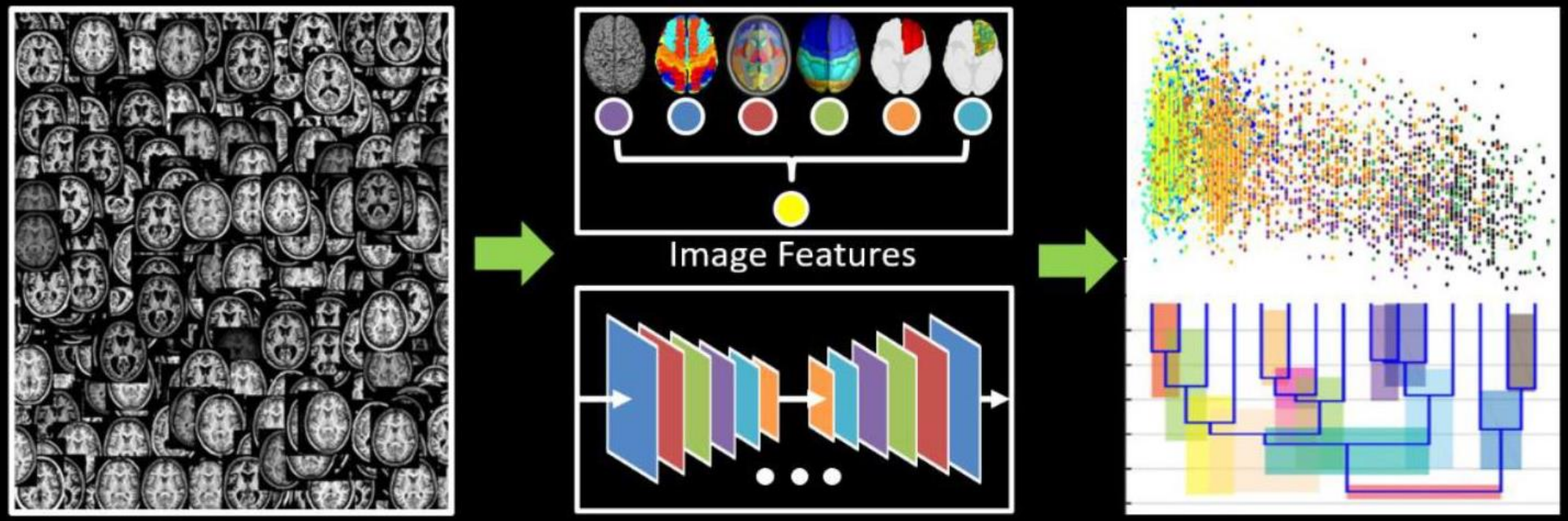


Semi-supervised and Weakly-supervised Learning

[Spring 2020 CS-8395 Deep Learning in Medical Image Computing]

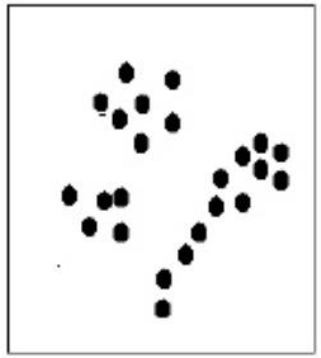
Instructor: Yuankai Huo, Ph.D.
Department of Electrical Engineering and Computer Science
Vanderbilt University

Topics

- Semi-supervised Learning
- Weakly-supervised Learning
- Papers

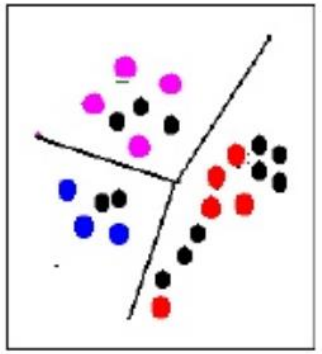
Learn from big unlabeled images

Unsupervised Learning



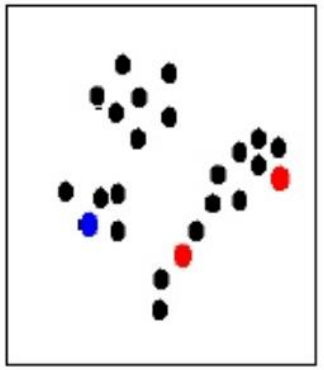
We know nothing about data sructure

Supervised Learning



We know well data sructure

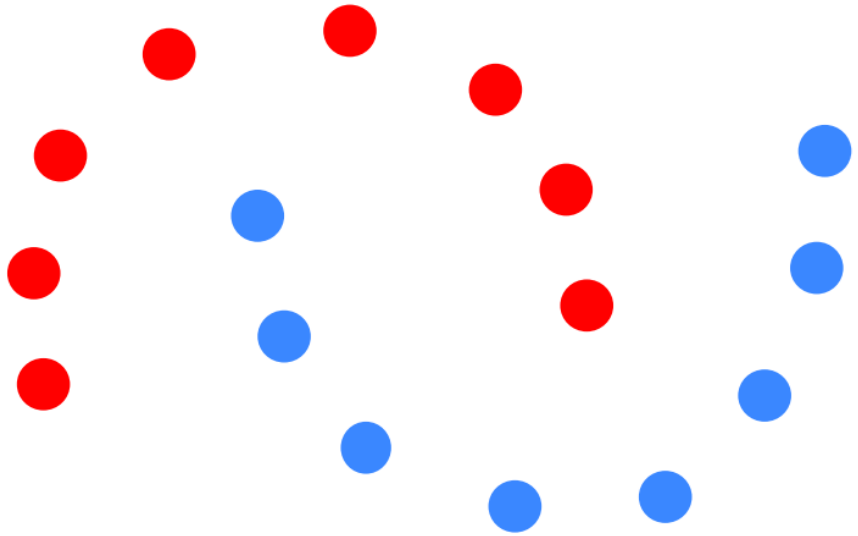
Semi-Supervised Learning



We know something about data sructure

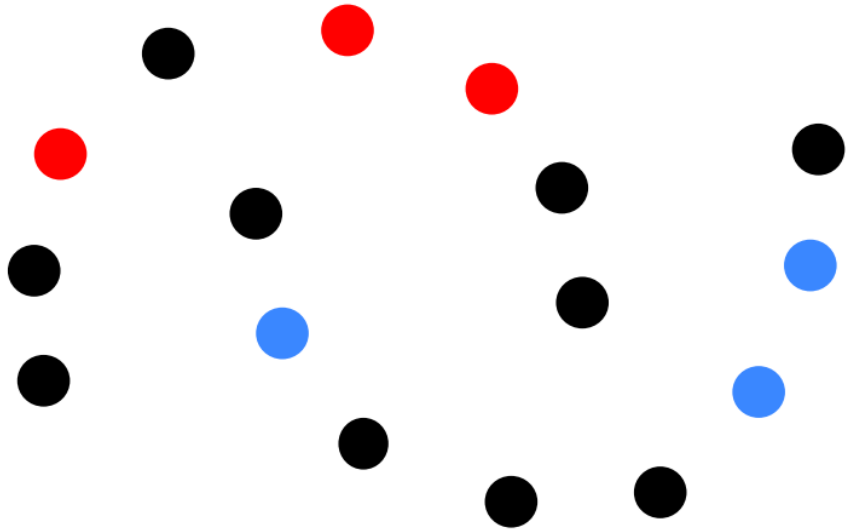
https://www.slideshare.net/Tatiana.lando/clustering-presentation

Two Moons Dataset



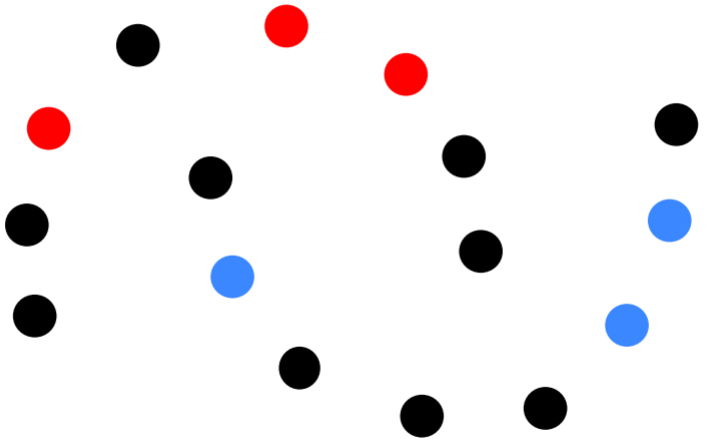
https://web.stanford.edu/class/cs224n/lectures/lecture17.pdf

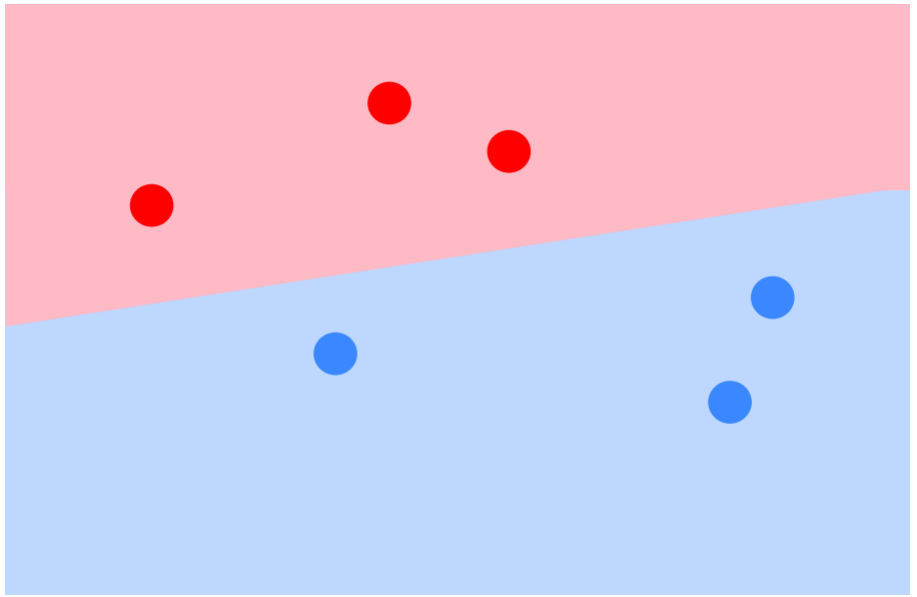
Don’t have all labels



https://web.stanford.edu/class/cs224n/lectures/lecture17.pdf

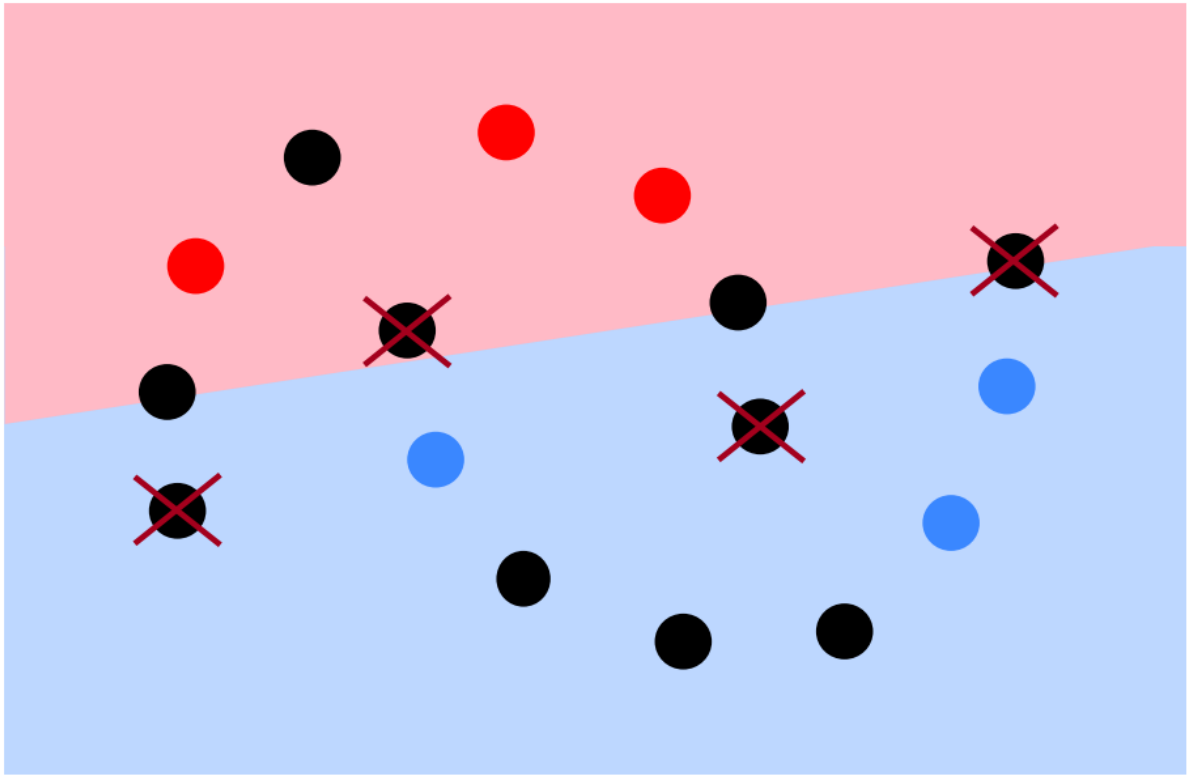
Supervised Learning





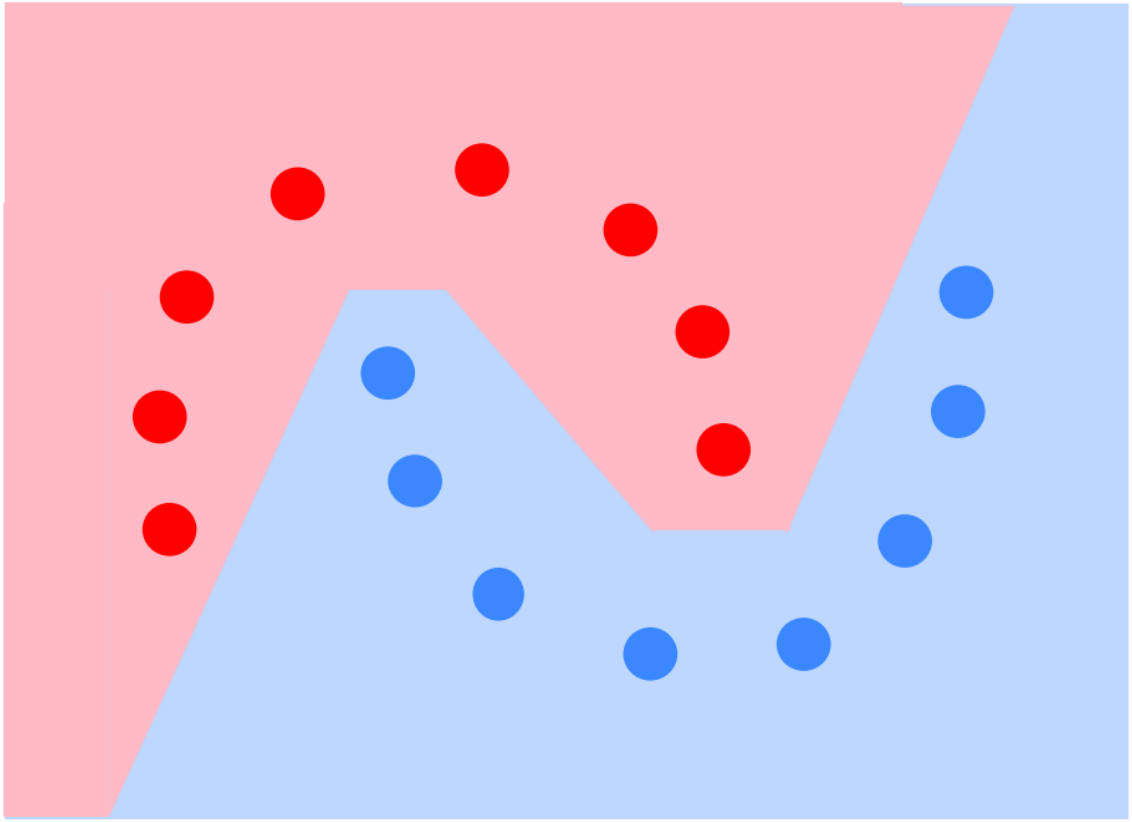
https://web.stanford.edu/class/cs224n/lectures/lecture17.pdf

Supervised Learning



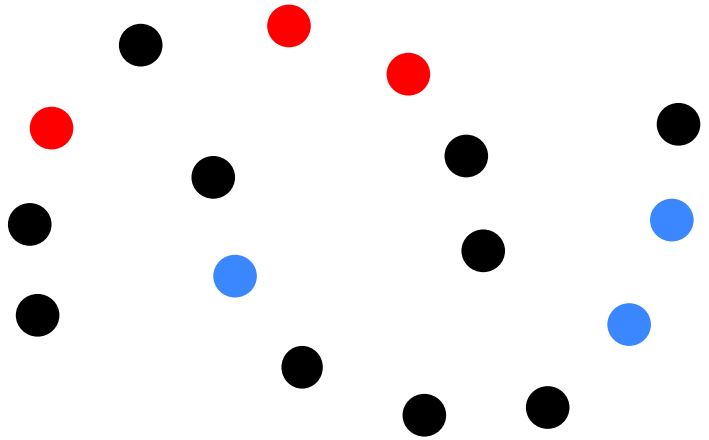
https://web.stanford.edu/class/cs224n/lectures/lecture17.pdf

Semi-supervised Learning



https://web.stanford.edu/class/cs224n/lectures/lecture17.pdf

Semi-supervised Learning



https://web.stanford.edu/class/cs224n/lectures/lecture17.pdf

Example

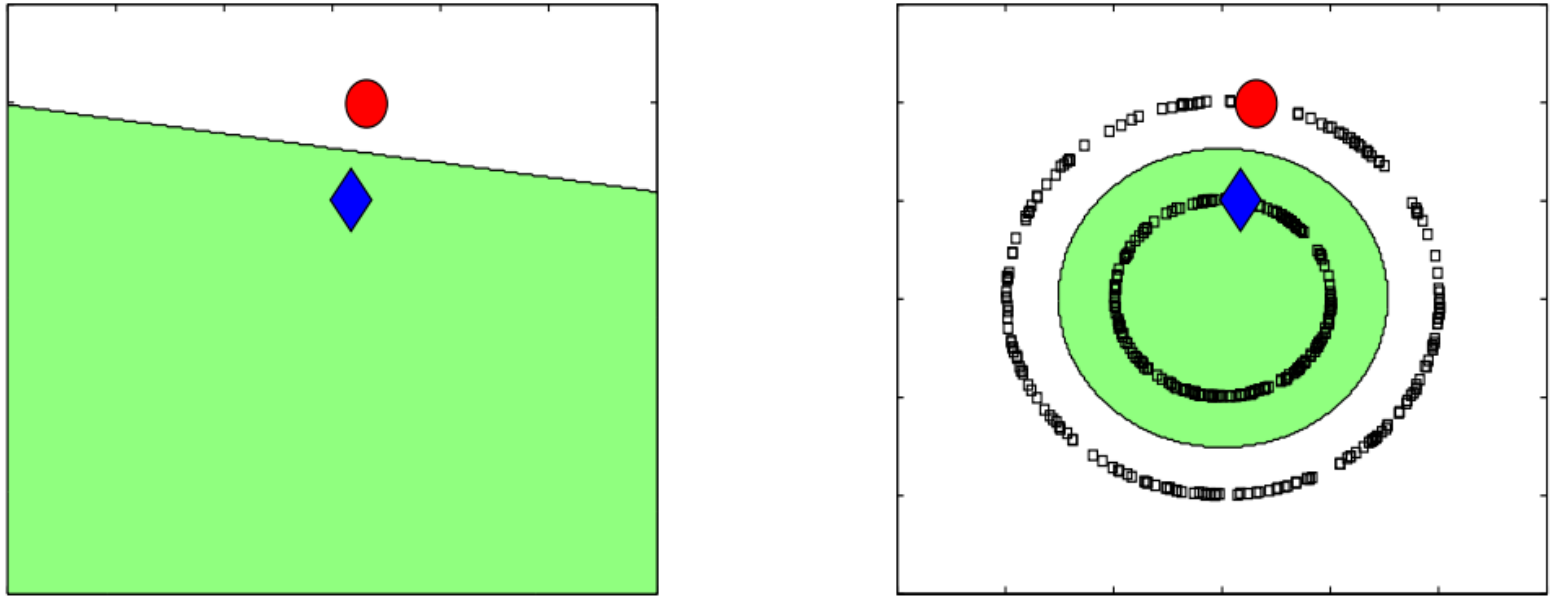


Figure 1: Unlabeled data and prior beliefs

http://people.cs.uchicago.edu/~niyogi/papersps/BNSJMLR.pdf

Traditional Method

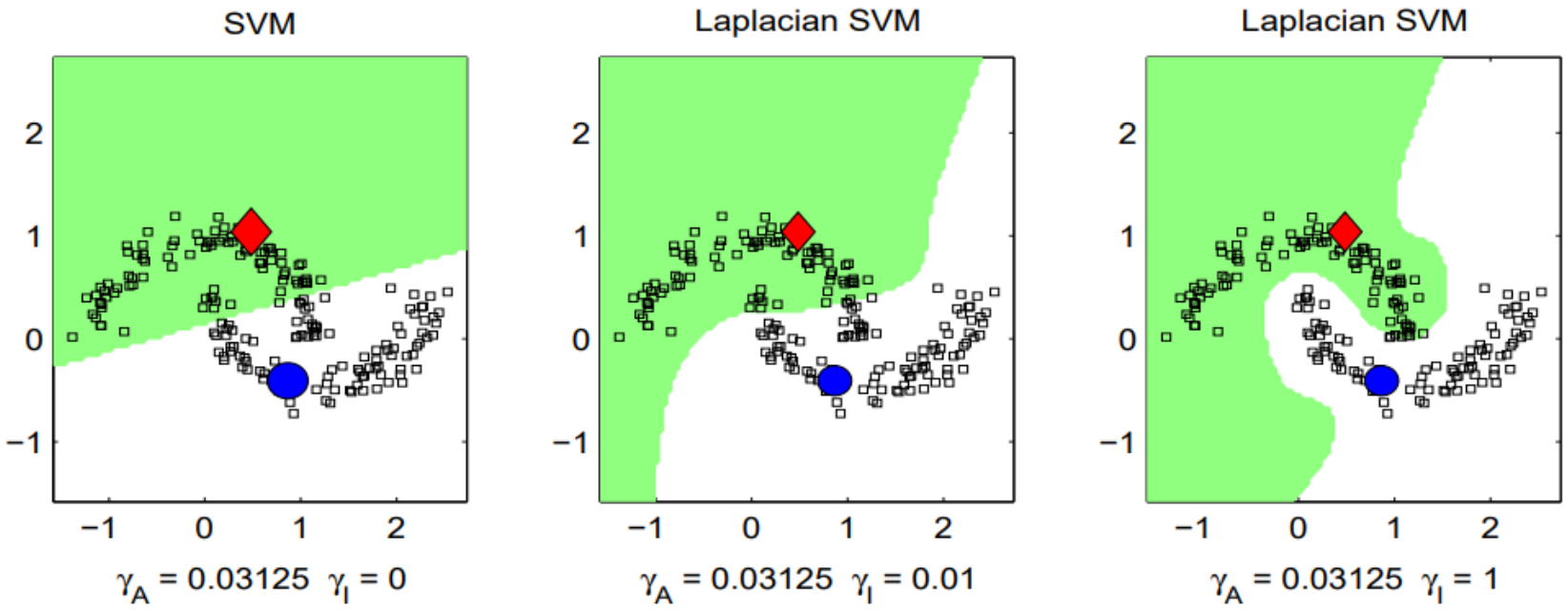


Figure 2: Laplacian SVM with RBF kernels for various values of γ_I. Labeled points are shown in color, other points are unlabeled.

http://people.cs.uchicago.edu/~niyogi/papersps/BNSJMLR.pdf

Major Categories

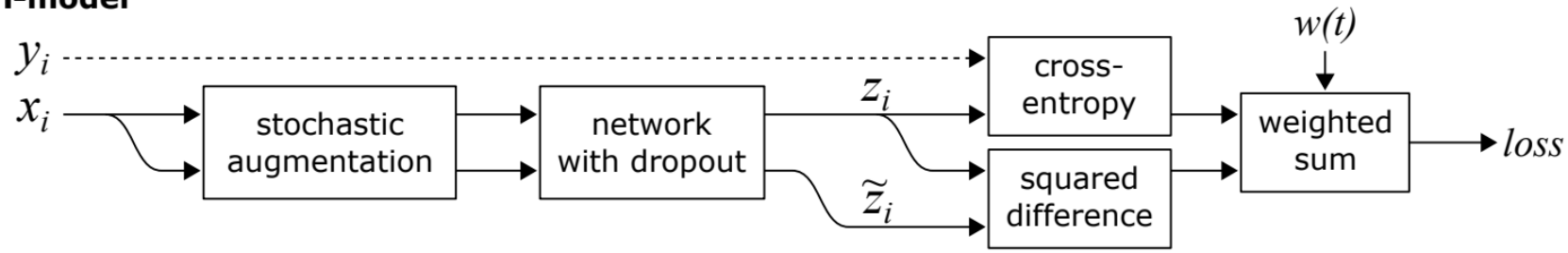
- Consistency Regularization
- Entropy-Based
- Pseudo-Labeling

Major Categories

- Consistency Regularization
- Entropy-Based
- Pseudo-Labeling

Π-MODEL

Π-model

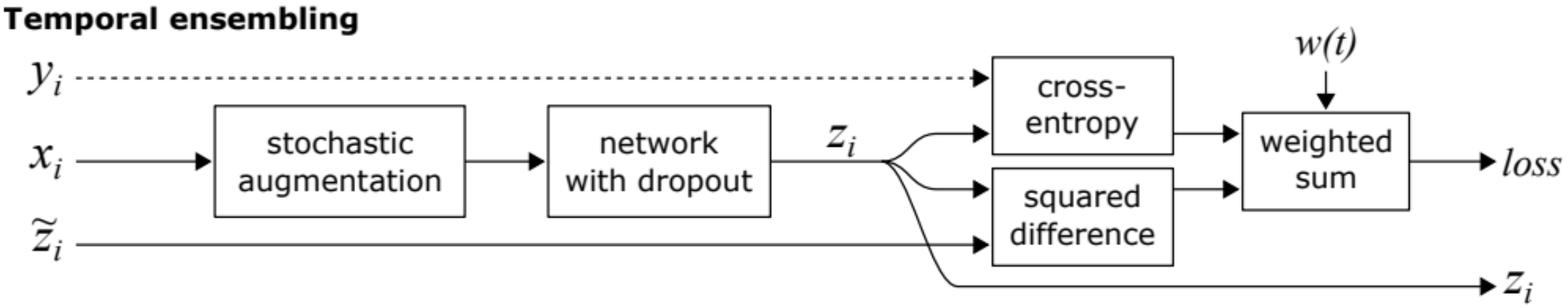


Algorithm 1 Π-model pseudocode.

Require: x_i = training stimuli
Require: L = set of training input indices with known labels
Require: y_i = labels for labeled inputs $i \in L$
Require: $w(t)$ = unsupervised weight ramp-up function
Require: $f_\theta(x)$ = stochastic neural network with trainable parameters θ
Require: $g(x)$ = stochastic input augmentation function
for t in $[1, num\_epochs]$ **do**
  for each minibatch B **do**
    $z_{i \in B} \leftarrow f_\theta(g(x_{i \in B}))$ ▷ evaluate network outputs for augmented inputs
    $\tilde{z}_{i \in B} \leftarrow f_\theta(g(x_{i \in B}))$ ▷ again, with different dropout and augmentation
    $loss \leftarrow -\frac{1}{|B|} \sum_{i \in (B \cap L)} \log z_i[y_i]$ ▷ supervised loss component
      $+ w(t) \frac{1}{C|B|} \sum_{i \in B} ||z_i - \tilde{z}_i||^2$ ▷ unsupervised loss component
    update θ using, e.g., ADAM ▷ update network parameters
  end for
end for
return θ

https://arxiv.org/pdf/1610.02242.pdf

Temporal-embedding

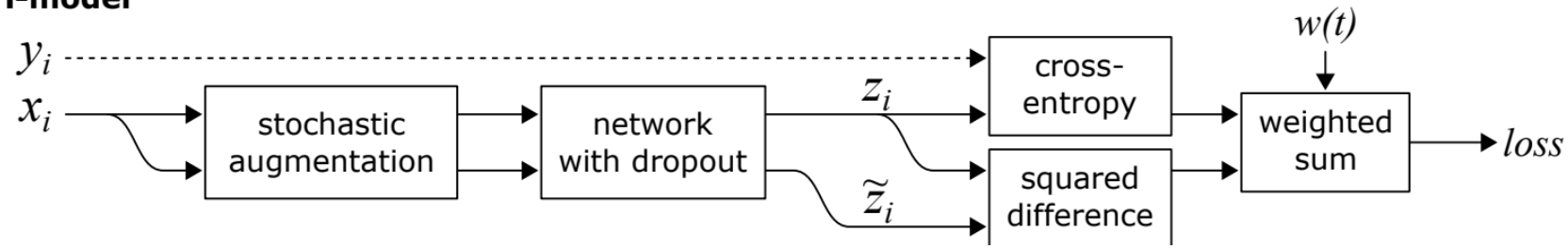


The structure of our temporal ensembling method is shown in Figure 1 (bottom), and the pseudocode in Algorithm 2. The main difference to the Π-model is that the network and augmentations are evaluated only once per input per epoch, and the target vectors $\tilde{z}$ for the unsupervised loss component are based on prior network evaluations instead of a second evaluation of the network.

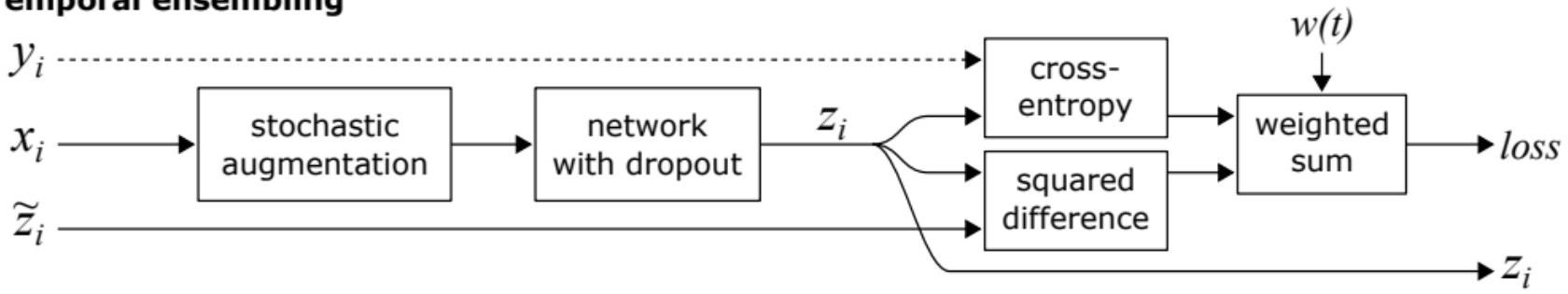
https://arxiv.org/pdf/1610.02242.pdf

Π-MODEL vs. Temporal-embedding

Π-model



Temporal ensembling



The benefits of temporal ensembling compared to Π-model are twofold. First, the training is faster because the network is evaluated only once per input on each epoch. Second, the training targets $\tilde{z}$ can be expected to be less noisy than with Π-model. As shown in Section 3, we indeed obtain

https://arxiv.org/pdf/1610.02242.pdf

Mean-teacher

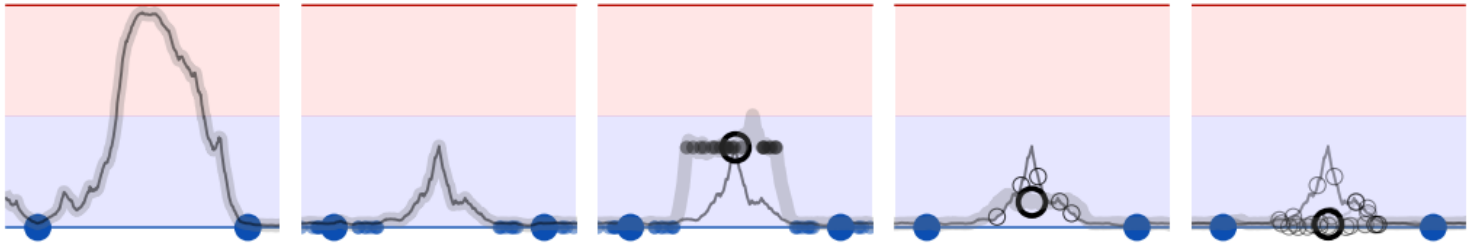


Figure 1: A sketch of a binary classification task with two labeled examples (large blue dots) and one unlabeled example, demonstrating how the choice of the unlabeled target (black circle) affects the fitted function (gray curve). **(a)** A model with no regularization is free to fit any function that predicts the labeled training examples well. **(b)** A model trained with noisy labeled data (small dots) learns to give consistent predictions around labeled data points. **(c)** Consistency to noise around unlabeled examples provides additional smoothing. For the clarity of illustration, the teacher model (gray curve) is first fitted to the labeled examples, and then left unchanged during the training of the student model. Also for clarity, we will omit the small dots in figures d and e. **(d)** Noise on the teacher model reduces the bias of the targets without additional training. The expected direction of stochastic gradient descent is towards the mean (large blue circle) of individual noisy targets (small blue circles). **(e)** An ensemble of models gives an even better expected target. Both Temporal Ensembling and the Mean Teacher method use this approach.

https://arxiv.org/pdf/1703.01780.pdf

Mean-teacher

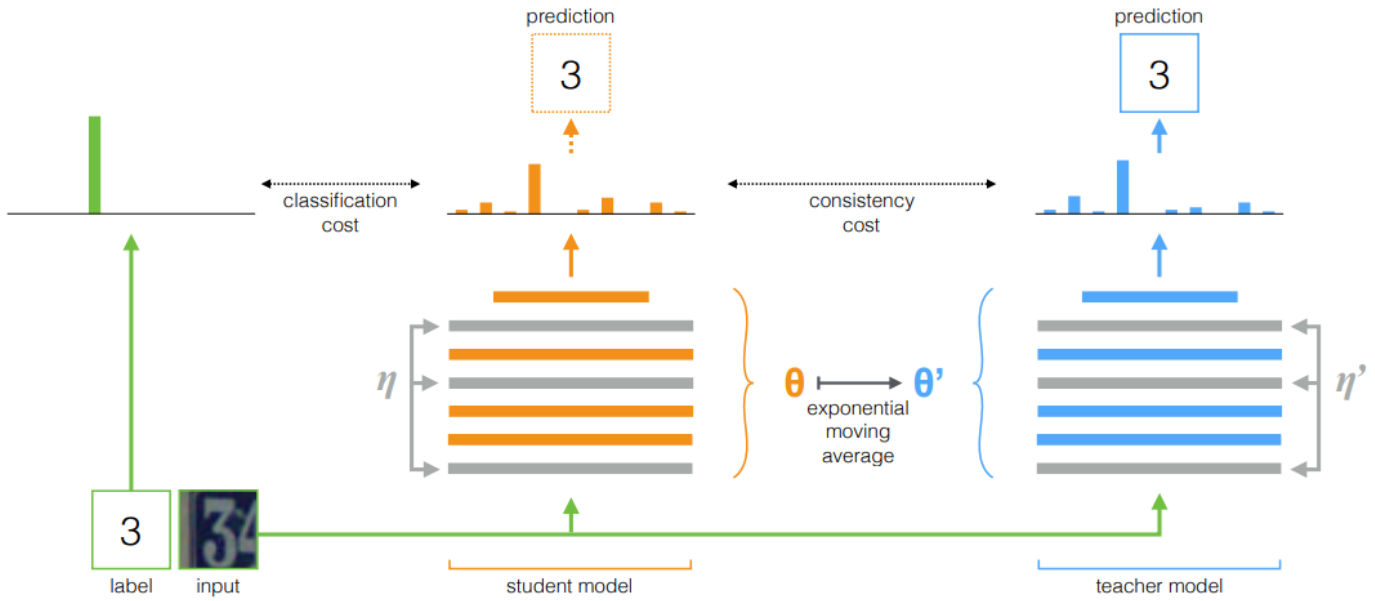


Figure 2: The Mean Teacher method. The figure depicts a training batch with a single labeled example. Both the student and the teacher model evaluate the input applying noise (η, η') within their computation. The softmax output of the student model is compared with the one-hot label using classification cost and with the teacher output using consistency cost. After the weights of the student model have been updated with gradient descent, the teacher model weights are updated as an exponential moving average of the student weights. Both model outputs can be used for prediction, but at the end of the training the teacher prediction is more likely to be correct. A training step with an unlabeled example would be similar, except no classification cost would be applied.

https://arxiv.org/pdf/1703.01780.pdf

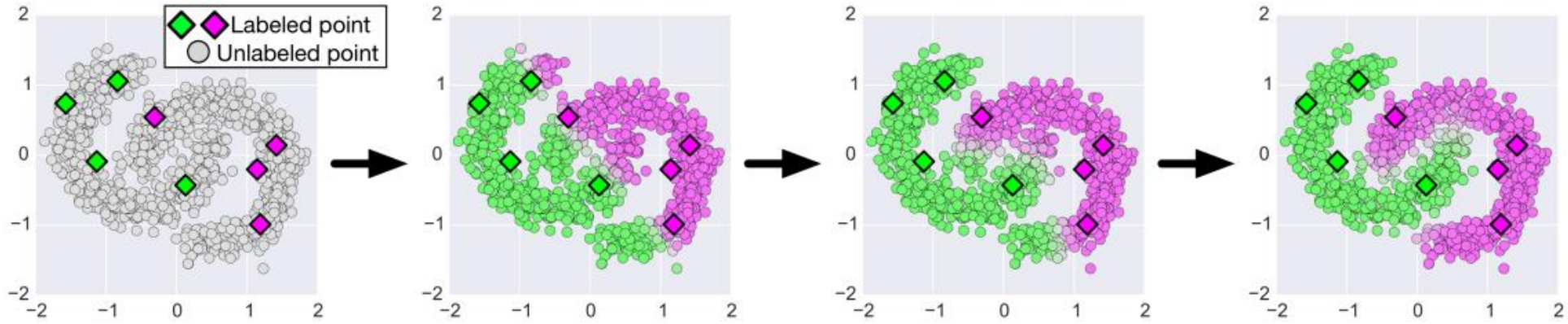
Virtual Adversarial Training (VAT)

Virtual Adversarial Training (VAT) (Miyato et al., 2017) directly approximates a tiny perturbation r_{adv} to add to x which would most significantly affect the output of the prediction function. The perturbation can be computed

$$r \sim \mathcal{N}\left(0, \frac{\xi}{\sqrt{\dim(x)}} I\right) \tag{1}$$

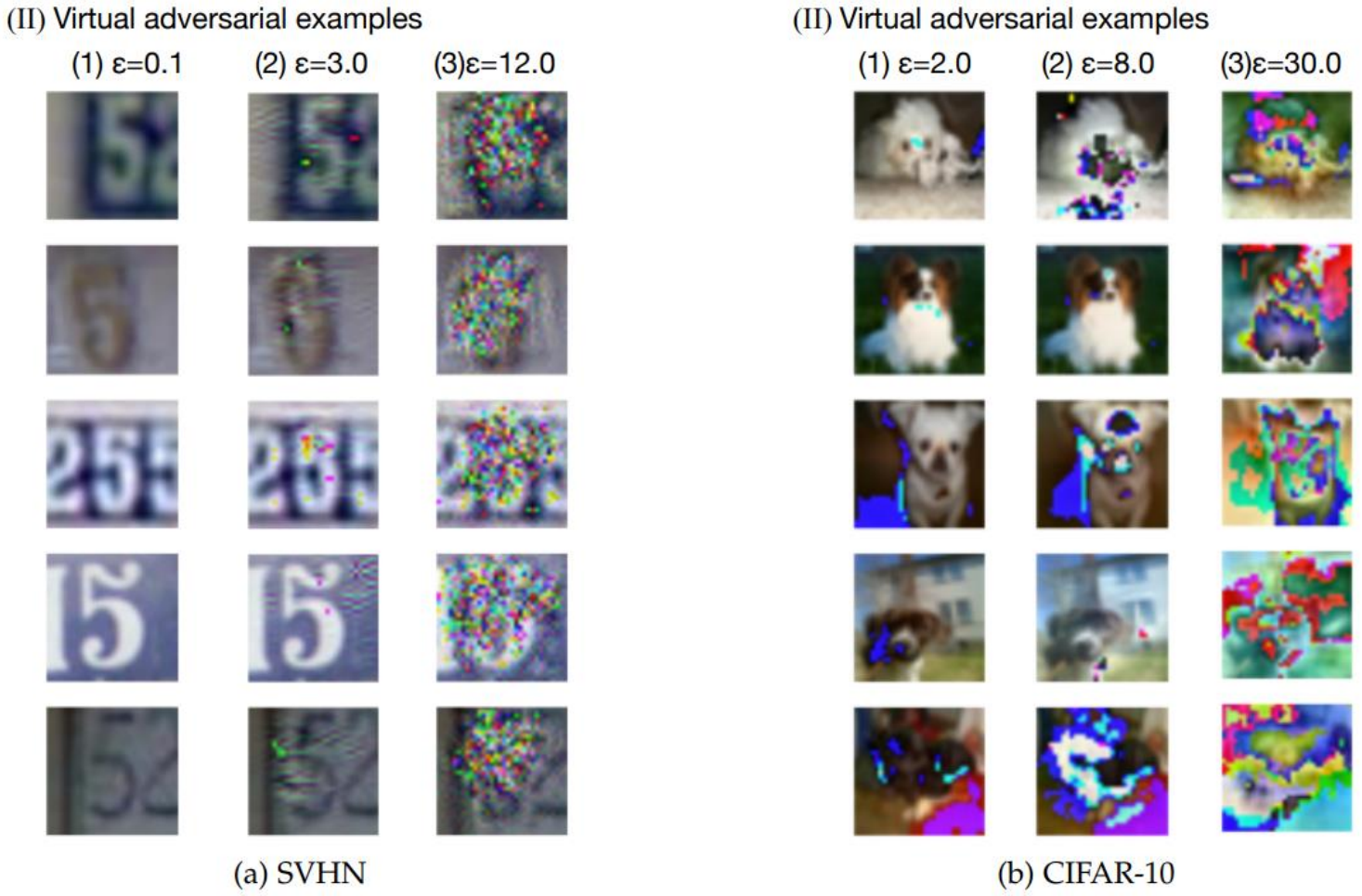
$$g = \nabla_r d(f_\theta(x), f_\theta(x + r)) \tag{2}$$

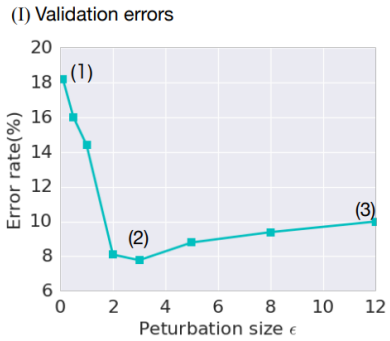
$$r_{adv} = \epsilon \frac{g}{||g||} \tag{3}$$

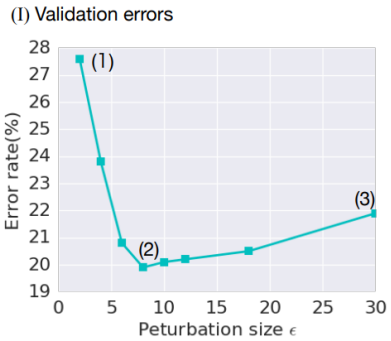


https://arxiv.org/pdf/1804.09170.pdf

VAT







https://arxiv.org/pdf/1704.03976.pdf

Major Categories

- Consistency Regularization
- Entropy-Based
- Pseudo-Labeling

Penalizing low entropy output distributions

A simple loss term which can be applied to unlabeled data is to encourage the network to make "confident" (low-entropy) predictions for all examples, regardless of the actual class predicted. Assuming a categorical output space with K possible classes (e.g. a K-dimensional softmax output), this gives rise to the "entropy minimization" term (Grandvalet & Bengio, 2005):

$$-\sum_{k=1}^{K} f_\theta(x)_k \log f_\theta(x)_k \qquad (4)$$

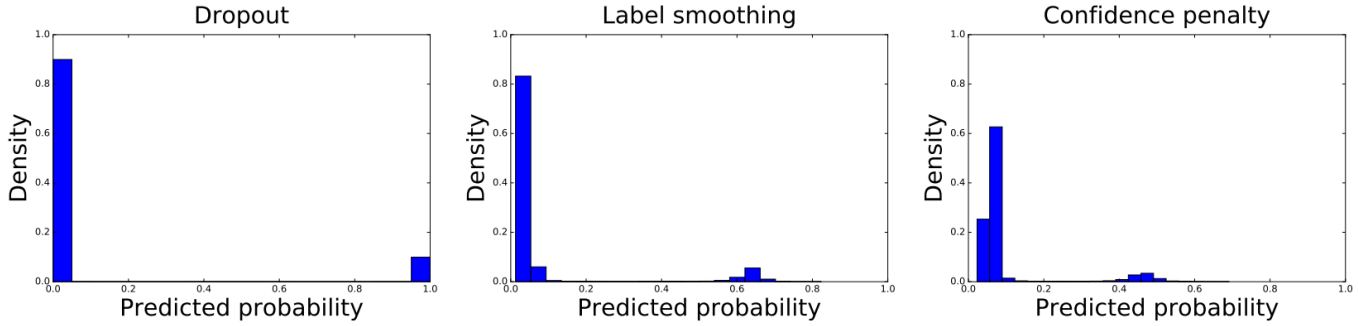


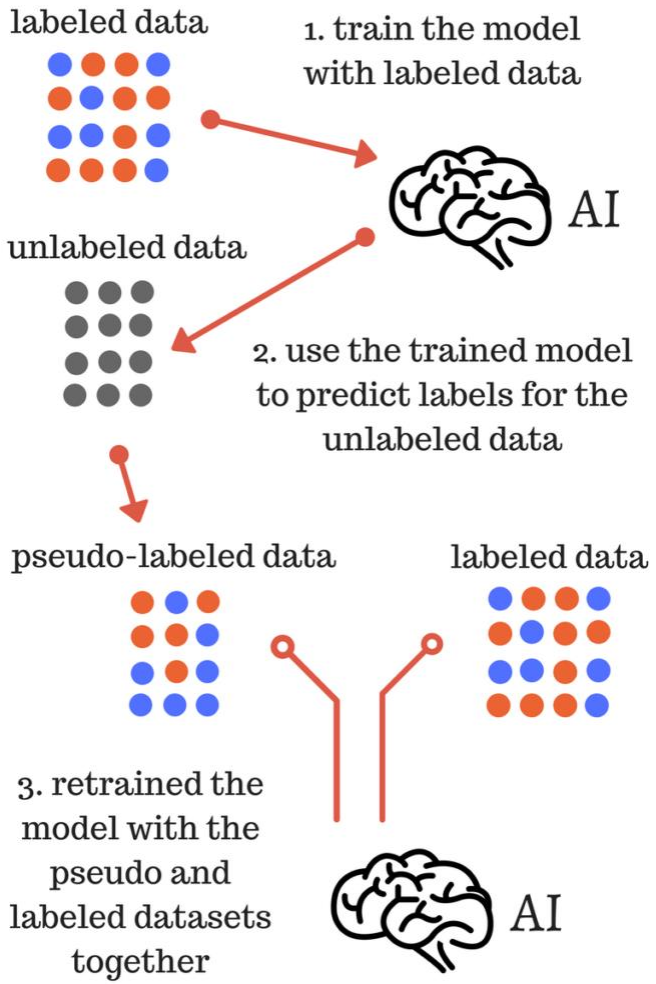
Figure 1: Distribution of the magnitude of softmax probabilities on the MNIST validation set. A fully-connected, 2-layer, 1024-unit neural network was trained with dropout (left), label smoothing (center), and the confidence penalty (right). Dropout leads to a softmax distribution where probabilities are either 0 or 1. By contrast, both label smoothing and the confidence penalty lead to smoother output distributions, which results in better generalization.

https://arxiv.org/pdf/1701.06548.pdf

Major Categories

- Consistency Regularization
- Entropy-Based
- Pseudo-Labeling

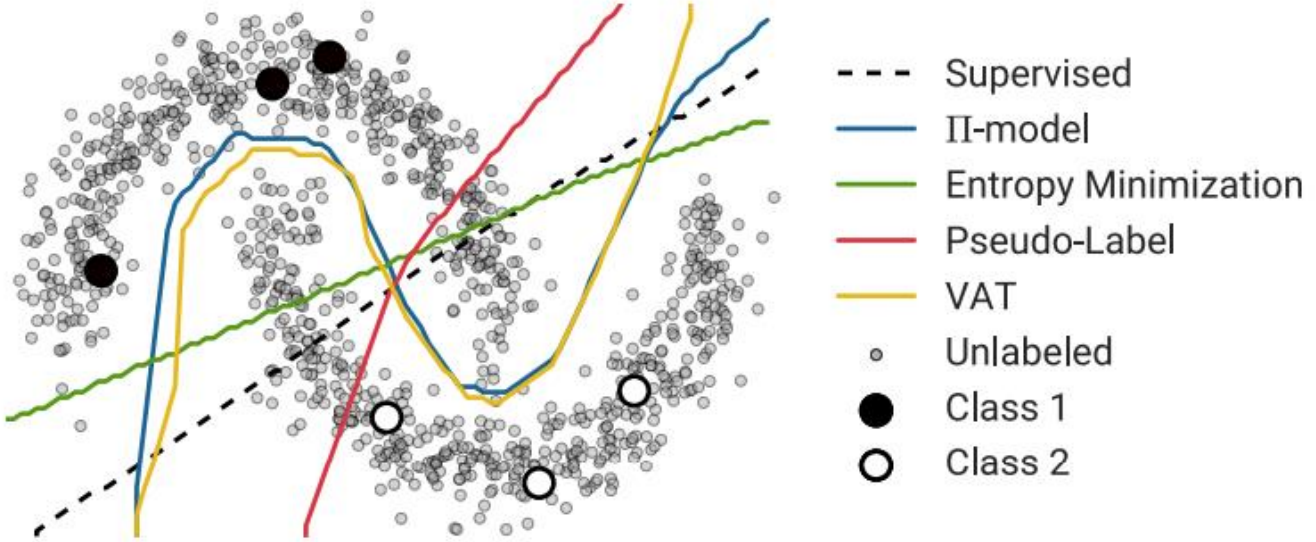
Pseudo-labeling



https://datawhatnow.com/pseudo-labeling-semi-supervised-learning/

Summarization

- Consistency Regularization
  - Π-MODEL
  - Temporal-embedding
  - Mean-teacher
- Entropy-Based
- Pseudo-Labeling



https://arxiv.org/pdf/1804.09170.pdf

Using Two Strategies

MixMatch: A Holistic Approach to Semi-Supervised Learning

David Berthelot
Google Research
dberth@google.com

Nicholas Carlini
Google Research
ncarlini@google.com

Ian Goodfellow
Work done at Google
ian-academic@mailfence.com

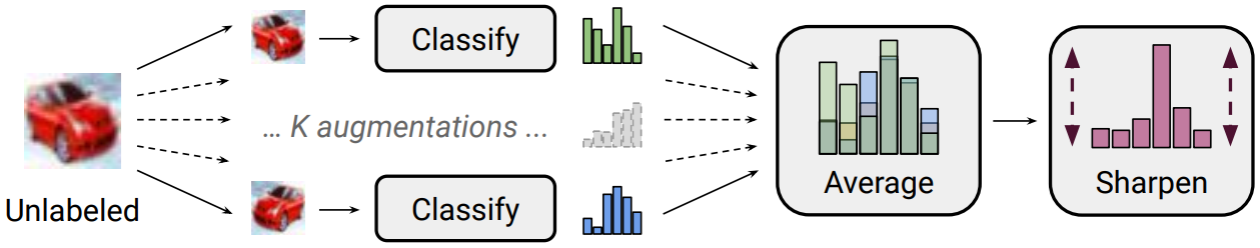
Avital Oliver
Google Research
avitalo@google.com

Nicolas Papernot
Google Research
papernot@google.com

Colin Raffel
Google Research
craffel@google.com

Abstract

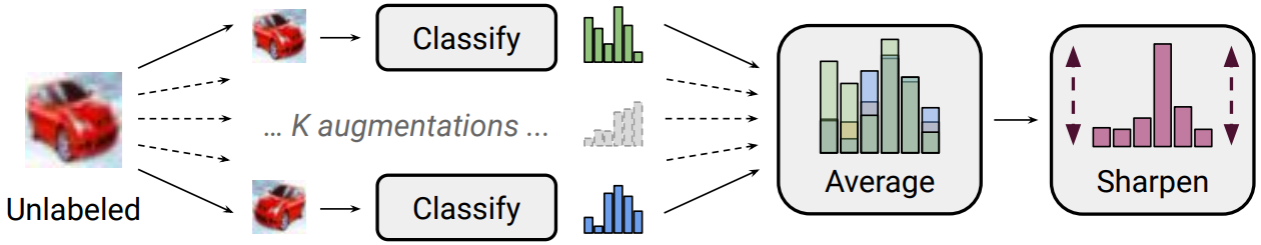
Semi-supervised learning has proven to be a powerful paradigm for leveraging unlabeled data to mitigate the reliance on large labeled datasets. In this work, we unify the current dominant approaches for semi-supervised learning to produce a new algorithm, MixMatch, that guesses low-entropy labels for data-augmented unlabeled examples and mixes labeled and unlabeled data using MixUp. MixMatch obtains state-of-the-art results by a large margin across many datasets and labeled data amounts. For example, on CIFAR-10 with 250 labels, we reduce error rate by a factor of 4 (from 38% to 11%) and by a factor of 2 on STL-10. We also demonstrate how MixMatch can help achieve a dramatically better accuracy-privacy trade-off for differential privacy. Finally, we perform an ablation study to tease apart which components of MixMatch are most important for its success. We release all code

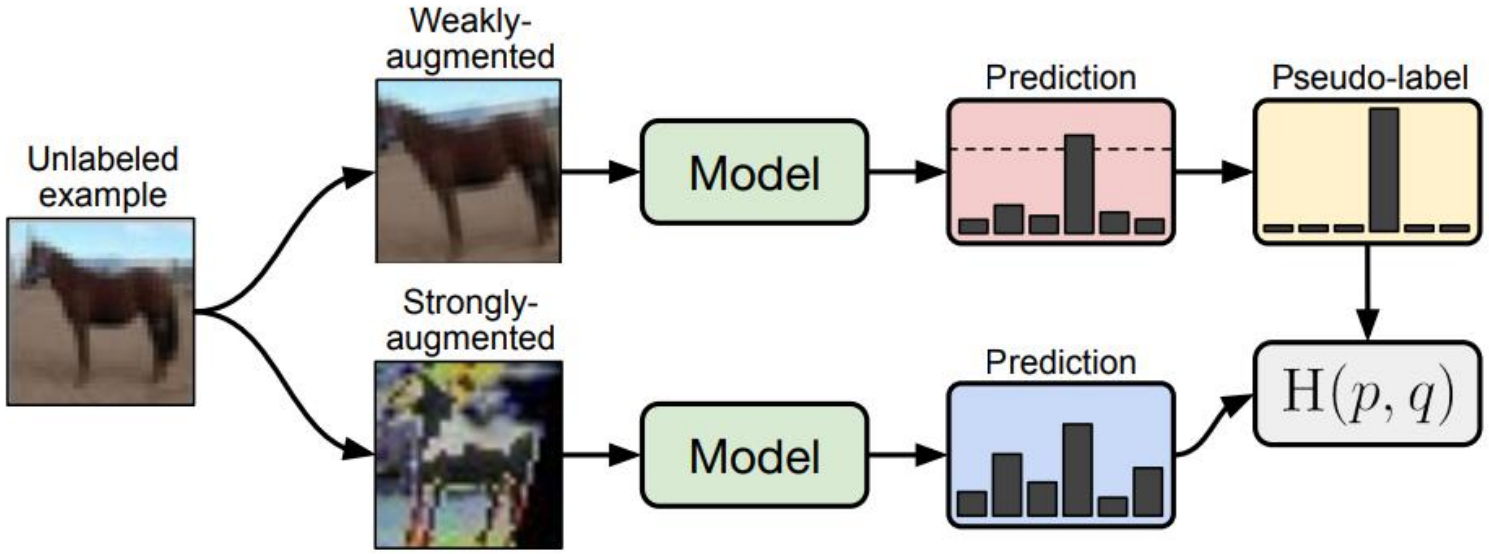


Using Three Strategies

FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence

Kihyuk Sohn* David Berthelot* Chun-Liang Li Zizhao Zhang Nicholas Carlini
Ekin D. Cubuk Alex Kurakin Han Zhang Colin Raffel
Google Research





Skin Lesion Classification in Dermoscopy Images Using Synergic Deep Learning

Jianpeng Zhang[1], Yutong Xie[1], Qi Wu[2], and Yong Xia[1(✉)]

[1] School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, People's Republic of China
yxia@nwpu.edu.cn
School of Computer Science, University of Adelaide, Adelaide, SA 5005, Australia

Abstract. Automated skin lesion classification in the dermoscopy images is an essential way to improve diagnostic performance and reduce melanoma deaths. Although deep learning has shown proven advantages over traditional methods, which rely on handcrafted features, in image classification, it remains challenging to classify skin lesions due to the significant intra-class variation and inter-class similarity. In this paper, we propose a synergic deep learning (SDL) model to address this issue, which not only uses dual deep convolutional neural networks (DCNNs) but also enables them to mutually learn from each other. Specifically, we concatenate the image representation learned by both DCNNs as the input of a synergic network, which has a fully connected structure and predicts whether the pair of input images belong to the same class. We train the SDL model in the end-to-end manner under the supervision of the classification error in each DCNN and the synergic error. We evaluated our SDL model on the ISIC 2016 Skin Lesion Classification dataset and achieved the state-of-the-art performance.

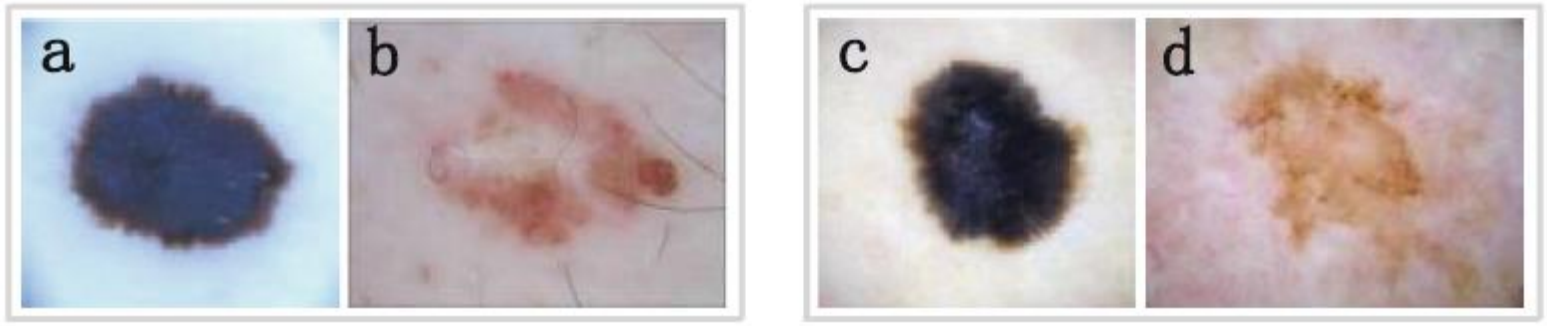


Fig. 1. Examples show the intra-class variation and inter-class similarity in skin lesion classification: (a, b) benign skin lesions, and (c, d) malignant skin lesions

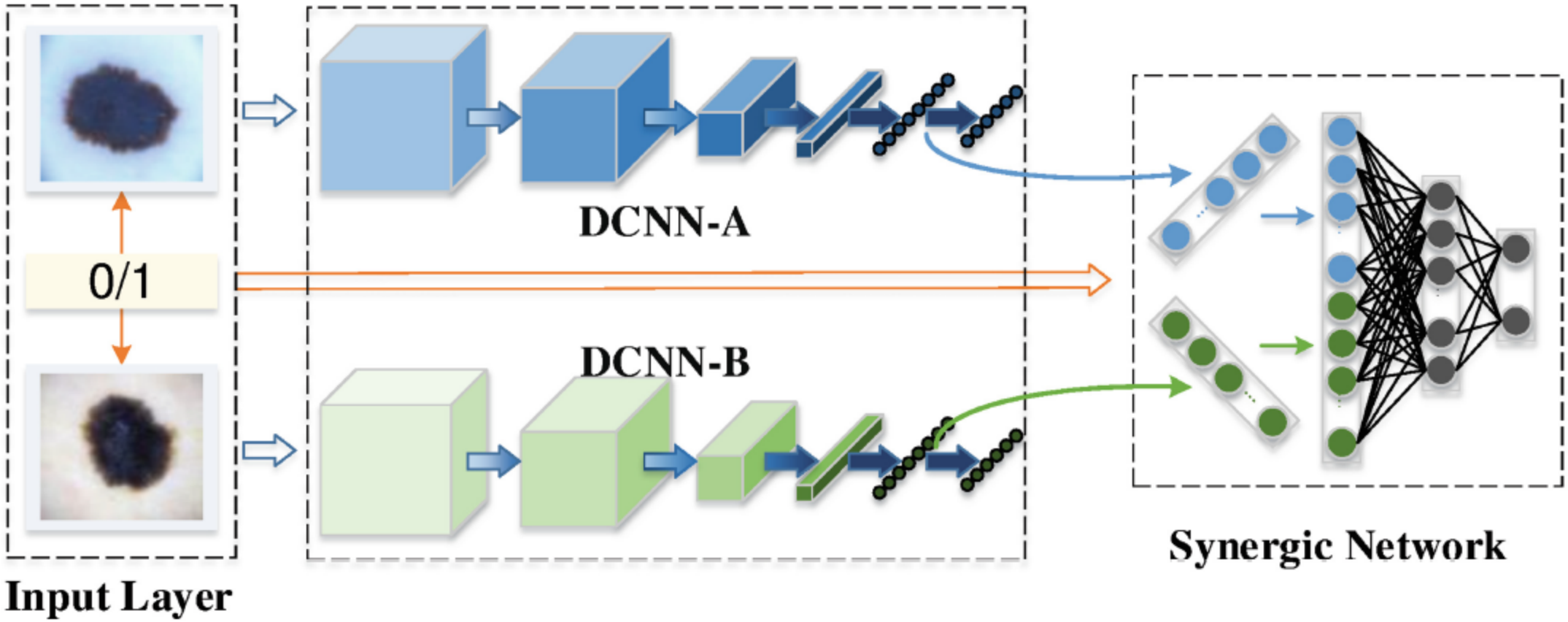
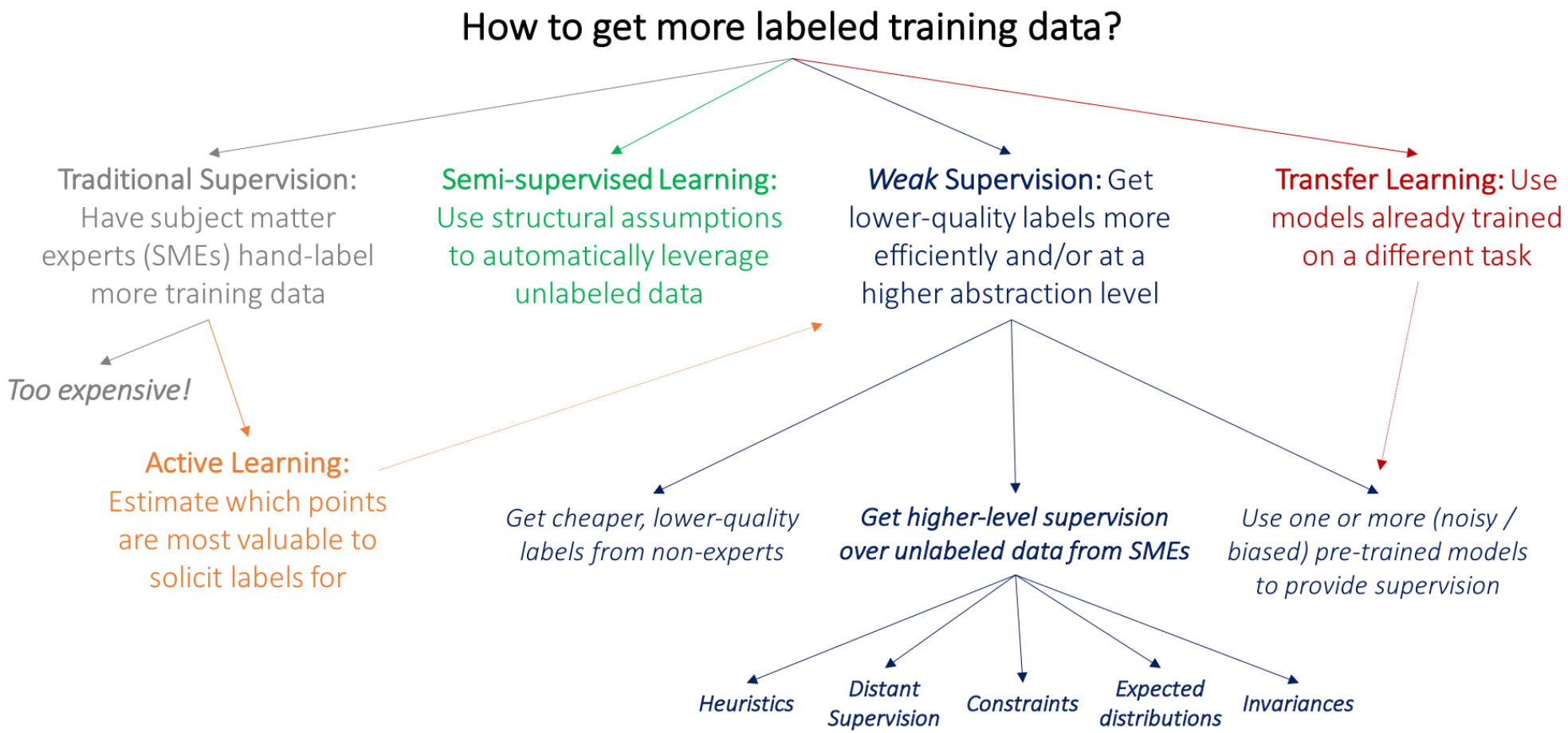


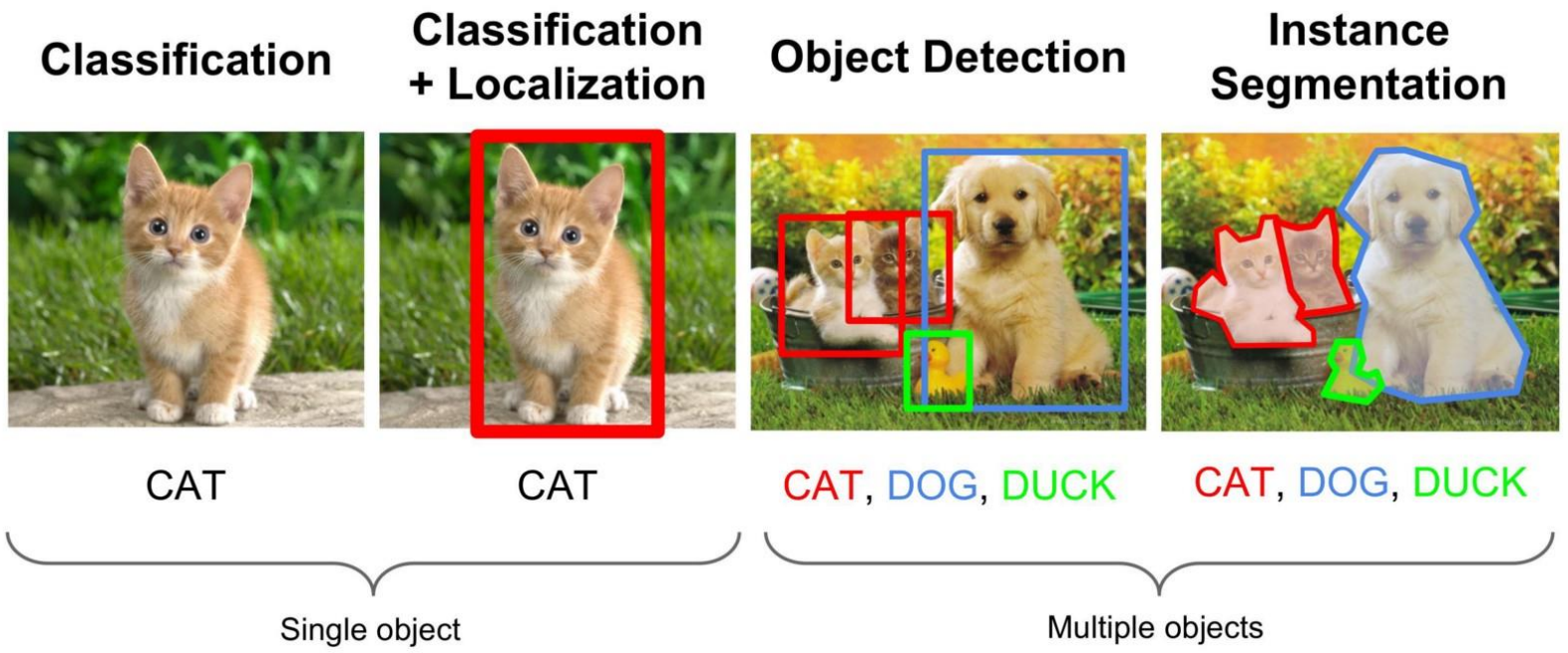
Fig. 2. Architecture of the proposed SDL model which has an input layer, dual DCNN components (DCNN-A/B) and a synergic network.

Weakly-supervised Learning



Major Categories

- Use classification for localization
- Use localization for segmentation

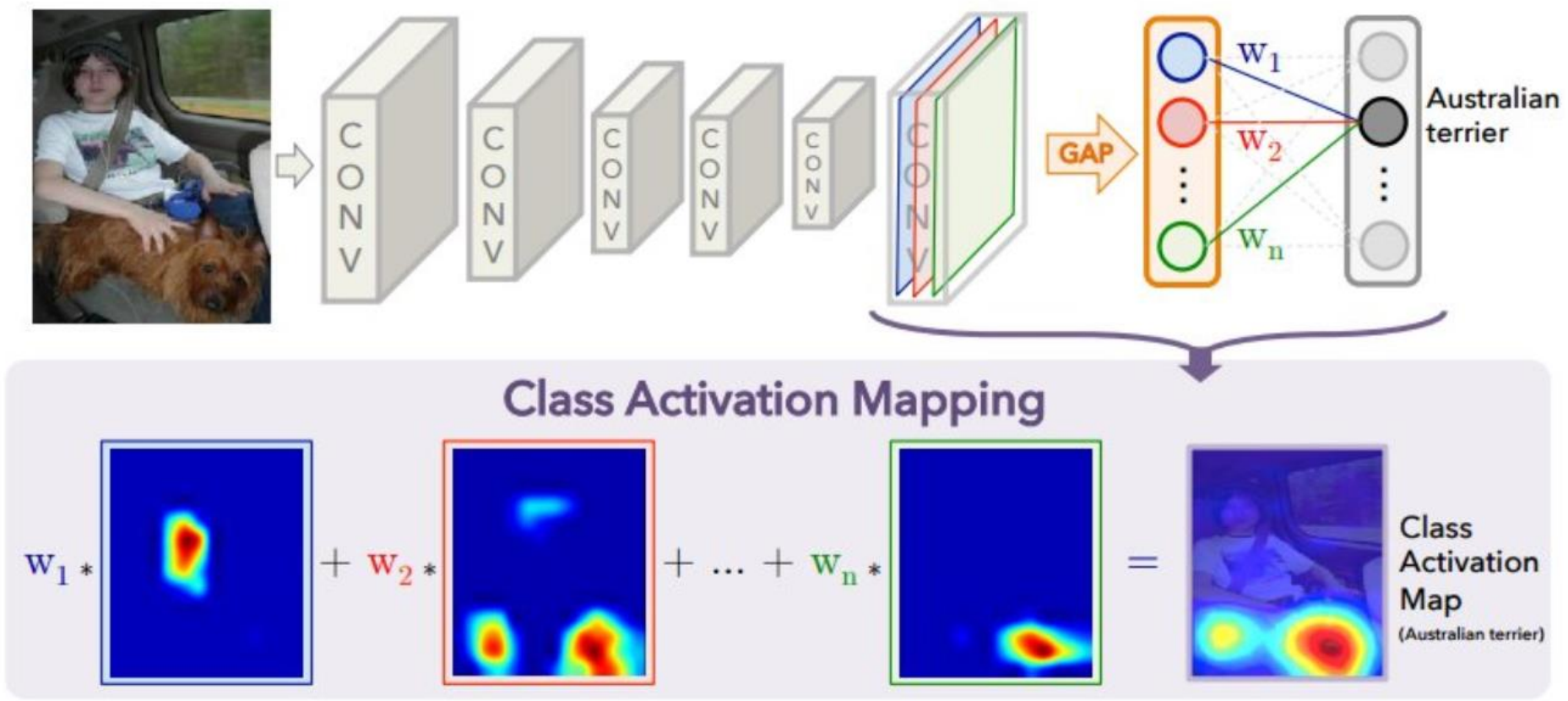


https://medium.com/comet-app/review-of-deep-learning-algorithms-for-object-detection-c1f3d437b852

Classification for Localization

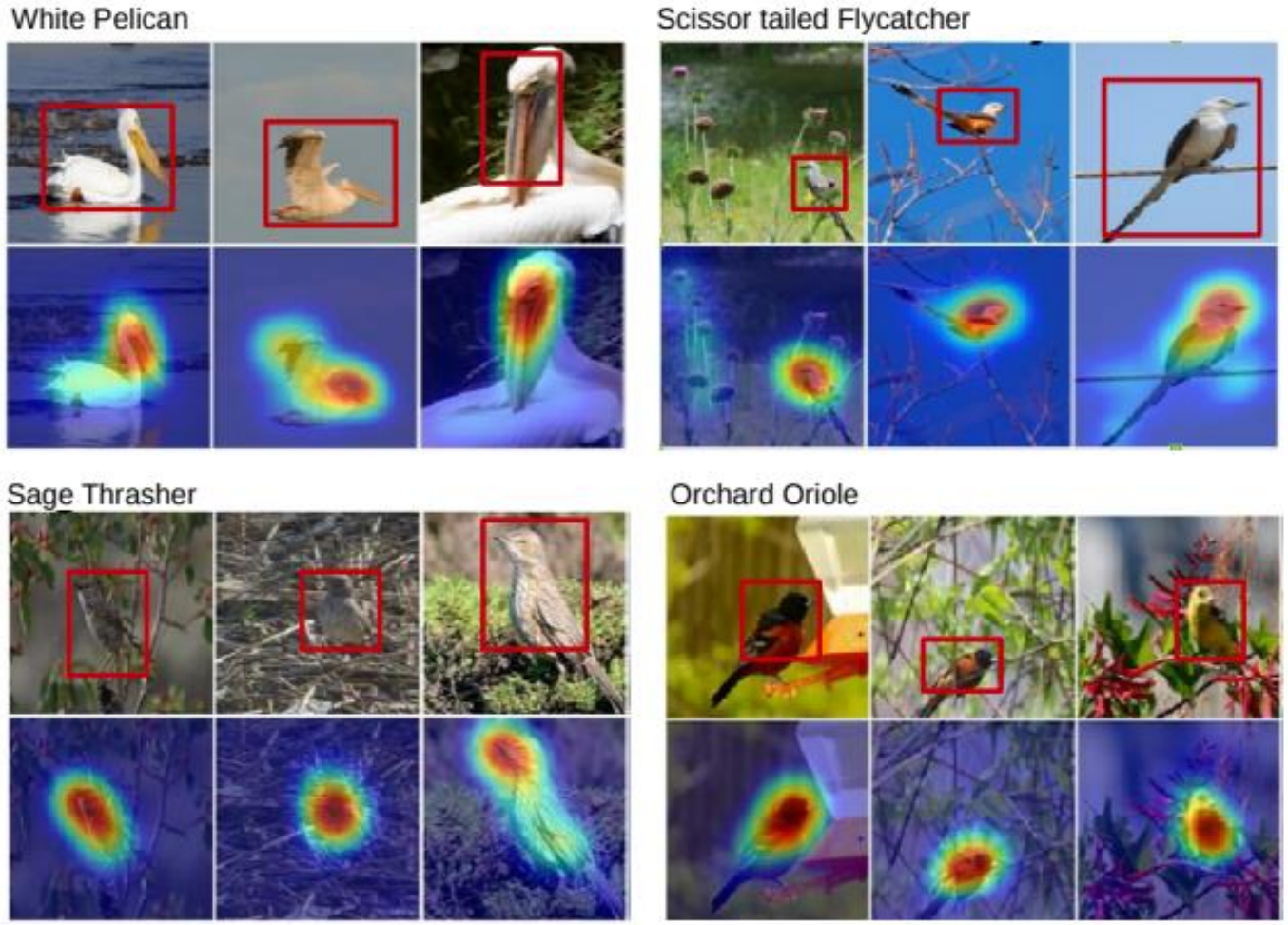
http://cnnlocalization.csail.mit.edu/

Attention



http://cnnlocalization.csail.mit.edu/

Classification for Localization

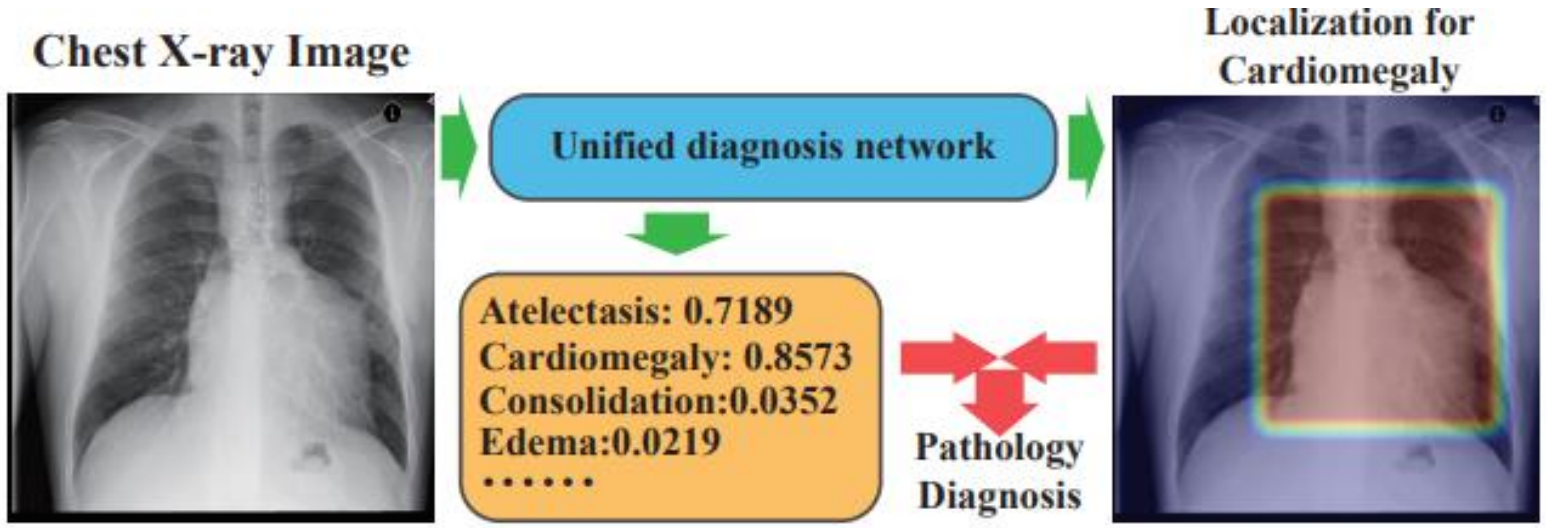


http://cnnlocalization.csail.mit.edu/

Medical Imaging

Thoracic Disease Identification and Localization with Limited Supervision

Zhe Li[1]*, Chong Wang[3], Mei Han[2]*, Yuan Xue[3], Wei Wei[3], Li-Jia Li[3], Li Fei-Fei[3]
[1]Syracuse University, [2]PingAn Technology, US Research Lab, [3]Google Inc.
[1]zli89@syr.edu, [2]ex-hanmei001@pingan.com.cn,
[3]{chongw, yuanxue, wewei, lijiali, feifeili}@google.com



http://openaccess.thecvf.com/content_cvpr_2018/papers/Li_Thoracic_Disease_Identification_CVPR_2018_paper.pdf

Medical Imaging

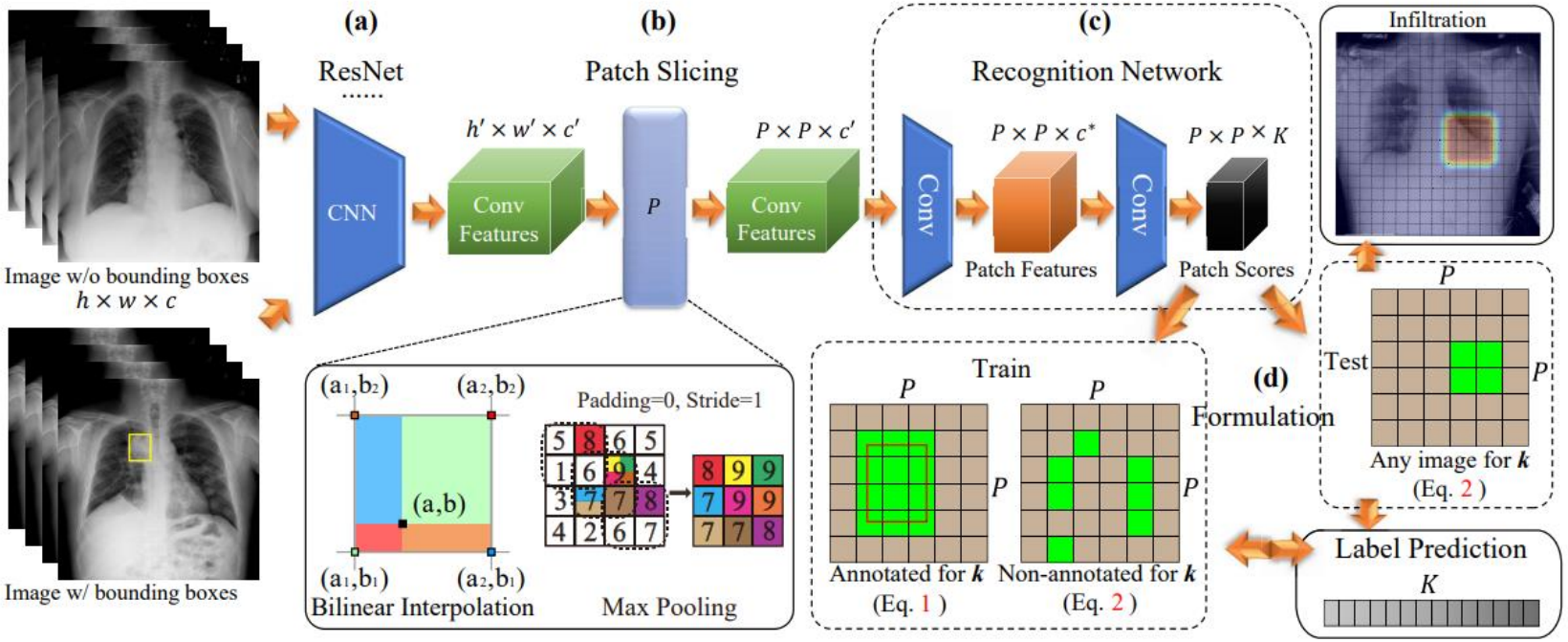
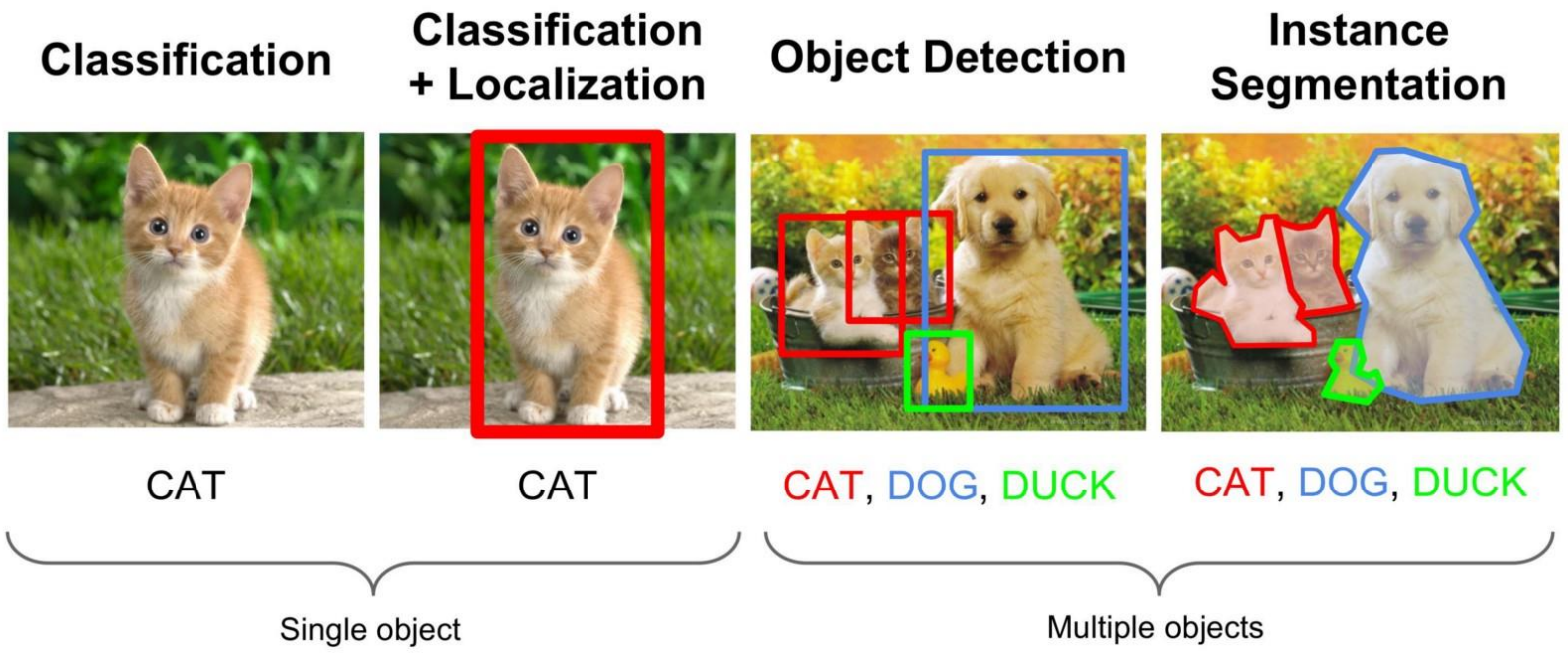


Figure 3. Model overview. (a) The input image is firstly processed by a CNN. (b) The patch slicing layer resizes the convolutional features from the CNN using max-pooling or bilinear interpolation. (c) These regions are then passed to a fully-convolutional recognition network. (d) During training, we use multi-instance learning assumption to formulate two types of images; during testing, the model predicts both labels and class-specific localizations. The red frame represents the ground truth bounding box. The green cells represent patches with positive labels, and brown is negative. Please note during training, for unannotated images, we assume there is at least one positive patch and the green cells shown in the figure are not deterministic.

http://openaccess.thecvf.com/content_cvpr_2018/papers/Li_Thoracic_Disease_Identification_CVPR_2018_paper.pdf

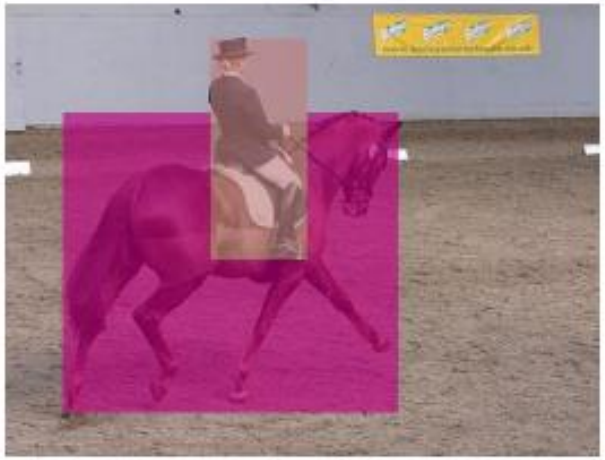
Major Categories

- Use classification for localization
- Use localization for segmentation



https://medium.com/comet-app/review-of-deep-learning-algorithms-for-object-detection-c1f3d437b852

Weakly Supervised Instance and Semantic Segmentation



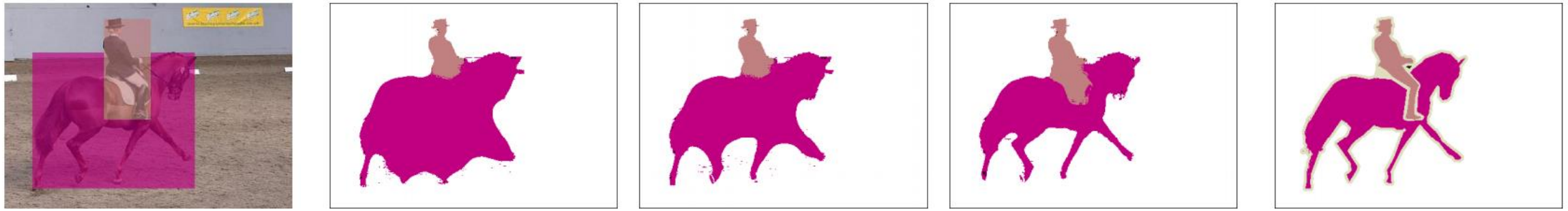
Training sample, with box annotations

Test image, fully supervised result

Test image, weakly supervised result

https://arxiv.org/pdf/1603.07485.pdf

Strategy



Example input rectangles | Output after 1 training round | After 5 rounds | After 10 rounds | Ground truth

Figure 2: Example results of using only rectangle segments and recursive training (using convnet predictions as supervision for the next round), see Section 3.1.

https://arxiv.org/pdf/1603.07485.pdf

Medical Imaging

Deep Learning Based Instance Segmentation in 3D Biomedical Images Using Weak Annotation

Zhuo Zhao[1], Lin Yang[1], Hao Zheng[1], Ian H. Guldner[2], Siyuan Zhang[2], and Danny Z. Chen[1(✉)]

Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA
dchen@nd.edu
[2] Department of Biological Sciences, Harper Cancer Research Institute, University of Notre Dame, Notre Dame, IN 46556, USA

Abstract. Instance segmentation in 3D images is a fundamental task in biomedical image analysis. While deep learning models often work well for 2D instance segmentation, 3D instance segmentation still faces critical challenges, such as insufficient training data due to various annotation difficulties in 3D biomedical images. Common 3D annotation methods (e.g., full voxel annotation) incur high workloads and costs for labeling enough instances for training deep learning 3D instance segmentation models. In this paper, we propose a new weak annotation approach for training a fast deep learning 3D instance segmentation model without using full voxel mask annotation. Our approach needs only 3D bounding boxes for all instances and full voxel annotation for a small fraction of the instances, and uses a novel two-stage 3D instance segmentation model utilizing these two kinds of annotation, respectively. We evaluate our approach on several biomedical image datasets, and the experimental results show that (1) with full annotated boxes and a small amount of masks, our approach can achieve similar performance as the best known methods using full annotation, and (2) with similar annotation time, our approach outperforms the best known methods that use full annotation.

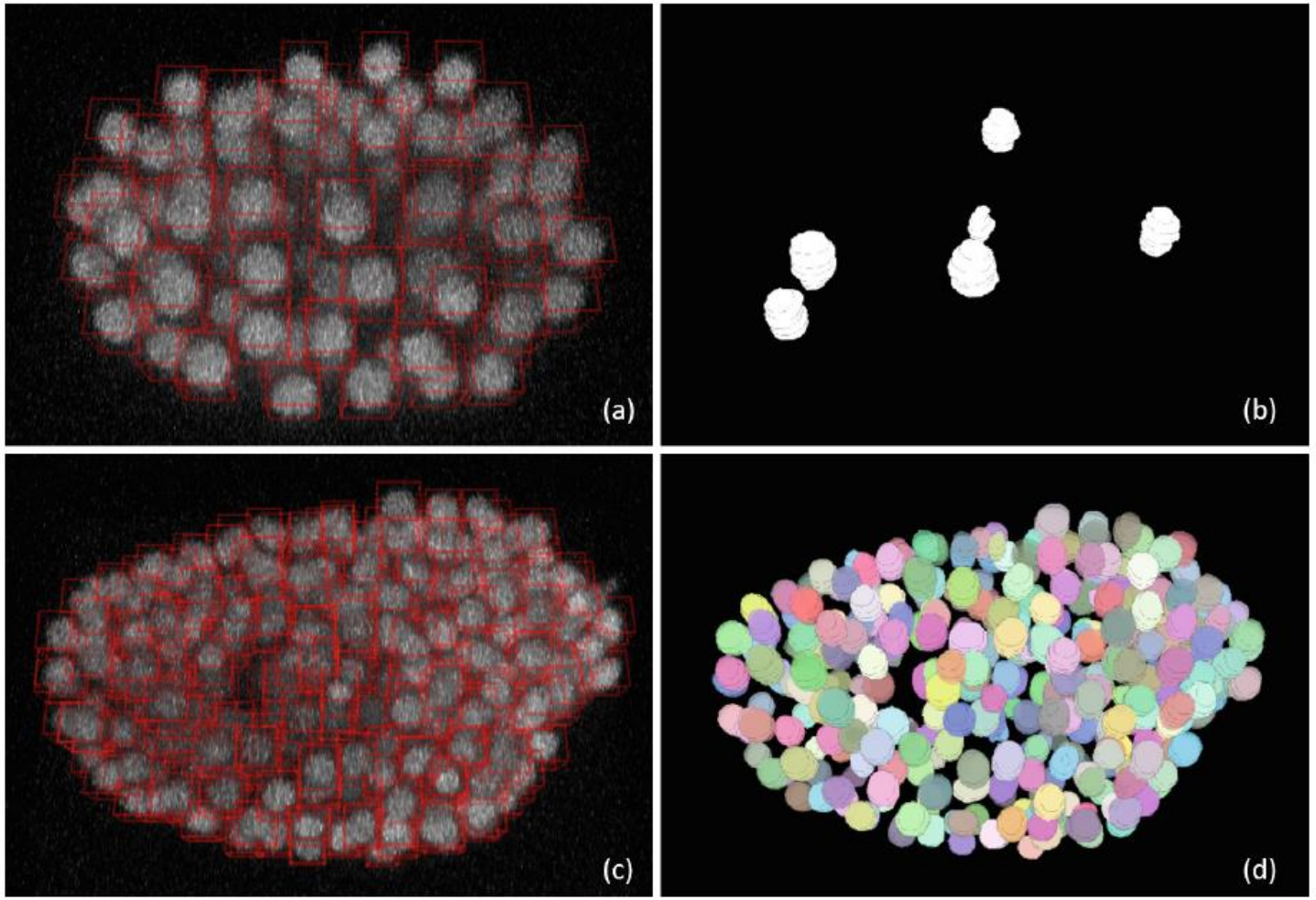


Fig. 1. A training image example and test results for C.elegans developing embryos [8]. (a) Each object instance is labeled by a 3D bounding box; (b) a small fraction of instances are labeled with voxel mask annotation; (c) in stage one, our model uses full box annotation to detect all instances; (d) in stage two, it uses full voxel mask annotation for a small fraction of the instances to segment each detected instance.

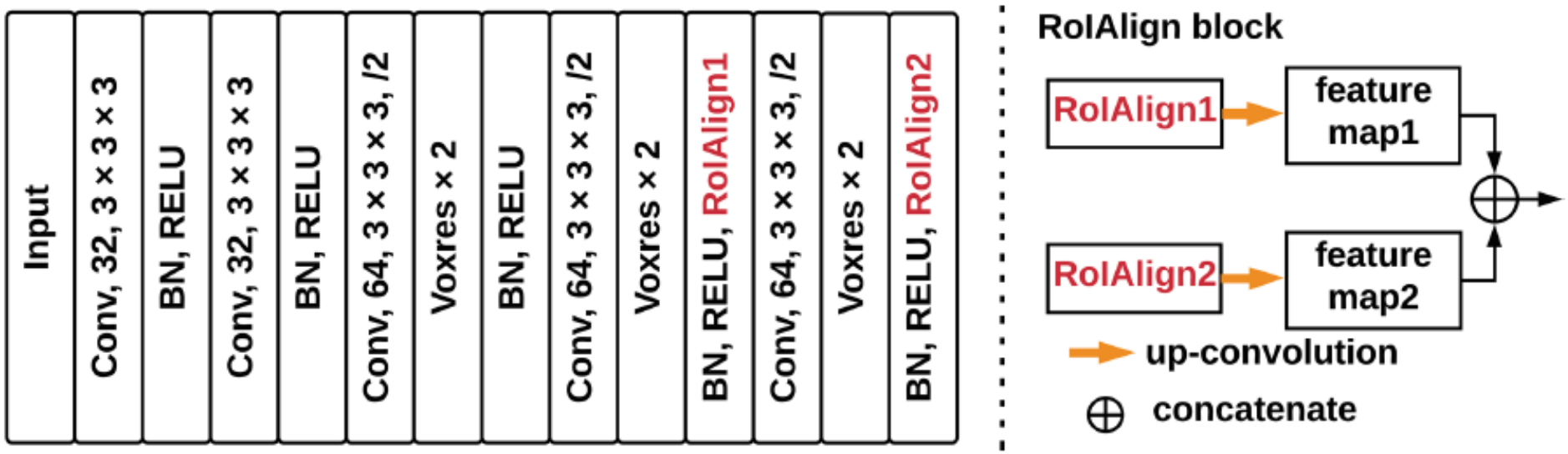


Fig. 2. The backbone and RoIAlign block of the model. RoIAlign is applied to two layers of the backbone. The feature maps after up-convolution are concatenated as the final feature maps.

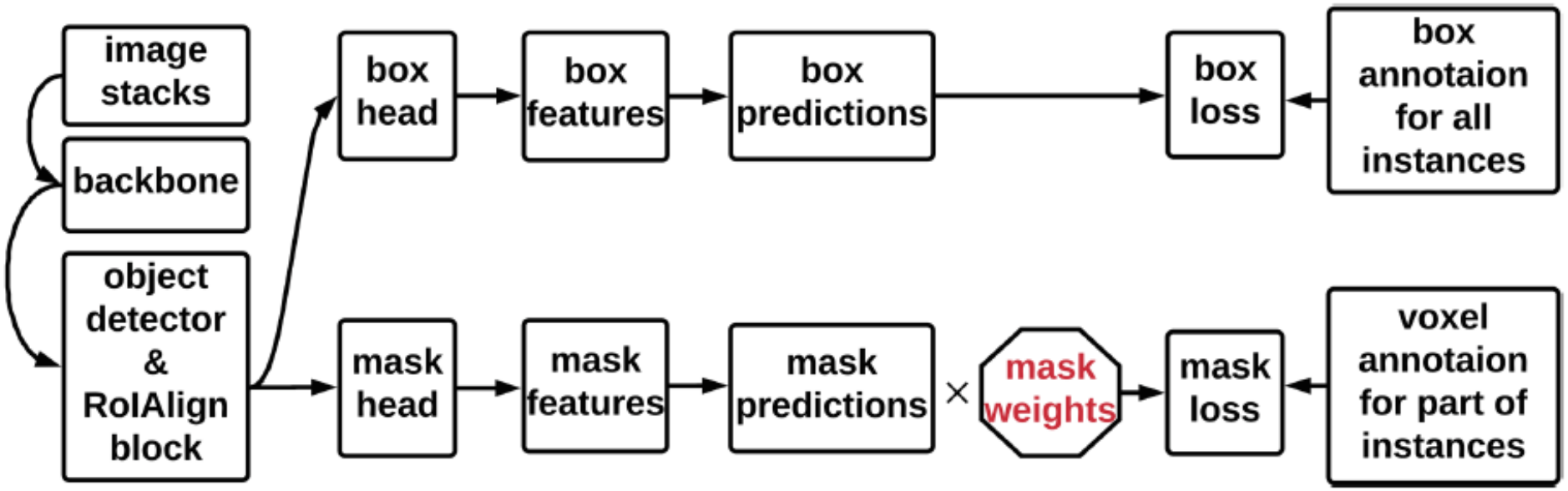


Fig. 3. Illustrating the flow of our method. All boxes contribute to the object detector; only the instances with voxel annotation contribute to the voxel segmentation model.

Learn Papers

Omni-Supervised Learning: Scaling Up to Large Unlabelled Medical Datasets

Ruobing Huang(✉), J. Alison Noble, and Ana I. L. Namburete

Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK
ruobing.huang@eng.ox.ac.uk

Abstract. Two major bottlenecks in increasing algorithmic performance in the field of medical imaging analysis are the typically limited size of datasets and the shortage of expert labels for large datasets. This paper investigates approaches to overcome the latter via *omni-supervised learning*: a special case of semi-supervised learning. Our approach seeks to exploit a small annotated dataset and iteratively increase model performance by scaling up to refine the model using a large set of unlabelled data. By fusing predictions of perturbed inputs, the method generates new training annotations without human intervention. We demonstrate the effectiveness of the proposed framework to localize multiple structures in a 3D US dataset of 4044 fetal brain volumes with an initial expert annotation of just 200 volumes (5% in total) in training. Results show that structure localization error was reduced from 2.07 ± 1.65 mm to 1.76 ± 1.35 mm on the hold-out validation set.

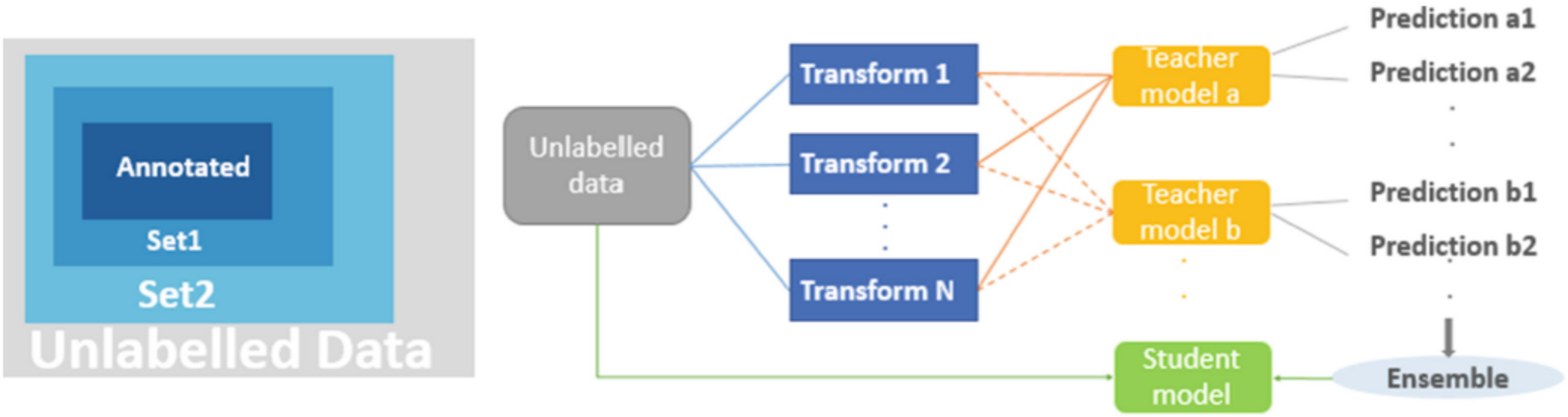


Fig. 1. Schematic of the proposed framework. The figure on the left shows the framework starts from a small annotated subset, to train the base models, and gradually expands to the full unlabelled set. The flow-chart shows unlabelled data are transformed to generate multiple of copies, and are sent to different base models for evaluation. The predictions are aggregated to generate new labels to train student models.

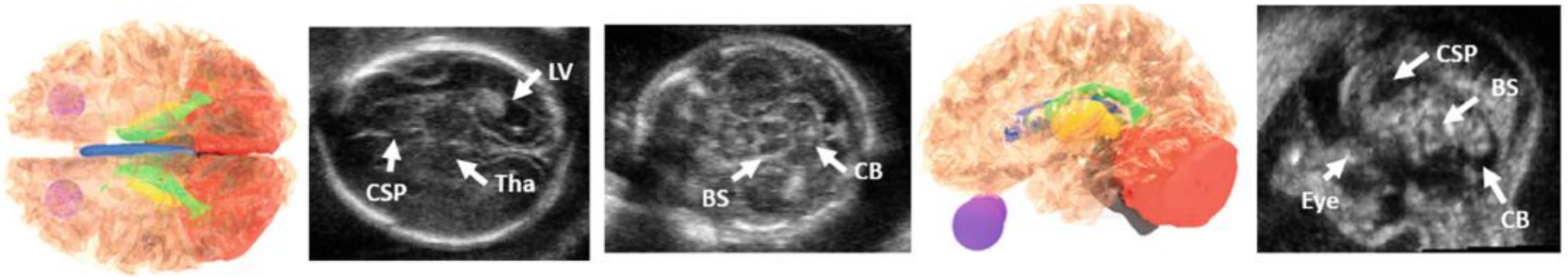


Fig. 2. Key brain structures. Schematics of CSP (blue), LV (green), TH (yellow), CB (red) and BS (grey), Eye (purple) are shown in axial and sagittal views. Examples of the structures shown in an US volume are displayed accordingly.

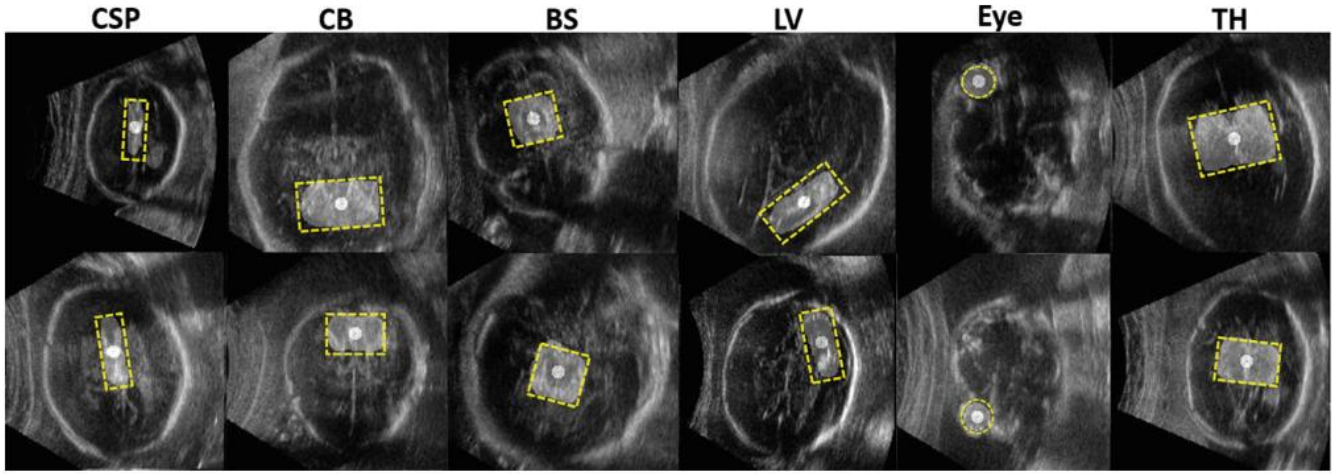


Fig. 3. Target structures viewed on US slices from random subjects (ground truth - yellow box, prediction - transparent overlay). The centre of each structure is plotted as a white dot. The image contrast, fetal head size, and orientation, vary dramatically. Speckle and acoustic shadows also influence structure visibility.

Semi-supervised Learning for Segmentation Under Semantic Constraint

Pierre-Antoine Ganaye, Michaël Sdika(✉), and Hugues Benoit-Cattin

Univ Lyon, INSA-Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm CREATIS UMR 5220, U1206, 69100 Lyon, France

Abstract. Image segmentation based on convolutional neural networks is proving to be a powerful and efficient solution for medical applications. However, the lack of annotated data, presence of artifacts and variability in appearance can still result in inconsistencies during the inference. We choose to take advantage of the invariant nature of anatomical structures, by enforcing a semantic constraint to improve the robustness of the segmentation. The proposed solution is applied on a brain structures segmentation task, where the output of the network is constrained to satisfy a known adjacency graph of the brain regions. This criteria is introduced during the training through an original penalization loss named NonAdjLoss. With the help of a new metric, we show that the proposed approach significantly reduces abnormalities produced during the segmentation. Additionally, we demonstrate that our framework can be used in a semi-supervised way, opening a path to better generalization to unseen data.

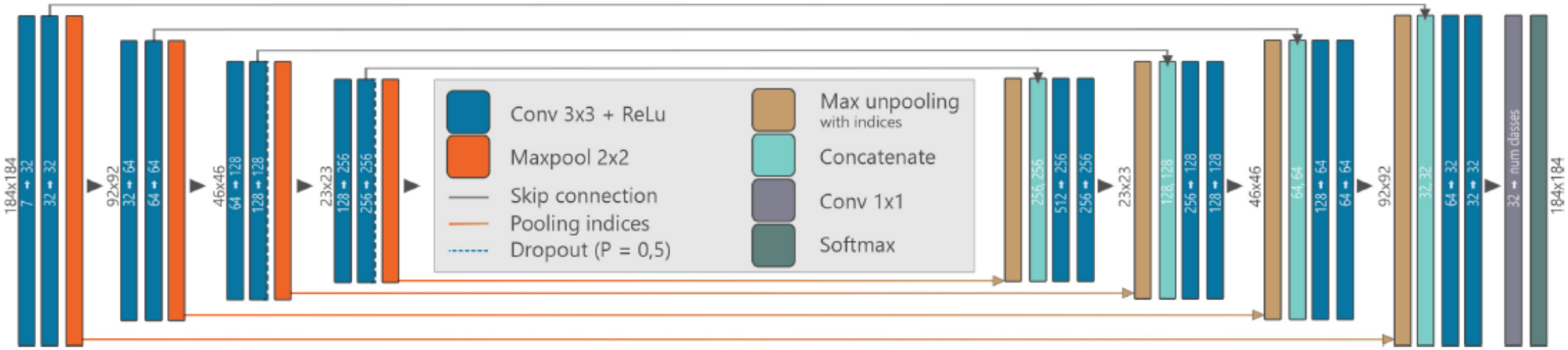


Fig. 1. Encoder-decoder architecture

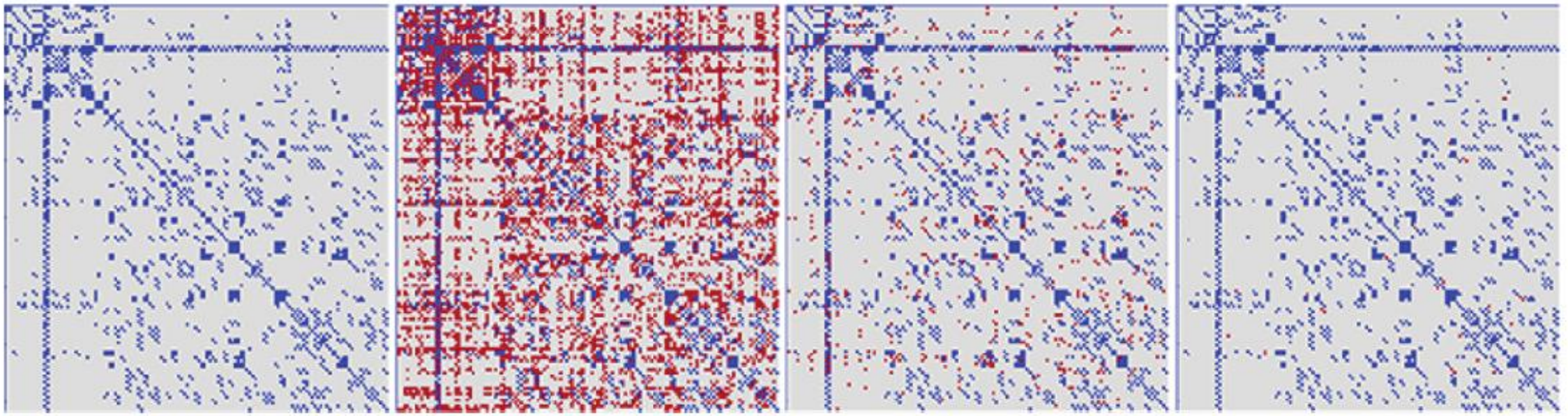


Fig. 2. Binary adjacency graphs matrices. Blue shows correct connections, red shows impossible adjacencies. From left to right: ground truth $\tilde{A}$ for the MICCAI12 dataset, after training without constraint, after training with NonAdjLoss, after semi-supervised training of NonAdjLoss with 100 images.

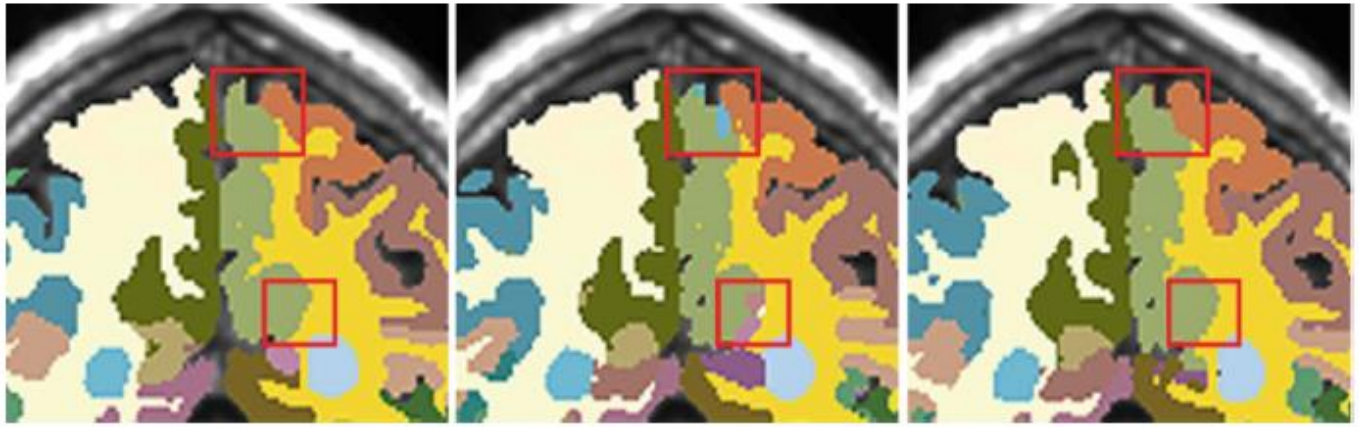


Fig. 3. Segmentation maps of one patient for: ground truth, model without loss, model with Non-AdjLoss (from left to right). Red boxes highlight areas where inconsistencies were corrected.

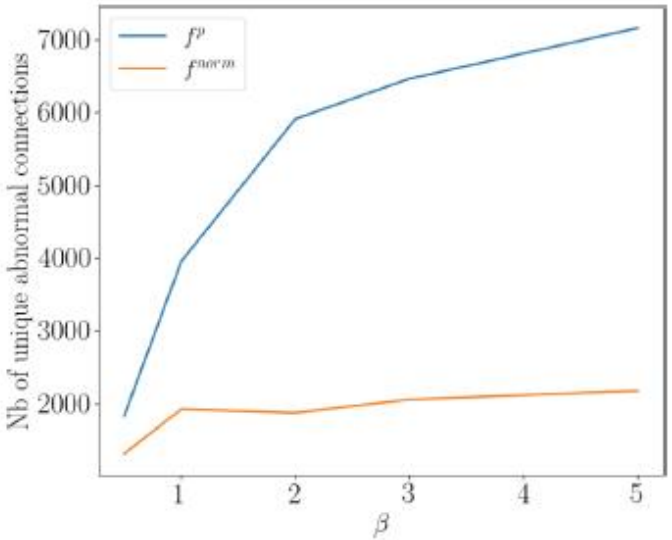


Fig. 4. Number of unique abnormal connections as a function of f^p, f^{norm} and β.

Improving Cytoarchitectonic Segmentation of Human Brain Areas with Self-supervised Siamese Networks

Hannah Spitzer[1(✉)], Kai Kiwitz[2], Katrin Amunts[1,2], Stefan Harmeling[3], and Timo Dickscheid[1]

[1] Institute of Neuroscience and Medicine INM-1, Forschungszentrum Jülich, Jülich, Germany
h.spitzer@fz-juelich.de
[2] C. and O. Vogt Institute of Brain Research, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany
[3] Institute of Computer Science, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany

Abstract. Cytoarchitectonic parcellations of the human brain serve as anatomical references in multimodal atlas frameworks. They are based on analysis of cell-body stained histological sections and the identification of borders between brain areas. The de-facto standard involves a semi-automatic, reproducible border detection, but does not scale with high-throughput imaging in large series of sections at microscopical resolution. Automatic parcellation, however, is extremely challenging due to high variation in the data, and the need for a large field of view at microscopic resolution. The performance of a recently proposed Convolutional Neural Network model that addresses this problem especially suffers from the naturally limited amount of expert annotations for training. To circumvent this limitation, we propose to pre-train neural networks on a self-supervised auxiliary task, predicting the 3D distance between two patches sampled from the same brain. Compared to a random initialization, fine-tuning from these networks results in significantly better segmentations. We show that the self-supervised model has implicitly learned to distinguish several cortical brain areas – a strong indicator that the proposed auxiliary task is appropriate for cytoarchitectonic mapping.

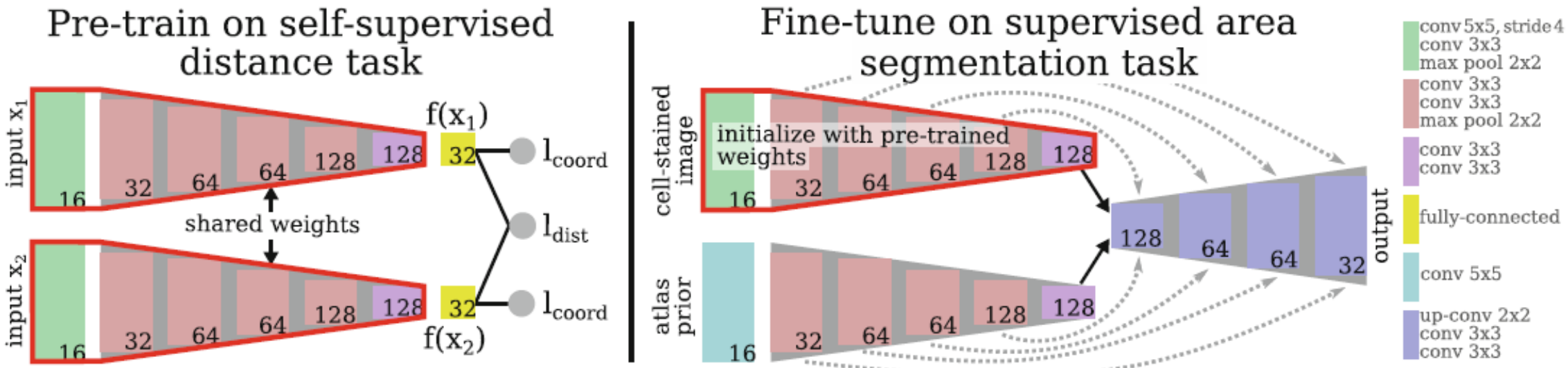


Fig. 2. Siamese network architecture for the auxiliary distance task (left) and extended U-net architecture for the area segmentation task (right). The network branches marked in red share the same architecture. Based on [10].

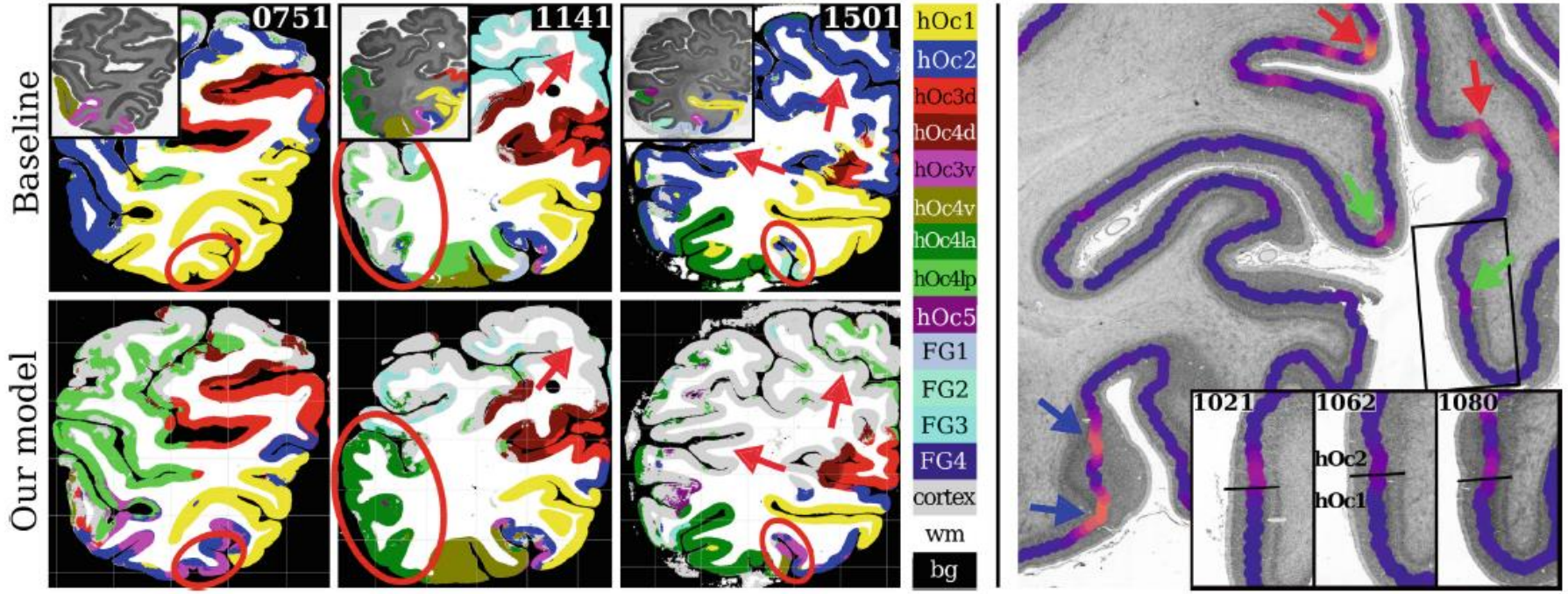


Fig. 4. Qualitative results. Left: Results on the area segmentation task with partial manual annotations in upper left corner. Compared to the baseline [10], our method predicts several areas significantly more accurate (circles) and has learned to deal much better with the "other cortex" class (arrows). Right: Squared Euclidean distances between averaged feature vectors of neighboring image patches, visualized by colored points along the cortical ribbon. Blue values indicate lower distances than yellowish values. Large distances occur at the border between hOc1/hOc2 over consecutive sections (green, enlarged boxes), at regions of high curvature (red) and at oblique regions (blue). This shows that the model has learned to recognize changes in cytoarchitecture.

MS-Net: Mixed-Supervision Fully-Convolutional Networks for Full-Resolution Segmentation

Meet P. Shah[1], S. N. Merchant[1], and Suyash P. Awate[2(✉)]

Electrical Engineering Department, Indian Institute of Technology (IIT) Bombay, Mumbai, India
[2] Computer Science and Engineering Department,
Indian Institute of Technology (IIT) Bombay, Mumbai, India

Abstract. For image segmentation, typical fully convolutional networks (FCNs) need strong supervision through a large sample of high-quality dense segmentations, entailing high costs in expert-raters' time and effort. We propose *MS-Net*, a new FCN to significantly reduce supervision cost, and improve performance, by coupling strong supervision with *weak supervision* through low-cost input in the form of *bounding boxes* and *landmarks*. Our MS-Net enables *instance-level segmentation* at *high spatial resolution*, with feature extraction using *dilated convolutions*. We propose a new loss function using *bootstrapped Dice* overlap for precise segmentation. Results on large datasets show that MS-Net segments more accurately at reduced supervision costs, compared to the state of the art.

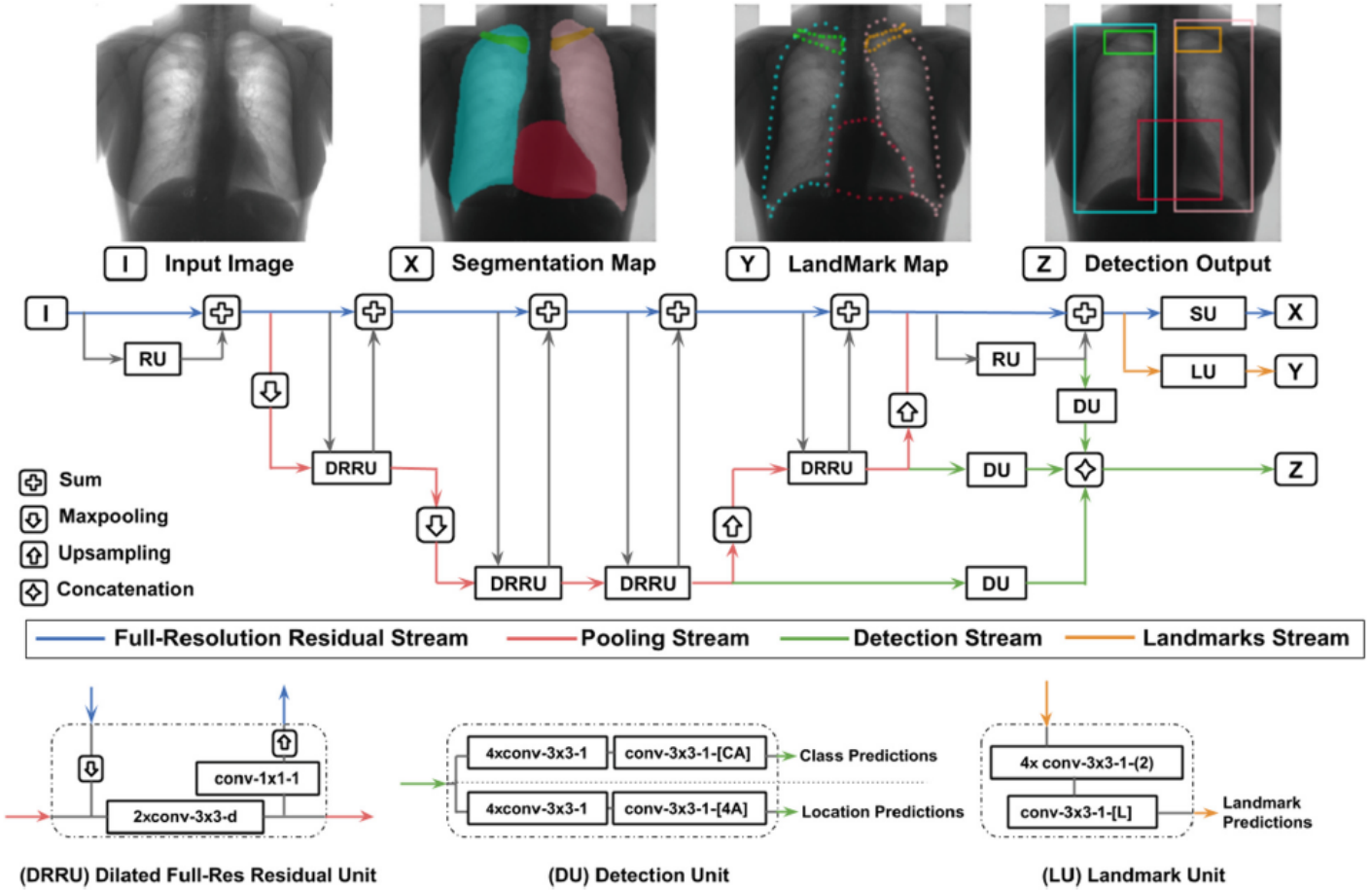


Fig. 1. Our MS-Net: Mixed-Supervision FCN for Full-Resolution Segmentation (abstract structure). We enable *mixed supervision* through a combination of:(i) high-quality *strong supervision* in the form of dense segmentation (per-pixel label) images, and (ii) low-cost *weak supervision* in the form of bounding boxes and landmarks. `N` × `conv-KxK-D-(S)-[P]` denotes: `N` sequential convolutional layers with kernels of spatial extent `(K, K)`, dilation factor `D`, spatial stride `S`, and `P` output feature maps.

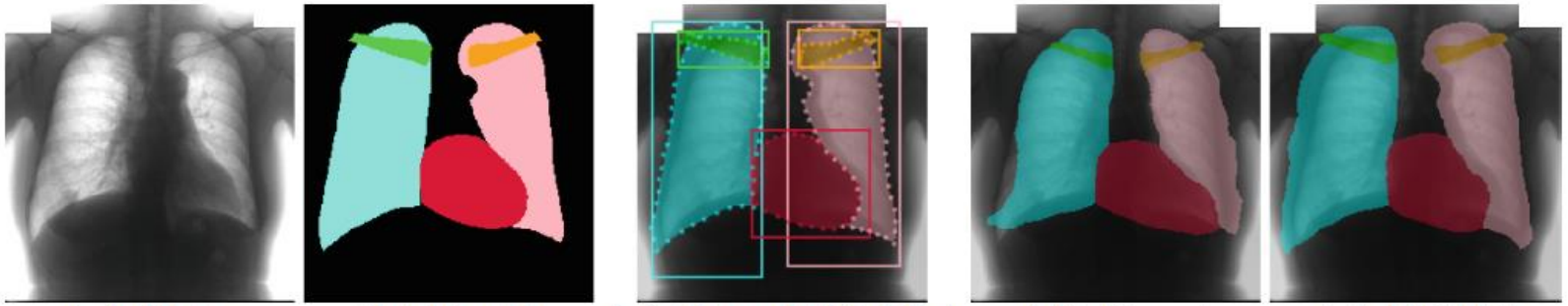


Fig. 3. Radiographs: Chest. (a) Data. **(b)** True segmentation. **(c)–(e)** Outputs for networks trained using all strong-supervision and weak-supervision data available.

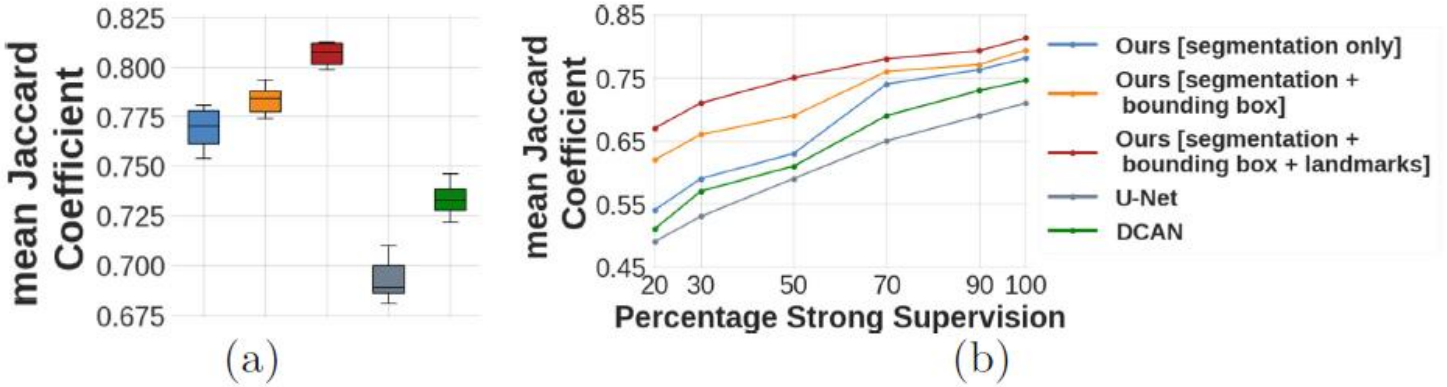


Fig. 4. Radiographs: Chest. (a) mJSC using all training data (strong + weak supervision). Box plots give variability over stochasticity in the optimization. **(b)** mJSC using different levels of strong supervision (remaining data with weak supervision).

Liver Lesion Detection from Weakly-Labeled Multi-phase CT Volumes with a Grouped Single Shot MultiBox Detector

Sang-gil Lee[1], Jae Seok Bae[2,3], Hyunjae Kim[1], Jung Hoon Kim[2,3,4], and Sungroh Yoon[1(✉)]

[1] Electrical and Computer Engineering, Seoul National University, Seoul, Korea
sryoon@snu.ac.kr
[2] Radiology, Seoul National University Hospital, Seoul, Korea
[3] Radiology, Seoul National University College of Medicine, Seoul, Korea
[4] Institute of Radiation Medicine, Seoul National University Medical Research Center, Seoul, Korea

Abstract. We present a focal liver lesion detection model leveraged by custom-designed multi-phase computed tomography (CT) volumes, which reflects real-world clinical lesion detection practice using a Single Shot MultiBox Detector (SSD). We show that grouped convolutions effectively harness richer information of the multi-phase data for the object detection model, while a naive application of SSD suffers from a generalization gap. We trained and evaluated the modified SSD model and recently proposed variants with our CT dataset of 64 subjects by five-fold cross validation. Our model achieved a 53.3% average precision score and ran in under three seconds per volume, outperforming the original model and state-of-the-art variants. Results show that the one-stage object detection model is a practical solution, which runs in near real-time and can learn an unbiased feature representation from a large-volume real-world detection dataset, which requires less tedious and time consuming construction of the weak phase-level bounding box labels.

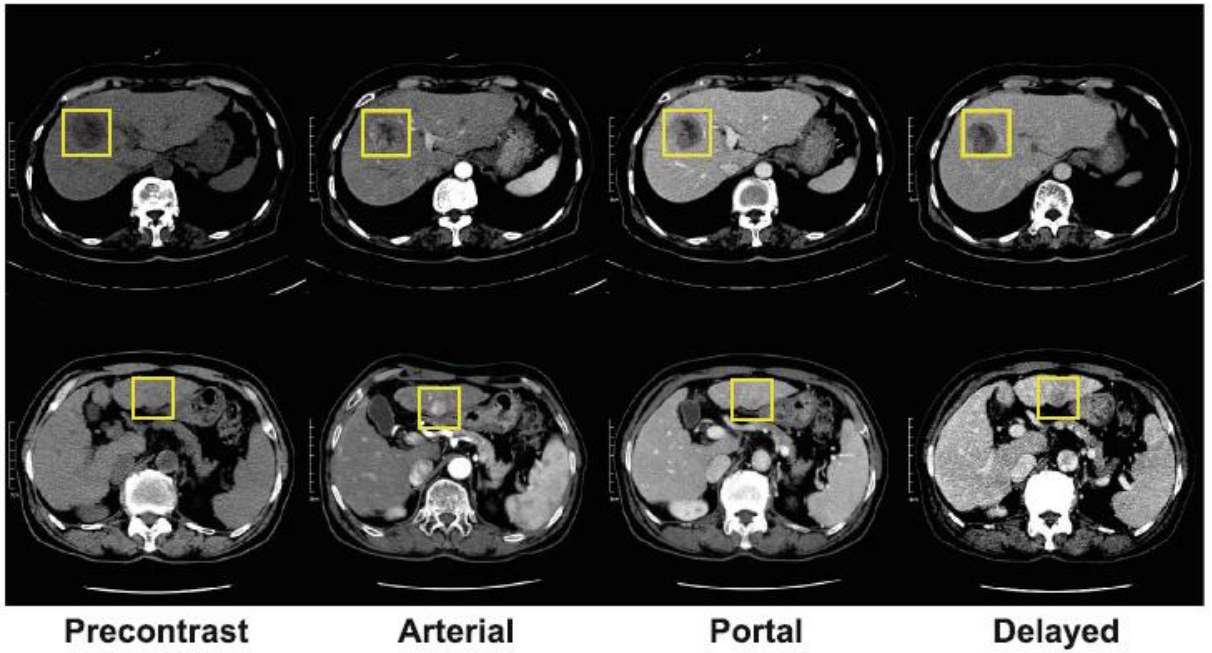


Fig. 1. Examples of the multi-phase CT dataset. Top: Lesions are visible from all phases. Bottom: Specific variants of lesions are visible only from specific phases. Note that the lesions are barely visible from the portal phase.

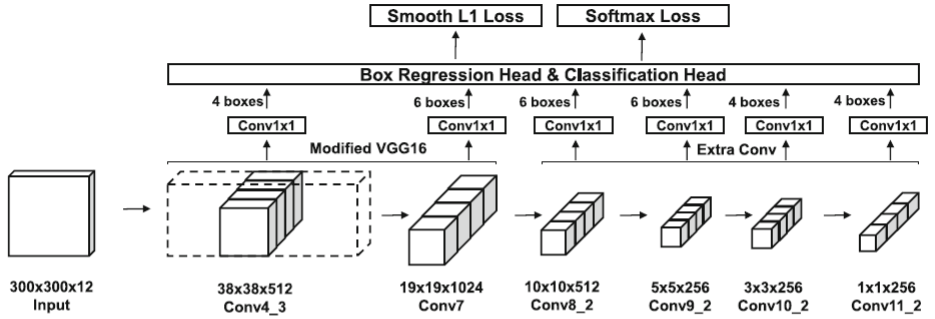


Fig. 2. Schematic diagram of grouped Single Shot MultiBox Detector. Solid lines of the convolutional feature map at the bottom indicate grouped convolutions. Digits next to upper arrows from the feature maps indicate the number of default boxes for each grid of the feature map. Intermediate layers and batch normalization are omitted for visual clarity.

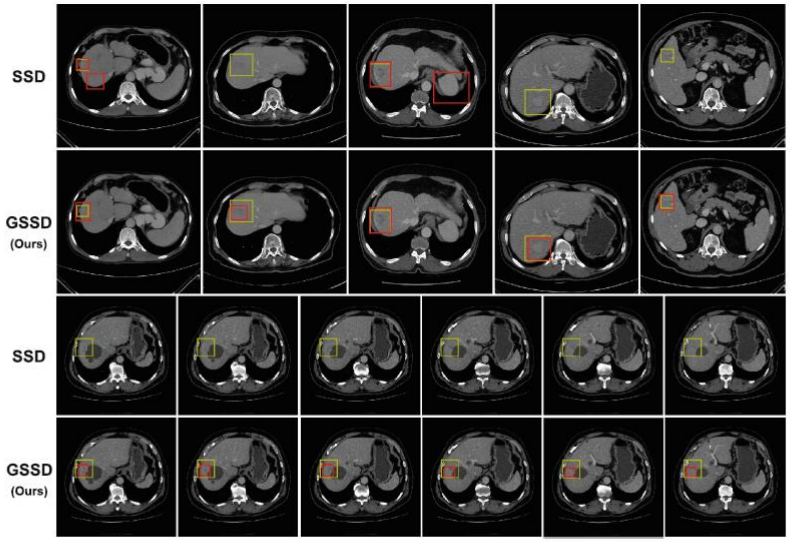


Fig. 4. Qualitative results from the validation set with a confidence threshold of 0.3. Yellow: Ground truth box. Red: Model predictions. Portal images shown. Top: GSSD accurately detects lesions, whereas the original model contains false positives or fails to detect. Bottom: A case of continuous slices. GSSD successfully tracks the exact location of lesions even with the given weak ground truth, whereas SSD completely fails.