Paper List

TMI2021, VOL. 40, NO. 10	A Unified Framework for Generalized Low-Shot Medical Image Segmentation with Scarce Data
CVPR2019	RepMet: Representative-based metric learning for classification and few-shot object detection

Information

A Unified Framework for Generalized Low-Shot Medical Image Segmentation with Scarce Data

Hengji Cui, Dong Wei, Kai Ma, Shi Gu, and Yefeng Zheng, *Senior Member, IEEE*

Manuscript received July 1, 2020; revised October 27, 2020; accepted December 14, 2020. This work was primarily supported by the NSFC General Program 61876032, the Key-Area Research and Development Program of Guangdong Province, China (No. 2018B010111001), the National Key R&D Program of China (2018YFC2000702), and the Scientific and Technical Innovation 2030-"New Generation Artificial Intelligence" Project (No. 2020AAA0104100). *(H. Cui and D. Wei contributed equally to this work.) (Corresponding author: S. Gu.)*

H. Cui and S. Gu are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China (email: chj@std.uestc.edu.cn, gus@uestc.edu.cn). H. Cui contributed to this work as an intern at Tencent Jarvis Lab.

D. Wei, K. Ma, and Y. Zheng are with Tencent Jarvis Lab (email: donwei@tencent.com, kylekma@tencent.com, yefengzheng@tencent.com).

Introduction

- Low-shot medical image segmentation
- Use only 1 or few training data to accomplish segmentation
- Both the annotations and data are scarce, *e.g.*, when developing algorithms for a rare disease, and when the emergence of a novel disease necessitating rapid algorithm development (COVID-19)

Difficulties

- Few-shot segmentation methods for natural natural images CANNOT be directly applied to medical images.
- Recent low-shot medical image segmentation methods relied on image synthesis. These methods utilized a considerable number of unlabeled data, thus inapplicable too.
- Registration-based segmentation(atlas) become less effective and time-consuming when large individual variations in appearance and/or structures are present.
- More challenging due to the shortage in both data and annotations

Contributions

1. Propose the MRE-Net, a unified framework for generalized low-shot medical image segmentation in case of scarcity of both annotations and samples.
2. Propose adaptive mixing coefficients for the modes of a category, a key innovation in multimodal representation learning to ensure that the modes more suitable for the current input are given more emphasis.
3. Conduct thorough experiments on two distinct datasets to demonstrate the superiority of the MRE-Net to both classical and modern methods in a variety of low-shot settings.

Methodology

- Conceptual overview : Segmentation via DML-Based Dense Prediction

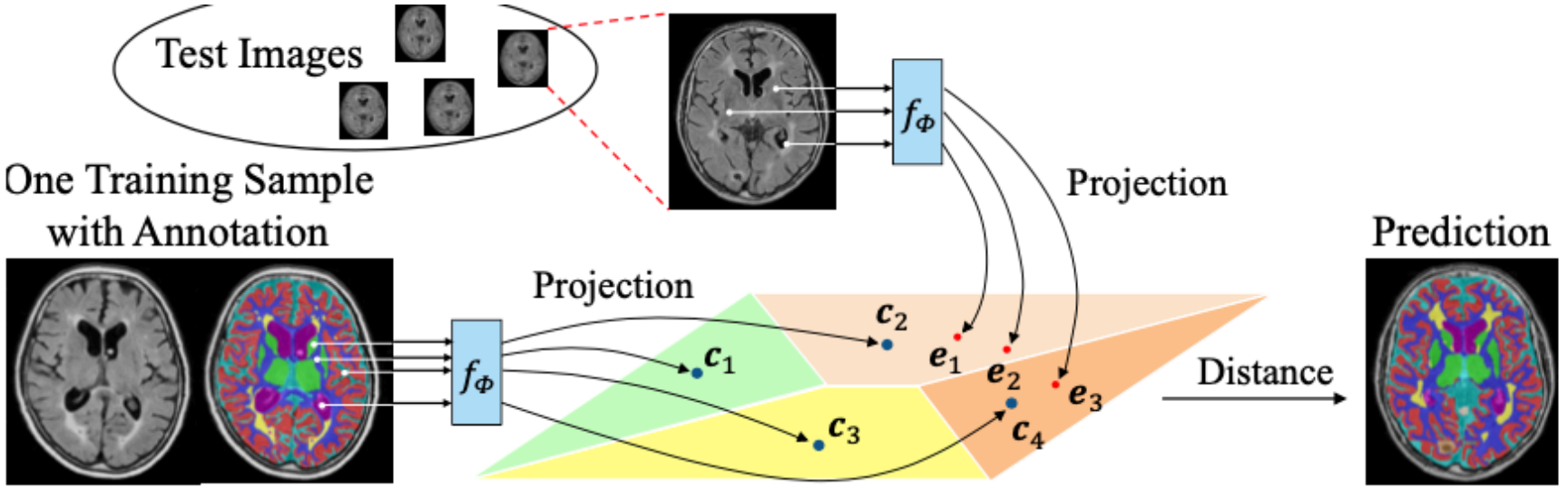


Fig. 1. Conceptual overview of our low-shot segmentation framework. Pixels of test images are projected in the embedding space, where dense predictions are made according to the distances between pixels' embeddings $\boldsymbol{e}_i$ and category prototypes $\boldsymbol{c}_k$.

Methodology

- Architecture of MRE-Net

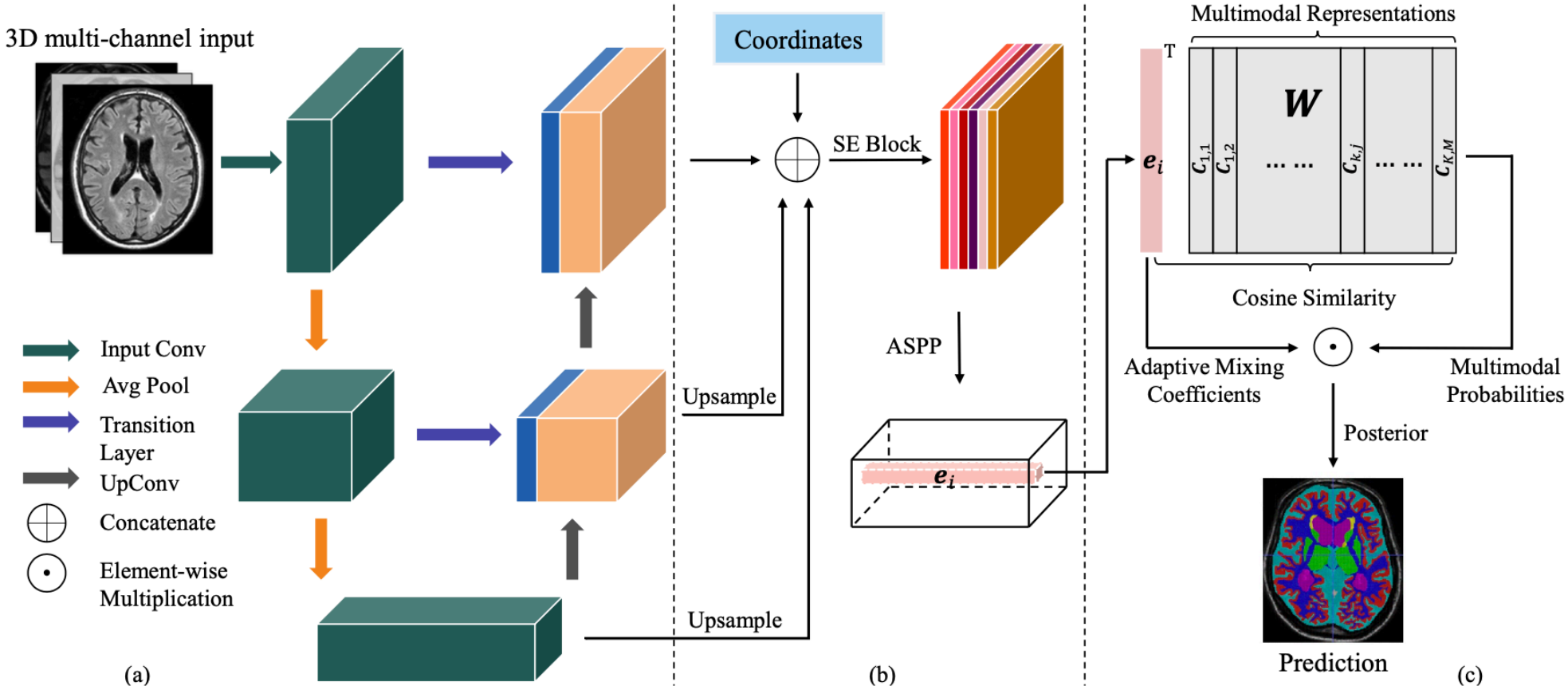


Fig. 2. Overview of the proposed framework: (a) backbone network, (b) dense feature embedding, and (c) cosine distance-based dense prediction. Note that 2D images are used for illustration purpose here, but the real data used in this work are 3D.

Methodology

- Architecture of MRE-Net:
  - Backbone network: a modified 3D-UNet
  - Dense Feature Embedding: Attentional Multi-Scale (AMS) embedding
    1. Concatenation of the (upsampled) multilevel feature maps + Location Map
    2. SE(Squeeze-and-Excitation) attention module
    3. Atrous Spatial Pyramid Pooling (ASPP) module

Methodology

- DML (Distance Metric Learning) of Multimodal Representation via Weight Embedding

    1. DML via Weight Embedding

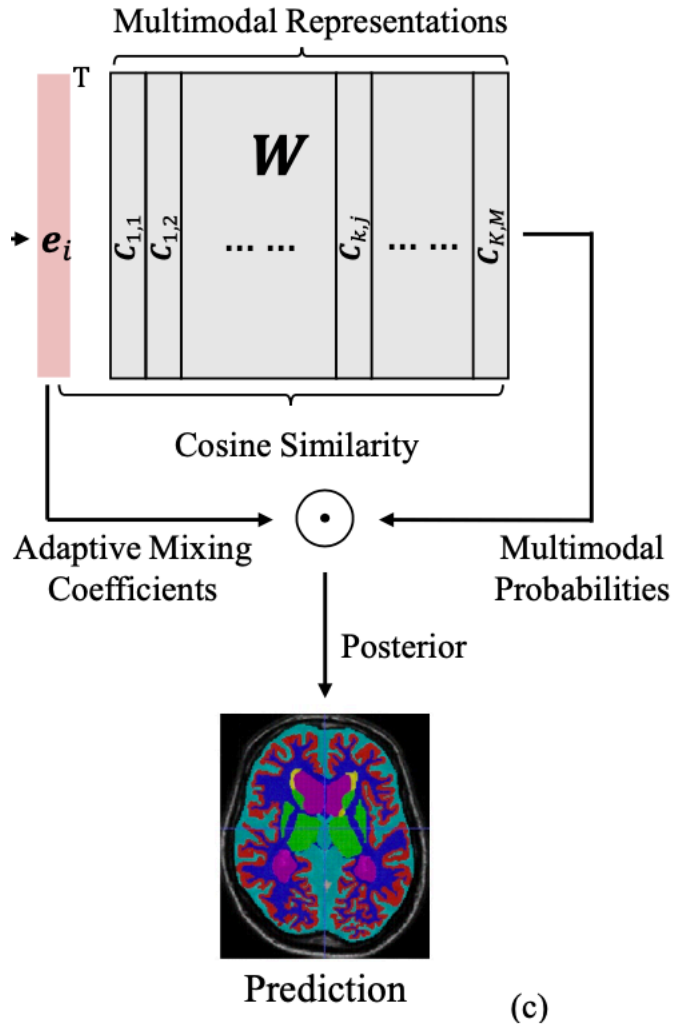
For unimodal category representation, the discriminative class posterior P ,

$$\mathbb{P}(s_i = k) = \frac{p_k(x_i)}{\sum_{k=1}^{K} p_k(x_i)} \triangleq \frac{\exp\left[-d(\boldsymbol{e}_i, \boldsymbol{c}_k)\right]}{\sum_{k=1}^{K} \exp\left[-d(\boldsymbol{e}_i, \boldsymbol{c}_k)\right]}, \quad (1)$$

According to the definition of cosine distance, after adding a learnable scaling factor ξ to adjust the exact values of the exponential function,

$$\mathbb{P}(s_i = k) = \exp\left(\xi \hat{\boldsymbol{e}}_i^{\mathrm{T}} \hat{\boldsymbol{c}}_k\right) \Big/ \sum_{k=1}^{K} \exp\left(\xi \hat{\boldsymbol{e}}_i^{\mathrm{T}} \hat{\boldsymbol{c}}_k\right). \quad (2)$$

,which can be implemented by fc layer with biases set to 0.



Methodology

- DML (Distance Metric Learning) of Multimodal Representation via Weight Embedding

  2. Multimodal Weight Embedding with Adaptive Mixing Coefficients

In multimodal, each category is represented by a mixture distribution model, where the centers of the modes are the category prototypes

(*i.e.*, a category is represented by a set of prototypes, instead of a single one)

The k-th category $R_k = \{\boldsymbol{c}_{k,j}\}_{j=1}^{M}$

The posterior P,

$$\begin{aligned}\mathbb{P}(s_i = k) &= \frac{p_k(x_i)}{\sum_{k=1}^{K} p_k(x_i)} = \frac{\sum_{j=1}^{M} \alpha_{k,j} p_{k,j}(x_i)}{\sum_{k=1}^{K}\sum_{j=1}^{M} \alpha_{k,j} p_{k,j}(x_i)} \\ &\triangleq \frac{\sum_{j=1}^{M} \alpha_{k,j} \exp\left(\xi \hat{\boldsymbol{e}}_i^{\mathrm{T}} \hat{\boldsymbol{c}}_{k,j}\right)}{\sum_{k=1}^{K}\sum_{j=1}^{M} \alpha_{k,j} \exp\left(\xi \hat{\boldsymbol{e}}_i^{\mathrm{T}} \hat{\boldsymbol{c}}_{k,j}\right)}.\end{aligned} \tag{3}$$

, where $\alpha_{k,j}$ is the mixing coefficient for the j[th] mode of the k[th] category and $\sum_{j=1}^{M} \alpha_{k,j} = 1$

Methodology

- DML (Distance Metric Learning) of Multimodal Representation via Weight Embedding

  2. Multimodal Weight Embedding with Adaptive Mixing Coefficients

  Adaptable mixing coefficients:

$\boldsymbol{e}_i$ (of dimension $N_{\boldsymbol{e}} \times 1$) —fc+ReLU→ vector (512*1) —fc+sigmoid→ $\boldsymbol{\beta} = [\beta_{k,j} | k = 1, \ldots, K, \text{and } j = 1, \ldots, M]$

  Finally,

$$\alpha_{k,j} = \exp(\beta_{k,j}) \Big/ \sum_{j=1}^{M} \exp(\beta_{k,j}). \qquad (4)$$

  Denote the parameter of these two fc layers by θ_{mix}

  If M is set to 1, the framework is degraded to unimodal, which is a unified frame work.

Methodology

Loss Function

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{k} \delta_k(a_i) \log \left[\mathbb{P}(s_i = k)\right] / N_k, \qquad (5)$$

where $a_i \in \{1, \ldots, K\}$ is the annotation label for pixel i,

$\delta_k(a_i)$ is the Dirac delta function, which equals 1 if k = a_i and 0 otherwise,

N_k is the number of pixels of the k-th category for loss computation in a mini- batch

Methodology

Training Procedure

Algorithm 1 Training procedure of the proposed MRE-Net.

Input: Annotated training set $\mathcal{X}^{\mathrm{tn}} = \{(x^{\mathrm{tn},(l)}, a^{(l)})\}_{l=1}^{L}$, where $L \leq 3$, and learning rate hyperparameter η

Output: Learned network parameters $\Omega = \{\Phi, \boldsymbol{W}, \theta_{\mathrm{mix}}\}$

1: Randomly initialize Ω
2: **for** number of training iterations **do**
3: Sample mini-batch of images (patches) from $\mathcal{X}^{\mathrm{tn}}$
4: Evaluate $\mathcal{L}_{\mathrm{CE}}$ in Equation (5) using the mini-batch
5: Update parameters with gradient descent:
6: $\Omega \leftarrow \Omega - \eta \nabla_{\Omega} \mathcal{L}_{\mathrm{CE}}$
7: **end for**

Dataset

1. MRBrainS18: 7 subjects, with fully annotated multi-sequence MRI scans acquired on a 3T scanner at the UMC Utrecht (the Nether- lands), including:
   A. T1-weighted (repetition time (TR): 7.9 ms and echo time (TE): 4.5 ms),
   B. T1-weighted inversion recovery (TR: 4416 ms, TE: 15 ms, and inversion time (TI): 400 ms)
   C. T2-FLAIR (TR: 11000 ms, TE: 125 ms, and TI: 2800 ms) sequences.

Officially the evaluation is carried out on 8 of totally 10 classes of labels: cortical gray matter (GM), basal ganglia (BG), white matter (WM), white matter hyper- intensities (WMH), cerebrospinal fluid (CSF), ventricles (VT), cerebellum (CB), and brain stem (BS).

Original Size: 240 × 240 × 48 voxels (voxel size: 0.958 × 0.958 × 3.0 mm^3)

For computational efficiency, crop the central region of size 216 × 216 × 48 voxels

Dataset

2. BTCV: comprises abdominal CT scans acquired at the Vanderbilt University Medical Center from metastatic liver cancer or post-operative ventral hernia patients.

   8 organs of totally 13 annotated abdominal organs are reported performances: spleen, right kidney (r.kid.), left kidney (l.kid.), gallbladder, esophagus, liver, stomach, and pancreas.

   Volume Size: 512 × 512 × 85–512 × 512 × 198 voxels

   Fields of views: approx. 280 × 280 × 280 – 500 × 500 × 650 mm^3

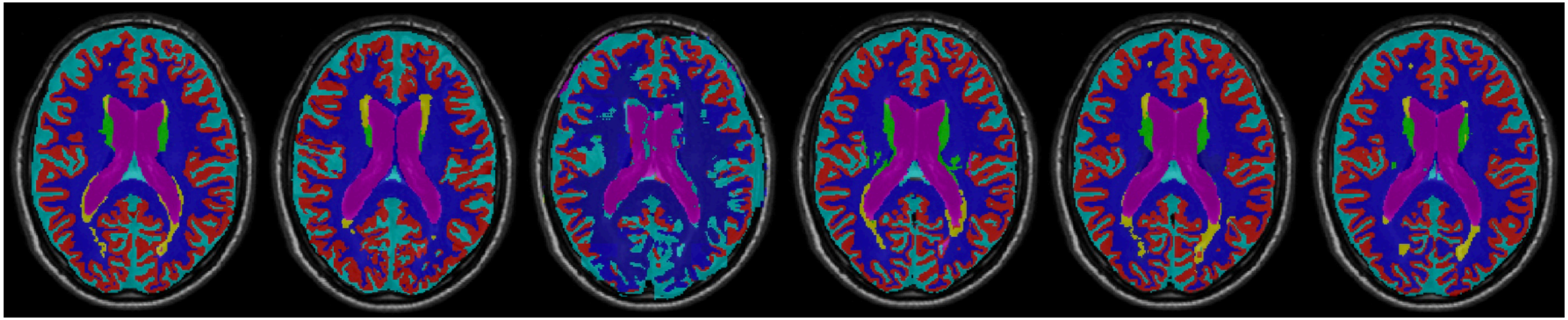
   In-plane resolution: 0.54 × 0.54 - 0.98 × 0.98 mm^2

   Slice thickness: 2.5 - 5.0mm

   Totally, 27 of the public 30 are finally used

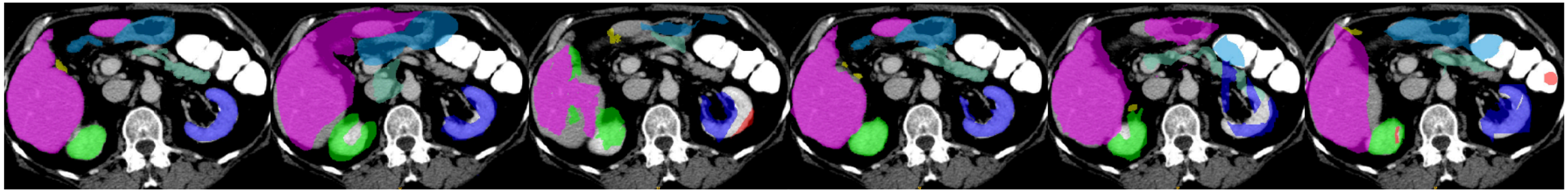
Results

1. Visualization Results



Ground truth | ANTs [15] | Aug. + U-Net-1 [26] | MRE-Net-1 (ours) | MRE-Net-3 (ours) | Aug. + U-Net-6 [26]

Fig. 3. Segmentation results of different methods on the MRBrainS18 dataset. Best viewed in color and zoomed-in.



Ground truth | ANTs-1 [15] | Aug. + U-Net-1 [26] | MRE-Net-1 (ours) | Aug. + U-Net-6 [26] | Aug. + U-Net-20 [26]

Fig. 4. Segmentation results of different methods on the BTCV dataset. Best viewed in color and zoomed-in.

Results

1. Feature Space Visualization Results

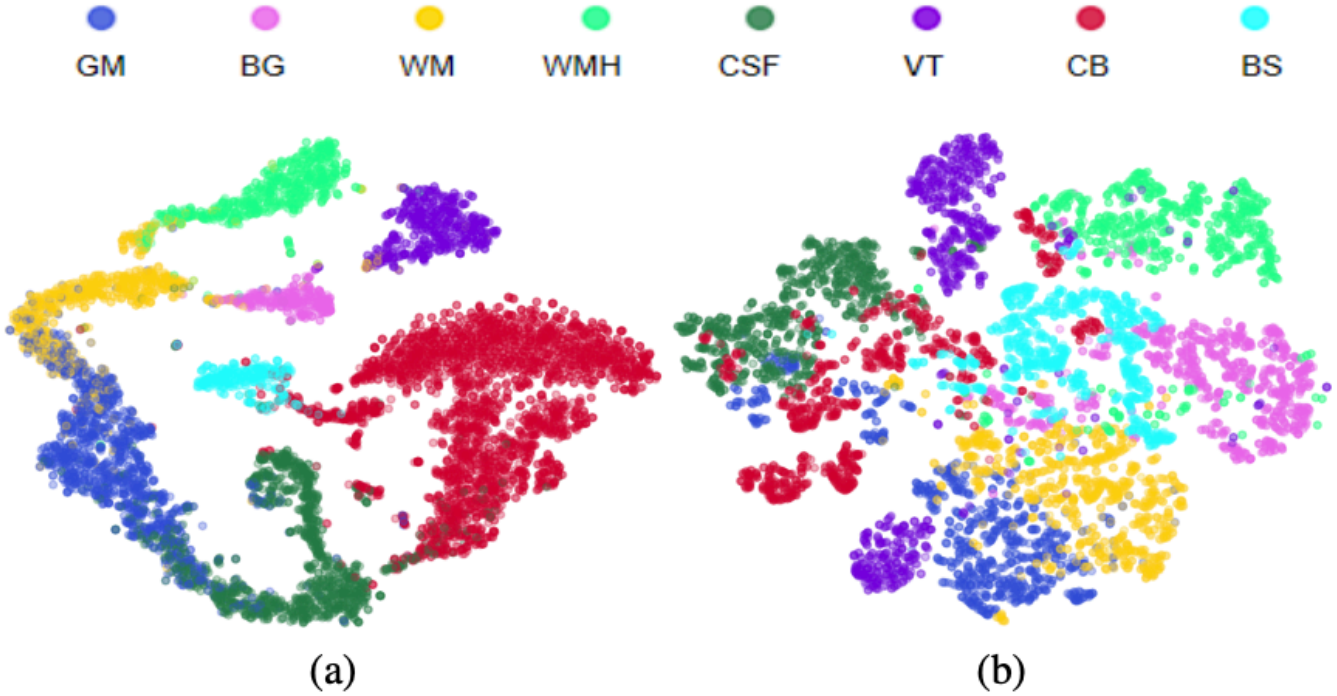


Fig. 5. Test sample visualization (t-SNE) of the feature spaces learned by (a) the proposed MRE-Net and (b) the 3D U-Net [26] (with online data augmentation) on the MRBrainS18 dataset in one-shot setting. Different colors represent different tissues (best viewed digitally).

Results

2. Evaluation Results on MRBrainS18

TABLE VII

EXPERIMENTAL RESULTS ON THE MRBRAINS18 DATASET ("AUG." STANDS FOR DATA AUGMENTATION; "-n" MEANS n-SHOT LEARNING; "MEAN EXCL." REPRESENTS RESULTS EXCLUDING WMH). VALUES ARE MEAN (STD.). BOLD FACES DENOTE BEST RESULTS PER COLUMN (GROUPED BY ONE- AND FEW-SHOT SETTINGS); ASTERISKS (*) DENOTE STATISTICALLY SIGNIFICANT DIFFERENCES FROM THE BEST RESULTS IN "MEAN" AND "MEAN EXCL." COLUMNS.

	GM	BG	WM	WMH	CSF	VT	CB	BS	Mean (std.)	Mean excl. (std.)
Dice Similarity Coefficient (%)										
ANTs-1 [15]	57.06 (3.18)	**77.27** (1.80)	63.98 (4.16)	26.36 (3.01)	57.29 (2.69)	85.13 (3.49)	84.07 (2.79)	**76.24** (11.25)	*65.93 (1.39)	*71.58 (1.29)
U-Net-1 [26]	14.06 (25.56)	6.44 (12.58)	22.22 (23.88)	0.18 (0.27)	24.63 (20.72)	18.77 (23.22)	16.35 (20.56)	16.21 (13.79)	*14.86 (19.11)	*16.96 (18.83)
Aug. + U-Net-1 [26]	58.84 (13.40)	41.54 (9.46)	56.56 (15.93)	26.38 (7.22)	63.13 (11.97)	56.53 (15.33)	55.50 (11.27)	34.72 (13.43)	*49.15 (11.91)	*52.40 (12.28)
MRE-Net-1 (ours)	**80.84** (1.61)	74.88 (2.71)	**83.70** (1.29)	58.42 (2.03)	**79.50** (1.58)	**90.57** (4.43)	87.57 (1.76)	71.66 (6.22)	**78.39** (1.07)	**81.25** (1.01)
Aug. + MRE-Net-1	80.60 (1.83)	75.16 (1.95)	83.21 (1.30)	**58.82** (1.39)	79.25 (1.63)	90.32 (3.06)	**87.60** (2.13)	71.61 (5.49)	78.32 (0.93)	81.11 (1.04)
MRE-Net-2 (ours)	82.34 (1.03)	78.94 (1.39)	84.78 (0.97)	68.67(14.26)	76.77 (0.66)	**92.23** (3.77)	88.61 (0.84)	72.06 (2.42)	80.55 (1.31)	82.25 (1.23)
Aug. + MRE-Net-2	82.46 (1.24)	78.82 (1.30)	84.92 (1.06)	68.38 (13.99)	76.67(0.63)	91.92 (3.72)	88.46 (0.75)	71.78 (2.72)	80.43 (1.78)	82.15 (1.61)
MRE-Net-3 (ours)	82.85 (1.00)	78.53 (1.20)	**85.21** (1.13)	72.64 (4.43)	**81.56** (0.24)	91.22 (0.75)	**90.51** (0.74)	**73.32** (3.01)	**81.98** (1.11)	**83.31** (1.15)
Aug. + MRE-Net-3	82.42 (0.88)	78.67 (1.32)	84.86 (1.01)	73.42 (4.17)	81.14 (0.22)	91.07 (0.79)	90.11 (0.75)	73.03 (2.52)	81.84 (1.28)	83.04 (1.33)
U-Net-6 [26]	81.26 (2.17)	78.49 (2.69)	81.98 (3.48)	74.10 (10.21)	79.20 (1.54)	86.80 (2.57)	88.53 (2.09)	70.39 (5.30)	80.09 (2.76)	80.95 (3.17)
Aug. + U-Net-6 [26]	**84.20** (1.94)	**78.96** (2.60)	84.80 (2.74)	**76.09** (10.25)	78.73 (1.39)	89.18 (3.77)	88.95 (1.78)	68.48 (5.69)	81.17 (2.60)	81.90 (2.80)
Non-low-shot [26][a]	86.0 (-)	83.4 (-)	88.2 (-)	65.2 (-)	83.7 (-)	93.1 (-)	93.9 (-)	90.5 (-)	-	-
95% Hausdorff Distance (mm)										
ANTs-1 [15]	3.04 (0.75)	**4.39** (0.40)	3.93 (0.98)	15.88 (0.86)	3.86 (0.51)	6.78 (6.59)	**5.01** (1.15)	**11.62** (9.86)	6.81 (1.09)	**5.52** (1.17)
U-Net-1 [26]	42.16 (28.74)	75.39 (29.41)	40.69 (22.27)	78.74 (23.61)	33.78 (27.65)	80.36 (20.30)	71.33 (17.21)	51.93 (4.55)	*59.30 (18.21)	*56.52 (17.97)
Aug. + U-Net-1 [26]	25.37 (20.17)	28.11 (18.52)	15.64 (16.96)	52.90 (19.16)	22.35 (11.33)	40.77 (13.36)	31.26 (9.09)	29.48 (10.34)	*30.74 (12.71)	*27.57 (11.96)
MRE-Net-1 (ours)	2.28 (0.12)	5.36 (1.58)	2.74 (0.19)	**10.33** (0.76)	**2.64** (0.22)	**6.31** (1.23)	7.56 (3.75)	13.16 (9.93)	**6.30** (1.73)	5.72 (1.76)
Aug. + MRE-Net-1	**2.10** (0.15)	5.32 (1.46)	**1.51** (0.21)	10.51 (1.51)	2.73 (0.18)	6.50 (1.45)	7.76 (3.78)	13.23 (8.80)	6.34 (1.87)	5.74 (1.79)
MRE-Net-2 (ours)	1.75 (0.13)	4.36 (0.32)	2.57 (0.20)	10.02 (2.50)	2.43 (0.14)	4.78 (0.53)	6.93 (2.45)	12.04 (2.01)	5.61 (1.06)	4.98 (1.24)
Aug. + MRE-Net-2	1.69 (0.09)	4.26 (0.38)	2.83 (0.21)	9.77 (1.46)	2.54 (0.12)	4.67 (0.79)	6.64 (2.93)	12.26 (4.90)	5.58 (1.62)	4.99 (1.57)
MRE-Net-3 (ours)	1.60 (0.11)	**4.03** (0.44)	2.49 (0.11)	9.87 (1.37)	2.31 (0.11)	3.26 (1.35)	6.17 (1.99)	12.32 (2.37)	5.26 (1.69)	4.60 (1.66)
Aug. + MRE-Net-3	**1.54** (0.09)	4.16 (0.30)	2.64 (0.15)	**9.61** (1.29)	**2.27** (0.14)	**3.00** (0.83)	**6.12** (1.96)	12.24 (2.00)	**5.20** (1.72)	**4.57** (1.75)
U-Net-6 [26]	1.86 (0.30)	4.42 (0.57)	2.91 (0.35)	17.25 (4.13)	2.95 (0.18)	8.24 (4.74)	7.16 (5.90)	15.31 (6.94)	7.51 (3.37)	6.12 (3.17)
Aug. + U-Net-6 [26]	1.65 (0.10)	4.12 (0.44)	**2.37** (0.19)	14.31 (5.32)	2.55 (0.09)	7.64 (3.91)	7.08 (4.77)	**10.32** (6.06)	6.26 (2.60)	5.10 (2.80)
Non-low-shot [26][a]	1.27 (-)	3.64 (-)	2.12 (-)	10.25 (-)	2.22 (-)	2.81 (-)	3.74 (-)	3.92 (-)	-	-

[a] As reported in [26] (missing numbers are represented by '-'). Note that different datasets and testing schemes preclude direction comparisons.

Results

3. Evaluation Results on BTCV

TABLE VIII

EXPERIMENTAL RESULTS ON THE BTCV DATASET ("AUG." STANDS FOR DATA AUGMENTATION, "-n" MEANS n-SHOT LEARNING). VALUES ARE MEAN (STD.). BOLD FACES DENOTE BEST RESULTS PER COLUMN; ASTERISKS (*) DENOTE STATISTICALLY SIGNIFICANT DIFFERENCES FROM THE MRE-NET-1 IN THE "MEAN" COLUMN.

	Spleen	R.kid.	L.kid.	Gallbladder	Esophagus	Liver	Stomach	Pancreas	Mean (std.)
Dice Similarity Coefficient (%)									
ANTs-1 [15]	51.62 (2.97)	33.60 (4.63)	17.65 (5.57)	15.25 (4.25)	56.30 (3.71)	76.48 (1.55)	24.60 (1.61)	4.31 (0.64)	*34.98 (3.73)
Aug. + U-Net-1 [26]	31.48 (3.75)	44.84 (4.11)	34.03 (4.62)	26.38 (2.98)	17.24 (2.80)	61.68 (1.41)	11.56 (0.59)	19.48 (0.82)	*30.84 (5.22)
MRE-Net-1 (ours)	**78.08** (1.92)	**75.92** (1.70)	**77.43** (2.98)	**52.15** (1.76)	**56.83** (1.62)	**91.35** (0.81)	**59.65** (1.53)	**61.26** (2.21)	**69.13** (1.43)
Aug. + U-Net-6 [26]	63.49 (1.56)	66.26 (1.89)	66.29 (0.80)	33.74 (2.32)	30.99 (2.45)	87.27 (0.88)	33.19 (0.89)	29.88 (1.02)	*51.39 (2.98)
Aug. + U-Net-20 [26]	64.81 (1.33)	72.73 (1.77)	69.26 (0.95)	28.63 (1.12)	25.75 (1.59)	85.51 (1.04)	45.95 (1.94)	25.80 (0.43)	*52.30 (2.57)
Non-low-shot[a]									
Zhou *et al.* [41]	92 (-)	-	91 (-)	65 (-)	43 (-)	95 (-)	60 (-)	62 (-)	-
DenseVNet [39]	95 (-)	-	93 (-)	73 (-)	71 (-)	95 (-)	87 (-)	75 (-)	-
95% Hausdorff Distance (mm)									
ANTs-1 [15]	63.30 (10.57)	42.44 (13.59)	105.26 (29.81)	41.56 (16.76)	43.78 (10.11)	31.72 (7.70)	33.09 (5.89)	70.33 (7.43)	*53.93 (15.47)
Aug. + U-Net-1 [26]	47.14 (8.52)	52.93 (10.71)	51.77 (8.09)	44.88 (7.98)	40.39 (10.30)	34.01 (5.27)	32.91 (5.25)	46.15 (6.17)	*43.77 (13.18)
MRE-Net-1 (ours)	**22.93** (3.37)	**18.54** (3.20)	**19.47** (5.66)	**27.61** (7.23)	**15.87** (6.78)	**10.62** (2.78)	**12.98** (2.36)	**24.10** (5.16)	**19.02** (8.63)
Aug. + U-Net-6 [26]	43.05 (3.82)	44.22 (5.00)	47.33 (6.57)	36.16 (9.39)	35.06 (10.44)	27.28 (7.17)	28.45 (10.57)	36.26 (6.69)	*37.23 (9.57)
Aug. + U-Net-20 [26]	42.57 (2.10)	39.08 (4.31)	46.65 (7.95)	32.12 (5.12)	34.87 (8.24)	21.35 (7.48)	20.74 (3.16)	31.17 (6.19)	*33.57 (6.91)

[a] As reported in [39] (missing numbers are represented by '-'); mean values for HD95 were not reported in [39]. Different datasets and testing schemes preclude direction comparisons.

Results

4. Comparison Results

TABLE XI

COMPARISON WITH THE SEMI-SUPERVISED DATAAUG METHOD [11] IN ONE-SHOT SETTING ON THE BTCV DATASET ("AUG." STANDS FOR ONLINE DATA AUGMENTATION; "-n" MEANS n UNLABELED SCANS WERE USED). VALUES ARE MEAN (STD.). BOLD FACES DENOTE BEST RESULTS PER COLUMN; ASTERISKS (*) DENOTE STATISTICALLY SIGNIFICANT DIFFERENCES FROM THE MRE-NET IN THE "MEAN" COLUMN.

	Spleen	R.kid.	L.kid.	Gallbladder	Esophagus	Liver	Stomach	Pancreas	Mean (std.)
				Dice Similarity Coefficient (%)					
Aug. + U-Net [26]	32.32 (6.38)	47.13 (18.40)	32.40 (8.54)	14.81 (4.14)	19.43 (3.69)	57.61 (11.90)	54.25 (12.08)	19.55 (5.88)	*34.69 (3.84)
MRE-Net (ours)	**78.76** (3.46)	**73.54** (4.42)	**78.36** (3.26)	**60.43** (9.77)	**55.44** (14.27)	**89.79** (3.08)	**62.43** (13.21)	**59.71** (15.94)	**69.81** (4.58)
Semi-supervised [11]									
SAS-aug-5	29.51 (3.44)	8.10 (4.57)	12.58 (4.91)	6.85 (1.98)	4.93 (3.08)	60.78 (7.51)	14.46 (9.68)	18.28 (9.19)	*19.44 (3.15)
IDS-aug-5	36.26 (6.51)	11.66 (6.33)	15.58 (11.16)	2.93 (4.49)	13.59 (5.01)	64.02 (9.29)	23.49 (10.17)	20.23 (8.23)	*23.47 (2.33)
SAS-aug-10	38.76 (6.20)	13.40 (3.93)	13.30 (4.08)	6.12 (3.72)	9.51 (4.29)	74.74 (3.12)	24.08 (3.15)	18.16 (8.87)	*24.76 (2.46)
IDS-aug-10	37.10 (8.11)	12.85 (6.78)	15.58 (11.94)	3.46 (4.17)	14.10 (5.65)	64.61 (9.23)	23.90 (9.83)	21.84 (9.29)	*24.18 (2.61)
SAS-aug-20	42.89 (8.13)	16.51 (13.79)	16.74 (8.27)	6.68 (5.56)	13.06 (2.98)	79.53 (3.79)	30.92 (3.88)	22.34 (8.33)	*28.58 (3.74)
IDS-aug-20	38.85 (8.18)	15.41 (6.86)	16.97 (11.80)	3.55 (2.86)	17.63 (5.05)	68.77 (7.59)	28.98 (10.25)	22.92 (10.20)	*26.64 (2.11)
				95% Hausdorff Distance (mm)					
Aug. + U-Net [26]	42.32 (15.49)	51.76 (14.31)	47.37 (24.62)	46.31 (6.96)	33.73 (11.83)	34.15 (11.78)	34.02 (9.10)	39.26 (18.33)	*41.12 (7.43)
MRE-Net (ours)	**19.89** (9.73)	**16.57** (4.36)	**18.49** (5.63)	**21.02** (5.75)	**13.42** (6.11)	**12.31** (7.87)	**9.30** (5.05)	**15.37** (5.28)	**15.80** (3.46)
Semi-supervised [11]									
SAS-aug-5	43.08 (10.20)	54.97 (7.92)	57.66 (12.72)	47.30 (13.55)	37.58 (7.73)	35.79 (7.21)	39.78 (8.01)	55.27 (13.45)	*46.43 (4.68)
IDS-aug-5	43.12 (7.36)	45.40 (12.10)	51.68 (5.49)	47.19 (6.57)	44.86 (7.08)	31.65 (7.25)	34.96 (12.72)	43.20 (13.37)	*42.76 (3.74)
SAS-aug-10	36.98 (9.83)	49.37 (7.20)	51.09 (12.16)	42.15 (13.24)	31.21 (7.94)	30.34 (5.95)	32.95 (7.75)	47.98 (18.94)	*40.26 (3.67)
IDS-aug-10	42.40 (6.60)	44.51 (12.52)	49.07 (5.96)	46.36 (9.74)	44.03 (7.86)	29.71 (6.56)	33.04 (12.69)	42.98 (13.82)	*41.51 (3.86)
SAS-aug-20	34.86 (15.31)	45.52 (13.15)	47.00 (17.51)	43.96 (6.13)	29.15 (12.97)	30.05 (14.22)	27.32 (10.09)	45.57 (24.33)	*37.93 (10.41)
IDS-aug-20	40.13 (6.12)	47.96 (7.93)	46.52 (6.71)	43.71 (12.65)	41.38 (8.77)	29.06 (7.63)	30.88 (11.74)	42.63 (13.63)	*40.28 (3.08)

Results

4. Comparison Results

TABLE X

COMPARISON WITH THE sSE METHOD [35] IN ONE-SHOT SETTING ("AUG." STANDS FOR DATA AUGMENTATION). VALUES ARE MEAN (STD.). BOLD FACES DENOTE BEST RESULTS PER COLUMN, AND ASTERISKS (*) DENOTE STATISTICALLY SIGNIFICANT DIFFERENCES FROM THE MRE-NET.

	MRBrainS18		BTCV	
	Dice (%)	HD95 (mm)	Dice (%)	HD95 (mm)
MRE-Net (ours)	**78.39** (1.07)	**6.30** (1.73)	**69.13** (1.43)	**19.02** (8.63)
Aug. + U-Net [26]	*49.15 (11.91)	*30.74 (12.71)	*30.84 (5.22)	*43.77 (13.18)
Aug. + sSE (k = 10)	*28.20 (2.26)	*31.58 (0.74)	*23.04 (4.05)	*61.53 (13.14)
Aug. + sSE (k = max.)	*22.66 (1.33)	*36.54 (0.90)	*25.37 (1.75)	*63.87 (3.31)

TABLE IX

COMPARISON OF THE TIME EFFICIENCY OF DIFFERENT METHODS.

	ANTs [15]	Aug. + U-Net [26]	MRE-Net (ours)
Training (one-shot)	-	∼120 min	∼150 min
Inference (per sample)	∼28 min[a]	∼3s	∼4s

[a] Using 64 threads on an Intel® Xeon® E5-2690 v4 @2.60GHz processor.

Ablation Study

TABLE II
RESULTS OF THE ABLATION STUDY ON DESIGN COMPONENTS IN ONE-SHOT SETTING. VALUES ARE MEAN (STD.).

	DML	Cartesian coordinate	AMS embedding	OHEM	MRBrainS18		BTCV	
					Dice (%)	HD95 (mm)	Dice (%)	HD95 (mm)
(a)	-	-	-	-	14.86 (19.11)	59.30 (18.21)	25.91 (5.53)	58.29 (16.08)
(b)	✓	-	-	-	65.38 (2.27)	14.47 (2.08)	53.36 (2.75)	35.47 (9.26)
(c)	✓	✓	-	-	69.41 (2.72)	9.36 (1.34)	59.08 (2.47)	26.59 (8.70)
(d)	✓	✓	✓	-	74.47 (1.65)	6.95 (1.77)	67.86 (1.90)	21.18 (8.99)
(e)	✓	✓	✓	✓	**78.39** (1.07)	**6.30** (1.73)	**69.13** (1.43)	**19.02** (8.63)
(f)	-	✓	✓	✓	38.17 (4.18)	24.56 (3.54)	31.23 (4.82)	42.63 (12.69)
(g)	✓	-	✓	✓	70.67 (2.34)	12.76 (2.18)	56.96 (2.67)	22.69 (10.22)

TABLE III
IMPACT OF THE EMBEDDING DIMENSION N_e (WITH $M = 3$). VALUES ARE MEAN (STD.), WHERE APPLICABLE.

N_e	512	1024	2048	4096
Dice (%)	75.77 (1.93)	77.56 (0.95)	**78.39** (1.07)	76.99 (1.26)
HD95 (mm)	6.59 (1.98)	6.62 (1.62)	**6.30** (1.73)	6.47 (1.86)
GPU memory (GB)	∼4×5.90	∼4×7.68	∼4×10.18	∼4×14.44

TABLE IV
IMPACT OF THE NUMBER OF MODES M FOR EACH CATEGORY (WITH $N_e = 2048$). VALUES ARE MEAN (STD.).

M	1	2	3	4	5
			One-shot		
Dice (%)	74.91 (1.98)	**78.42** (1.13)	78.39 (**1.07**)	78.05 (1.30)	77.68 (1.62)
HD95 (mm)	7.08 (1.81)	**6.21** (1.55)	6.30 (1.73)	6.54 (1.83)	6.93 (**1.40**)
			Three-shot		
Dice (%)	79.21 (1.25)	81.39 (1.43)	**81.98** (**1.11**)	80.99 (1.39)	81.23 (1.91)
HD95 (mm)	5.88 (1.83)	**5.25** (1.79)	5.26 (**1.69**)	5.44 (1.76)	5.50 (1.75)

Ablation Study

TABLE V
IMPACT OF MIXING COEFFICIENTS. VALUES ARE MEAN (STD.).

	One-hot [16]	Average	Adapt (proposed)
Dice (%)	77.04 (1.28)	76.08 (1.38)	**78.39** (1.07)
HD95 (mm)	7.59 (1.96)	6.54 (1.78)	**6.30** (1.73)

TABLE VI
RESULTS OF THE ABLATION STUDY ON DISTANCE METRICS IN ONE-SHOT SETTING. VALUES ARE MEAN (STD.).

	MRBrainS18				BTCV			
Distance	Dice (%)	HD95 (mm)	GPU memory	Training time	Dice (%)	HD95 (mm)	GPU memory	Training time
Cosine	78.39 (1.07)	6.30 (1.73)	∼4×10.18 GB	∼150 min	69.13 (1.43)	19.02 (8.63)	∼4×9.81 GB	∼140 min
Euclidean	77.57 (1.01)	6.52 (1.91)	∼8×11.30 GB	∼260 min	67.68 (1.48)	18.95 (9.17)	∼8×10.95 GB	∼250 min

Information

RepMet: Representative-based metric learning for classification and few-shot object detection

Leonid Karlinsky[*], Joseph Shtok[*], Sivan Harary[*], Eli Schwartz[*], Amit Aides, Rogerio Feris
IBM Research AI
Haifa, Israel

Raja Giryes
School of Electrical Engineering, Tel-Aviv University
Tel-Aviv, Israel

Alex M. Bronstein
Department of Computer Science, Technion
Haifa, Israel

Introduction

- Few-shot learning:

An N -way, M -shot episode is an instance of the few-shot task

Description of an N -way, M -shot episode:

A set of M training examples from each of the N categories,

and 1 query image of an object from one of the categories.

The goal is to determine the correct category for the query.

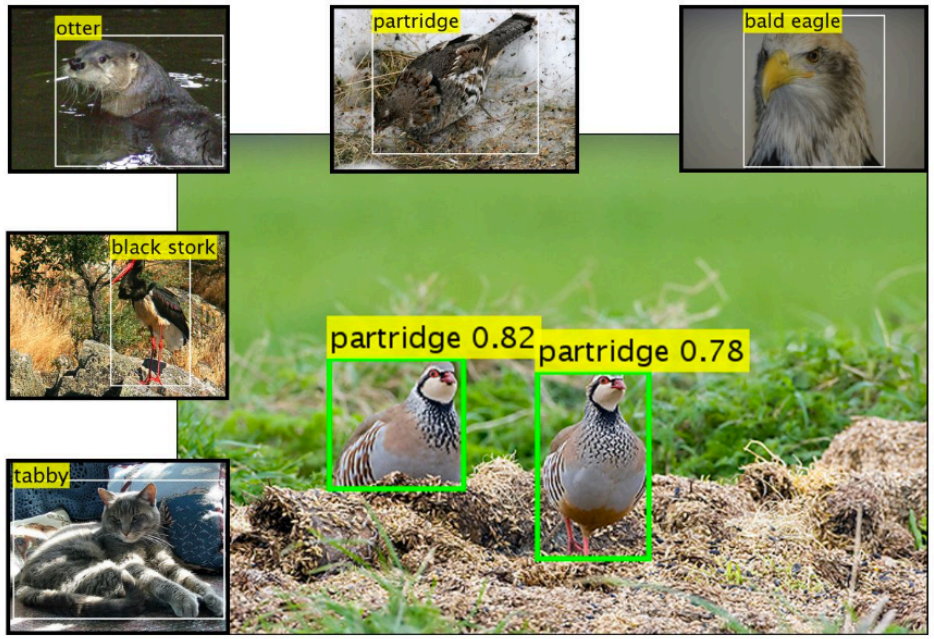
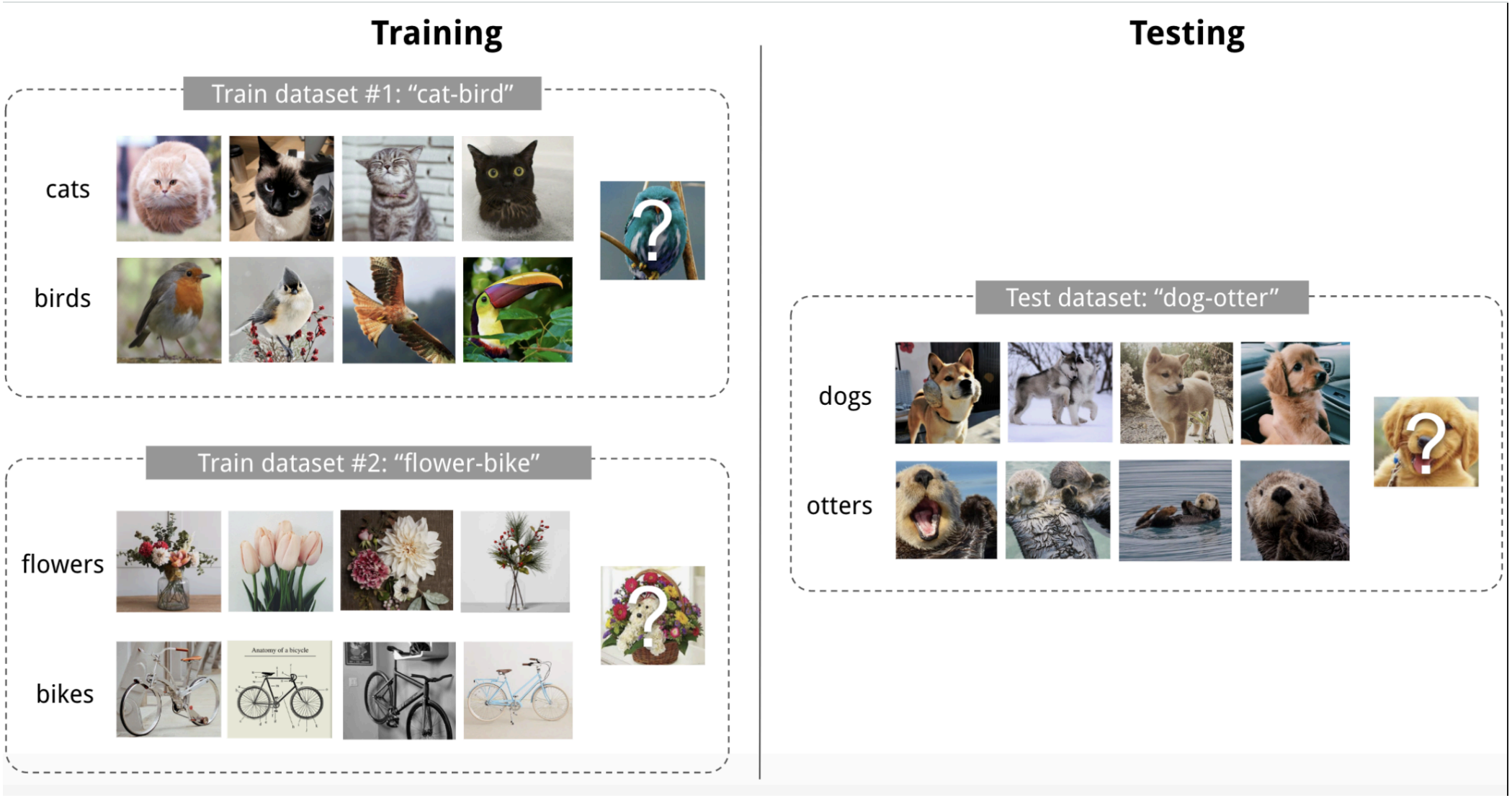


Figure 1. **One-shot detection example**. Surrounding images: examples of new categories unseen in training. Center image: detection result for the one-shot detector on an image containing instances of partridge, which is one of the new categories.

Introduction



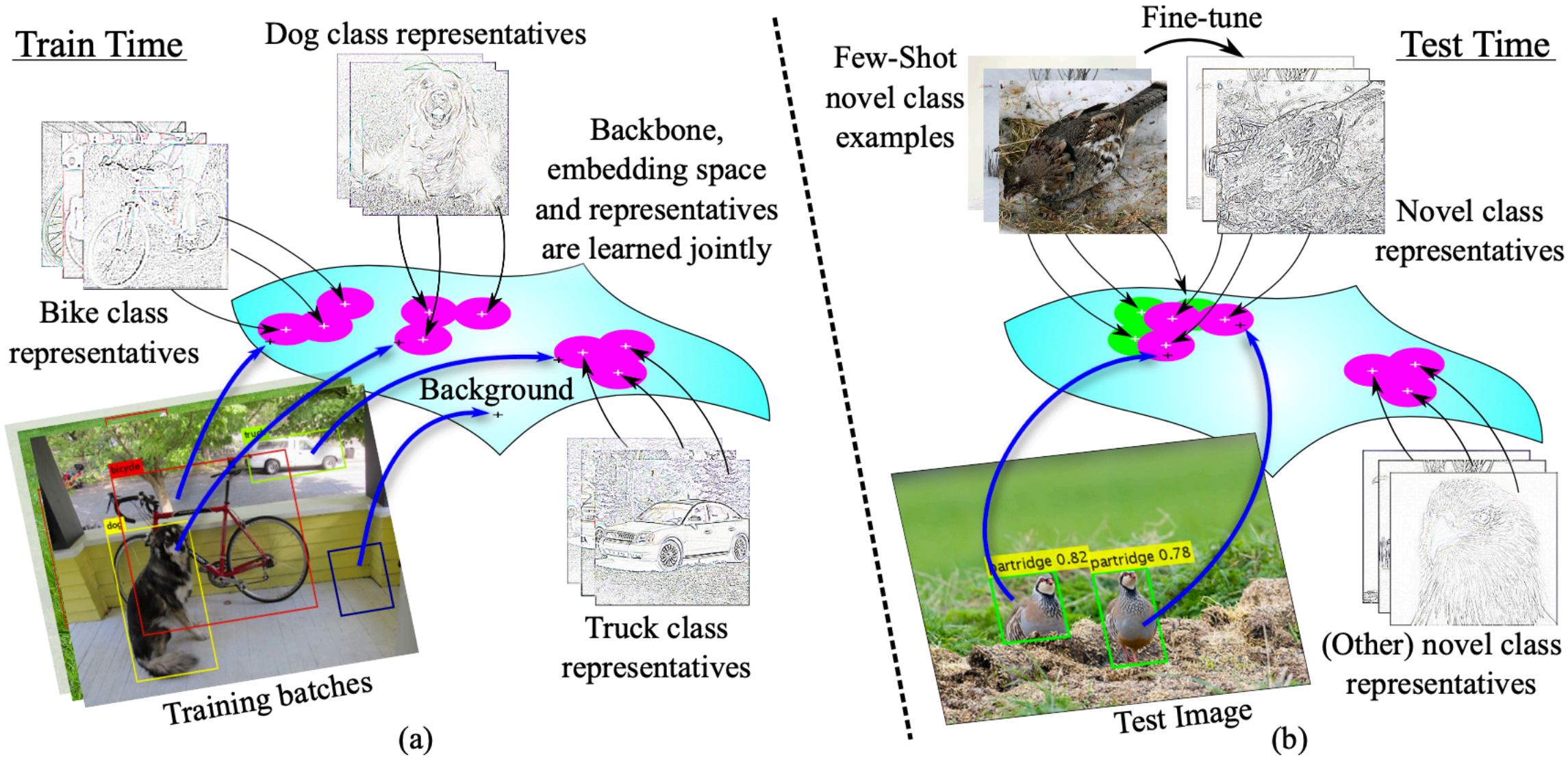


Figure 2. **Overview of our approach.** (a) *Train time:* backbone, embedding space and mixture models for the classes are learned jointly, class representatives are mixture mode centers in the embedding space; (b) *Test time:* new (unseen during training) classes are introduced to the detector in the learned embedding space using just one or a few examples. Fine tuning the representatives and the embedding (on the episode train data) can be used to further improve performance (Section 5). For brevity, only two novel classes are illustrated in the test. The class posteriors are computed by measuring the distances of the input features to the representatives of each of the classes.

Contributions

1. Propose a novel sub-net architecture for jointly training an embedding space together with the set of mixture distributions in this space, having one (multi-modal) mixture for each of the categories.
2. First to propose a method to equip an object detector with a DML classifier head that can admit new categories, and thus transform it into a few-shot detector
3. Offer an episodic benchmark for the few-shot detection problem, built on a challenging fine-grained few-shot detection task.

Methodology

- Network architecture.

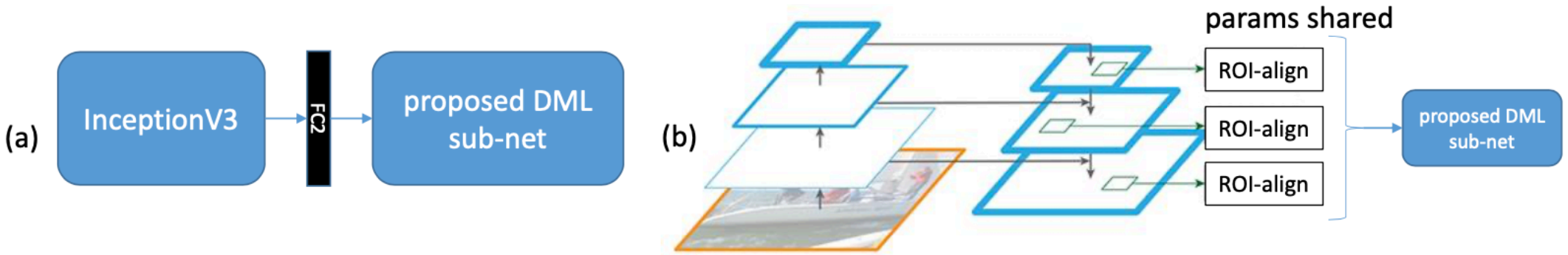


Figure 4. **Network architectures used.** (a) Network for DML based classification. (b) Network for few-shot detection; its backbone is FPN+DCN with deformable ROI-align [7].

Methodology

- RepMet DML sub-net architecture.

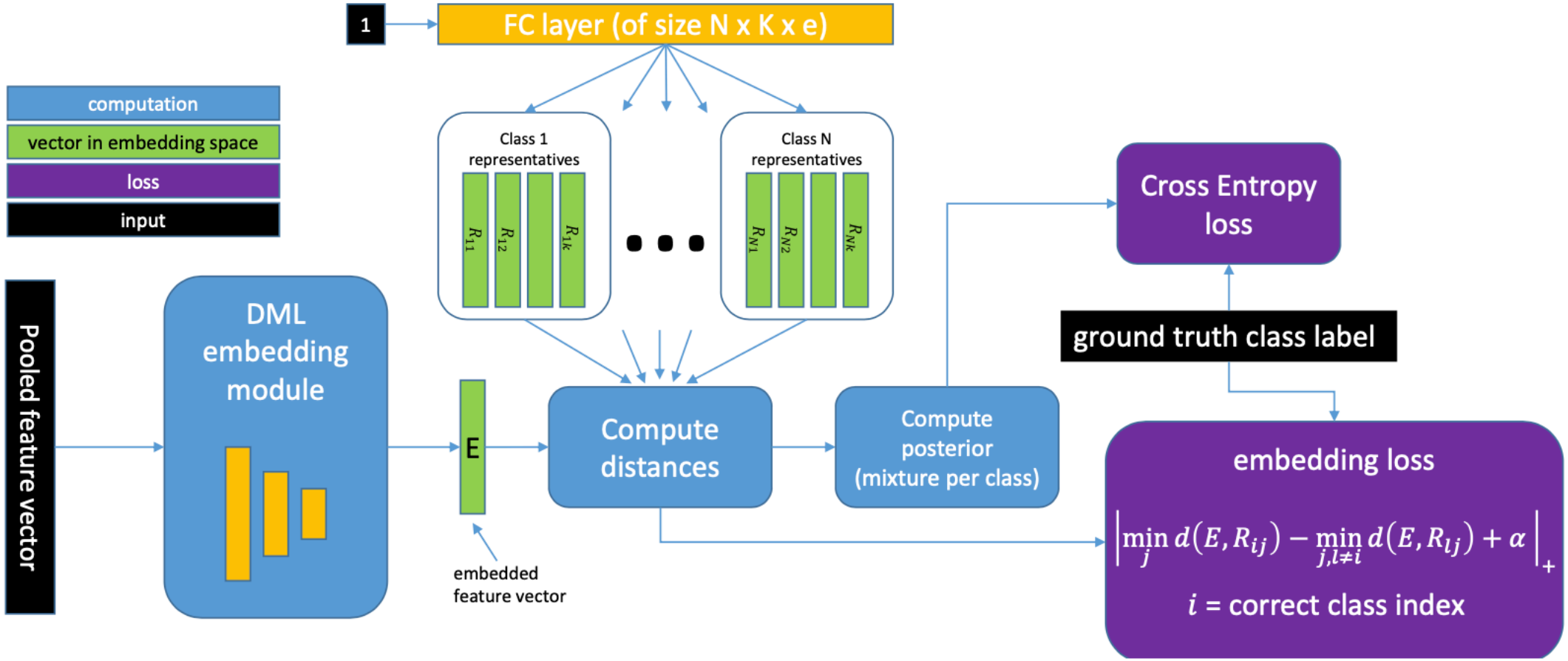


Figure 3. **The proposed RepMet DML sub-net architecture** performs joint end-to-end training of the DML embedding together with the modes of the class posterior distribution.

Methodology

- RepMet DML sub-net architecture.
- Similar to the former paper,
- Different in the distance computation:

$$p_{ij}(E) \propto \exp\left(-\frac{d_{ij}^2(E)}{2\sigma^2}\right). \qquad (1)$$

- Simplify the mixing coefficients by using the upper bound on the actual class posterior

$$\mathbb{P}(\mathcal{C}=i|X) = \mathbb{P}(\mathcal{C}=i|E) \equiv \max_{j=1,\ldots,K} p_{ij}(E). \qquad (2)$$

- Reason: Mixture coefficients are associated with specific modes, and since the modes change at test time, learning the mixture coefficients becomes highly non-trivial.
- Differently, for the former paper, in medical images, all the categories are in the training set.

Methodology

- RepMet DML sub-net architecture.
- For background,

$$\mathbb{P}(\mathcal{B}|X) = \mathbb{P}(\mathcal{B}|E) = 1 - \max_{ij} p_{ij}(E). \qquad (3)$$

- Loss: CE-Loss + eq.4

$$L(E, R) = \left| \min_{j} d_{i^* j}(E) - \min_{j, i \neq i^*} d_{ij}(E) + \alpha \right|_+ , \qquad (4)$$

- where i^* is the correct class index for the current example and $|\cdot|_+$ is the ReLU function

Results

	method			
dataset	MsML [22]	Magnet [25]	VMF [42]	Ours
Stanford Dogs	29.7	24.9	24.0	**13.7**
Oxford Flowers	10.5	8.6	**4.4**	11
Oxford Pet	18.8	10.6	9.9	**6.9**
ImageNet Attributes	–	15.9	—	**13.2**

Table 1. Comparison of **test error (in %)** with the state-of-the-art DML classifier approaches on different fine-grained classification datasets (lower is better).

	1-shot	5-shot	10-shot
LSTD [5]	19.2	37.4	44.3
ours	**24.1**	**39.6**	**49.2**

Table 2. Comparison to LSTD [5] on their Task 1 experiment: 50-way detection on 50 ImageNet categories (as mAP %).

Results

dataset	method	no episode fine-tuning			with episode fine-tuning		
		1-shot	5-shot	10-shot	1-shot	5-shot	10-shot
ImageNet-LOC (214 unseen animal classes)	baseline-FT (FPN-DCN [7])	—	—	—	35.0	51.0	59.7
	baseline-DML	41.3	58.2	61.6	41.3	59.7	66.5
	baseline-DML-external	19.0	30.2	30.4	32.1	37.2	38.1
	Ours	**56.9**	**68.8**	**71.5**	**59.2**	**73.9**	**79.2**
ImageNet-LOC (100 seen animal classes)	Ours - trained representatives	—	86.3	—	—	—	—
	Ours - episode representatives	64.5	79.4	82.6	—	—	—

Table 3. Few-shot 5-way detection test performance on ImageNet-LOC. Reported as mAP in %.

Results

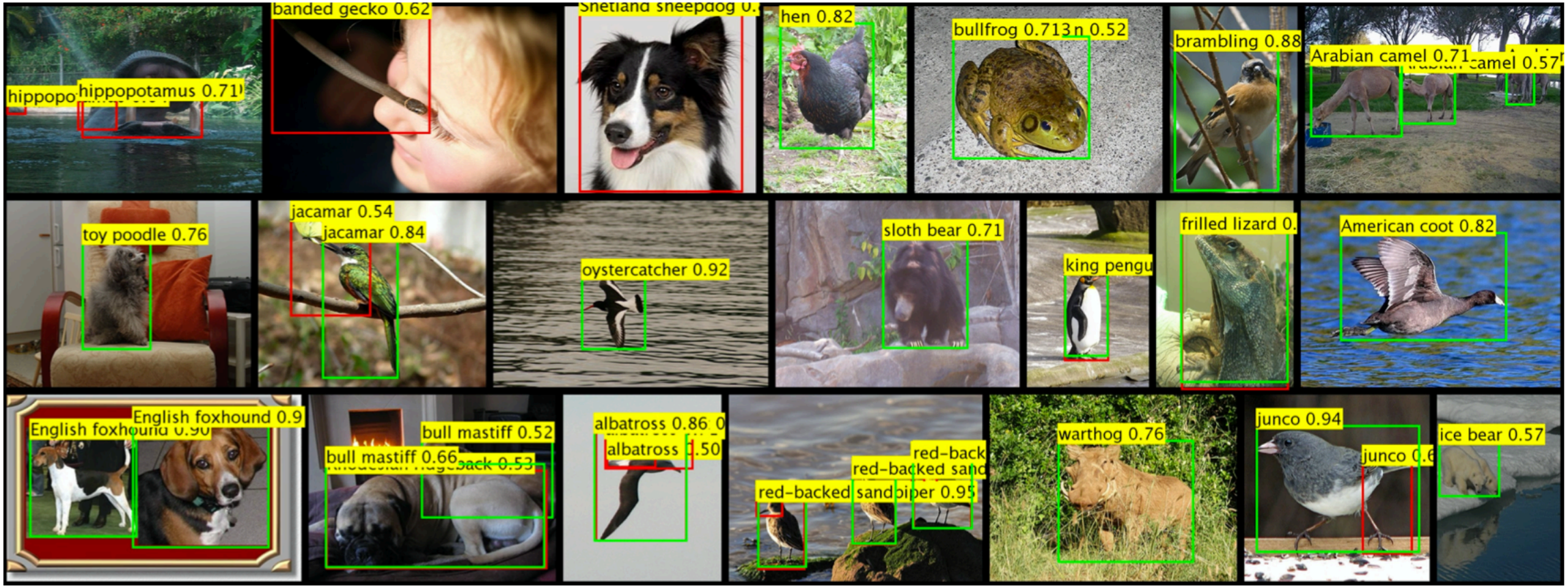


Figure 7. **Example one-shot detection results.** Green frames indicate correctly detected objects and red frames indicate wrong detections. A threshold of 0.5 on the detection score is used throughout. Detections with higher scores are drawn on top of those with lower scores.