Paper Presentation: Multi-organ Segmentation over Partially Labeled Datasets with Multi-scale Feature Abstraction

Benjamin Albert from Group 9 on
Predicting Hemorrhage Related Outcomes with CT Volumetry for Traumatic Hemothorax

https://ieeexplore.ieee.org/document/9112221
Fang X, Yan P. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE Transactions on Medical Imaging. 2020 Jun 9;39(11):3619-29.

Reason for choosing this paper

1. Open-sourced PyTorch implementation
2. Focuses on multiscale feature fusion, which our mentor believes will improve our current results
3. Evaluated on abdominal CT scans
4. Recent (published less than a year ago
5. Good journal (IEEE Transactions on Medical Imaging)

Authors’ Objectives and Hypothesis

The authors develop a novel network architecture to compete with benchmark models on four distinct organ segmentation challenges.

To do so, the authors’ architecture focuses on new ways to fuse hierarchically learned features so as to maintain local and global contexts.

The authors hypothesize “that the semantic information in various depths can be further enhanced by utilizing hierarchical contextual features. PIPO-FAN [Pyramid-Input Pyramid-Output Feature Abstract Network] aims to effectively extract multi-scale features for medical image segmentation, on top of the multi-scale nature of U-net.”

Jargon

- **Multi-scale**: pertaining to layer inputs, multiple scales arise from skip connections and explicit rescaling/pooling, both of which combine local with contextual information
- **Pyramid structure**: reducing an image size through convolution, pooling, etc. though typically adding many channels / learned filters
- **Deeply supervised**: evaluation and prioritization of discriminative latent features introduced by [1]
- **Semantic gap**: bridge between low-level latent features and high-level/human features
- **Attention mechanism**: a component of a network to assign importance to particular features or regions of features, similar to low-level human visual attention (e.g. superior colliculus)

[1] Lee CY, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-supervised nets.
In Artificial intelligence and statistics 2015 Feb 21 (pp. 562-570). PMLR.

Claimed Contributions

1. Pyramid-Input Pyramid-Output network to address the semantic gap that arises in multiscale features
2. Adaptive weighting layer to combine multiscale features
3. Adaptive loss to enable learning from partially labeled datasets
4. Good performance on public datasets

PIPO Architecture

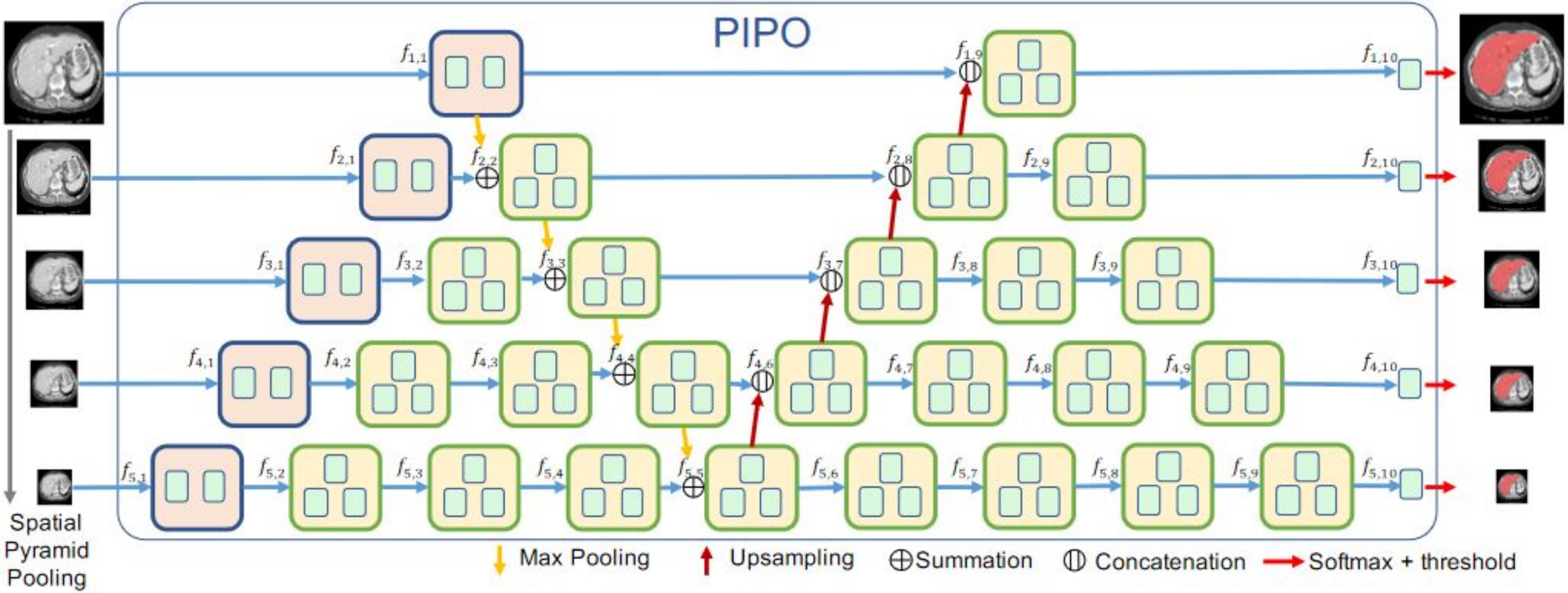

Fig. 3. Overview of the PIPO architecture. With the designed architecture, image information propagates from pyramid input to pyramid output through hierarchical abstraction and combination at each level.

PIPO-FAN architecture

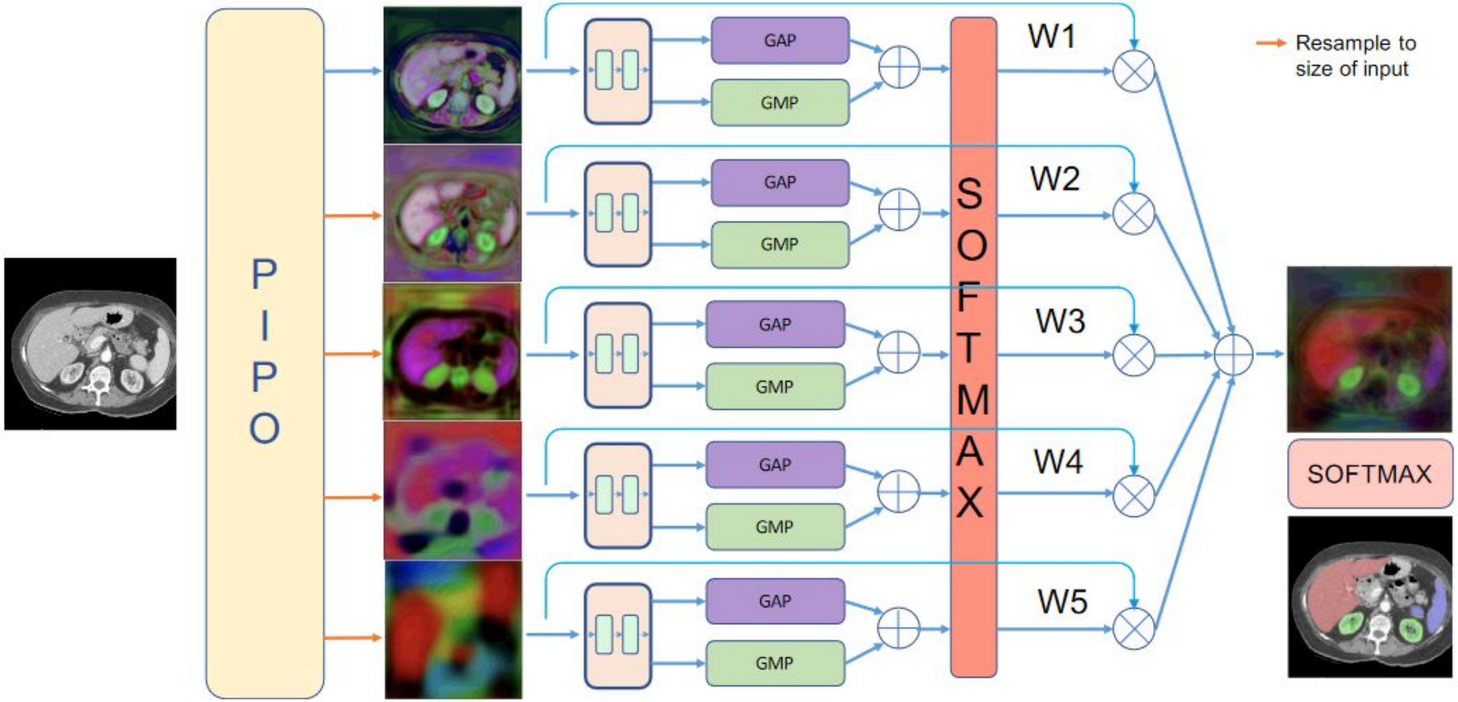

Fig. 4. Adaptive fusion of the multi-scale output segmentation features from PIPO-FAN. Features from lower scales tend to represent specific local segmentation, while features from higher scales are blurry but carry class information. Adaptive weights are computed by applying a shared convolutional module to the pyramid output features.

Implementation Highlights

Equal convolutional depth (ECD): features that are fused have passed through the same number of convolutional filters

Adaptive Fusion (AF): attention mechanism to indicate importance at each scale

Target adaptive loss (TAL): treats unknown labels as background and the final layer is branched to segment multiple organs

Results

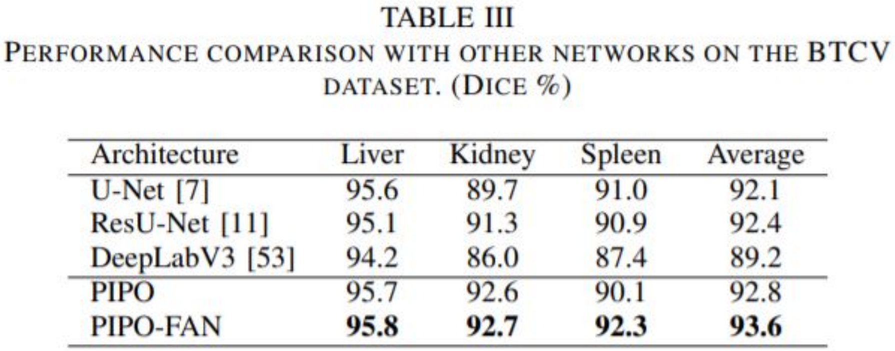

TABLE III
PERFORMANCE COMPARISON WITH OTHER NETWORKS ON THE BTCV DATASET. (DICE %)

Architecture	Liver	Kidney	Spleen	Average
U-Net [7]	95.6	89.7	91.0	92.1
ResU-Net [11]	95.1	91.3	90.9	92.4
DeepLabV3 [53]	94.2	86.0	87.4	89.2
PIPO	95.7	92.6	90.1	92.8
PIPO-FAN	**95.8**	**92.7**	**92.3**	**93.6**

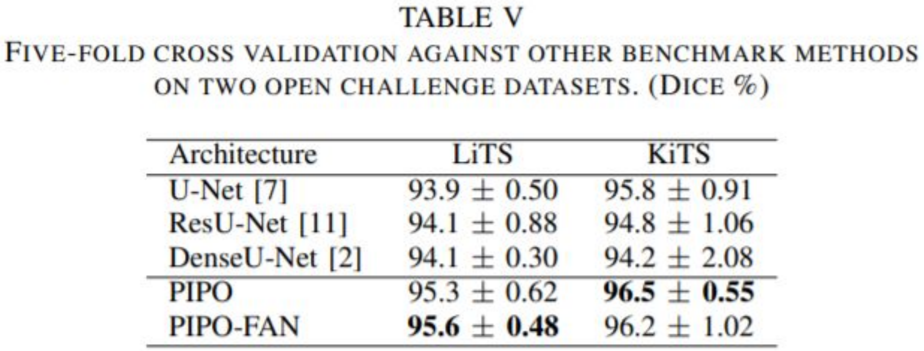

TABLE V
FIVE-FOLD CROSS VALIDATION AGAINST OTHER BENCHMARK METHODS ON TWO OPEN CHALLENGE DATASETS. (DICE %)

Architecture	LiTS	KiTS
U-Net [7]	93.9 ± 0.50	95.8 ± 0.91
ResU-Net [11]	94.1 ± 0.88	94.8 ± 1.06
DenseU-Net [2]	94.1 ± 0.30	94.2 ± 2.08
PIPO	95.3 ± 0.62	**96.5 ± 0.55**
PIPO-FAN	**95.6 ± 0.48**	96.2 ± 1.02

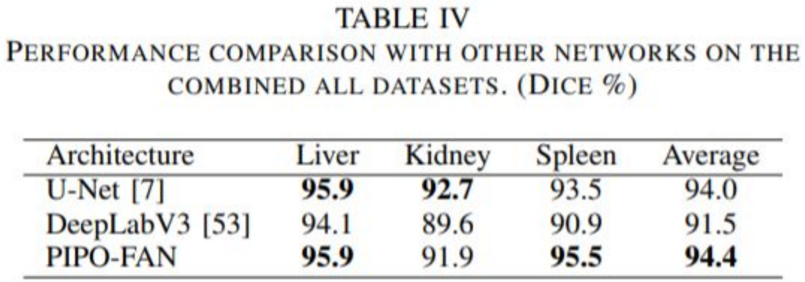

TABLE IV
PERFORMANCE COMPARISON WITH OTHER NETWORKS ON THE COMBINED ALL DATASETS. (DICE %)

Architecture	Liver	Kidney	Spleen	Average
U-Net [7]	**95.9**	**92.7**	93.5	94.0
DeepLabV3 [53]	94.1	89.6	90.9	91.5
PIPO-FAN	**95.9**	91.9	**95.5**	**94.4**

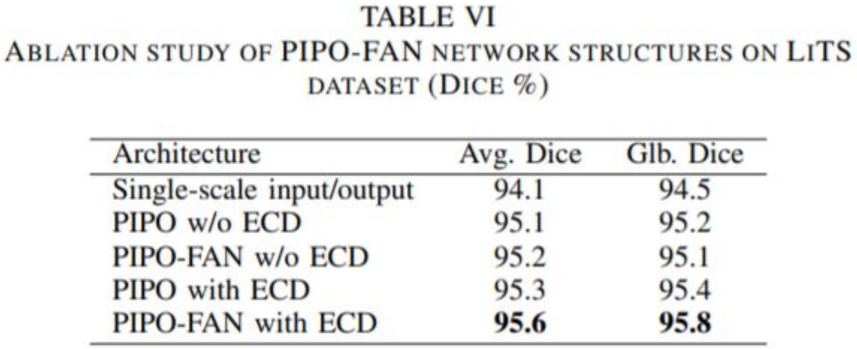

TABLE VI
ABLATION STUDY OF PIPO-FAN NETWORK STRUCTURES ON LITS DATASET (DICE %)

Architecture	Avg. Dice	Glb. Dice
Single-scale input/output	94.1	94.5
PIPO w/o ECD	95.1	95.2
PIPO-FAN w/o ECD	95.2	95.1
PIPO with ECD	95.3	95.4
PIPO-FAN with ECD	**95.6**	**95.8**

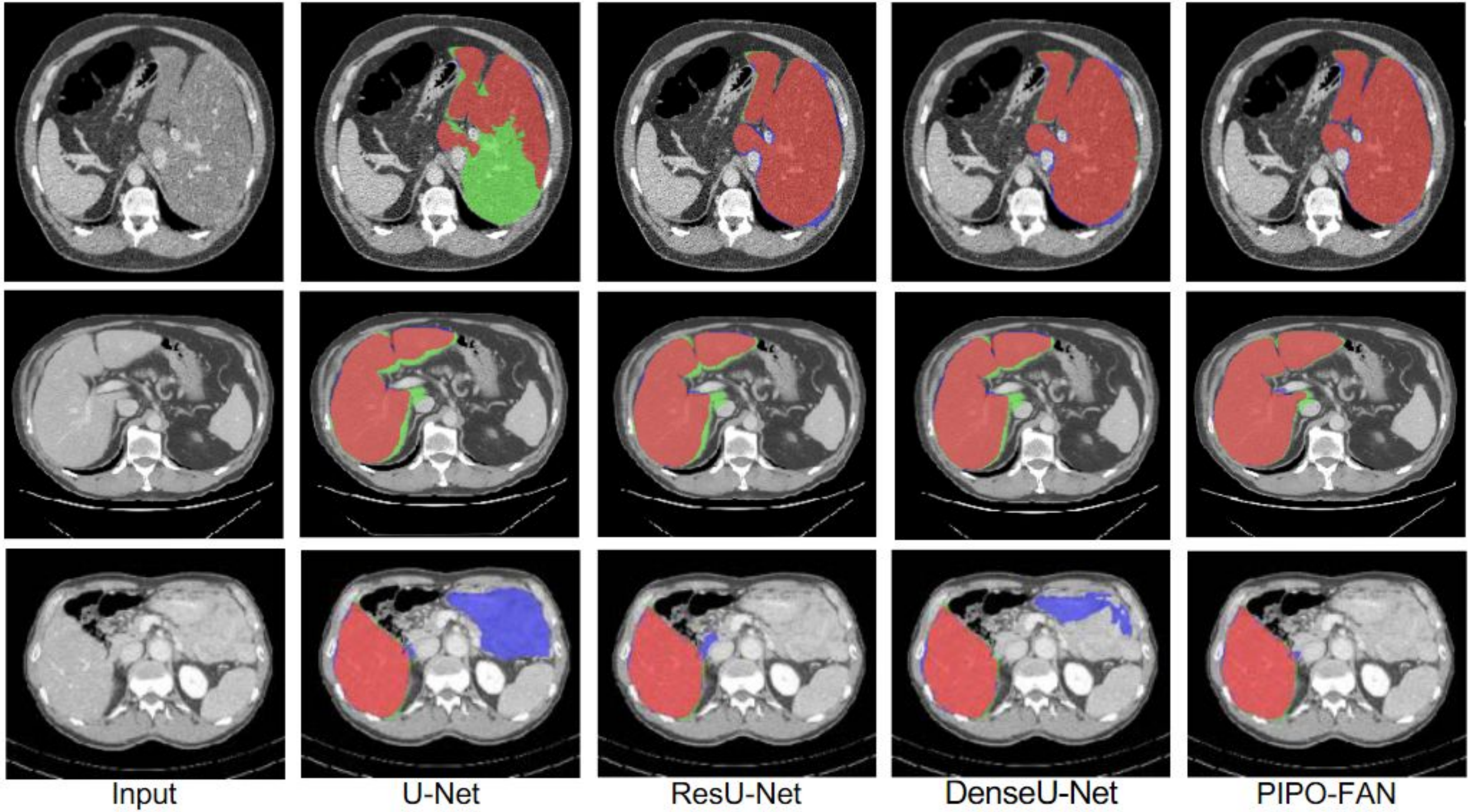

Fig. 6. Segmentation examples of different methods on LiTS data. From left to right are the raw image, results of U-Net, ResU-Net, DenseU-Net and our proposed PIPO-FAN, the red depicts correctly predicted liver segmentation, the blue shows false positive, green shows false negative.

The good

- Results (dice score) were generally better than benchmark models
- Ablation study
- Open-sourced and written in PyTorch

The good: ablation study

The ablation study helps to demonstrate the efficacy of the different modules in the network

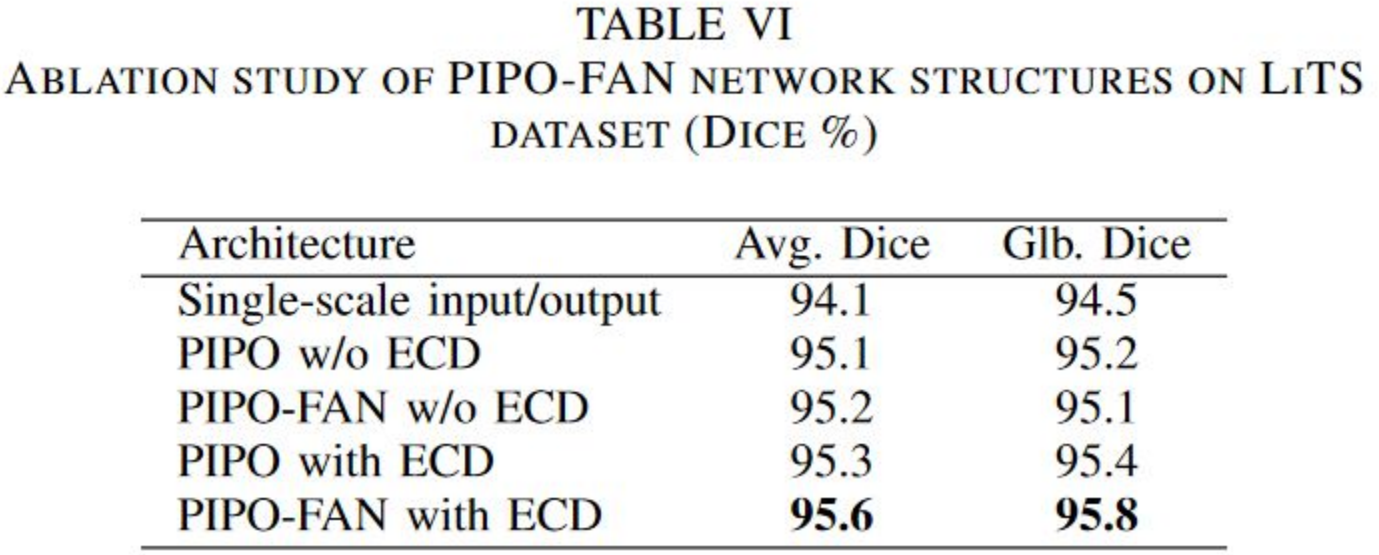

TABLE VI
ABLATION STUDY OF PIPO-FAN NETWORK STRUCTURES ON LITS DATASET (DICE %)

Architecture	Avg. Dice	Glb. Dice
Single-scale input/output	94.1	94.5
PIPO w/o ECD	95.1	95.2
PIPO-FAN w/o ECD	95.2	95.1
PIPO with ECD	95.3	95.4
PIPO-FAN with ECD	**95.6**	**95.8**

The bad

- Most of the claimed contributions are not novel concepts
- Undocumented and uncommented code
- Unsupported claims
- Inconsistent evaluation criteria
- Suboptimal evaluation of statistical significance

Criticism: claimed contributions

1. Pyramid-Input Pyramid-Output network to address the semantic gap that arises in multiscale features

    (novel)

2. Adaptive weighting layer to combine multiscale features

    (an attention mechanism)

3. Adaptive loss to enable learning from partially labeled datasets

    (already existed in the form of multiclass SVM)

4. Good performance on public datasets

    (not exactly a contribution)

Criticism: undocumented code

Virtually no comments (though many blocks of code are commented out)

From the abstract:

“The source code of this work is publicly shared at https://github.com/DIAL-RPI/PIPO-FAN to facilitate others to reproduce the work and build their own models using the introduced mechanisms.”

Criticism: unsupported claims

For example: “DPS can help relieve the problem of gradient vanishing in deep neural networks and learn deep level features with hierarchical contexts. It also enforces the outputs in all scales to maintain structural information.”

Criticism: inconsistent evaluation criteria

In total, PIPO-FAN was benchmarked against four state-of-the-art networks on four separate challenge datasets. However, the authors selectively chose a subset of these four networks against which to compare for each challenge dataset.

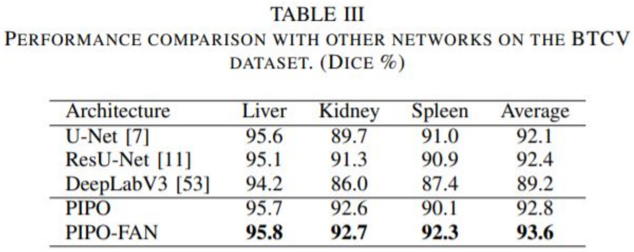

TABLE III
PERFORMANCE COMPARISON WITH OTHER NETWORKS ON THE BTCV DATASET. (DICE %)

Architecture	Liver	Kidney	Spleen	Average
U-Net [7]	95.6	89.7	91.0	92.1
ResU-Net [11]	95.1	91.3	90.9	92.4
DeepLabV3 [53]	94.2	86.0	87.4	89.2
PIPO	95.7	92.6	90.1	92.8
PIPO-FAN	**95.8**	**92.7**	**92.3**	**93.6**

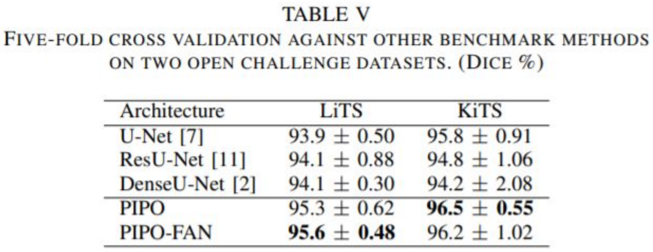

TABLE V
FIVE-FOLD CROSS VALIDATION AGAINST OTHER BENCHMARK METHODS ON TWO OPEN CHALLENGE DATASETS. (DICE %)

Architecture	LiTS	KiTS
U-Net [7]	93.9 ± 0.50	95.8 ± 0.91
ResU-Net [11]	94.1 ± 0.88	94.8 ± 1.06
DenseU-Net [2]	94.1 ± 0.30	94.2 ± 2.08
PIPO	95.3 ± 0.62	$\mathbf{96.5 \pm 0.55}$
PIPO-FAN	$\mathbf{95.6 \pm 0.48}$	96.2 ± 1.02

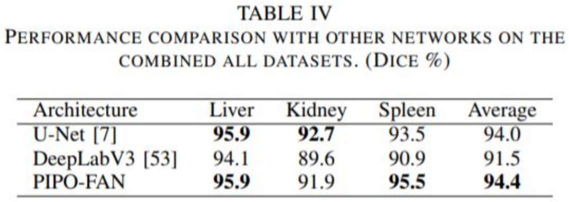

TABLE IV
PERFORMANCE COMPARISON WITH OTHER NETWORKS ON THE COMBINED ALL DATASETS. (DICE %)

Architecture	Liver	Kidney	Spleen	Average
U-Net [7]	**95.9**	**92.7**	93.5	94.0
DeepLabV3 [53]	94.1	89.6	90.9	91.5
PIPO-FAN	**95.9**	91.9	**95.5**	**94.4**

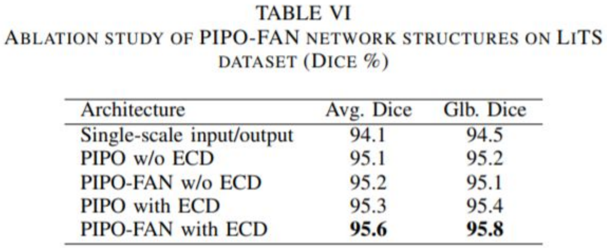

TABLE VI
ABLATION STUDY OF PIPO-FAN NETWORK STRUCTURES ON LITS DATASET (DICE %)

Architecture	Avg. Dice	Glb. Dice
Single-scale input/output	94.1	94.5
PIPO w/o ECD	95.1	95.2
PIPO-FAN w/o ECD	95.2	95.1
PIPO with ECD	95.3	95.4
PIPO-FAN with ECD	**95.6**	**95.8**

Criticism: suboptimal evaluation of statistical significance

The authors used the t-test to evaluate whether PIPO-FAN significantly outperforms the benchmark models.

The authors do not demonstrate nor state that the underlying data distribution is normally distributed, which is one of the assumptions of the test.

It would have been better for them to have used the Wilcoxon signed rank test because:

1. The data are paired
2. Does not assume normal distribution

Minor criticisms

- Table that enumerates the total number of parameters in PIPO-FAN does not compare against the benchmark models
    - This is listed as minor because it was not the focus of the research
- Some acronyms are used before defined (e.g. GAP/GMP in Fig. 4)
- Materials section does not describe nor list the GPUs, but the authors eventually mention them