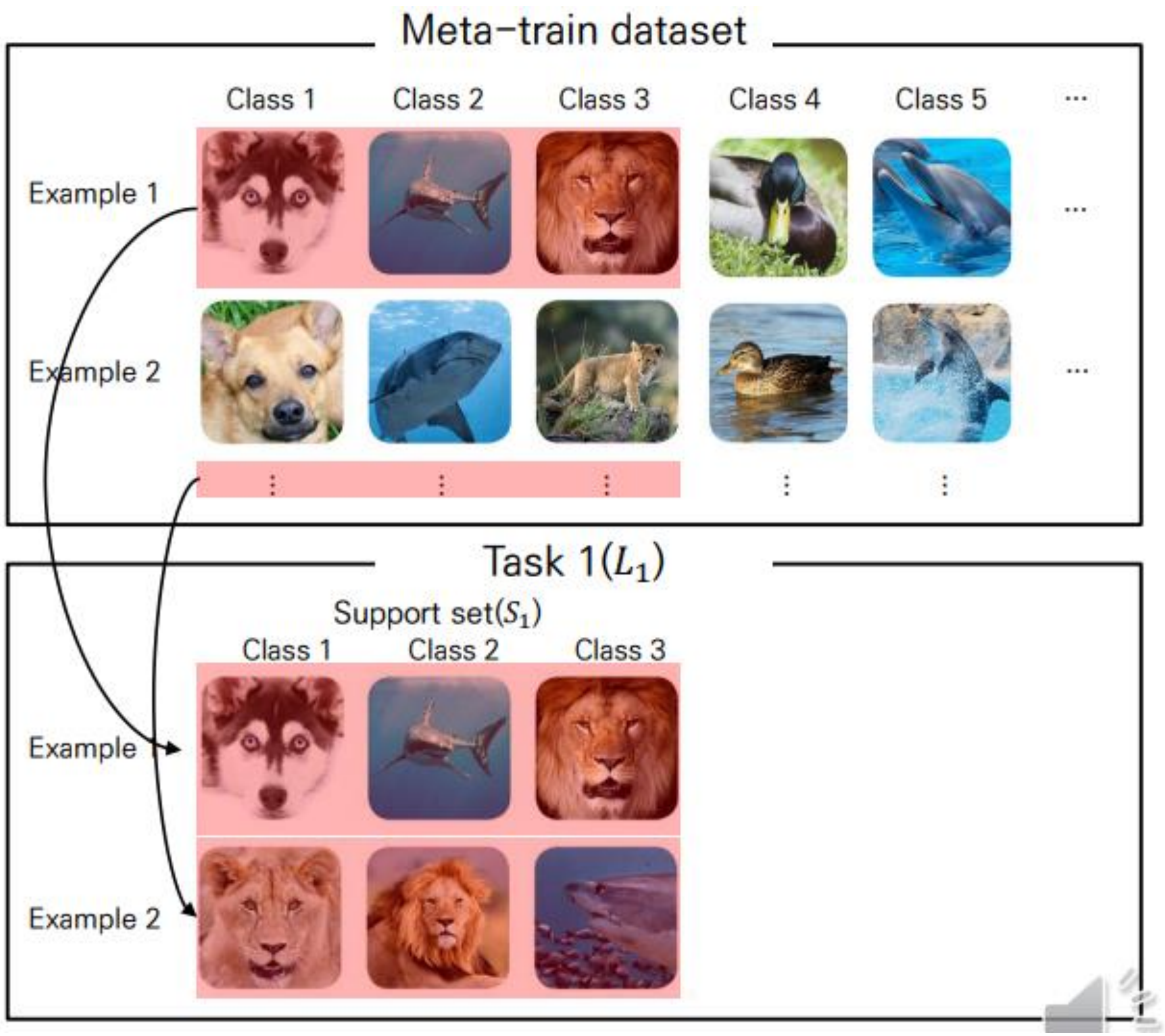
T(task): 가능한 label set L에 대한 분포

Ex) L:{cat, shark, lion}

❖ Episode training

$$\theta = \underset{\theta}{\operatorname{argmax}}\ E_{L\sim T}\left[E_{S\sim L,B\sim L}\left[\sum_{(x,y)\in B} logP_\theta(y|x,S)\right]\right]$$

- Task sampling($E_{L\sim T}$)
  - 모든 class 중 학습시킬 N개 class 샘플링
- Support set sampling($E_{S\sim L}$)
  - N개 class 별 각각 t개의 examples 샘플링

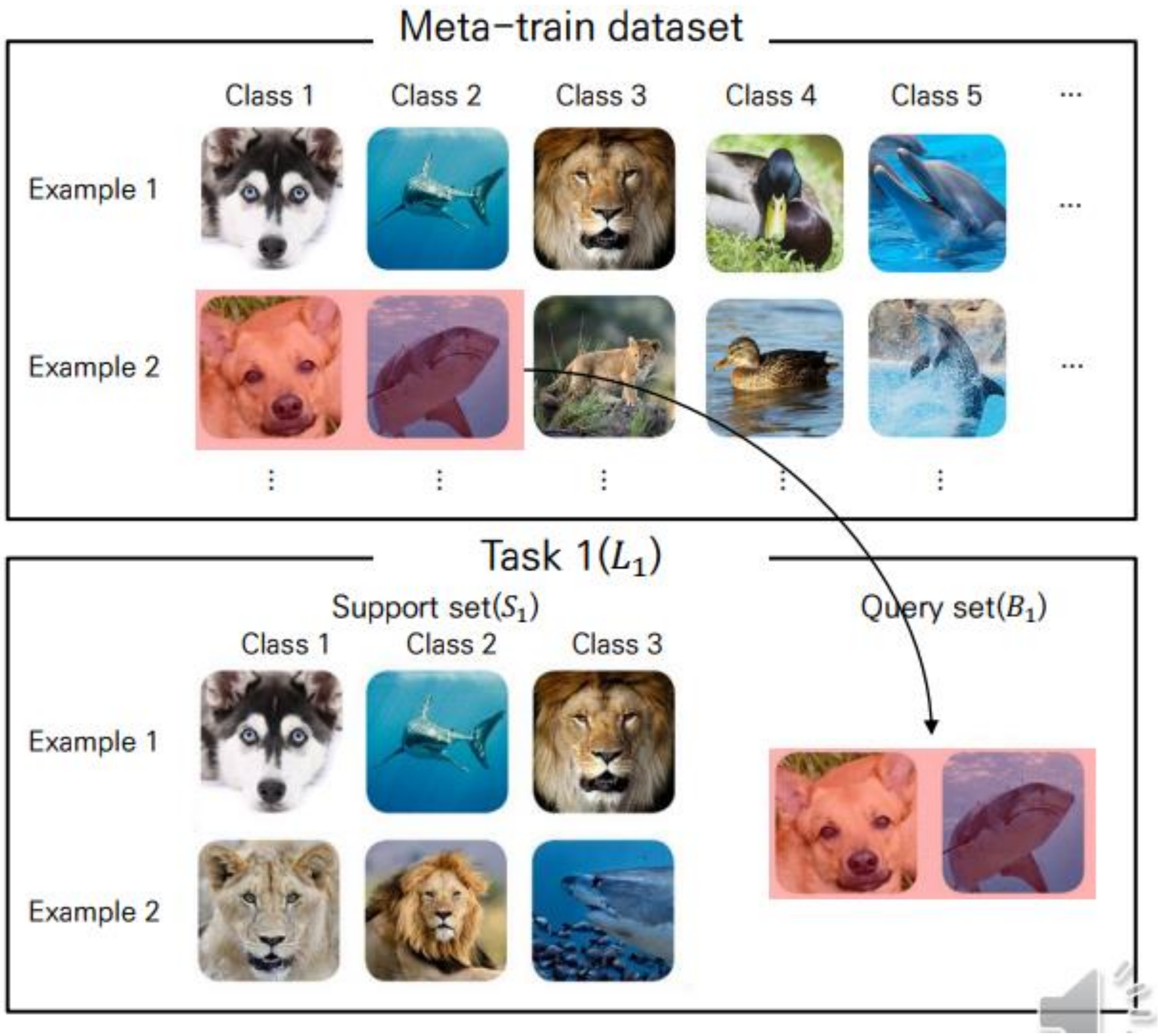


T(task): 가능한 label set L에 대한 분포
S : support set

❖ Episode training

$$\theta = \underset{\theta}{\operatorname{argmax}}\ E_{L\sim T}\left[E_{S\sim L,B\sim L}\left[\sum_{(x,y)\in B} logP_\theta(y|x,S)\right]\right]$$

- Task sampling($E_{L\sim T}$)
  - 모든 class 중 학습시킬 N개 class 샘플링
- Support set sampling($E_{S\sim L}$)
  - N개 class 별 각각 t개의 examples 샘플링
- Query set sampling($E_{B\sim L}$)
  - N개 class 별 각각 u개의 examples 샘플링

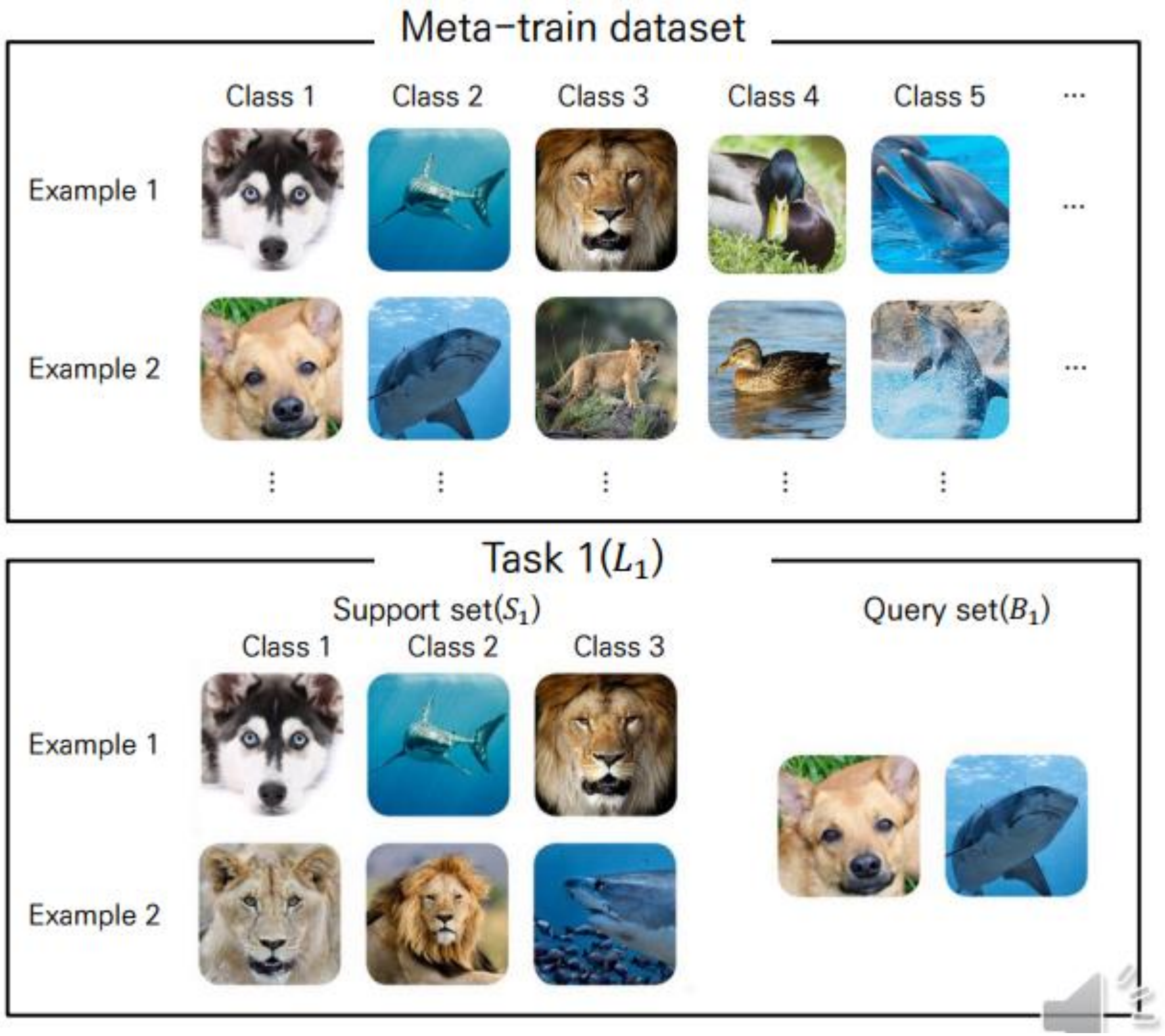


T(task): 가능한 label set L에 대한 분포
S : support set
B : batch

❖ Episode training

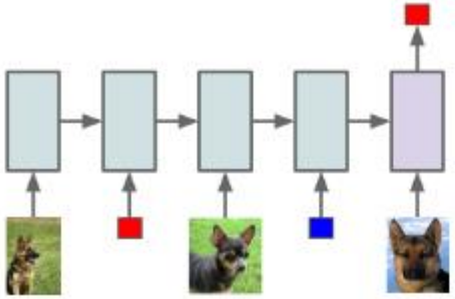
$$\theta = \underset{\theta}{\operatorname{argmax}}\ E_{L\sim T}\left[E_{S\sim L,B\sim L}\left[\textstyle\sum_{(x,y)\in B} \log P_\theta(y|x,S)\right]\right]$$

- Task sampling($E_{L\sim T}$)
  - 모든 class 중 학습시킬 N개 class 샘플링
- Support set sampling($E_{S\sim L}$)
  - N개 class 별 각각 t개의 examples 샘플링
- Query set sampling($E_{B\sim L}$)
  - N개 class 별 각각 u개의 examples 샘플링
- Query set의 class를 가장 잘 맞추는 파라미터 θ 학습



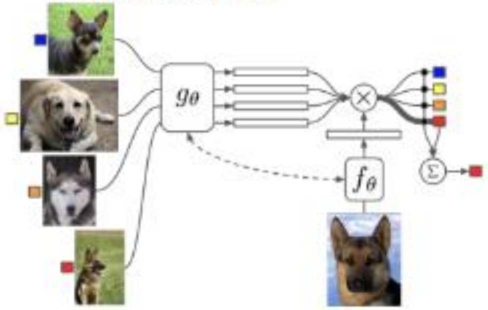
Meta Learning Models Taxonomy

Model Based



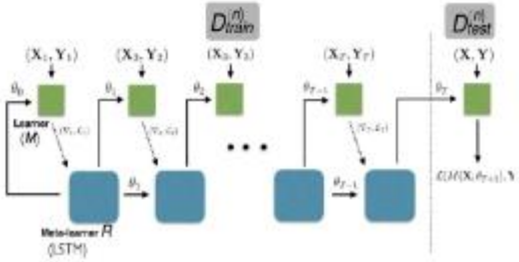
- Santoro et al. '16
- Duan et al. '17
- Wang et al. '17
- Munkhdalai & Yu '17
- Mishra et al. '17

Metric Based



- Koch '15
- Vinyals et al. '16
- Snell et al. '17
- Shyam et al. '17
- Sung et al. '17

Optimization Based

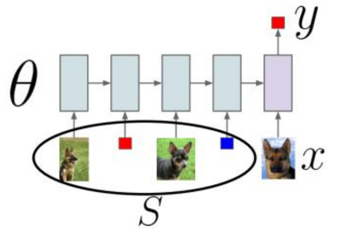


- Schmidhuber '87, '92
- Bengio et al. '90, '92
- Hochreiter et al. '01
- Li & Malik '16
- Andrychowicz et al. '16
- Ravi & Larochelle '17
- Finn et al. '17

Adapted from Finn '17

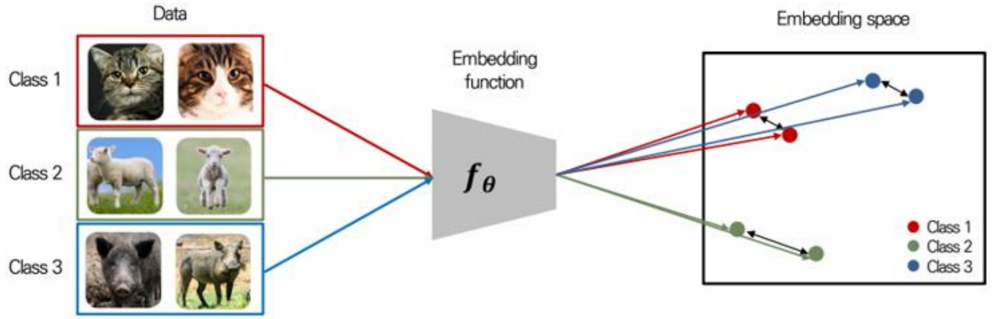
Oriol Vinyals, NIPS 17

Likelihood를 모델 f로 정의

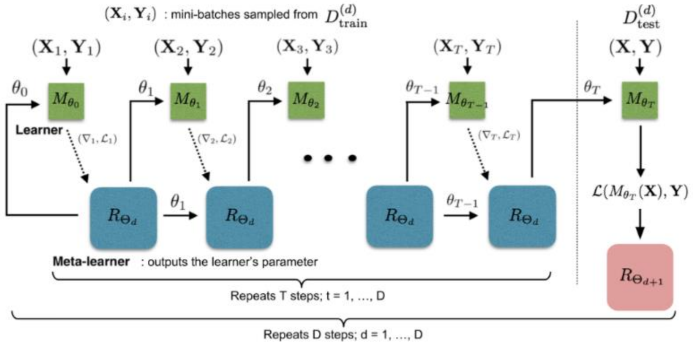


$$P_\theta(y|x,S) = f_\theta(x,S)$$

거리(metric) 기반



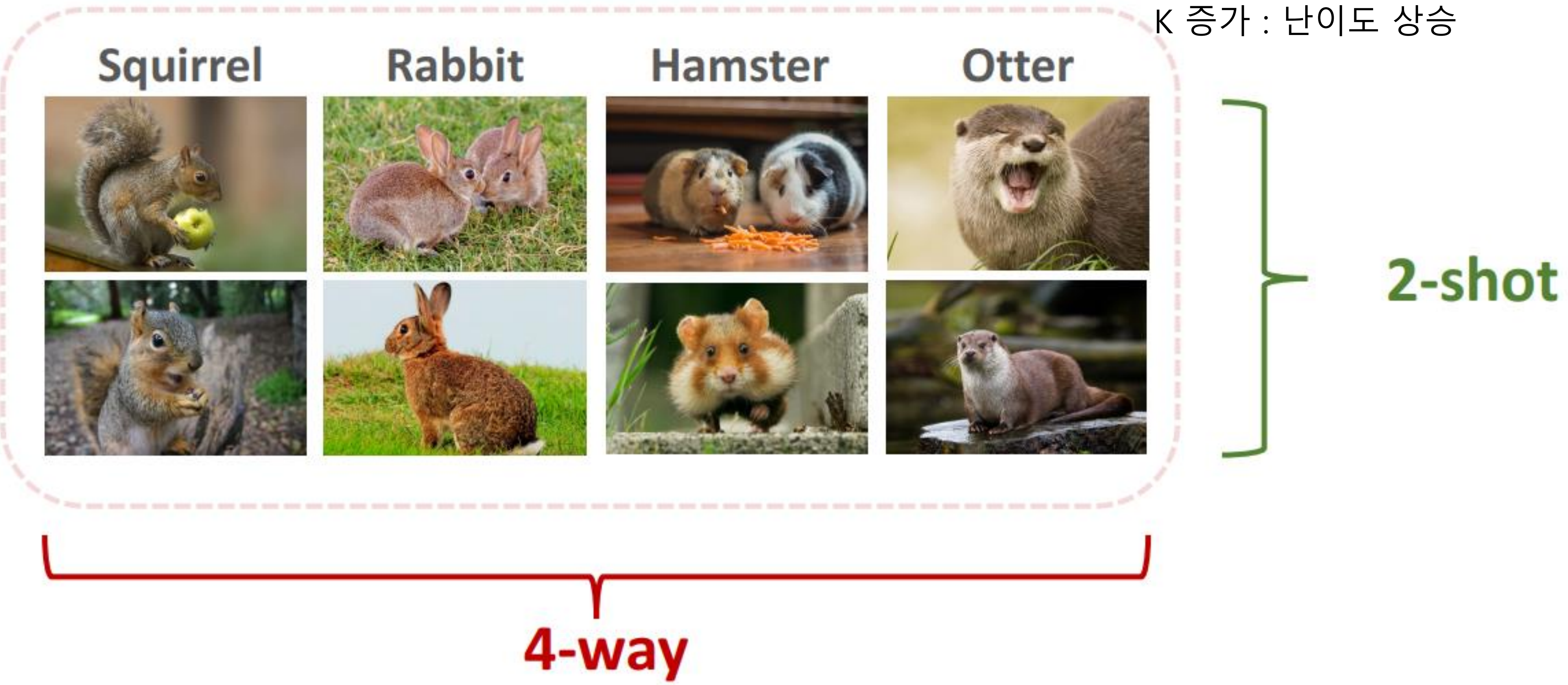
여러 개의 Learner M이 각각 task에서 optimization을 수행하고, Meta-learner R이 최종적으로 조절



k-way *n*-shot Support Set

K-way : class 개수
n-shot : class가 가진 샘플 개수
N 증가 : 난이도 감소
K 증가 : 난이도 상승

Support Set:



- Learn a similarity function: $\text{sim}(\mathbf{x}, \mathbf{x}')$.
- Ideally, $\text{sim}(\mathbf{x}_1, \mathbf{x}_2) = 1$, $\text{sim}(\mathbf{x}_1, \mathbf{x}_3) = 0$, and $\text{sim}(\mathbf{x}_2, \mathbf{x}_3) = 0$.

Bulldog

$\mathbf{x}_1$

Bulldog

$\mathbf{x}_2$

Fox

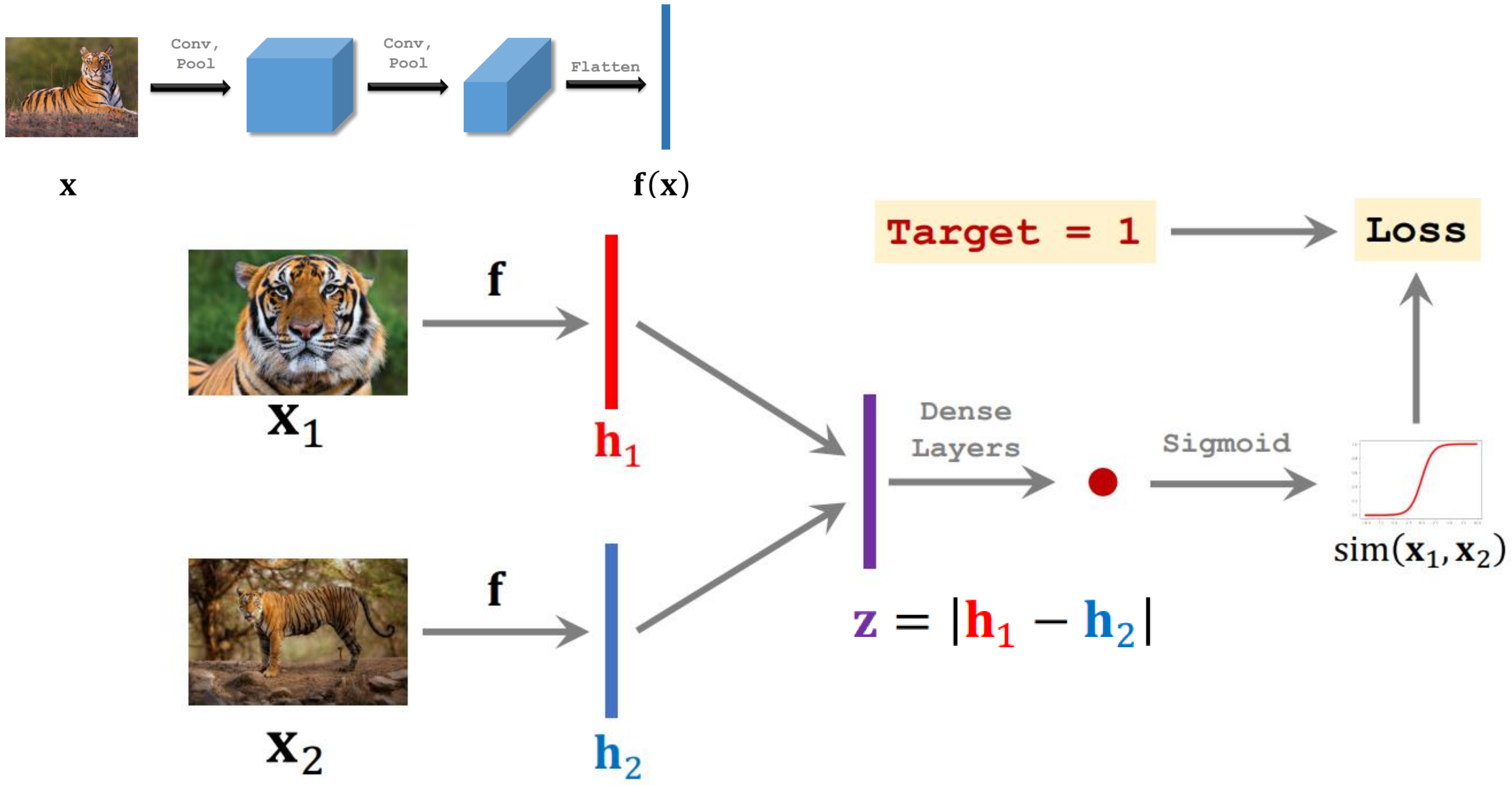
$\mathbf{x}_3$

What is in the image?

Query:

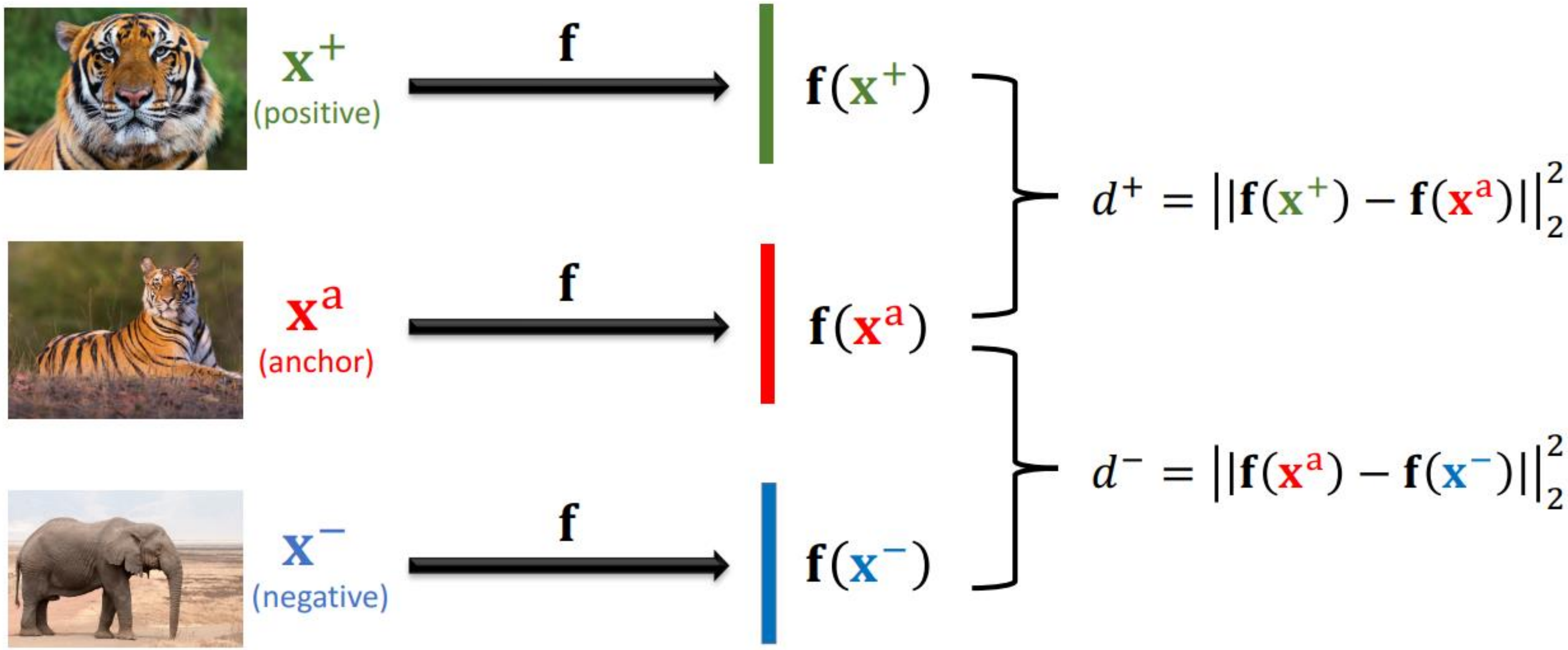
sim = 0.2	sim = 0.1	sim = 0.03	sim = 0.05	sim = 0.7	sim = 0.5
Greyhound	**Bulldog**	**Armadillo**	**Pangolin**	**Otter**	**Beaver**

Siamese network for pairwise similarity



Siamese network with triplet loss

Siamese network with triplet loss



$\mathbf{x}^{+}$ (positive)

$\mathbf{x}^{\mathrm{a}}$ (anchor)

$\mathbf{x}^{-}$ (negative)

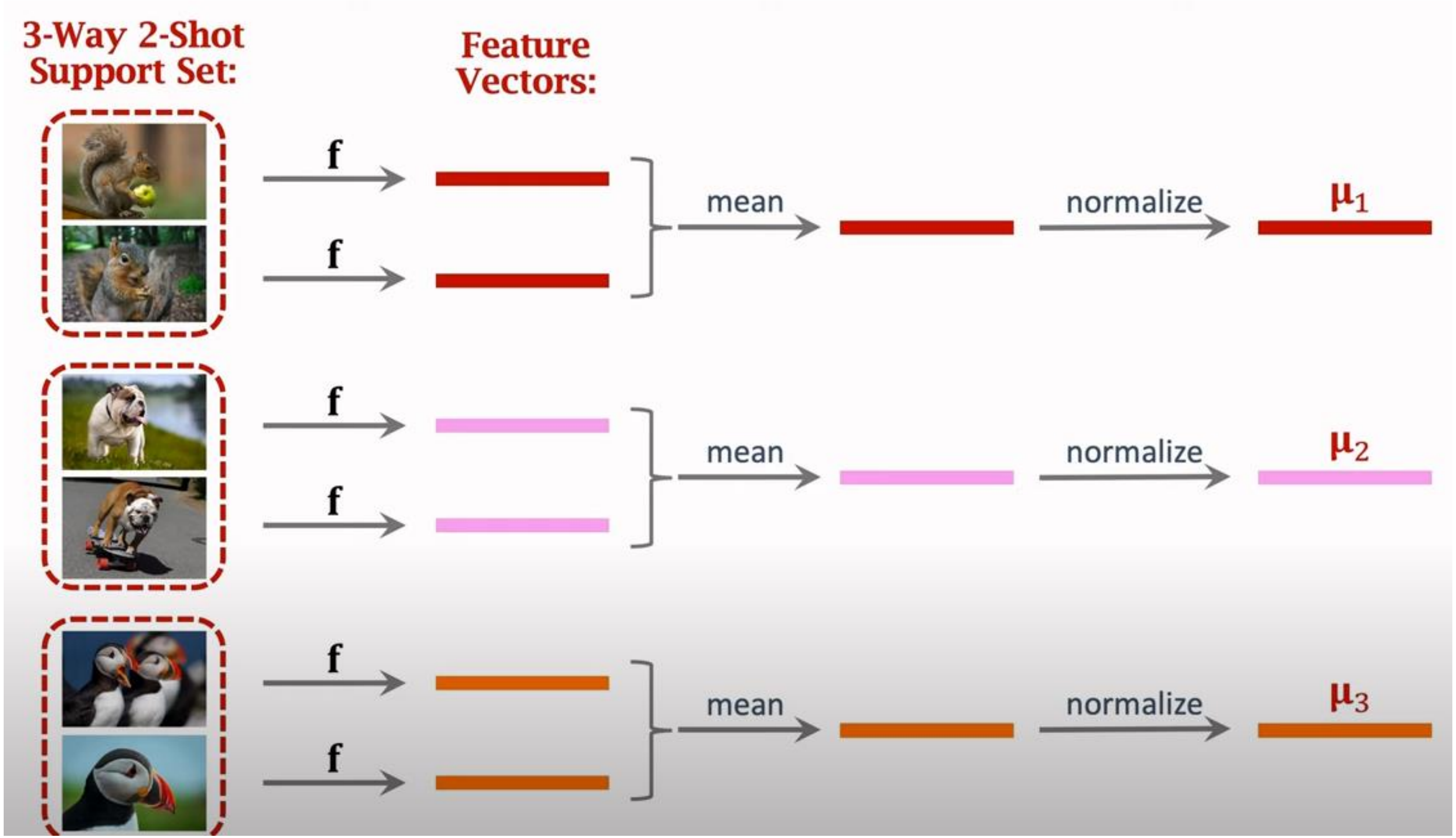
- Encourage $d^{+} = \left\|\mathbf{f}(\mathbf{x}^{+}) - \mathbf{f}(\mathbf{x}^{\mathrm{a}})\right\|_2^2$ to be small.
- Encourage $d^{-} = \left\|\mathbf{f}(\mathbf{x}^{\mathrm{a}}) - \mathbf{f}(\mathbf{x}^{-})\right\|_2^2$ to be big.
- If $d^{-} \geq d^{+} + \alpha$, then no loss. ($\alpha > 0$ is margin.)
- Otherwise, the loss is $d^{+} + \alpha - d^{-}$.
- $\mathrm{Loss}(\mathbf{x}^{\mathrm{a}}, \mathbf{x}^{+}, \mathbf{x}^{-}) = \max\{0,\ d^{+} + \alpha - d^{-}\}$.
- Update the CNN (function $\mathbf{f}$) to decrease the loss.

Query:

dist = 231	dist = 19	dist = 138	dist = 76	dist = 122	dist = 94
Fox	Squirrel	Rabbit	Hamster	Otter	Beaver





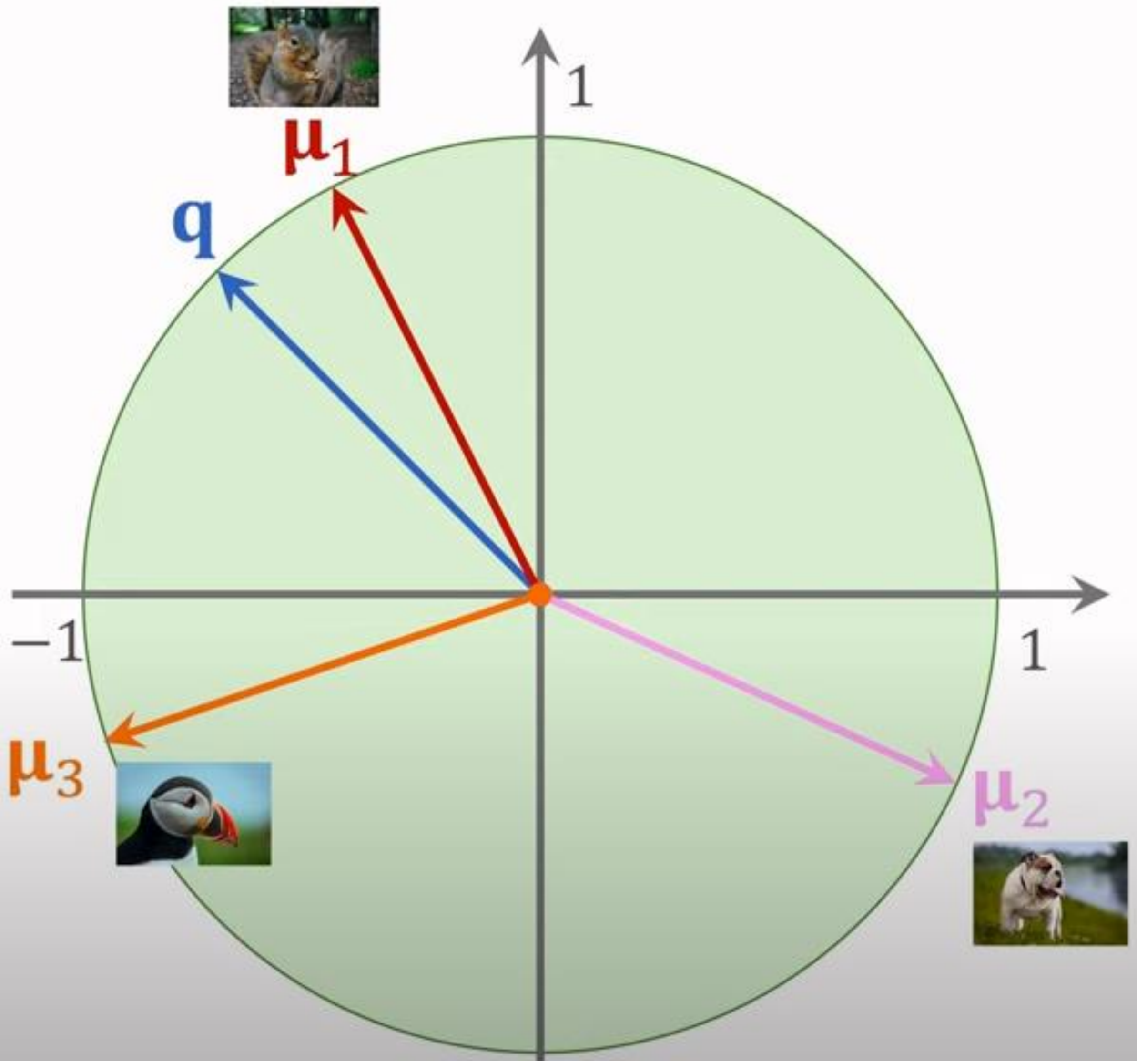
f: feature extractor(pretrained siamese network or supervised cnn or ...)

- Make prediction:

$$\mathbf{p} = \mathrm{Softmax}(\mathbf{M}\mathbf{q})$$

$$= \mathrm{Softmax}\left(\begin{bmatrix} \boxed{\boldsymbol{\mu}_1^T \mathbf{q}} \\ \boldsymbol{\mu}_2^T \mathbf{q} \\ \boldsymbol{\mu}_3^T \mathbf{q} \end{bmatrix}\right).$$

- Which entry of **p** is the biggest?

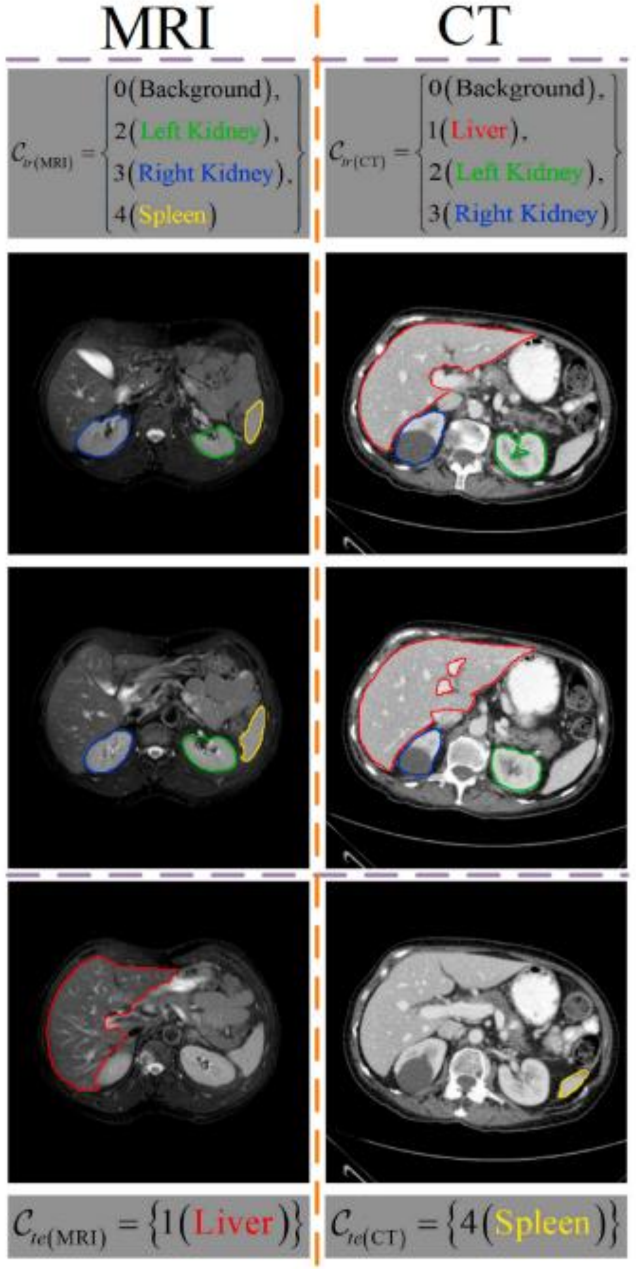


Few-shot learning in medical imaging

[Few-shot medical image segmentation using a global correlation network with discriminative embedding (clinicalkey.com)]

Sun, Liyan, et al. "Few-shot medical image segmentation using a global correlation network with discriminative embedding." *Computers in biology and medicine* 140 (2022): 105067.

Ex)
MRI에서는 간(liver)를 unseen class로
CT에서는 비장(spleen)을 unseen class로 놓음



train

test

Fig. 1. Axial abdomen multi-organ annotated CT/MRI dataset. The MRI data are shown in the left column, and the CT data are in the right. For effective illustration, we kept the liver as the unseen class in the testing for MRI and the spleen as the unseen class in the testing for CT. For MRI/CT, two images drawn from the training dataset and one image from the testing dataset are presented.

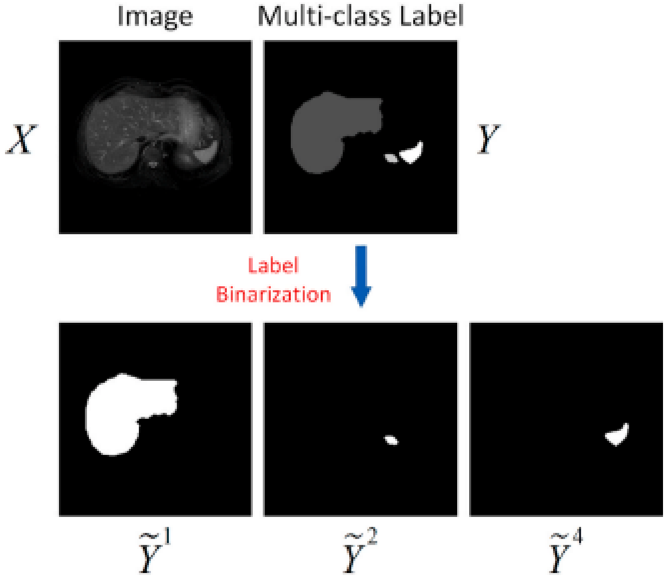


Fig. 2. Label binarization of an annotated slice.

GCN-DE

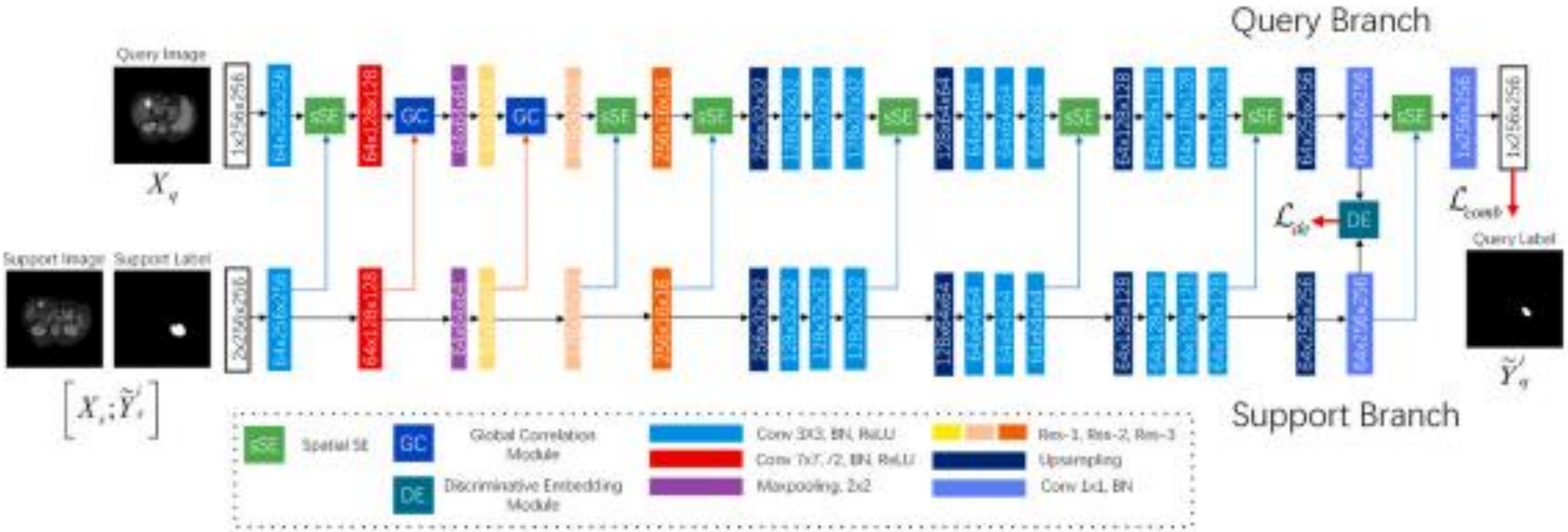


Fig. 3. Network architecture of the proposed global correlation network with discriminative embedding. The support image and its annotation on one foreground class are concatenated as a two-channel input for the support branch. The support features are leveraged to produce spatial attention in the spatial squeeze-and-excitation and global correlation modules. The discriminative embedding module takes the query and support features as input for annotations of different organs.

Global correlation network with discriminative embedding

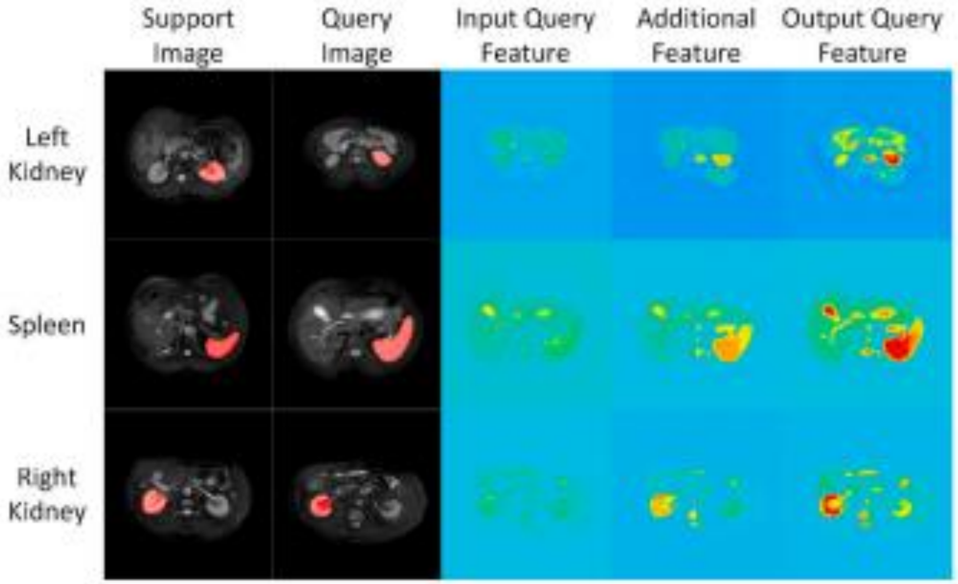
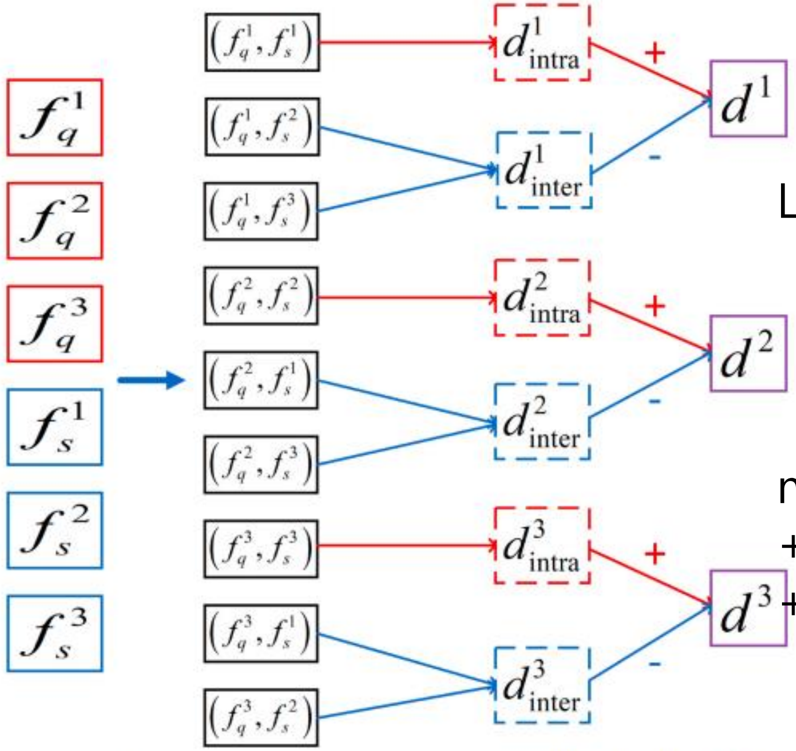


Fig. 6. Some visual examples from the global correlation module. The input query feature, its supplemented additional feature generated by global correlation, and the output query feature are presented. The foreground objects in the query images are highlighted are highlighted.



L2 distance

max(0,d1)
+max(0,d2)
+max(0,d3)

Fig. 8. Illustration of the constraint for discriminative embedding when $\mathcal{C}_{Y_q} = \mathcal{C}_{Y_s} = \{1, 2, 3\}$. We have $\{f_q^1, f_q^2, f_q^3\}$ and $\{f_s^1, f_s^2, f_s^3\}$. The operation $(\cdot, \cdot)$ represents the computation of L2-norm distance.

$$\mathcal{L}_{dice}^j = 1 - \frac{2\sum_i P_q^j(i)\widetilde{Y}_q^j(i)}{\sum_i P_q^j(i) + \sum_i \widetilde{Y}_q^j(i)},$$

$$\mathcal{L}_{bce}^j = -\frac{1}{HW}\sum_i \widetilde{Y}_q^j(i)\log\left(P_q^j(i)\right),$$

$$\mathcal{L}_{comb} = \frac{1}{|\mathcal{C}_{Y_q}|}\sum_j (\mathcal{L}_{dice}^j + \mathcal{L}_{bce}^j), \quad j \in \mathcal{C}_{Y_q}.$$

The overall loss function $\mathcal{L}_{overall}$ is stated as

$$\mathcal{L}_{overall} = \mathcal{L}_{comb} + \lambda_{de}\mathcal{L}_{de}.$$

Quantitative results measured in DC scores in percentiles of the ablation study and compared SOTA methods.

Modality	MRI					CT				
Organ	Liver	Spleen	Left Kidney	Right Kidney	Mean	Liver	Spleen	Left Kidney	Right Kidney	Mean
OSLSM	25.73	34.66	29.21	22.61	28.00	29.65	19.40	15.82	7.54	18.08
co-FCN	**53.74**	57.41	60.62	71.13	60.70	**47.50**	43.86	41.30	33.51	41.53
PANet	51.37	43.59	25.54	26.45	36.74	44.25	30.49	25.30	22.95	30.75
SG-One	50.33	42.41	26.79	24.16	35.92	44.98	30.88	26.79	20.88	30.88
SE-FSS	40.32	48.93	62.56	65.81	54.38	44.51	40.52	40.10	34.80	39.97
GCN w/o GC Module & Pretraining (Baseline w/o Pretraining)	44.88	57.26	60.20	65.89	57.06	42.55	43.88	44.15	33.95	41.13
GCN w/o GC Module (Baseline)	45.67	58.67	61.33	67.33	58.25	44.67	45.67	45.67	35.33	42.83
GCN	51.33	58.67	63.67	70.33	61.00	47.00	46.67	42.33	35.00	42.75
GCN-DE (Ours)	49.47	**60.63**	**76.07**	**83.03**	**67.30**	46.77	**56.53**	**68.13**	**75.50**	**61.73**

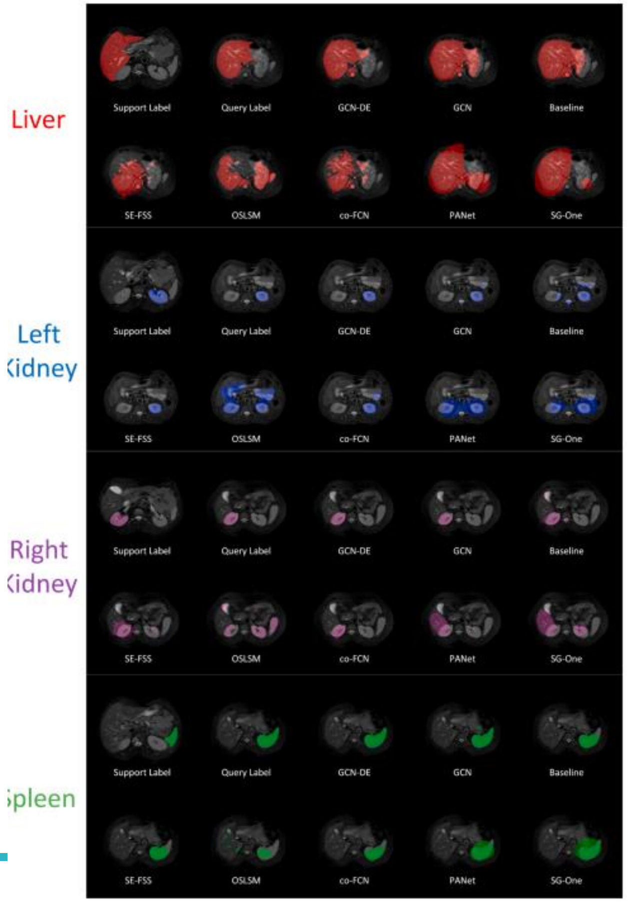


Fig. 10. Segmentation prediction of compared methods regarding MR images.

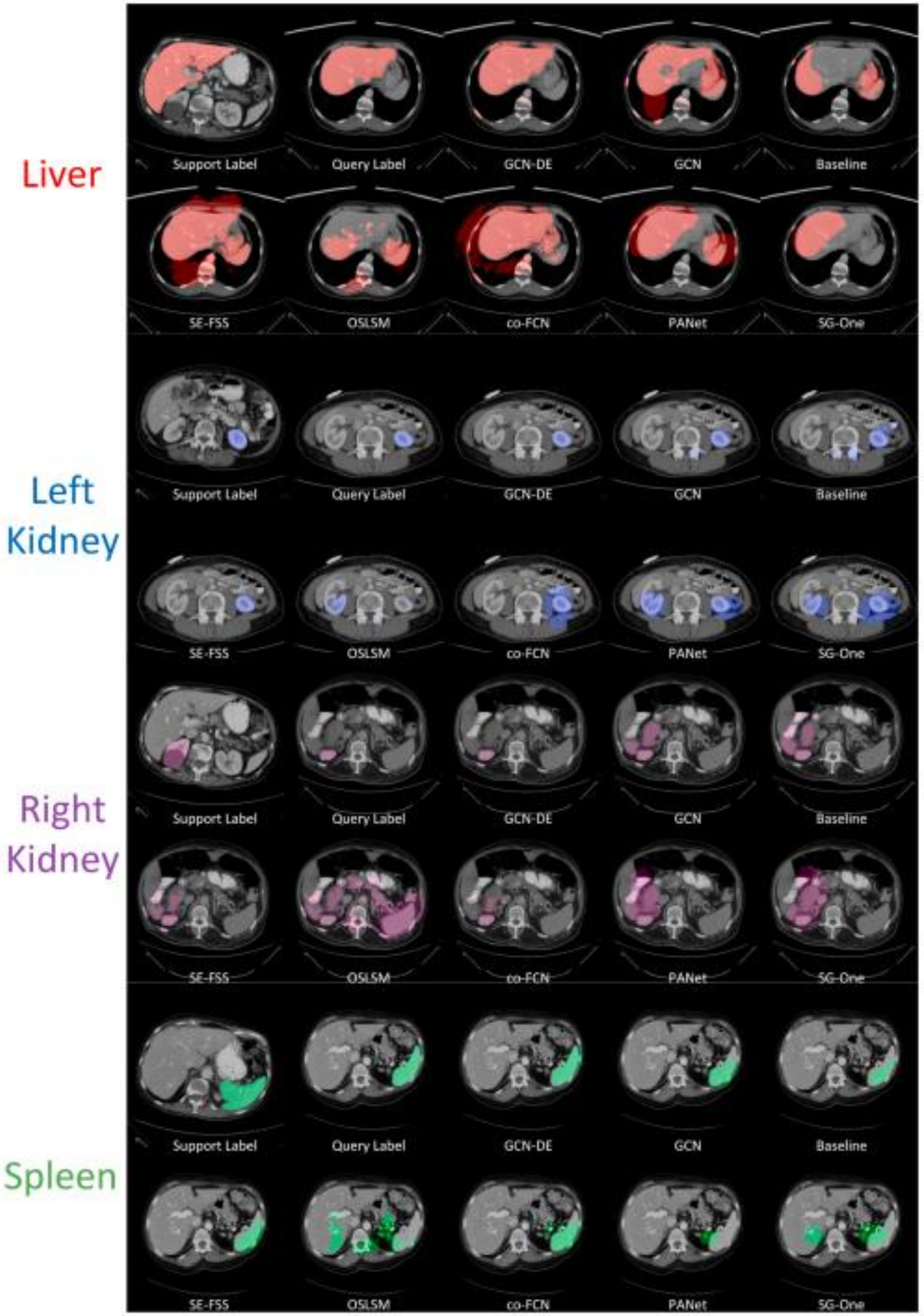


Fig. 9. Segmentation prediction of compared methods regarding CT images.

Zero-shot learning

+Semantic information

Stripped,
Long nose,
…