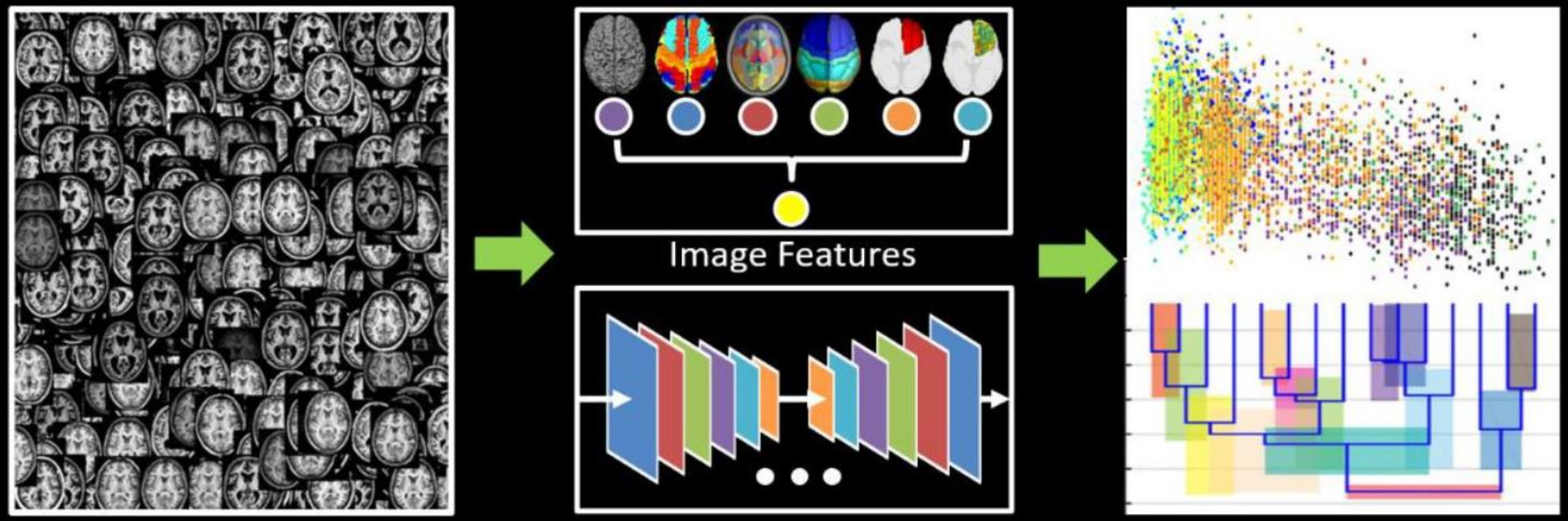


Multi-Task Learning

[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

- Review
- Multi-task Learning
- Papers

Organization of Course

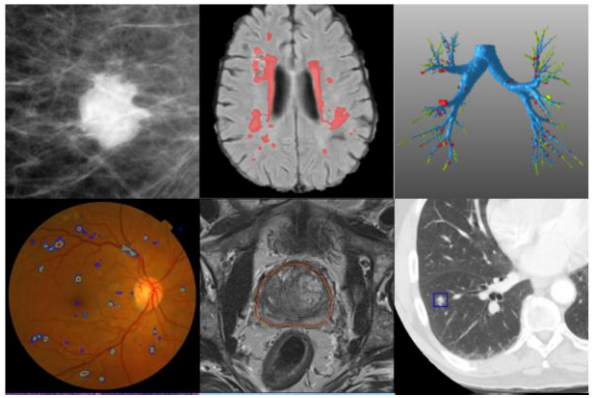
Overview

Overview of Deep Learning in Medical Image Computing
Neural Networks and CNN

Key Techs

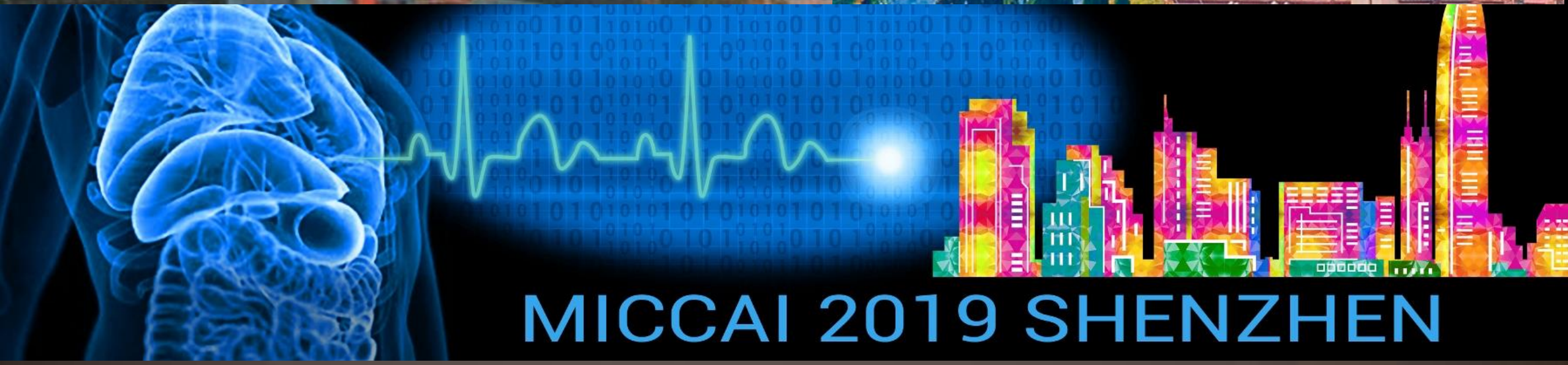
Classification (Medical Image Diagnosis)
Detection (Landmark Localization and Detection)
Segmentation (Medical Image Segmentation)
GAN (Medical Image Synthesis)

Topics in Medical Image Computing

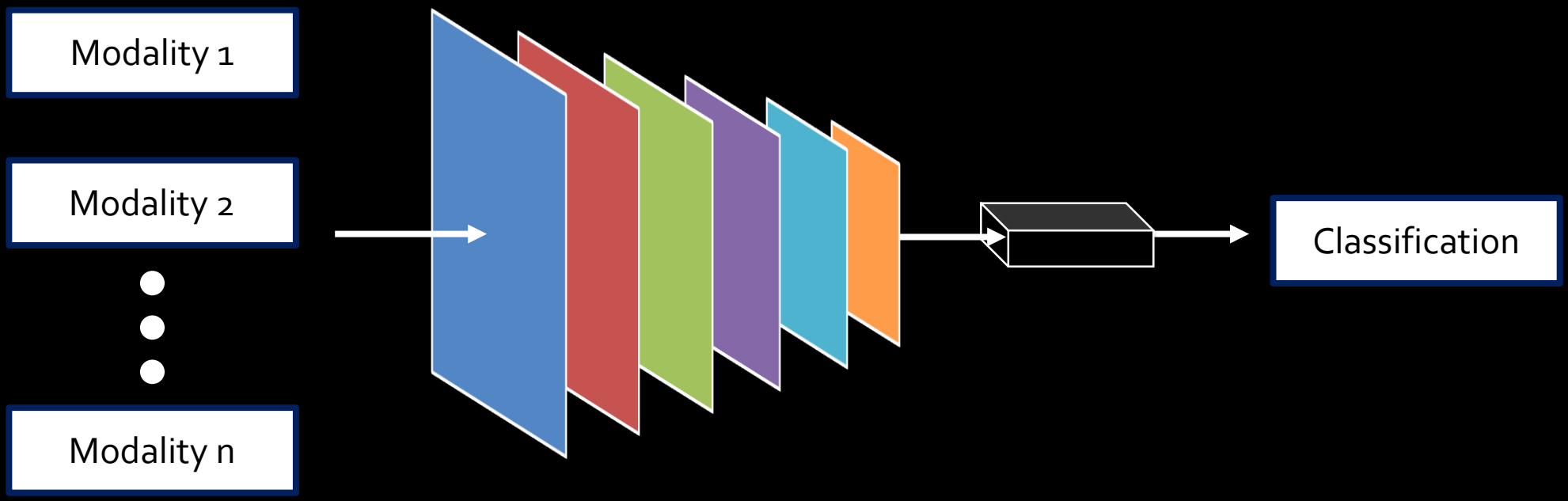



MICCAI
Sept 10-14, 2017 in Quebec City, Quebec, Canada
2017

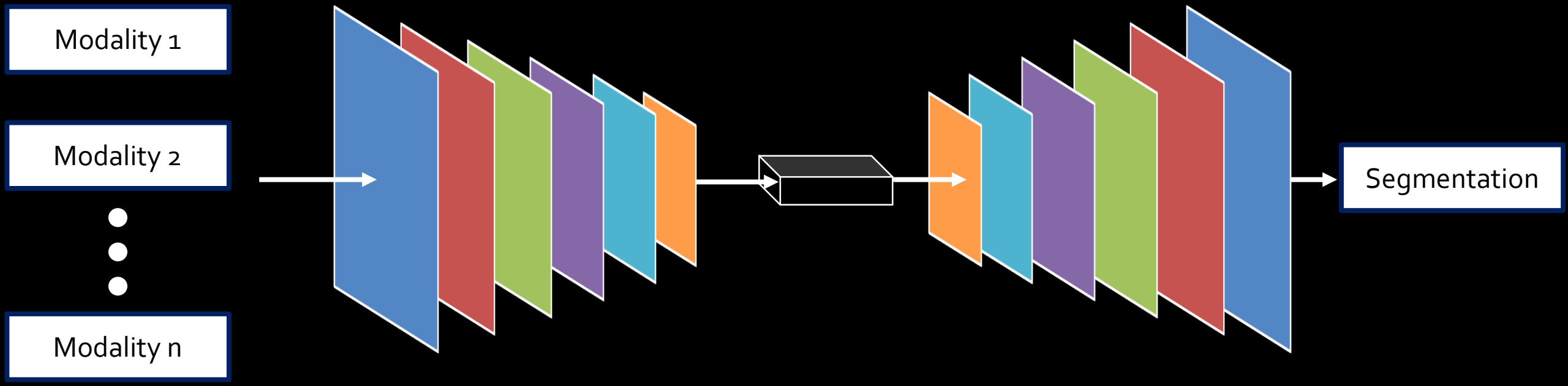

MICCAI2018
Granada
SPAIN


MICCAI 2019 SHENZHEN

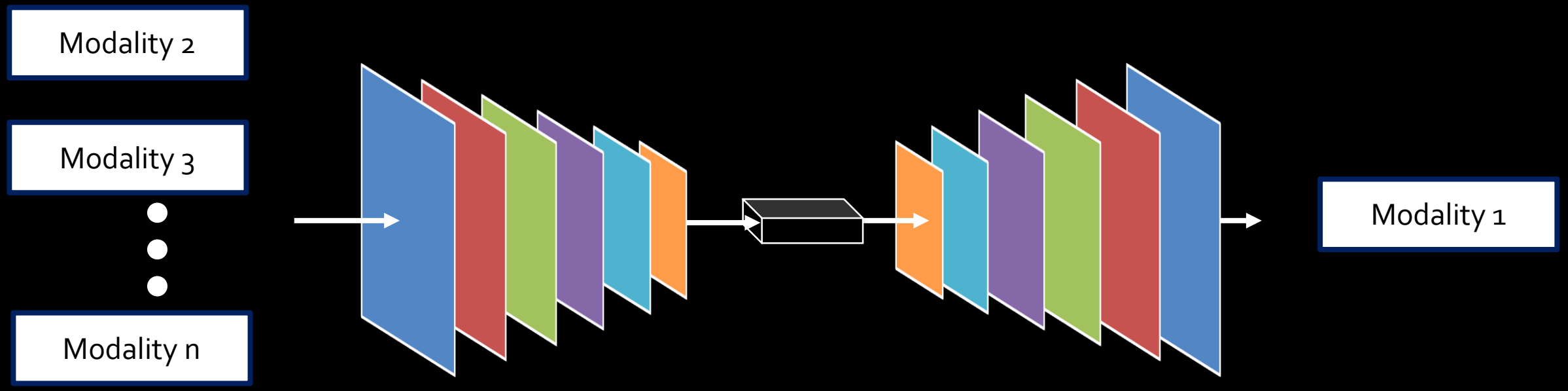
Multi-modal Learning



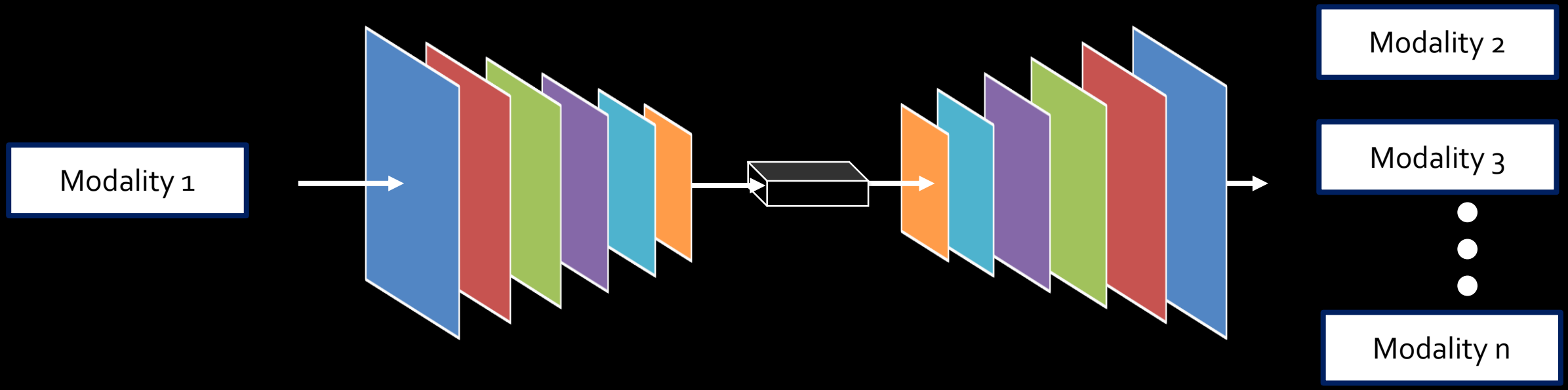
Multi-modal Learning



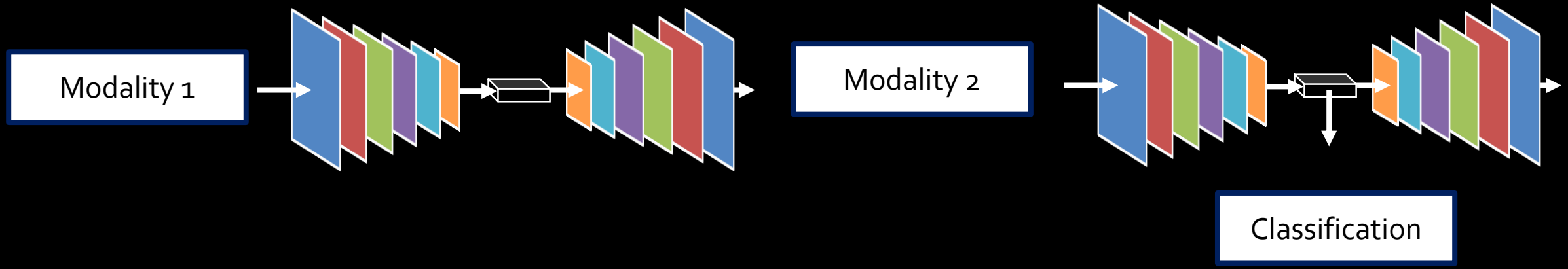
Multi-modal Learning



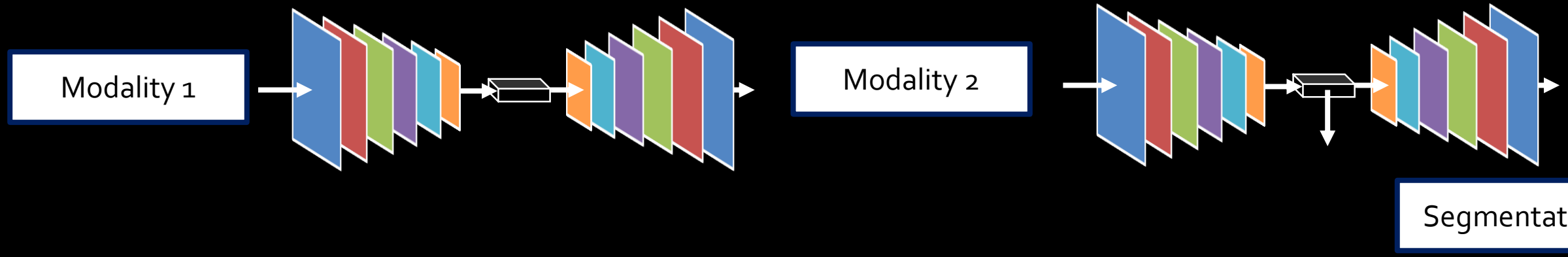
Multi-modal Learning



Multi-modal Learning

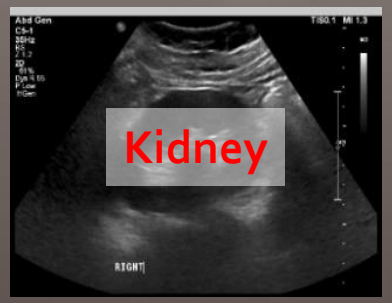


Multi-modal Learning

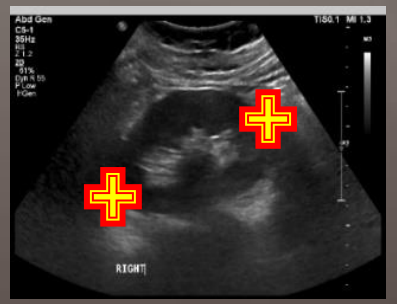


Multi-task Learning

Input

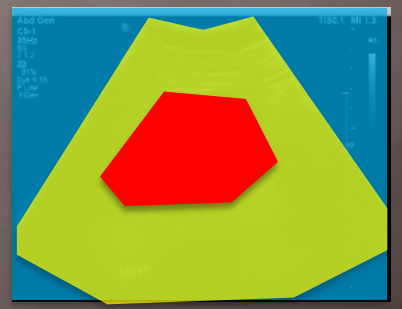


Classification



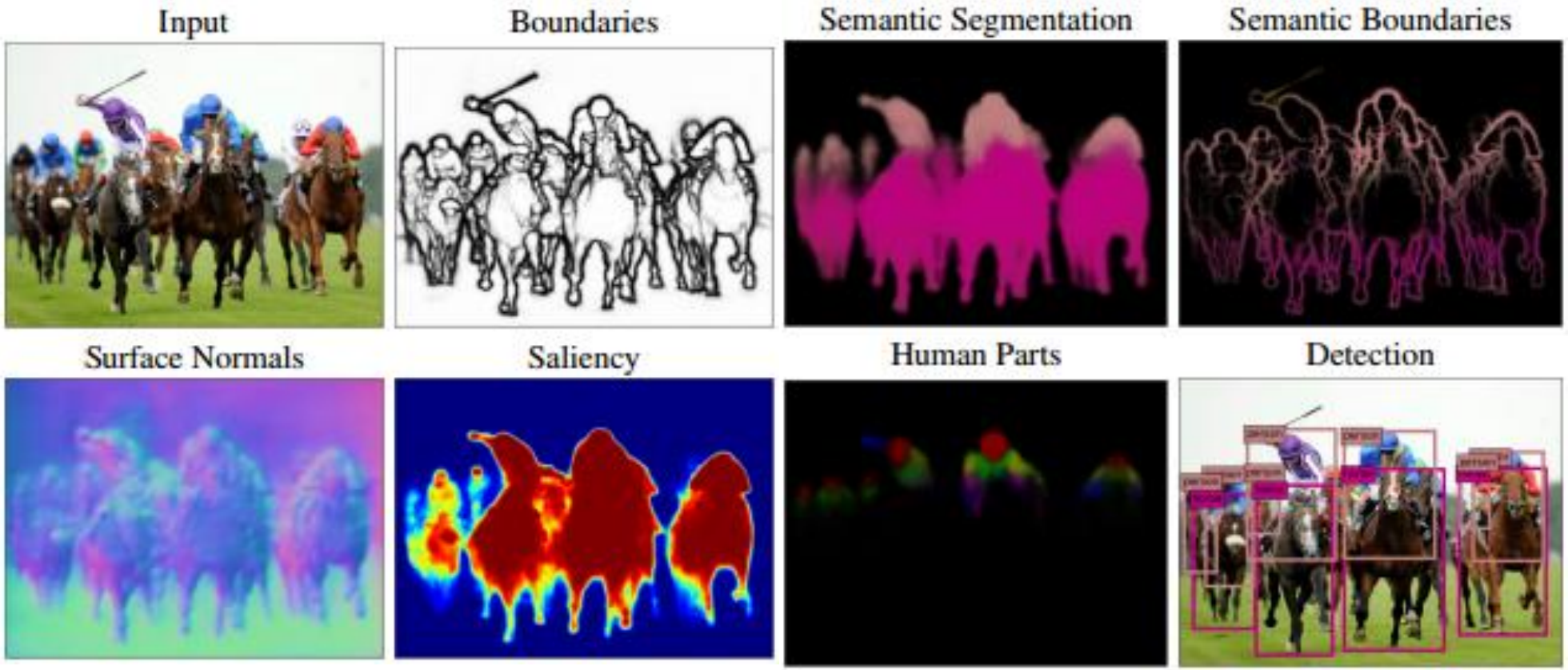
Detection

Synthesis



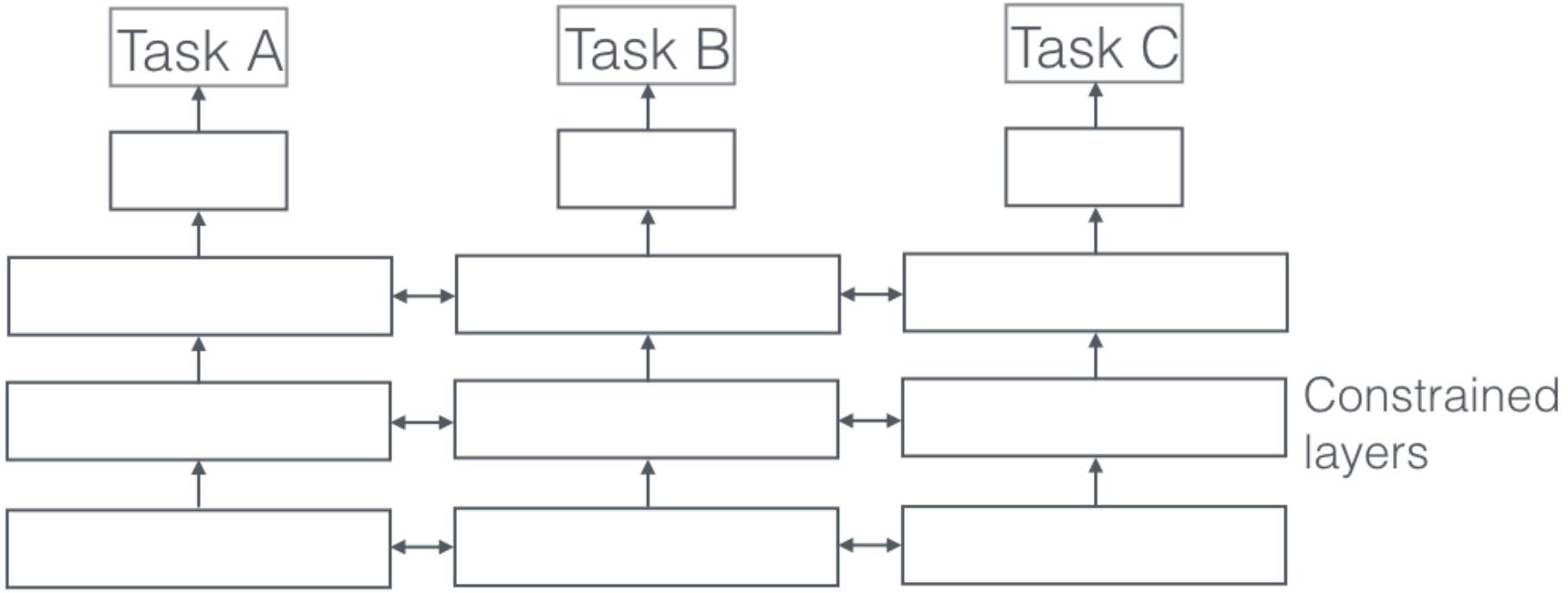
Segmentation

Introduction

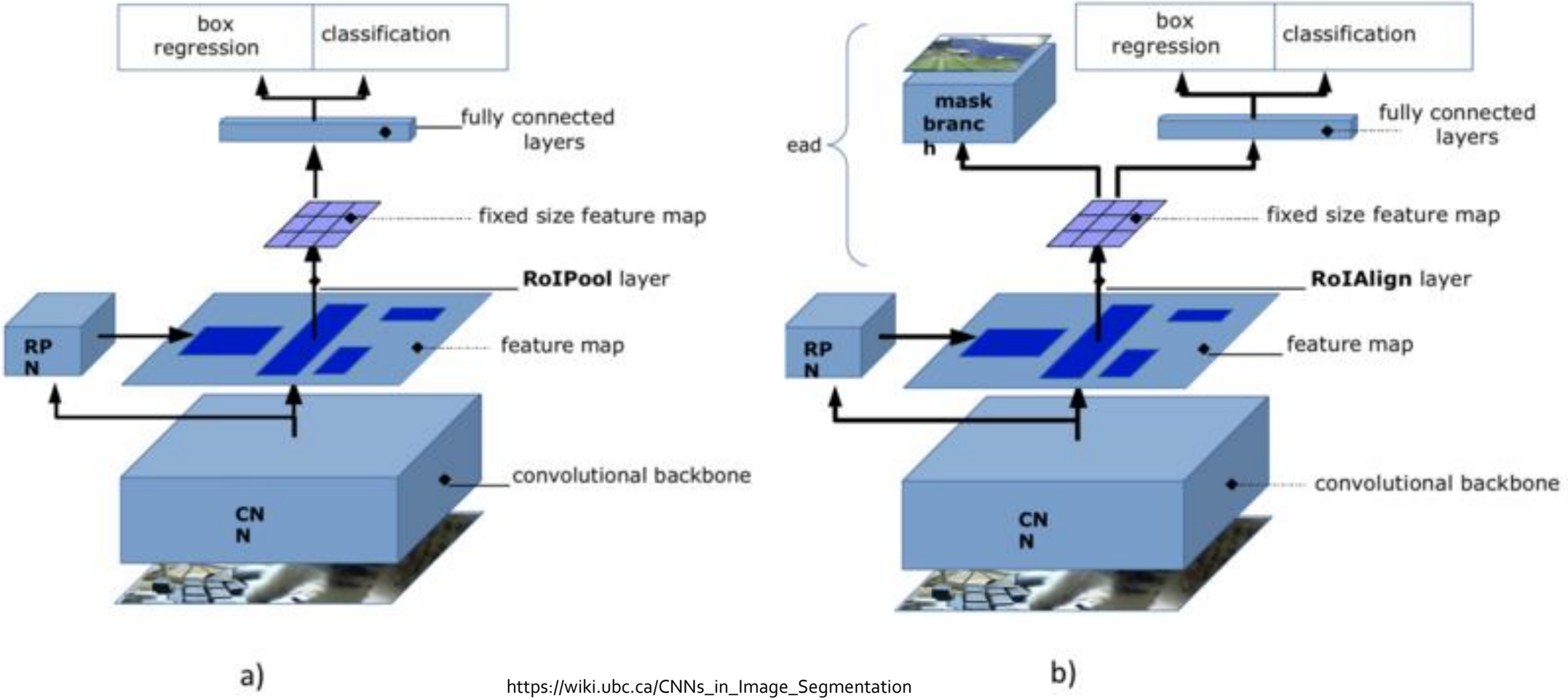


Train a single model → Fine-tune tweak the parameters → Single-Task Learning (STL)

Sharing representations between related tasks → optimizing more than one loss function → Multi-Task Learning (MTL)



Example



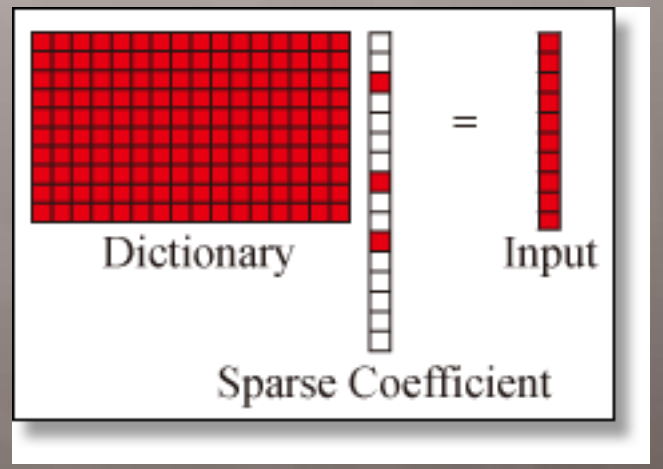
https://wiki.ubc.ca/CNNs_in_Image_Segmentation

Motivation

human
learning

Machine Learning

$$= \underset{\beta \in \mathbf{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$



The inductive bias is provided by the auxiliary tasks, which cause the model to prefer hypotheses that explain more than one task

Why MTL Work?

Multitask Learning - Springer
link.springer.com/chapter/10.1007%2F978-1-4615-5529-2_5
by R Caruana - 1998 - Cited by 2401 - Related articles
Rich **Caruana** ... **Multitask** Learning is an approach to inductive transfer that improves ... In this paper we demonstrate **multitask** learning in three domains. ..... Pages: pp 95-133; Copyright: **1998**; DOI: 10.1007/978-1-4615-5529-2_5; Print ISBN ...

"MTL improves generalization by leveraging the domain-specific information contained in the training signals of related tasks". -- Rich Caruana

Implicit data augmentation

Attention focusing

Eavesdropping

Representation bias

Regularization

Implicit data augmentation | Attention focusing

Eavesdropping | Representation bias | Regularization

MTL effectively increases the sample size that we are using for training our model.

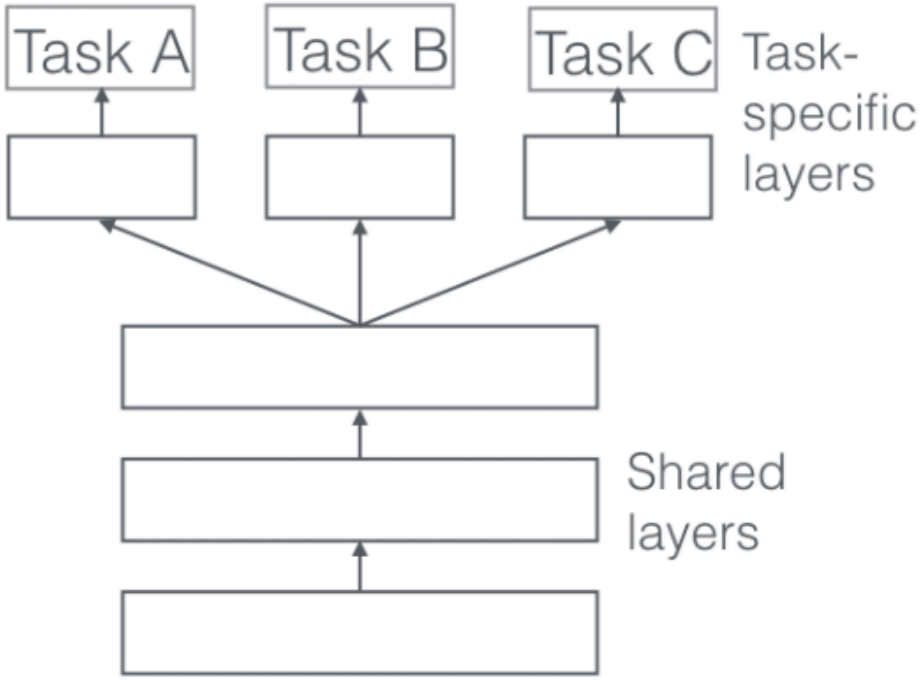
Learning just task ***A*** bears the risk of overfitting to task ***A***

Learning ***A*** and ***B*** jointly enables the model to obtain a better representation ***F***

Two MTL Frameworks For Deep Learning

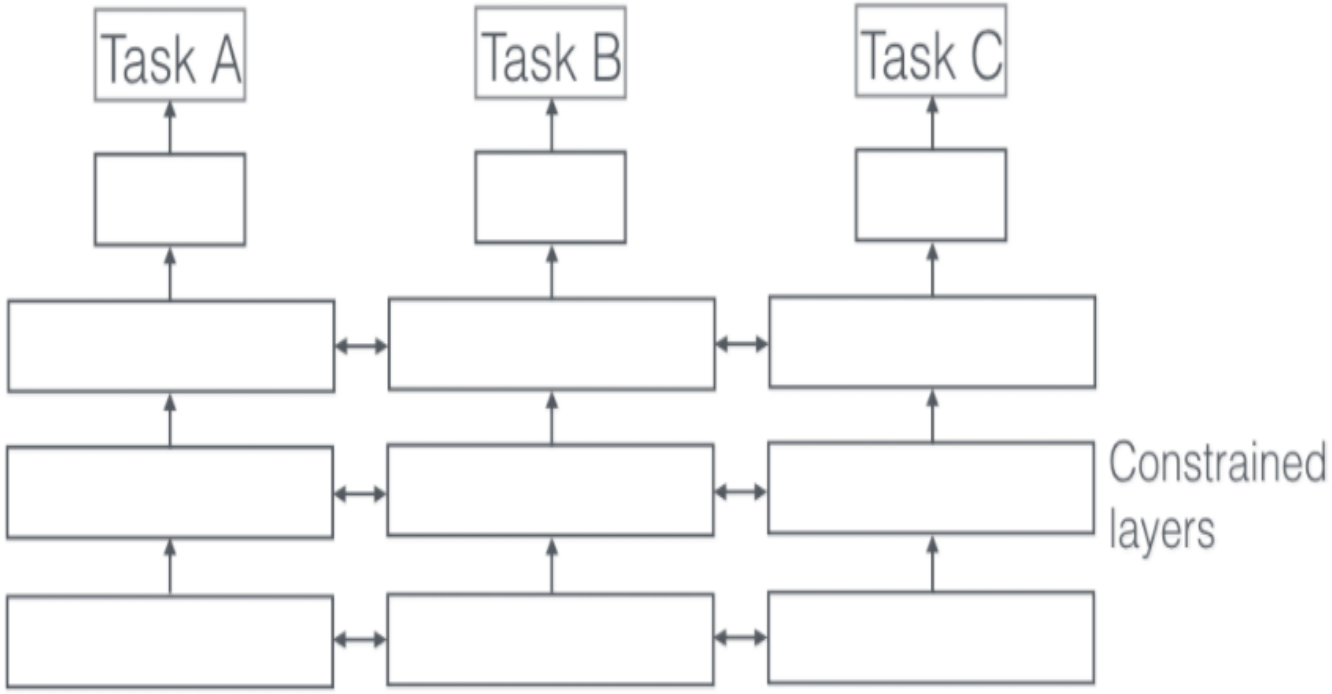
How to share the parameters of layers?

hard



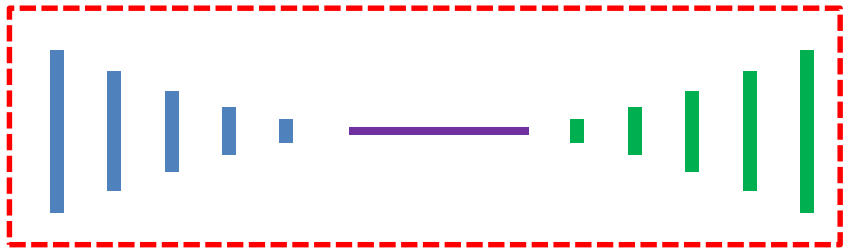
Hard parameter sharing greatly reduces the risk of overfitting

soft

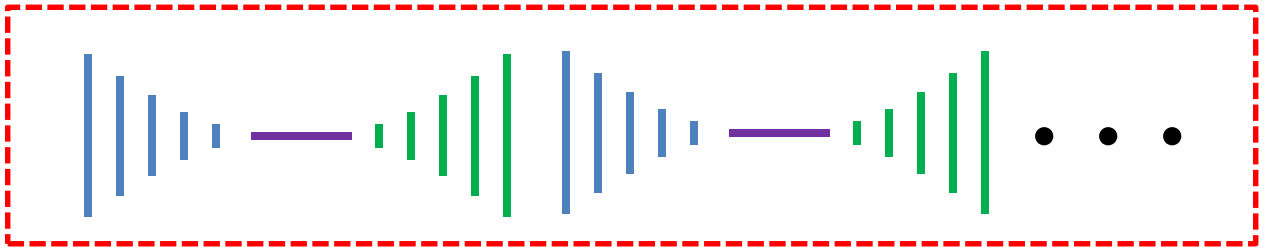


Distance between the parameters of the model is regularized in order to encourage the parameters to be similar

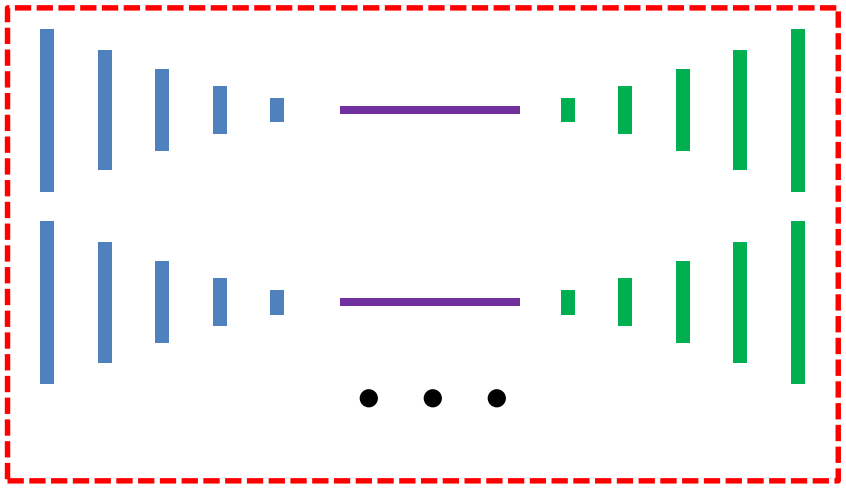
Recent Works of Using MTL in Deep Learning



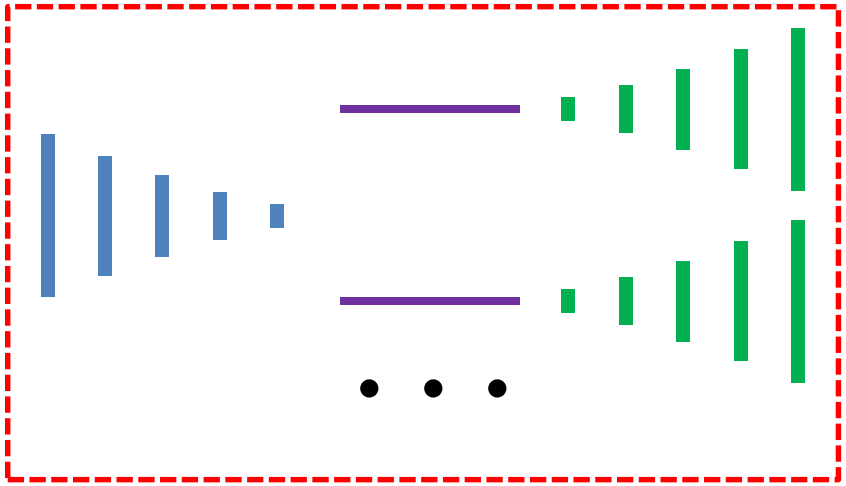
Single



Consequent

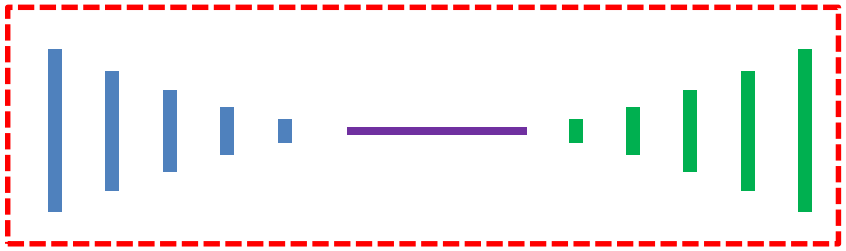


Soft



Hard

- Model Dependency Among Different Tasks
- Characterize the Commonalities and Differences Between Tasks



Single

Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture

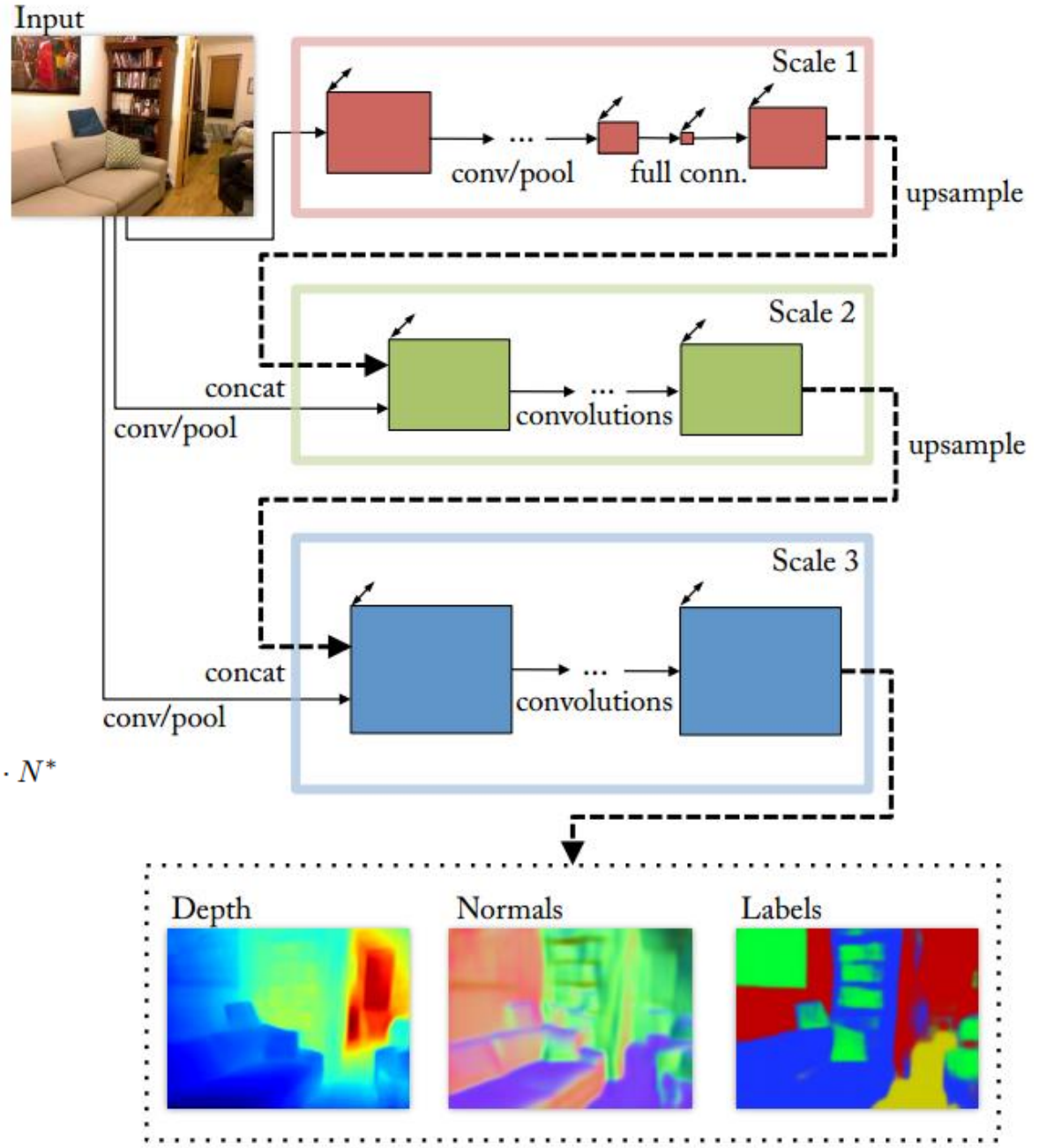
David Eigen[1] Rob Fergus[1,2]
[1] Dept. of Computer Science, Courant Institute, New York University
[2] Facebook AI Research
{deigen,fergus}@cs.nyu.edu

$$L_{depth}(D, D^*) = \frac{1}{n}\sum_i d_i^2 - \frac{1}{2n^2}\left(\sum_i d_i\right)^2 + \frac{1}{n}\sum_i[(\nabla_x d_i)^2 + (\nabla_y d_i)^2]$$

$$L_{normals}(N, N^*) = -\frac{1}{n}\sum_i N_i \cdot N_i^* = -\frac{1}{n} N \cdot N^*$$

$$L_{semantic}(C, C^*) = -\frac{1}{n}\sum_i C_i^* \log(C_i)$$

ICCV 2015 [X]
https://arxiv.org/pdf/1411.4734.pdf



Deep Learning for Multi-Task Medical Image Segmentation in Multiple Modalities

Pim Moeskops[1,2⋆], Jelmer M. Wolterink[1⋆], Bas H.M. van der Velden[1], Kenneth G.A. Gilhuijs[1], Tim Leiner[3], Max A. Viergever[1], and Ivana Išgum[1]

[1] Image Sciences Institute, University Medical Center Utrecht, The Netherlands
[2] Medical Image Analysis, Eindhoven University of Technology, The Netherlands
[3] Department of Radiology, University Medical Center Utrecht, The Netherlands

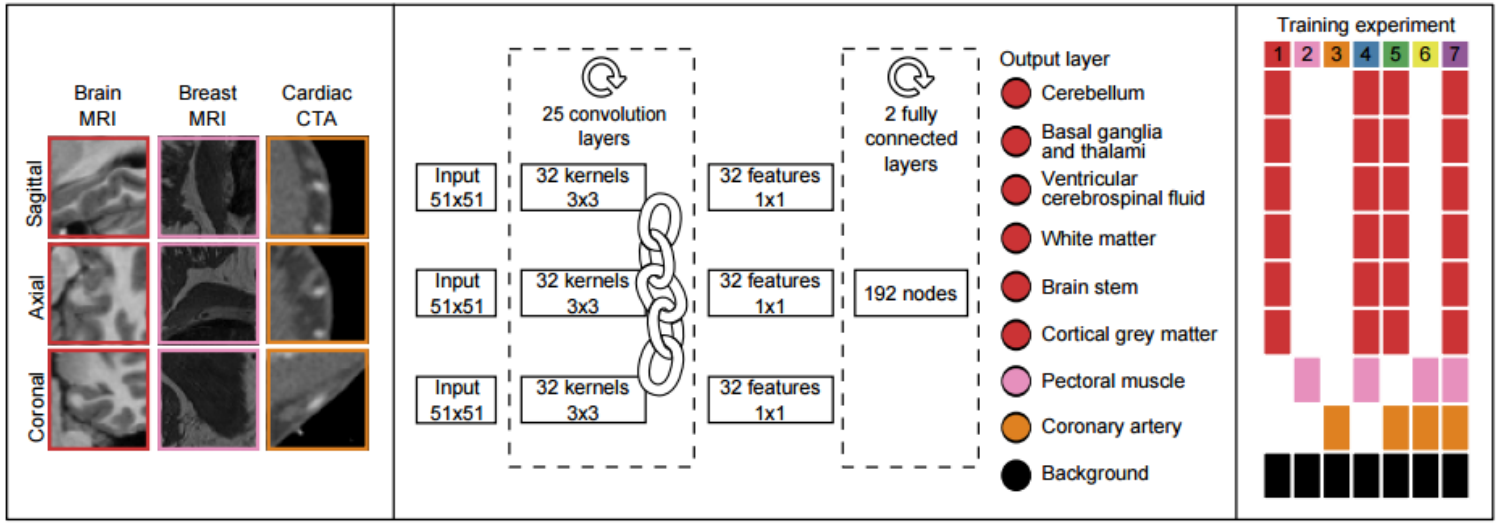
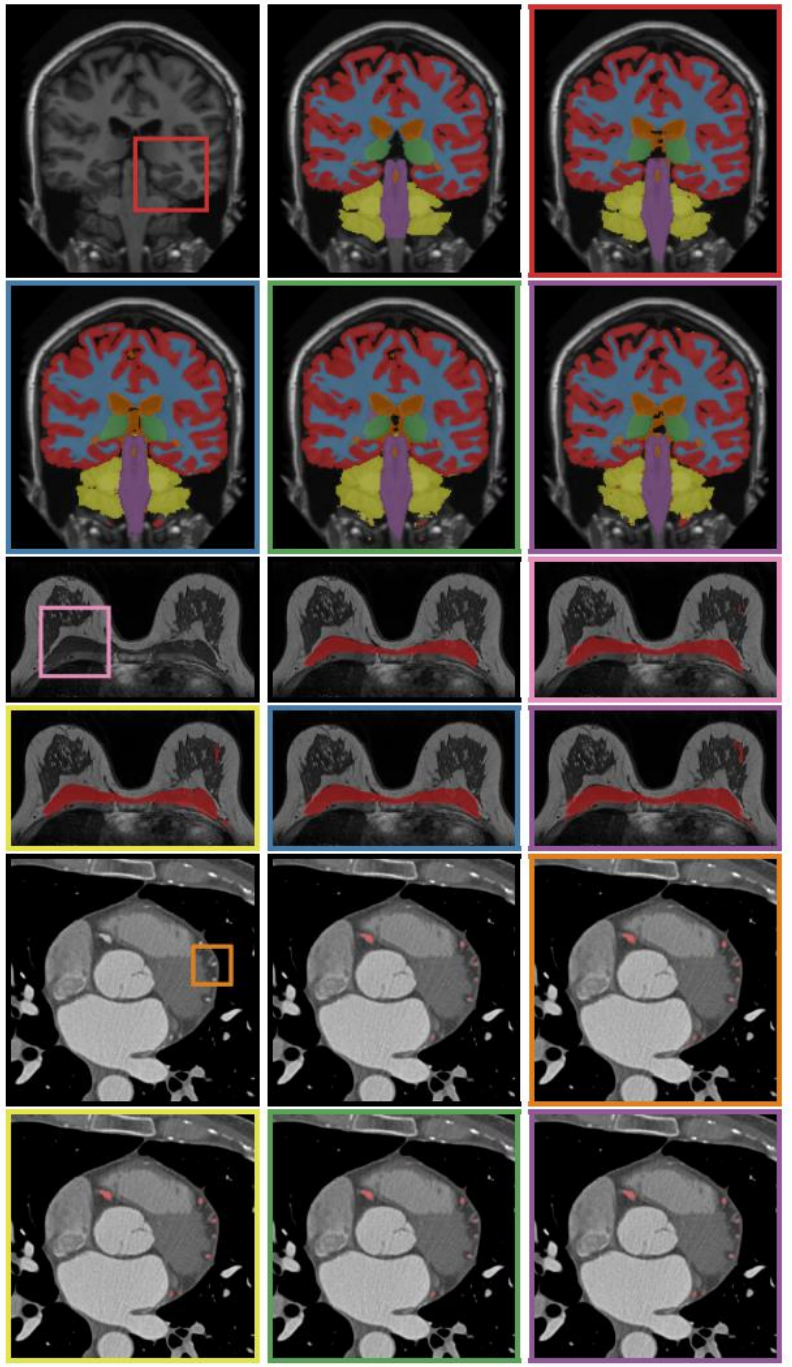
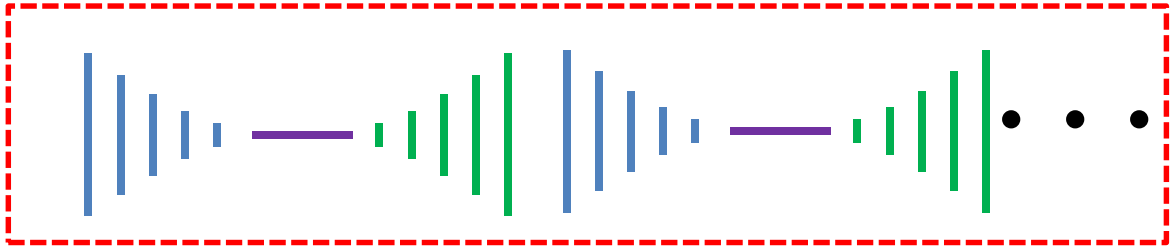


Fig. 1. Example 51×51 triplanar input patches (*left*). CNN architecture with 25 shared convolution layers, 2 fully connected layers and an output layer with at most 9 classes, including a background class common among tasks (*centre*). Output classes included in each training experiment (*right*).

MICCAI 2016
https://arxiv.org/pdf/1704.03379.pdf





Consequent

A Joint Many-Task Model:
Growing a Neural Network for Multiple NLP Tasks

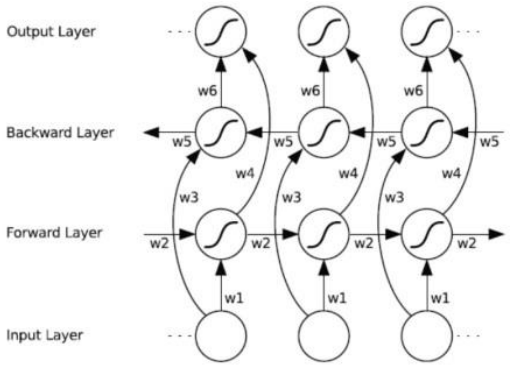
Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka & Richard Socher
The University of Tokyo
{hassy, tsuruoka}@logos.t.u-tokyo.ac.jp
Salesforce Research
{cxiong, rsocher}@salesforce.com

$$J_1(\theta_{\text{POS}}) = -\sum_s \sum_t \log p(y_t^{(1)} = \alpha | h_t^{(1)}) + \lambda \|W_{\text{POS}}\|^2 + \delta \|\theta_e - \theta'_e\|^2,$$

$$J_2(\theta_{\text{chk}}) = -\sum_s \sum_t \log p(y_t^{(2)} = \alpha | h_t^{(2)}) + \lambda \|W_{\text{chk}}\|^2 + \delta \|\theta_{\text{POS}} - \theta'_{\text{POS}}\|^2,$$

• • •

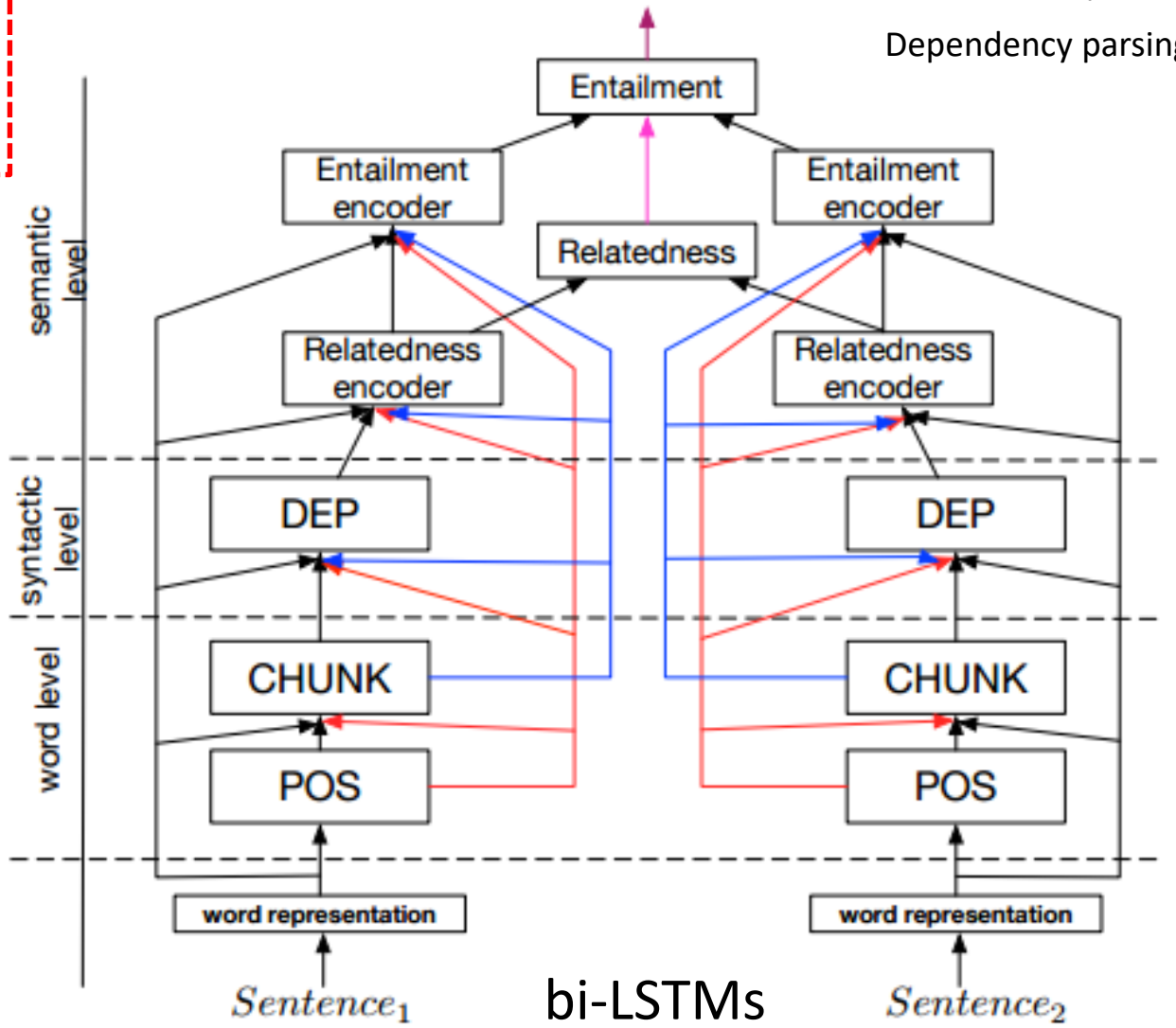
$$J_5(\theta_{\text{ent}}) = -\sum_{(s,s')} \log p(y_{(s,s')}^{(5)} = \alpha | h_s^{(5)}, h_{s'}^{(5)}) + \lambda \|W_{\text{ent}}\|^2 + \delta \|\theta_{\text{rel}} - \theta'_{\text{rel}}\|^2,$$



bi-LSTM

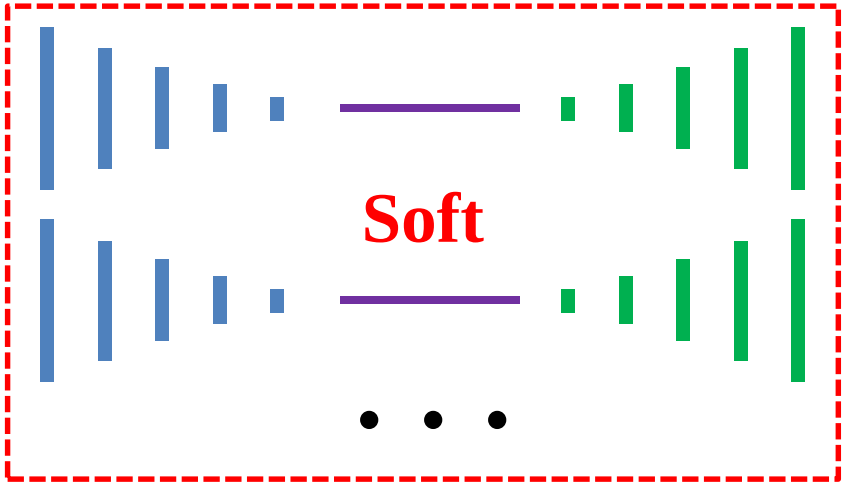
Part-Of-Speech (POS)

Dependency parsing (DEP)



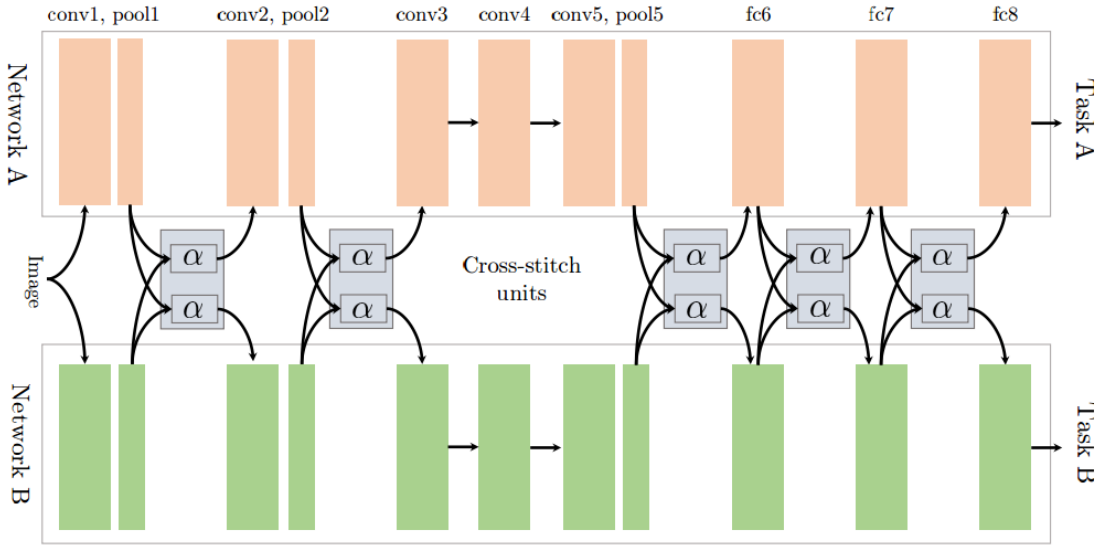
bi-LSTMs

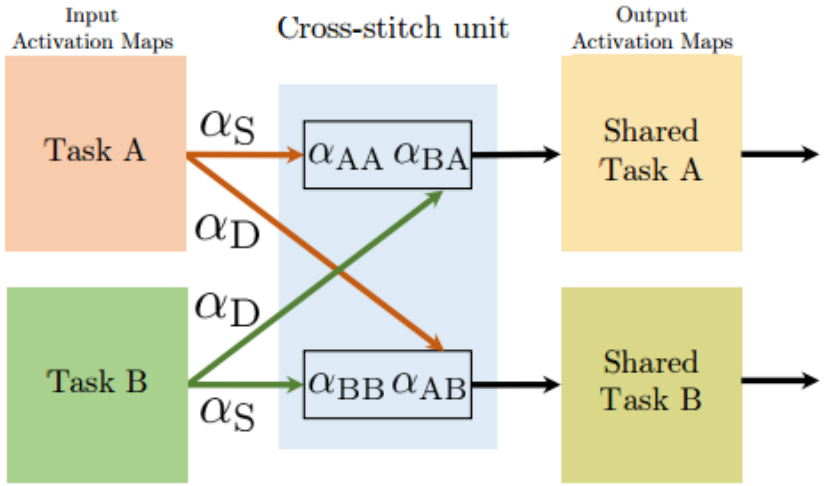
arXiv 2017
https://arxiv.org/pdf/1611.01587.pdf



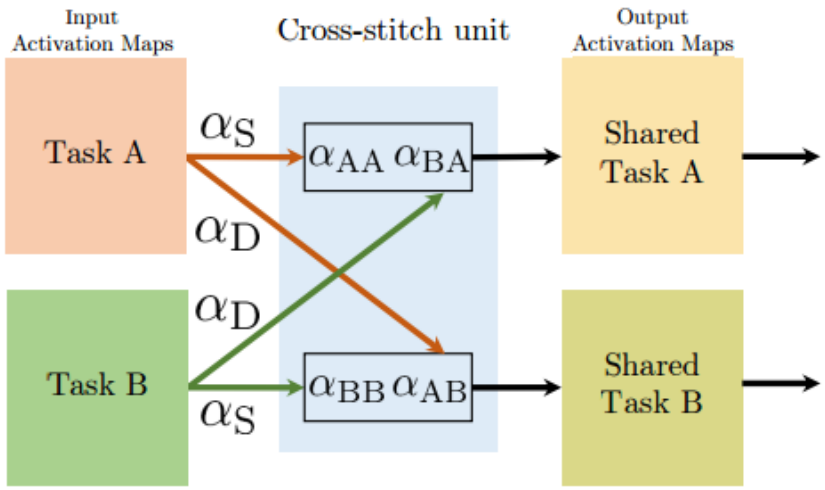
Cross-stitch Networks for Multi-task Learning

Ishan Misra* Abhinav Shrivastava* Abhinav Gupta Martial Hebert
The Robotics Institute, Carnegie Mellon University





CVPR 2016
https://arxiv.org/pdf/1604.03539.pdf



$$\begin{bmatrix} \tilde{x}_{\mathrm{A}}^{ij} \\ \tilde{x}_{\mathrm{B}}^{ij} \end{bmatrix} = \begin{bmatrix} \alpha_{\mathrm{AA}} & \alpha_{\mathrm{AB}} \\ \alpha_{BA} & \alpha_{\mathrm{BB}} \end{bmatrix} \begin{bmatrix} x_{\mathrm{A}}^{ij} \\ x_{\mathrm{B}}^{ij} \end{bmatrix}$$

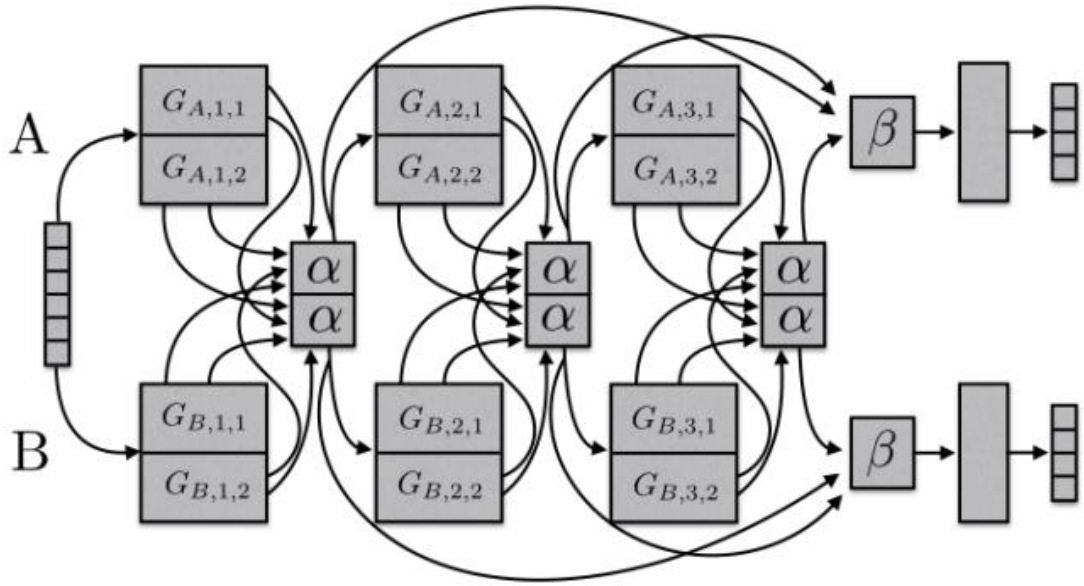
Backpropagating through cross-stitch units

$$\begin{bmatrix} \frac{\partial L}{\partial x_{\mathrm{A}}^{ij}} \\ \frac{\partial L}{\partial x_{\mathrm{B}}^{ij}} \end{bmatrix} = \begin{bmatrix} \alpha_{\mathrm{AA}} & \alpha_{\mathrm{BA}} \\ \alpha_{\mathrm{AB}} & \alpha_{\mathrm{BB}} \end{bmatrix} \begin{bmatrix} \frac{\partial L}{\partial \tilde{x}_{\mathrm{A}}^{ij}} \\ \frac{\partial L}{\partial \tilde{x}_{\mathrm{B}}^{ij}} \end{bmatrix}$$

$$\frac{\partial L}{\partial \alpha_{\mathrm{AB}}} = \frac{\partial L}{\partial \tilde{x}_{\mathrm{B}}^{ij}} x_{\mathrm{A}}^{ij}, \quad \frac{\partial L}{\partial \alpha_{\mathrm{AA}}} = \frac{\partial L}{\partial \tilde{x}_{\mathrm{A}}^{ij}} x_{\mathrm{A}}^{ij}$$

Sluice networks: Learning what to share between loosely related tasks

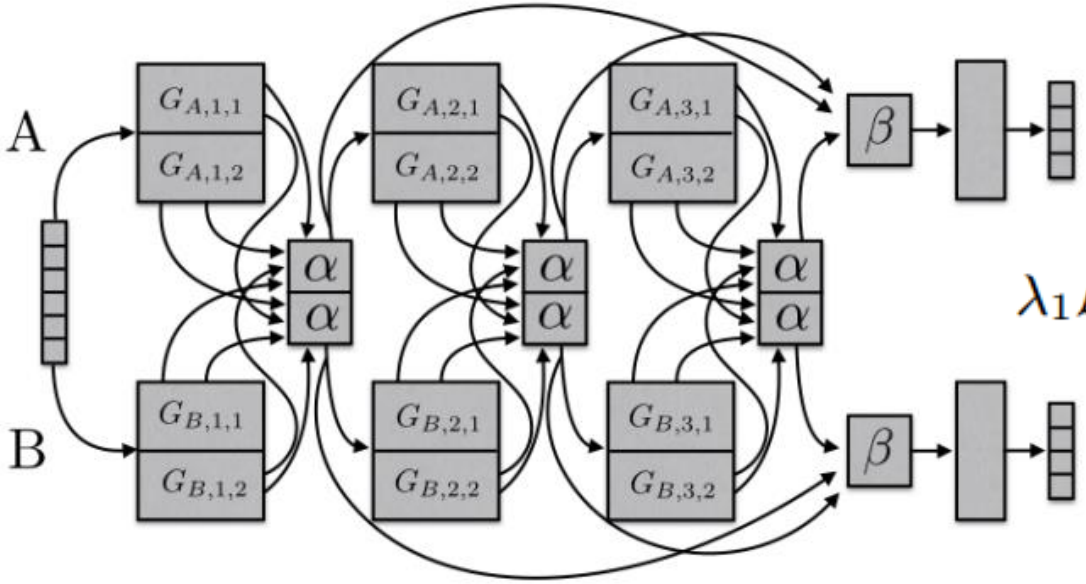
Sebastian Ruder[1,2]*, Joachim Bingel[3], Isabelle Augenstein[4]*, Anders Søgaard[3]
[1]Insight Research Centre, National University of Ireland, Galway
[2]Aylien Ltd., Dublin, Ireland
[3]Department of Computer Science, University of Copenhagen, Denmark
[4]Department of Computer Science, UCL, UK
s.ruder1@nuigalway.ie, {bingel|soegaard}@di.ku.dk, i.augenstein@ucl.ac.uk



In machine learning, it is hard to estimate if sharing will lead to improvements; especially if tasks are only loosely related

This paper introduce SLUICE NETWORKS, a general framework for multi-task learning where trainable parameters control the amount of sharing – including which parts of the models to share

arXiv 2017 May 23
https://arxiv.org/pdf/1705.08142.pdf



$$\lambda_1 \mathcal{L}_1(\mathbf{f}(x;\theta_1), z) + \ldots + \lambda_M \mathcal{L}_M(\mathbf{f}(x;\theta_M), z) + \Omega(W))$$

$$\mathcal{L}_c = \sum_{m=1}^{M} \sum_{k=1}^{K} \| {G_{m,k,1}}^{\top} G_{m,k,2} \|_F^2$$

$$\begin{bmatrix} \widetilde{h}_{\mathrm{A}_1,k} \\ \vdots \\ \widetilde{h}_{\mathrm{B}_2,k} \end{bmatrix} = \begin{bmatrix} \alpha_{\mathrm{A}_1\mathrm{A}_1} & \cdots & \alpha_{\mathrm{B}_2\mathrm{A}_1} \\ \vdots & \ddots & \vdots \\ \alpha_{\mathrm{A}_1\mathrm{B}_2} & \cdots & \alpha_{\mathrm{B}_2\mathrm{B}_2} \end{bmatrix} \left[{h_{\mathrm{A}_1,k}}^{\top}, \quad \ldots, \quad {h_{\mathrm{B}_2,k}}^{\top} \right]$$

$$\widetilde{h}_{\mathrm{A}}^{\top} = \begin{bmatrix} \beta_{\mathrm{A},1} \\ \cdots \\ \beta_{\mathrm{A},k} \end{bmatrix}^{\top} \left[{h_{\mathrm{A},1}}^{\top}, \quad \ldots \quad {h_{\mathrm{A},k}}^{\top} \right]$$

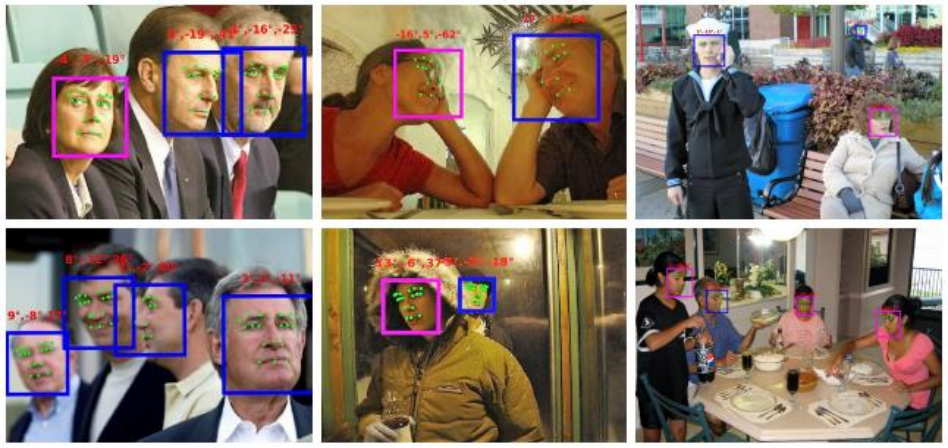
NLP tasks

Hard I

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE , VOL. XX, NO. XX, 2016 1

HyperFace: A Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition

Rajeev Ranjan, *Member, IEEE*, Vishal M. Patel, *Senior Member, IEEE*, and Rama Chellappa, *Fellow, IEEE*



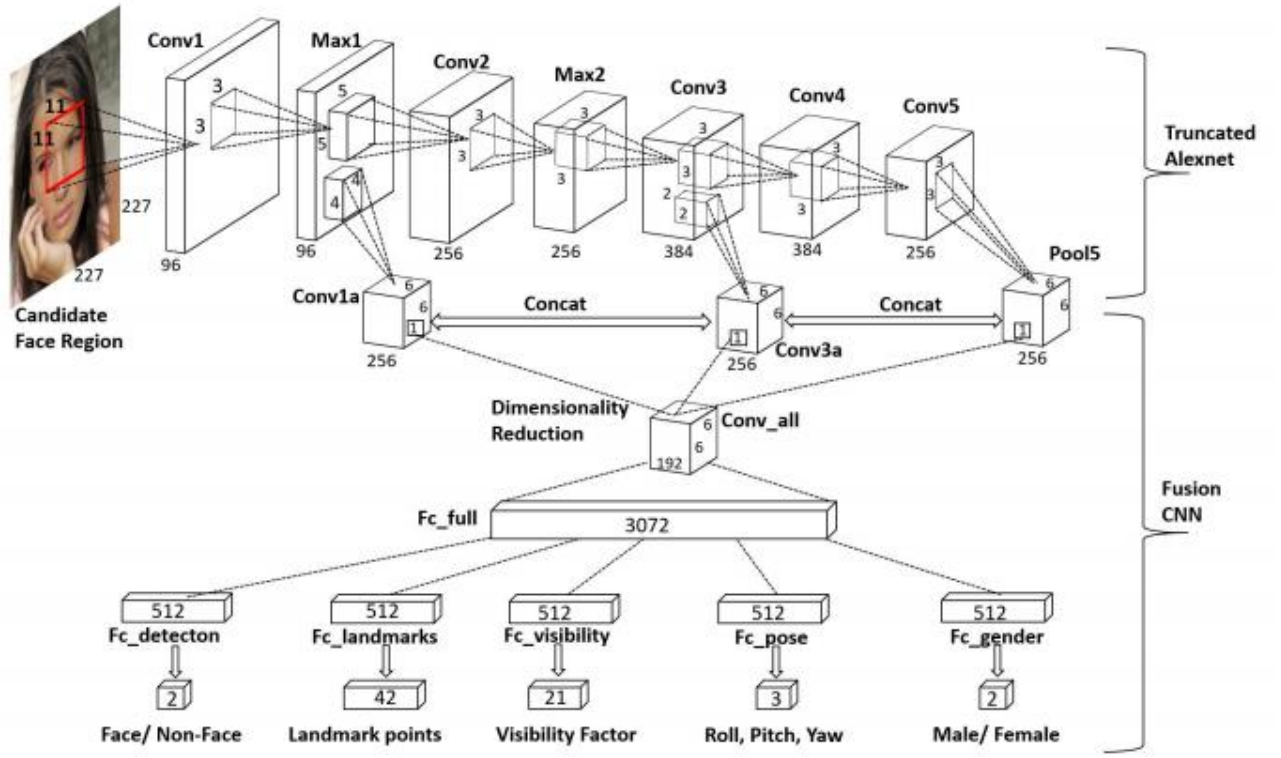


Fig. 2. The architecture of the proposed HyperFace. The network is able to classify a given image region as face or non-face, estimate the head pose, locate face landmarks and recognize gender.

PAMI 2016
https://arxiv.org/pdf/1603.01249.pdf

$$loss_D = -(1-l)\cdot\log(1-p) - l\cdot\log(p),$$

$$loss_L = \frac{1}{2N}\sum_{i=1}^{N} v_i((\hat{x}_i - a_i)^2 + ((\hat{y}_i - b_i)^2),$$

$$loss_V = \frac{1}{N}\sum_{i=1}^{N} (\hat{v}_i - v_i)^2,$$

$$loss_P = \frac{(\hat{p}_1 - p_1)^2 + (\hat{p}_2 - p_2)^2 + (\hat{p}_3 - p_3)^2}{3},$$

$$loss_G = -(1-g)\cdot\log(1-p_0) - g\cdot\log(p_1),$$

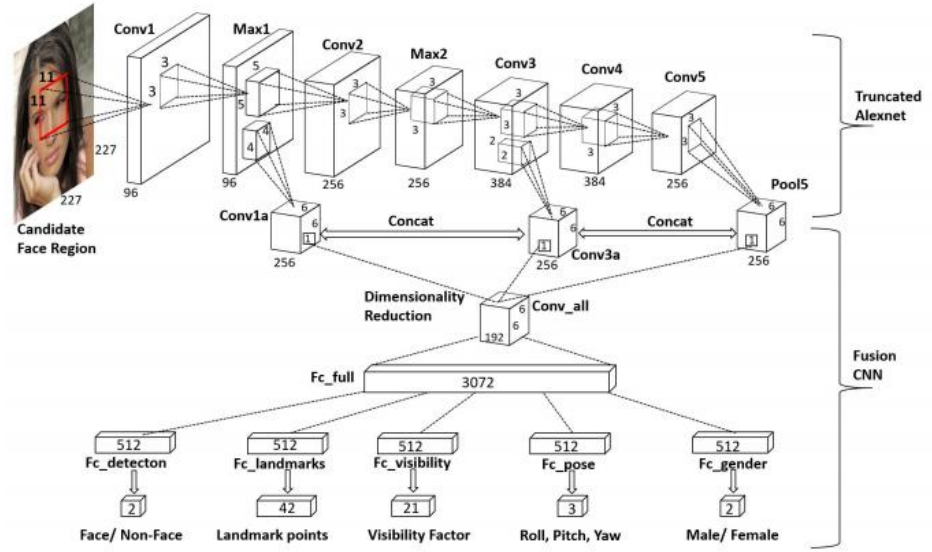


Fig. 2. The architecture of the proposed HyperFace. The network is able to classify a given image region as face or non-face, estimate the head pose, locate face landmarks and recognize gender.

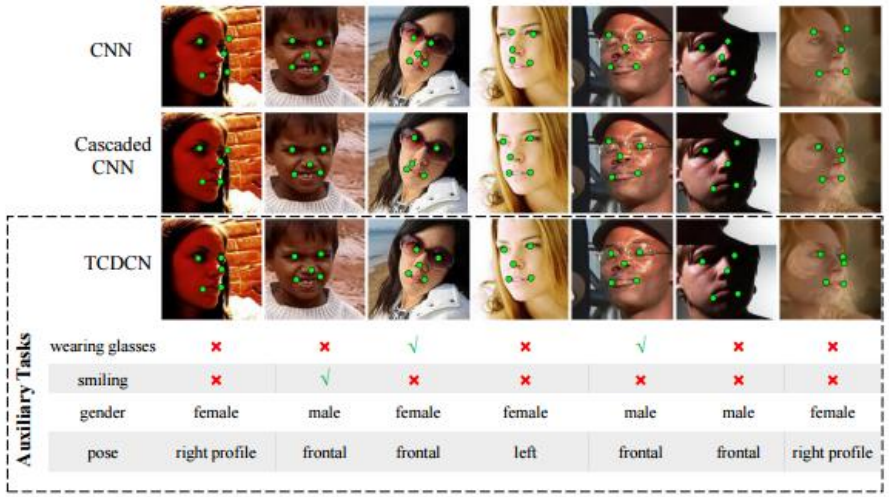
$$loss_{full} = \sum_{t=1}^{t=5} \lambda_t loss_t,$$

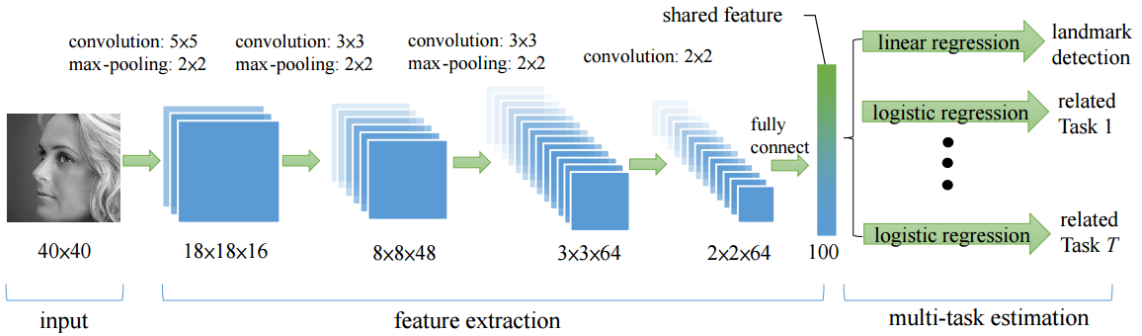
$$(\lambda_D = 1, \lambda_L = 5, \lambda_V = 0.5, \lambda_P = 5, \lambda_G = 2)$$

Facial Landmark Detection by Deep Multi-task Learning

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang

Dept. of Information Engineering, The Chinese University of Hong Kong





ECCV 2014

http://personal.ie.cuhk.edu.hk/~ccloy/files/eccv_2014_deepfacealign.pdf

$$\underset{\{\mathbf{w}^t\}_{t=1}^T}{\operatorname{argmin}} \sum_{t=1}^{T}\sum_{i=1}^{N} \ell(y_i^t, f(\mathbf{x}_i^t; \mathbf{w}^t)) + \Phi(\mathbf{w}^t),$$

In contrast to conventional MTL that maximizes the performance of all tasks, our aim is to optimize the main task r, which is facial landmark detection, with the assistances of arbitrary number of related/auxiliary tasks a ∈ A.

$$\underset{\mathbf{W}^r,\{\mathbf{W}^a\}_{a\in A}}{\operatorname{argmin}} \sum_{i=1}^{N} \ell^r(y_i^r, f(\mathbf{x}_i; \mathbf{W}^r)) + \sum_{i=1}^{N}\sum_{a\in A} \lambda^a \ell^a(y_i^a, f(\mathbf{x}_i; \mathbf{W}^a)),$$

$$\underset{\mathbf{W}^r,\{\mathbf{W}^a\}}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{N} \|\mathbf{y}_i^r - f(\mathbf{x}_i; \mathbf{W}^r)\|^2 - \sum_{i=1}^{N}\sum_{a\in A} \lambda^a y_i^a \log(p(y_i^a|\mathbf{x}_i; \mathbf{W}^a)) + \sum_{t=1}^{T} \|\mathbf{W}\|_2^2,$$

The second term is a softmax function

$$p(y_i = m|\mathbf{x}_i) = \frac{\exp\{(\mathbf{W}_m^a)^\top \mathbf{x}_i\}}{\sum_j \exp\{(\mathbf{W}_j^a)^\top \mathbf{x}_i\}}$$

Third term penalizes large weights

$$(W = \{\mathbf{W}^r, \{\mathbf{W}^a\}\})$$

stop the task if

$$\frac{k \cdot \mathrm{med}_{j=t-k}^{t} E_{tr}^{a}(j)}{\sum_{j=t-k}^{t} E_{tr}^{a}(j) - k \cdot \mathrm{med}_{j=t-k}^{t} E_{tr}^{a}(j)} \cdot \frac{E_{val}^{a}(t) - \min_{j=1..t} E_{tr}^{a}(j)}{\lambda^{a} \cdot \min_{j=1..t} E_{tr}^{a}(j)} > \epsilon,$$

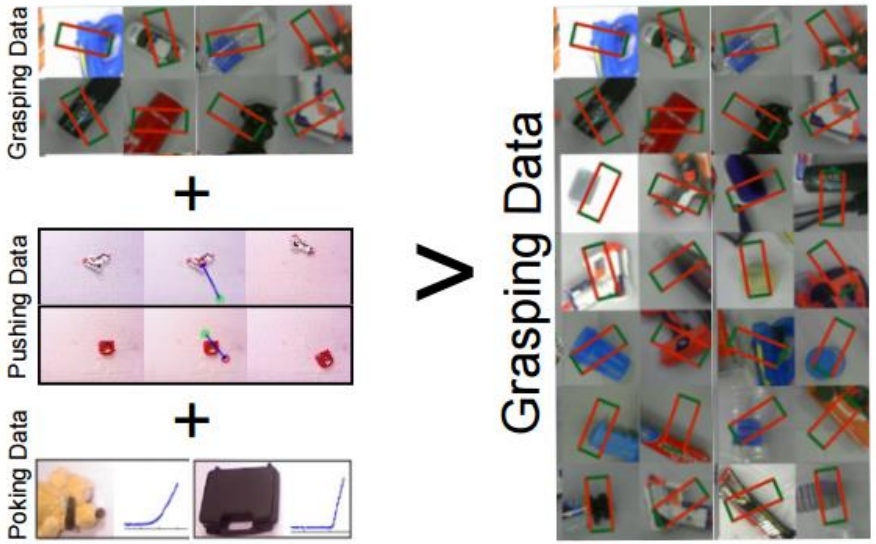
where t denotes the current iteration and k controls a training strip of length k. The 'med' denotes the function for calculating median value.

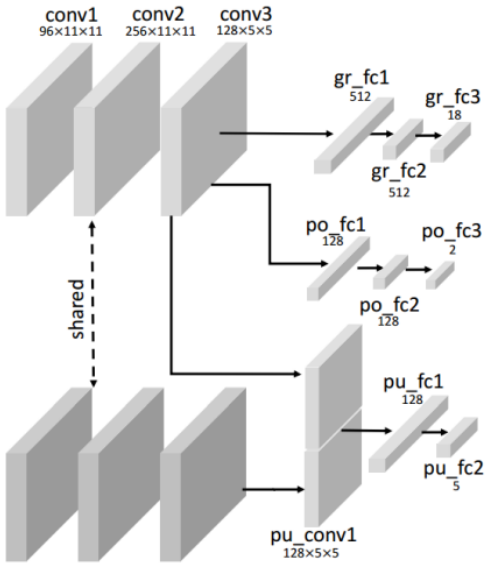
The first term represents the tendency of the training error. If the training error drops rapidly within a period of length k, the value of the first term is small, indicating that training can be continued as the task is still valuable; otherwise, the first term is large, then the task is more likely to be stopped.

The second term measures the generalization error compared to the training error. The λ a is the importance coefficient of a-th task's error, which can be learned through gradient descent. Its magnitude reveals that more important task tends to have longer impact. This strategy achieves satisfactory results for learning deep convolution network given multiple tasks.

Learning to Push by Grasping: Using multiple tasks for effective learning

Lerrel Pinto and Abhinav Gupta
Carnegie Mellon University





$$L_G = -y \times log(sig(y')) - (1-y) \times log(1 - sig(y'))$$

$$L_P = (p_A - p'_A)^2$$

$$L_{Poke} = (p_R - p'_R)^2$$

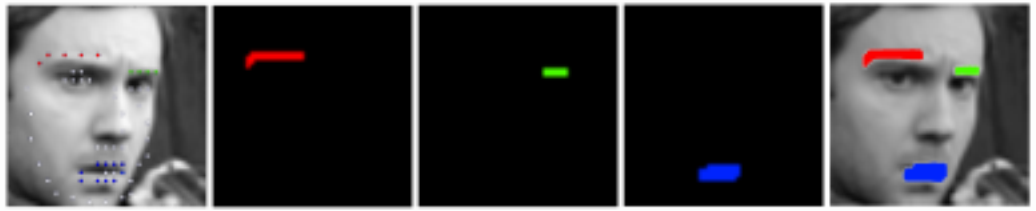
$$\frac{\partial(L_B G + L_B P + L_B Poke)}{\partial W_S}$$

ICRA 2017
https://arxiv.org/pdf/1609.09025.pdf

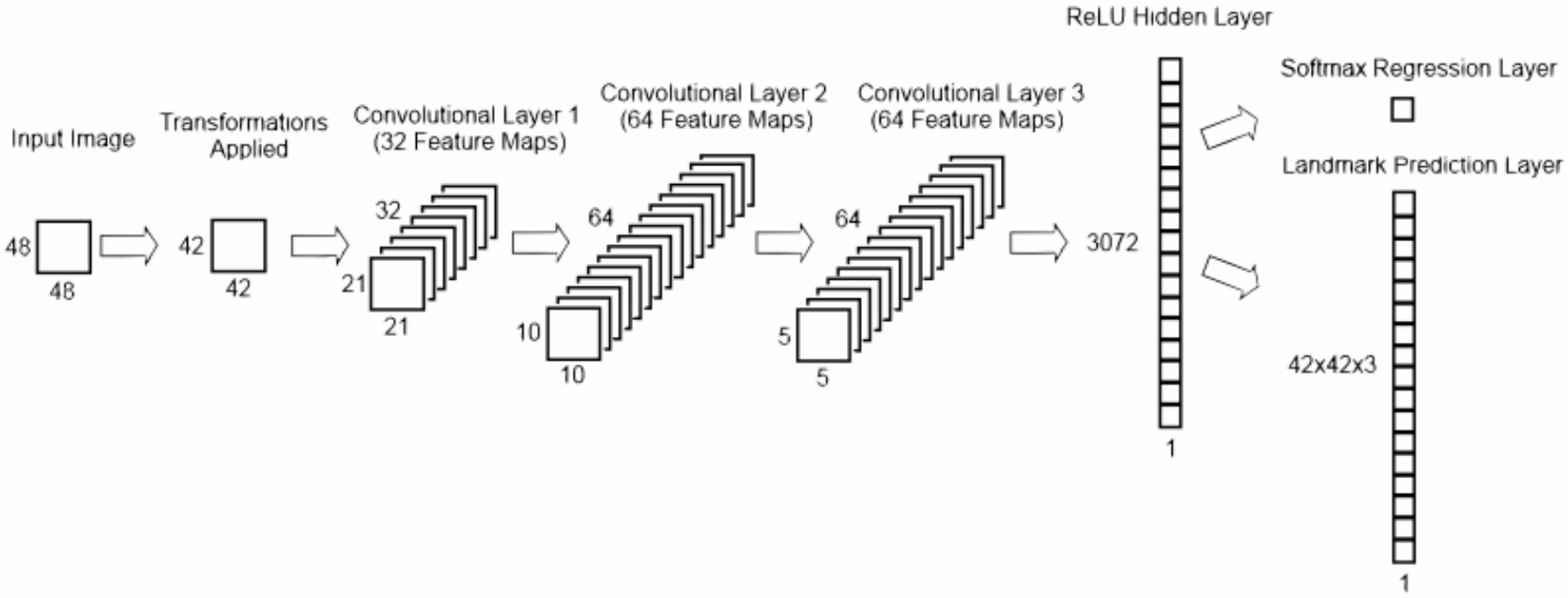
Multi-Task Learning of Facial Landmarks and Expression

Terrance Devries[1], Kumar Biswaranjan[2], and Graham W. Taylor[1]

[1]School of Engineering, University of Guelph, Guelph, Canada N1G 2W1
[2]Department of Electronics and Electrical Engineering, IIT Guwahati, Guwahati, Assam India 781039



a) b) c) d) e)

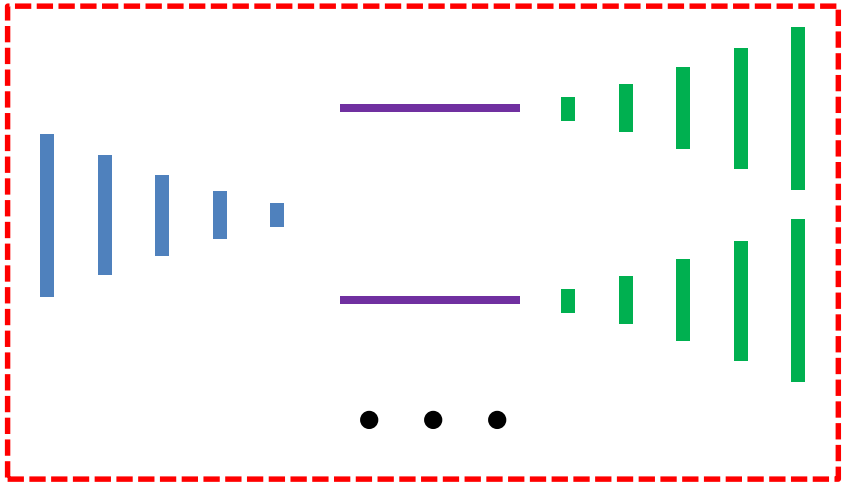


$$\mathcal{C}_e = \frac{1}{|\mathcal{D}|}\mathcal{L}(\theta, \mathcal{D}) = -\frac{1}{|\mathcal{D}|}\sum_{n=0}^{|\mathcal{D}|-1} \log\left(p(y_n = t_n | x_n, \theta)\right)$$

$$\mathcal{C}_l = \frac{1}{|\mathcal{D}|}\sum_{n=0}^{|\mathcal{D}|-1}\left[\frac{1}{\mathcal{K}}\sum_{k=0}^{\mathcal{K}-1} -\left(l_n^{(k)} \log\left(u_n^{(k)}\right) + \left(1 - l_n^{(k)}\right)\log\left(1 - u_n^{(k)}\right)\right)\right] \times v_n$$

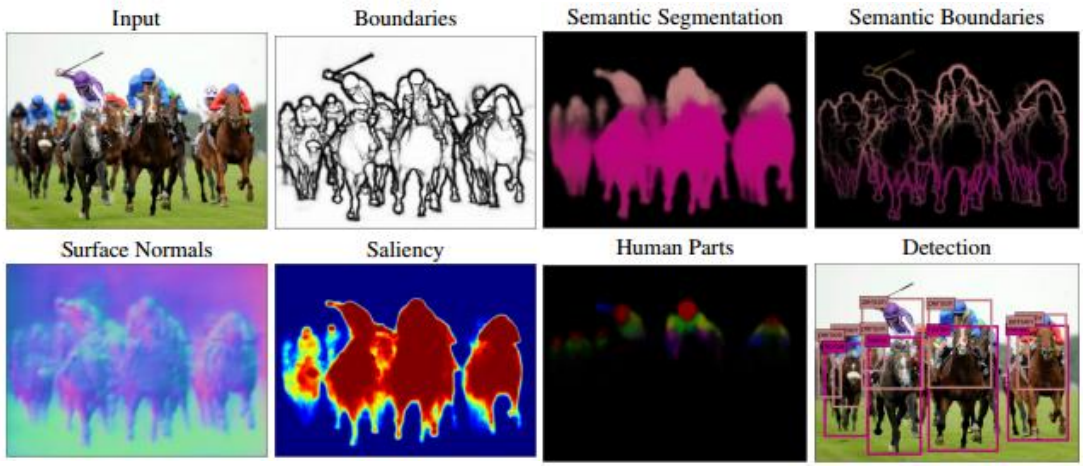
$$\mathcal{C} = \mathcal{C}_e + \lambda \mathcal{C}_l$$

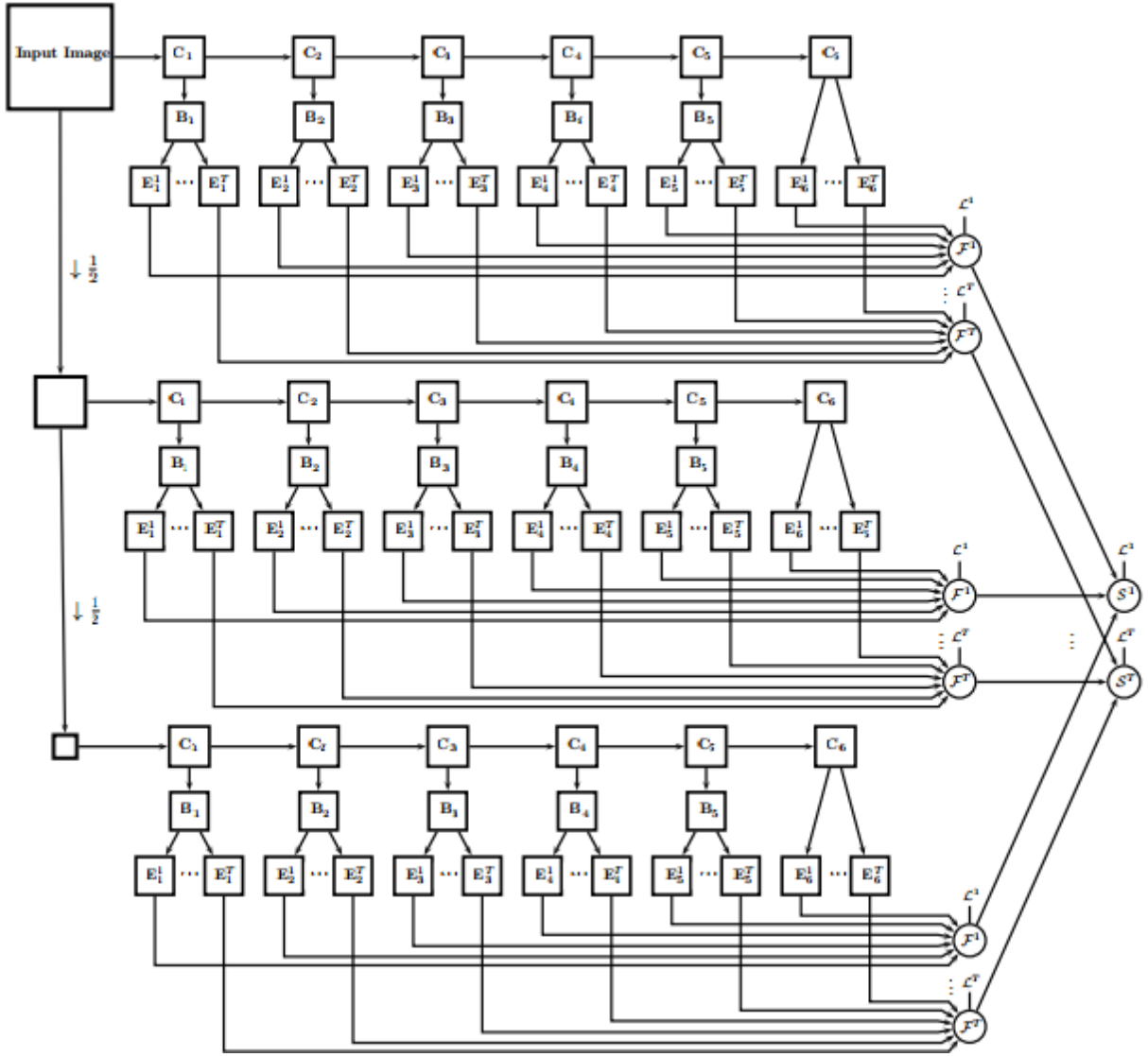
CRV 2014
http://www.uoguelph.ca/~gwtaylor/publications/devries2014multi-task.pdf



UberNet : Training a 'Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory

Iasonas Kokkinos
iasonas.kokkinos@ecp.fr
CentraleSupélec - INRIA





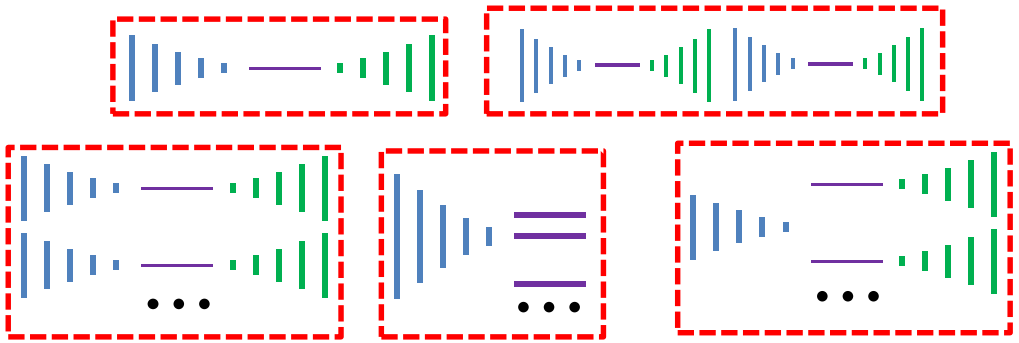
CVPR 2017
https://arxiv.org/pdf/1609.02132.pdf

$$\mathcal{L}(\mathbf{w}_{0,t_1,\ldots,t_T}) = \mathcal{R}(\mathbf{w}_0) + \sum_{t=1}^{T} \gamma_{t_k} \left(\mathcal{R}(\mathbf{w}_t) + L_t\left(\mathbf{w}_0, \mathbf{w}_t\right)\right)$$

In Eq. 3 we use t to index tasks; $\mathbf{w}_0$ denotes the weights of the common CNN trunk, and $\mathbf{w}_t$ are task-specific weights; γ_t is an hyperparameter that determines the relative importance of task t, $\mathcal{R}(\mathbf{w}_*) = \frac{\lambda}{2}\|\mathbf{w}_*\|^2$ is an ℓ_2 regularization on the relevant network weights, and $L_t\left(\mathbf{w}_0, \mathbf{w}_t\right)$ is the task-specific loss function.

$$L_t\left(\mathbf{w}_0, \mathbf{w}_t\right) = \frac{1}{N} \sum_{i=1}^{N} \delta_{t,i} L_t\left(\mathbf{f}_t^i(\mathbf{w}_0, \mathbf{w}_t), \mathbf{y}_t^i\right)$$

Short Sum



Network Structures

$$L_{total} = \sum_i \lambda_i L_i.$$

Loss Function

How to model dependency among different tasks?
How to characterize the commonalities and differences between different tasks?

Empirical ➡ Adaptive?

Fully-adaptive Feature Sharing in Multi-Task Networks with Applications in Person Attribute Classification

Yongxi Lu
UC San Diego
yol070@ucsd.edu

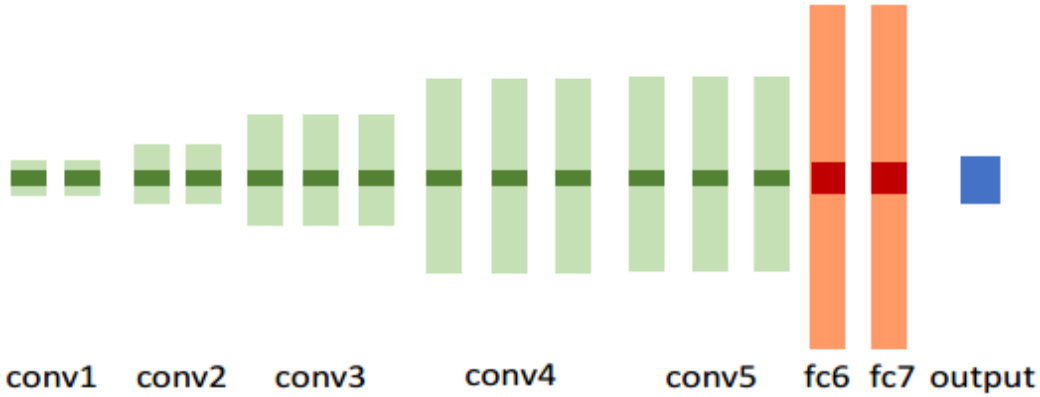
Abhishek Kumar
IBM Research
abhishk@us.ibm.com

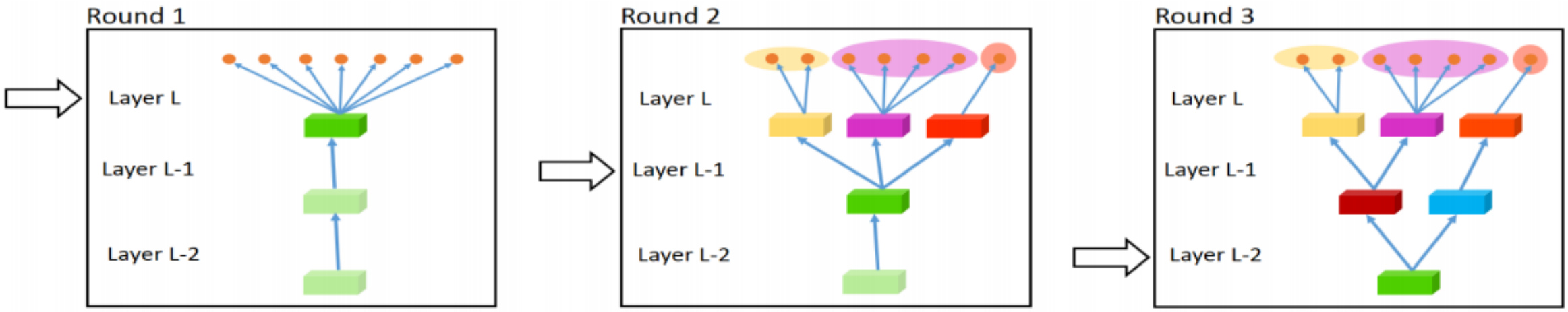
Shuangfei Zhai
Binghamton Univeristy, SUNY
szhai2@binghamton.edu

Yu Cheng
IBM Research
chengyu@us.ibm.com

Tara Javidi
UC San Diego
tjavidi@eng.ucsd.edu

Rogerio Feris
IBM Research
rsferis@us.ibm.com

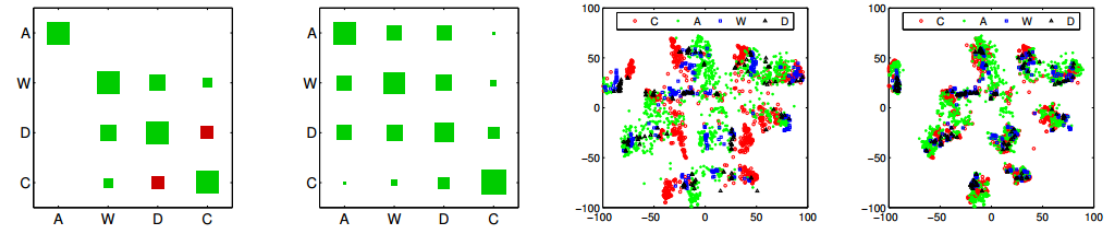




CVPR 2016
https://arxiv.org/pdf/1611.05377.pdf

Learning Multiple Tasks with Deep Relationship Networks

Mingsheng Long[†], **Jianmin Wang**[†], **Philip S. Yu**[‡]
[†]School of Software, Tsinghua University, Beijing 100084, China
[‡]University of Illinois at Chicago, Chicago, IL 60607, USA
mingsheng@tsinghua.edu.cn, jimwang@tsinghua.edu.cn, psyu@uic.edu



(a) MTRL Relationship (b) DRN Relationship (c) MTDL Representation (d) DRN Representation

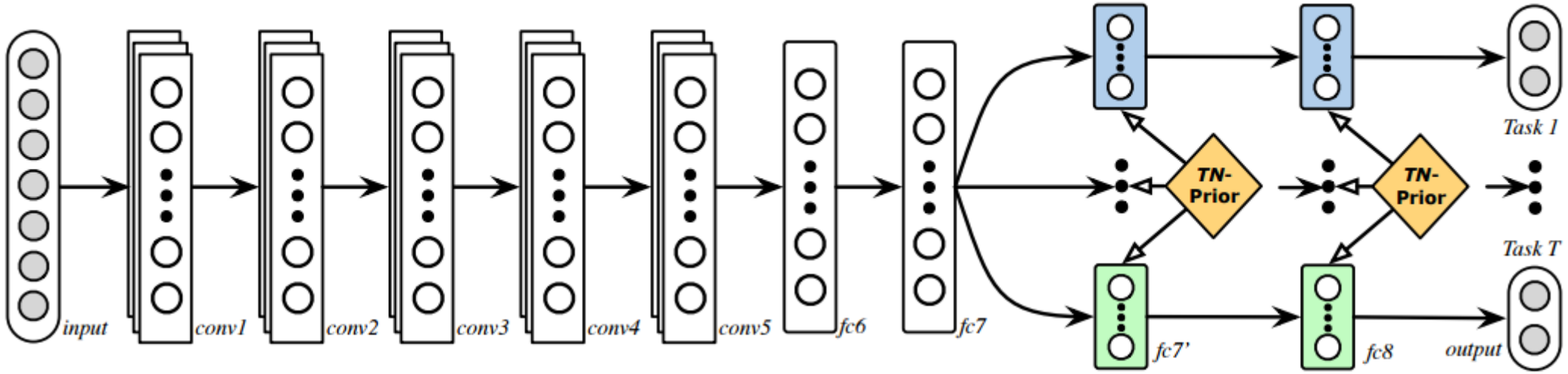


Figure 1: Deep relationship network (DRN) for multi-task learning: (1) convolutional layers $conv1$–$conv5$ and fully connected layers $fc6$–$fc7$ learn transferable features, and their parameters are shared across tasks; (2) full connected layers $fc7'$–$fc8$ are tailored to fit task-specific variations, and their parameters are modeled via tensor normal priors for learning task relationships.

arXiv 2017
https://arxiv.org/pdf/1506.02117.pdf

The Maximum a Posteriori (MAP) estimation of network parameters W given training data {X , Y} to learn multiple tasks

$$p(\mathcal{W}|\mathcal{X},\mathcal{Y}) \propto p(\mathcal{W}) \cdot p(\mathcal{Y}|\mathcal{X},\mathcal{W})$$

$$= \prod_{\ell \in \mathcal{L}} p(\mathcal{W}_\ell) \cdot \prod_{t=1}^{T} \prod_{n=1}^{N_t} p\left(\mathbf{y}_n^t | \mathbf{x}_n^t, \mathcal{W}\right),$$

Prior

MLE

Kronecker product

$$p(\mathcal{W}_\ell) = \mathcal{TN}_{D_\ell^i \times D_\ell^o \times T}\left(\mathbf{O}, \mathbf{\Sigma}_\ell^i, \mathbf{\Sigma}_\ell^o, \mathbf{\Sigma}_\ell\right),$$

$$\mathbf{\Sigma}_\ell^i \in \mathbb{R}^{D_\ell^i \times D_\ell^i}, \mathbf{\Sigma}_\ell^o \in \mathbb{R}^{D_\ell^o \times D_\ell^o}, \text{ and } \mathbf{\Sigma}_\ell \in \mathbb{R}^{T \times T}$$

$$p(\mathbf{x}) = (2\pi)^{-d/2} \prod_{k=1}^{K} |\mathbf{\Sigma}_k|^{-d/(2d_k)}$$
$$\times \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}_{1:K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$

$$\min_{f_t} \sum_{n=1}^{N_t} J\left(f_t\left(\mathbf{x}_n^t\right), \mathbf{y}_n^t\right),$$

J is the cross-entropy loss function

$f_t\left(\mathbf{x}_n^t\right)$ is the conditional probability that the CNN assigns $\mathbf{x}_n^t$ to label $\mathbf{y}_n^t$

Loss function = -log $p(\mathcal{W}|\mathcal{X},\mathcal{Y})$

In MICCAI

Less is More: Simultaneous View Classification and Landmark Detection for Abdominal Ultrasound Images

Zhoubing Xu[1(✉)], Yuankai Huo[2], JinHyeong Park[1], Bennett Landman[2], Andy Milkowski[3], Sasa Grbic[1], and Shaohua Zhou[1]

[1] Medical Imaging Technologies, Siemens Healthineers, Princeton, NJ, USA
zhoubing.xu@siemens-healthineers.com
[2] Electrical Engineering, Vanderbilt University, Nashville, TN, USA
[3] Ultrasound, Siemens Healthineers, Issaquah, WA, USA

Abstract. An abdominal ultrasound examination, which is the most common ultrasound examination, requires substantial manual efforts to acquire standard abdominal organ views, annotate the views in texts, and record clinically relevant organ measurements. Hence, automatic view classification and landmark detection of the organs can be instrumental to streamline the examination workflow. However, this is a challenging problem given not only the inherent difficulties from the ultrasound modality, e.g., low contrast and large variations, but also the heterogeneity across tasks, i.e., one classification task for all views, and then one landmark detection task for each relevant view. While convolutional neural networks (CNN) have demonstrated more promising outcomes on ultrasound image analytics than traditional machine learning approaches, it becomes impractical to deploy multiple networks (one for each task) due to the limited computational and memory resources on most existing ultrasound scanners. To overcome such limits, we propose a multi-task learning framework to handle all the tasks by a single network. This network is integrated to perform view classification and landmark detection simultaneously; it is also equipped with global convolutional kernels, coordinate constraints, and a conditional adversarial module to leverage the performances. In an experimental study based on 187,219 ultrasound images, with the proposed simplified approach we achieve (1) view classification accuracy better than the agreement between two clinical experts and (2) landmark-based measurement errors on par with inter-user variability. The multi-task approach also benefits from sharing the feature extraction during the training process across all tasks and, as a result, outperforms the approaches that address each task individually.

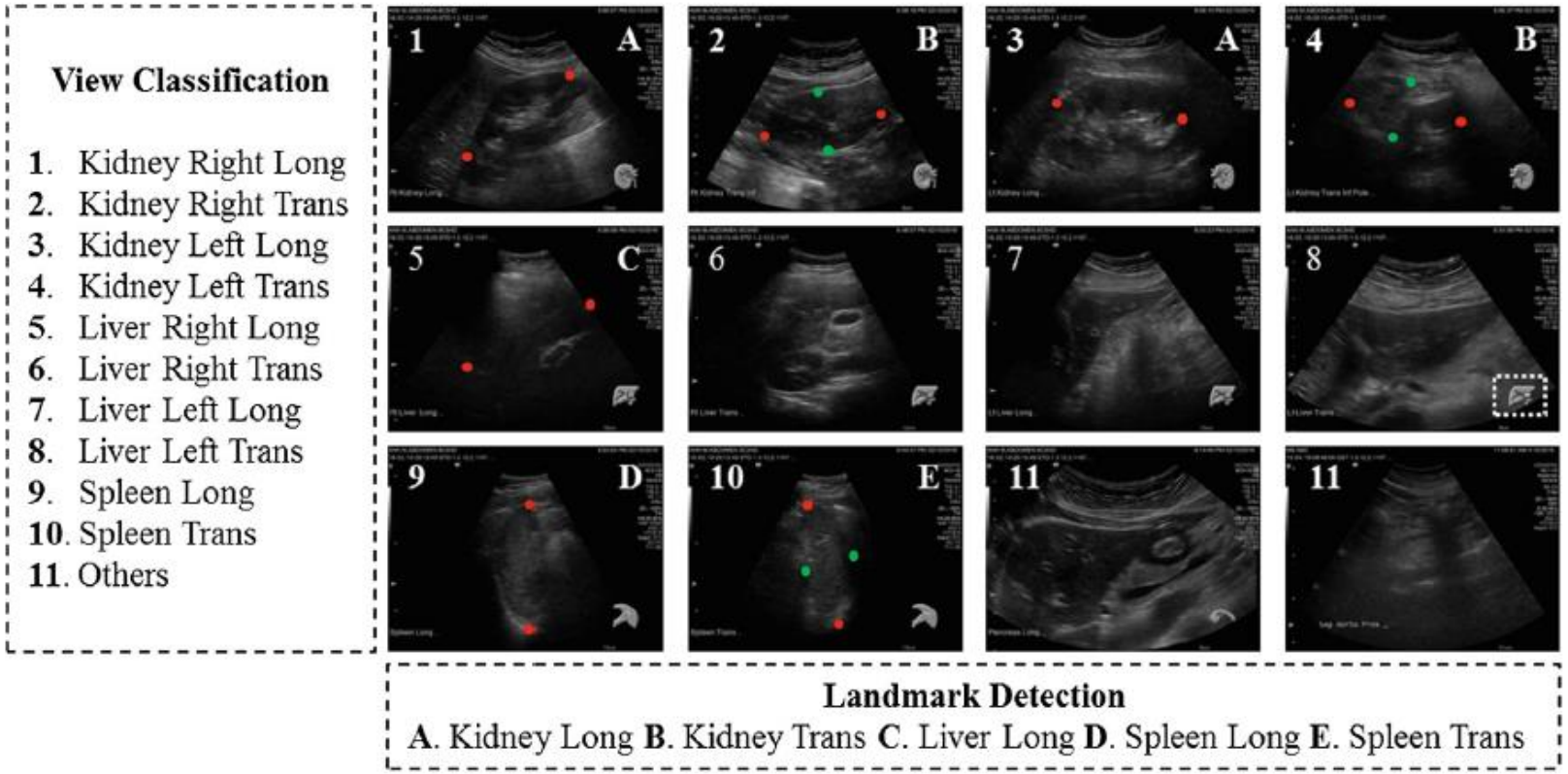


Fig. 1. An overview of the tasks for abdominal ultrasound analytics. In each image, the upper left corner indicates its view type. If present, the upper right corner indicates the associated landmark detection task, and the pairs of long- and short-axis landmarks are colored in red and green, respectively. An icon is circled on one image; such icons are masked out when training the view classification.

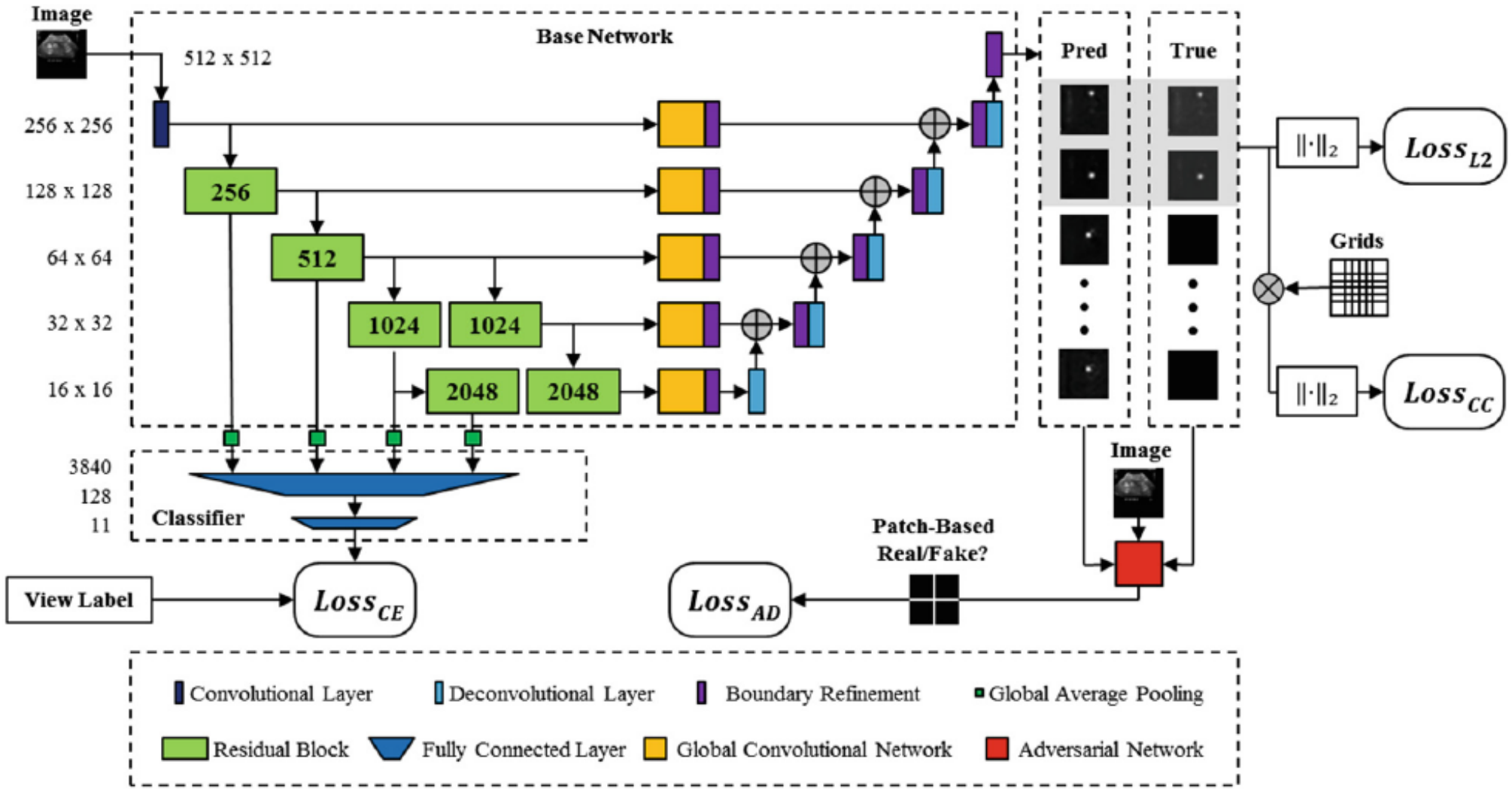


Fig. 2. An illustration of the proposed MTL Framework.

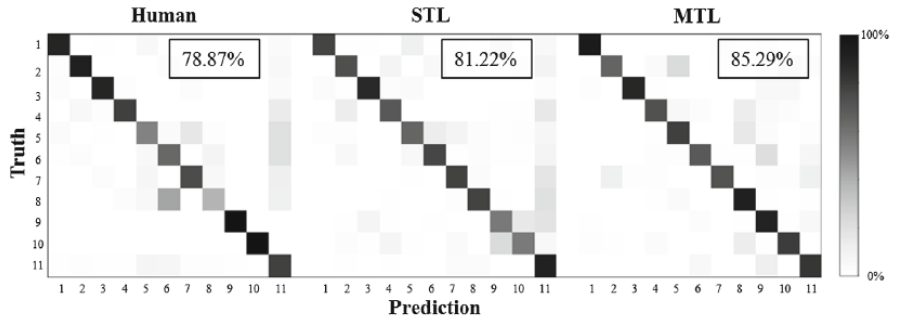


Fig. 3. Confusion matrices for the view classification task of STL, MTL, and between humans. The numbers on x and y axes follow the view definitions in Fig. 1. The overall classification accuracy is overlaid for each approach. The diagonal entries indicate correct classifications.

Table 1. Quantitative comparison for landmark-based measurements in mm.

	KL_LA	KT_LA	KT_SA	LL_LA	SL_LA	ST_LA	ST_SA
Human	4.500	5.431	4.283	5.687	6.104	4.578	4.543
PBT [12]	11.036	9.147	8.393	11.083	7.289	9.359	12.308
SFCN	7.044	7.332	5.189	10.731	8.693	91.309	43.773
MFCN	10.582	16.508	15.942	17.561	8.856	49.887	29.167
MGCN	10.628	5.535	5.348	9.46	7.718	12.284	19.79
MGCN_R	**4.278**	**4.426**	**3.437**	**6.989**	**3.610**	**7.923**	**7.224**

Note that LA and SA represent for long- and short-axis measurements. KL, KT, LL, SL, and ST stand for Kidney Long, Kidney Trans, Liver Long, Spleen Long, and Spleen Trans, respectively. For the methods, the prefix S and M represent single-task and multi-task, respectively, while FCN and GCN are both based on ResNet50 except that the later embeds large kernels and boundary refinement in skip connection. MGCN_R is the proposed method that includes two additional regularization modules. PBT is a traditional machine learning approach. The human statistics are computed on a subset of images for reference.

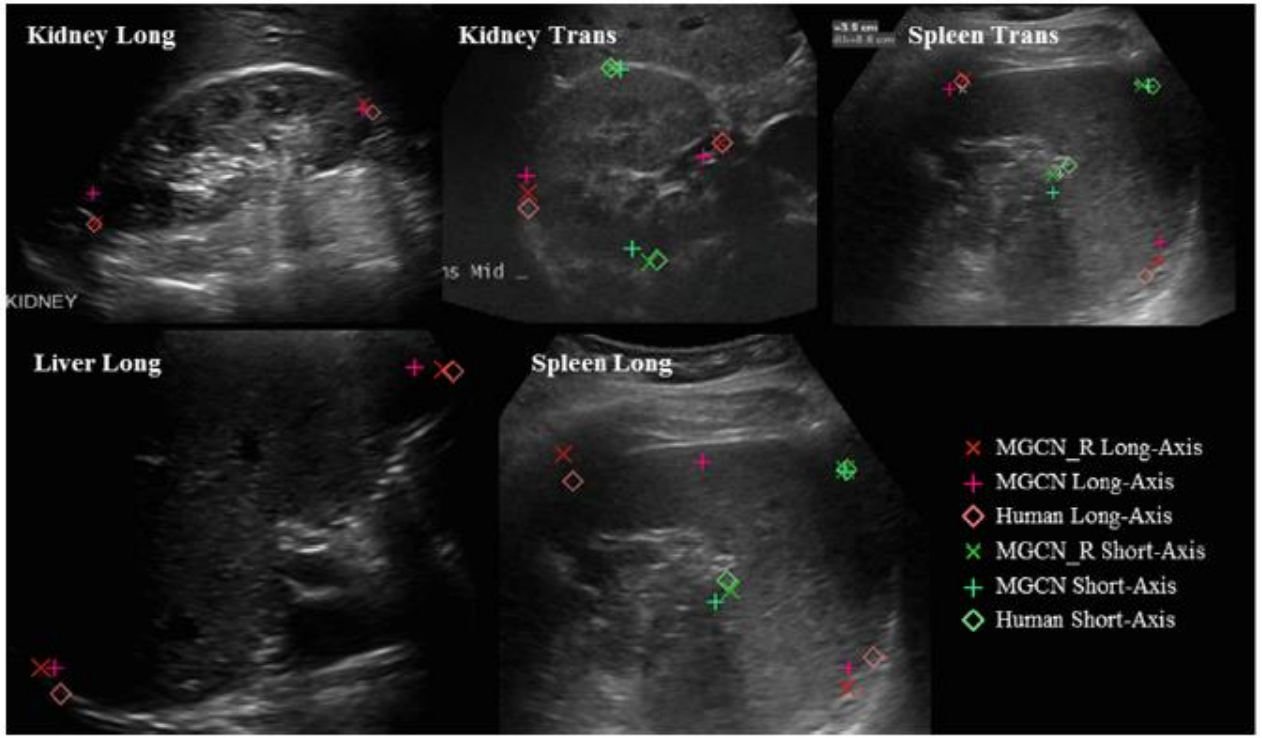


Fig. 4. Qualitative comparison of landmark detection results with and without regularization for the proposed approach. Images are zoomed into region of interest for better visualization

Integrate Domain Knowledge in Training CNN for Ultrasonography Breast Cancer Diagnosis

Jiali Liu[1,2], Wanyu Li[3], Ningbo Zhao[4], Kunlin Cao[3(✉)], Youbing Yin[3(✉)], Qi Song[3], Hanbo Chen[3], and Xuehao Gong[1,2(✉)]

[1] Shenzhen Second People's Hospital, Shenzhen, Guangdong, China
fox_gxh@sina.com
[2] Anhui Medical University, Heifei, Anhui, China
[3] Shenzhen Keya Medical Technology Corporation, Shenzhen, Guangdong, China
{cao,yin}@keyayun.com
[4] The Third People's Hospital of Shenzhen, Shenzhen, Guangdong, China

Abstract. Breast cancer is the most common cancer in women, and ultrasound imaging is one of the most widely used approach for diagnosis. In this paper, we proposed to adopt Convolutional Neural Network (CNN) to classify ultrasound images and predict tumor malignancy. CNN is a successful algorithm for image recognition tasks and has achieved human-level performance in real applications. To improve the performance of CNN in breast cancer diagnosis, we integrated domain knowledge and conducted multi-task learning in the training process. After training, a radiologist visually inspected the class activation map of the last convolutional layer of trained network to evaluate the result. Our result showed that CNN classifier can not only give reasonable performance in predicting breast cancer, but also propose potential lesion regions which can be integrated into the breast ultrasound system in the future.

Keywords: Breast cancer · Ultrasound · BI-RADS assessments Convolutional neural network · Multi-task learning

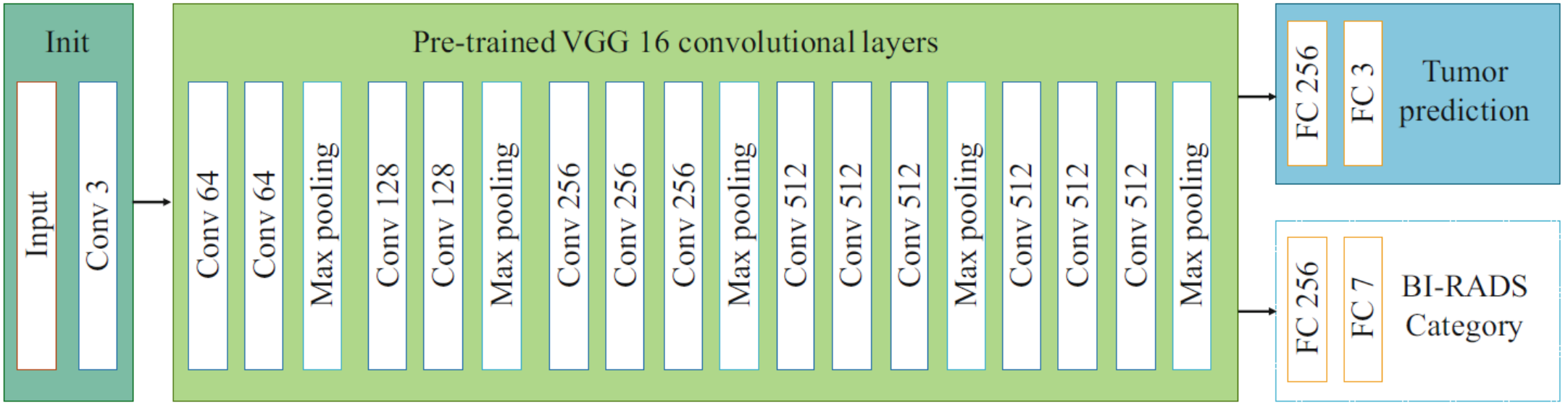


Fig. 1. Illustration of the multi-task CNN applied. The numbers of hidden units/convolutional kernels applied in each layer were shown in the figure.

A Multitask Learning Architecture for Simultaneous Segmentation of Bright and Red Lesions in Fundus Images

Clément Playout[1(✉)], Renaud Duval[2], and Farida Cheriet[1]

[1] LIV4D, École Polytechnique de Montréal, Montreal, Canada
clement.playout@polymtl.ca

[2] CUO-Hôpital Maisonneuve Rosemont, Montreal, Canada

Abstract. Recent CNN architectures have established state-of-the-art results in a large range of medical imaging applications. We propose an extension to the U-Net architecture relying on multi-task learning: while keeping a single encoding module, multiple decoding modules are used for concurrent segmentation tasks. We propose improvements of the encoding module based on the latest CNN developments: residual connections at every scale, mixed pooling for spatial compression and large kernels for convolutions at the lowest scale. We also use dense connections within the different scales based on multi-size pooling regions. We use this new architecture to jointly detect and segment red and bright retinal lesions which are essential biomarkers of diabetic retinopathy. Each of the two categories is handled by a specialized decoding module. Segmentation outputs are refined with conditional random fields (CRF) as RNN and the network is trained end-to-end with an effective Kappa-based function loss. Preliminary results on a public dataset in the segmentation task on red (resp. bright) lesions shows a sensitivity of 66,9% (resp. 75,3%) and a specificity of 99,8% (resp. 99,9%).

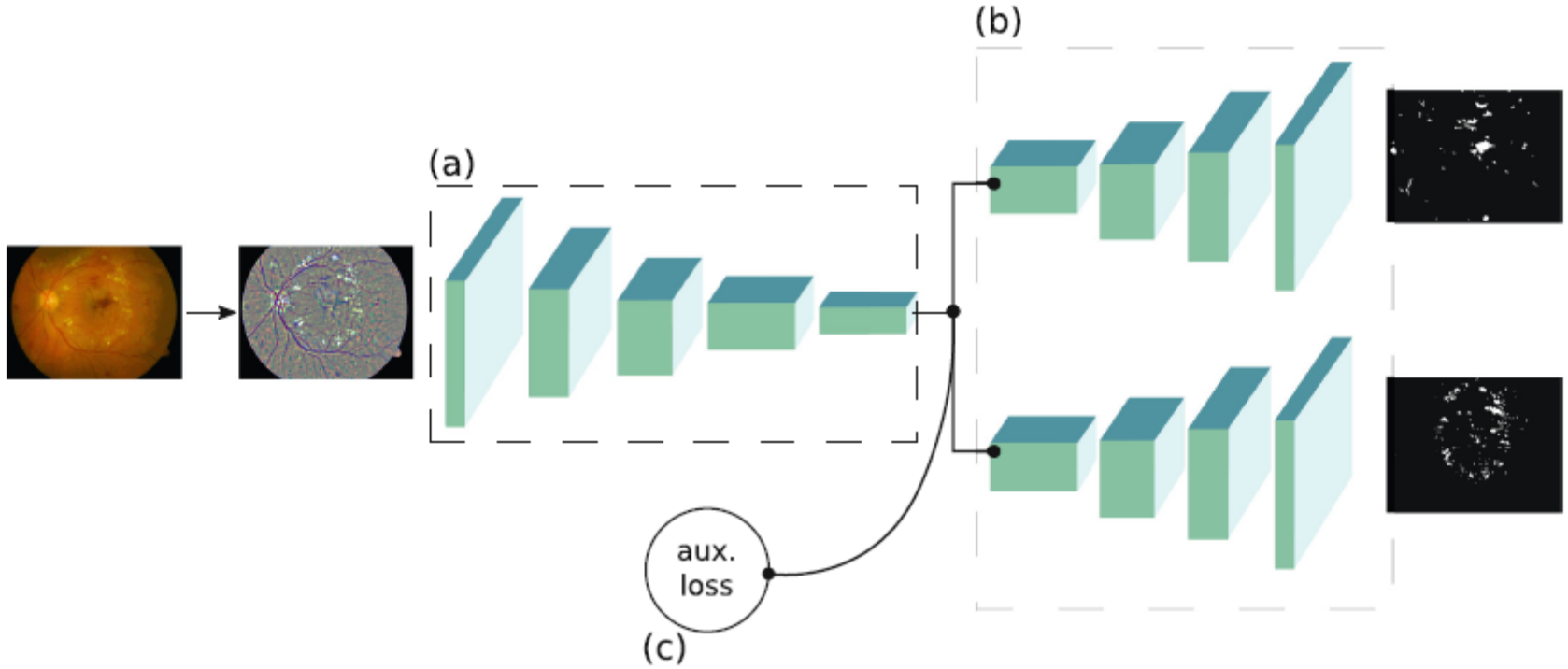


Fig. 1. The network is fed patches from the normalized images. (a) The *encoding module* uses a generic set of parameters shared by the two tasks. (b) The *decoding modules* are task-specific. An auxiliary cost (c) is added at the end of the encoding module; it is trained only to predict the presence of lesions.

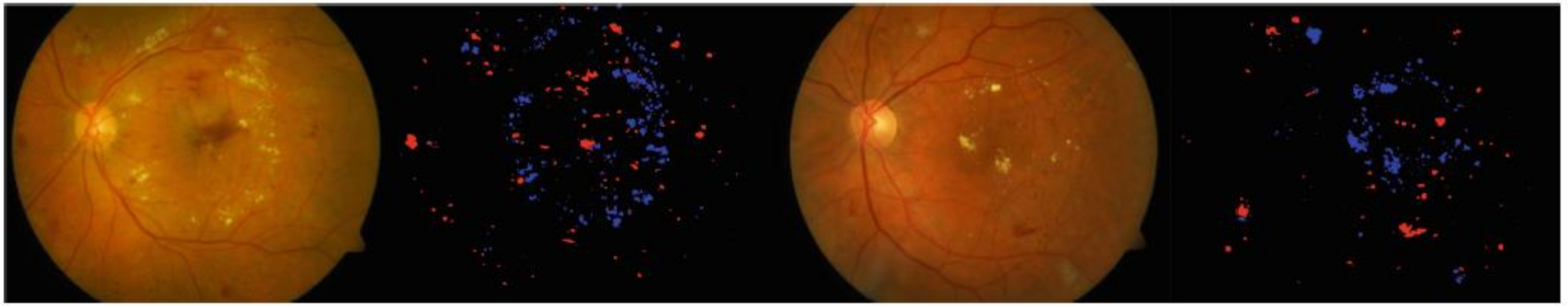


Fig. 4. Some results showing good performance overall but with over-segmentation of red lesions (false positives). One source of errors (observable in the first image) comes from laser coagulation marks, similar to small hemorrhages.

Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images

Sachin Mehta[1(✉)], Ezgi Mercan[1], Jamen Bartlett[2], Donald Weaver[2], Joann G. Elmore[3], and Linda Shapiro[1]

[1] University of Washington, Seattle, WA 98195, USA
{sacmehta,ezgi,shapiro}@cs.washington.edu
[2] University of Vermont, Burlington 05405, USA
{jamen.bartlett,donald.weaver}@uvmhealth.org
[3] University of California, Los Angeles, CA 90095, USA
jelmore@mednet.ucla.edu

Abstract. In this paper, we introduce a conceptually simple network for generating discriminative tissue-level segmentation masks for the purpose of breast cancer diagnosis. Our method efficiently segments different types of tissues in breast biopsy images while simultaneously predicting a discriminative map for identifying important areas in an image. Our network, Y-Net, extends and generalizes U-Net by adding a parallel branch for discriminative map generation and by supporting convolutional block modularity, which allows the user to adjust network efficiency without altering the network topology. Y-Net delivers state-of-the-art segmentation accuracy while learning 6.6× fewer parameters than its closest competitors. The addition of descriptive power from Y-Net's discriminative segmentation masks improve diagnostic classification accuracy by 7% over state-of-the-art methods for diagnostic classification. Source code is available at: https://sacmehta.github.io/YNet.

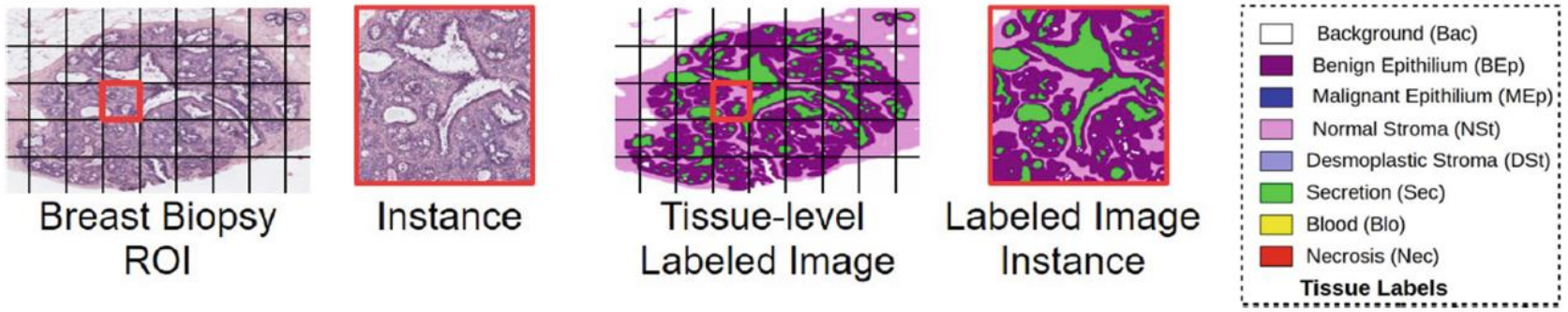


Fig. 1. This figure shows (at left) the breast biopsy ROI with H&E staining broken into multiple instances with one instance enlarged to show more detail. On the right are the pixel-wise tissue-level labelings of the ROI and the instance.

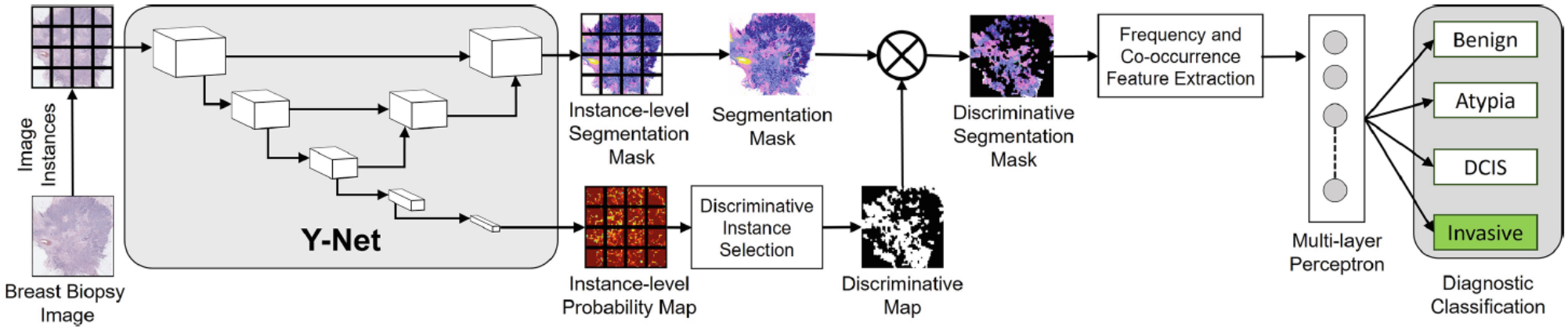
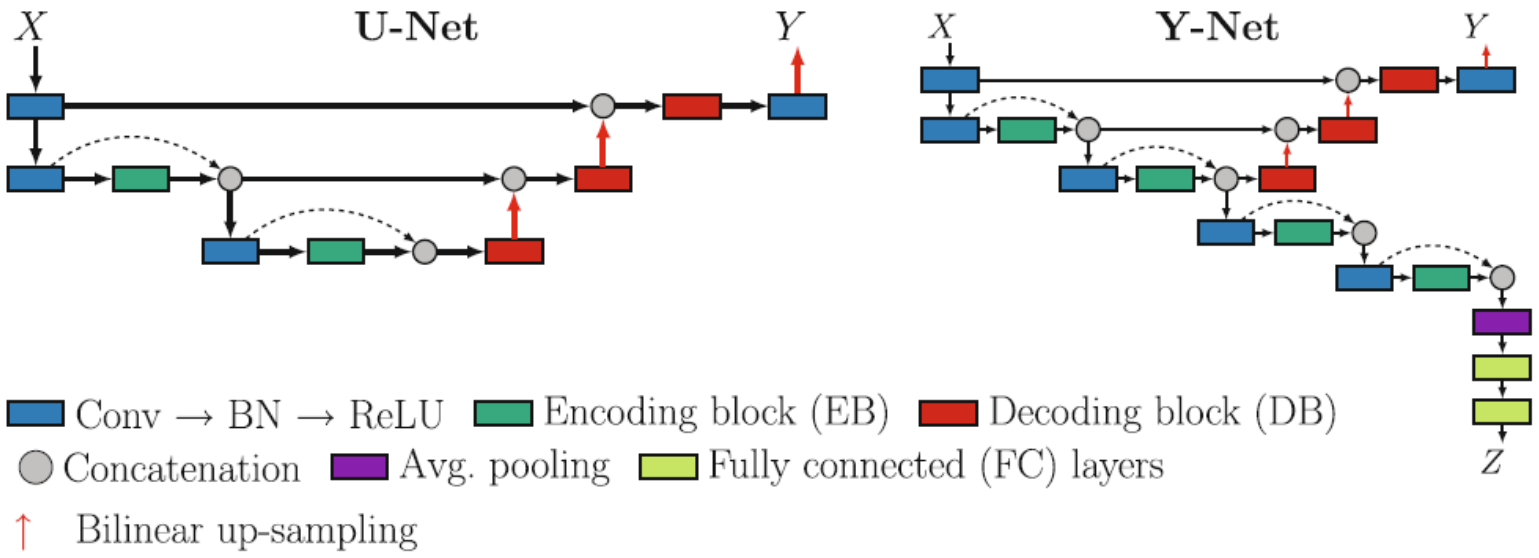


Fig. 2. Overview of our method for detecting breast cancer.



(a) U-Net vs. Y-Net

Spatial Level	Block Name	# Blocks (depth)	# Channels (width)
1	Conv	1	16
2	Conv	1	w
	EB	2	$2w$
3	Conv	1	$2w$
	EB	d	$4w$
4	Conv	1	w
	EB	2	$2w$
5	Conv	1	$w/2$
	EB	2	w
6	Avg. Pool	1	w
	FC	1	64
	FC	1	Classes

(b) Encoding Network

Fig. 3. (a) Comparison between U-Net and Y-Net architectures. (b) The encoding network architecture used in (a). U-Net in (a) is a generalized version of U-Net [11].

Training Medical Image Analysis Systems like Radiologists

Gabriel Maicas[1(✉)], Andrew P. Bradley[2], Jacinto C. Nascimento[3], Ian Reid[1] and Gustavo Carneiro[1]

[1] Australian Institute for Machine Learning, School of Computer Science, The University of Adelaide, Adelaide, Australia
gabriel.maicas@adelaide.edu.au

[2] Science and Engineering Faculty, Queensland University of Technology, Brisbane, Australia

[3] Institute for Systems and Robotics, Instituto Superior Tecnico, Lisbon, Portugal

Abstract. The training of medical image analysis systems using machine learning approaches follows a common script: collect and annotate a large dataset, train the classifier on the training set, and test it on a hold-out test set. This process bears no direct resemblance with radiologist training, which is based on solving a series of tasks of increasing difficulty, where each task involves the use of significantly smaller datasets than those used in machine learning. In this paper, we propose a novel training approach inspired by how radiologists are trained. In particular, we explore the use of meta-training that models a classifier based on a series of tasks. Tasks are selected using teacher-student curriculum learning, where each task consists of simple classification problems containing small training sets. We hypothesize that our proposed meta-training approach can be used to pre-train medical image analysis models. This hypothesis is tested on the automatic breast screening classification from DCE-MRI trained with weakly labeled datasets. The classification performance achieved by our approach is shown to be the best in the field for that application, compared to state of art baseline approaches: DenseNet, multiple instance learning and multi-task learning.

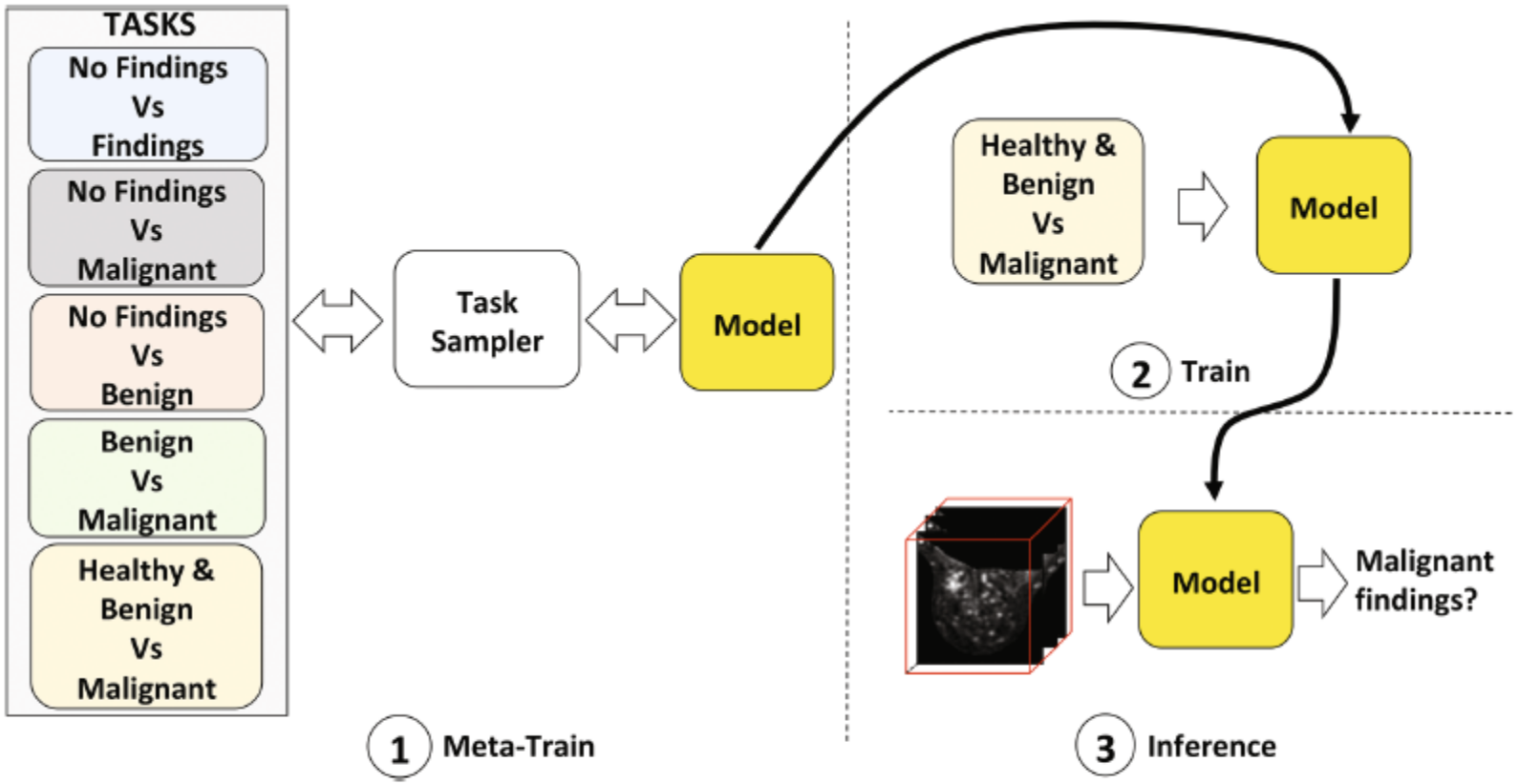


Fig. 1. The model is first meta-trained using several tasks containing relatively small training sets. The meta-trained model is then used to initialize the usual training process for breast screening (i.e., healthy and benign versus malignant). The probability of malignancy is estimated from a forward pass during the inference process.

Joint Learning of Motion Estimation and Segmentation for Cardiac MR Image Sequences

Chen Qin[1(✉)], Wenjia Bai[1], Jo Schlemper[1], Steffen E. Petersen[2], Stefan K. Piechnik[3], Stefan Neubauer[3], and Daniel Rueckert[1]

[1] Department of Computing, Imperial College London, London, UK
c.qin15@imperial.ac.uk
IHR Biomedical Research Centre at Barts, Queen Mary University of London, London, UK
[3] Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

Abstract. Cardiac motion estimation and segmentation play important roles in quantitatively assessing cardiac function and diagnosing cardiovascular diseases. In this paper, we propose a novel deep learning method for joint estimation of motion and segmentation from cardiac MR image sequences. The proposed network consists of two branches: a cardiac motion estimation branch which is built on a novel unsupervised Siamese style recurrent spatial transformer network, and a cardiac segmentation branch that is based on a fully convolutional network. In particular, a joint multi-scale feature encoder is learned by optimizing the segmentation branch and the motion estimation branch simultaneously. This enables the weakly-supervised segmentation by taking advantage of features that are unsupervisedly learned in the motion estimation branch from a large amount of unannotated data. Experimental results using cardiac MlRI images from 220 subjects show that the joint learning of both tasks is complementary and the proposed models outperform the competing methods significantly in terms of accuracy and speed.

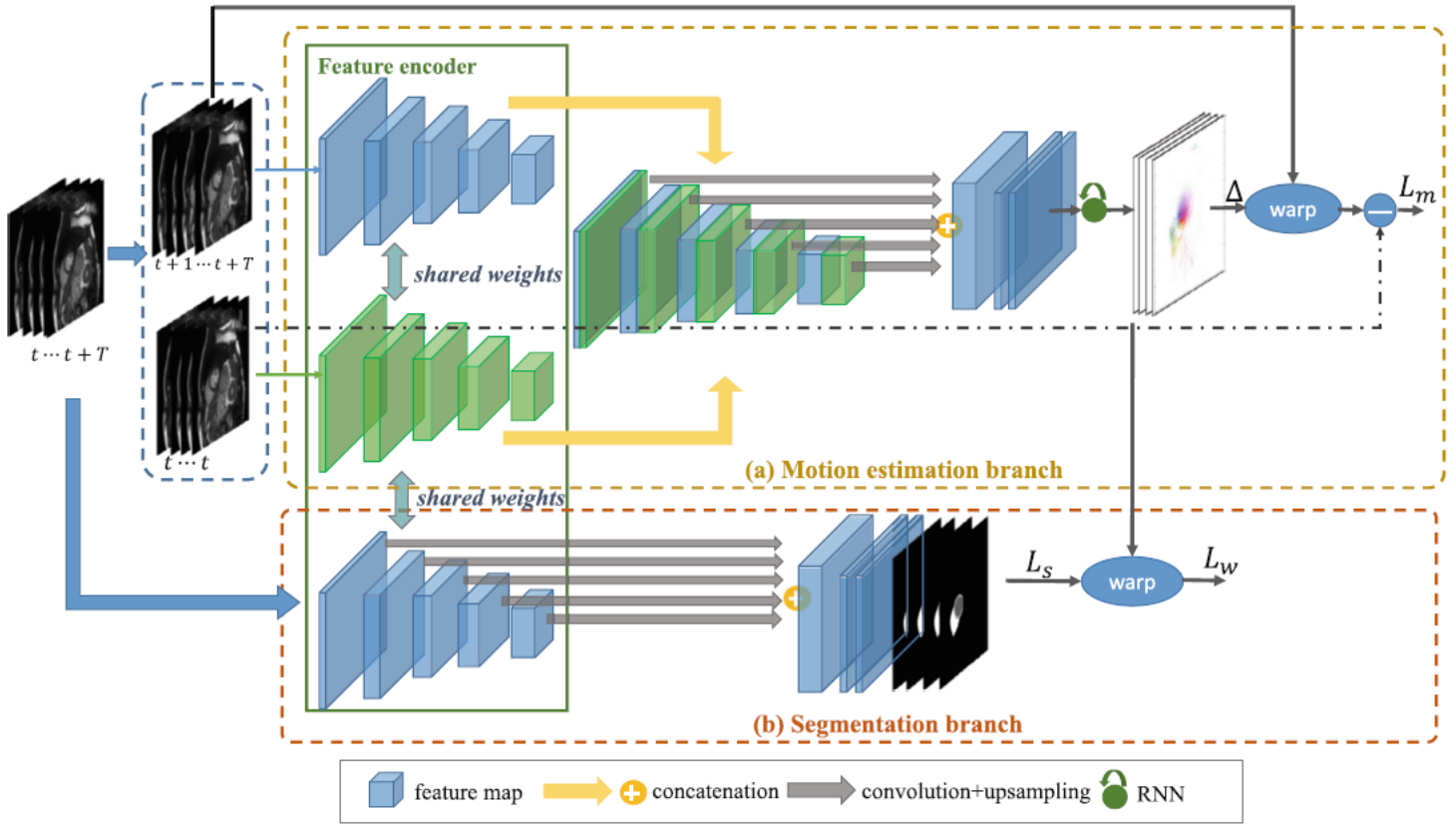


Fig. 1. The overall schematic architecture of proposed network for joint estimation of cardiac motion and segmentation. (a) The proposed Siamese style multi-scale recurrent motion estimation branch. (b) The segmentation branch which shares the joint feature encoder with motion estimation branch. The architecture for feature encoder is adopted from VGG-16 net before FC layer. Both branches have the same head architecture as the one proposed in [4], and the concatenation layers of motion estimation branch are from last layers at different scales of the feature encoder. For detailed architecture, please refer to supplementary material.

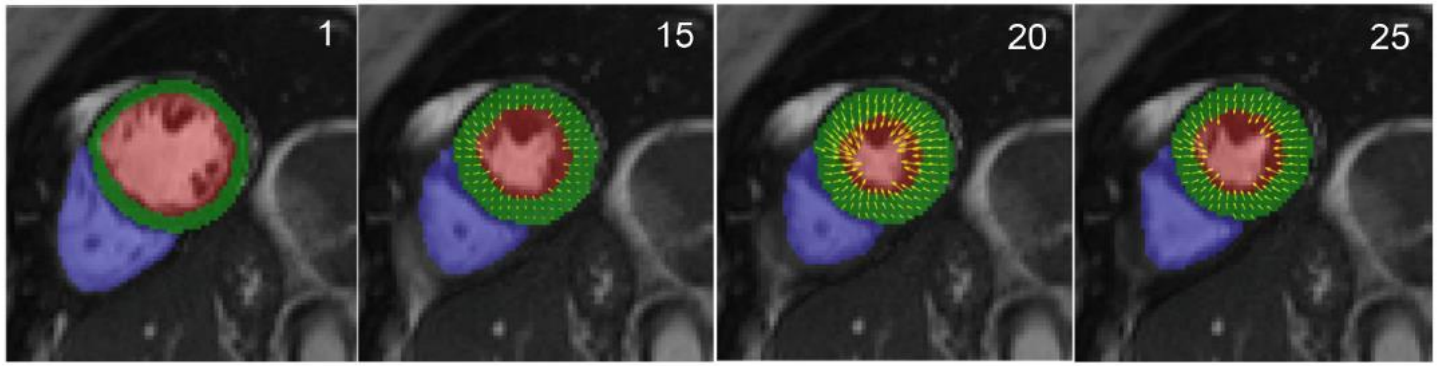


Fig. 2. Visualization results for simultaneous prediction of motion estimation and segmentation. Myocardial motions are from ED to other time points. Please refer to supplementary material for a dynamic video of a cardiac cycle.

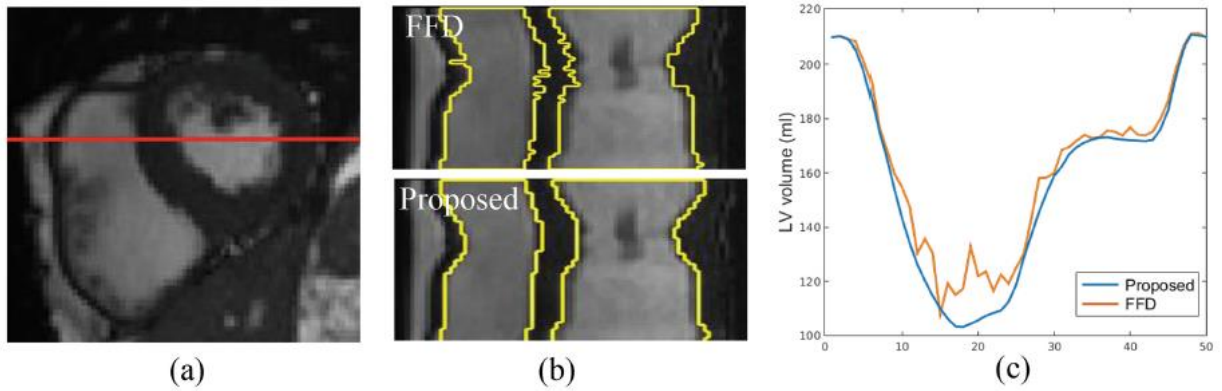


Fig. 3. (a) (b) Labeling results obtained by warping the ED frame segmentation to other time points using FFD and the proposed joint model. Results are shown in temporal views of the red short-axis line. (c) Left ventricular volume (ml) of the subject by warping the ED frame segmentation to other time points in a cardiac cycle. (Color figure online)