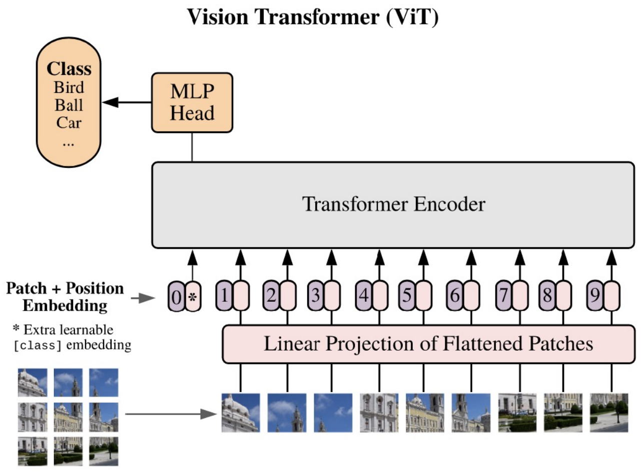
Safety in AI

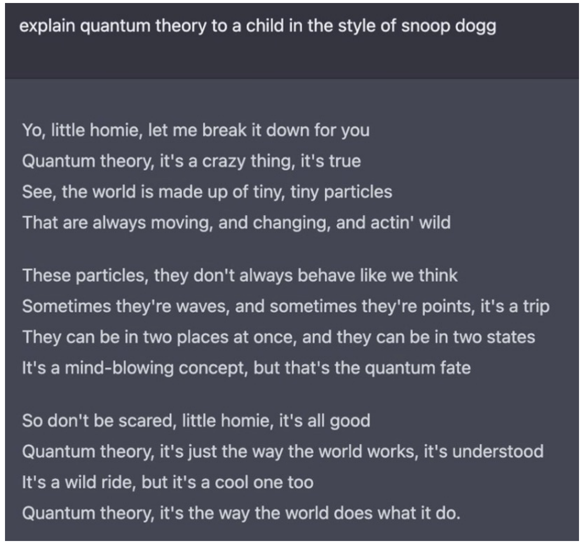
Nicolas THOME – Prof at Sorbonne University

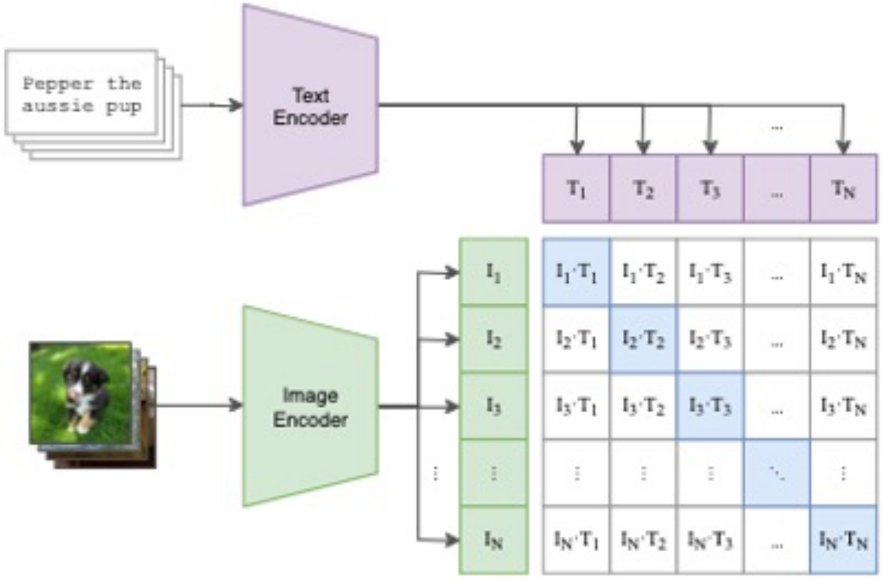
ISIR Lab, MLIA Team

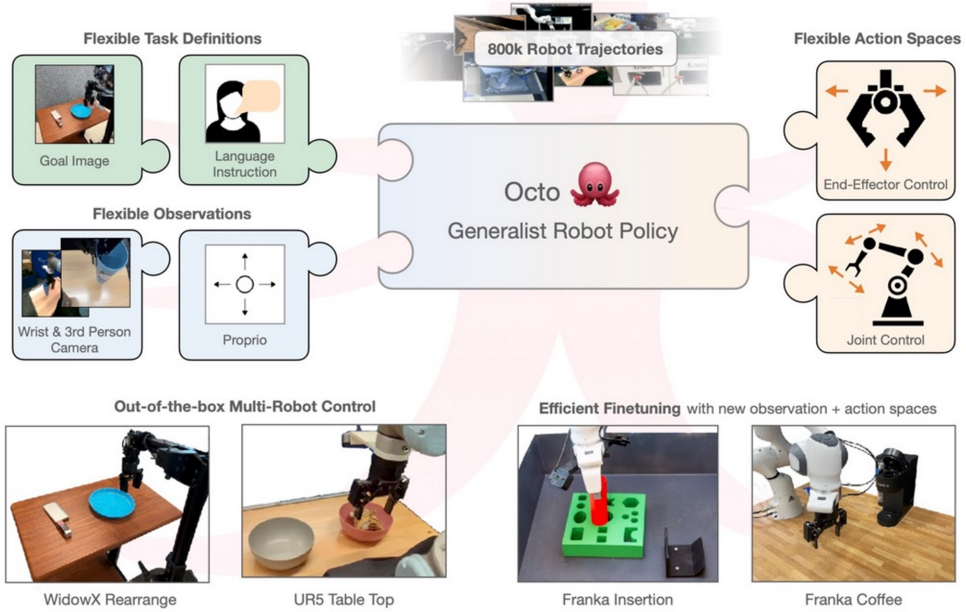
Context: AI/ML summer

- AI in the last decade: huge performance boost
  - Vision, NLP, multi-modal prediction robotics







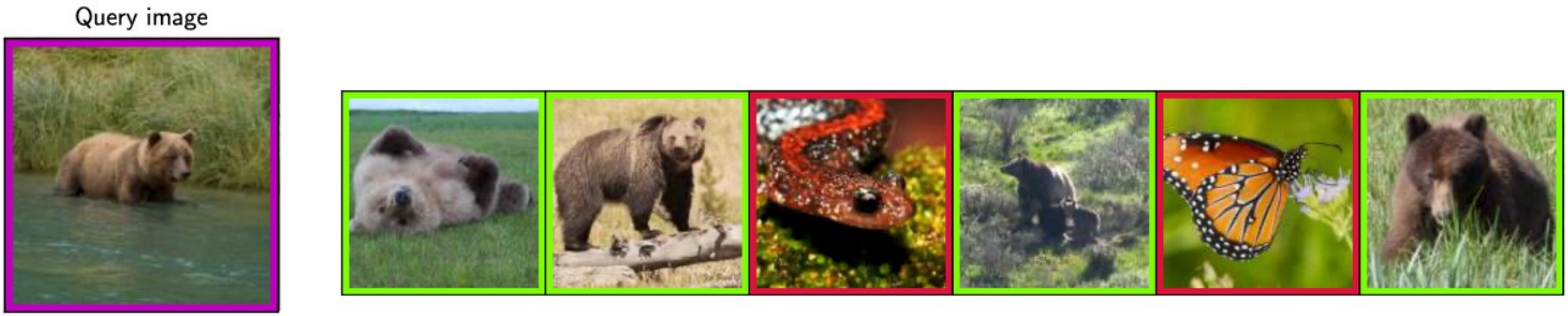


Context: safety in deep learning

Robustness: several brittleness aspects in deep learning

- Explainability, biases & shortcuts, fairness, *etc*

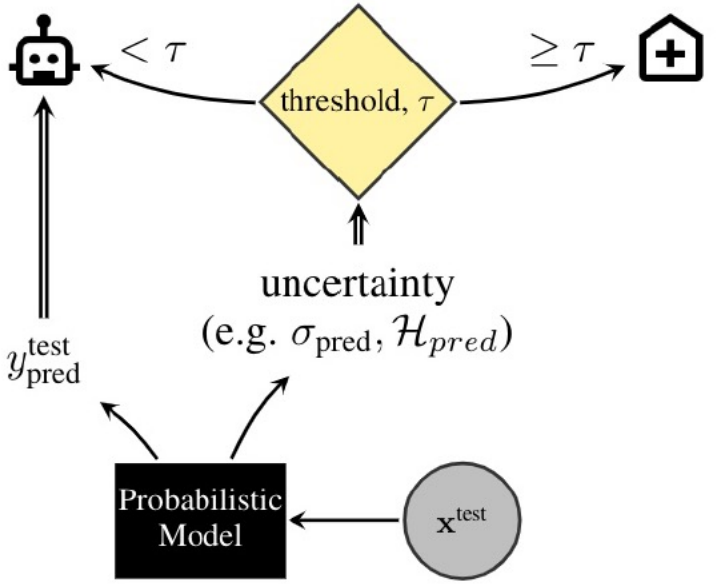
i) Stability: adversarial examples, mistake severity



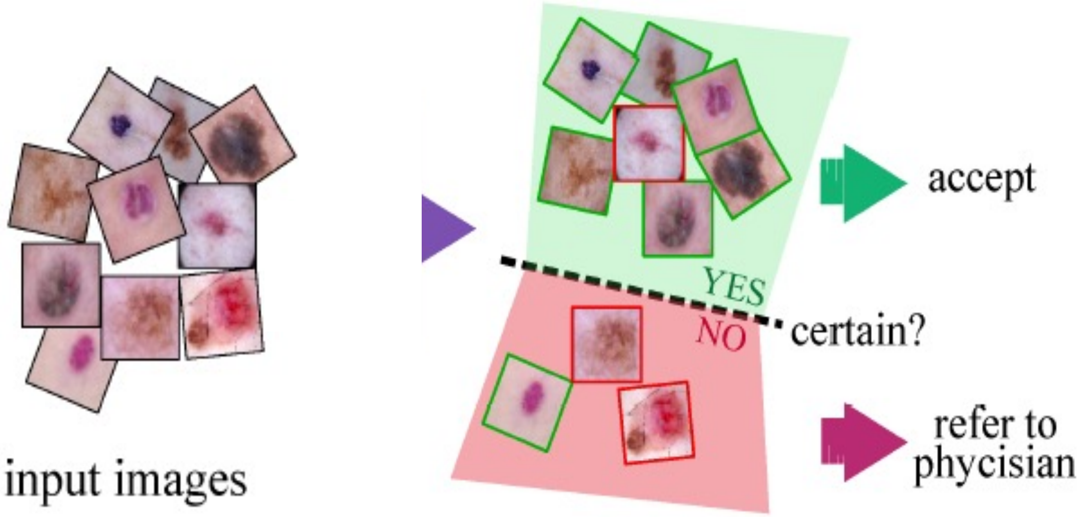
Context: robustness in deep learning

ii) Uncertainty Quantification (UQ)

“Know when you do not know”



Abstain to make a prediction



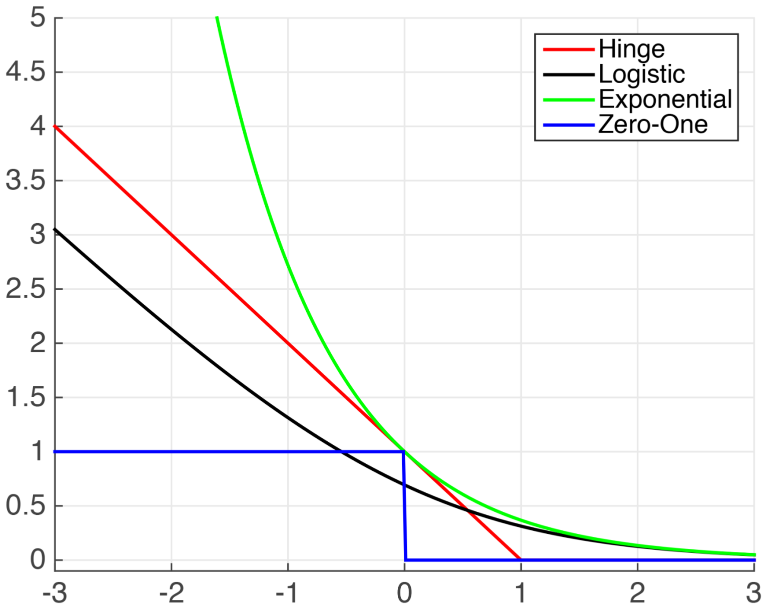
UQ: a challenge in DL

- **Which uncertainty score?**
- **Calibration, ranking correct/incorrect prediction**

Context: robustness in deep learning

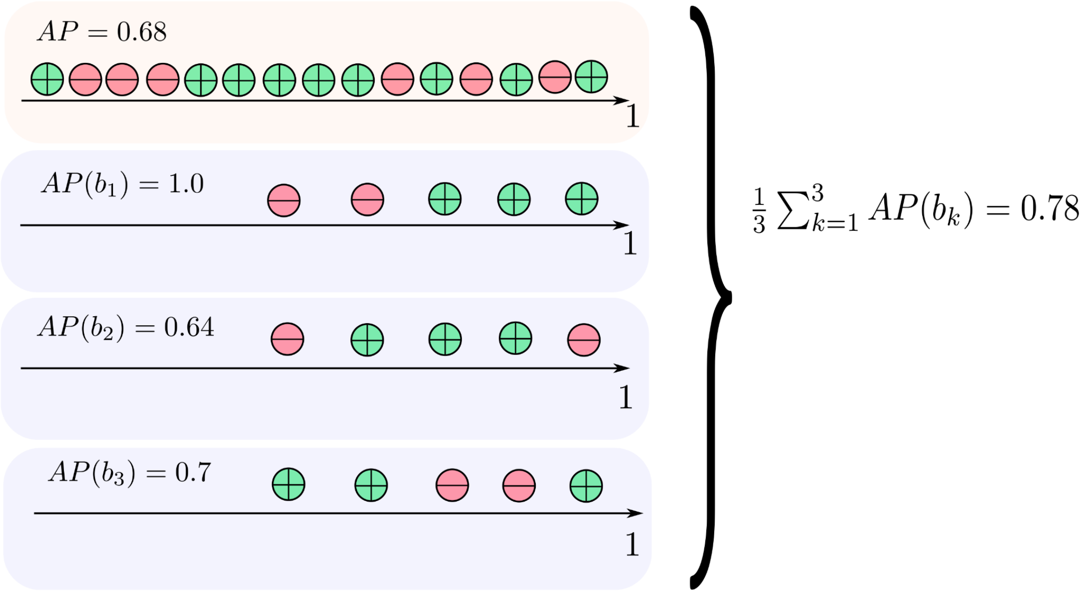
iii) Training: direct optimization of target metrics, image retrieval

- Non-differentiable losses



"Good" (with guarantees) surrogates for classification, what about other metrics, rank losses?

- Non-decomposable losses
  - Rank losses, e.g. Average Precision



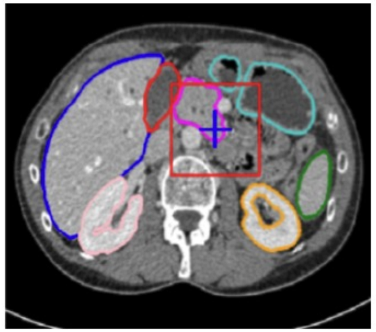
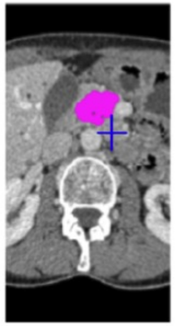
  - Many other losses (IoU, Dice in segmentation), including global constraints (fairness)

Improving robustness in deep learning

1. Uncertainty quantification
2. Direct optimization of rank losses
3. Controlling mistake severity

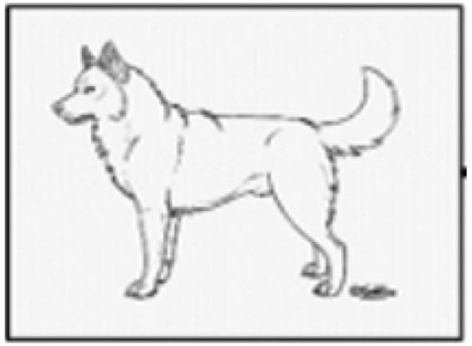
Sources of uncertainty

- Aleatoric uncertainty: data
  - Class confusion, ambiguous data, sensor noise


- Epistemic uncertainty: model
  - Distribution shift in p(x,y), *e.g.* x (snow, image->cartoon), or y (open set, new classes)



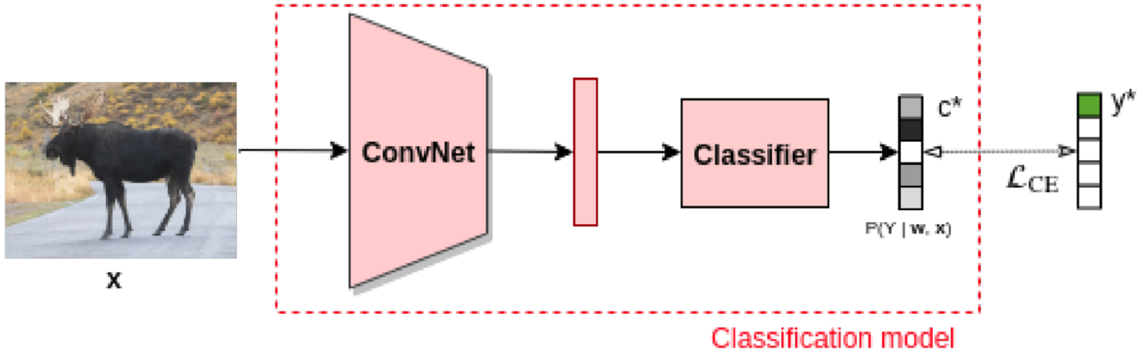



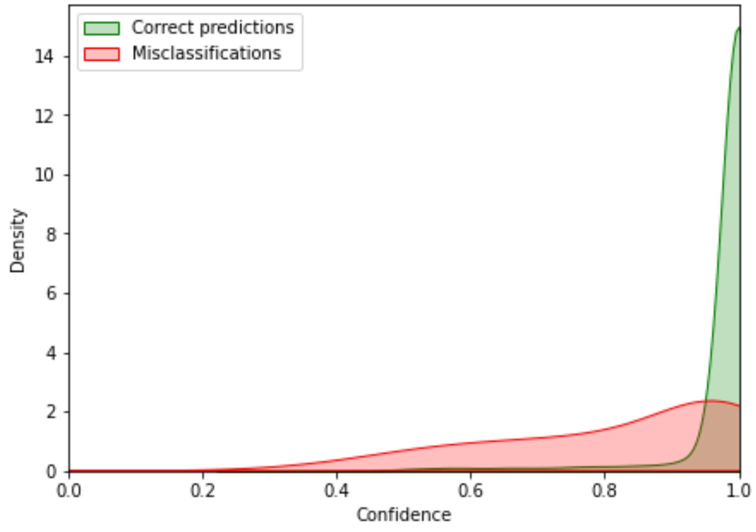
Uncertainty quantification in deep learning

- **Uncertainty for failure prediction [CBT+19]:** correct vs incorrect predictions

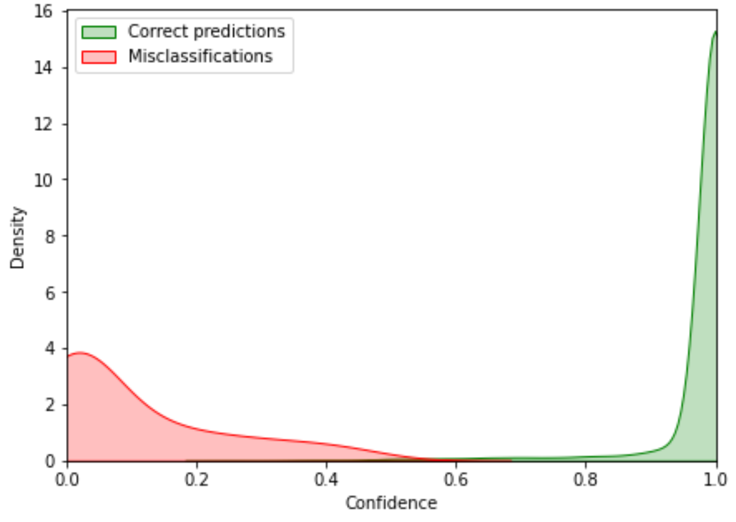
- **<u>Our proposal:</u> True Class probability (TCP)** vs Maximum Class Probability (MCP)

- TCP better than MCP for failure prediction





MCP

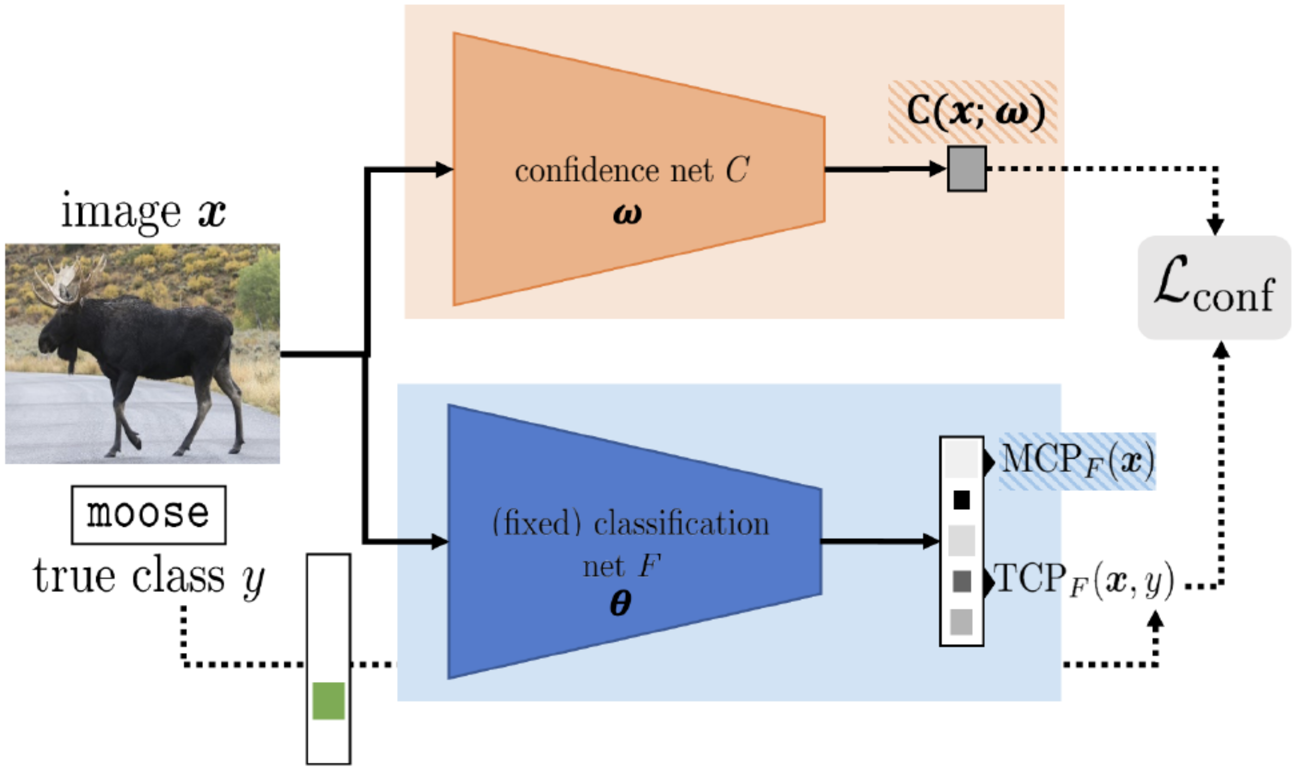


TCP

[CTB+19] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, P. Pérez. Addressing Failure Detection by Learning Model Confidence. NeurIPS 2019.

Uncertainty quantification in deep learning

<u>TCP unknown at test time:</u> learning it! => ConfidNet

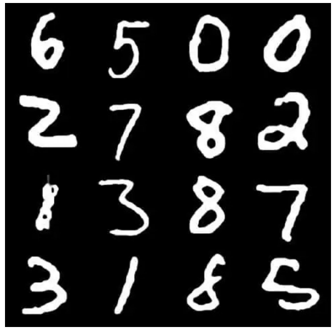


- Pre-trained prediction model (blue)
- Learning to regress TCP with an auxiliary model (orange)

$$\mathcal{L}_{\boldsymbol{conf}}(\theta; \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} (\hat{c}(\mathbf{x}_i, \theta) - c^*(\mathbf{x}_i, y_i^*))^2$$

[CTB+19] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, P. Pérez. Addressing Failure Detection by Learning Model Confidence. NeurIPS 2019.

Results

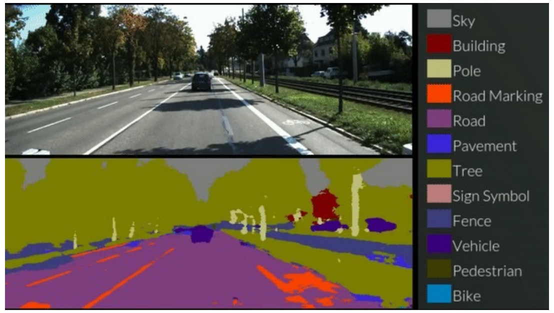


MNIST

SVHN

CIFAR-10

CIFAR-100



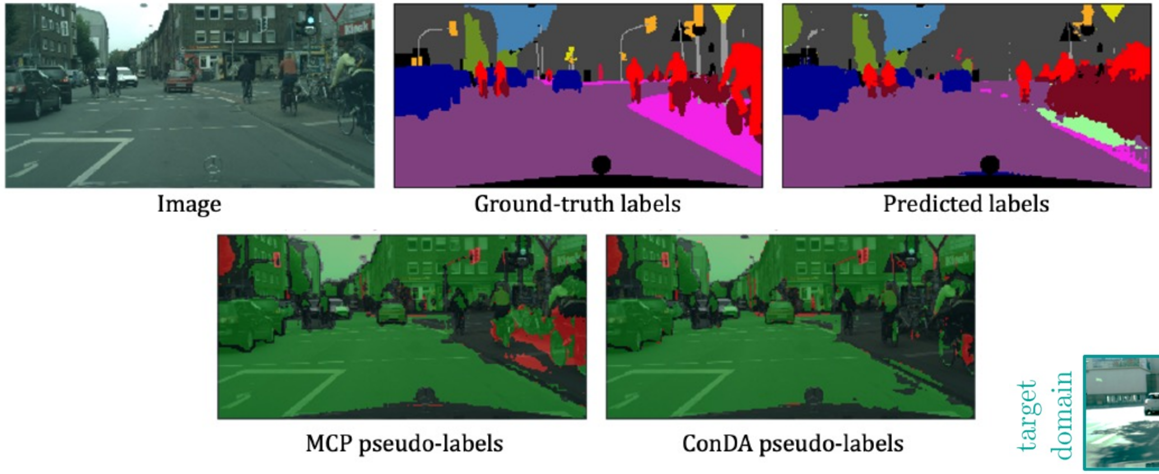
CamVid

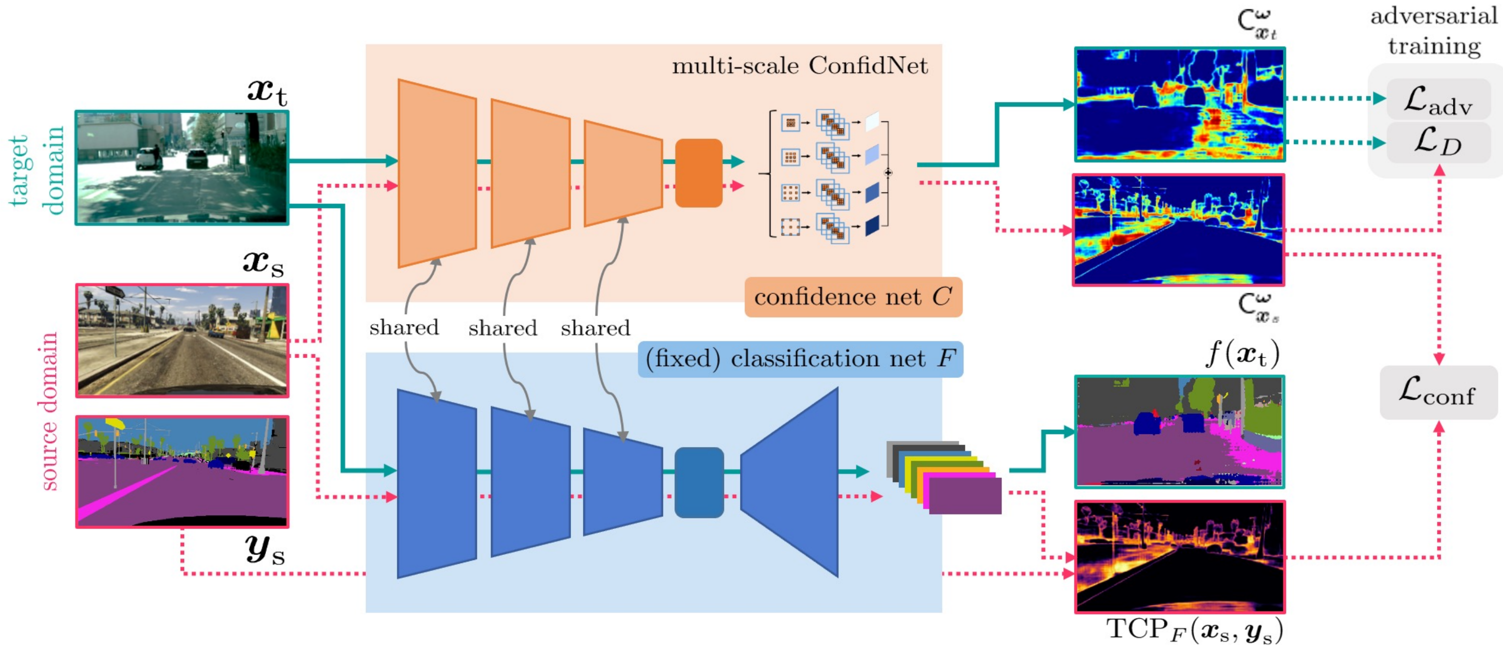
	MNIST		**SVHN**	**CIFAR-10**	**CIFAR-100**	**CamVid**
	MLP	*LeNet-5*	*LeNet-5*	*VGG-16*	*VGG-16*	*SegNet*
MCP [Hendrycks & Gimpel, 2017]	47.3 ± 1.7	36.1 ± 3.6	46.2 ± 0.5	48.4 ± 0.7	71.3 ± 0.4	48.5 ± 0.3
MC Dropout [Gal et Ghahramani, 2015]	41.0 ± 1.2	42.1 ± 5.5	45.2 ± 1.3	48.1 ± 1.0	71.9 ± 0.7	49.4 ± 0.3
TrustScore [Jiang *et al.*, 2019]	52.1 ± 1.8	33.5 ± 3.8	44.8 ± 1.3	41.8 ± 2.0	66.8 ± 0.5	20.4 ± 1.0
ConfidNet	**59.7** ± 1.9	**45.5** ± 3.8	**48.6** ± 1.0	**53.7** ± 0.6	**73.6** ± 0.6	**50.5** ± 0.3

AP errors (%)

Learning confidence for self-labelling

- Extension for domain adaptation [CTS+21]

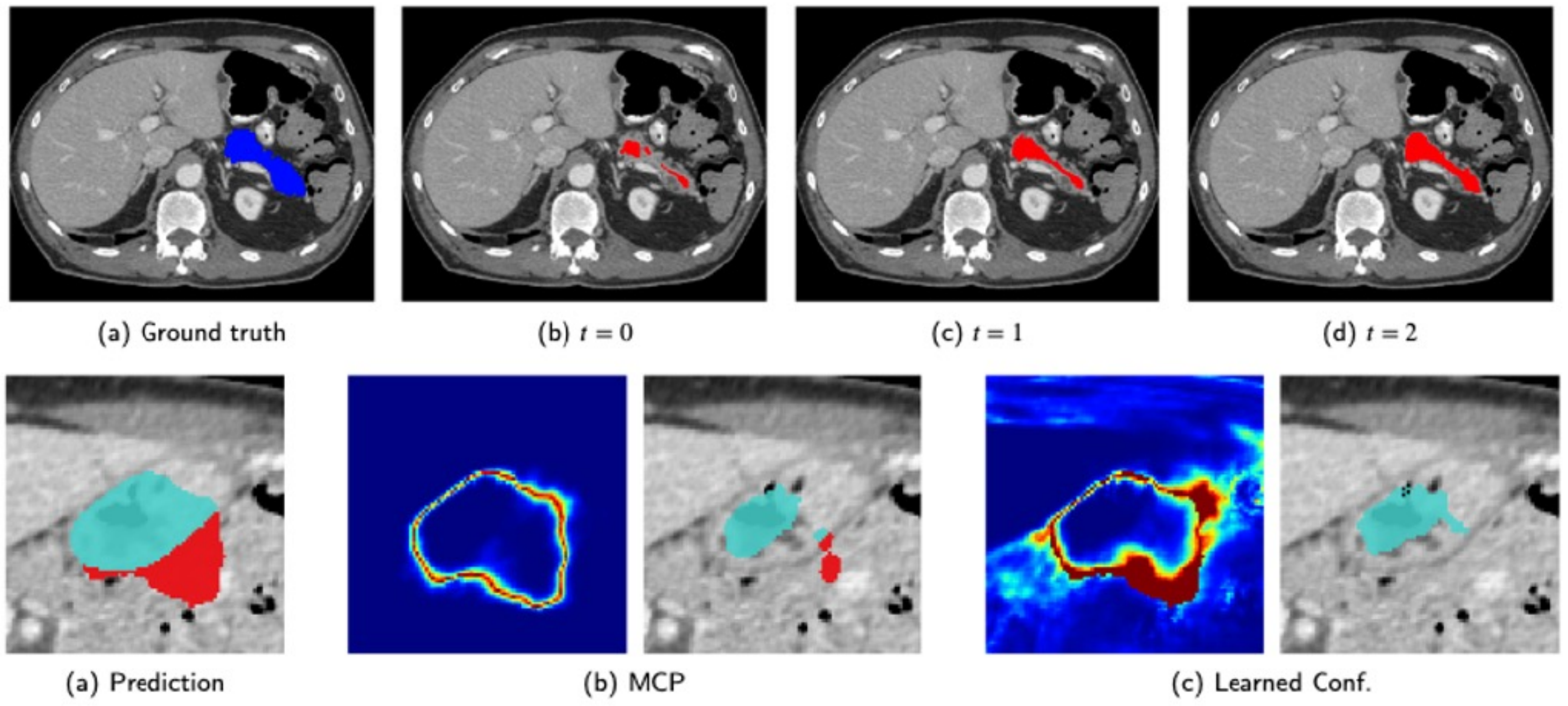




[CTS+21] C. Corbière, N. Thome, A. Saporta, T-H. Vu, M. Cord, P. Pérez. Confidence Estimation via Auxiliary Models. IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI), vol. 44, no. 10, pp. 6043-6055, June 2021.

Learning confidence for self-labelling

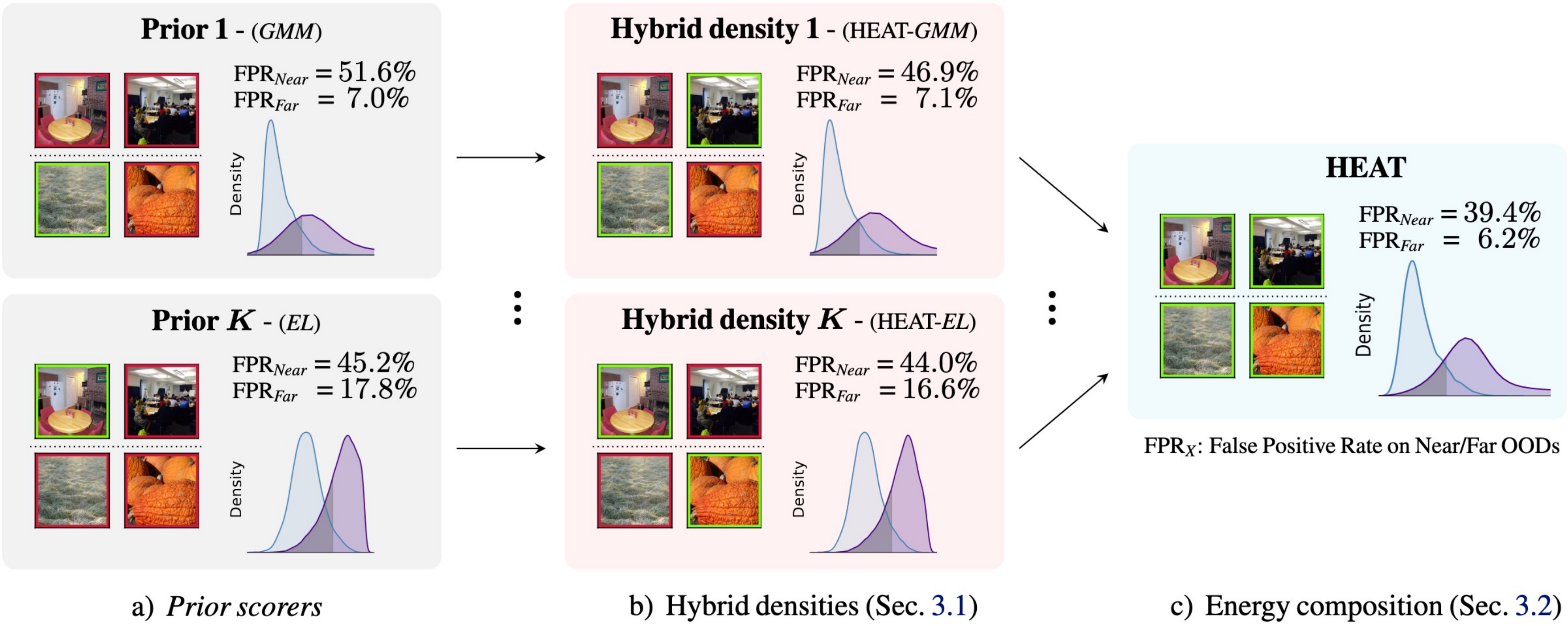
- Extension for Medical image segmentation [PTS21]



[PTS21] O. Petit, N. Thome, L. Soler. 3D Spatial Priors for Semi-Supervised Organ Segmentation with Deep Convolutional Neural Networks. International Journal of Computer Assisted Radiology and Surgery, Springer Verlag, In press, 2021.

Out-Of-Distribution (OOD) detection

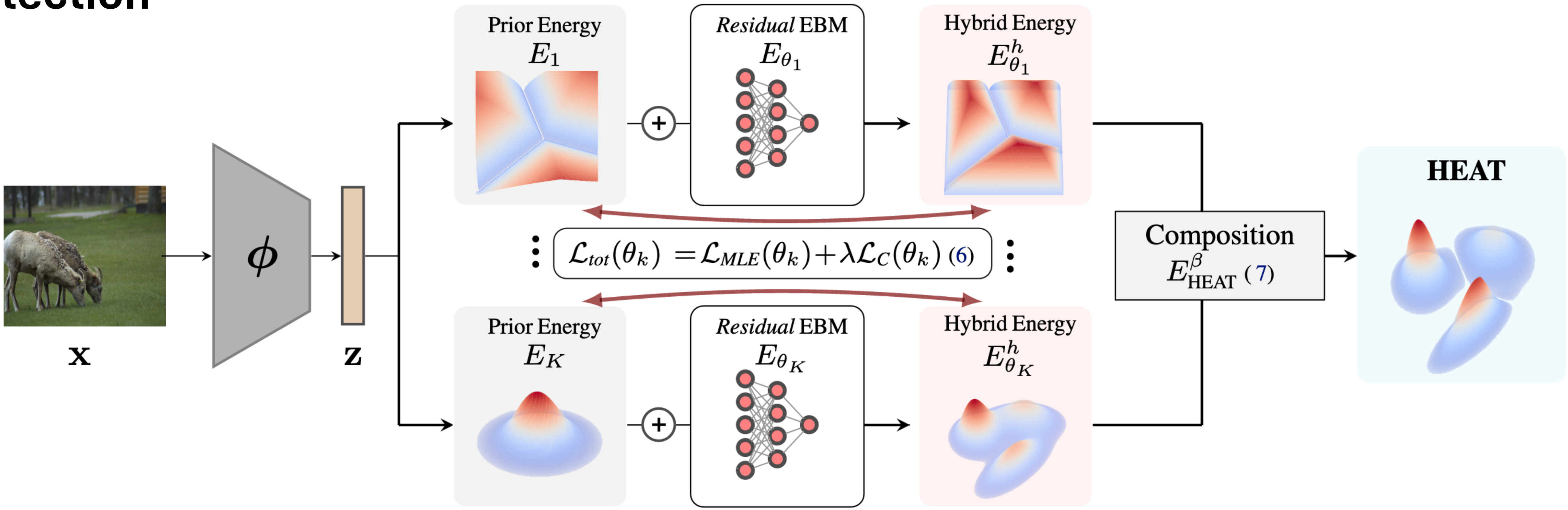
- Post-hoc OOD detection: leveraging any state-of-the-art prediction model
- Accurate OOD detection ⇔ accurate in-distribution (ID) density estimation
  - State-of-the-art ID density estimation: prior densities, *e.g.*, GMM, Energy Logits (EL)



- Prior density: not accurate **=> Energy correction**
- GMM good for far-OOD, EL for near-OOD **=>Energy composition**

OOD detection

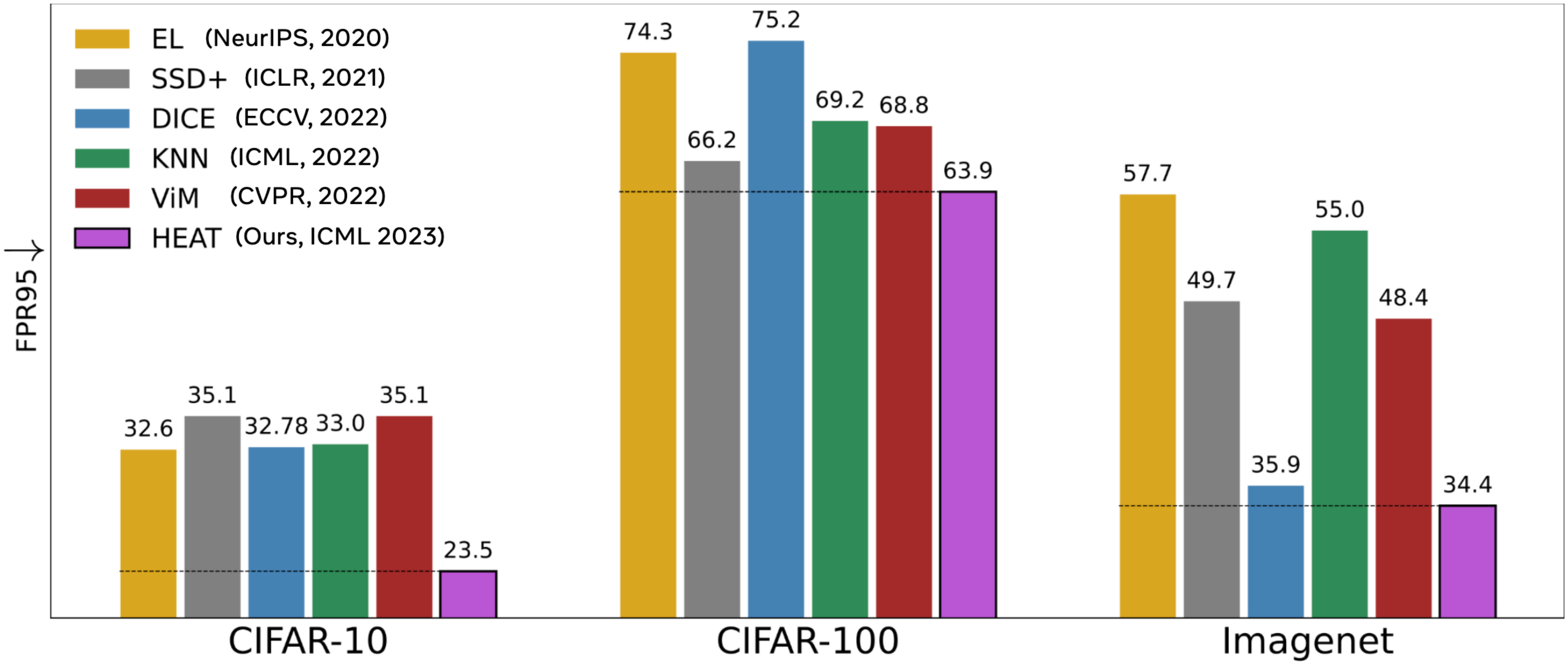
- **HEAT** [LRR+23]: **Hybrid Energy Based Model (EBM) in the feature space for OOD detection**



- **Energy-based correction** of prior energy terms, *e.g.* Gaussians
- **Energy composition** of several terms (Gaussian, Energy Logits, std for style)

[LRR+23] M. Lafon, E. Ramzi, C. Rambour, N. Thome. Addressing Failure Detection by Learning Model Confidence. ICML 2023.

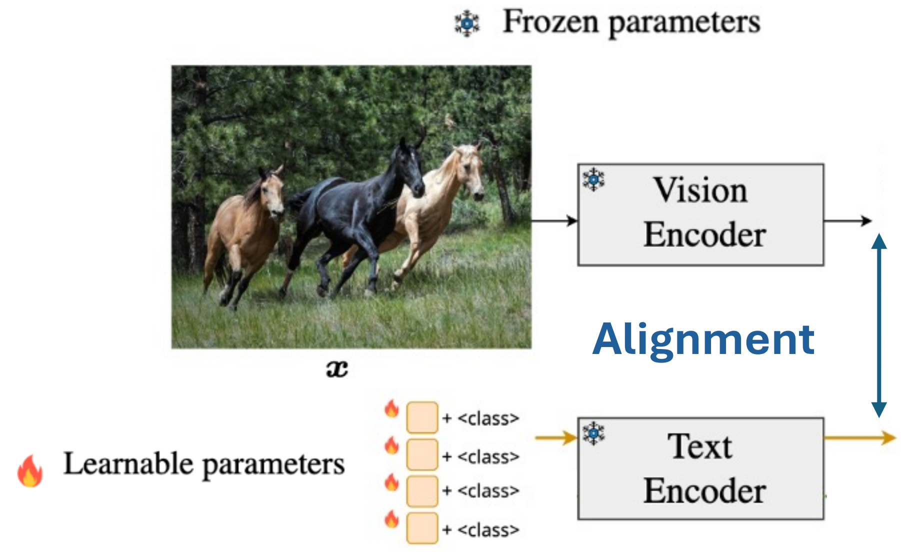
Results



EL (NeurIPS, 2020)
SSD+ (ICLR, 2021)
DICE (ECCV, 2022)
KNN (ICML, 2022)
ViM (CVPR, 2022)
HEAT (Ours, ICML 2023)
FPR95
32.6
35.1
32.78
33.0
35.1
23.5
74.3
66.2
75.2
69.2
68.8
63.9
57.7
49.7
35.9
55.0
48.4
34.4
CIFAR-10
CIFAR-100
Imagenet

Robustness in prompt learning

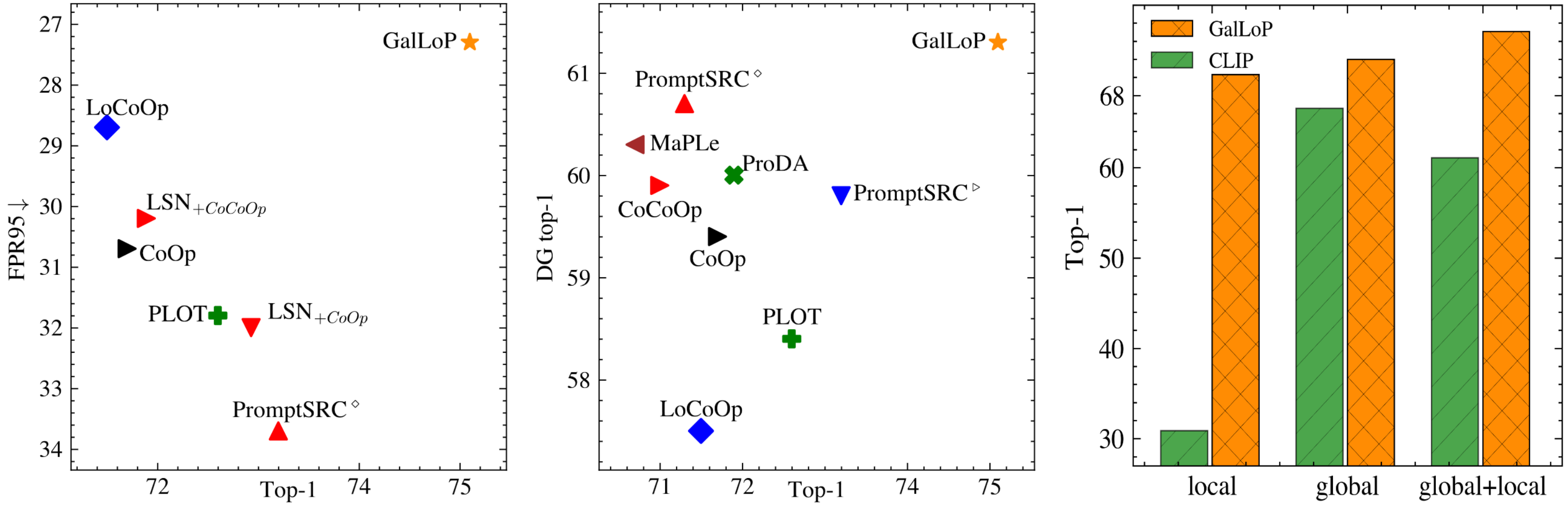
- Learning prompts from frozen vision-language models (VLMs), e.g., CLIP



<u>State-of-the-art methods' shortcomings:</u>

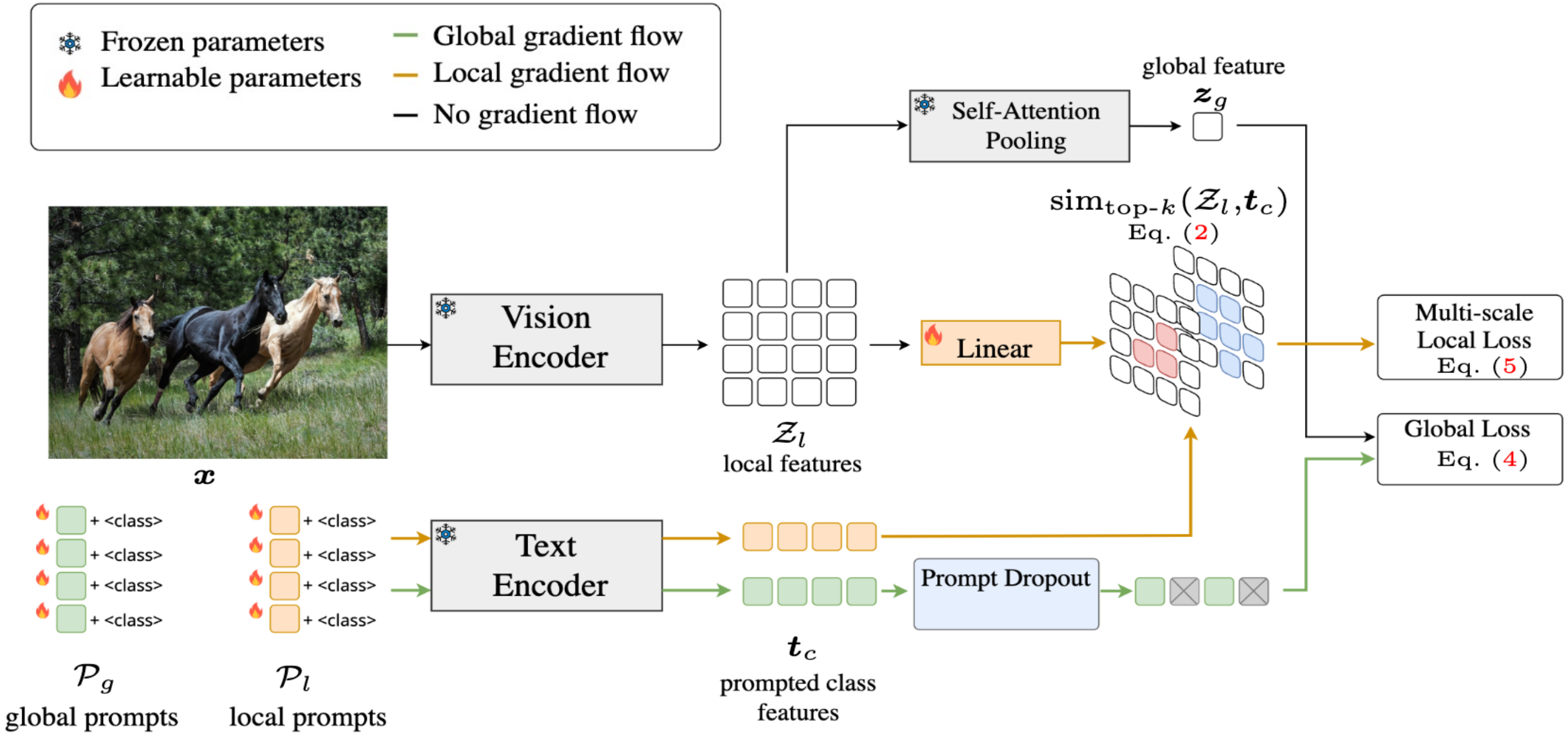
- Local prompt: aligning local features
- Accuracy and robustness, e.g., OOD detection, domain generalization

<u>Our method:</u> Global and Local Prompts for VLMs **GalLoP**



(a) OOD detection (b) Domain Generalization (c) Global-local

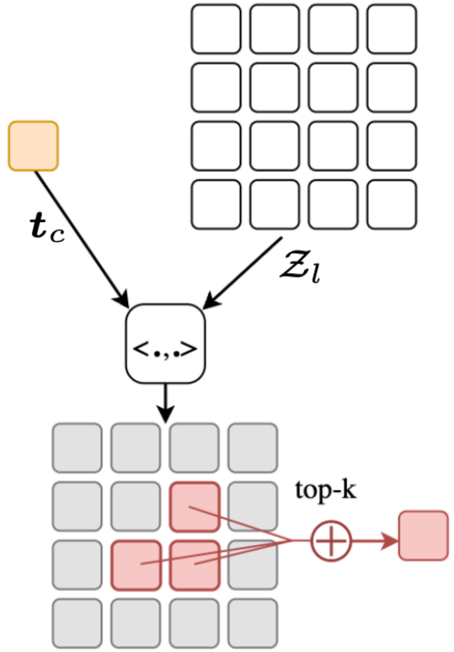
Learning Global and Local Prompts for VLMs (GalLoP)



- **<u>Local prompts:</u>** sparse local matching + linear alignment
- **<u>Global prompts diversity:</u>** prompt dropout = multiscale

Local prompts with GalLoP

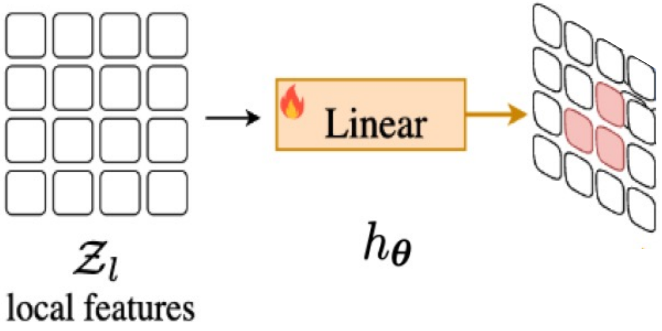
- **Sparse local alignment**



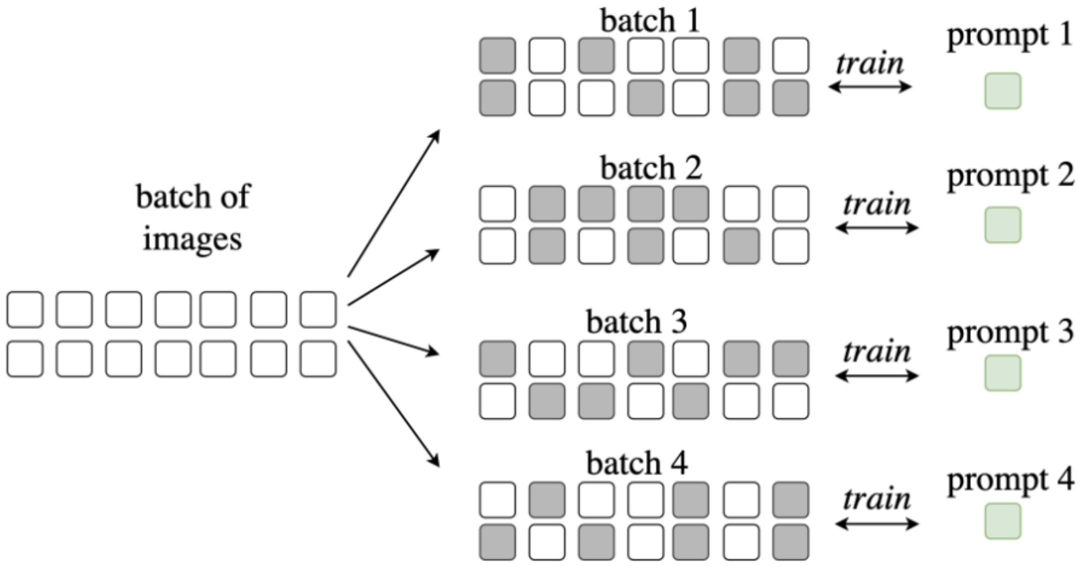
$$\mathrm{sim}_{\text{top-}k}(\mathcal{Z}_l, \boldsymbol{t}_c) := \frac{1}{k}\sum_{i=1}^{L} \mathbb{1}_{\text{top-}k}(i)\cdot\langle \boldsymbol{z}_i^l, \boldsymbol{t}_c\rangle$$

$$\text{where} \qquad \mathbb{1}_{\text{top-}k}(i) = \begin{cases} 1 & \text{if} \quad \mathrm{rank}_i(\langle \boldsymbol{z}_i^l, \boldsymbol{t}_c\rangle) \leq k, \\ 0 & \text{otherwise.} \end{cases}$$

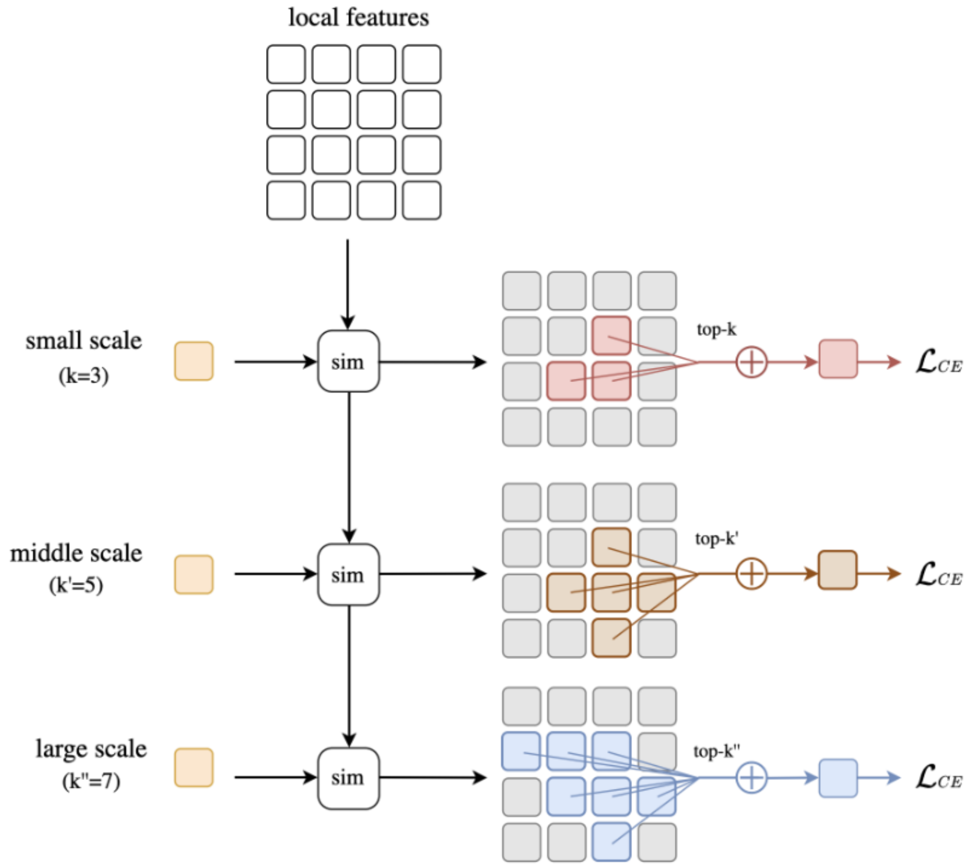
- **Learned alignment**



Prompts diversity with GalLoP

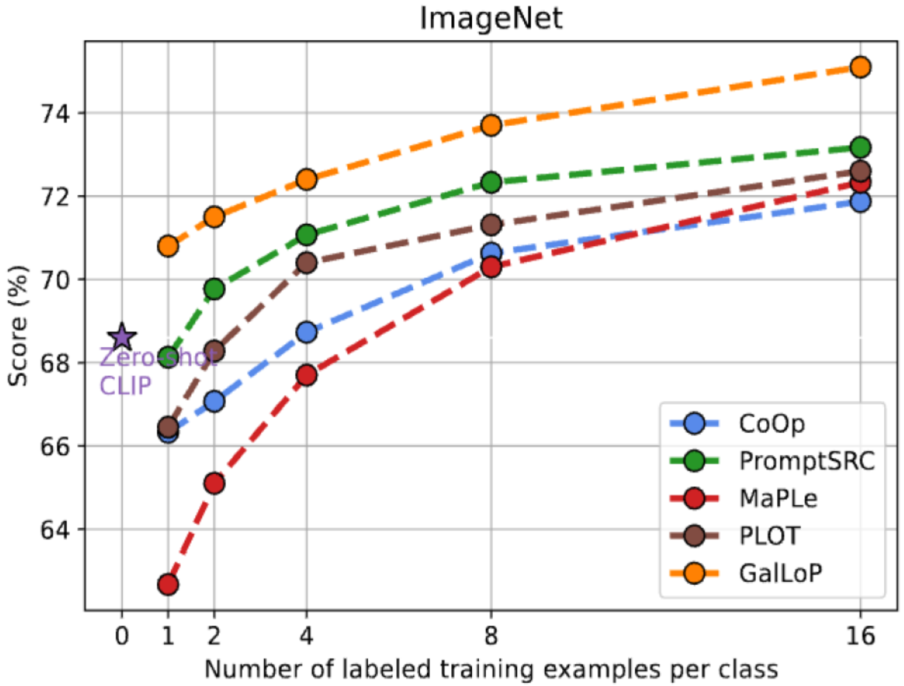


(a) Prompt dropout



(b) Multiscale loss

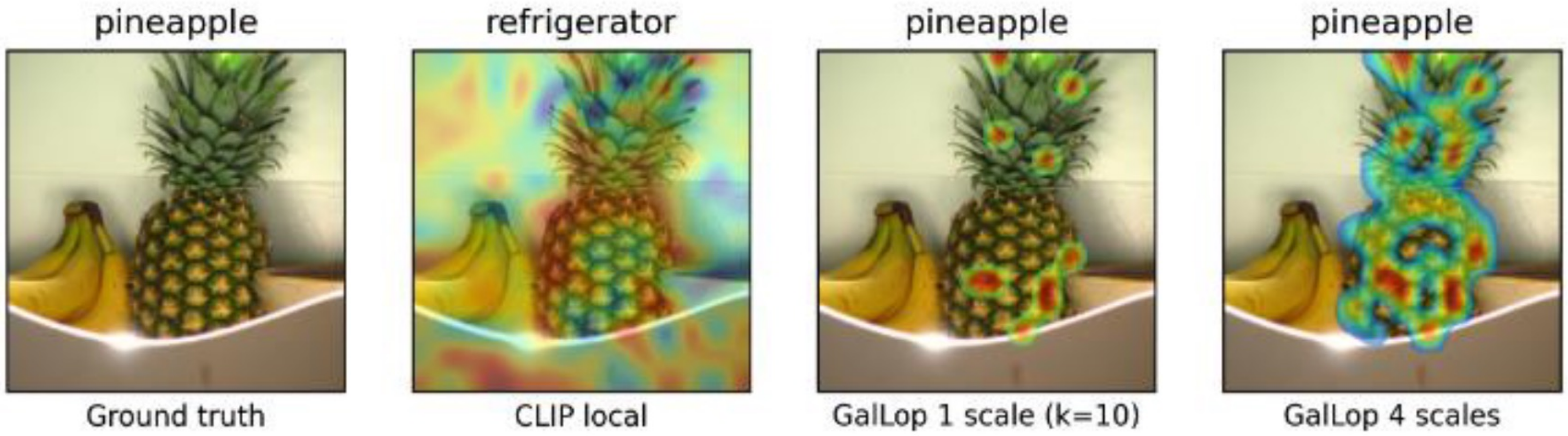
GalLoP Results



	Top-1	DG	FPR95↓	AUC
$\mathrm{CLIP}_{\mathrm{Global}}$	66.6	57.2	42.8	90.8
$\mathrm{CLIP}_{\mathrm{Local}}$	12.5	9.49	73.3	73.7
$\mathrm{CLIP}_{\mathrm{GL}}$	61.1	49.3	35.5	90.8
$\mathrm{CoOp}_{\mathrm{Global}}$	71.4	59.2	39.1	91.1
$\mathrm{CoOp}_{\mathrm{Local}}$	41.2	30.1	65.2	78.3
$\mathrm{CoOp}_{\mathrm{GL}}$	69.5	55.6	33.7	90.5
$\mathrm{GalLoP}_{\mathrm{Global}}$	72.0	60.4	37.0	91.7
$\mathrm{GalLoP}_{\mathrm{Local}}$	70.9	54.1	36.0	90.1
GalLoP	**75.1**	**61.3**	**27.3**	**93.2**

Improved local prompts => effective combination with global prompts

Better accuracy and robustness (OOD detection and domain generalization)



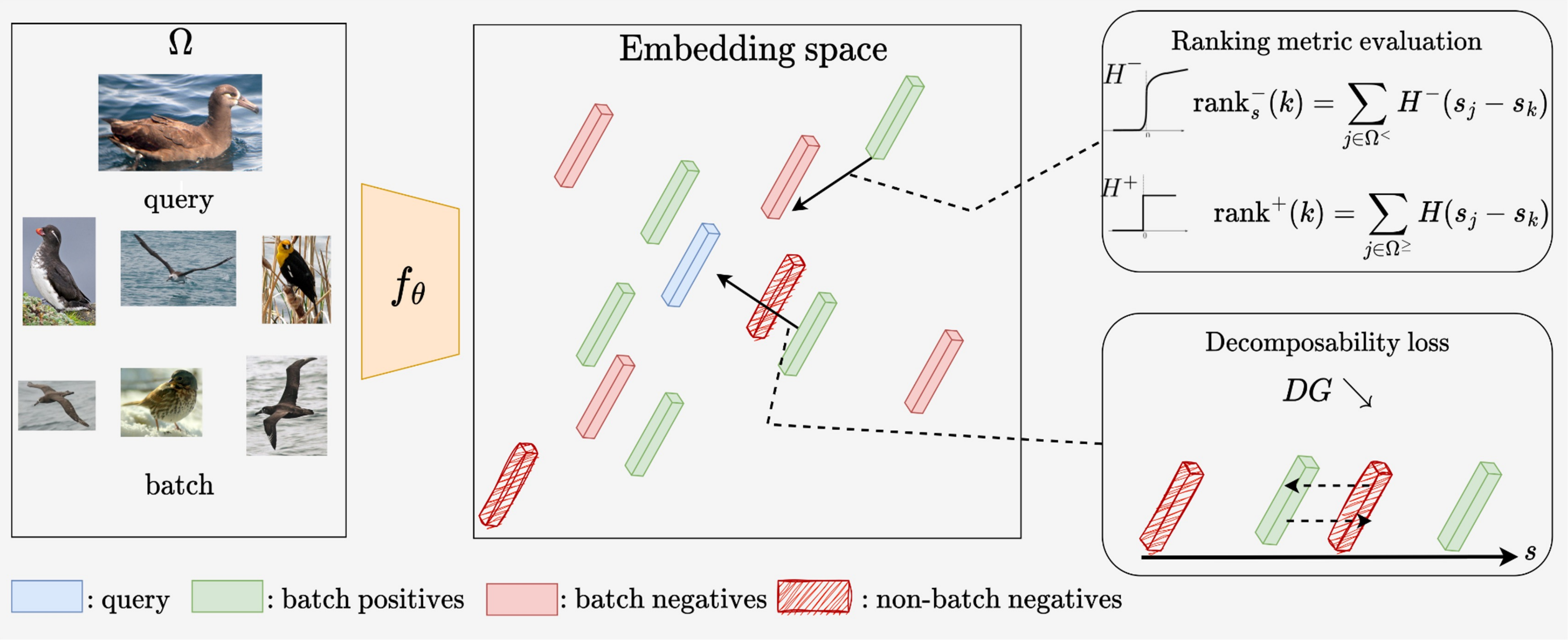
CLIP *vs* GalLop with local features

Robustness: recent contributions

1. Uncertainty quantification
2. Direct optimization of rank losses
3. Robustness

Direct optimization of rank losses for image retrieval

1. Theoretically sound surrogates for non-differentiable rank losses, *e.g.*, Average Precision (AP)
2. Reducing the decomposability gap

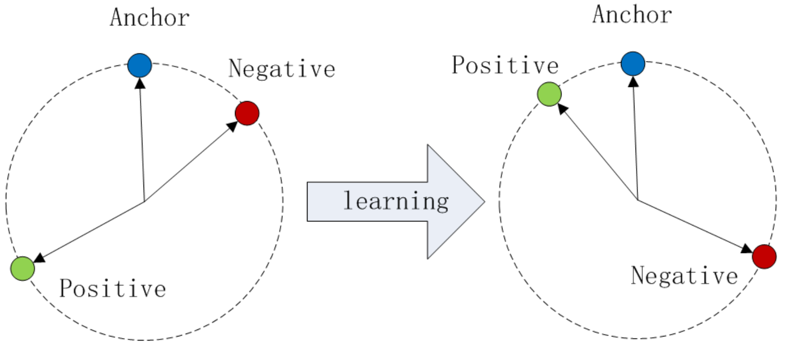


$\mathcal{L}_{\mathrm{SupAP}}(\boldsymbol{\theta})$

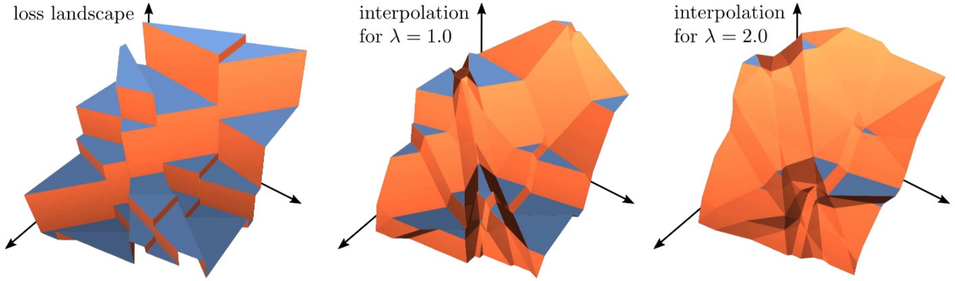
$\mathcal{L}_{\mathrm{DG}}(\boldsymbol{\theta})$

$$\mathcal{L}_{\mathrm{ROADMAP}}(\boldsymbol{\theta}) = (1-\lambda)\cdot\mathcal{L}_{\mathrm{SupAP}}(\boldsymbol{\theta}) + \lambda\cdot\mathcal{L}_{\mathrm{DG}}(\boldsymbol{\theta})$$

Image retrieval: non-smooth metrics & losses

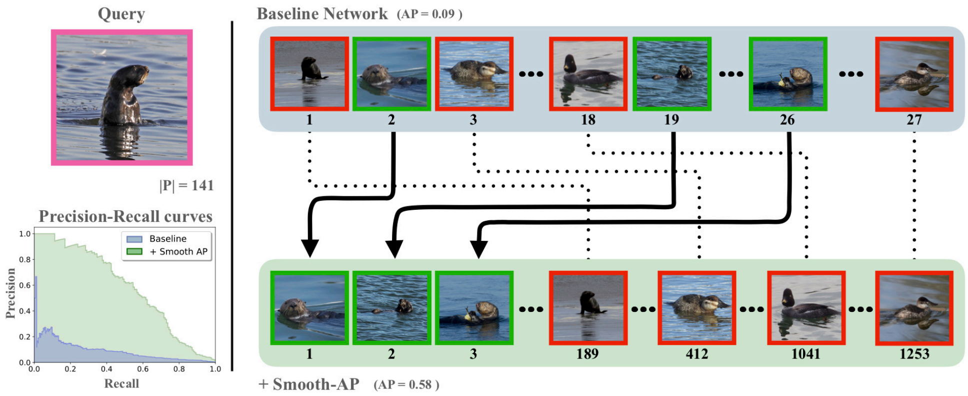


- Standard losses, e.g., : triplet loss, NSM [A]
  ⊖ Coarse upper-bounds, not well-aligned with metrics: supports bottom *vs.* top of the ranking



- Upper bounds: structural SVMs, Blackbox optim [B]
  ⊕ General methods, theoretical guarantees
  ⊖ Coarse upper bounds

- Rank approximation: binning approches, smoothAP [C,D]
  ⊕ Tighter approximations
  ⊖ No theoretical guarantees

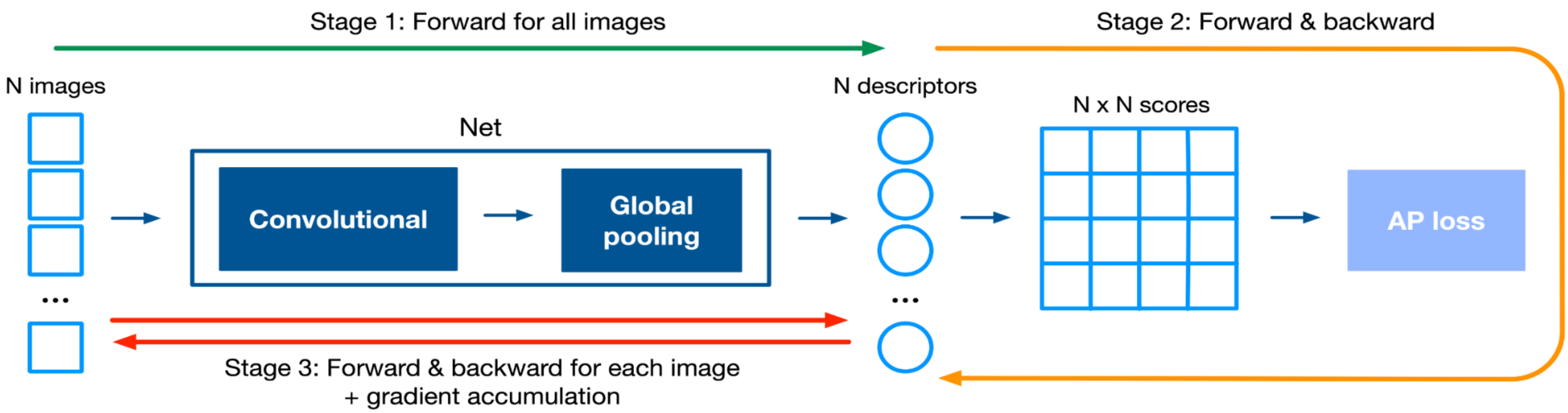


[A] A. Zhai, and H.Y Wu. Classification is a strong baseline for deep metric learning. BMVC 2018
[B] M. Rolínek, V. Musil, A. Paulus, M. Vlastelica, C. Michaelis, G. Martius. Optimizing rank-based metrics with blackbox differentiation. CVPR 2020
[C] A. Brown, W. Xie, V. Kalogeiton, A. Zisserman. Smooth-ap: Smoothing the path towards large-scale image retrieval. ECCV 2020
[D] Y. Patel, G. Tolias, and J. Matas, “Recall@ k surrogate loss with large batches and similarity mixup,” in CVPR, 2022

Image retrieval: addressing non-decomposability

Fewer works, brute-force approaches

- Sampling informative batches or constraints in batch
- Storing the datasets, e.g., x-batch memory [D]: **increased in memory**
- Large batches + 2-step approach for AP and back-prop [E] **: increase in training time**

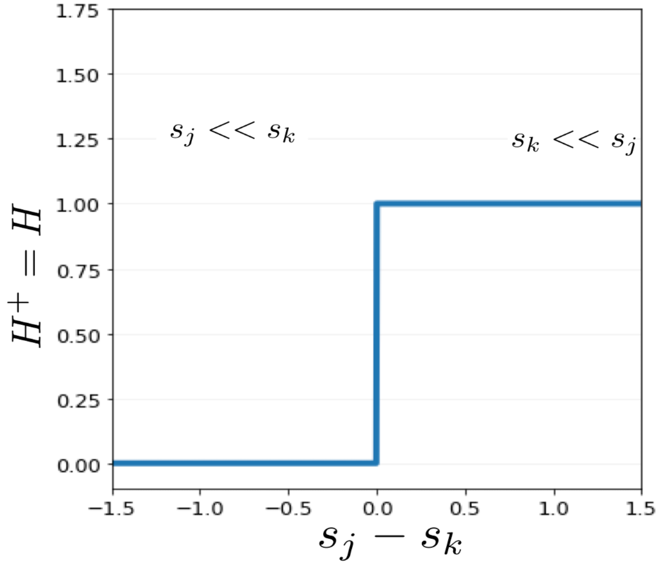


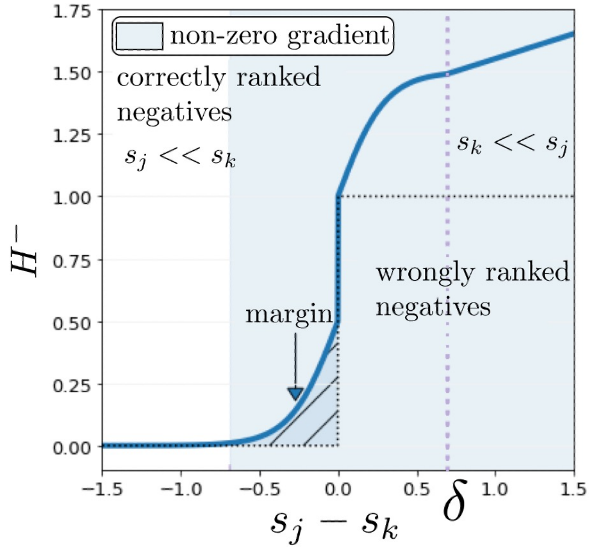
[D] X. Wang, H. Zhang, W. Huang, and M. R. Scott, "Cross-batch memory for embedding learning," in CVPR, 2020
[E] J. Revaud, J. Almazan, R. S. Rezende, and C. R. d. Souza, "Learning with average precision: Training image retrieval with a listwise loss," in ICCV, 2019.

Robust and decomposable AP (ROADMAP)

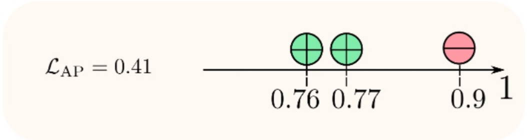
$$\text{AP} = \frac{1}{|\Omega^+|} \sum_{k \in \Omega^+} \frac{\text{rank}^+(k)}{\text{rank}(k)} = \frac{1}{|\Omega^+|} \sum_{k \in \Omega^+} \frac{\text{rank}^+(k)}{\text{rank}^+(k) + \text{rank}^-(k)}$$

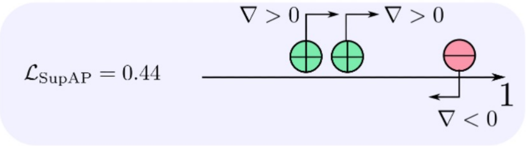
$$\text{rank}(k) = 1 + \sum_{j \in \Omega} H(s_j - s_k)$$



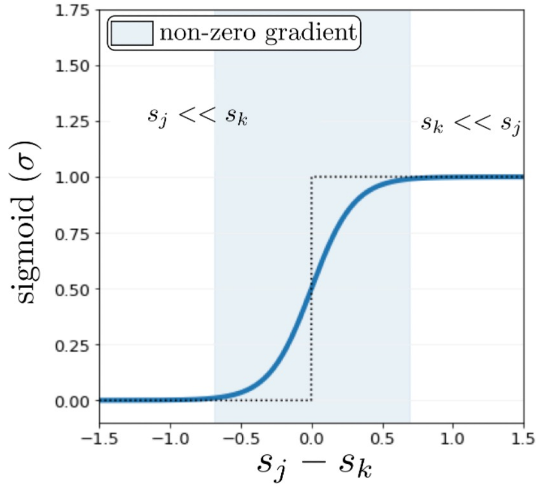


- Optimizing rank^- → smooth approximation & **upper bound of AP**
- Not optimizing rank^+ → well-behaved gradients.



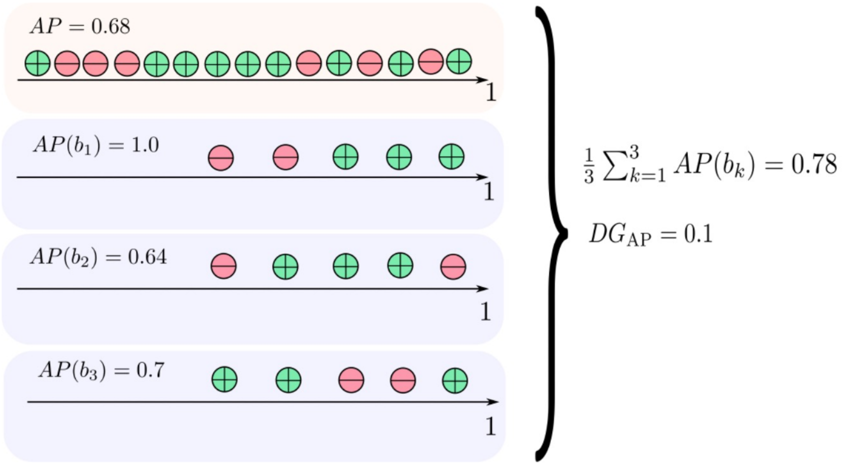


Better than



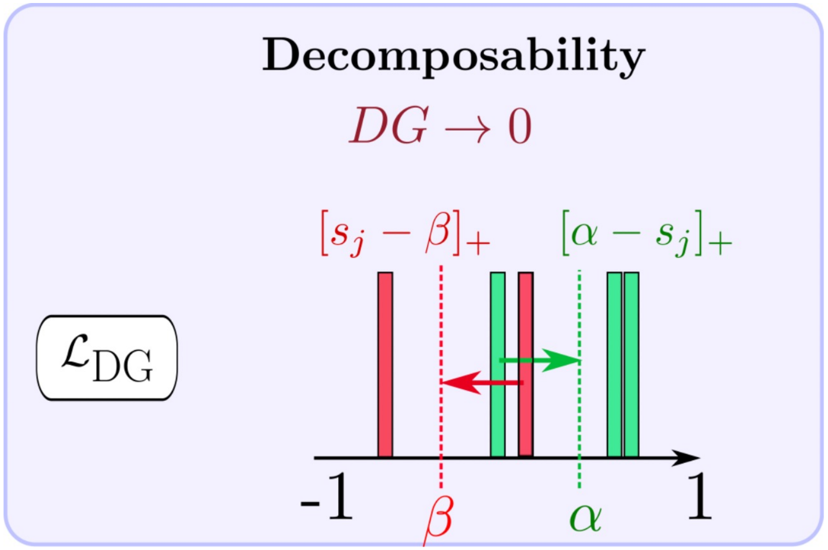
$$\boxed{\mathcal{L}_{\text{SupAP}} = 1 - \frac{1}{|\Omega^+|} \sum_{k \in \Omega^+} \frac{\text{rank}^+(k)}{\text{rank}^+(k) + \text{rank}_s^-(k)}}$$

Improving decomposability



Decomposability gap:

$$\mathrm{DG}(\Omega) = \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} \mathcal{M}(b) - \mathcal{M}(\Omega)$$

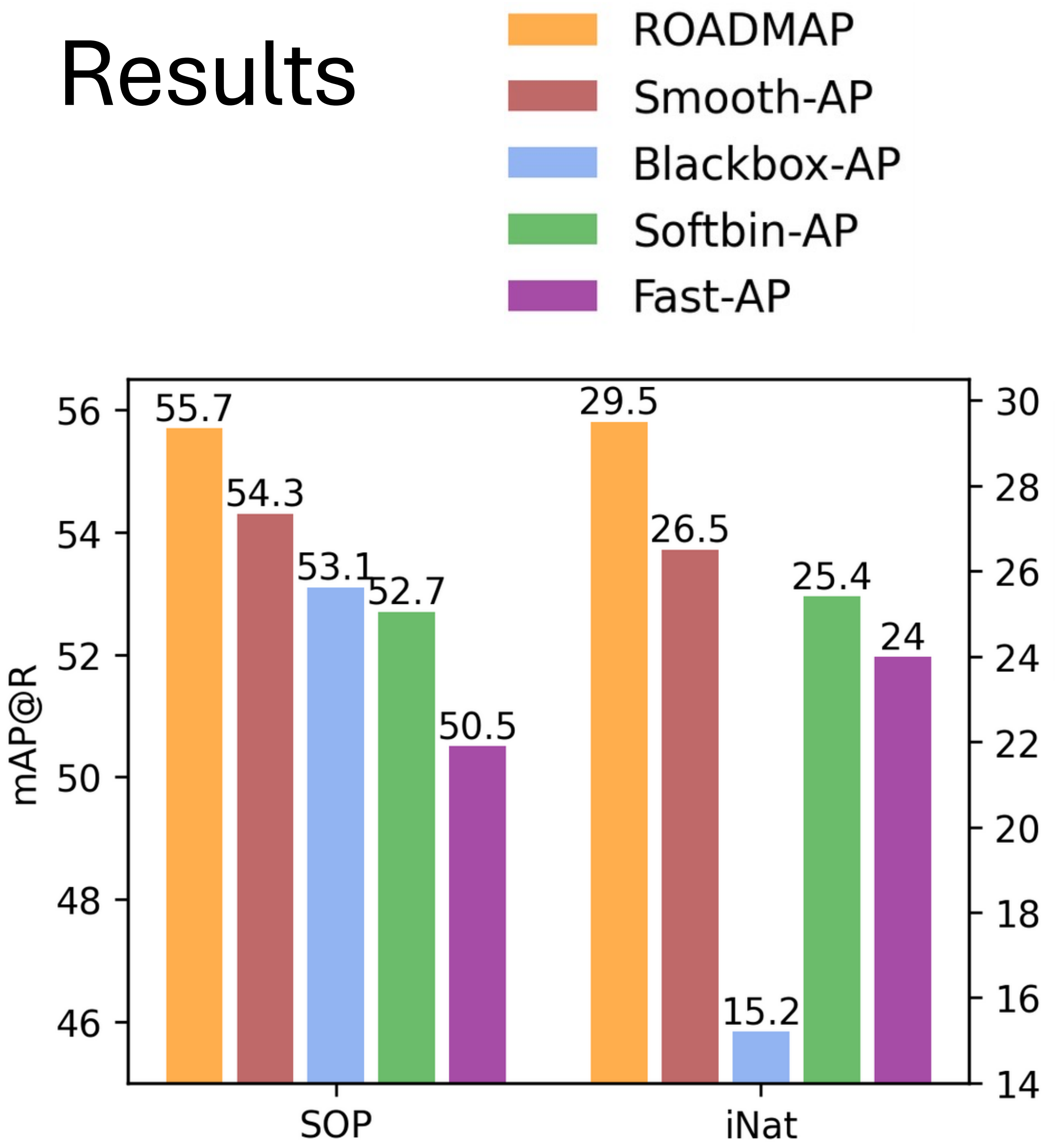


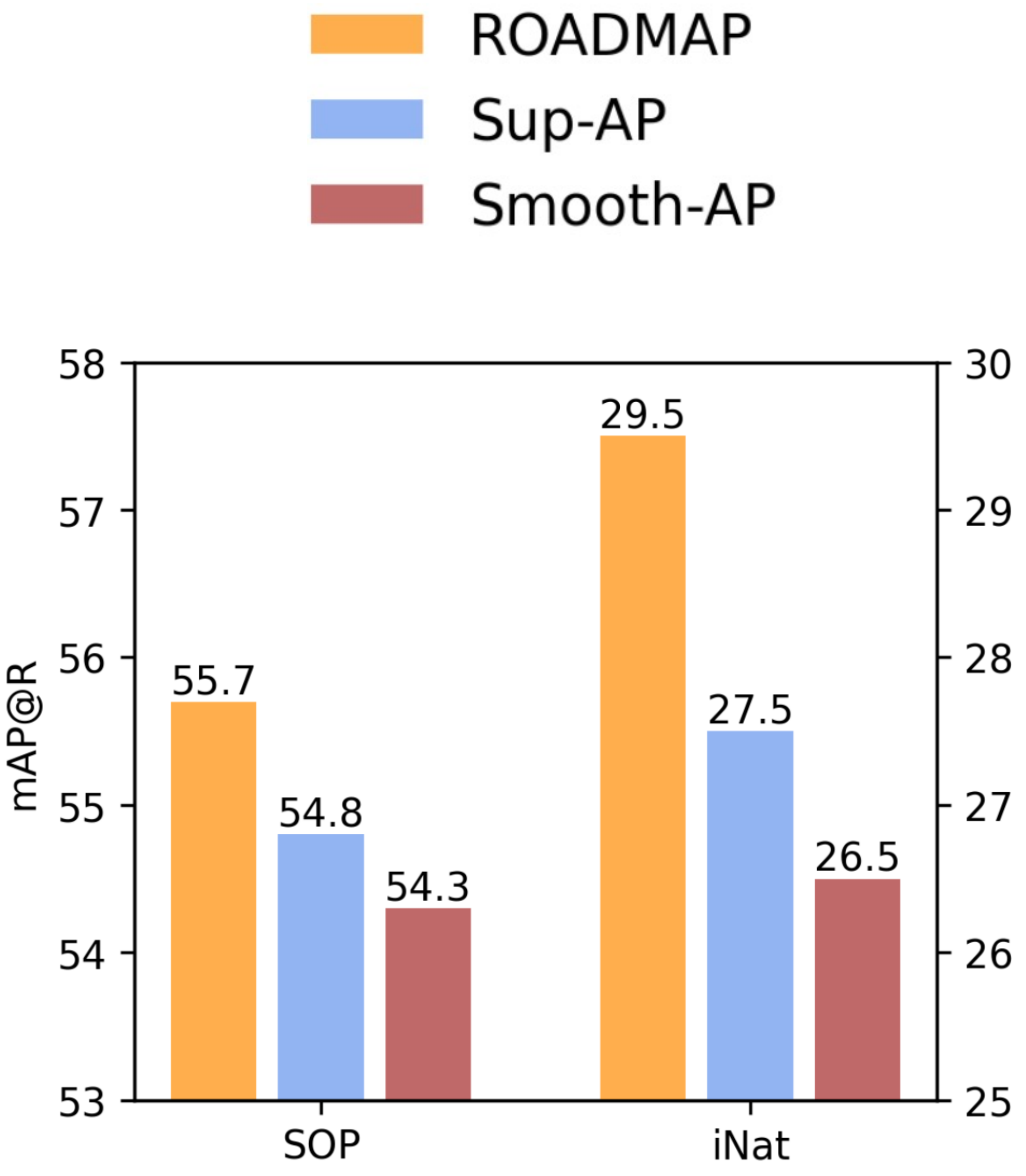
$$\mathcal{L}_{DG}(\boldsymbol{\theta}) = \frac{1}{|\Omega^+|}\sum_{\boldsymbol{x}_j\in\Omega^+}[\alpha - s_j]_+ + \frac{1}{|\Omega^-|}\sum_{\boldsymbol{x}_j\in\Omega^-}[s_j - \beta]_+$$

- Calibrates scores across batches
  - Positive scores $\geq \alpha$
  - Negative scores < β
- **Proof: $\mathcal{L}_{DG}$ reduces decomposability gap**

$$\mathcal{L}_{\mathrm{ROADMAP}}(\boldsymbol{\theta}) = (1-\lambda)\cdot\mathcal{L}_{\mathrm{SupAP}}(\boldsymbol{\theta}) + \lambda\cdot\mathcal{L}_{\mathrm{DG}}(\boldsymbol{\theta})$$

Results


ROADMAP
Smooth-AP
Blackbox-AP
Softbin-AP
Fast-AP
mAP@R
55.7
54.3
53.1
52.7
50.5
29.5
26.5
15.2
25.4
24
SOP
iNat

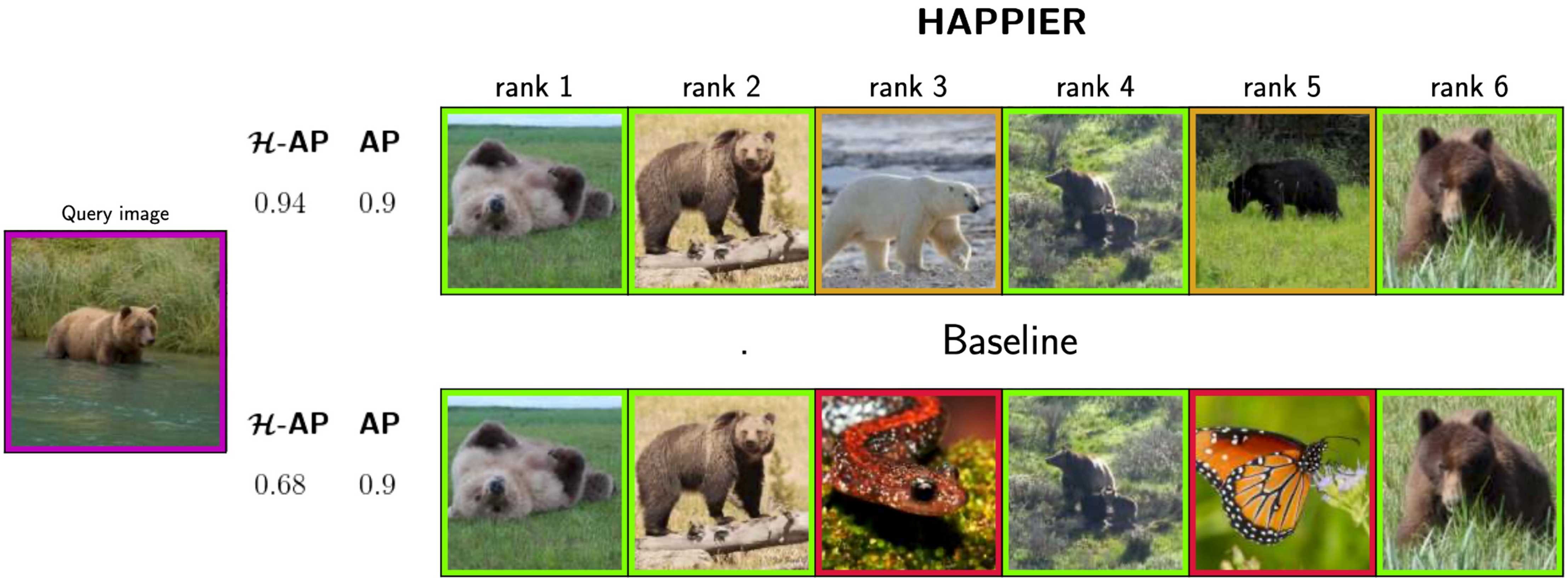

ROADMAP
Sup-AP
Smooth-AP
mAP@R
55.7
54.8
54.3
29.5
27.5
26.5
SOP
iNat

Robustness: recent contributions

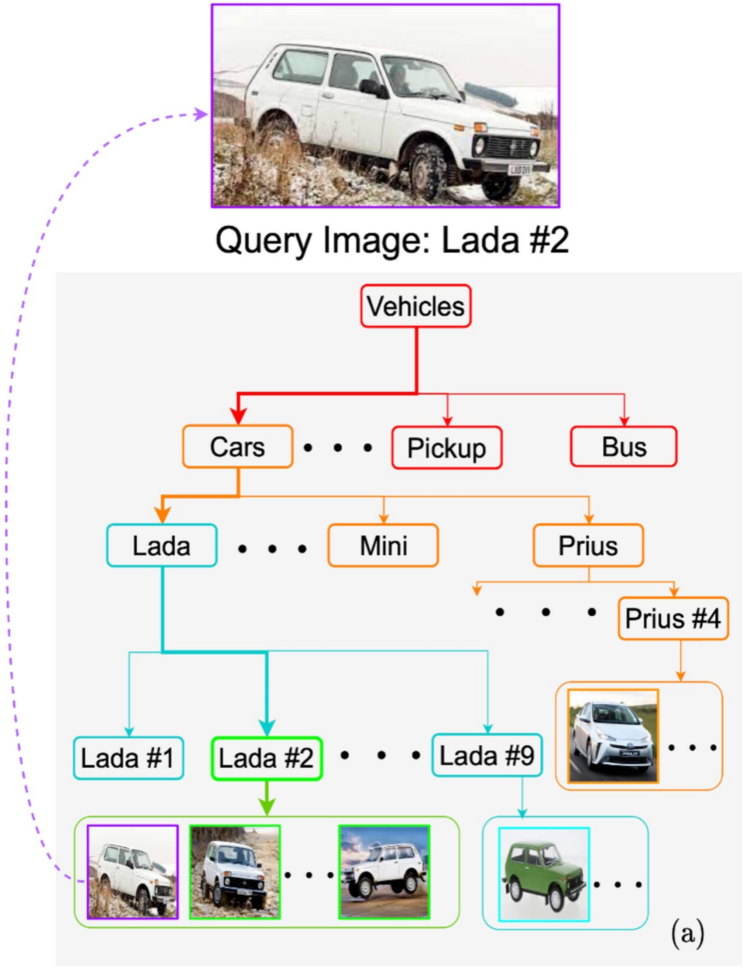
1. Uncertainty quantification
2. Direct optimization of rank losses
3. Controlling mistake severity

Hierarchical Image Retrieval for Robust Ranking

- Binary image retrieval → do not take into account mistake severity
- HAPPIER: Hierarchical Average Precision training for Pertinent Image Retrieval
  Extending AP to graded setting to take importance of errors into account



Relevance function: graded similarities



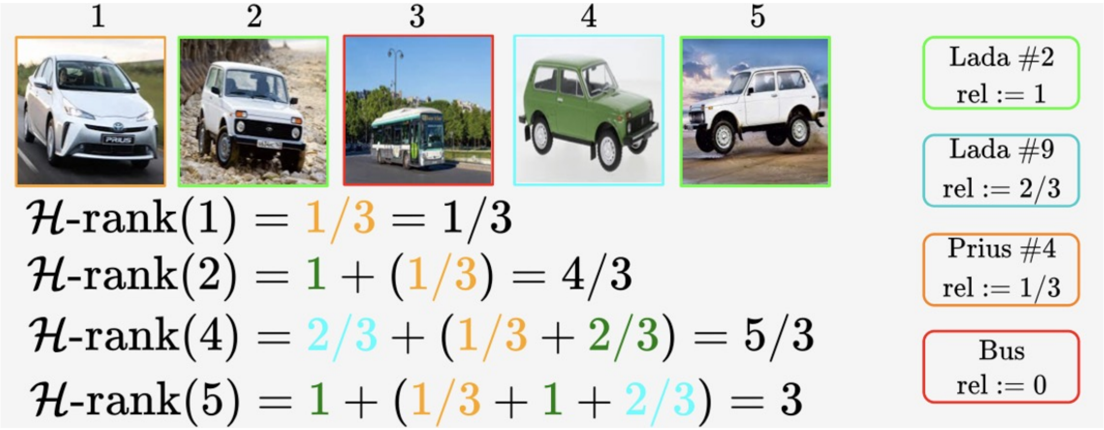
- **Relations between categories → proxy for mistake severity.**
- **Decreasing function of the distance** in the hierarchical tree.

$$\mathrm{rel}(k) = \frac{l/L}{|\Omega^{(l)}|}$$

- l: level of the closest ancestor in the tree.
- L total number of levels.

Hierarchical average precision ($\mathcal{H}$-AP)

$$\mathcal{H}\text{-rank}(k) = \text{rel}(k) + \sum_{j\in\Omega^+} \min(\text{rel}(k), \text{rel}(j)) \cdot H(s_j - s_k)$$

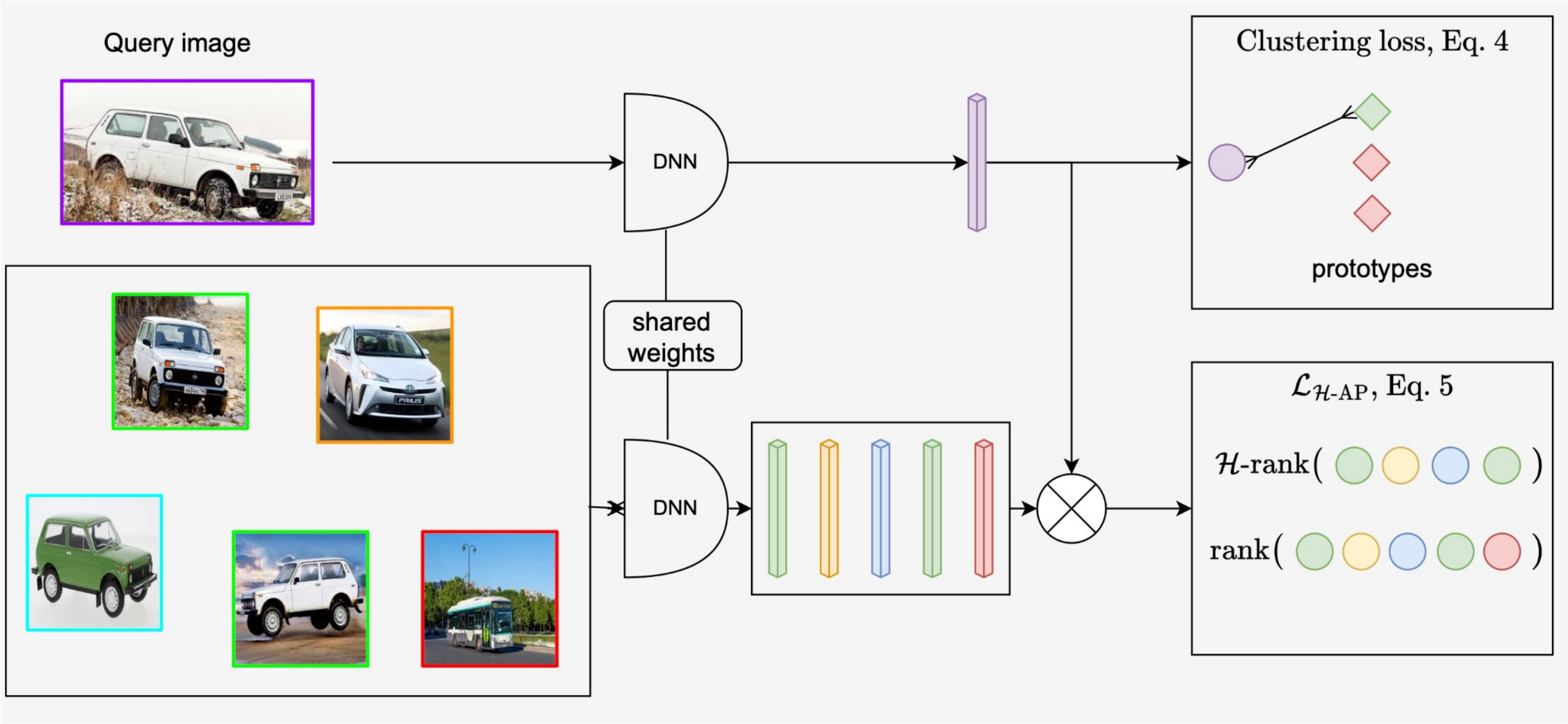


- Errors in ranking → **weighted by relevance**
- Correct $\mathcal{H}$-rank → decreasing order of relevance

$$\mathcal{H}\text{-AP} = \frac{1}{\sum_{k\in\Omega^+} \text{rel}(k)} \sum_{k\in\Omega^+} \frac{\mathcal{H}\text{-rank}(k)}{\text{rank}(k)}$$

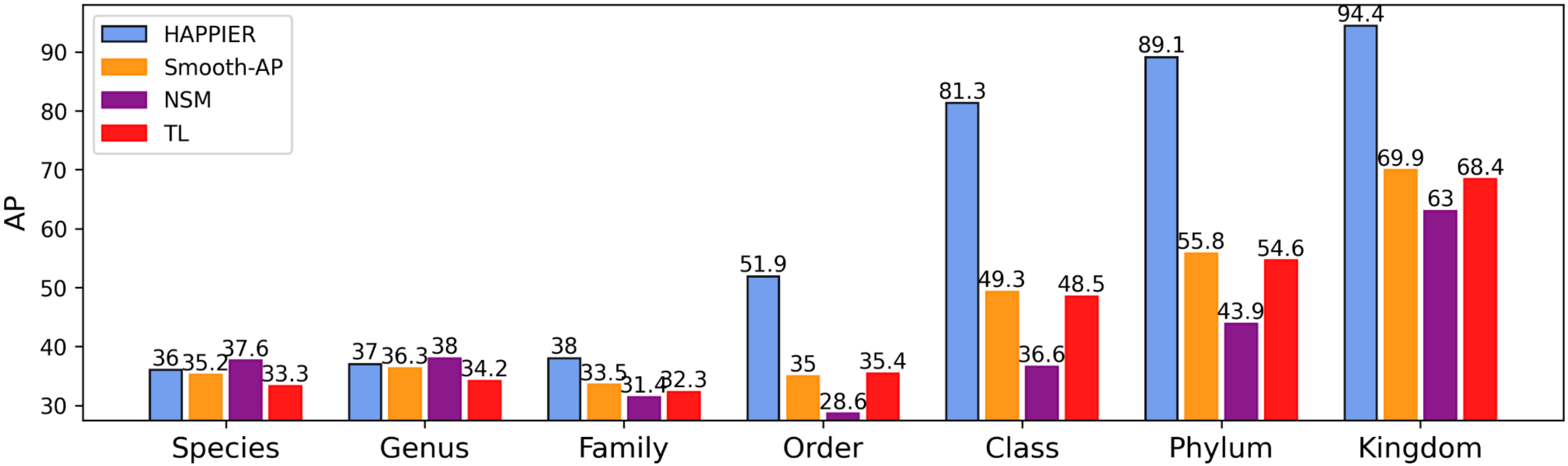
- Consistent generalization of AP.
- Flexible wrt. the relevance

HAPPIER training



$$\mathcal{L}_{\text{Sup-}\mathcal{H}\text{-AP}}(\theta) = 1 - \frac{1}{\sum_{k \in \Omega^+} \text{rel}(k)} \sum_{k \in \Omega^+} \frac{\mathcal{H}\text{-rank}(k)}{\text{rank}^+(k) + \text{rank}_s^-(k)}$$

Results



- On par for fine-grained retrieval ("Species")
- **Large gains** on other hierarchical levels from "Family"

$\mathcal{H}$-GLDv2: a hierarchical landmark dataset

GLDv2 → **large scale** landmarks retrieval dataset [F]

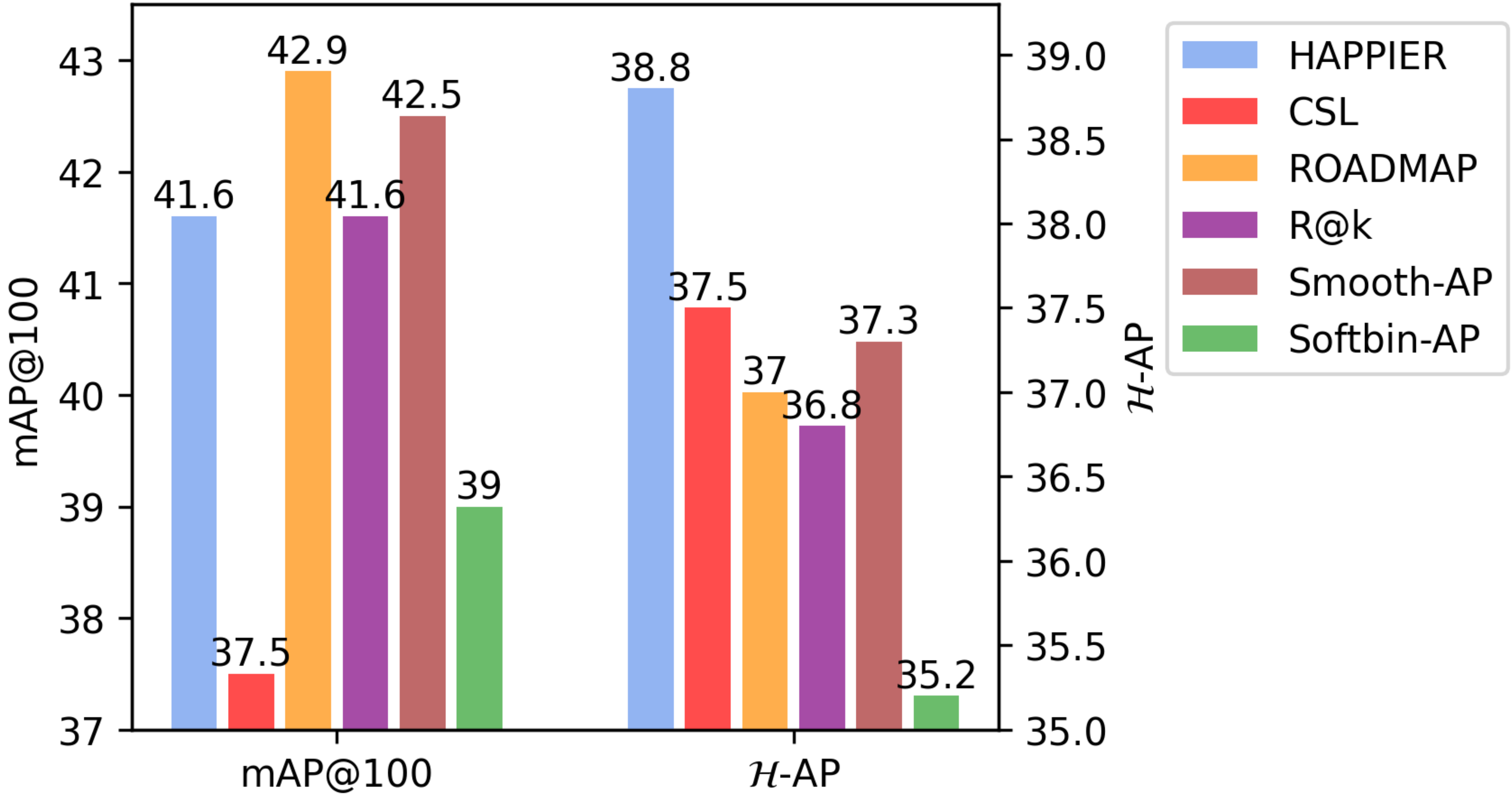
No hierarchical annotations
→ how difficult is it to create hierarchical annotations?

$\mathcal{H}$-GLDv2

1. Scraping Wikimedia Commons
2. Post-processing

[F] Weyand, Tobias, et al. "Google landmarks dataset v2 - a large-scale benchmark for instance-level recognition and retrieval." *CVPR*, 2020.

Results



HAPPIER
CSL
ROADMAP
R@k
Smooth-AP
Softbin-AP
mAP@100
$\mathcal{H}$-AP
41.6
37.5
42.9
41.6
42.5
39
38.8
37.5
37
36.8
37.3
35.2
43
42
41
40
39
38
37
39.0
38.5
38.0
37.5
37.0
36.5
36.0
35.5
35.0

Perspectives

- Uncertainty quantification:
  - Global measure of uncertainty (aleatoric, epistemic)
  - For foundation models, *e.g.*, CLIP
  - Test-time adaptation
- Non-smooth & non-decomposable metrics beyond ranking
- Mistake severity robustness
  - Adaptation to multi-modal models

Thank you for you attention!

C. Corbière, N. Thome, A. Bar-Hen, M. Cord, P. Pérez. Addressing Failure Detection by Learning Model Confidence. NeurIPS 2019.https://github.com/valeoai/ConfidNet

C. Corbière, N. Thome, A. Saporta, T-H. Vu, M. Cord, P. Pérez. Confidence Estimation via Auxiliary Models. IEEE T-PAMI, vol. 44, no. 10, pp. 6043-6055, June 2021.

O. Petit, N. Thome, L. Soler. 3D Spatial Priors for Semi-Supervised Organ Segmentation with Deep Convolutional Neural Networks. IJCARS, Springer Verlag, In press, 2021.

E.Ramzi, N. Thome, C. Rambour, N. Audebert, X. Bitot. "Robust and Decomposable Average Precision for Image Retrieval." *NeurIPS* 2021 https://github.com/elias-ramzi/ROADMAP

E.Ramzi, , N. Audebert, C. Rambour, N. Thome, X. Bitot. "Hierarchical Average Precision Training for Pertinent Image Retrieval." *ECCV* ,2022. https://github.com/elias-ramzi/HAPPIER

Lafon, Marc, E. Ramzi, C. Rambour, N. Thome. "Hybrid Energy Based Model in the Feature Space for Out-of-Distribution Detection." *ICML,* 2023. https://github.com/MarcLafon/heatood

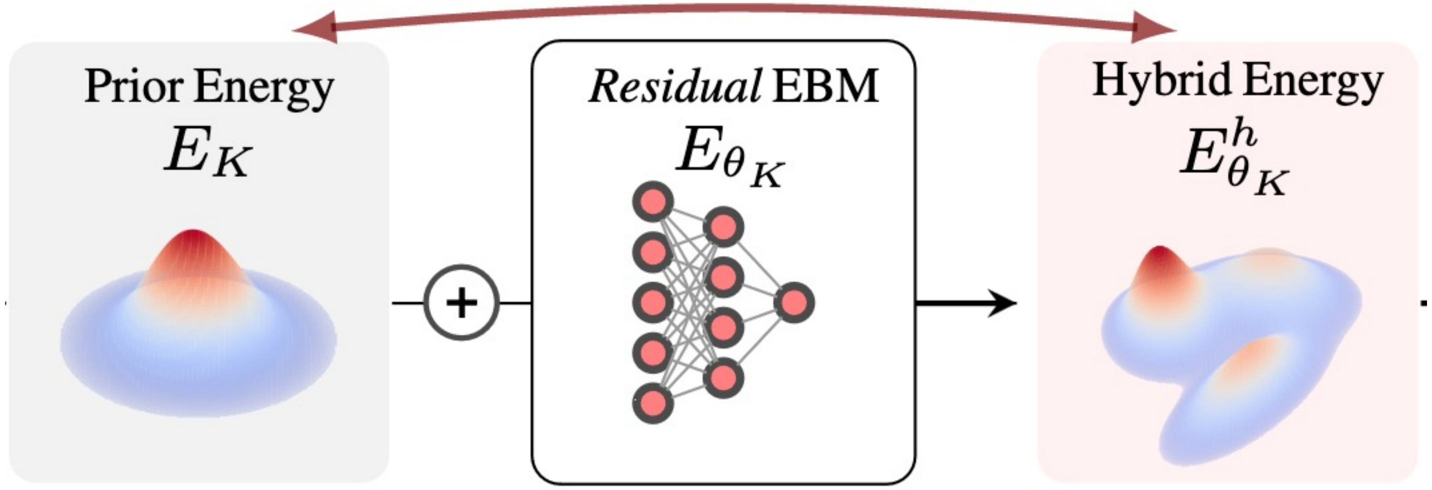
Lafon, Marc, E. Ramzi, C. Rambour, N. Audebert. N. Thome. "GalLoP: Learning Global and Local Prompts for Vision-Language Models." *ECCV,* 2024.

Ramzi, Elias, et al. "Optimization of Rank Losses for Image Retrieval." *under-review TPAMI,* 2024. https://github.com/cvdfoundation/google-landmark

HEAT: Energy correction

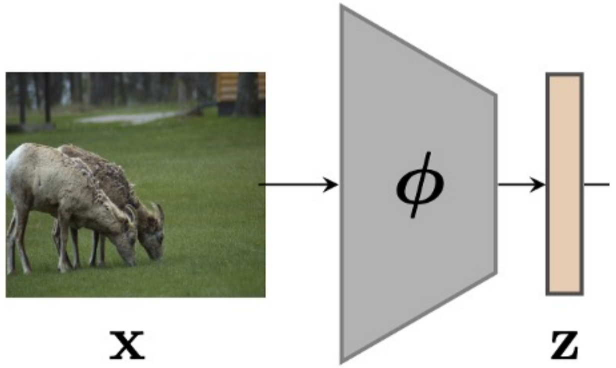
- Hybrid energy: $E^h_{\theta_k}(\mathbf{z}) = E_{q_k}(\mathbf{z}) + E_{\theta_k}(\mathbf{z})$ $\quad p^h_{\theta_k}(\mathbf{z}) = \frac{1}{Z(\theta_k)}\exp(-E^h_{\theta_k}(\mathbf{z}))$
- Controlling the residual $\mathcal{L}_C(\theta_k) = \mathbb{E}_{p_{in}, p^h_{\theta_k}}\left[(E^h_{\theta_k} - E_{q_k})^2\right]$
  - Correction with mimical norm

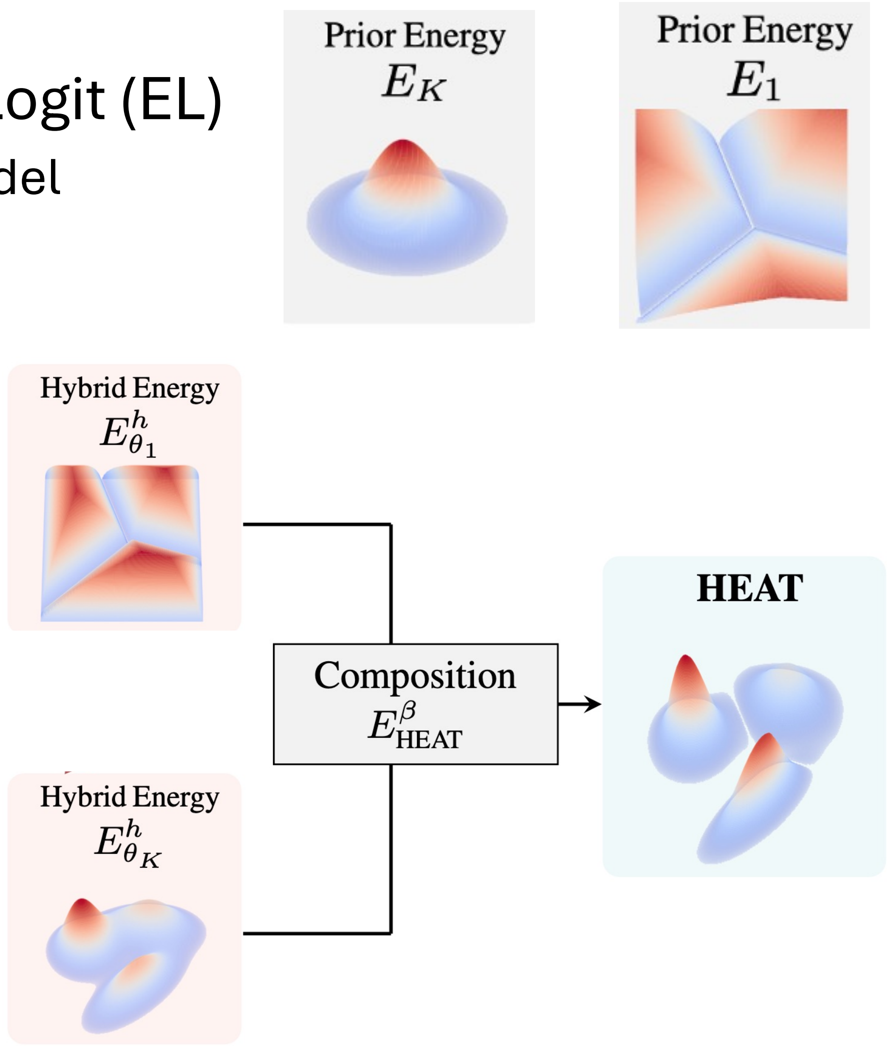
$$\boxed{\mathcal{L}_{tot}(\theta_k) = \mathcal{L}_{MLE}(\theta_k) + \lambda \mathcal{L}_C(\theta_k) \quad (6)}$$



HEAT: Energy composition

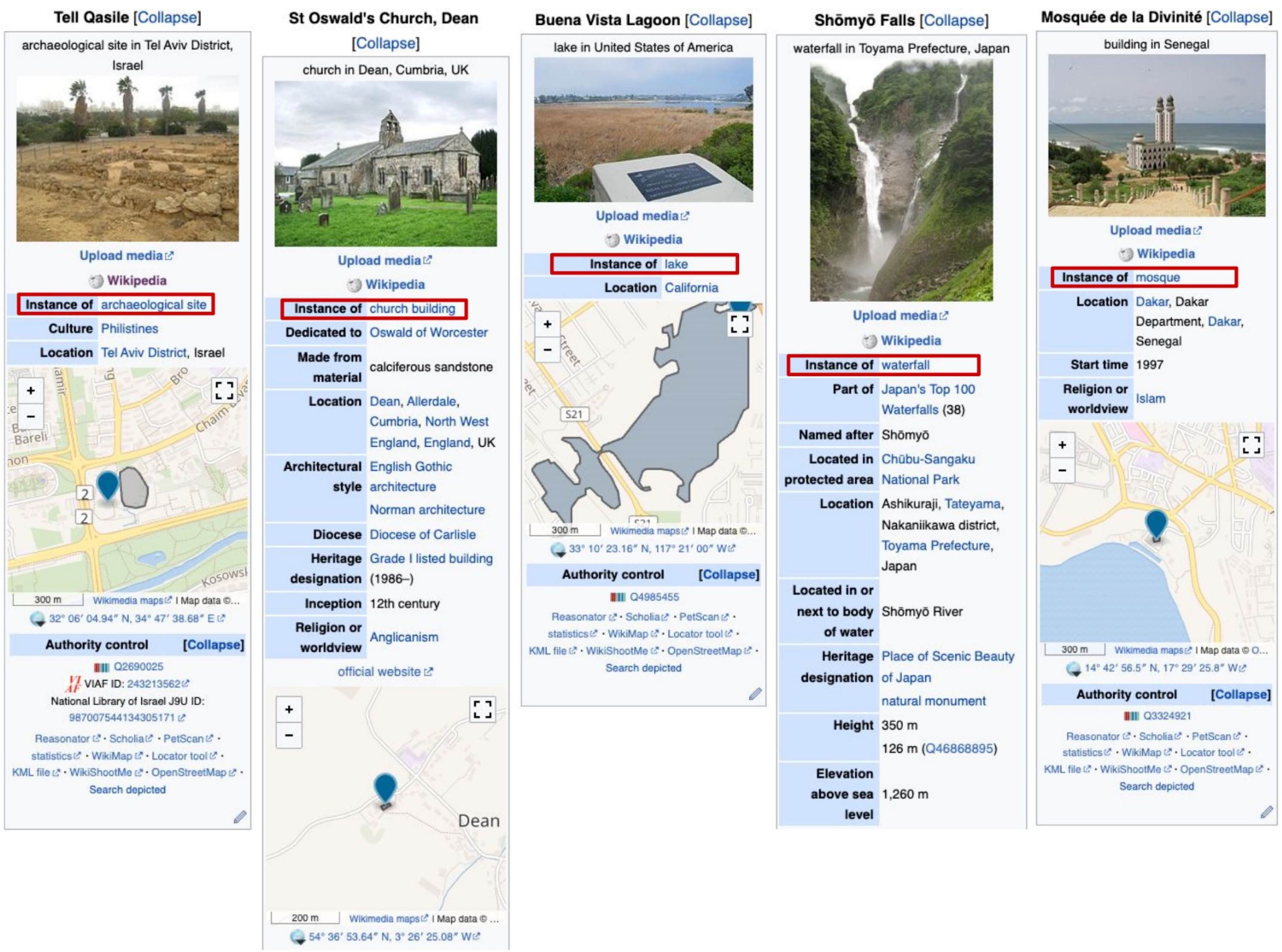
- Prior energy function, *e.g.* Gaussians or Energy Logit (EL)
  - Avg or Std (style) features as inputs for the energy model





- Energy composition: $$E^\beta_{\text{HEAT}} = \frac{1}{\beta}\log\sum_{k=1}^{K} e^{\beta E^h_{\theta_k}}$$
  - $\beta \rightarrow +/-\infty$, max/
  - $\beta = -1$, logsumexp

1. Scraping Wikimedia Commons



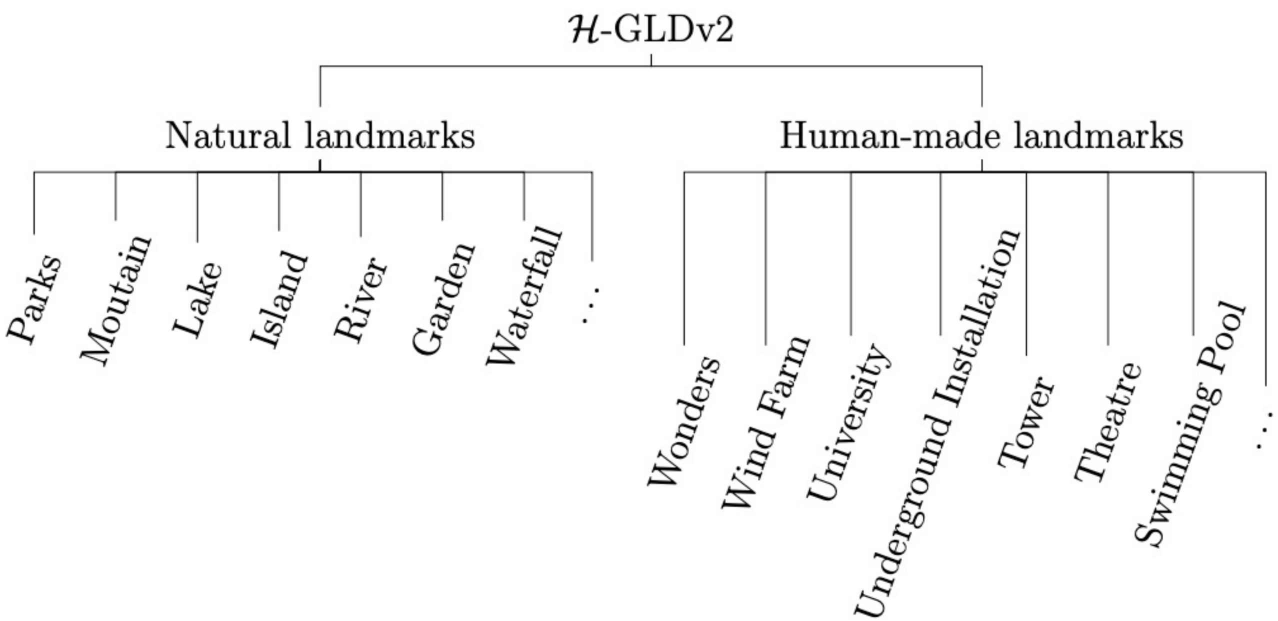
Wikimedia Commons → largest open database of landmarks.

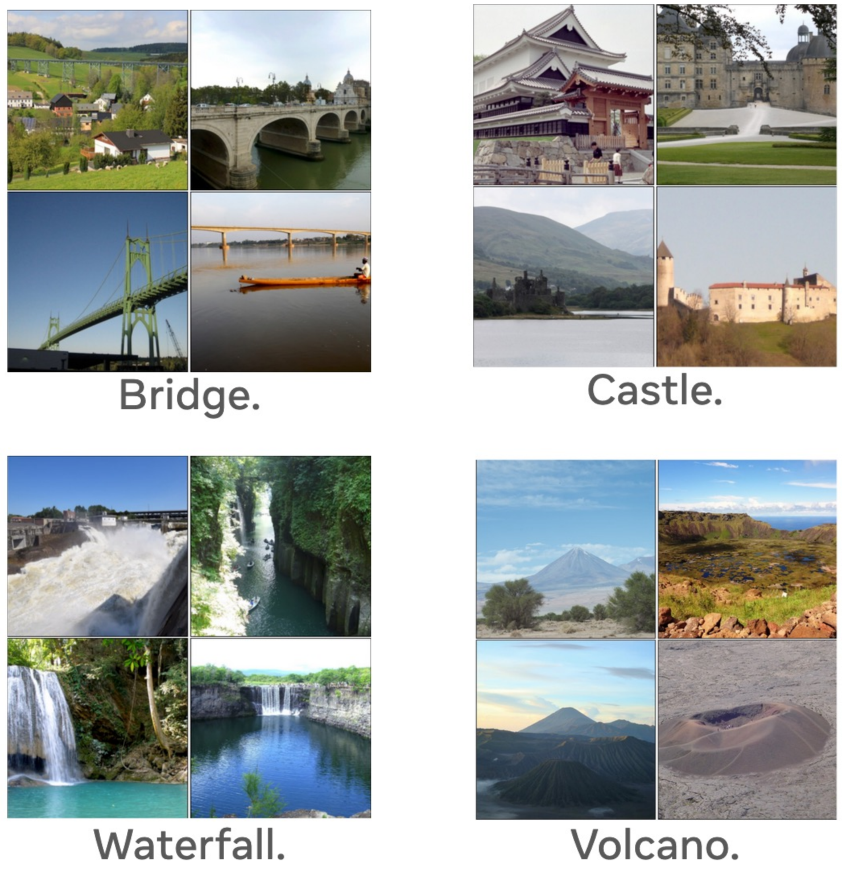
- GLDv2 sourced from wikimedia commons
- « Instance of » => super-category

Ex of scraped labels:

- Church building.
- Church building (1172–1954)
- Cathedral
- Castle
- Corsican nature reserve
- New Zealand great walks
- Waterfall
- Arch-gravity dam
- Canal
- Association football venue
- Astronomical observatory
- Village

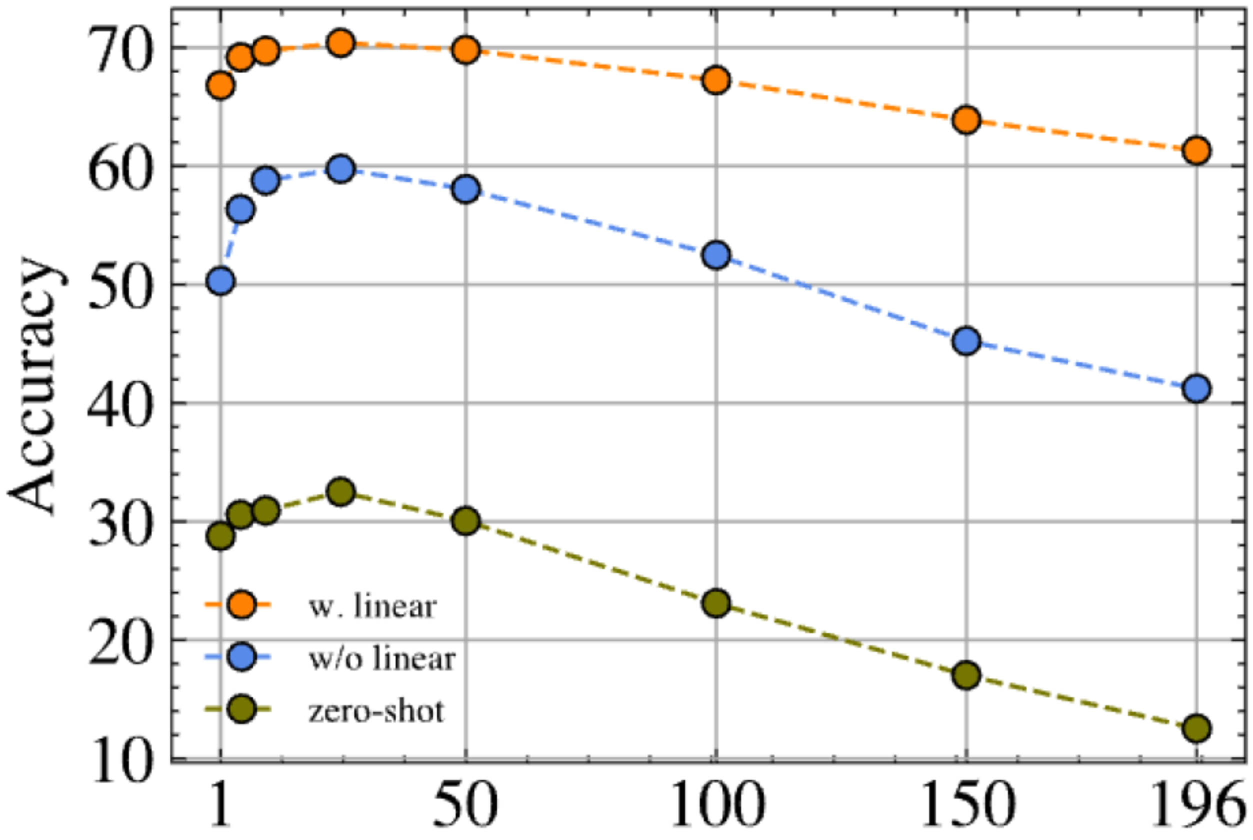
2. Post-processing (manual + automatic)





- K-Means clustering from CLIP’s textual encoder
- Manual verification + adding natural/man made in hierarchy
- 78 super categories

GalLoP Results



Impact of number of regions k

Local Method	Sparse	Dense
$\text{CLIP}_{\text{Local}}$	30.9	12.5
Local Align.	67.9	59.4
$\text{Prompt}_{\text{Local}}$	59.8	41.2
$\text{GalLoP}_{\text{Local}}$	69.8	61.1

Main GalLop's components