POLYTECHNIQUE MONTRÉAL

UNIVERSITÉ D'INGÉNIERIE

CRCHUM

Summer school on deep learning for medical imaging

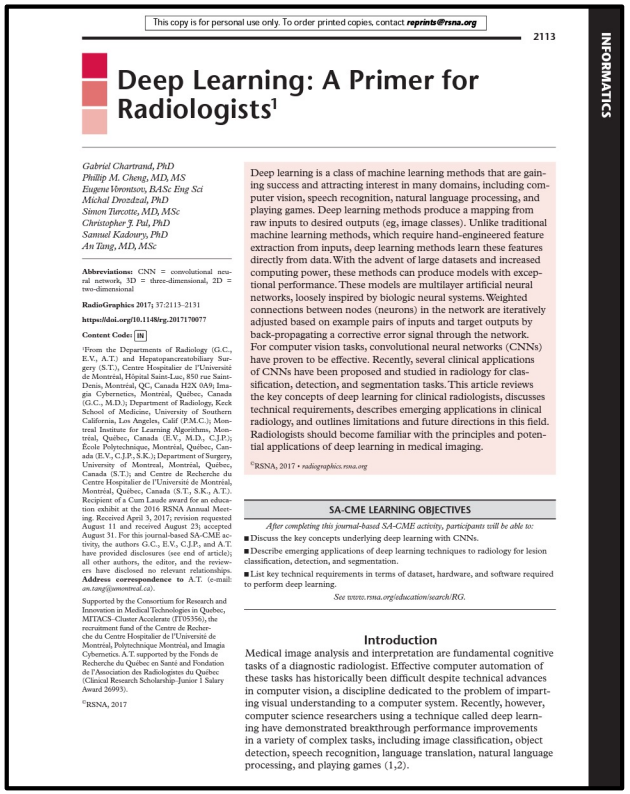
Samuel Kadoury Ph.D., ing.
Polytechnique Montréal / Centre de recherche du CHUM
Canada Research Chair in medical imaging and assisted interventions

July 6th, 2022

Topics

1) Data harmonization and normalization
2) Label noise
3) Image synthesis
4) Deformable registration
5) Motion modeling

Supporting materials



Deep Learning: A Primer for Radiologists[1]

Gabriel Chartrand, PhD
Phillip M. Cheng, MD, MS
Eugene Vorontsov, BASc Eng Sci
Michal Drozdzal, PhD
Simon Turcotte, MD, MSc
Christopher J. Pal, PhD
Samuel Kadoury, PhD
An Tang, MD, MSc

Abbreviations: CNN = convolutional neural network, 3D = three-dimensional, 2D = two-dimensional

RadioGraphics 2017; 37:2113–2131

https://doi.org/10.1148/rg.2017170077

Content Code: IN

[1]From the Departments of Radiology (G.C., E.V., A.T.) and Hepatopancreatobiliary Surgery (S.T.), Centre Hospitalier de l'Université de Montréal, Hôpital Saint-Luc, 850 rue Saint-Denis, Montréal, QC, Canada H2X 0A9; Imagia Cybernetics, Montréal, Québec, Canada (G.C., M.D.); Department of Radiology, Keck School of Medicine, University of Southern California, Los Angeles, Calif (P.M.C.); Montreal Institute for Learning Algorithms, Montréal, Québec, Canada (E.V., M.D., C.J.P.); École Polytechnique, Montréal, Québec, Canada (E.V., C.J.P., S.K.); Department of Surgery, University of Montreal, Montréal, Québec, Canada (S.T.); and Centre de Recherche du Centre Hospitalier de l'Université de Montréal, Montréal, Québec, Canada (S.T., S.K., A.T.). Recipient of a Cum Laude award for an education exhibit at the 2016 RSNA Annual Meeting. Received April 3, 2017; revision requested August 11 and received August 23; accepted August 31. For this journal-based SA-CME activity, the authors G.C., E.V., C.J.P., and A.T. have provided disclosures (see end of article); all other authors, the editor, and the reviewers have disclosed no relevant relationships. **Address correspondence to** A.T. (e-mail: *an.tang@umontreal.ca*).

Supported by the Consortium for Research and Innovation in Medical Technologies in Quebec, MITACS–Cluster Accelerate (IT05356), the recruitment fund of the Centre de Recherche du Centre Hospitalier de l'Université de Montréal, Polytechnique Montréal, and Imagia Cybernetics. A.T. supported by the Fonds de Recherche du Québec en Santé and Fondation de l'Association des Radiologistes du Québec (Clinical Research Scholarship–Junior 1 Salary Award 26993).

Deep learning is a class of machine learning methods that are gaining success and attracting interest in many domains, including computer vision, speech recognition, natural language processing, and playing games. Deep learning methods produce a mapping from raw inputs to desired outputs (eg, image classes). Unlike traditional machine learning methods, which require hand-engineered feature extraction from inputs, deep learning methods learn these features directly from data. With the advent of large datasets and increased computing power, these methods can produce models with exceptional performance. These models are multilayer artificial neural networks, loosely inspired by biologic neural systems. Weighted connections between nodes (neurons) in the network are iteratively adjusted based on example pairs of inputs and target outputs by back-propagating a corrective error signal through the network. For computer vision tasks, convolutional neural networks (CNNs) have proven to be effective. Recently, several clinical applications of CNNs have been proposed and studied in radiology for classification, detection, and segmentation tasks. This article reviews the key concepts of deep learning for clinical radiologists, discusses technical requirements, describes emerging applications in clinical radiology, and outlines limitations and future directions in this field. Radiologists should become familiar with the principles and potential applications of deep learning in medical imaging.

SA-CME LEARNING OBJECTIVES

After completing this journal-based SA-CME activity, participants will be able to:

- Discuss the key concepts underlying deep learning with CNNs.
- Describe emerging applications of deep learning techniques to radiology for lesion classification, detection, and segmentation.
- List key technical requirements in terms of dataset, hardware, and software required to perform deep learning.

See www.rsna.org/education/search/RG.

Introduction

Medical image analysis and interpretation are fundamental cognitive tasks of a diagnostic radiologist. Effective computer automation of these tasks has historically been difficult despite technical advances in computer vision, a discipline dedicated to the problem of imparting visual understanding to a computer system. Recently, however, computer science researchers using a technique called deep learning have demonstrated breakthrough performance improvements in a variety of complex tasks, including image classification, object detection, speech recognition, language translation, natural language processing, and playing games (1,2).

Insights into Imaging

EDUCATIONAL REVIEW Open Access

Deep learning workflow in radiology: a primer

Emmanuel Montagnon[1], Milena Cerny[1], Alexandre Cadrin-Chênevert[2], Vincent Hamilton[1], Thomas Derennes[1], André Ilinca[1], Franck Vandenbroucke-Menu[4], Simon Turcotte[1,4], Samuel Kadoury[5] and An Tang[1,3*]

Abstract

Interest for deep learning in radiology has increased tremendously in the past decade due to the high achievable performance for various computer vision tasks such as detection, segmentation, classification, monitoring, and prediction. This article provides step-by-step practical guidance for conducting a project that involves deep learning in radiology, from defining specifications, to deployment and scaling. Specifically, the objectives of this article are to provide an overview of clinical use cases of deep learning, describe the composition of multi-disciplinary team, and summarize current approaches to patient, data, model, and hardware selection. Key ideas will be illustrated by examples from a prototypical project on imaging of colorectal liver metastasis. This article illustrates the workflow for liver lesion detection, segmentation, classification, monitoring, and prediction of tumor recurrence and patient survival. Challenges are discussed, including ethical considerations, cohorting, data collection, anonymization, and availability of expert annotations. The practical guidance may be adapted to any project that requires automated medical image analysis.

Keywords: Review article, Deep learning, Medical imaging, Cohorting, Convolutional neural network

Key points

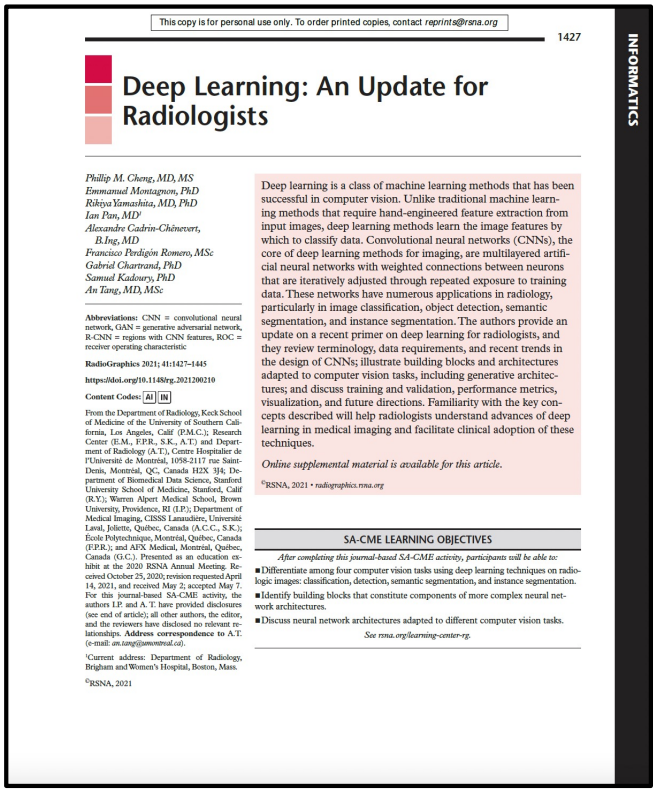
- Deep learning provides state-of-the-art performance for detection, segmentation, classification, and prediction.
- A multi-disciplinary team with clinical, imaging, and technical expertise is recommended.
- Data collection and curation constitute the most time-consuming steps.
- Several open-source deep learning frameworks with permissive licenses are available.
- Cloud computing leverages third-party hardware, storage, and technical resources.

Introduction

Deep learning is a subtype of representation learning which aims to describe complex data representations using simpler hierarchized structures defined from a set of specific features. With the advent of powerful parallel computing hardware based on graphical processing units (GPUs) and the availability of large datasets, deep learning has become a state-of-the-art technique in computer vision [1]. In the context of healthcare, deep learning shows great promise for analyzing structured (e.g., databases, tables) and unstructured data (e.g., images, text) [2]. Over the past decade, medical image analysis has greatly benefited from the application of deep learning (DL) techniques to various imaging modalities and organs [3].

Several tasks traditionally performed by radiologists such as lesion detection, segmentation, classification, and monitoring may be automated using deep learning techniques [4]. In abdominal radiology, deep learning has been applied to diverse tasks [3], organs [5, 6], and pathologies [7–9]. Despite the emerging application of deep learning techniques [1, 10], few articles have described the workflow to execute projects in abdominal radiology which require a broad range of steps, ranging from selection of patient population, choice of index test and reference standard, model selection, and assessment of performance.

* Correspondence:
[1]Centre de recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM), Montréal, Québec, Canada
[3]Department of Radiology, Radio-Oncology and Nuclear Medicine, Université Montréal and CRCHUM, 1058 rue Saint-Denis, Montréal, Québec H2X 3 J4, Canada
Full list of author information is available at the end of the article



Deep Learning: An Update for Radiologists

Phillip M. Cheng, MD, MS
Emmanuel Montagnon, PhD
Rikiya Yamashita, MD, PhD
Ian Pan, MD[1]
Alexandre Cadrin-Chênevert, B.Ing, MD
Francisco Perdigón Romero, MSc
Gabriel Chartrand, PhD
Samuel Kadoury, PhD
An Tang, MD, MSc

Abbreviations: CNN = convolutional neural network, GAN = generative adversarial network, R-CNN = regions with CNN features, ROC = receiver operating characteristic

RadioGraphics 2021; 41:1427–1445

https://doi.org/10.1148/rg.2021200210

Content Codes: AI IN

From the Department of Radiology, Keck School of Medicine of the University of Southern California, Los Angeles, Calif (P.M.C.); Research Center (E.M., F.P.R., S.K., A.T.) and Department of Radiology (A.T.), Centre Hospitalier de l'Université de Montréal, 1058-2117 rue Saint-Denis, Montréal, QC, Canada H2X 3J4; Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, Calif (R.Y.); Warren Alpert Medical School, Brown University, Providence, RI (I.P.); Department of Medical Imaging, CISSS Lanaudière, Université Laval, Joliette, Québec, Canada (A.C.C., S.K.); École Polytechnique, Montréal, Québec, Canada (F.P.R.); and AFX Medical, Montréal, Québec, Canada (G.C.). Presented as an education exhibit at the 2020 RSNA Annual Meeting. Received October 25, 2020; revision requested April 14, 2021, and received May 2; accepted May 7. For this journal-based SA-CME activity, the authors I.P. and A.T. have provided disclosures (see end of article); all other authors, the editor, and the reviewers have disclosed no relevant relationships. **Address correspondence to** A.T. (e-mail: *an.tang@umontreal.ca*).

[1]Current address: Department of Radiology, Brigham and Women's Hospital, Boston, Mass.

Deep learning is a class of machine learning methods that has been successful in computer vision. Unlike traditional machine learning methods that require hand-engineered feature extraction from input images, deep learning methods learn the image features by which to classify data. Convolutional neural networks (CNNs), the core of deep learning methods for imaging, are multilayered artificial neural networks with weighted connections between neurons that are iteratively adjusted through repeated exposure to training data. These networks have numerous applications in radiology, particularly in image classification, object detection, semantic segmentation, and instance segmentation. The authors provide an update on a recent primer on deep learning for radiologists, and they review terminology, data requirements, and recent trends in the design of CNNs; illustrate building blocks and architectures adapted to computer vision tasks, including generative architectures; and discuss training and validation, performance metrics, visualization, and future directions. Familiarity with the key concepts described will help radiologists understand advances of deep learning in medical imaging and facilitate clinical adoption of these techniques.

Online supplemental material is available for this article.

SA-CME LEARNING OBJECTIVES

After completing this journal-based SA-CME activity, participants will be able to:

- Differentiate among four computer vision tasks using deep learning techniques on radiologic images: classification, detection, semantic segmentation, and instance segmentation.
- Identify building blocks that constitute components of more complex neural network architectures.
- Discuss neural network architectures adapted to different computer vision tasks.

See rsna.org/learning-center-rg.

Deep learning: A primer for radiologists

INFORMATICS

Deep Learning: A Primer for Radiologists[1]

Gabriel Chartrand, PhD
Phillip M. Cheng, MD, MS
Eugene Vorontsov, BASc Eng Sci
Michal Drozdzal, PhD
Simon Turcotte, MD, MSc
Christopher J. Pal, PhD
Samuel Kadoury, PhD
An Tang, MD, MSc

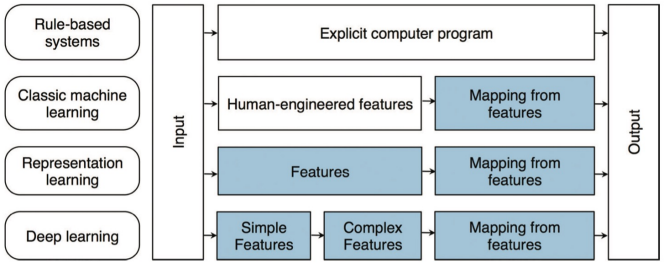
Deep learning is a class of machine learning methods that are gaining success and attracting interest in many domains, including computer vision, speech recognition, natural language processing, and playing games. Deep learning methods produce a mapping from raw inputs to desired outputs (eg, image classes). Unlike traditional machine learning methods, which require hand-engineered feature extraction from inputs, deep learning methods learn these features

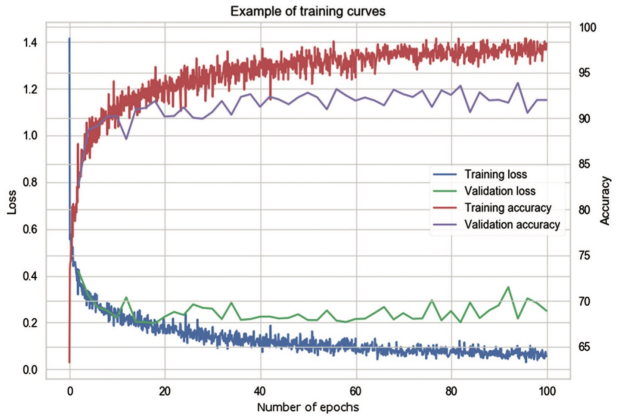
SA-CME LEARNING OBJECTIVES

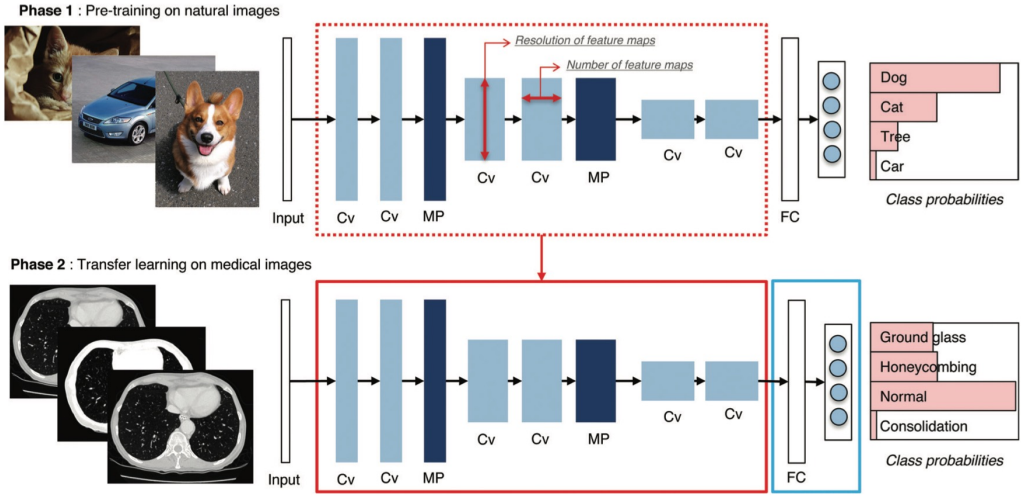
After completing this journal-based SA-CME activity, participants will be able to:

- Discuss the key concepts underlying deep learning with CNNs.
- Describe emerging applications of deep learning techniques to radiology for lesion classification, detection, and segmentation.
- List key technical requirements in terms of dataset, hardware, and software required to perform deep learning.

See www.rsna.org/education/search/RG.



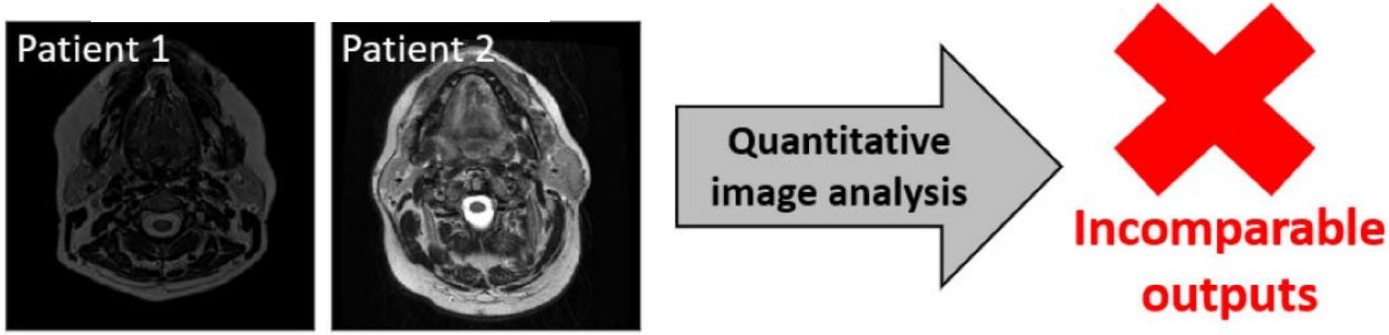




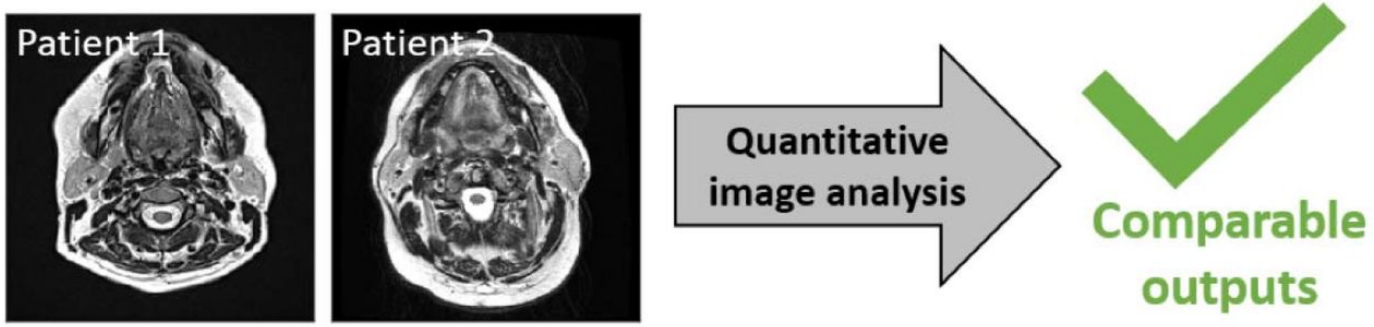
1) Data harmonization and normalization

Effect of normalization

No normalization



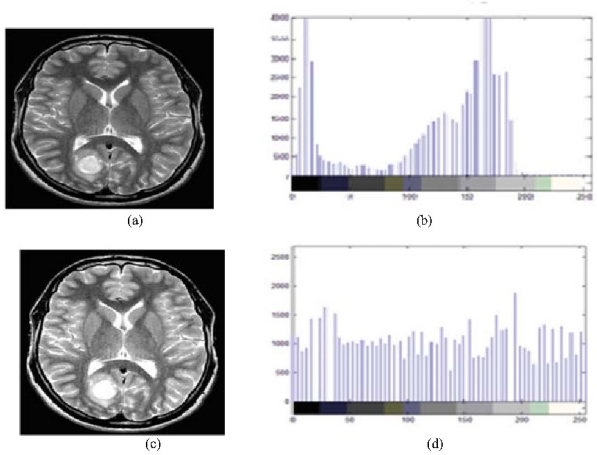
Normalization



From Wahid et al. 2021

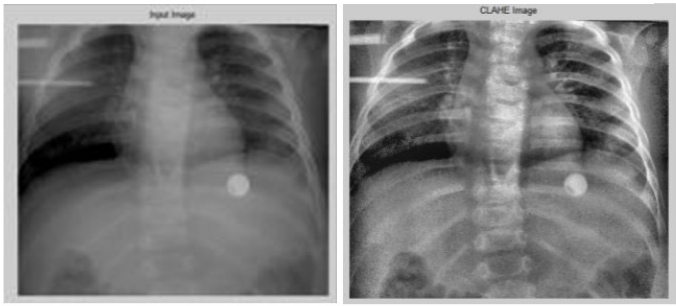
Image pre-processing techniques

- Histogram equalization
- Hist. eq. + Gaussian blur
- Hist. eq. + bilateral filter



- Adaptive masking
  - (max - min) followed by binary thresholding

- Adaptive masking + hist. eq. + Gaussian blur



(a) Input Image (b) CLAHE Image

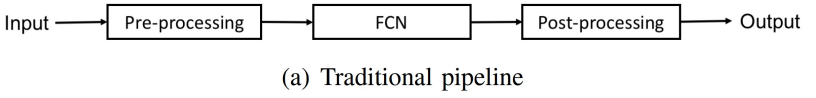
Normalization techniques for deep learning

- **Min-max normalization**: in this method, pixel gray level values are brought into a specific range (usually [0-1]). Alternatively, a [10^{th} percentile, 90^{th} percentile] range is chosen in some cases to avoid imaging outliers.

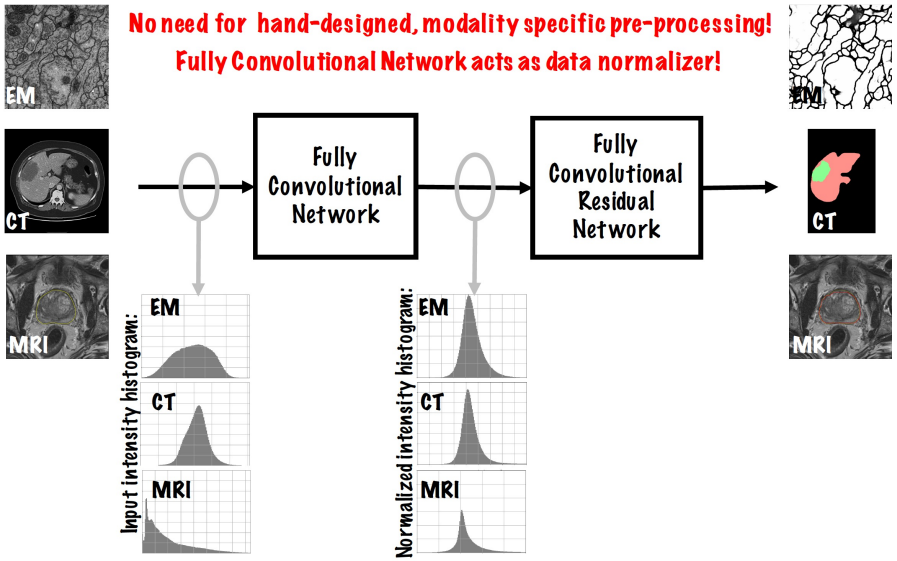
- **Volume standardization**: this technique consists of computing the mean and standard deviation of the entire 3D image. Each voxel is then subtracted by the volume mean and divided by the voxel standard deviation.

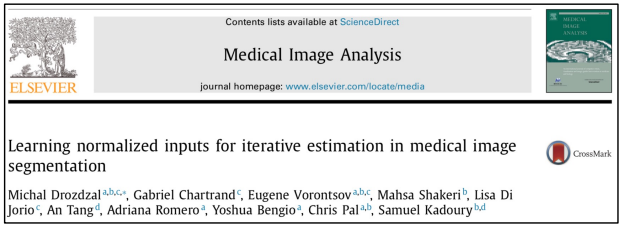
- **Slice-by-slice standardization**: this strategy is similar to the previous one. However, instead of computing a global mean and standard deviation for the entire volume, slice-specific mean and standard deviation are computed, and pixels are standardized using the slice statistics instead of the volume statistics.

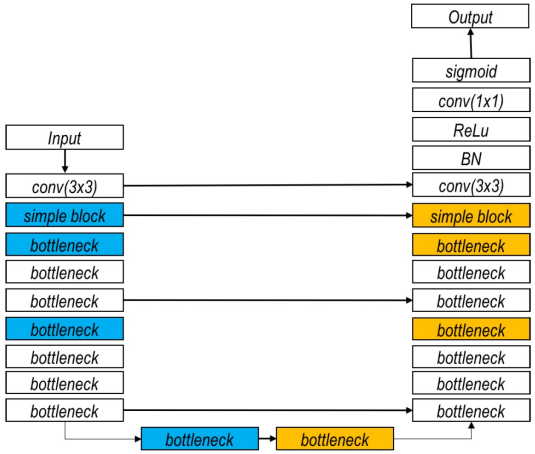
Learning normalized input for image segmentation



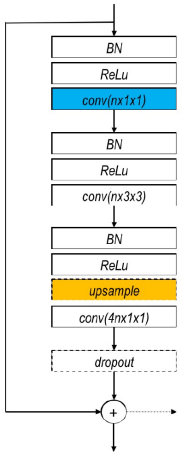
(a) Traditional pipeline



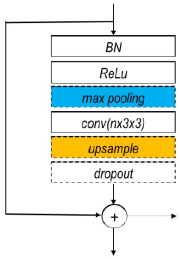




(a) FC-ResNet



(b) Bottleneck block

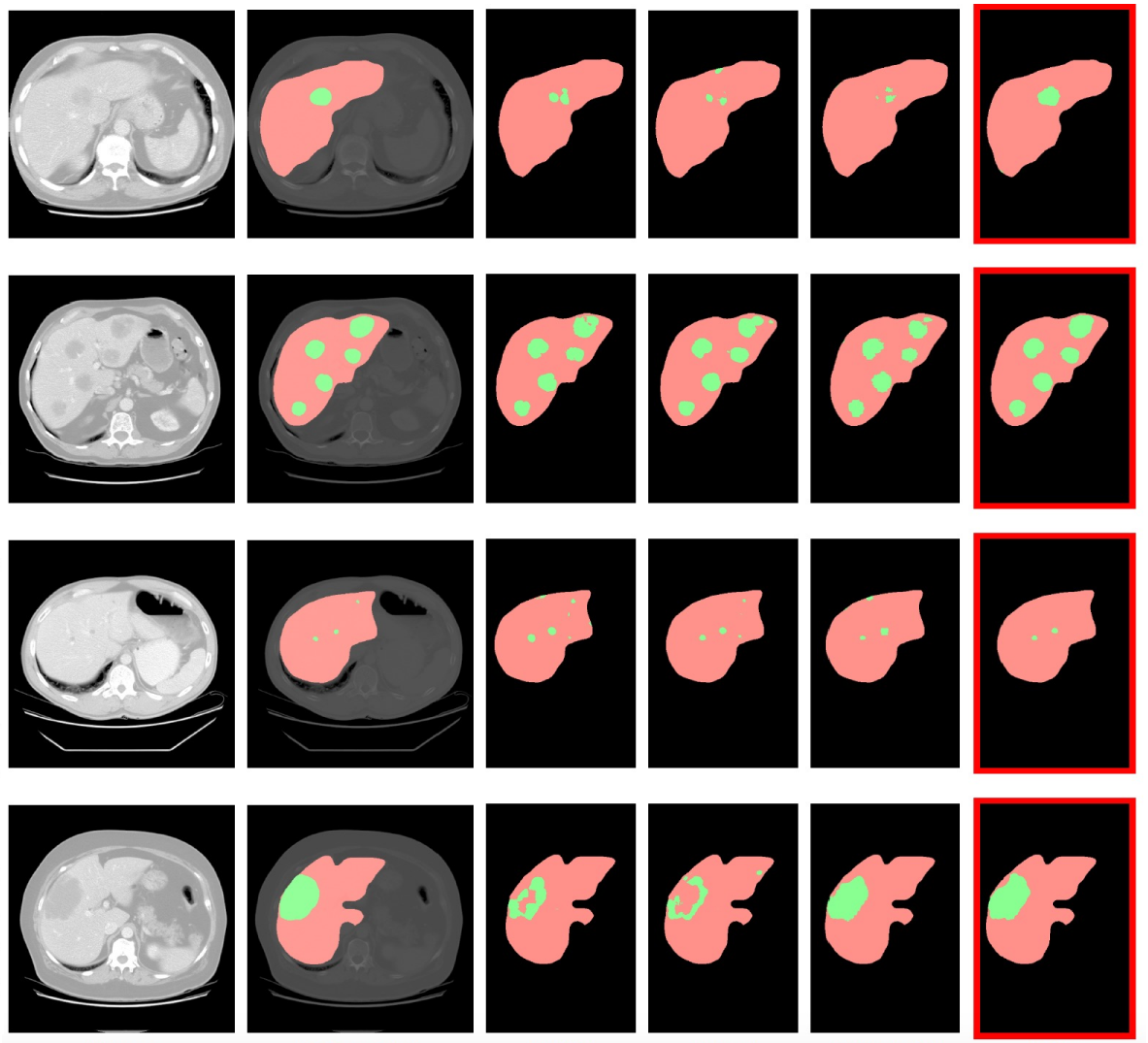


(c) Simple block

Data:
> CT data (512 x 512 x 200)
> 105 images
- 77 training
- 28 validation
> 30 test

Method	Validation		
	loss	$Dice_{lesion}$	$Dice_{liver}$
FCN8 [36]	-0.419	0.589	0.994
Unet [43]	-0.451	0.553	0.994
FC-ResNet [20]	-0.223	0.551	0.993
Ours	**-0.795**	**0.771**	**0.997**

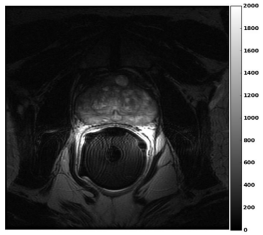
Method	Test		
	loss	$Dice_{lesion}$	$Dice_{liver}$
FCN8 [36]	-0.437	0.535	0.989
Unet [43]	-0.396	0.570	0.990
FC-ResNet [20]	-0.224	0.617	0.990
Ours	**-0.796**	**0.711**	**0.993**



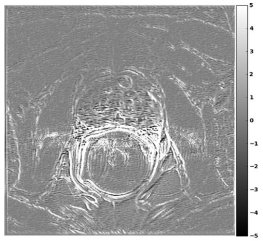
(a) Image (b) Ground Truth (c) FCN8 (d) Unet (e) FC-ResNet (f) Ours

Prostate segmentation

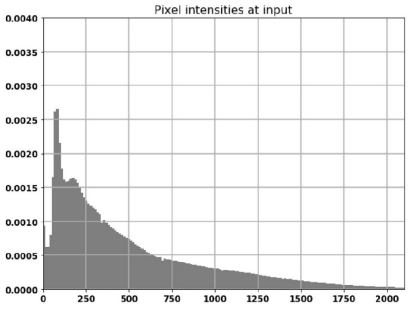
PROMISE12 challenge data:
> T2-w MRI (256 x 256 x 30)
> 50 images
- 40 training
- 10 validation
> 30 test images

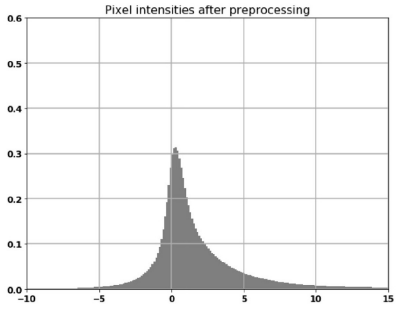


a) Input image

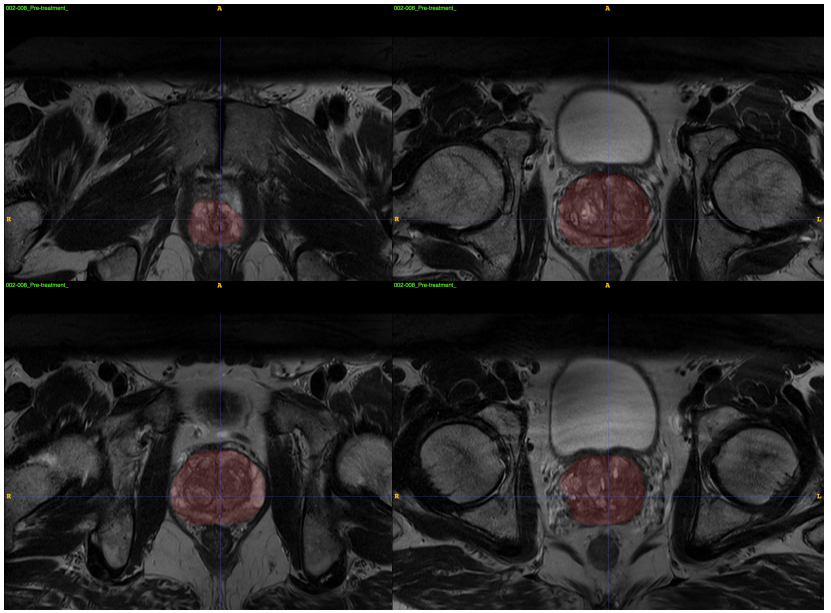


b) Pre-procesor output





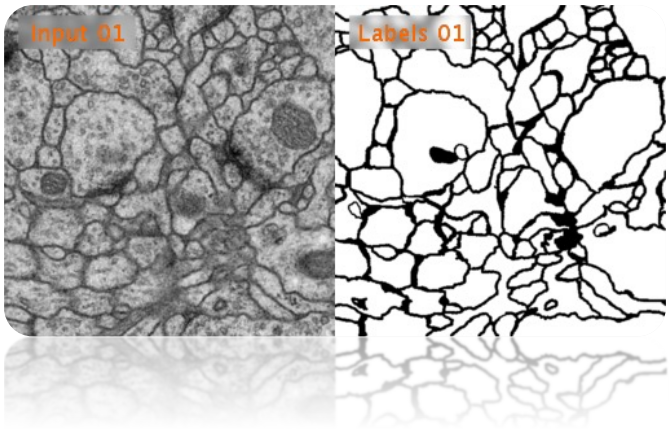
Test set (n=7) Princess Margaret Hospital



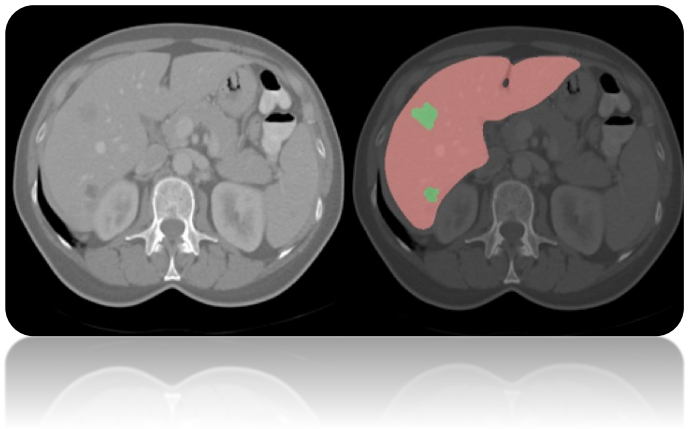
Method	Score [-]	Dice [%]	Avg. Dist. [mm]	Vol. Diff. [%]
2D FCNs				
Ours	**83.02**	87.4	2.17	12.37
SITUS	79.92	84.13	2.96	23.00
3D FCNs				
CUMED [50]	86.65	89.43	1.95	6.95
CAMP-TUM2 [35]	82.39	86.91	2.23	14.98
SRIBHME	74.17	74.46	2.83	34.89

Semantic image segmentation

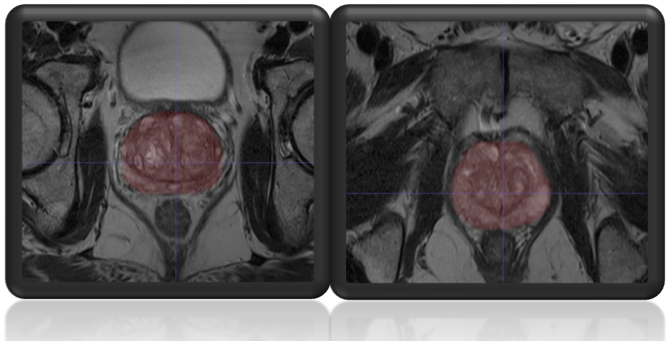
Electron microscopy



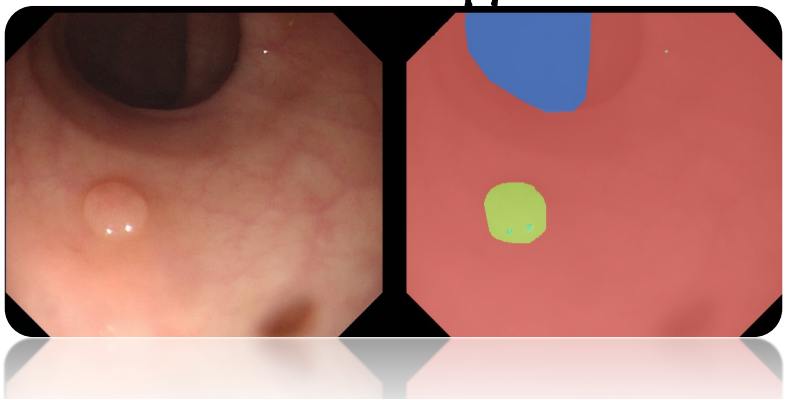
CT



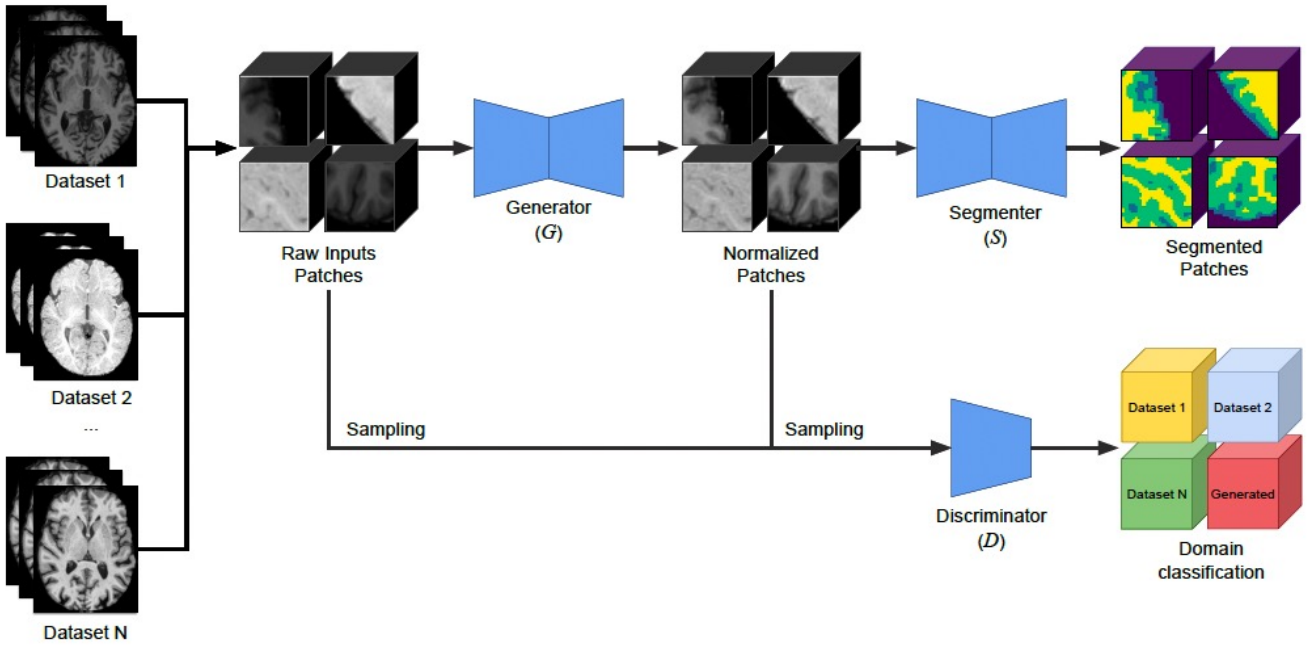
MRI

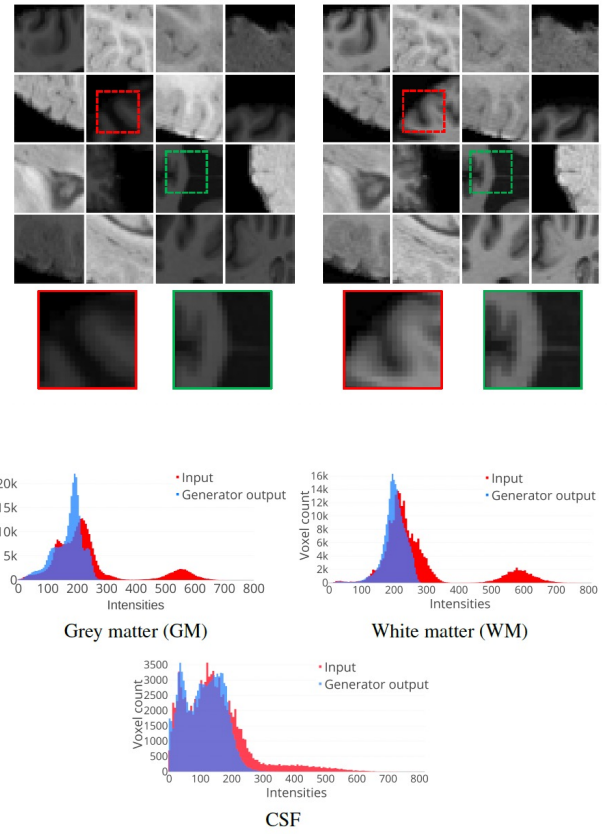


Endoscopy



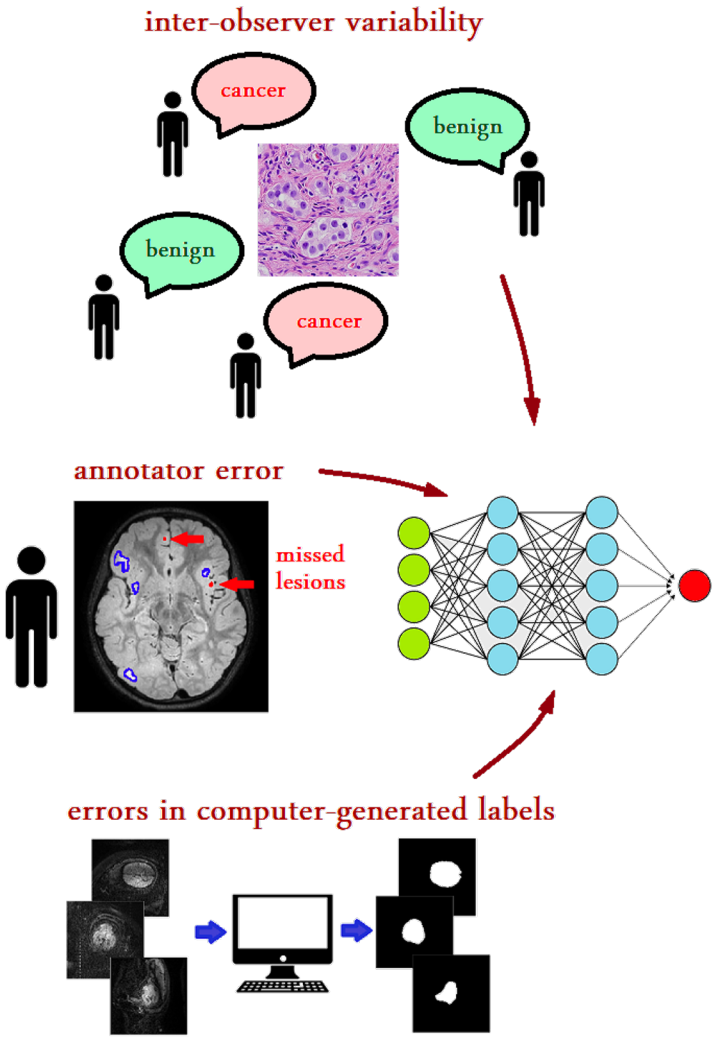
Realistic Image Normalization for Multi-Domain Segmentation





From Deslile et al. Medical Image analysis, 2020

2) Label noise



Label cleaning and pre-processing
- Identifying and removing samples with incorrect labels
- Iterative label cleaning in parallel with training
- Label smoothing
- Label cleaning with generative models

Network architectures
- Estimating and utilizing noise statistics using
  - Noise layers
  - Probabilistic graphical models

Label-noise-robust loss functions
- Mean-absolute-error loss
- Cross entropy with abstention
- "Correcting" standard loss functions using label noise statistics

Data re-weighting
- Meta-learning
- Gradient scaling
- Learned re-weighting
- Learning to weight labels from different annotators

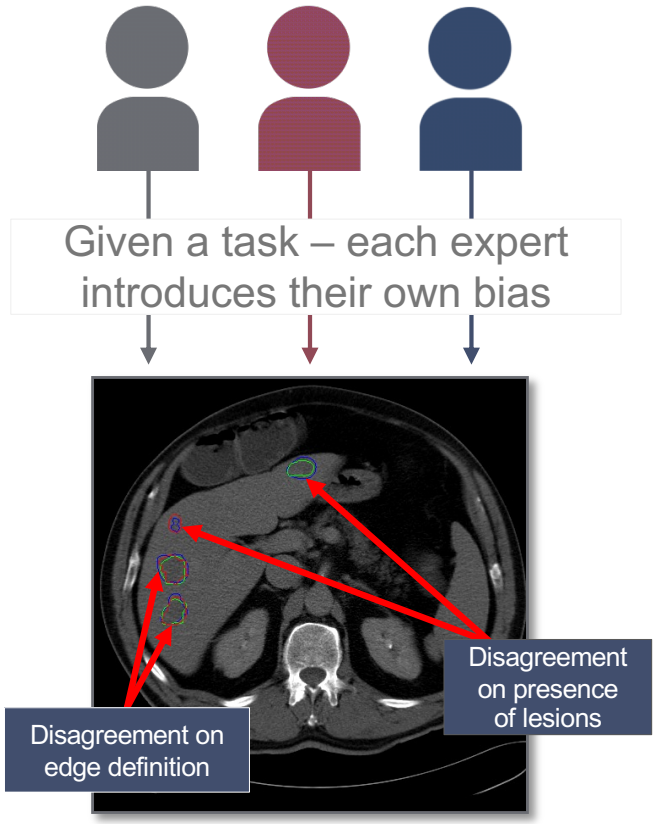
Data and label consistency
- Exploiting data similarity to identify incorrect labels
- Auxiliary image regularization
- Manifold regularization

Training procedures
- Curriculum learning
- Knowledge distillation
- Co-teaching
- Data augmentation (mixup)

From Karimi et al. 2020

Sources of error

Inter-Rater* Variability

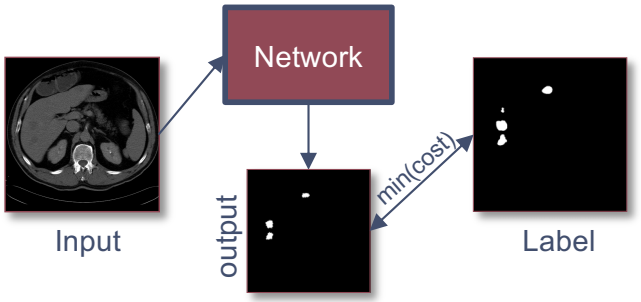


Lack of Ground Truth

- Lack of agreement between experts results in a lack of availability of a single "ground truth" in the medical field
- Highly biased towards the single rater
- **Impact:**
  - limits the performance of supervised deep learning models
  - Poor generalizability due to biased input data
- No consensus on who is right
- Variability cannot be represented in traditional segmentation approaches

Segmentation

- Deep learning segmentation models are mostly trained for a single output – a deterministic viewpoint based on a single mask/label



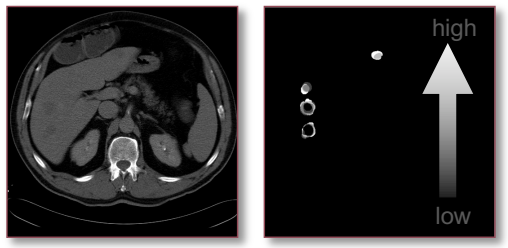
+ Deterministic → output is consistent post optimization
− Does not capture variability of opinions

*rater = annotator = expert = observer

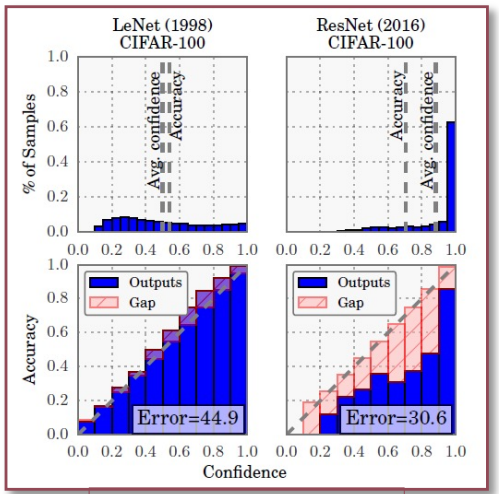
How to work with noisy labels? Probabilistic approach

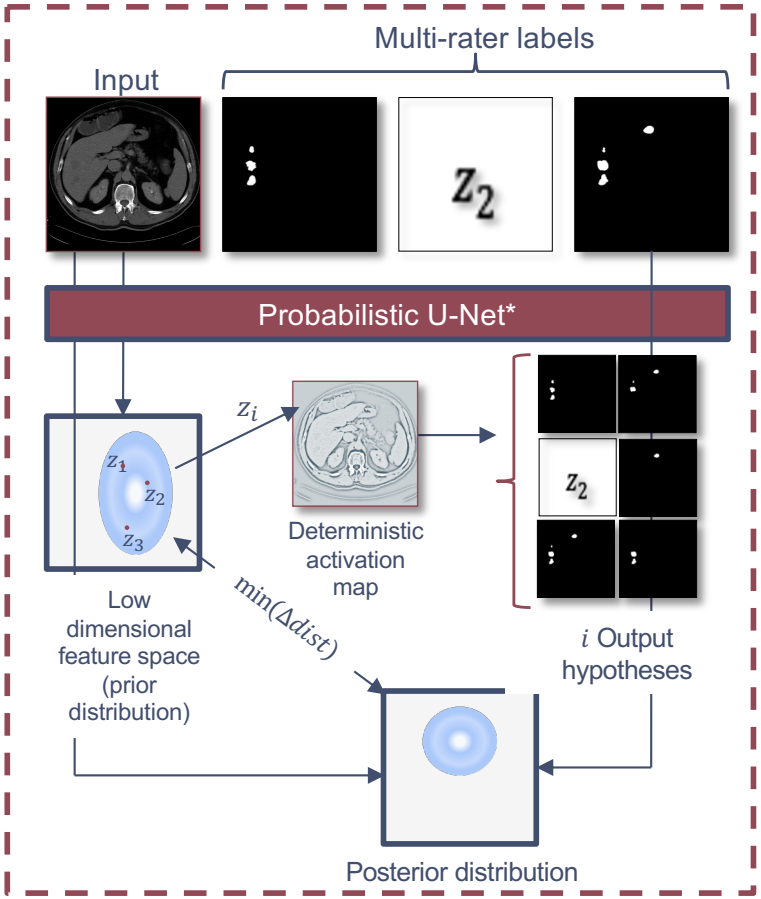
- Take a probabilistic approach to represent a distribution of hypothetical outputs (e.g. probabilistic U-Net)
  - A combination of a deterministic U-Net with a **c**onditional **V**ariational **Auto-Encoder** (**cVAE**: generative, probabilistic)
  - **cVAE** assumes data is generated from a random process → able to generate multiple hypotheses.
- Quantify the **uncertainty** in the output – i.e. how confident is the model with its prediction?
- Based on this confidence, verify whether the model is **calibrated** (Guo, et al)– i.e. does the accuracy match model confidence? Is the model overconfident in regions where it has low accuracy?
  - ➢ the latter is an issue to be addressed

to be able to trust predictions from DL models



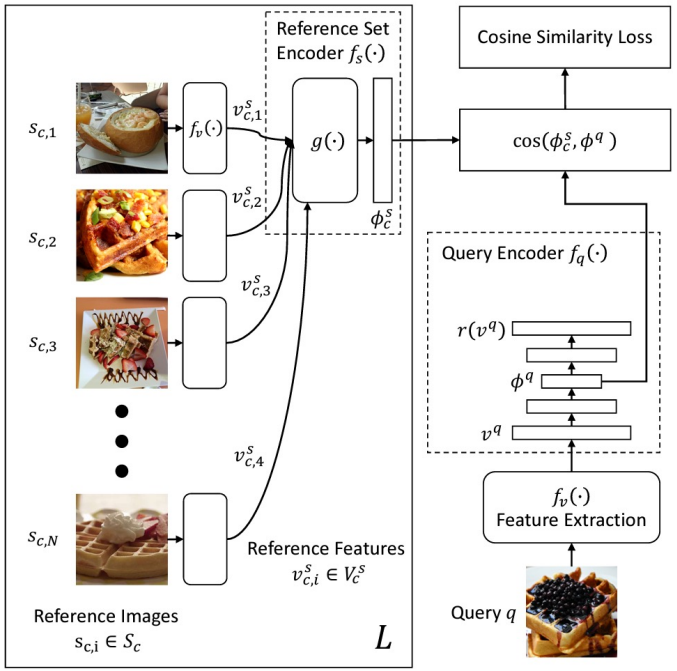
Uncertainty Map





*Kohl, et al. NeurIPS 2018

Label cleaning and pre-processing



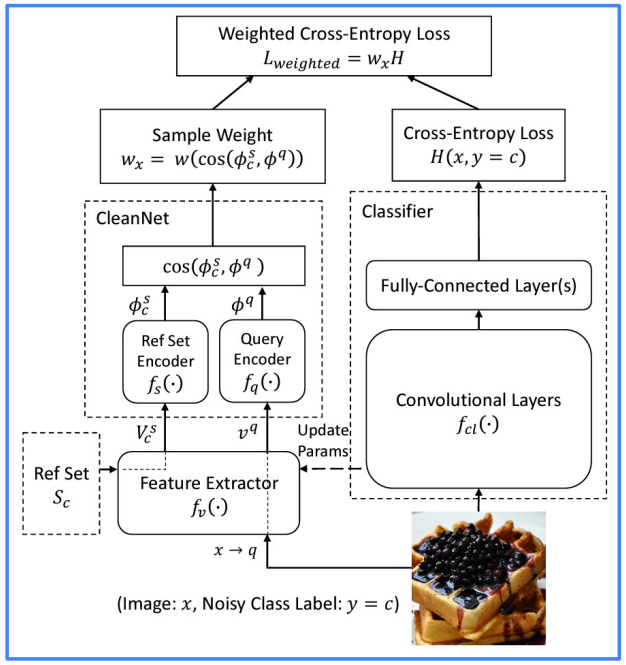
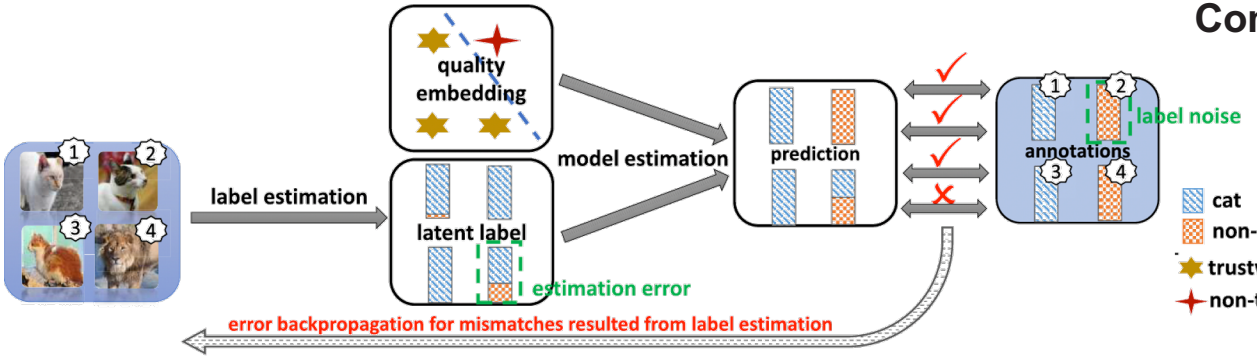


Illustration of integrating CleanNet for training the CNN based image classifier with label noise

F -> cosine similarity by model w/ verification labels
D -> cosine similarity under transfer learning

Lee et al. “Cleannet: Transfer learning for scalable image classifier training with label noise”, CVPR 2018.

Network architecture



Contrastive-additive noise network (CAN)

	V07TE				
P_{noise}	Resnet-N	LearnQ	ICNM	Bootstrap	CAN
1.0	6.4	9.1	**9.2**	8.9	8.6
0.8	33.4	28.0	28.5	30.1	**36.1**
0.6	53.0	56.4	57	59.3	**63.2**
0.4	70.2	72.0	71.6	73.3	**79.4**
0.2	78.2	80.1	79.6	81.0	**83.6**
0.0	**86.8**	85.4	85.4	85.5	85.3

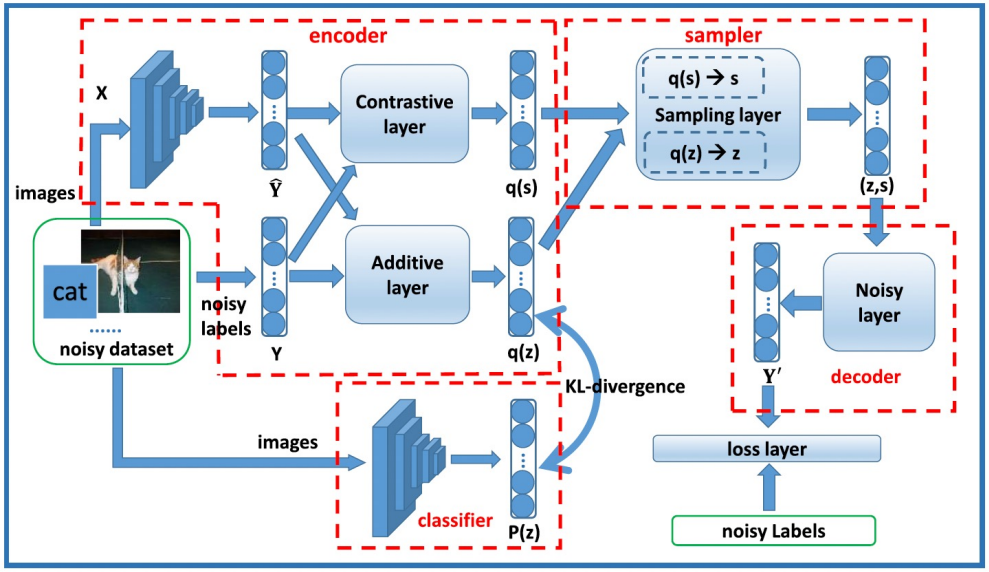
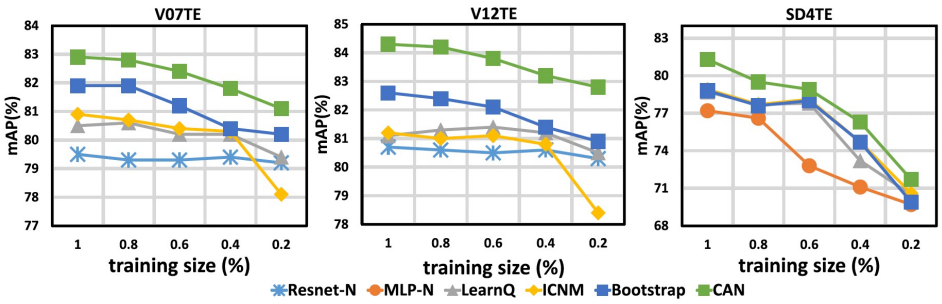
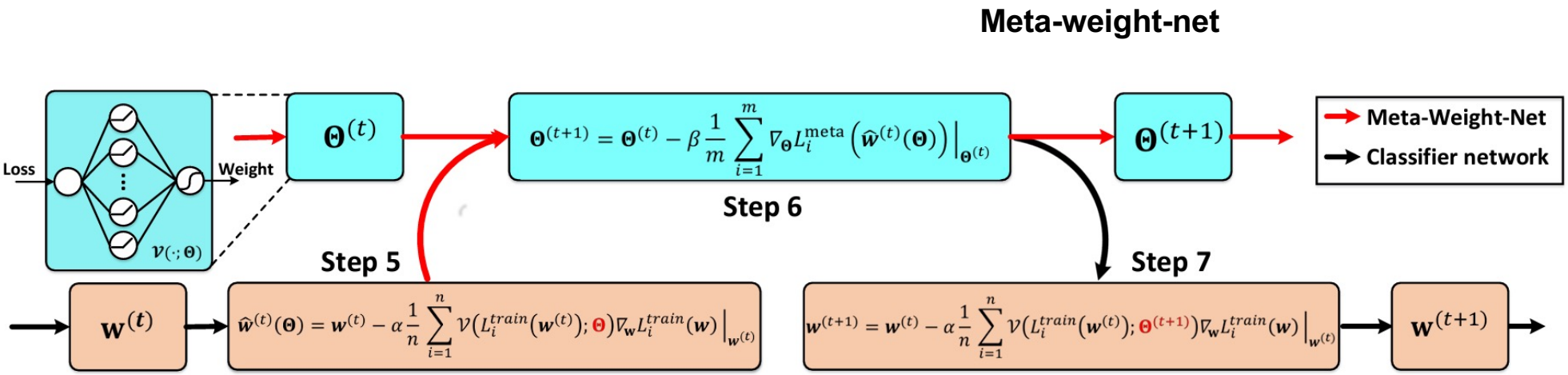


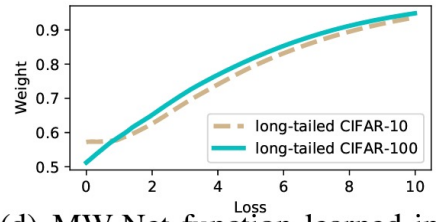
Illustration of integrating CleanNet for training the CNN based image classifier with label noise



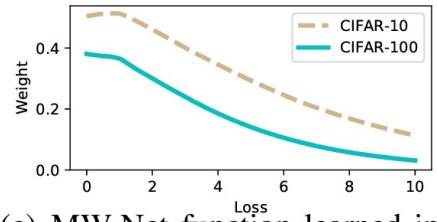
Yao, J. et al. 2018. Deep learning from noisy image labels with quality embedding. IEEE TIP.

Data re-weighting

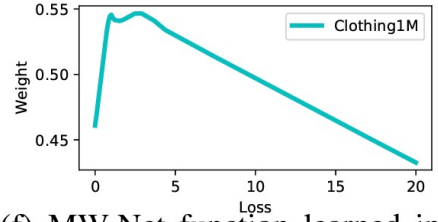




(d) MW-Net function learned in class imbalance case



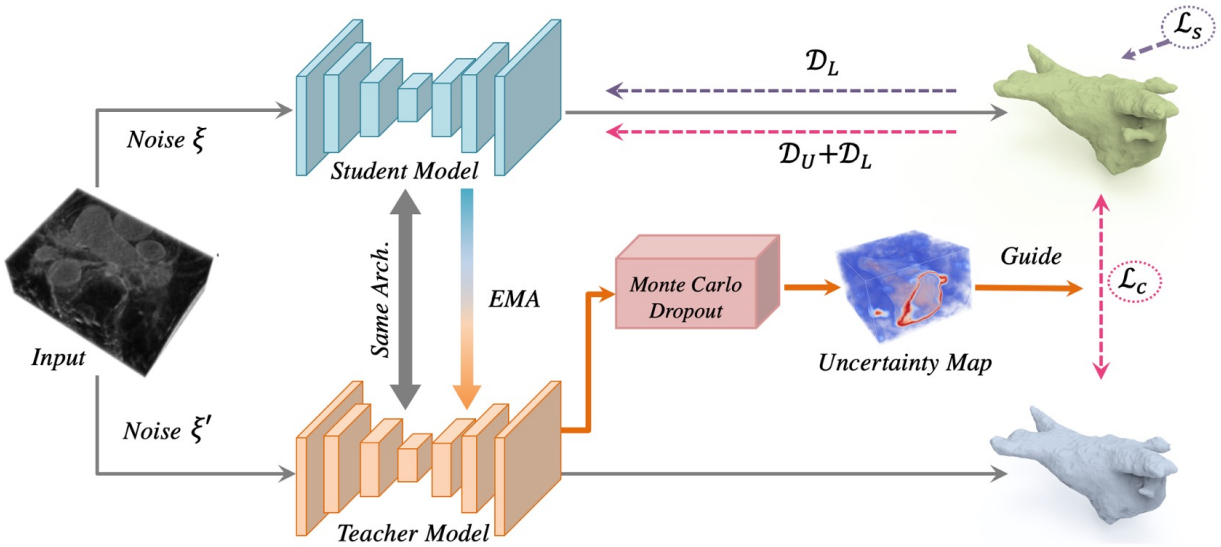
(e) MW-Net function learned in corrupter labels case



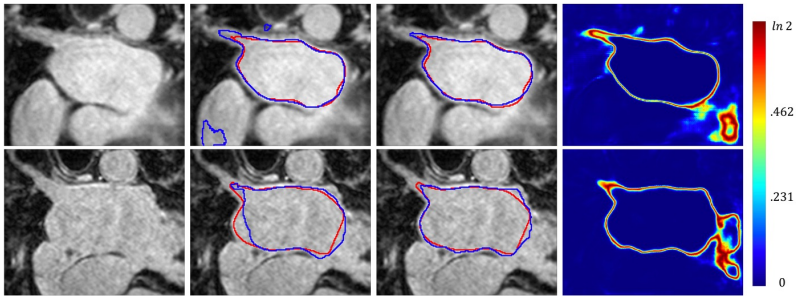
(f) MW-Net function learned in real Clothing1M dataset

Shu, J. , et al. , 2019. Meta-weight-net: learning an explicit mapping for sample weighting.

Data and label consistency



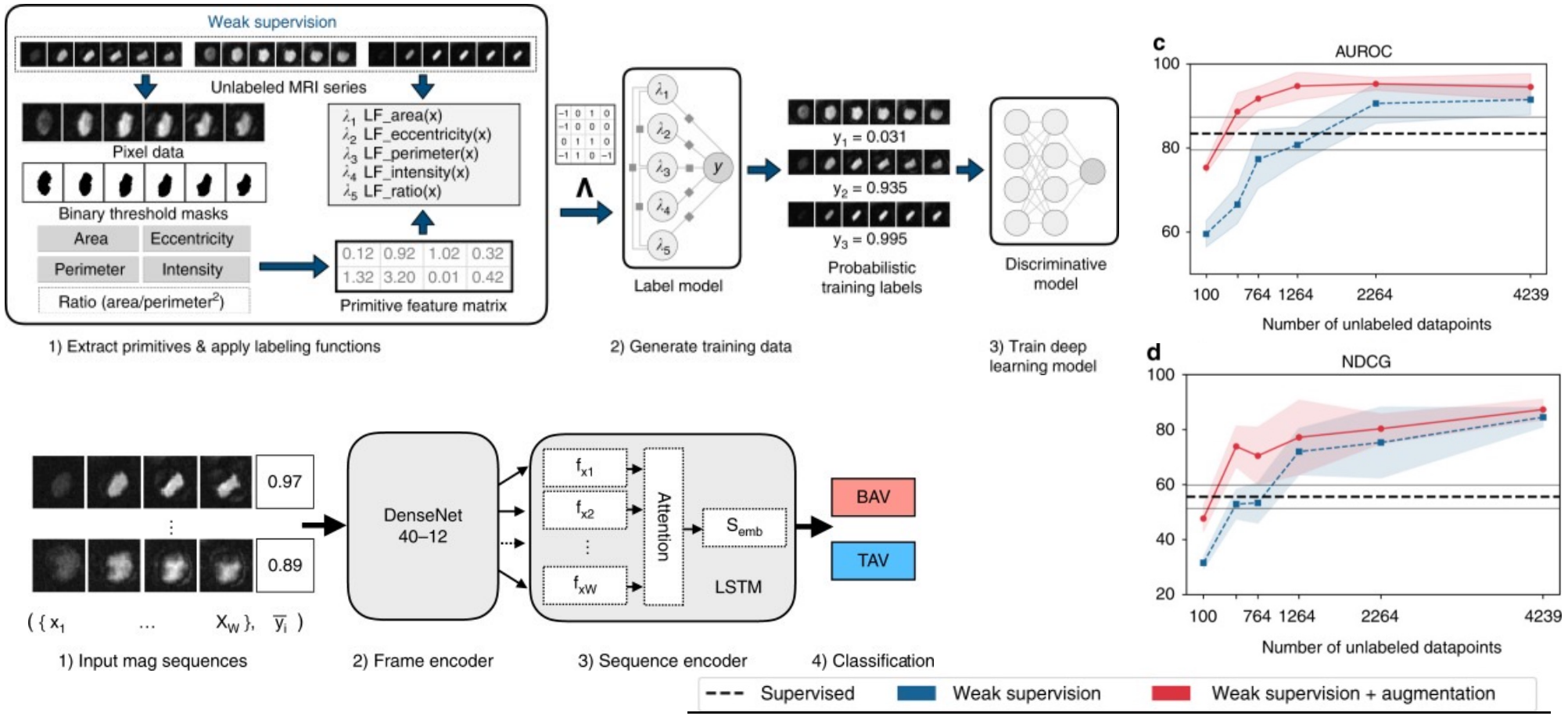
Contrastive-additive noise network (CAN)



Method	# scans used		Metrics			
	Labeled	Unlabeled	Dice[%]	Jaccard[%]	ASD[voxel]	95HD[voxel]
MT	16	64	88.23	79.29	2.73	10.64
MT-Dice [5]	16	64	88.32	79.37	2.76	10.50
Our UA-MT	16	64	88.88	80.21	2.26	7.32
Bayesian V-Net	8	0	79.99	68.12	5.48	21.11
Our UA-MT	8	72	84.25	73.48	3.36	13.84
Bayesian V-Net	24	0	88.52	79.70	2.60	10.45
Our UA-MT	24	56	90.16	82.18	2.73	8.90

Yu, L. et al. 2019. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. MICCAI.

Training procedure

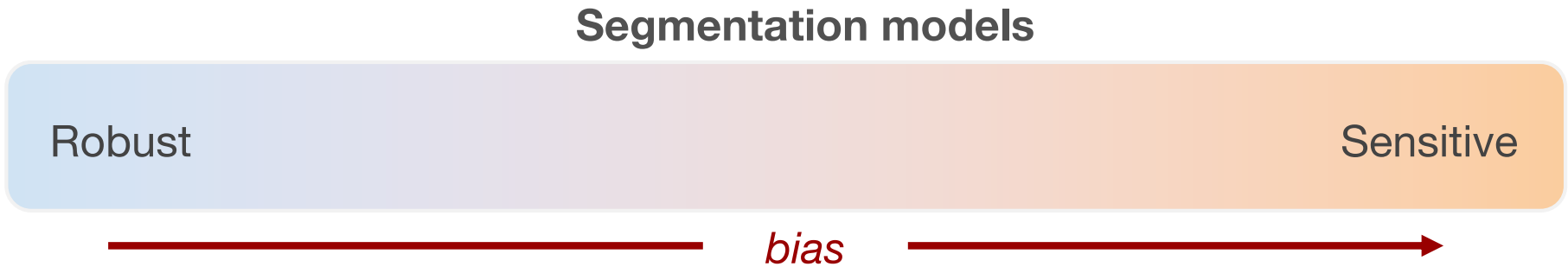


Fries, J.A. et al., 2019. Nature Communications.

Ground-truth annotation quality

Loosely define **biased errors** in segmentation labels as:

A consistent error such that the annotations are always wrong in the same way.



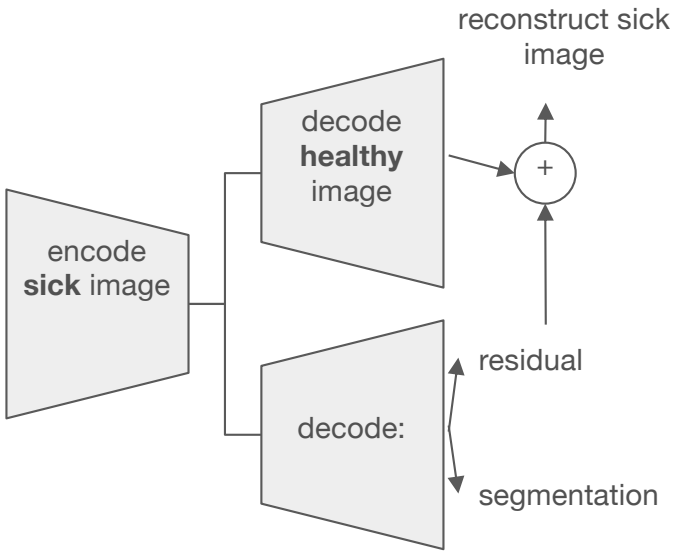
Experimental setup

Perturbations

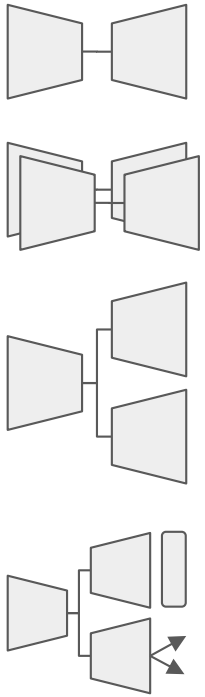
- Random crop
- Constant shift
- Random warp
- Permutation

Method description

Image-to-image translator



Summary



Seg

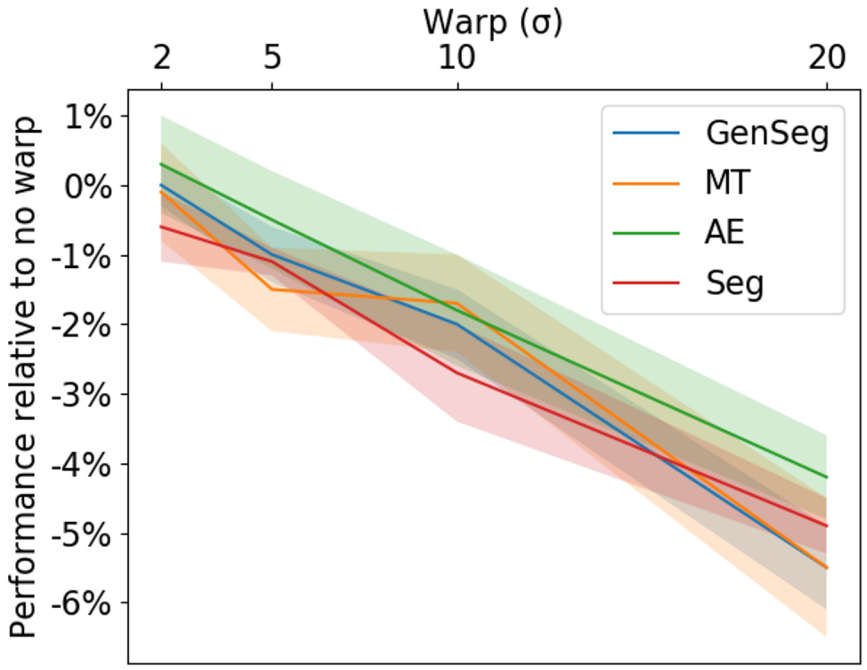
MT

AE

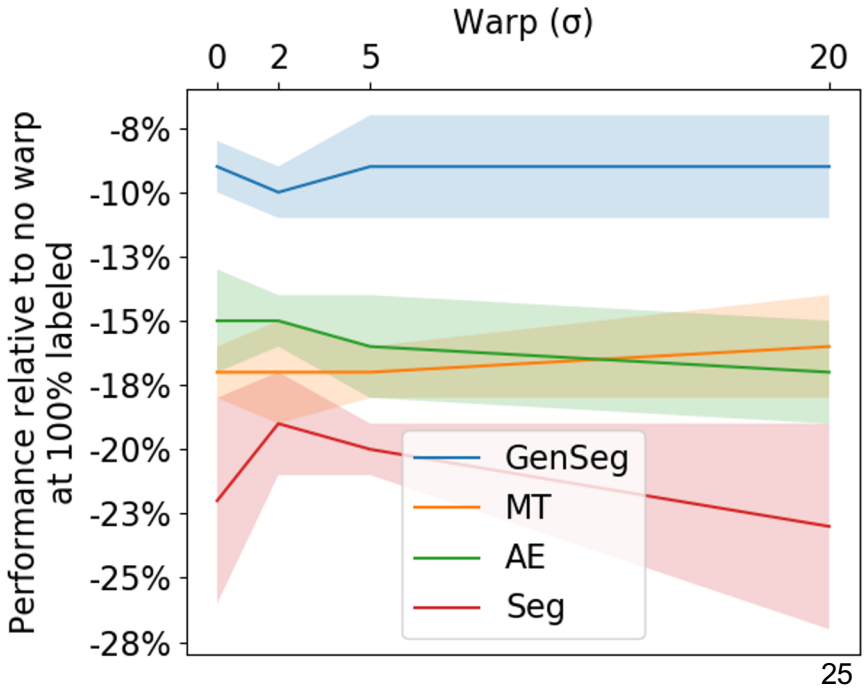
GenSeg

Random warp

All data annotated

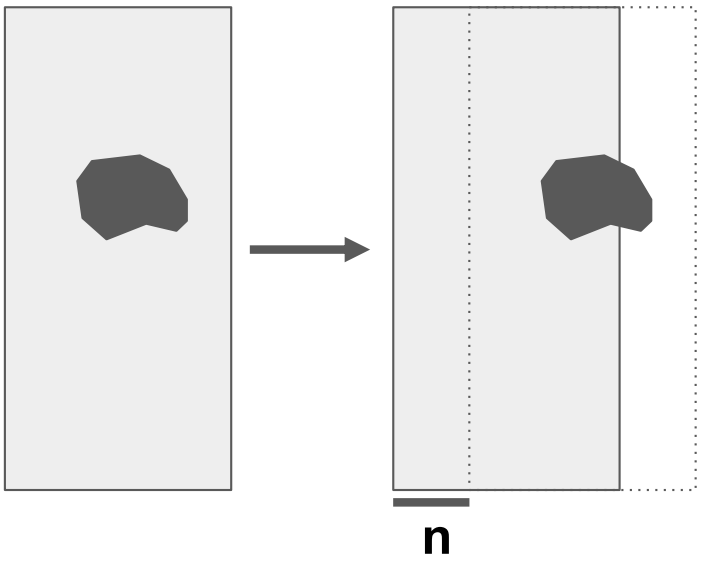


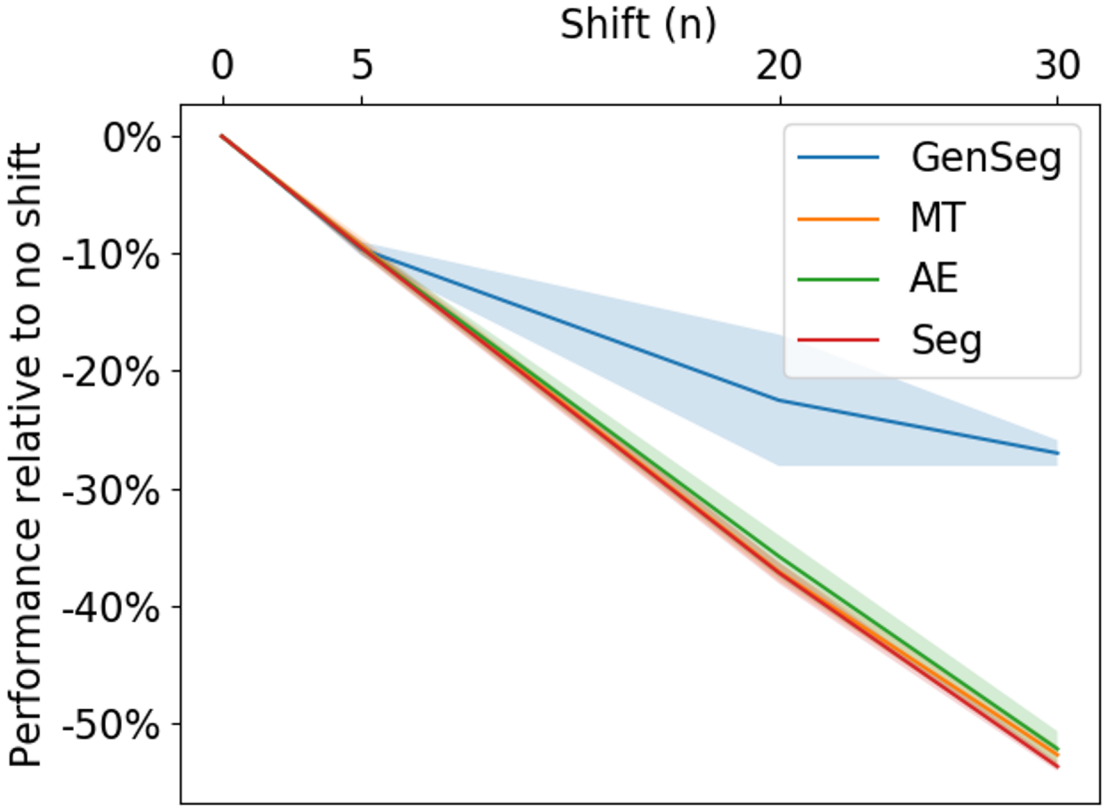
1% of data annotated



Constant shift

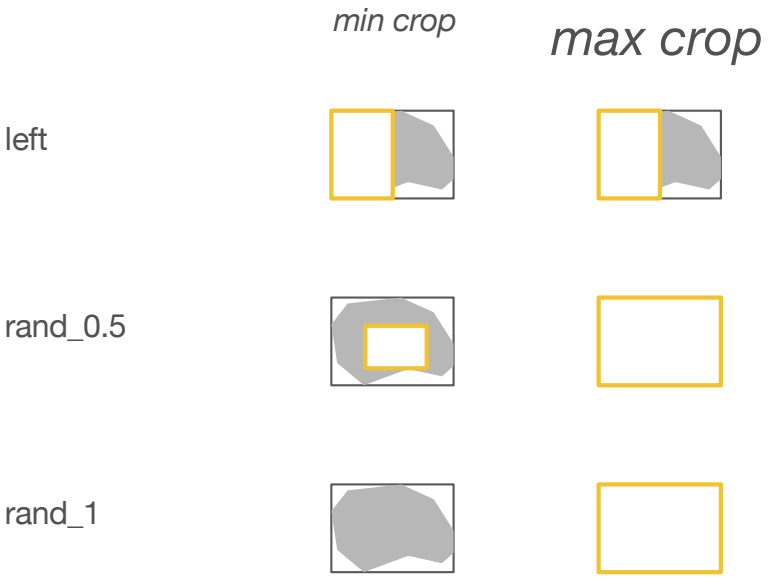
Shift mask by **n** pixels

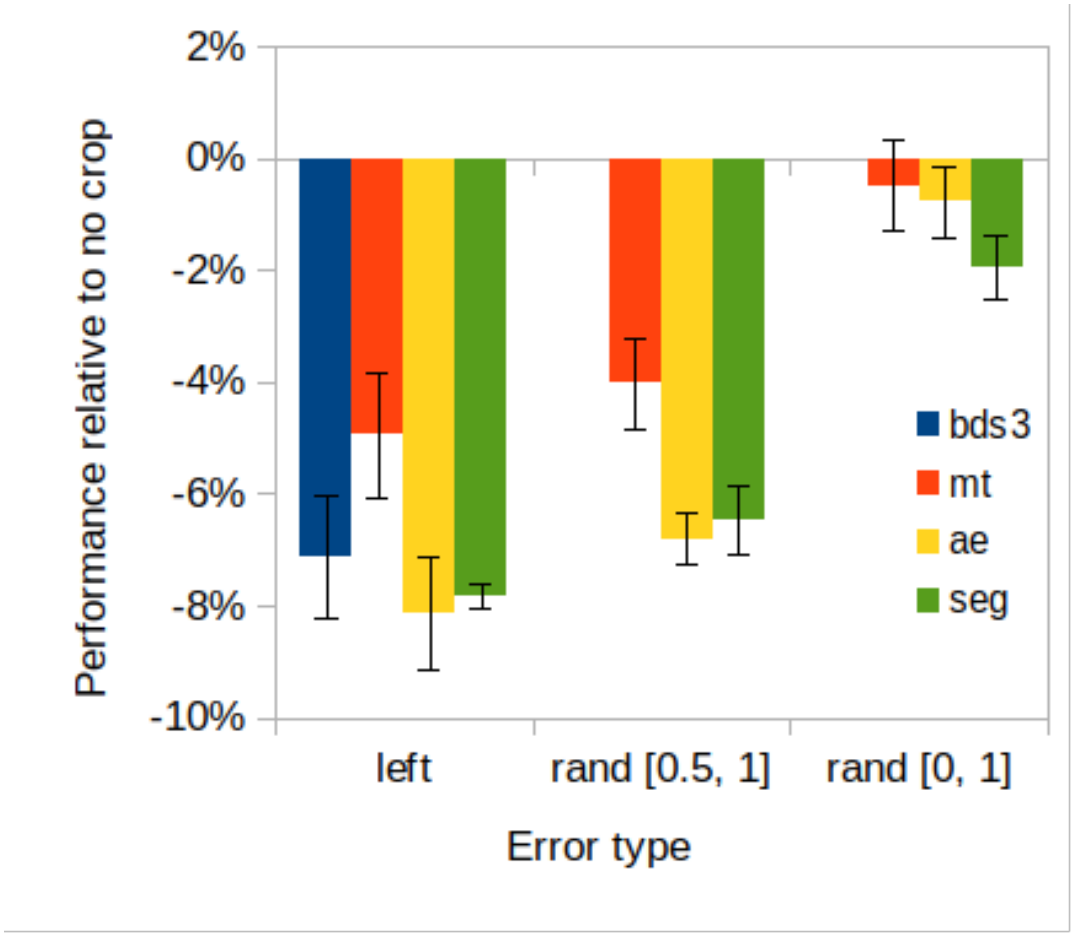




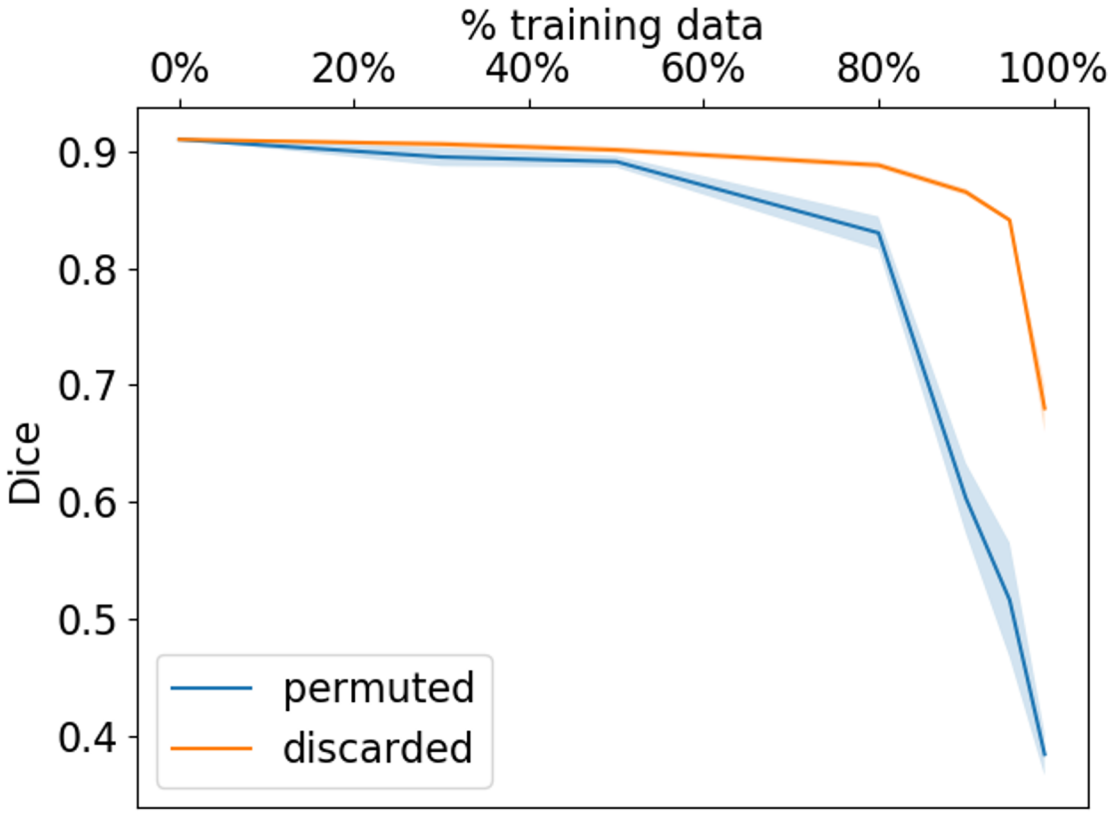
Random crop

On average, crop half the tumour





Permutation



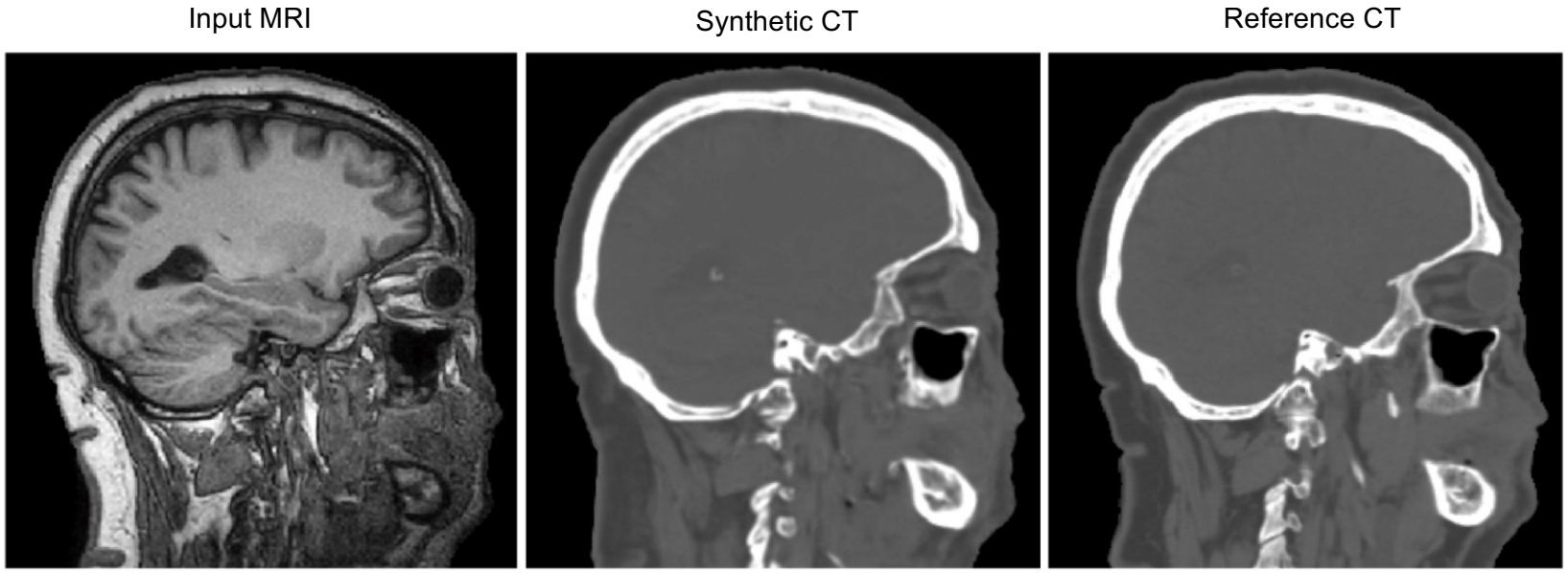
- Permute labels across *% training data*
- VS discard *% training data*
- **Permutation is worse**

Summary

Random warp	Constant shift	Random crop	Permutation
Unbiased	*Biased*	*Varying bias*	*Biased*
Low error	High error	Low bias: low error Med bias: med error High bias: high error	High error

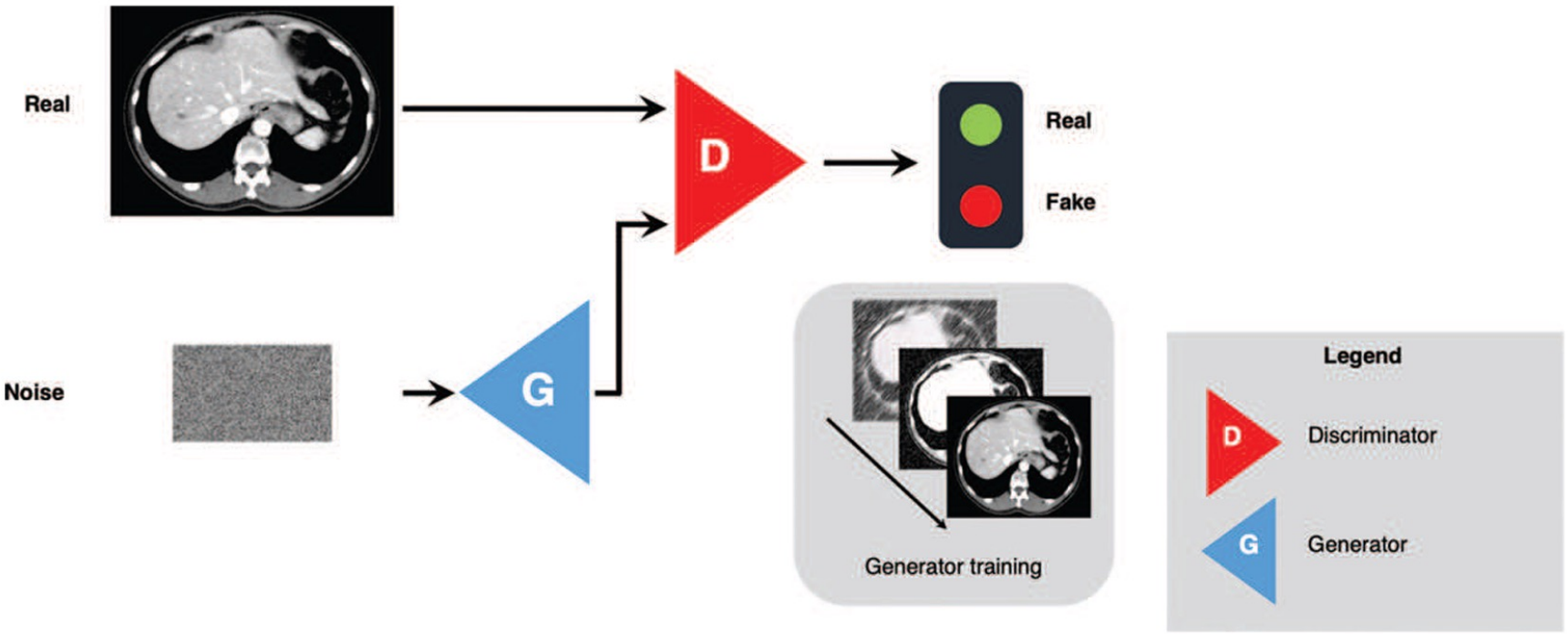
3) Image synthesis

Image synthesis example



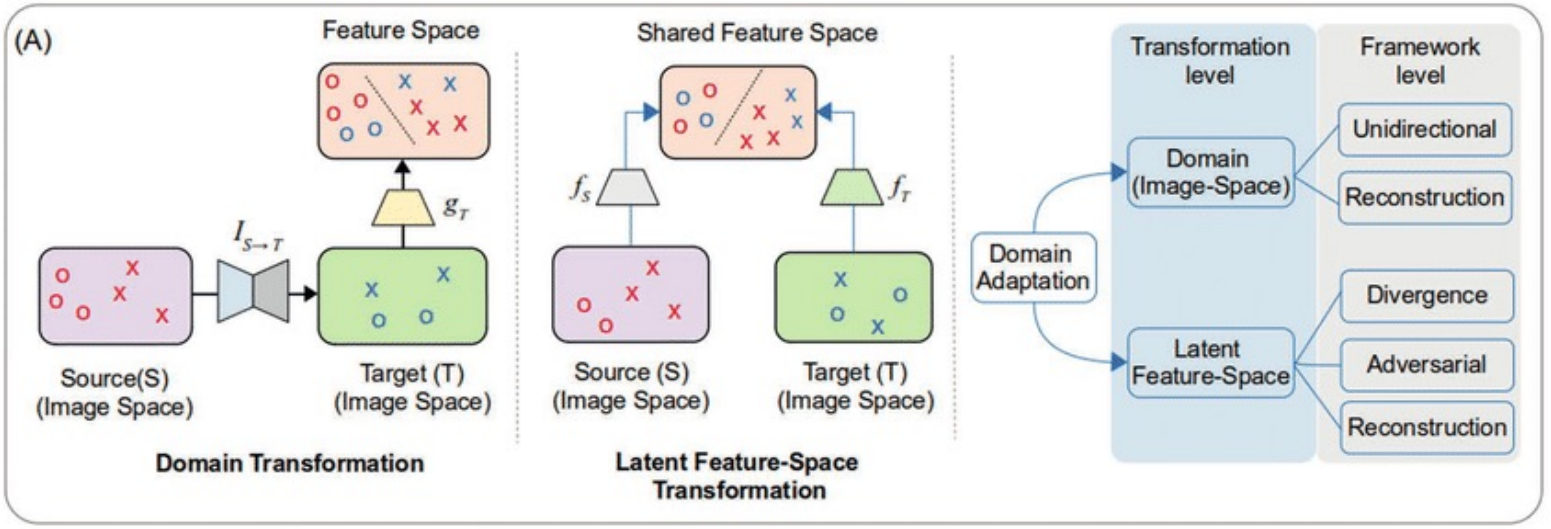
MRI to CT Synthesis with a CycleGAN
Source: Wolterink et al. 2017

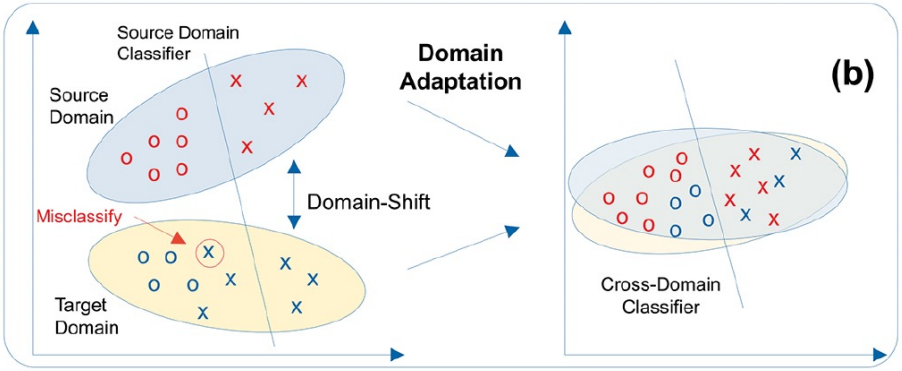
Generative adversarial networks



Standard GAN architecture diagram. A first network, known as the generator (G), aims to transform a random input into a realistic image to fool a second network, known as the discriminator (D). During training, the generator learns from the response of the discriminator.

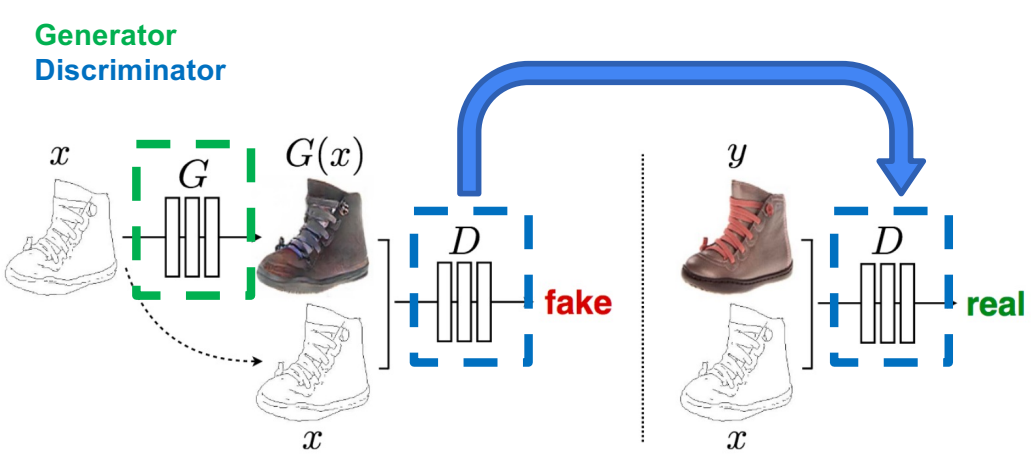
Domain transformation scenarios



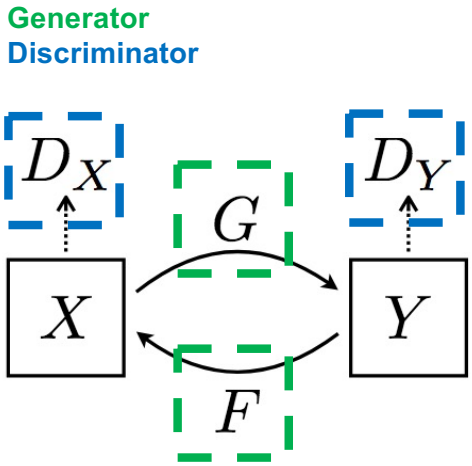


Choudhary et al. 2020

Generative adversarial networks

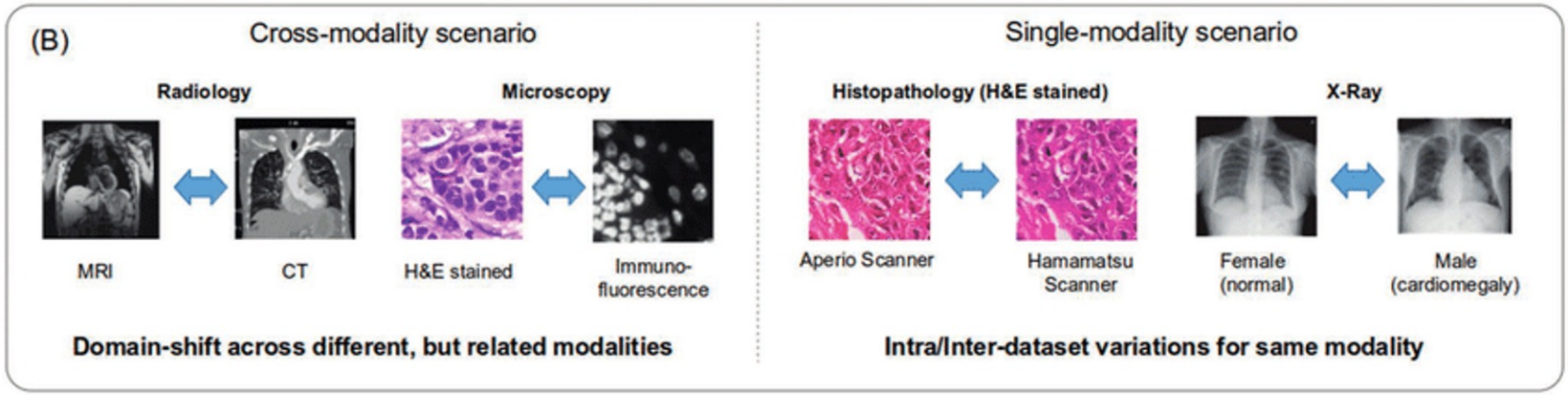


*Sketch-to-picture synthesis with **conditional generative adversarial network** (cGAN)*
Source: Isola et al. 2017



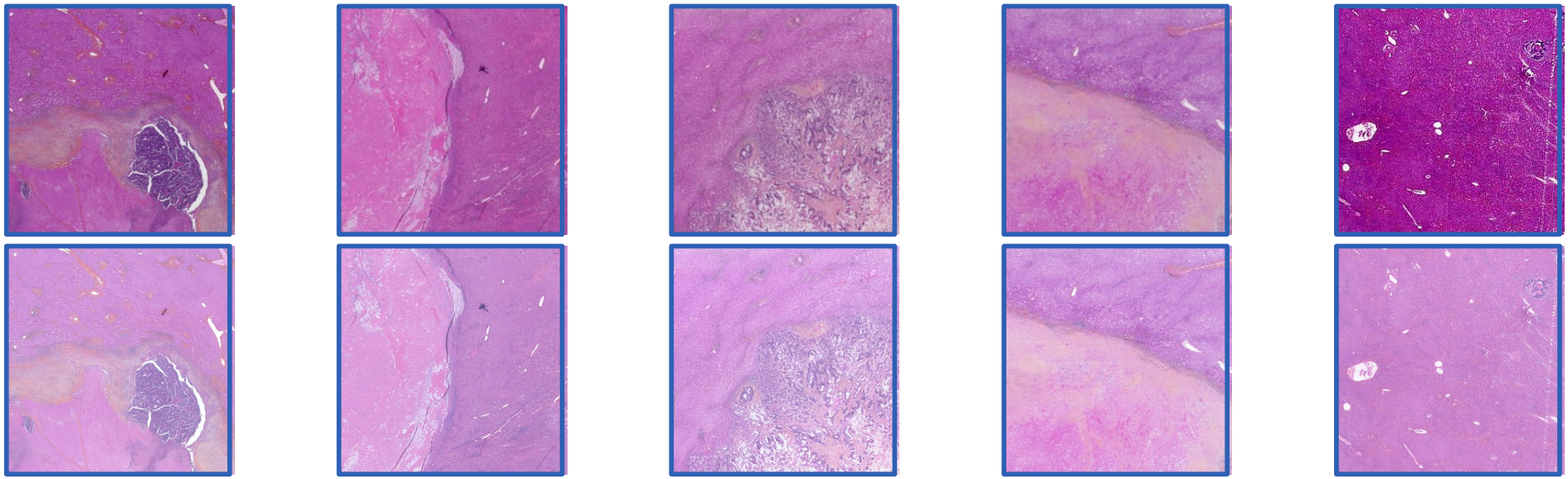
*Architecture of a **cycle-consistent generative adversarial network** (CycleGAN)*
Source: Zhu et al. 2017

Examples of normalized slides with the GAN model

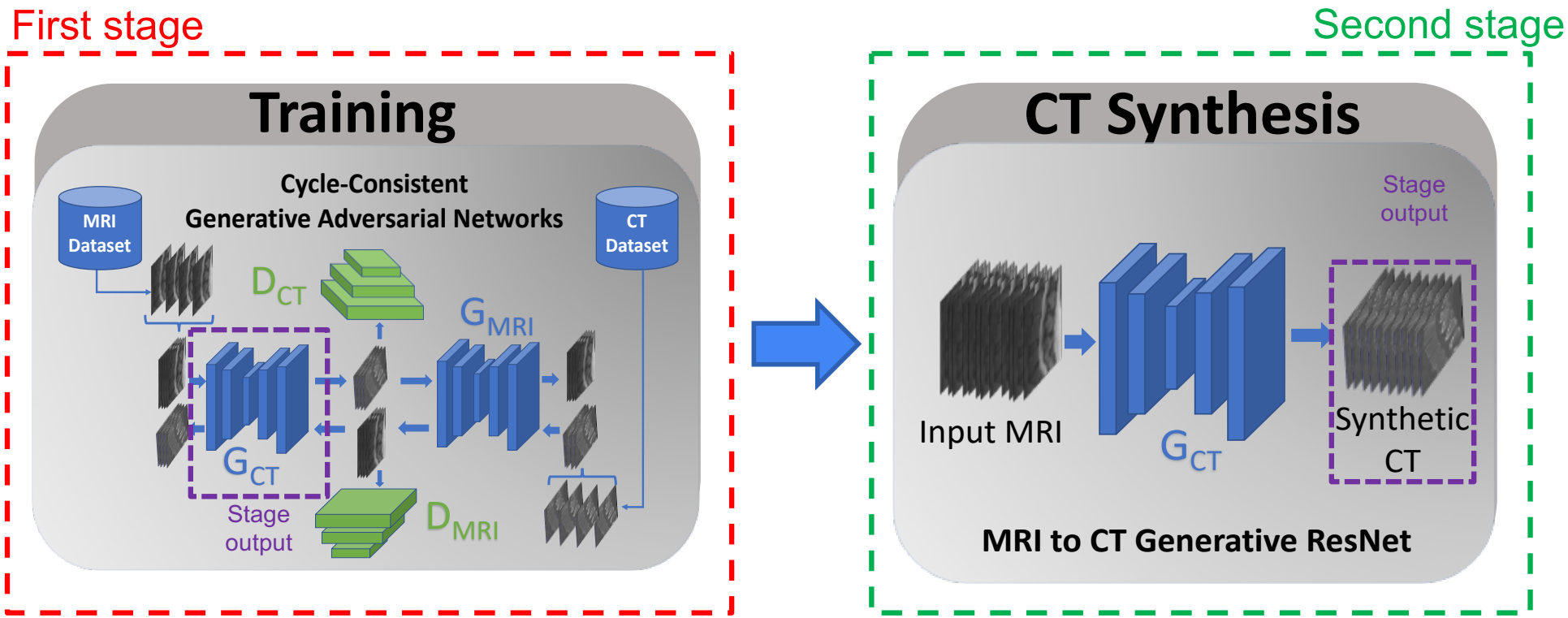


Original

Normalized

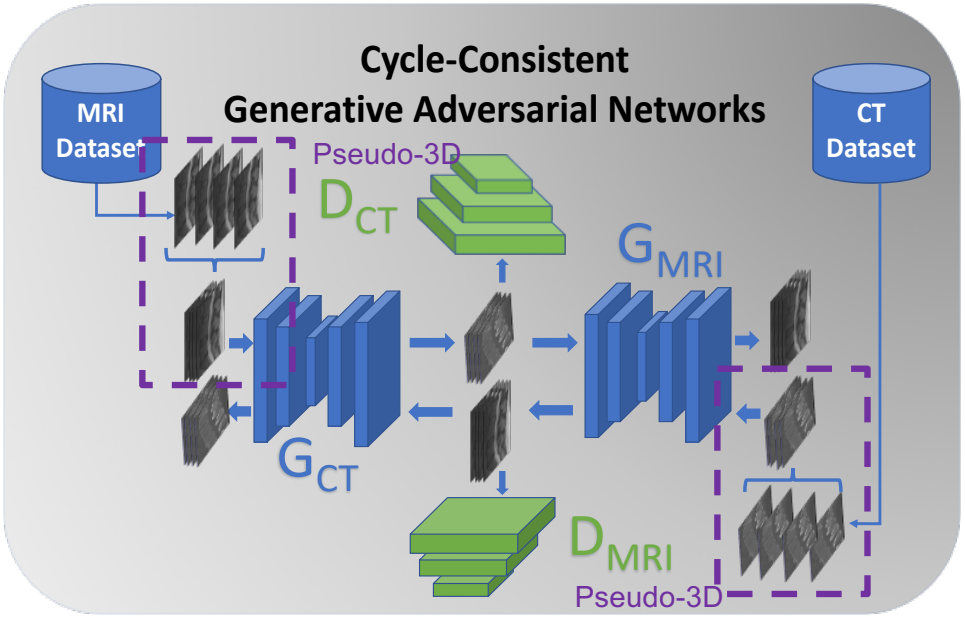


MRI to CT synthesis pipeline



Oulbacha et al. Med. Phys. 2020

Cycle-Consistent GAN

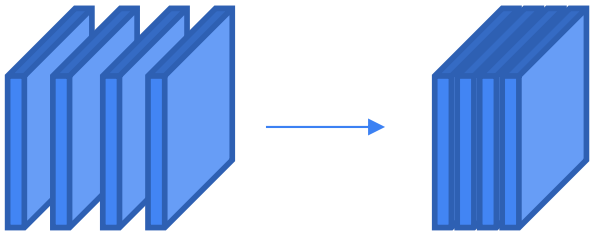


CycleGAN MRI and CT architecture

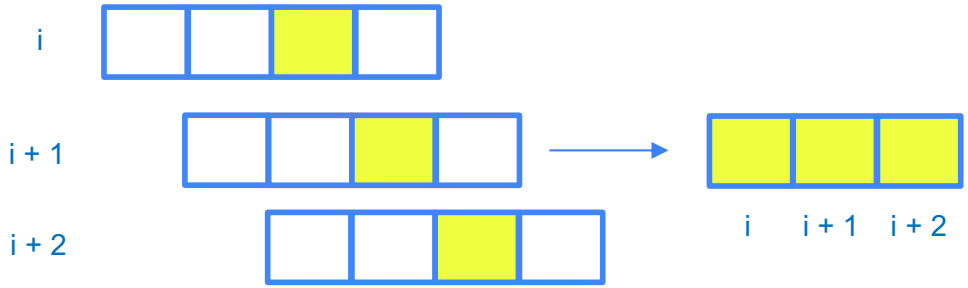
- **Image synthesis using CycleGAN model**, jointly learning both **MRI → CT and CT → MRI** mappings.
- We use a **pseudo-3D** approach, contrary to the more classical single-slice 2D approach
- Pseudo-3D consists of **stacking neighboring slices** along the channel dimension, forming a **thick slice**

Oulbacha et al. Med. Phys. 2020

Pseudo-3D recombination



Slice thickening strategy.

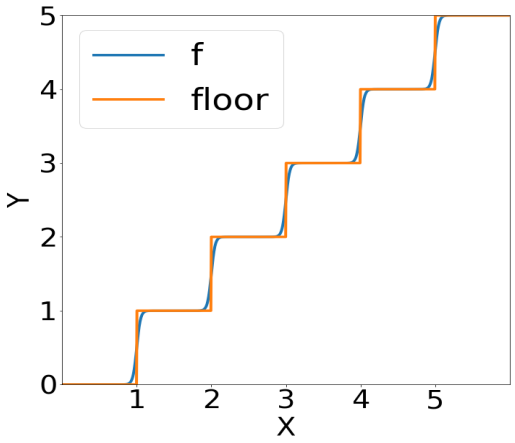


Recombining strategy.
Left, quadruplet slices. Right, output volume slices.

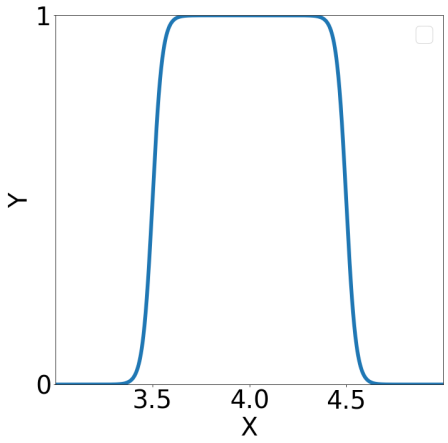
- Pseudo-3D is performed for each **quadruplet** of neighbouring slices
- **Recombining quadruplets** into a single volume is done selecting the third slice from each quadruplet and adding it to the output volume
- With this recombining strategy, slices at indices i and i + 1 in the output volume necessarily come from **quadruplets with 3 slices in common**

Oulbacha et al. Med. Phys. 2020

Differential histograms



Floor function and smooth counterpart.



Smooth indicator function for the (4-0.5, 4+0.5) interval.

- **Differentiable histograms** for image synthesis, to be used with **backpropagation**
- Regular histogram density computation has **nil gradient** due to the **non-differentiable floor and Dirac indicator functions.** It is thus impossible to use backpropagation.
- Proposed solution: **smooth out** those functions

Oulbacha et al. Med. Phys. 2020

CycleGAN loss

- CycleGAN loss function

$$L\left(G_{CT}, G_{MRI}\right) = \left\|D_{CT}\left(CT_{syn}\right) - 1\right\|_2^2 + \left\|D_{MRI}\left(MRI_{syn}\right) - 1\right\|_2^2 + \lambda_{cycle}\left[\left|MRI_{recons} - MRI\right|_1 + \left|CT_{recons} - CT\right|_1\right] + \lambda_{hist}\left[\left|H^{MRI_{recons}} - H^{MRI}\right|_1 + \left|H^{CT_{recons}} - H^{CT}\right|_1\right]$$

Adversarial loss

Cycle-consistency loss

Histogram loss

$$L\left(D_{CT}, D_{MRI}\right) = \left\|D_{CT}(CT) - 1\right\|_2^2 + \left\|D_{MRI}(MRI) - 1\right\|_2^2 + \left\|D_{CT}\left(CT_{syn}\right)\right\|_2^2 + \left\|D_{MRI}\left(MRI_{syn}\right)\right\|_2^2$$

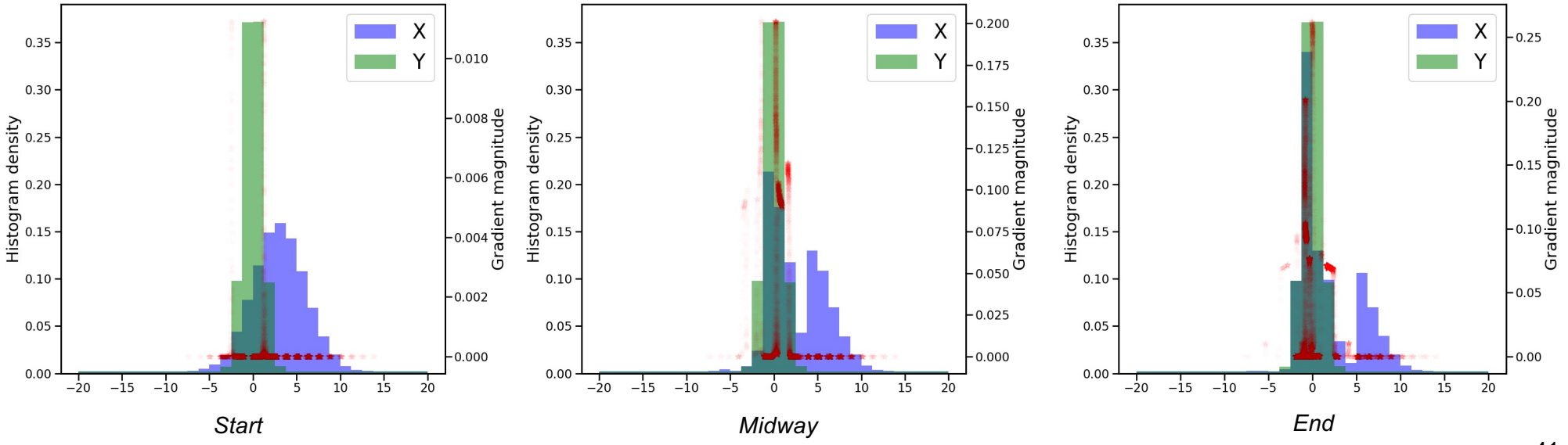
$\lambda_{cycle} = 10$
$\lambda_{hist} = 0.1$
$n_{bins} = 32$

Oulbacha et al. Med. Phys. 2020

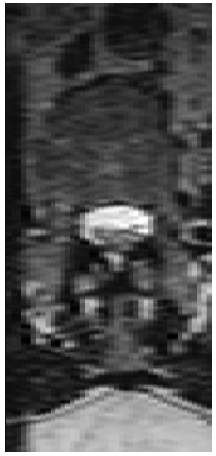
Histogram similarity

- Toy experiment, with X being a vector sampled from N(3,3), and Y being a vector sampled from N(0,1). We **minimize the difference of histograms** with respect to X

Gradient magnitudes -

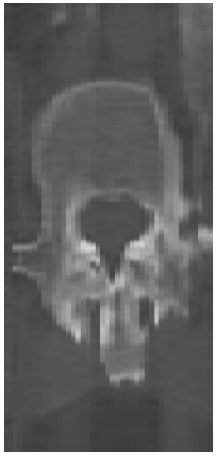


Samples from the image synthesis stage

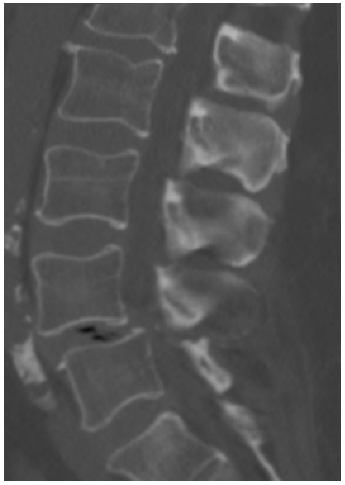


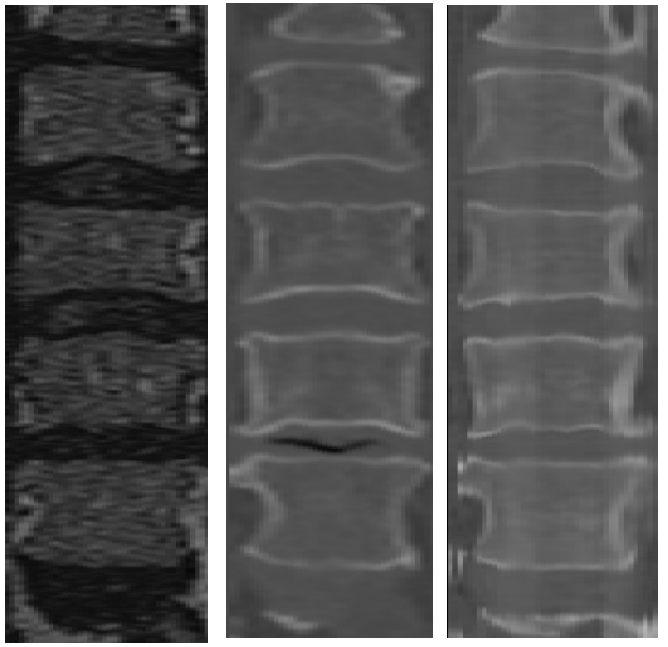
MRI

CT



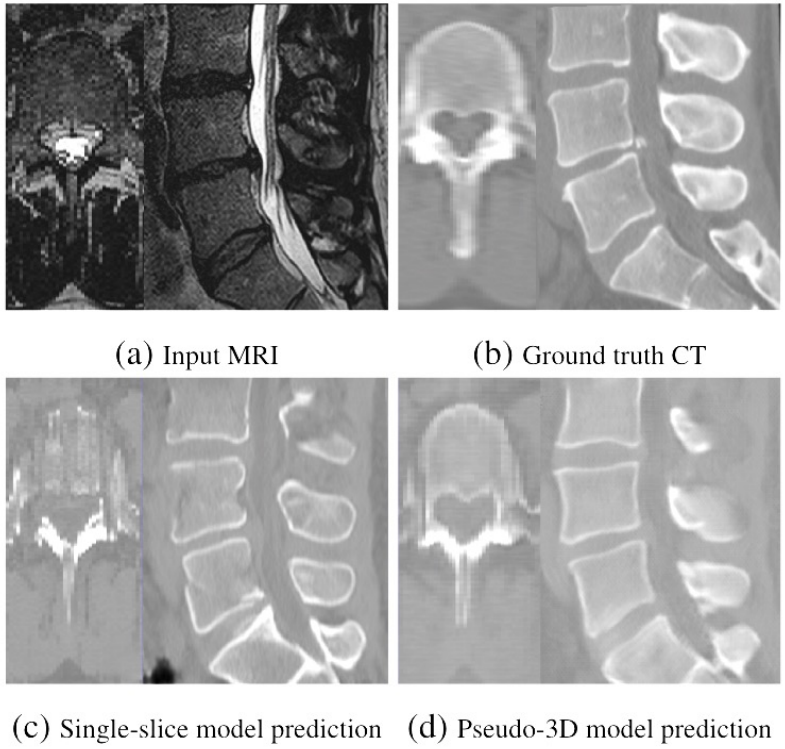
Synthetic CT





MRI *CT* *Synthetic CT*

Axial resolution

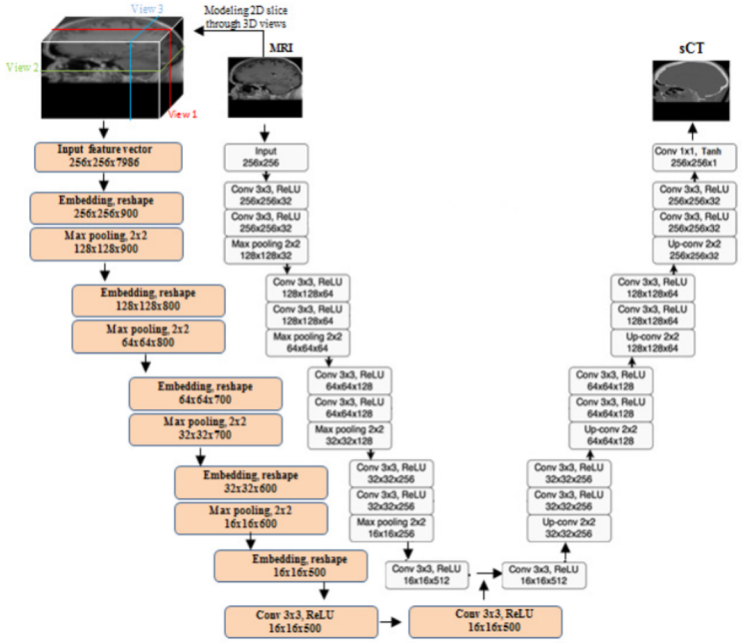


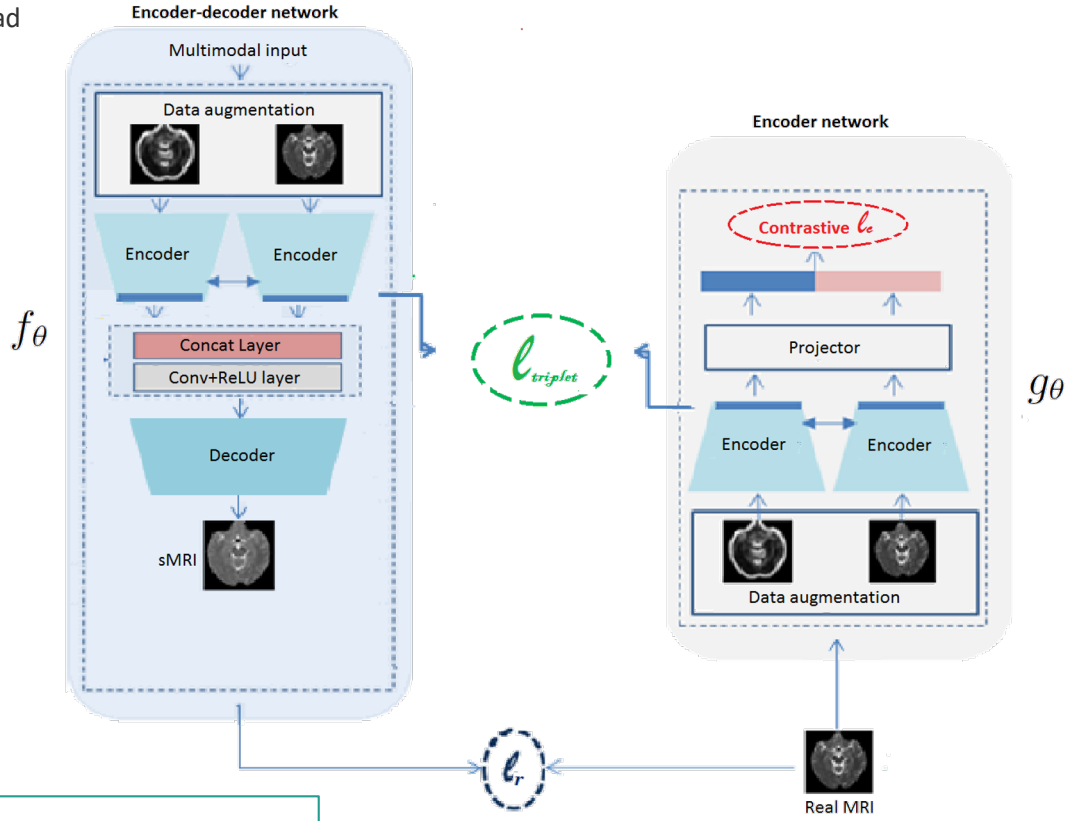
Sample axial and sagittal slices for one patient.

- **Axial resolution improvement** with the **pseudo-3D** approach, opposed to a single-slice 2D approach

Contrastive feature GAN

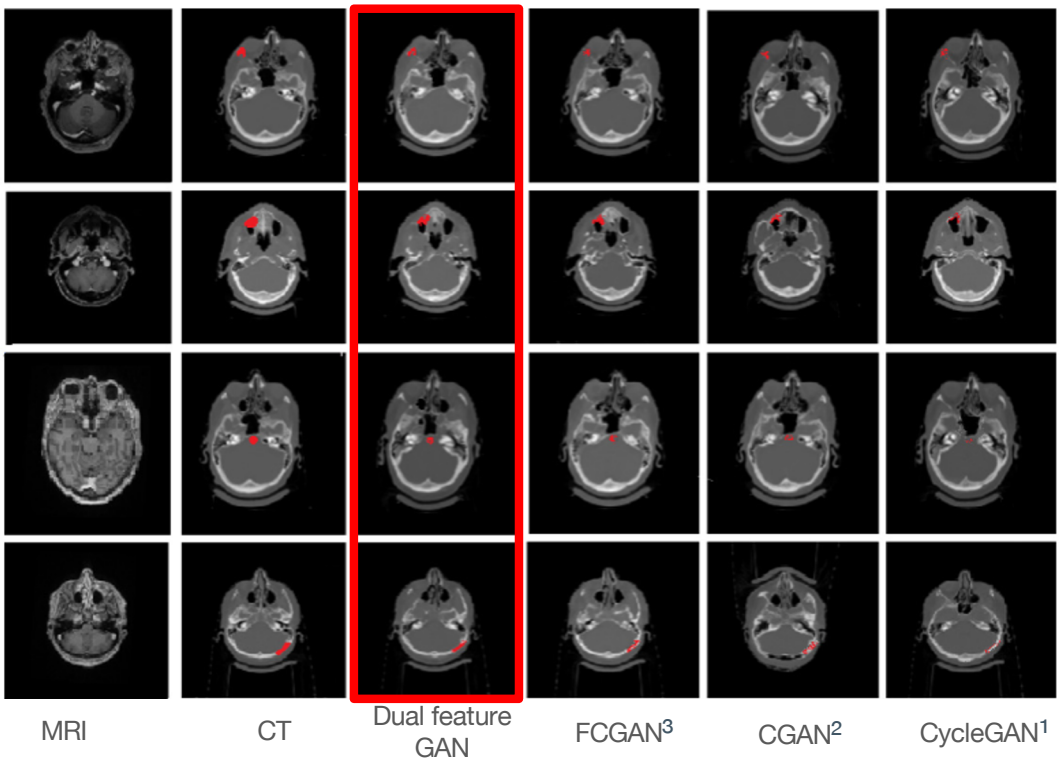
R. Touati et al. “A feature invariant generative adversarial network for head and neck MRI/CT image synthesis”, **Phys. Med. Biology,** 2021.

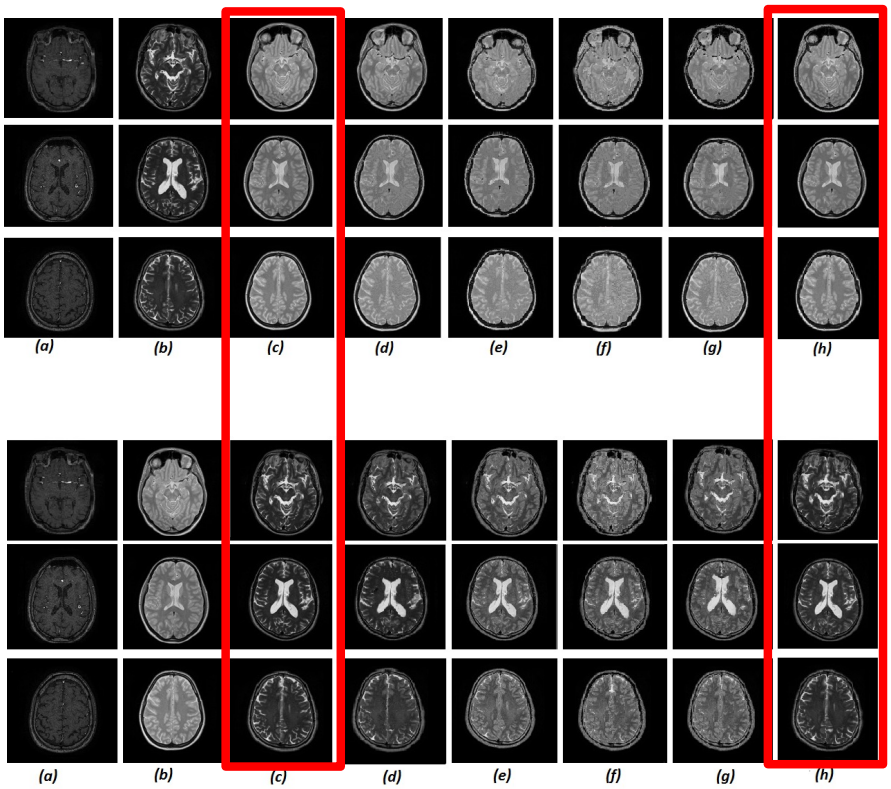




Contrastive feature GAN. Generative adversarial network augmented with multi-planar dynamic features to improve spatial positioning and preserve anatomical structures.

MRI-CT synthesis results



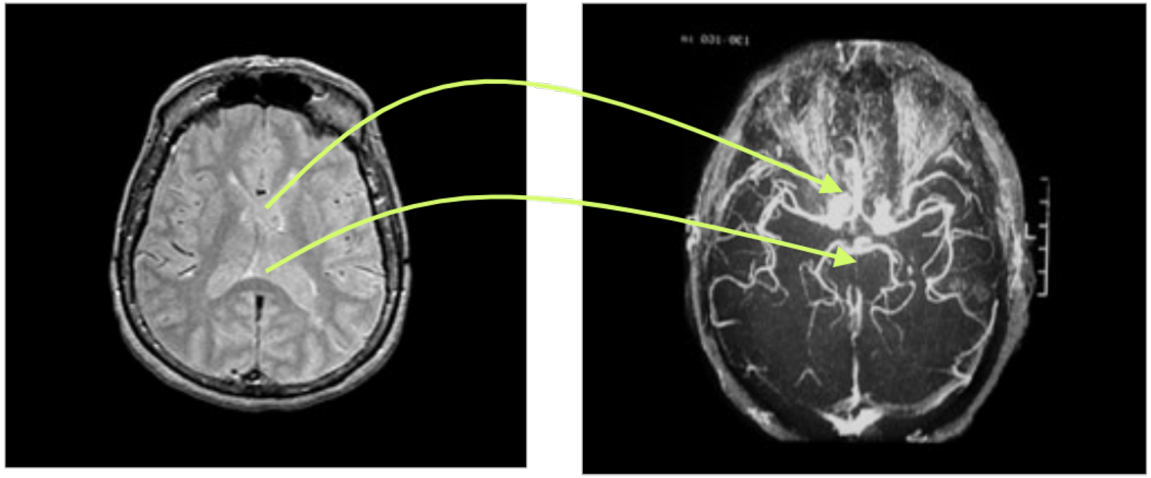


1. Isola P, Zhu JY, Zhou T, Efros AA. *Image-to-image translation with conditional adversarial networks*. CVPR 2017.
2. Oulbacha R, Kadoury S. *MRI to CT Synthesis of the Lumbar Spine from a Pseudo-3D Cycle GAN*. ISBI 2020.

4) Deformable registration

Deformable registration

" Image registration is about determining a spatial transformation – or mapping – that relates positions in one image, to corresponding positions in one or more other images"

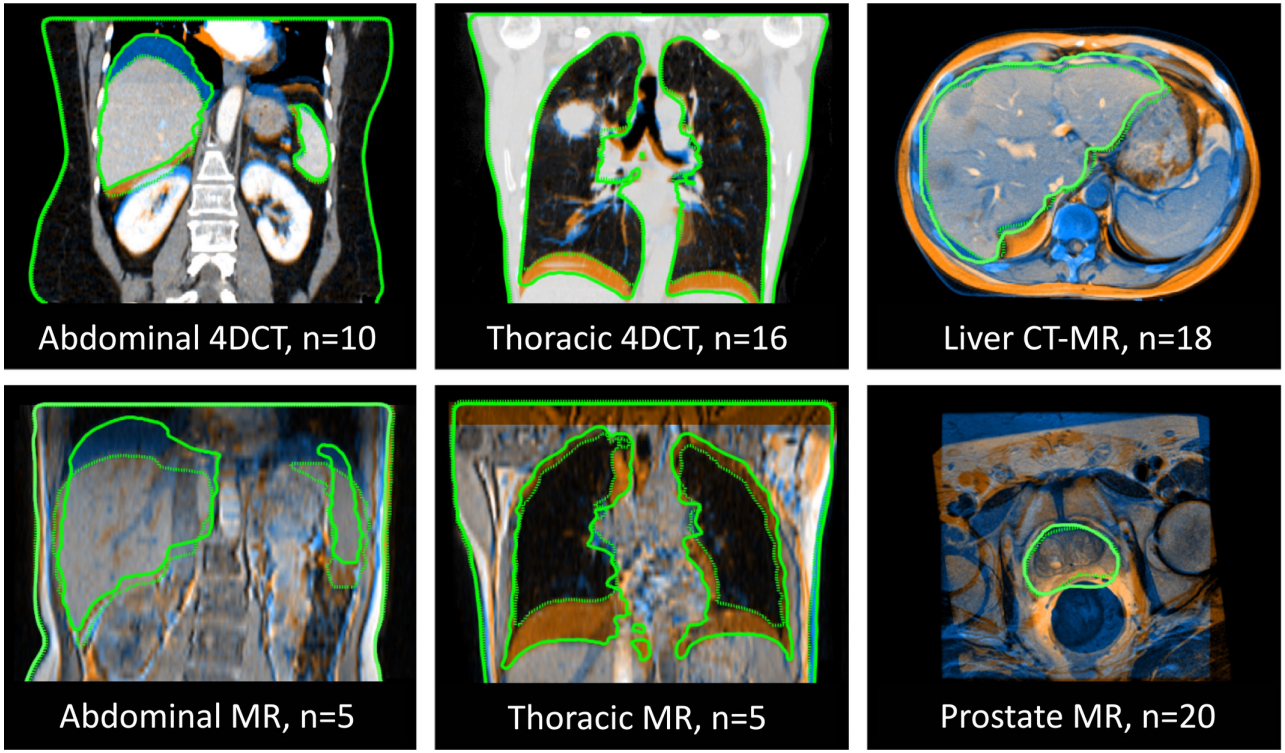


Source image

Target image

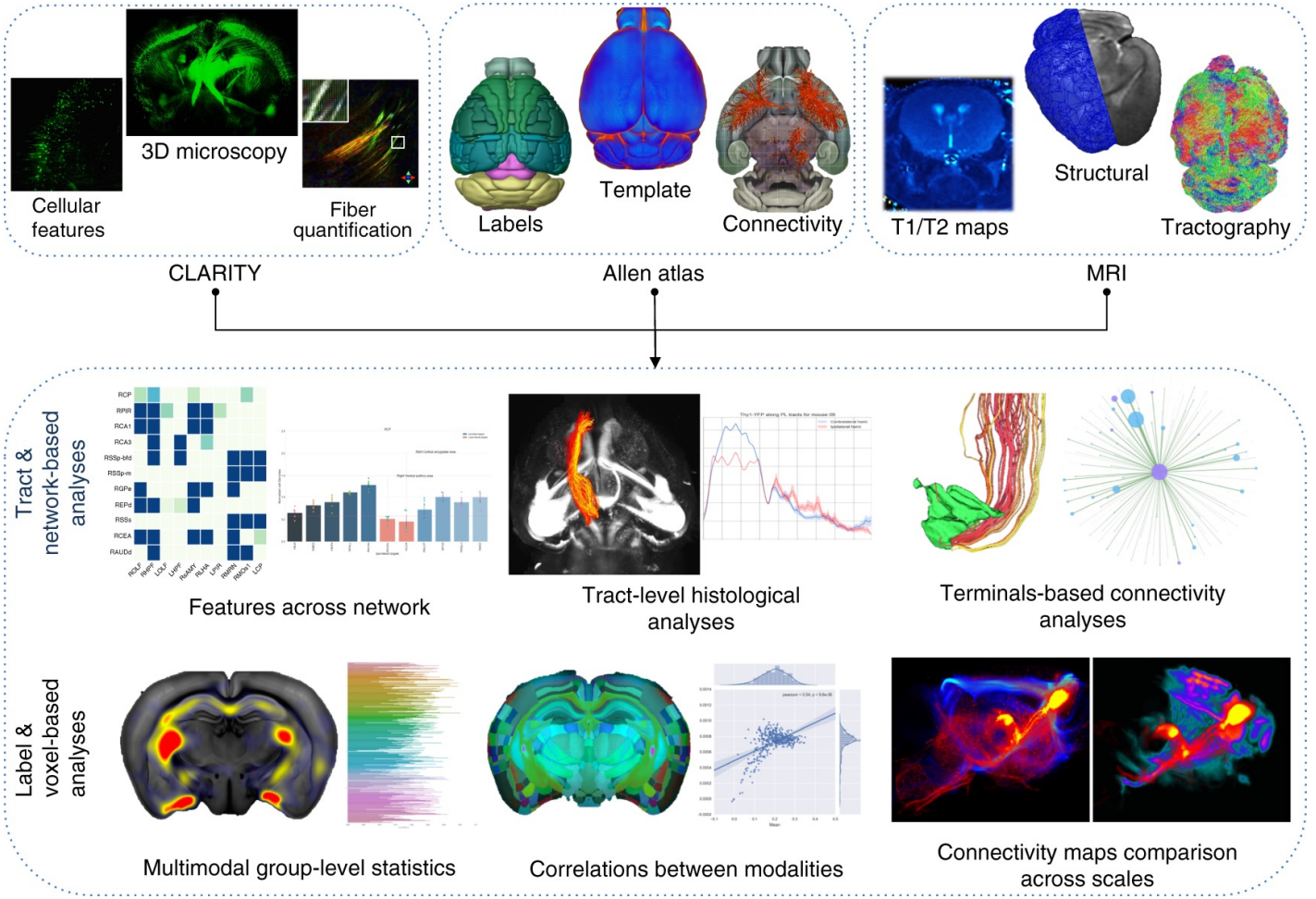
- 3D - 3D
- 3D - 2D
- 3D/2D - patient

Organ shift and deformations



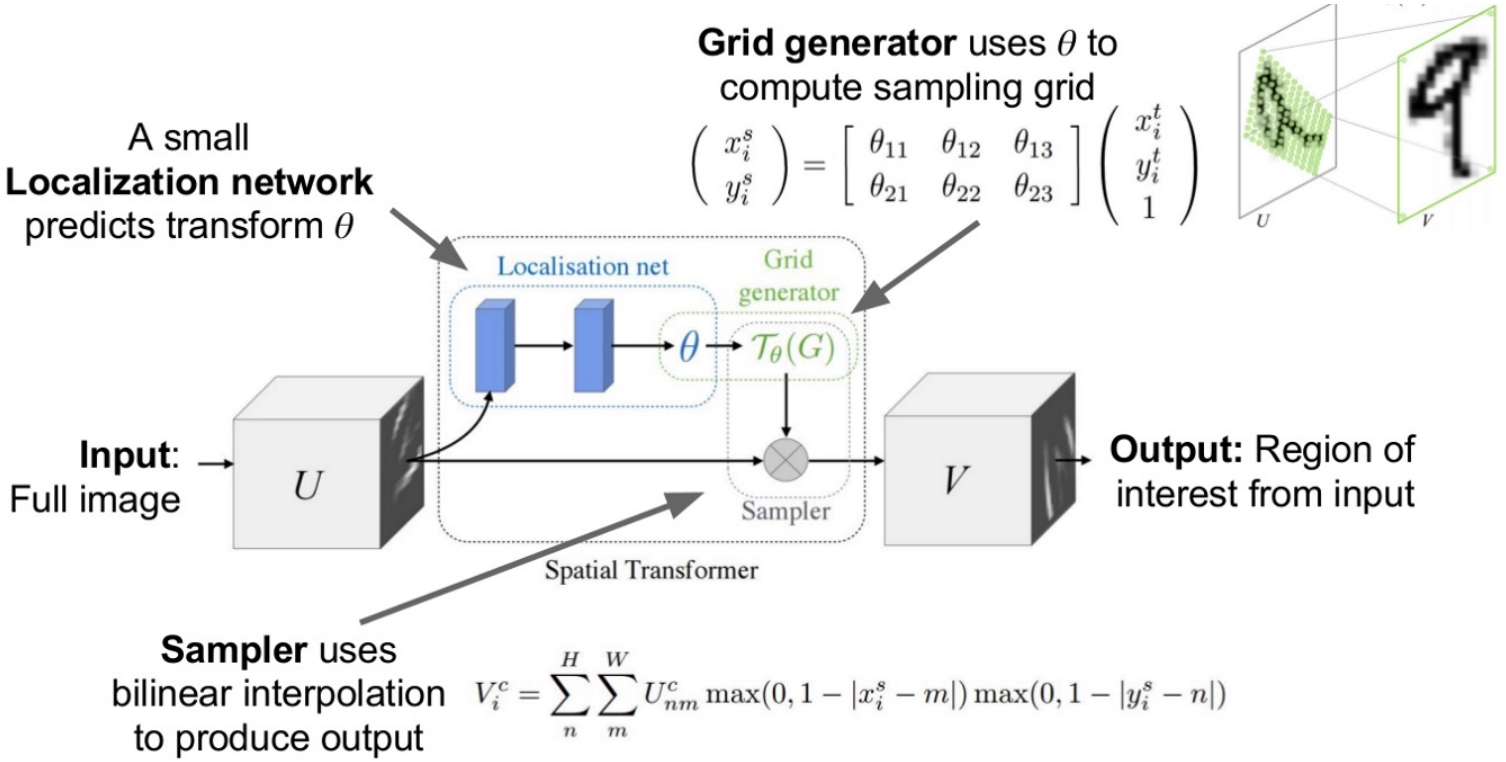
Velec et al.2020, Med. Phys.

Applications in deformable image registration



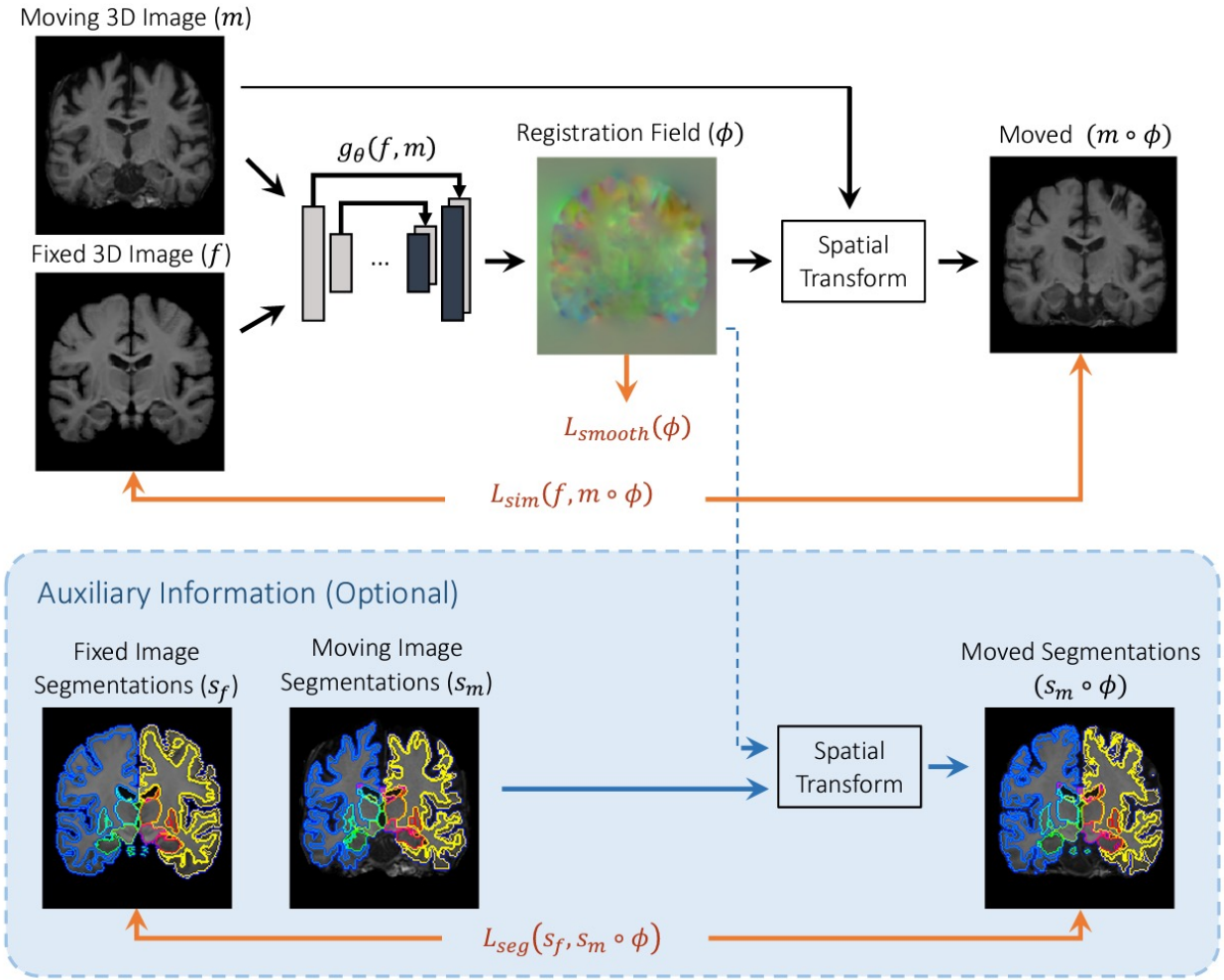
Goubran et al.2020, Nature

Spatial transformer networks



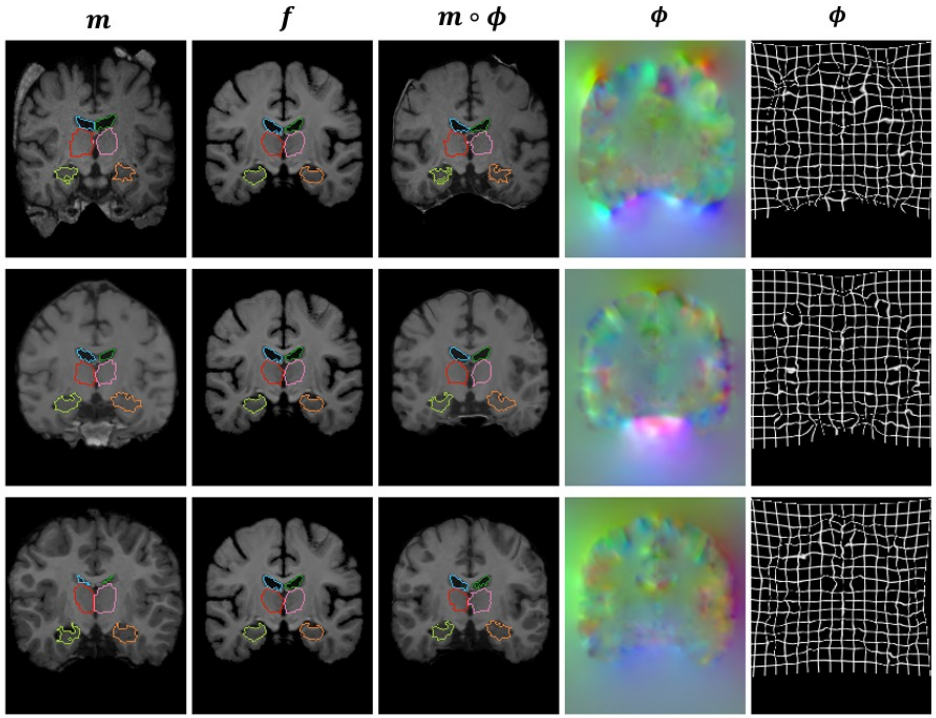
Jaderberg et al.2015, NeurIPS

VoxelMorph



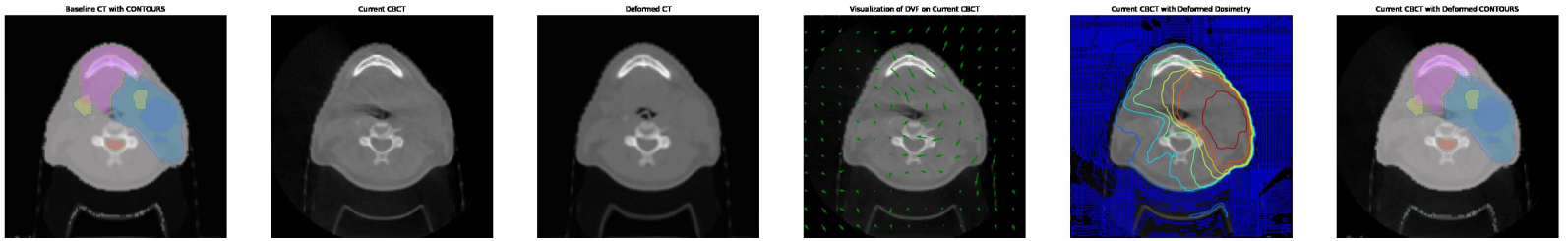
Balakrishnan et al.2019, IEEE TMI

Results with Voxelmorph



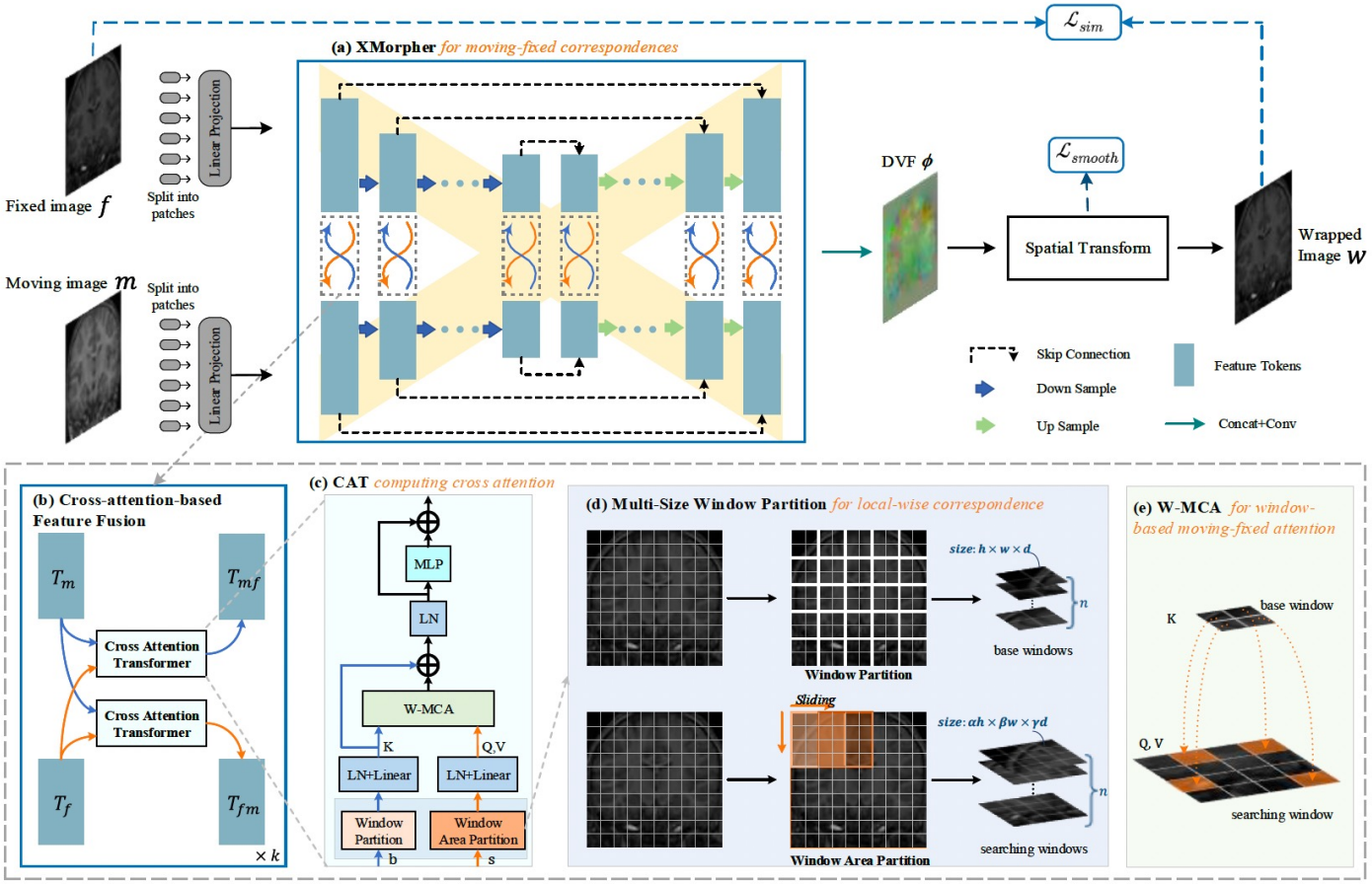
Method	Dice
Affine only	0.579 (0.173)
ANTs SyN (CC)	0.761 (0.117)
NiftyReg (CC)	0.772 (0.117)
VoxelMorph (MSE)	0.727 (0.146)
VoxelMorph x2 (MSE)	0.750 (0.058)
VoxelMorph x2 (MSE) inst.	0.764 (0.048)
VoxelMorph (CC)	0.737 (0.139)
VoxelMorph x2 (CC)	0.763 (0.049)
VoxelMorph x2 (CC) inst.	0.772 (0.119)

TABLE III: Results for subject-to-subject alignment using affine, ANTs, and VoxelMorph variants, where "x2" refers to a model where we doubled the number of features to account for the increased inherent variability of the task, and "inst." indicates additional instance-specific optimization.



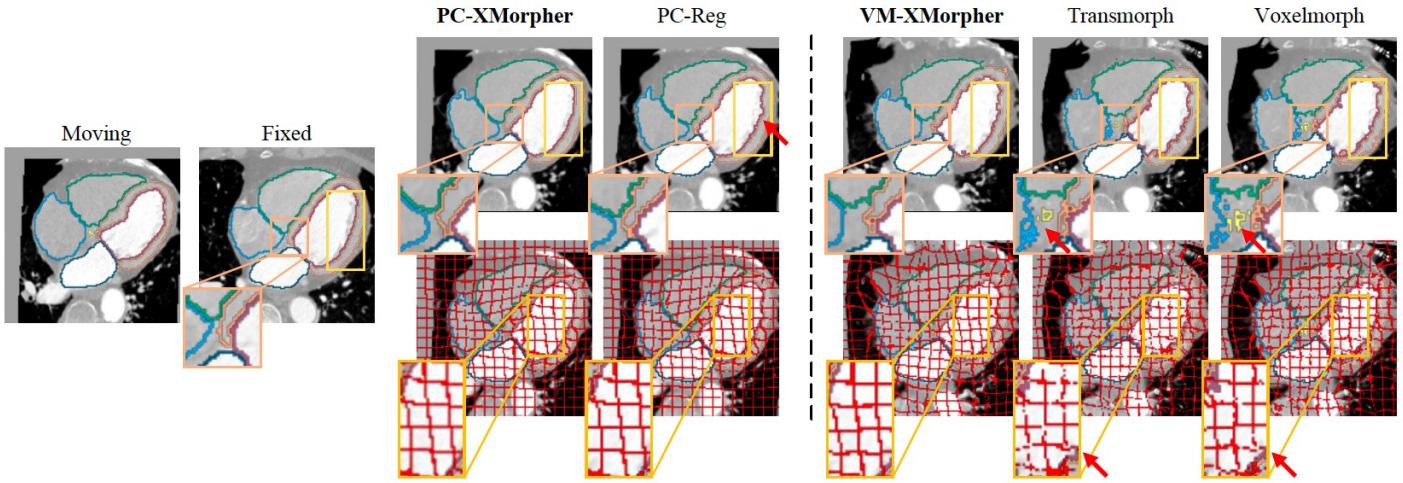
Balakrishnan et al.2019, IEEE TMI

XMorpher



Shi et al. MICCAI 2022

XMorpher



Method	Un-/Semi-	Backbone	DSC	$\lvert J_\phi \rvert \leq 0$ (%)
Affine initialization	-	-	69.2 ± 7.2	-
BSpline [14]	Unsup	-	80.9 ± 7.6	5.25 ± 3.27
Voxelmorph [1]		CNN	80.2 ± 5.5	4.02 ± 0.82
Transmorph [4]		CNN+Transformer	81.1 ± 5.2	3.46 ± 0.75
Our no-cross XMorpher		Full Transformer	81.5 ± 5.4	**0.94±0.26**
Our VM-XMorpher		Full Transformer	**83.0±4.7**	3.15 ± 0.79
PC-Reg[8]	Semi	CNN	86.0 ± 2.5	0.36 ± 0.20
Our PC-XMorpher		Full Transformer	**86.9±2.4**	**0.32±0.18**

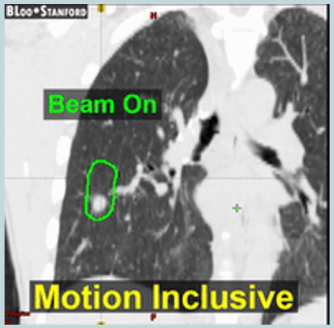
Shi et al. MICCAI 2022

5) Motion modeling

Motion management strategies

Non-adaptive

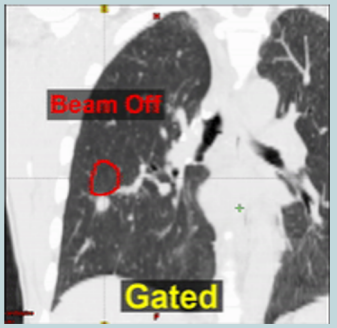
Large margins



Abdominal compression

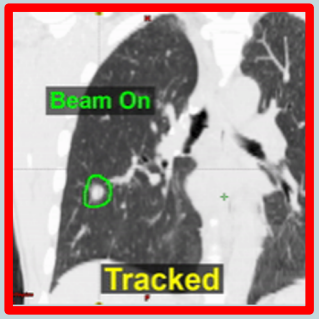
Breath-hold

Gating



Real-time adaptive tracking

- Move or shape the radiation beam **dynamically**
- **Accuracy** depends on the system **adapting** to the moving anatomy
- Driven by some **surrogate** signal to estimate the target position

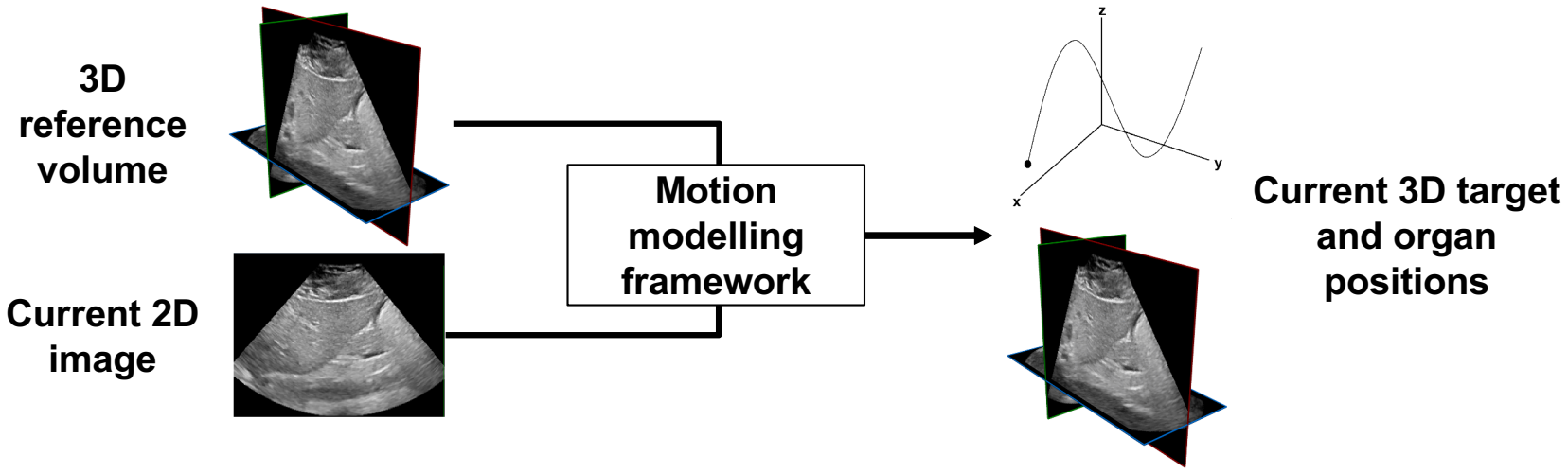


Source: https://tibaray.com/medical.html

Real-time 3D ultrasound generation and tracking

Motion modelling framework to generate up-to-date 3D organ and target positions from input 2D US images and a limited number of reference volumes

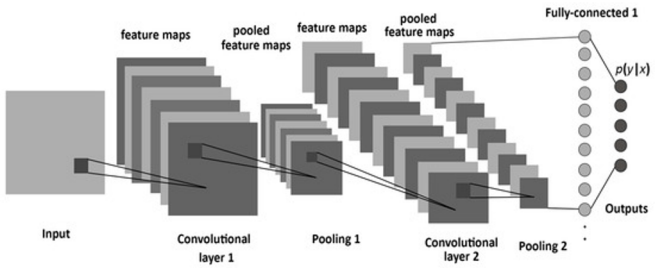
- Train DNN with 4D US dataset of the liver
- Develop motion modelling framework
- Evaluate framework on 3D target tracking in US volumes



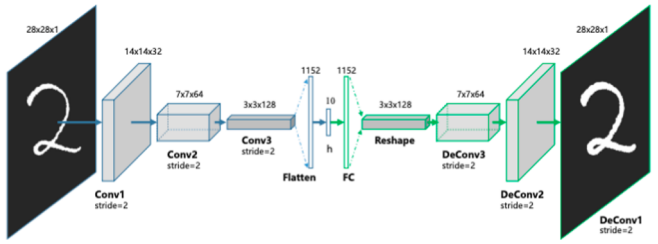
Deep learning for motion modelling

- Architectures
  - CNN ⟶ Feature extraction
  - Convolutional AE ⟶ Representational power
  - STN ⟶ Spatial warping
  - GAN ⟶ Generative capacity
- Advantages
  - Good generalization
  - No inter-subject correspondences
- Disadvantages
  - Requires large amount of data

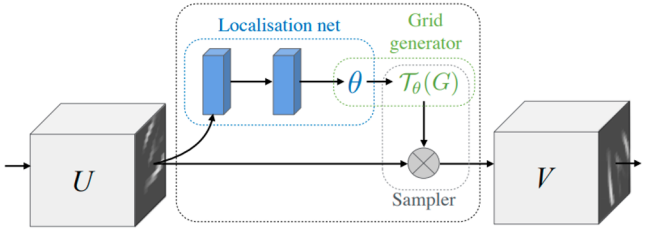
Convolutional neural network (CNN)[1]



Convolutional autoencoders[2]



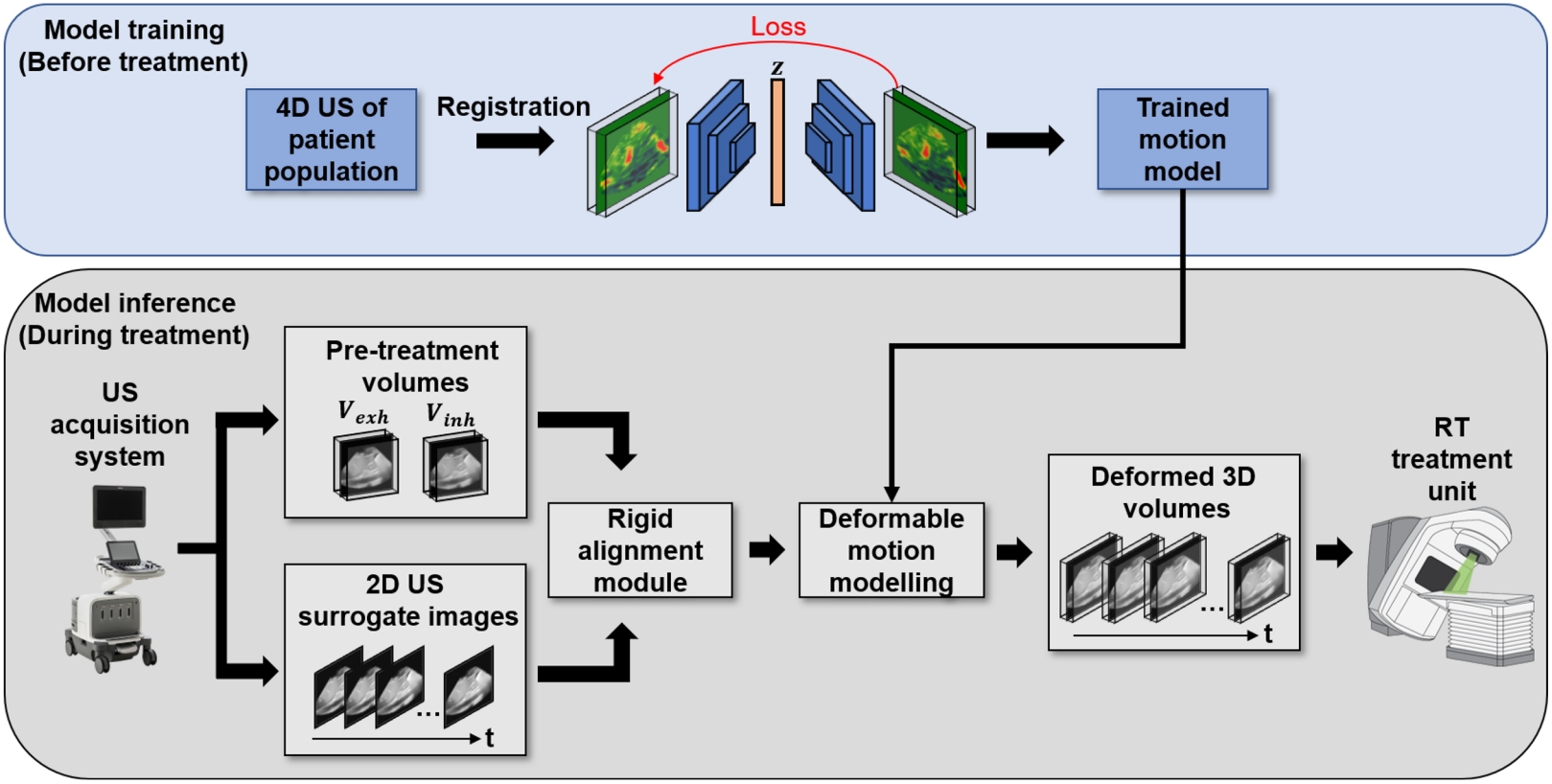
Spatial transformation network (STN)[3]



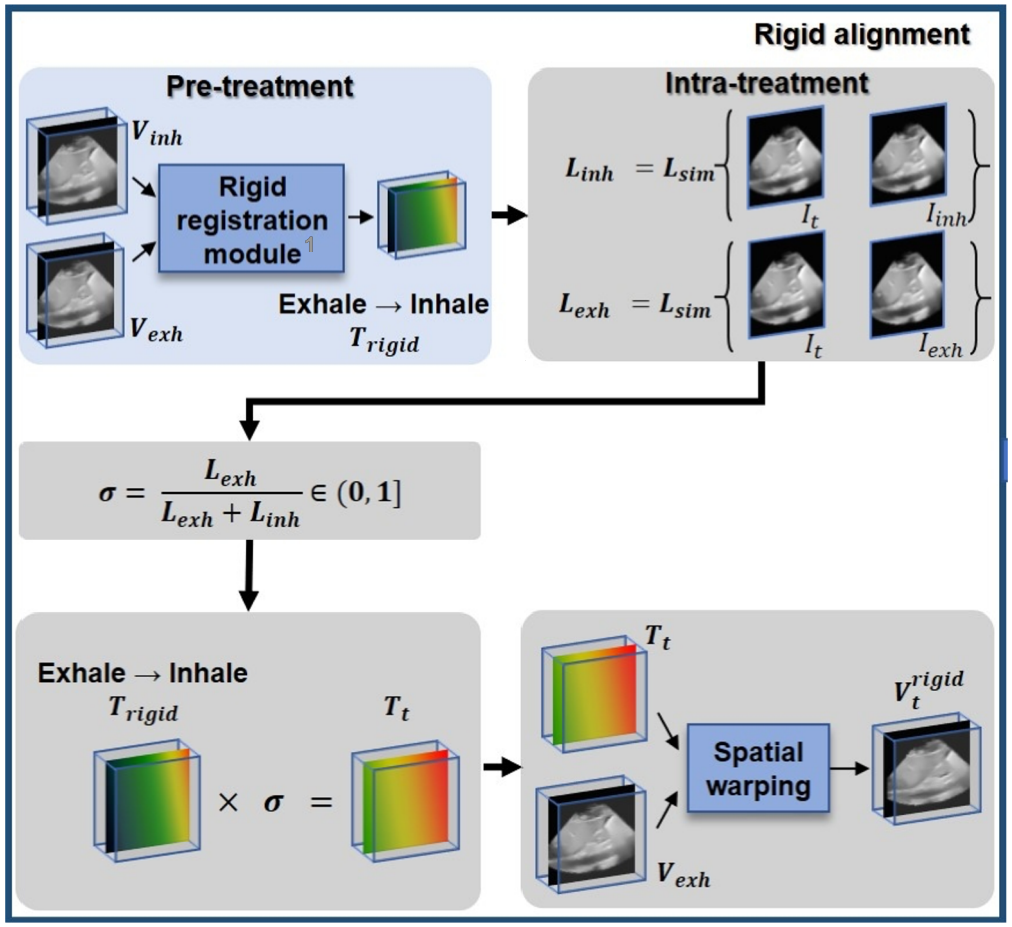
1. S. Albelwi et A. Mahmood, “A framework for designing the architectures of deep convolutional neural networks,”Entropy, vol. 19, no. 6, 2017.
2. Towards Data Science. (2019) Convolutional autoencoders for image noise reduction. [Online].
3. M. Jaderberg, K. Simonyan, A. Zissermanet al., “Spatial transformer networks,” in Advances in neural information processing systems, 2015, p. 2017–2025.

Overview of US-based motion model

Tal Mezheritsky et al., “Population-based 3D motion modelling from convolutional autoencoders for 2D ultrasound-guided radiotherapy” **Medical Image Analysis**, 2022.



Rigid alignment module



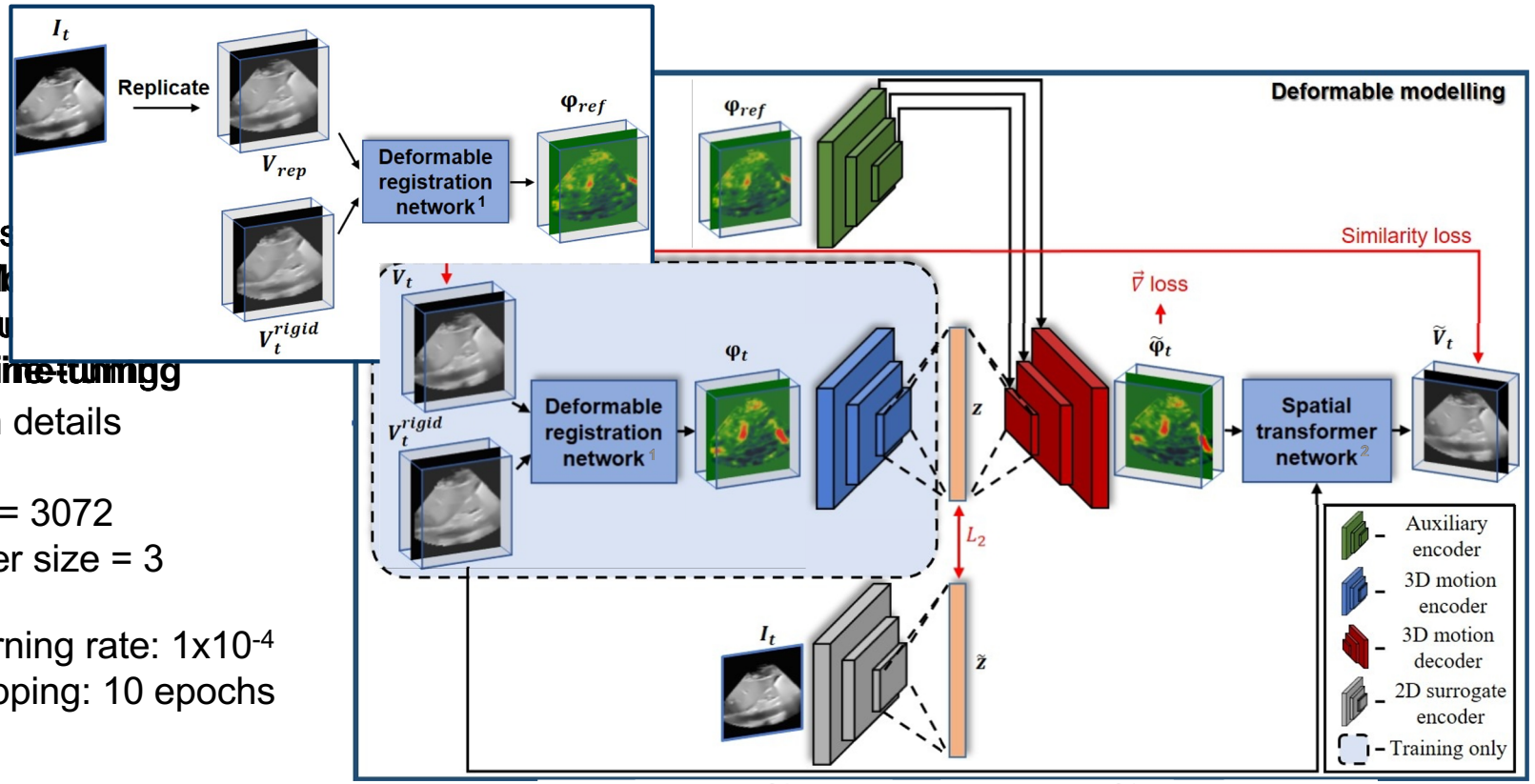
1. S. Klein, M. Staring et al., “elastix : a toolbox for intensity-based medical image registration,”IEEE Transactions on Medical Imaging,vol. 29, no. 1, p. 196 – 205, January 2010.

Deformable motion model

Training progres
- **Step 1: M**
- **Step 2: S**
- **Step 3: Fine-tuning**

Implementation details
- β = 0.01
- Size of z = 3072
- Conv. filter size = 3

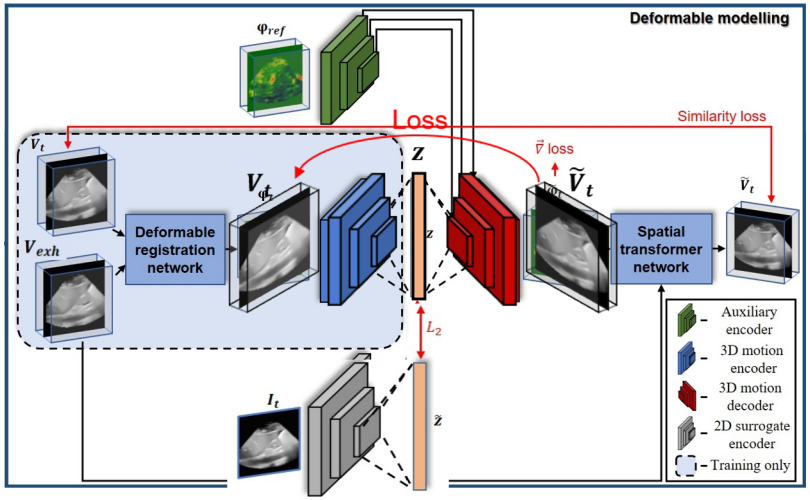
Training details
- Initial learning rate: $1x10^{-4}$
- Early stopping: 10 epochs



$$\mathcal{L} = \mathcal{L}_{sim}(\tilde{V}_t, V_t) + \beta \mathcal{L}_{grad}(\tilde{\phi}_t) + ||z, \tilde{z}||_2^2$$

1. A. V. Dalca, G. Balakrishnan, J. Guttag et al., "Unsupervised learning for fast probabilistic diffeomorphic registration,"Lecture Notes in Computer Science, p.729–738, 2018.
2. M. Jaderberg, K. Simonyan, A. Zisserman et al., "Spatial transformer networks," in Advances in neural information processing systems, 2015, p. 2017–2025.

Ablation study



- Deformation properties
  - |J| < 0 ⇒ 1.1%
- Inference time
  - 0.47s CPU
  - 0.09s GPU

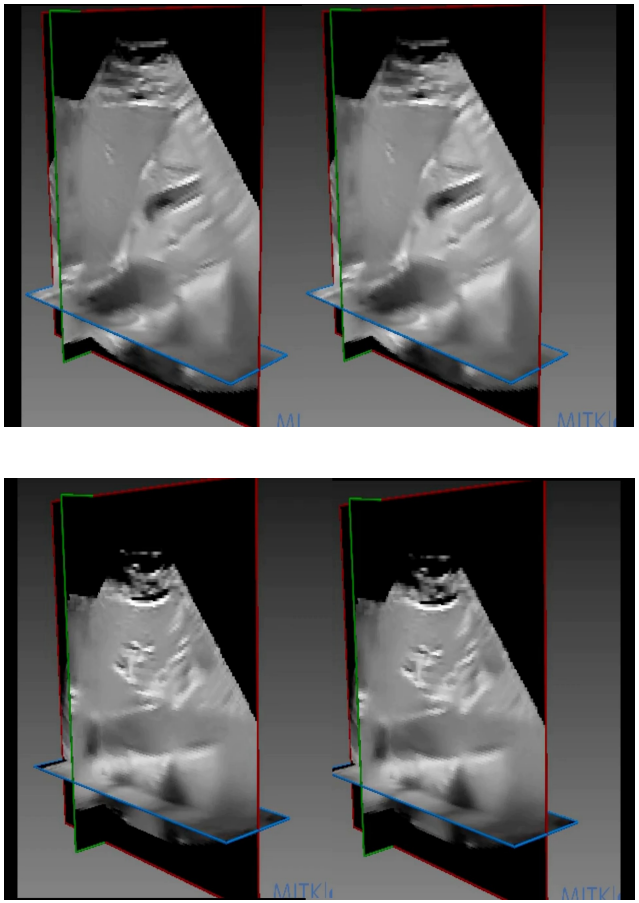
Image similarity metrics for different model configurations.
Values are mean ± std

Model	MSE	NCC	SSIM
Baseline	0.15 ± 0.04	0.42 ± 0.05	0.29 ± 0.05
Baseline + STN[1]	0.10 ± 0.06	0.57 ± 0.10	0.54 ± 0.11
Baseline + STN[1] + ϕ_{ref}	0.07 ± 0.04	0.62 ± 0.10	0.60 ± 0.10
Rigid	0.10 ± 0.06	0.61 ± 0.11	0.60 ± 0.12
Proposed (axi.)	0.07 ± 0.04	0.63 ± 0.10	0.61 ± 0.10
Proposed (sag.)	**0.06 ± 0.03**	**0.66 ± 0.09**	**0.65 ± 0.08**
P-value	3x10^{-3}	2x10^{-6}	9x10^{-7}

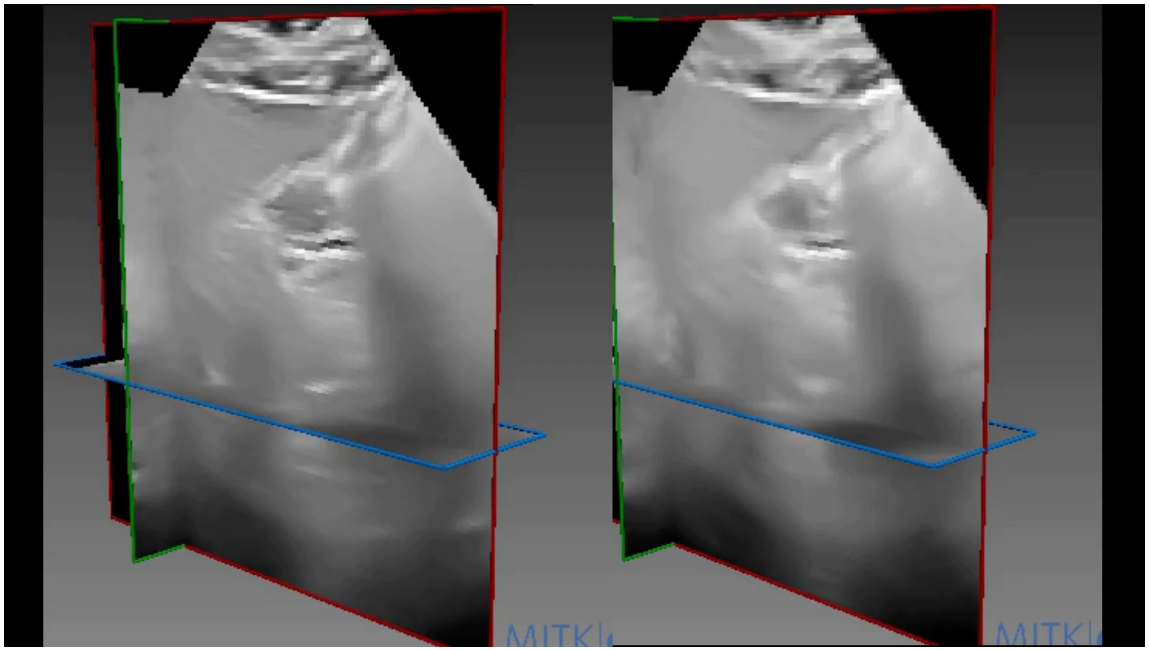
1. M. Jaderberg, K. Simonyan, A. Zisserman et al., "Spatial transformer networks," in Advances in neural information processing systems, 2015, p. 2017–2025.

Qualitative results

Ground truth Output

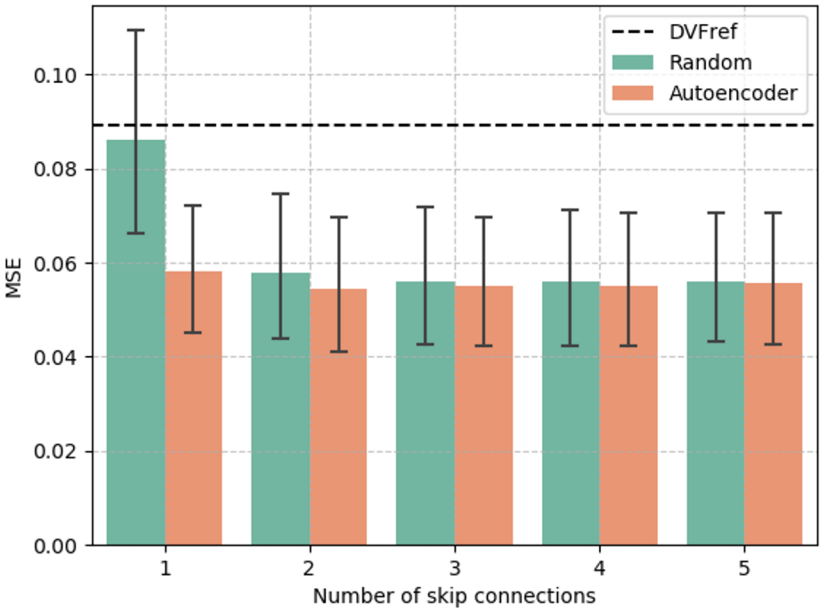


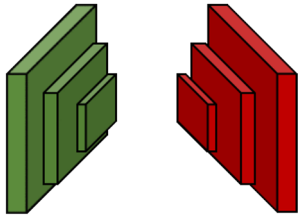
Ground truth Output



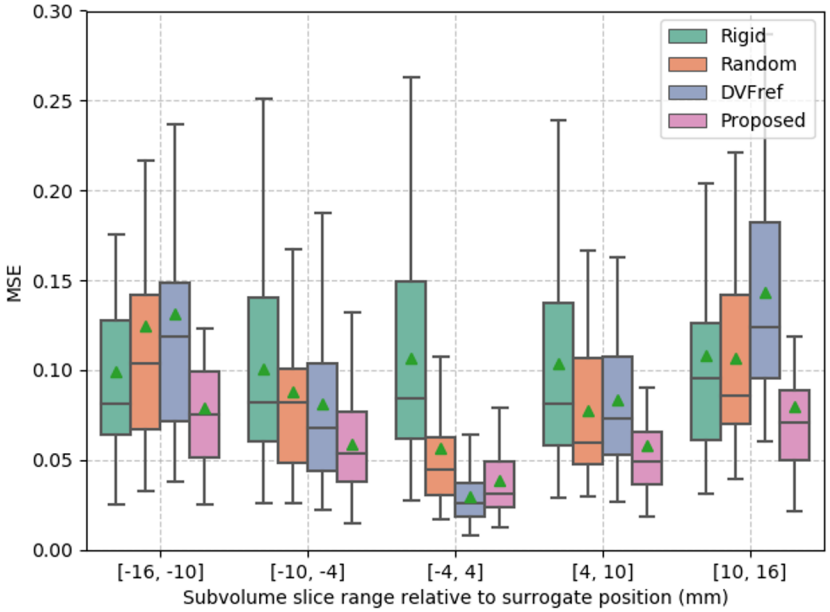
Proposed model: Auxiliary encoder

Motion autoencoder performance when varying the number of skip connections



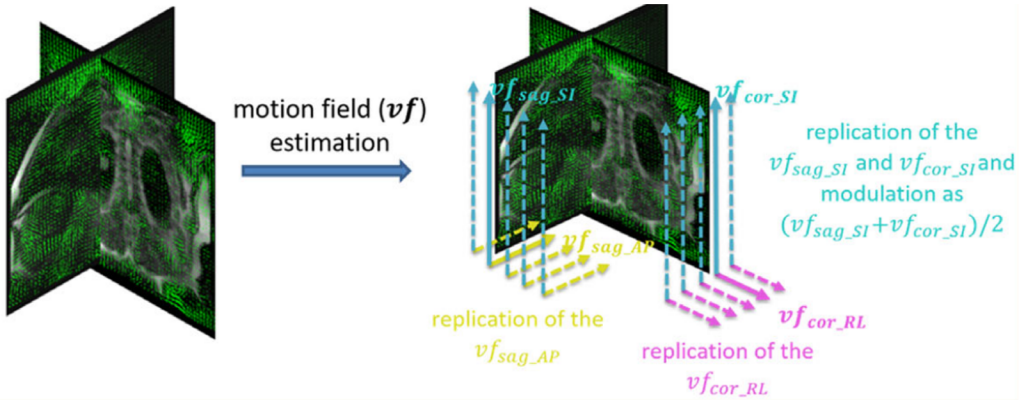


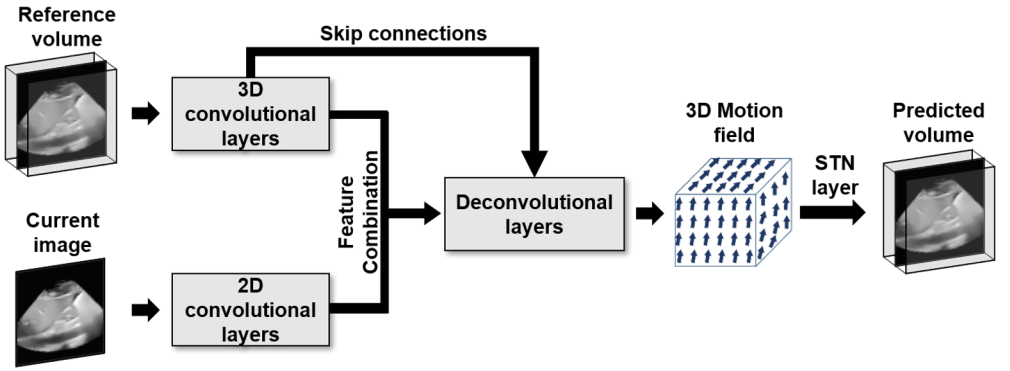
MSE distributions between ground-truth and predicted sub-volumes



Proposed model: comparative methods

- Motion extrapolation (ME)[1]
  - Compute displacement between current and reference orthogonal slices
  - Extrapolate motion field to volume
- Feature combination (FC)[2]
  - CNN based model from preliminary work
  - Combined with rigid module





1. C. Paganelli et al. “Feasibility study on 3d image reconstruction from 2d orthogonal cine-mri for mri-guided radiotherapy,”Journal of Medical Imaging and Radiation Oncology,2018.
2. T. Mezheritsky et al., “3d ultrasound generation from partial 2d observations using fully convolutional and spatial transformation networks,” in 2020 IEEE (ISBI), 2020

Comparative results

Image similarity and target tracking metrics between ground-truth and predicted volumes for different comparative methods. Values are mean±std.

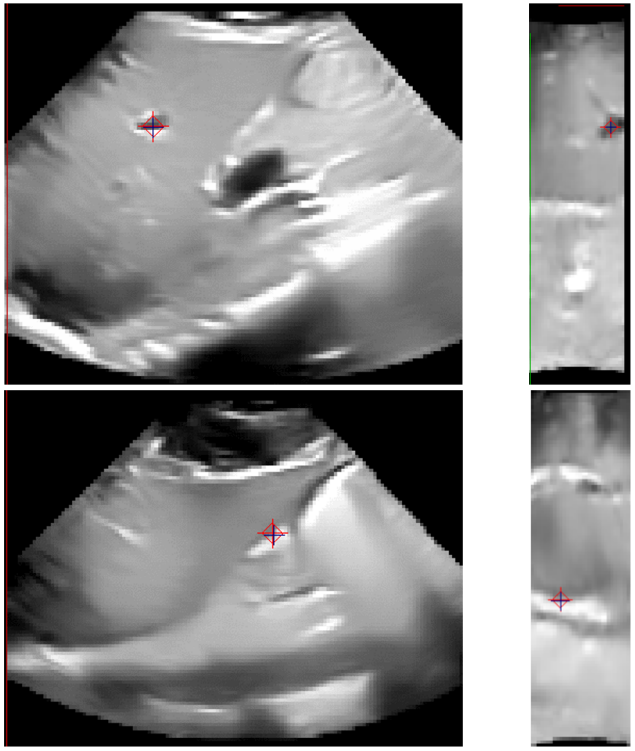
Image similarity				TRE Mean ± STD (mm)			
Model	**MSE**	**NCC**	**SSIM**	**Exhale**	**Mid-cycle**	**Inhale**	**Overall**
Reference	0.09 ± 0.06	0.59 ± 0.11	0.55 ± 0.13	---	9.8 ± 8.2	18.0 ± 13.4	10.7 ± 9.7
Rigid	0.10 ± 0.06	0.61 ± 0.11	0.60 ± 0.12	3.5 ± 1.3	3.9 ± 1.7	6.3 ± 4.3	4.6 ± 3.2
ME[1]	0.21 ± 0.08	0.59 ± 0.08	0.53 ± 0.10	**2.7 ± 1.4**	5.9 ± 2.8	10.9 ± 7.9	6.5 ± 6.4
FC[2]	0.09 ± 0.04	0.57 ± 0.09	0.54 ± 0.10	5.0 ± 3.3	7.9 ± 4.3	13.8 ± 10.7	8.9 ± 7.5
FC+rigid	0.08 ± 0.05	0.63 ± 0.10	**0.63 ± 0.10**	**3.1 ± 0.5**	4.5 ± 2.2	7.2 ± 4.4	4.9 ± 3.9
Proposed	**0.06 ± 0.03**	**0.66 ± 0.09**	**0.65 ± 0.08**	**2.8 ± 1.6**	**3.2 ± 0.8**	**4.5 ± 2.5**	**3.5 ± 2.4**
P-value	$2x10^{-3}$	$6x10^{-5}$	$4x10^{-3}$	$9x10^{-2}$	$2x10^{-5}$	$9x10^{-6}$	$2x10^{-6}$

1. C. Paganelli et al. “Feasibility study on 3d image reconstruction from 2d orthogonal cine-mri for mri-guided radiotherapy,”Journal of Medical Imaging and Radiation Oncology,2018.
2. T. Mezheritsky et al., “3d ultrasound generation from partial 2d observations using fully convolutional and spatial transformation networks,” in 2020 IEEE (ISBI), 2020

Comparative results

Sagittal

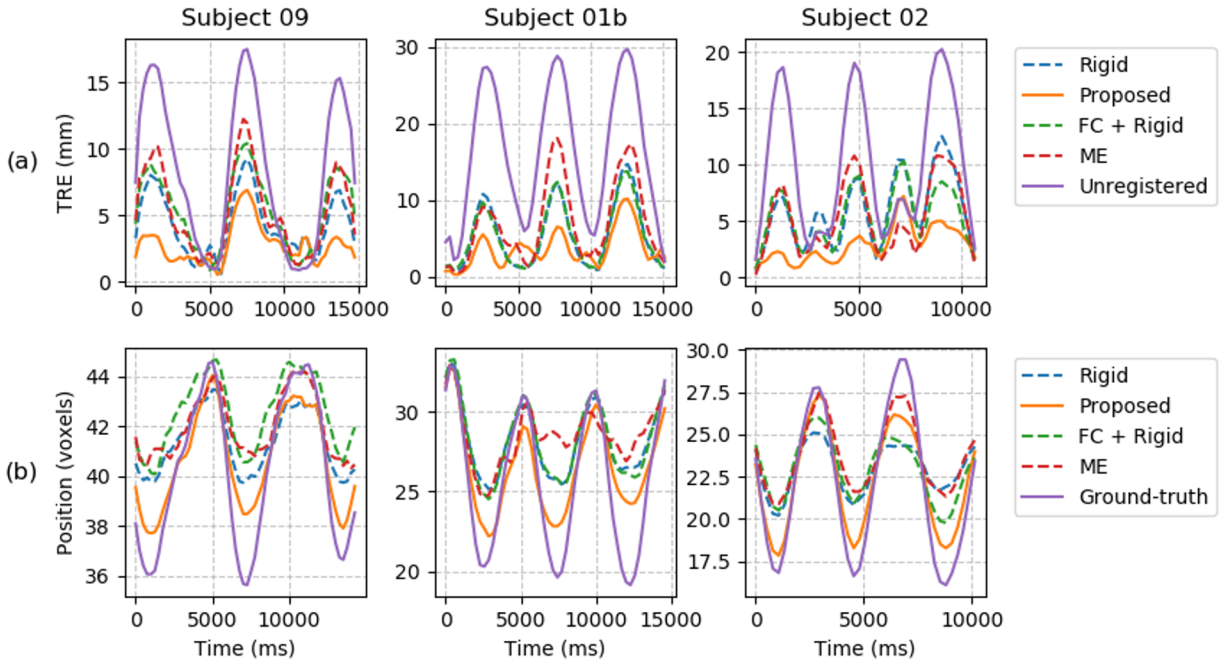
Coronal



\+ - Ground-truth

\+ - Prediction

Comparative plots showing (a) Evolution of TRE through time and (b) target trajectories for 3 cases over 3 respiratory cycles



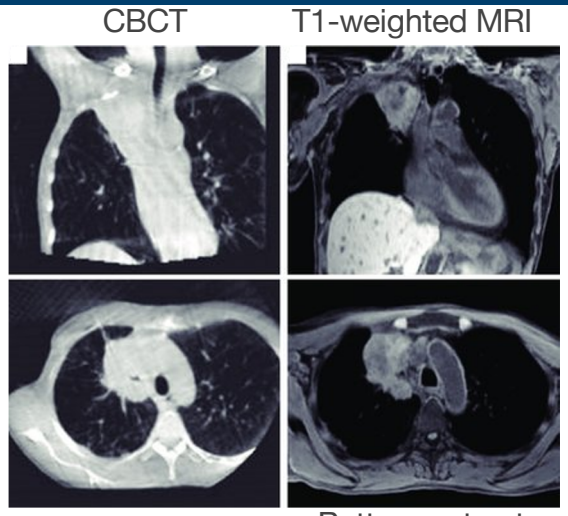
Summary

- Proposed convolutional AE based motion modelling framework
- Framework specific experiments
  - Patient specific features
  - Split motion into rigid and deformable components
- Comparative experiments
  - Achieved average tracking error of 3.5mm
  - Better than rigid alignment alone (4.6 mm)
  - Better than comparative approaches (ME: 6.5mm, FC+rigid: 4.9)
- Rigid alignment module can be used with other models to reduce tracking errors

MR-guided radiotherapy (IGRT)

- CBCT is the clinical standard, but increasing use of MRI
- Simultaneous beam delivery and organ monitoring with real-time 2D images → lacks volumetric information
- Image acquisition, target localization, beam modulation → system latency
- By the time a gating decision is made, the patient anatomy has already changed shape/position

4D motion modelling framework using deep learning to allow the forecasting of temporal volumes from surrogate 2D images

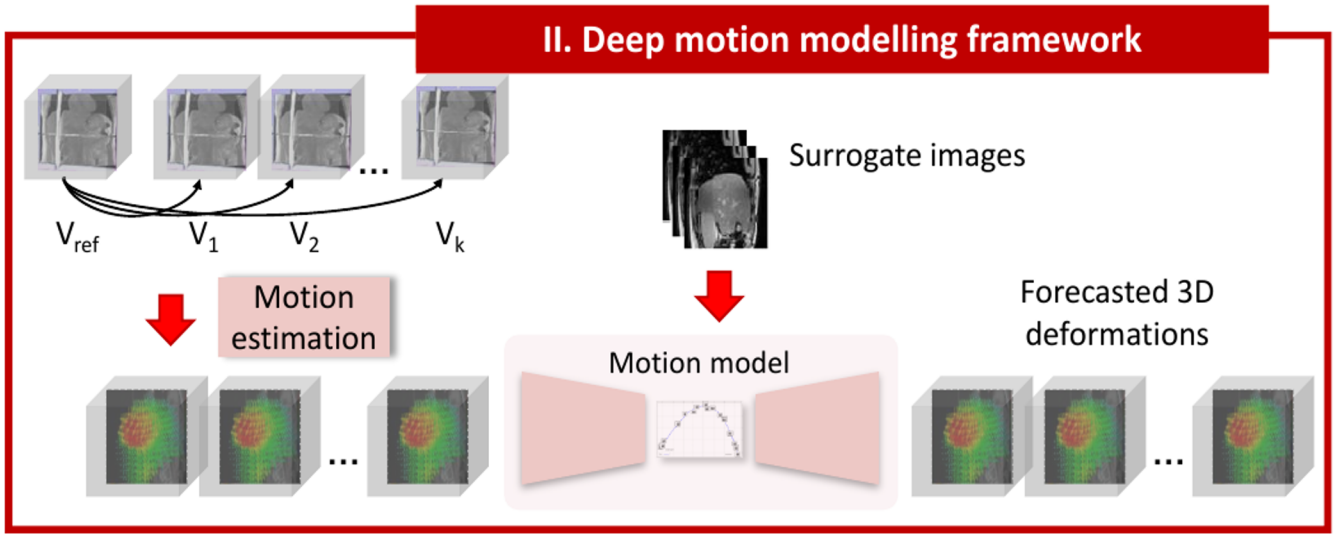


Better contrast
non-ionizing

Elekta-Unity MRI-linac

Viewray - MRIdian

Deep motion modelling framework



- Romaguera et al. International Journal of Computer Assisted Radiology and Surgery. June 2021.
- Romaguera et al. Int' Conf Medical Imaging and Computer Assisted Interventions 2021.
- Romaguera et al. Medical image analysis Dec 2021.

Probabilistic motion model

Model creation

Ground-truth motion data $\Phi = \{\phi_{t+1}, \phi_{t+2}, \ldots, \phi_{t+n}\}$

Sequential 2D images $I_{seq} = \{I_t, I_{t-1}, \ldots, I_{t-m}\}$

Reference volume V_{ref}

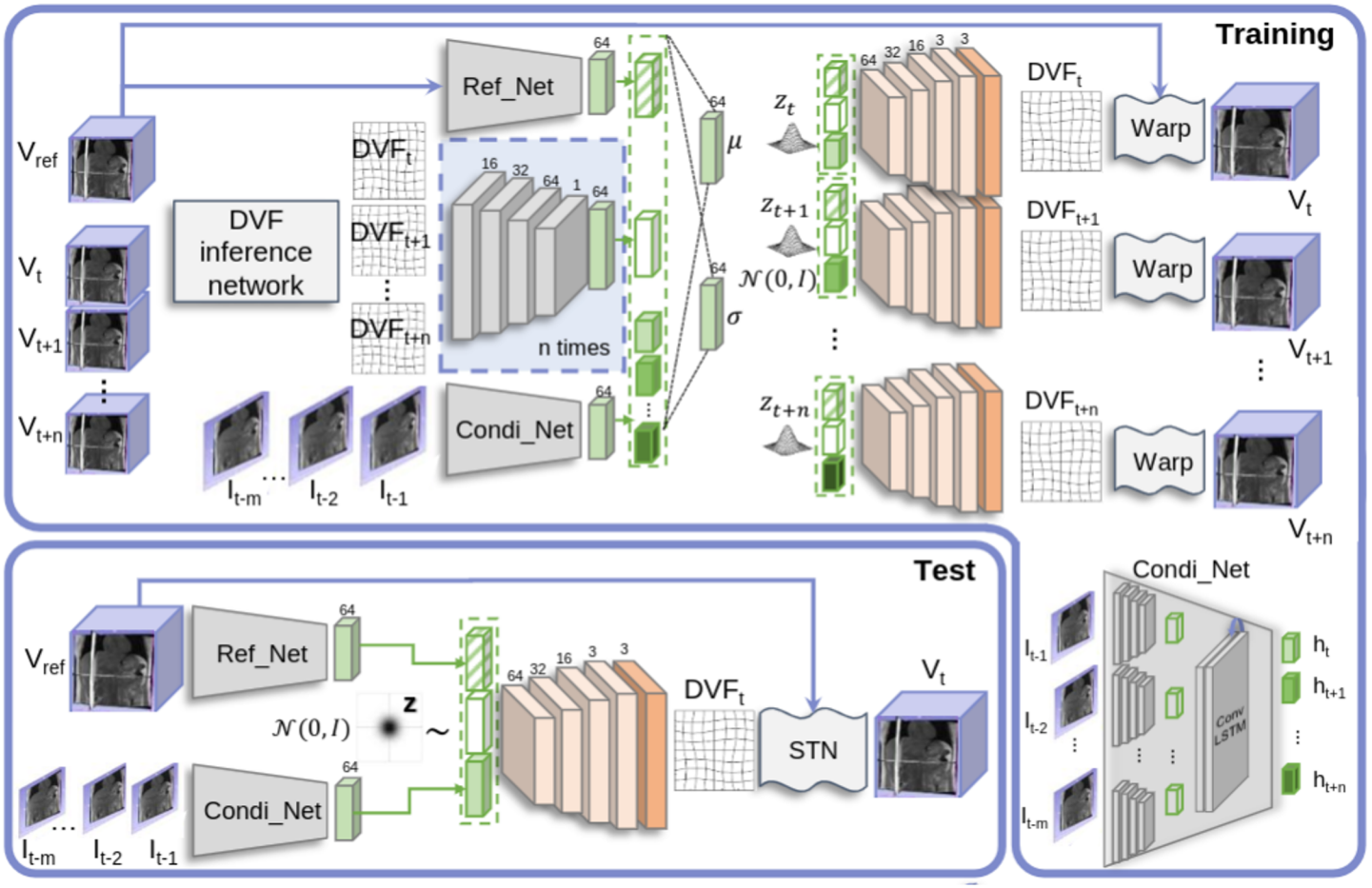
- Goal: to maximize the conditional probability distribution $p_\theta(\Phi|I_{seq}, V_{ref})$
- Law of total probability → relates the conditional and marginal probabilities

$$p_\theta(\Phi|I_{seq}, V_{ref}) = \int_z \underbrace{p_\theta(\Phi|z, I_{seq}, V_{ref})}_{\text{Likelihood}} p(z)\, dz$$

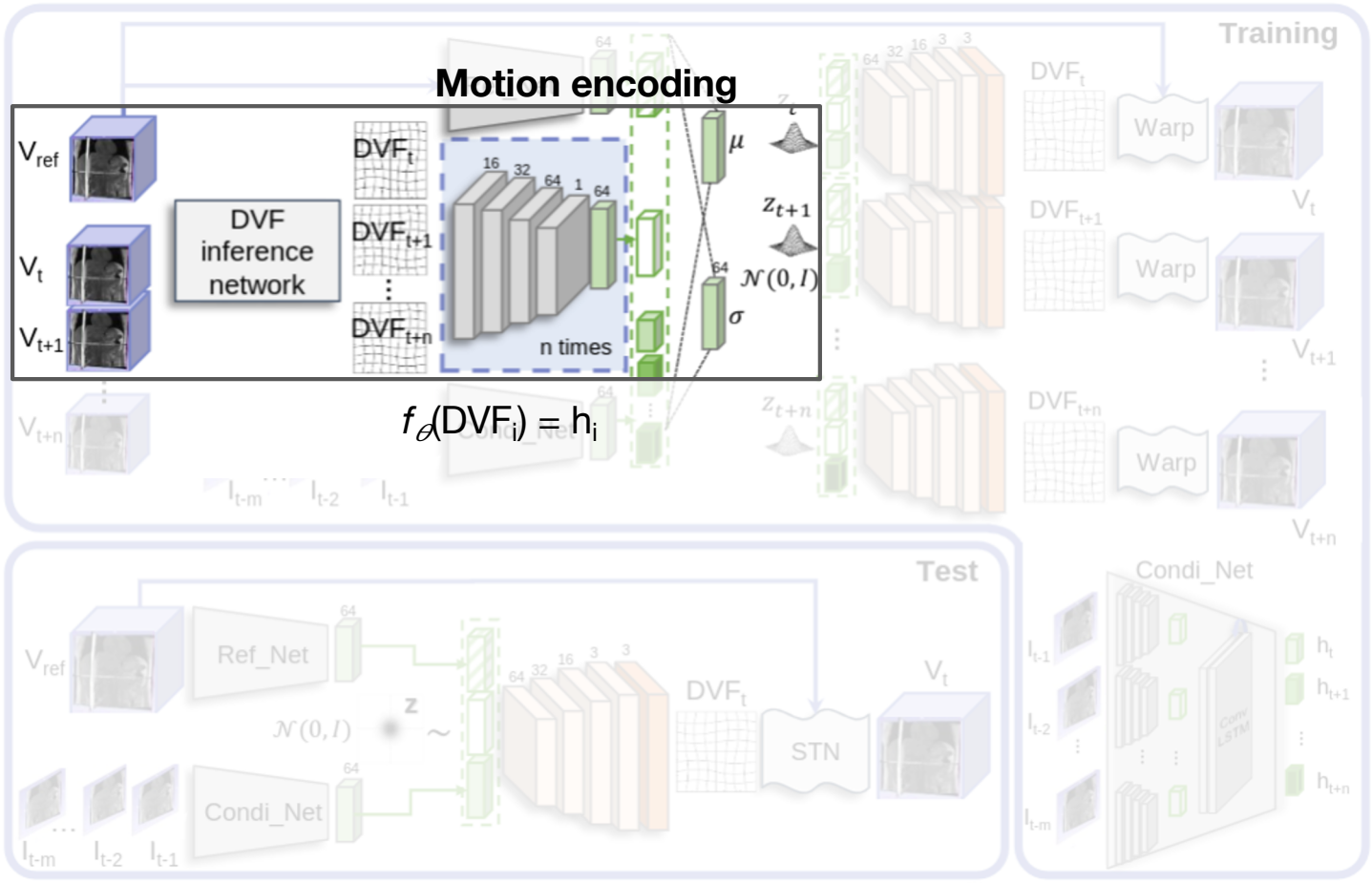
Large # of samplings!

- Only z values likely to produce $p_\theta(\Phi|I_{seq}, V_{ref})$ are considered, i.e. the posterior $p_\theta(z|\Phi, I_{seq}, V_{ref})$

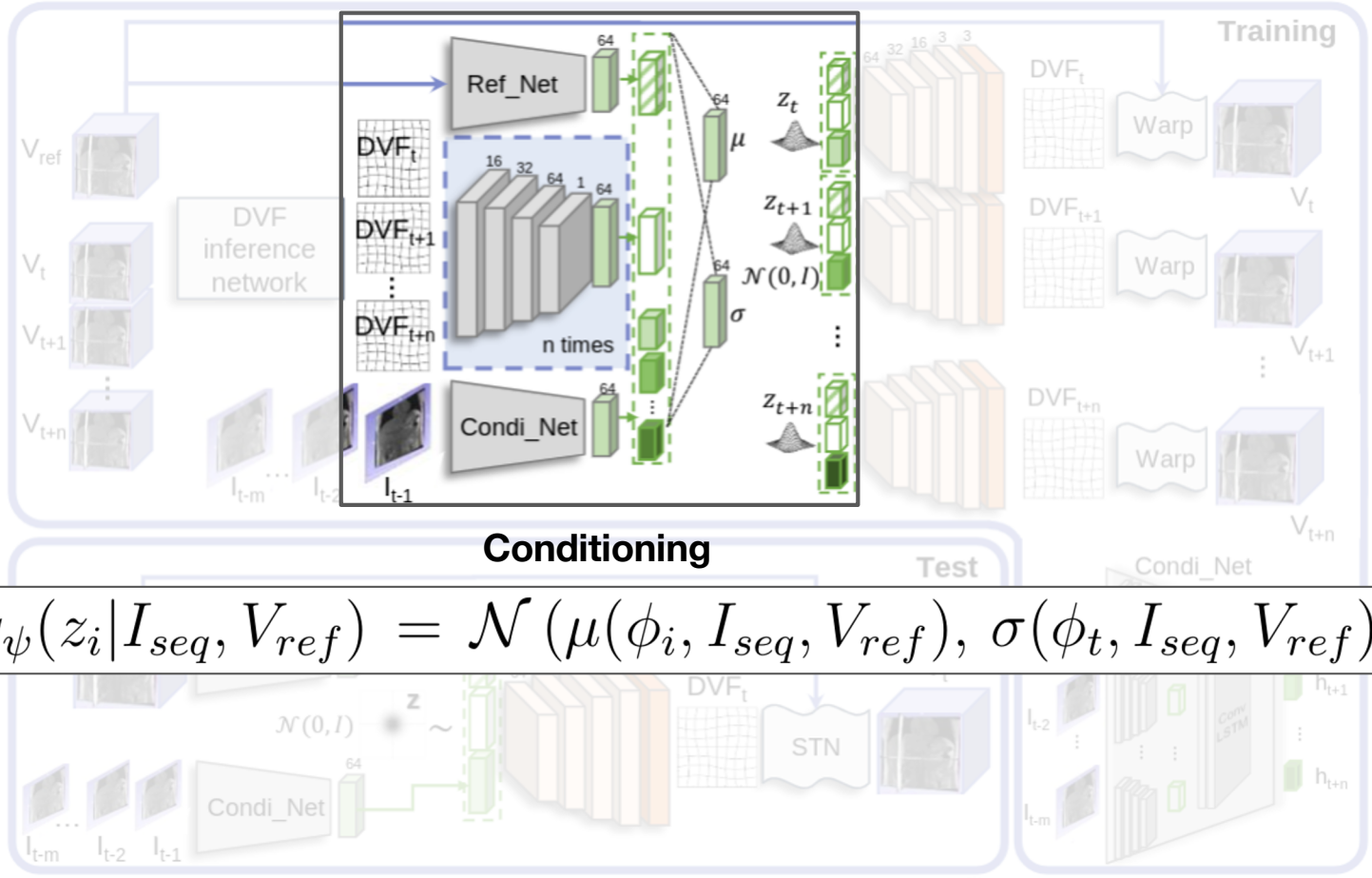
Probabilistic motion model



Probabilistic motion model



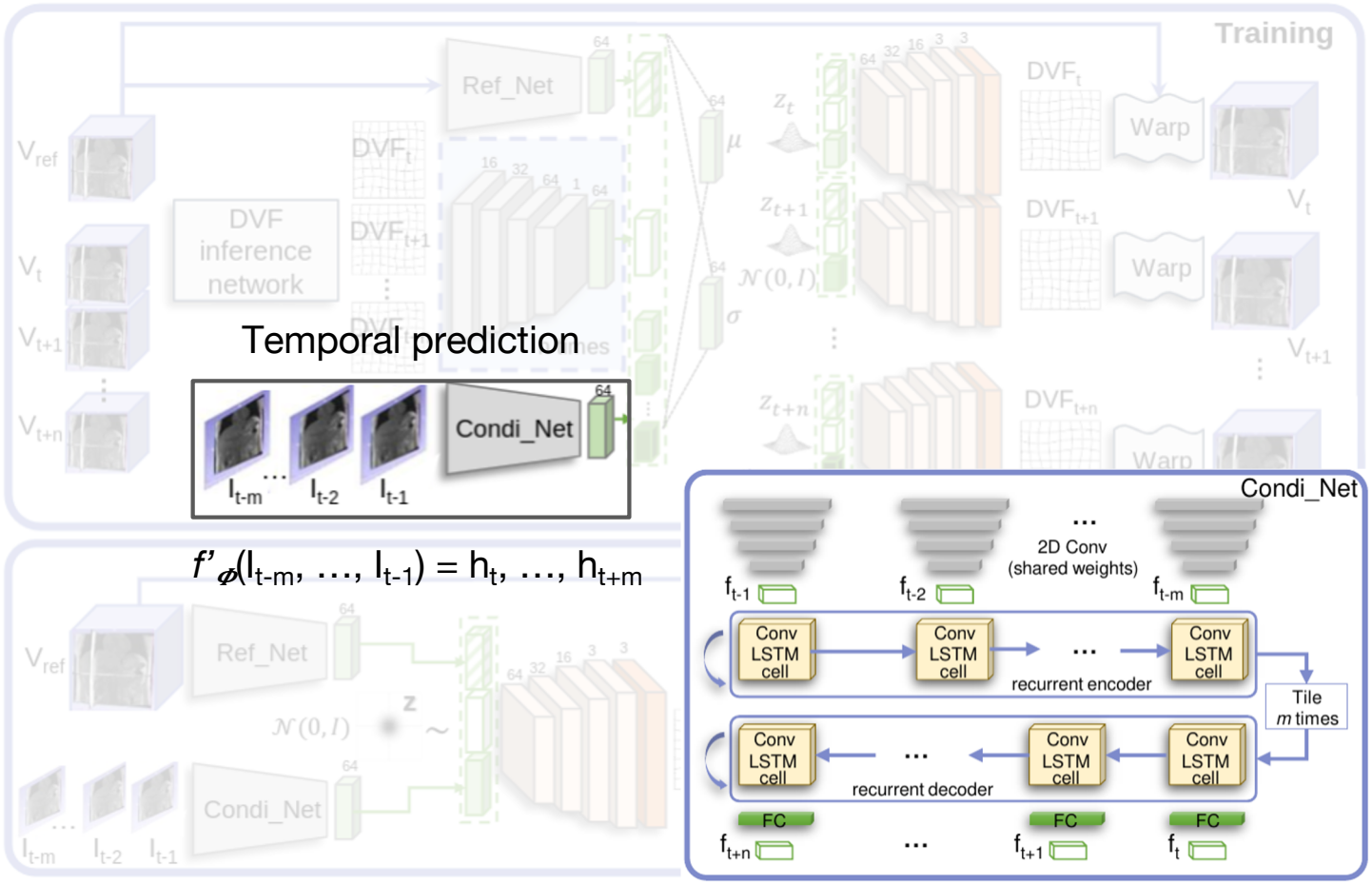
Probabilistic motion model



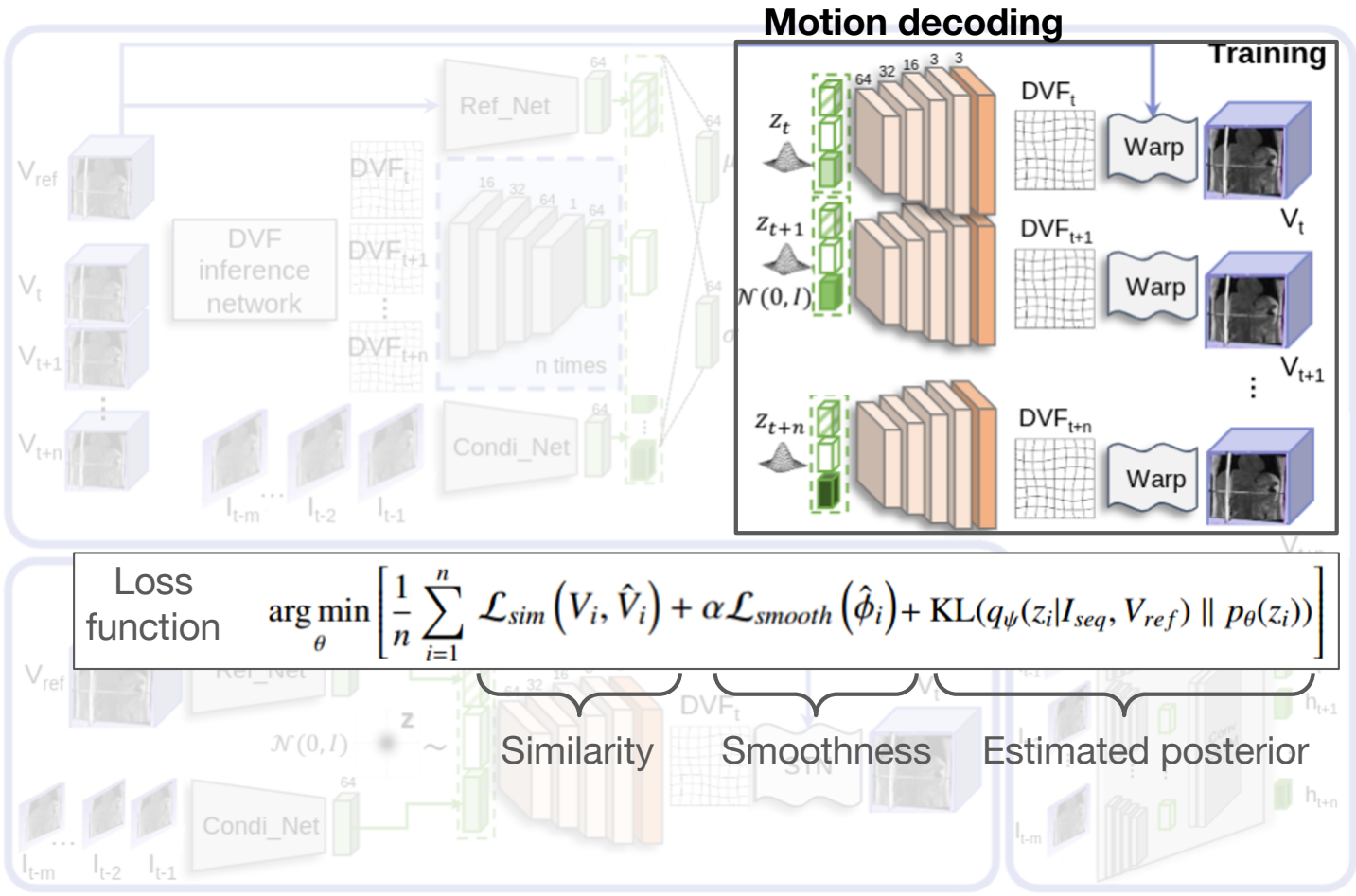
Conditioning

$$q_\psi(z_i|I_{seq}, V_{ref}) = \mathcal{N}\left(\mu(\phi_i, I_{seq}, V_{ref}),\ \sigma(\phi_t, I_{seq}, V_{ref})\right)$$

Probabilistic motion model



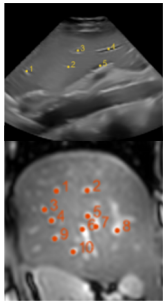
Probabilistic motion model



$$\arg\min_{\theta}\left[\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{sim}\left(V_i,\hat{V}_i\right)+\alpha\mathcal{L}_{smooth}\left(\hat{\phi}_i\right)+\mathrm{KL}(q_\psi(z_i|I_{seq},V_{ref})\parallel p_\theta(z_i))\right]$$

Geometrical errors

V-MRI



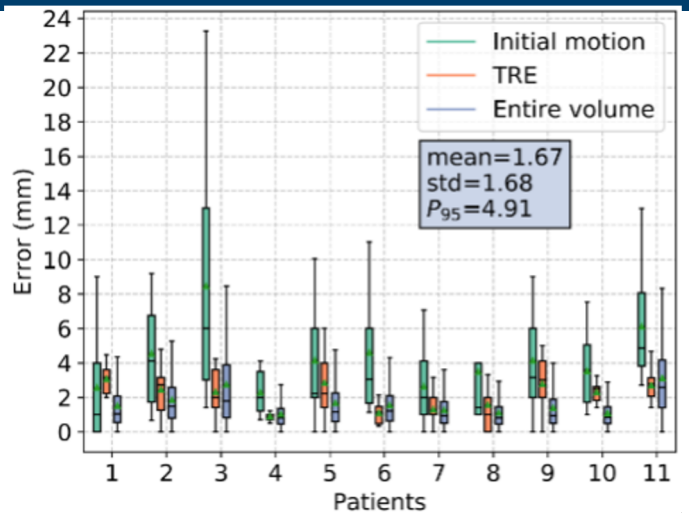
Target tracking errors (mm) for the V-MRI dataset. Values are mean ± std (95th percentile)

Model	Overall
Initial motion	6.7 ± 4.4 (13.7)
DN (Mezheritsky et al., 2020)	3.9 ± 2.7 (8.2)
ME (Paganelli et al., 2018)	2.4 ± 2.0 (5.8)
AE (Romaguera et al., 2021)	2.6 ± 2.1 (4.8)
Proposed (sag, P)	2.6 ± 2.2 (4.7)
Proposed (cor, P)	2.3 ± 1.9 (3.6)
PCA+AB (Pham et al., 2019)	1.8 ± 1.6 (3.6)
Proposed (cor, SS)	**1.4 ± 1.1 (3.3)**

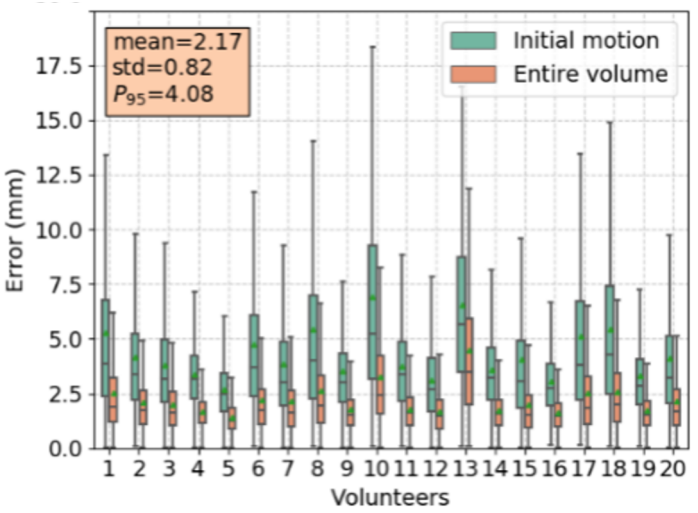
Geometrical errors (mm) for the V-MRI dataset obtained with different predictive mechanisms. Values are mean ± std (P_{95})

Predictor	Δt = 450 ms	Δt = 900 ms	Δt = 1350 ms
ConvGRU	1.6 ± 0.9 (3.7)	1.7 ± 1.1 (3.9)	1.3 ± 1.0 (3.2)
3D Conv	1.4 ± 1.0 (3.2)	1.6 ± 1.2 (4.2)	1.3 ± 0.9 (3.3)
ConvLSTM	**1.2 ± 0.6 (2.6)**	**1.4 ± 0.9 (3.3)**	**1.3 ± 0.9 (3.1)**

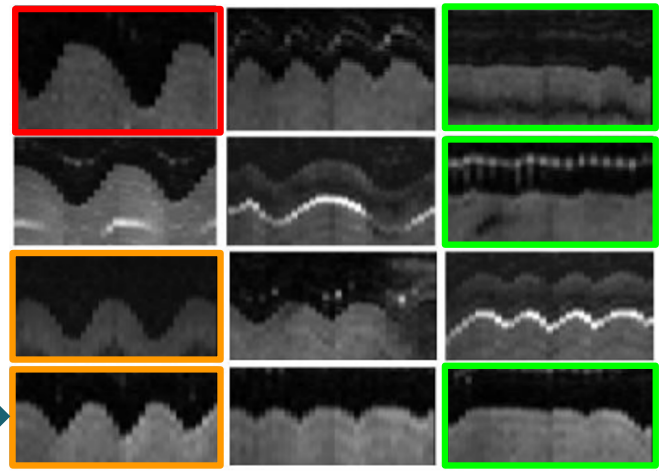
P-MRI



US



Inter-cycle variability

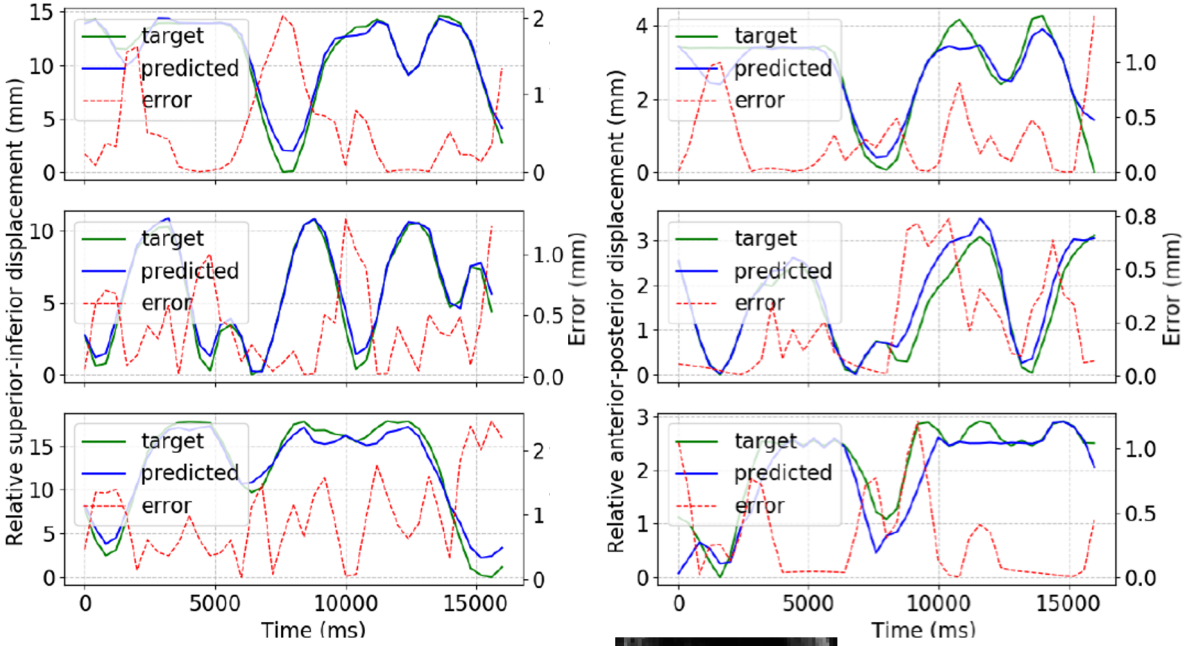


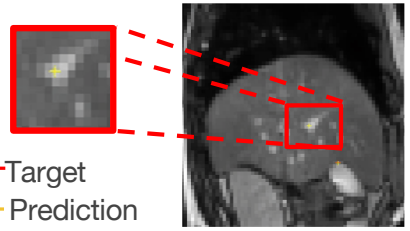
Different amplitudes and frequencies

Shallow Regular Deep

Same subject

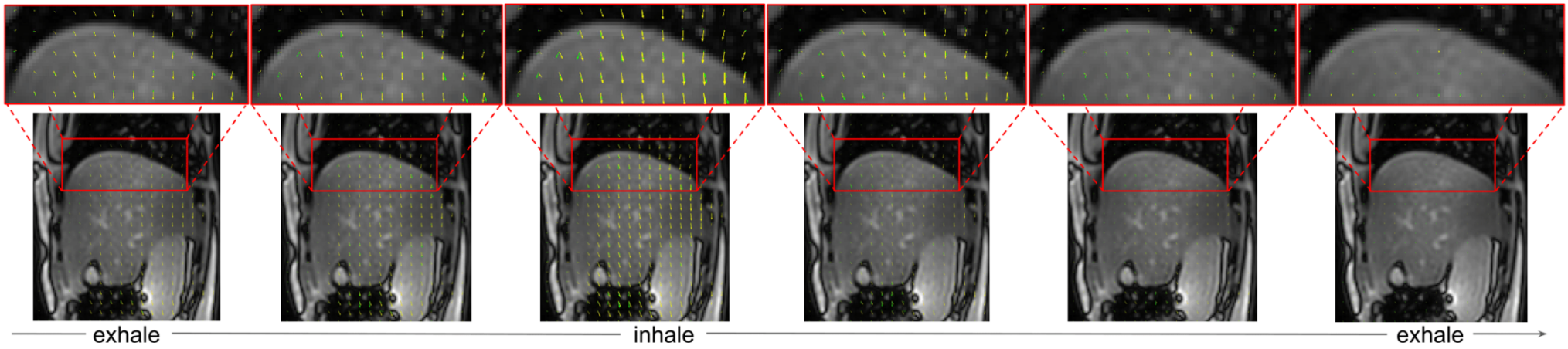
Vessel trajectories observed in 3 subjects with irregular breathing



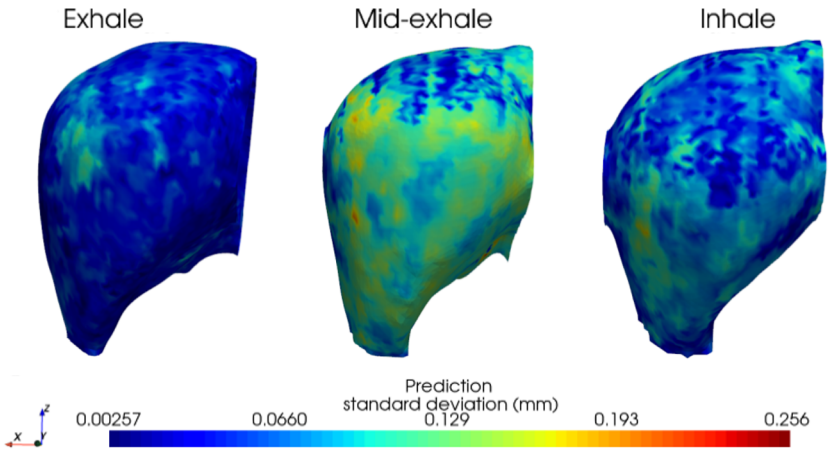


Target
Prediction

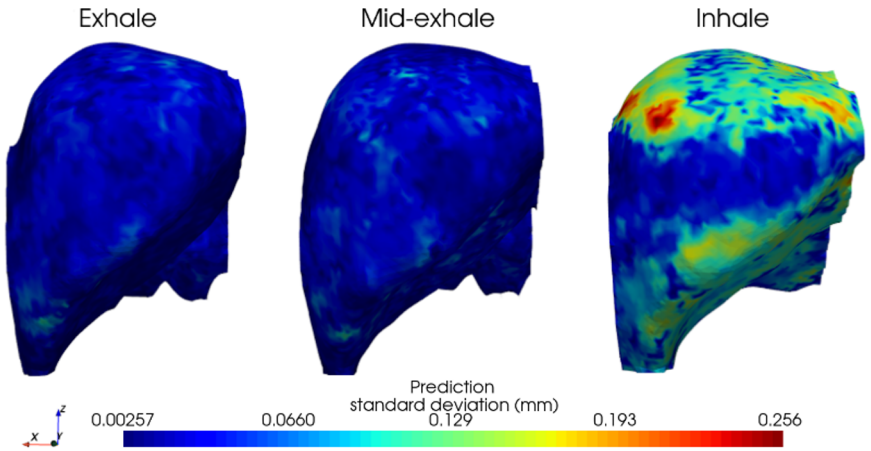
Sampling the latent space



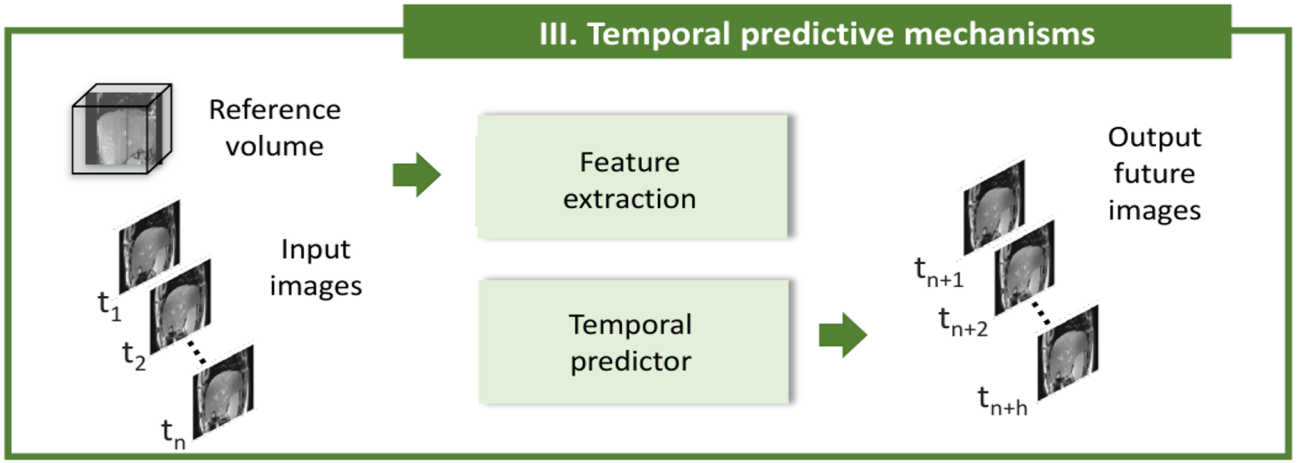
Subject 1



Subject 2



Temporal predictive mechanisms



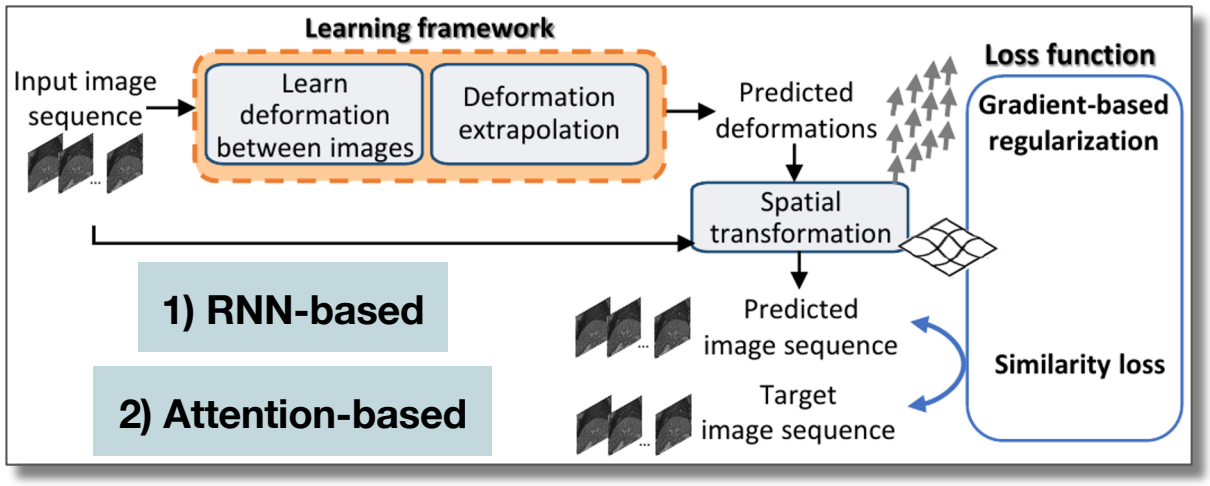
- Romaguera et al.. Medical image analysis, Dec. 2020
- Romaguera et al.. IEEE Trans. Medical Imaging. 2022.

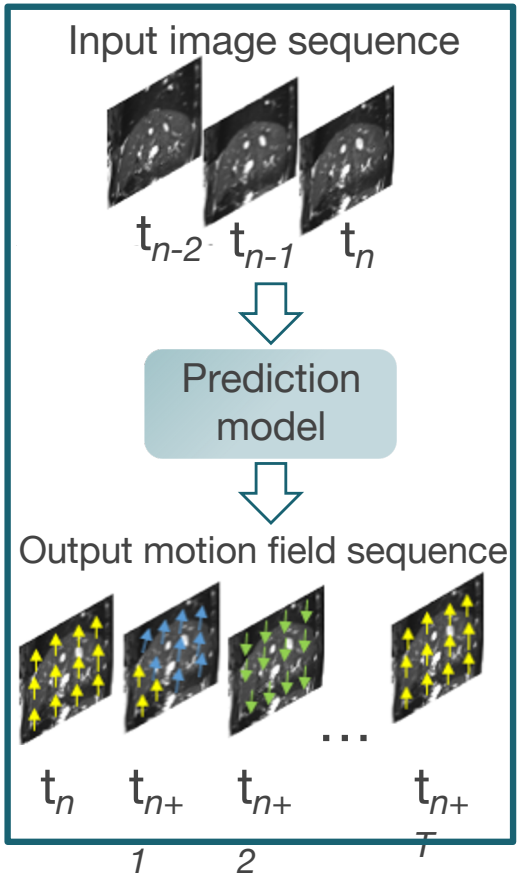
Temporal prediction

The goal is to predict a motion field sequence $\mathbf{Y_{out}}$ in *n* future times:

$$\mathbf{Y_{out}} = [\Phi_t, \Phi_{t+1}, \ldots, \Phi_{t+n}]$$

given an input image sequence $\mathbf{X_{in}} = [I_{t-m}, \ldots, I_{t-1}, I_t]$ in *m* observed times





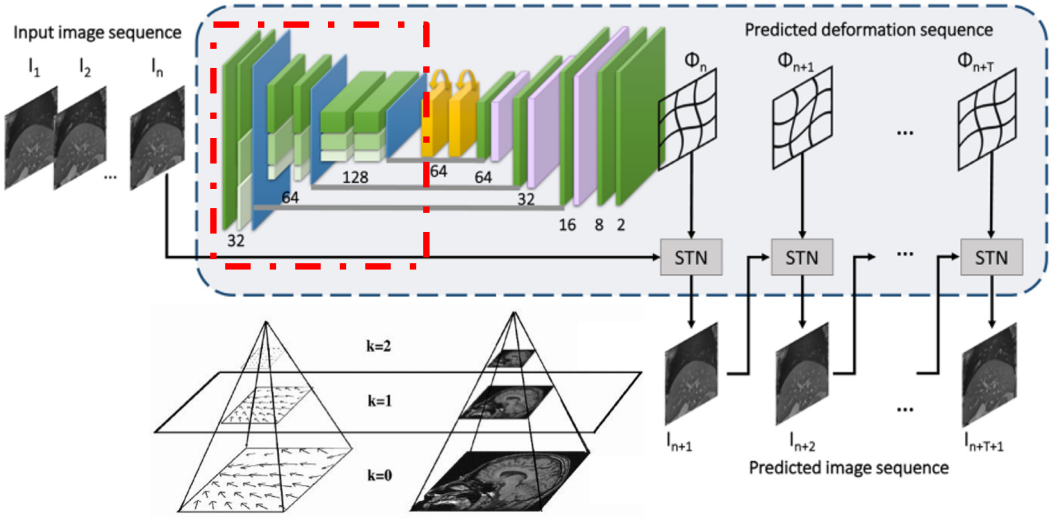
RNN-based proposed model

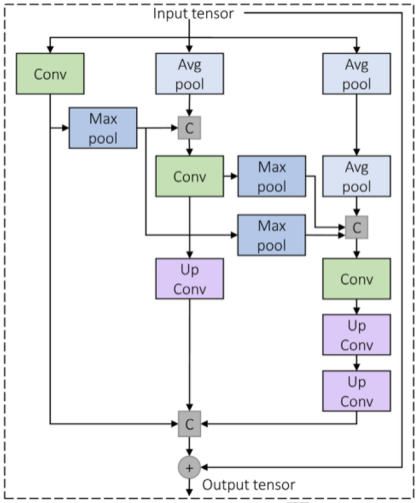
- Pixel intensities → Motion fields
- Pyramidal feature extraction
- Spatio-temporal prediction (ConvLSTM)
- Spatial transformations to implicitly regress future images

Task: To maximize $\mathrm{P}(\mathbf{Y}|\mathbf{X})$

$$\mathbf{X} = \left\langle I_1^i, I_2^i, \ldots, I_n^i \right\rangle \qquad \mathbf{Y} = \left\langle I_{n+1}^o, I_{n+2}^o, \ldots, I_{n+T+1}^o \right\rangle$$

$$\mathcal{L}_{total} = \frac{1}{T} \sum_{k=1}^{T} \mathcal{L}_{sim}\left(I_k^t, I_k^p\right) + \lambda \mathcal{L}_{smooth}\left(\phi_k\right)$$





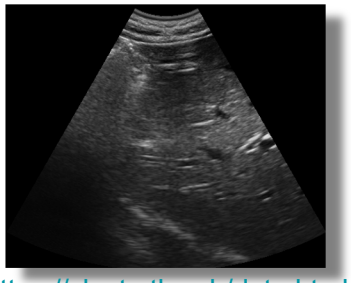
RNN-based | Experiments

- Comparison
  - Principal Component Analysis
  - Classification-based approach

- Datasets

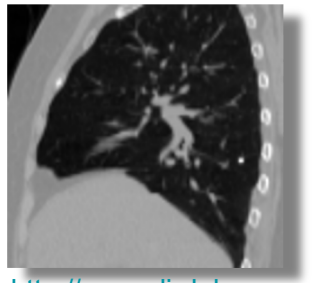
MRI (12 volunteers)
Free-breathing sagittal slices
Siemens Skyra 3T scanner
2D T2w true FISP sequence
Spatial res. = 1.7 × 1.7 mm^2
Temporal res. ~ 320 ms
Leave-one-out validation

US (63 volunteers)
Multicenter dataset
63 free-breathing sequences
Spatial res. = 0.5 × 0.5 mm^2
Temporal res. = 200 ms
Split: 62% - 13% - 25%
(train-val-test)



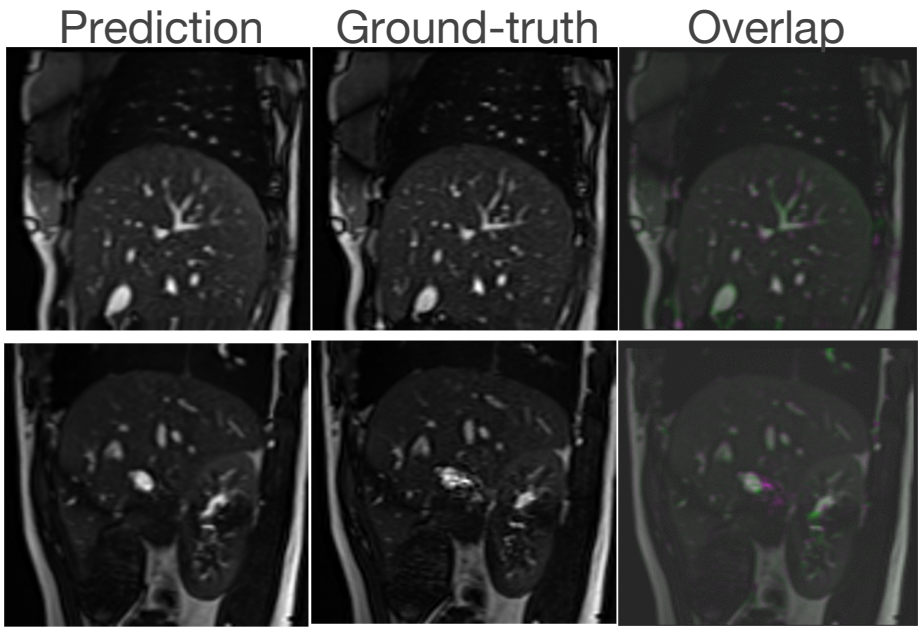
https://clust.ethz.ch/data.html

CT (10 cancer patients)
4D thoracic planning CT
Slices covering the right hemidiaphragm
Spatial res. = 1.0 × 1.0 mm^2
Temporal res. ~ 400 ms
Leave-one-out validation

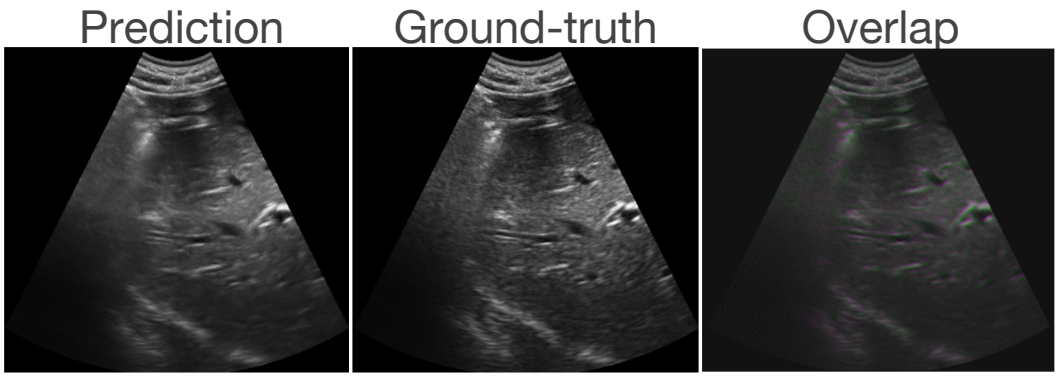


http://www.dir-lab.com

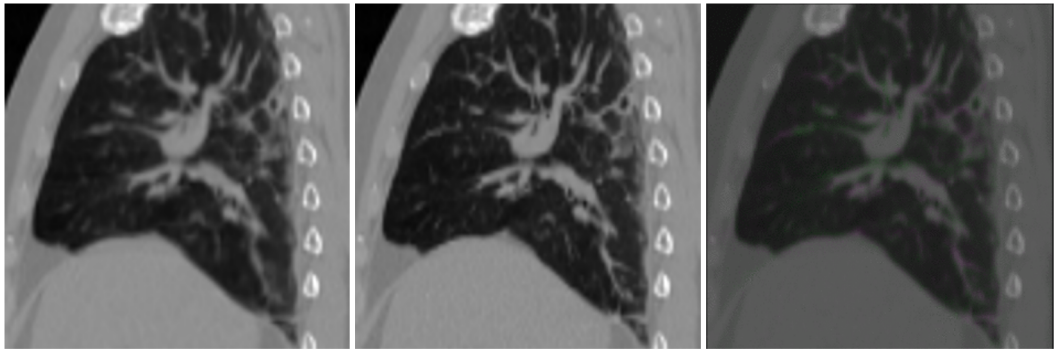
RNN-based | Qualitative results



horizon = 320 ms



horizon = 200 ms



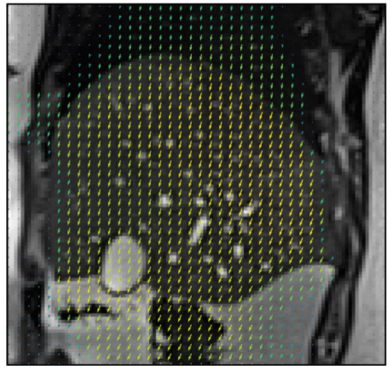
horizon = 400 ms

target prediction

RNN-based | Tracking results

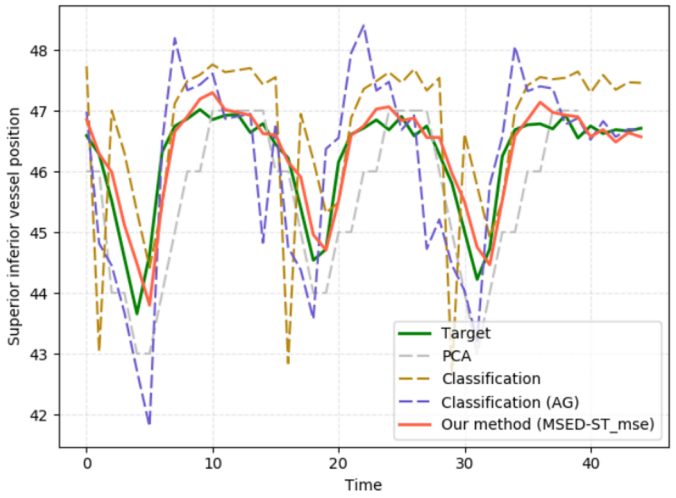
Vessel tracking error position (in mm) in the MRI dataset; median (IQR)

Model	t_p=1 (320 ms)	t_p=2 (640 ms)	t_p=3 (960 ms)	t_p=4 (1280 ms)	t_p=5 (1600 ms)
Classification	1.63 (2.29)	2.32 (2.58)	2.96 (2.88)	3.23 (2.65)	3.55 (2.52)
Classif. (AG)	1.55 (1.45)	2.33 (2.10)	2.77 (2.64)	3.16 (2.63)	3.20 (2.82)
PCA	1.36 (2.73)	1.85 (2.98)	2.37 (2.88)	2.72 (2.67)	3.01 (2.49)
ED-ST(ncc)	0.54 (0.66)	0.74 (0.98)	1.03 (1.26)	1.17 (1.42)	1.30 (1.66)
MSED-ST(ncc)	0.43 (0.54)	0.72 (0.91)	0.88 (1.22)	1.01 (1.36)	1.21 (1.57)
ED-ST(mse)	0.56 (0.65)	0.77 (0.96)	0.94 (1.15)	1.00 (1.15)	1.28 (1.61)
MSED-ST(mse)	**0.45 (0.55)**	**0.57 (0.75)**	**0.80 (0.99)**	**0.88 (1.25)**	**0.77 (1.36)**

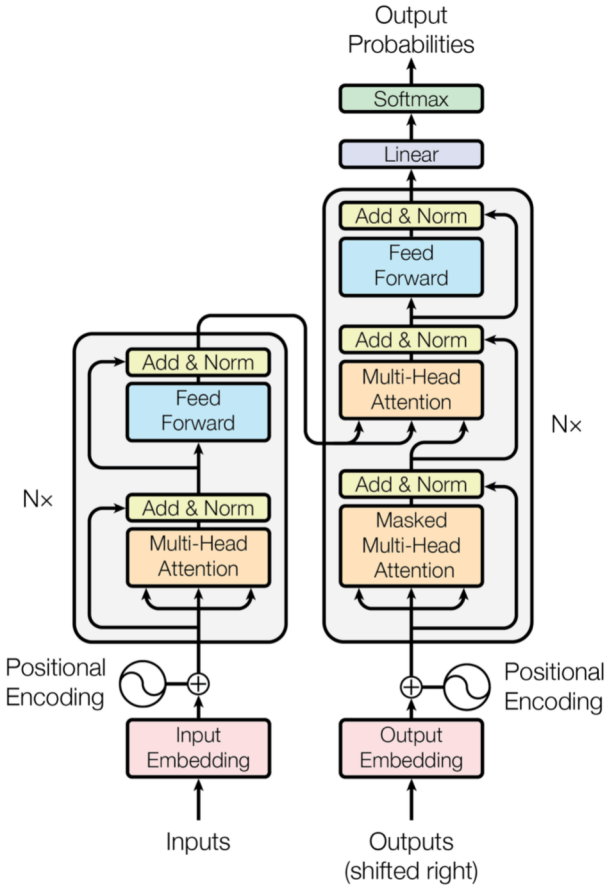


US dataset
0.45 (0.74) mm < error < 1.28 (1.31) mm

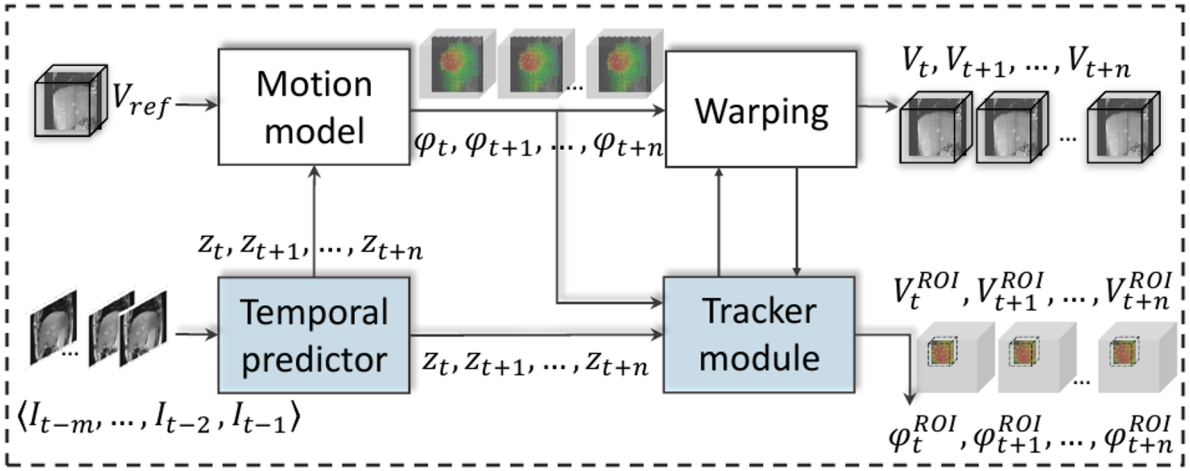
CT dataset
0.28 (0.58) mm < error < 0.42 (0.49) mm



Attention-based 4D deformation forecasting and tracking



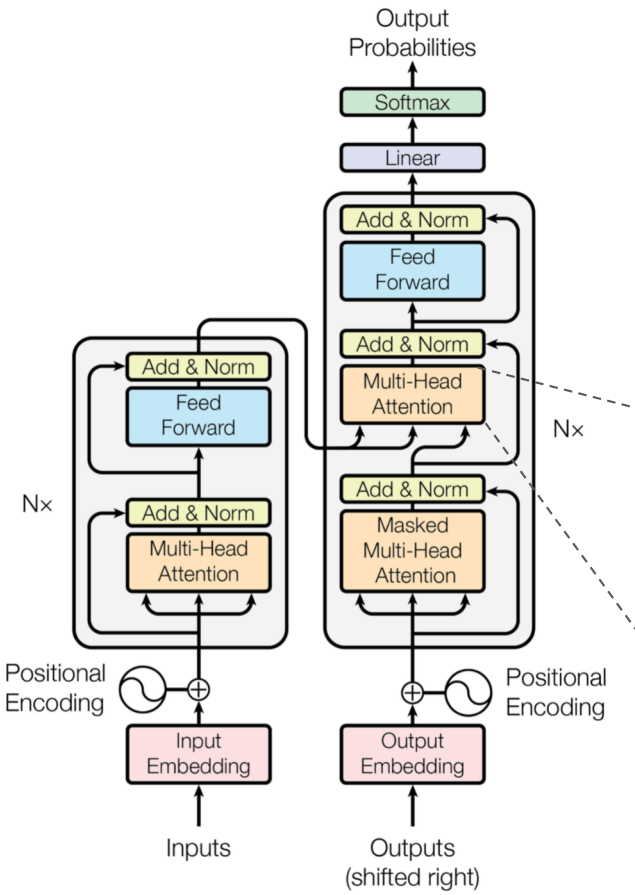
Source: Vaswani et al. (2017)



- Potential use of **attention models for temporal prediction** and its integration within the 4D motion model

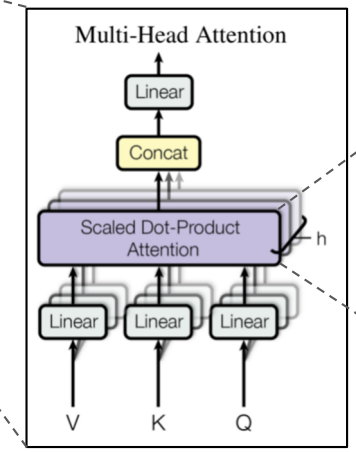
- 4D motion model for **local tracking**

Transformer

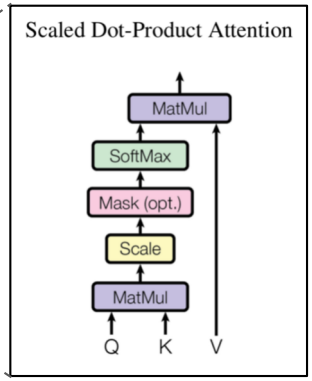


Source: Vaswani et al. (2017)

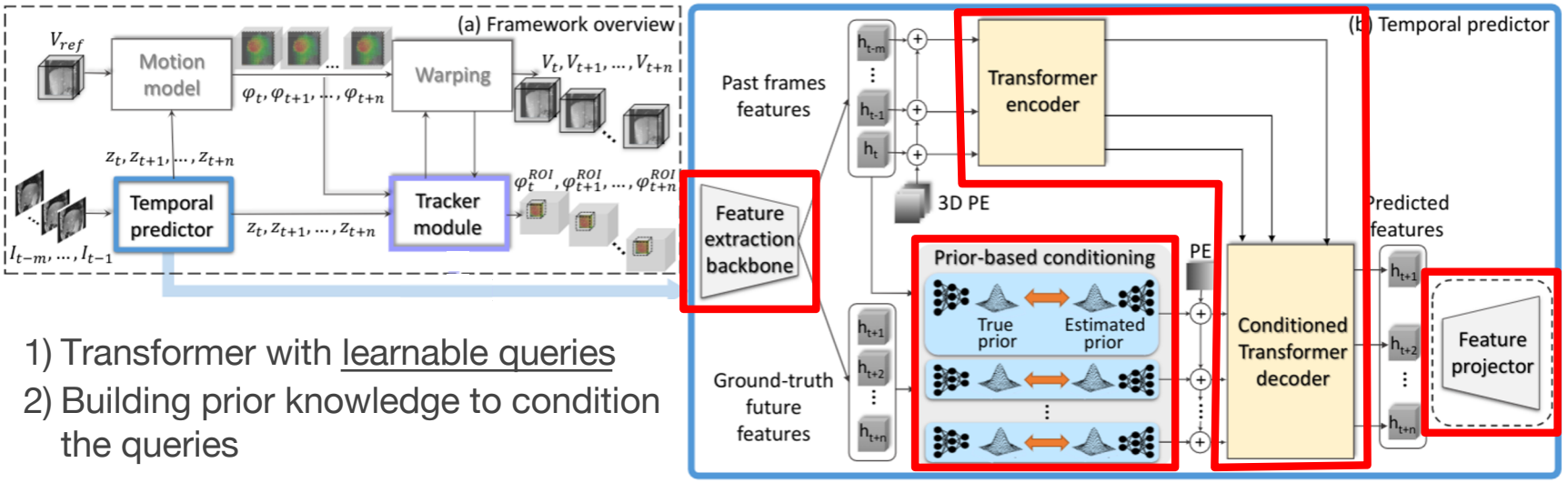
- No recurrency nor convolutions
- Attend to complete sequences → learns long-range dependencies and avoids loss of information
- Input vectors are linearly projected to *Keys*, *Queries* and *Values*
- Based on dot-product multi-head attention



$$A(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



Attention-based deformation forecasting

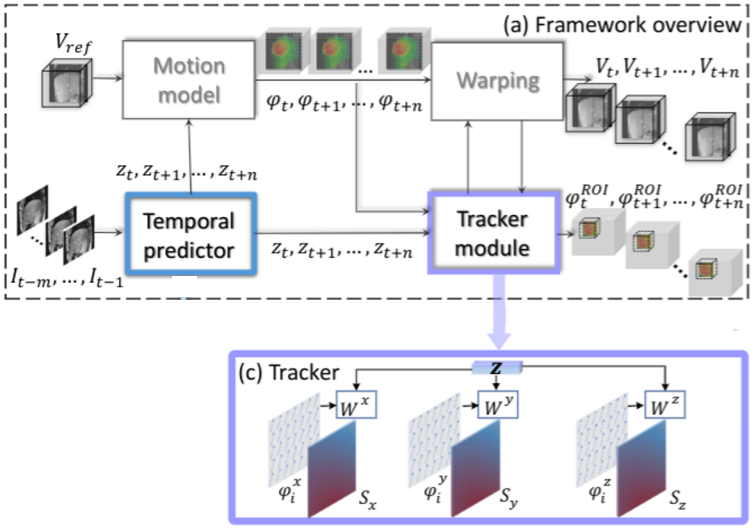


1) Transformer with learnable queries
2) Building prior knowledge to condition the queries

$$\arg\min_{\theta} \left[\sum_{i=1}^{n} \mathcal{L}_{rec} + \alpha \mathcal{L}_{KL} \right]$$

$$\mathcal{L}_{KL} = \sum_{i=0}^{n} KL\left[\overbrace{p_\theta(h_{t+i}|z_{t-m:t+i})}^{\text{True prior}} | \overbrace{r_\psi(h_{t+i}|z_{t-m:t-1})}^{\text{Estimated prior}} \right]$$

Tracker



Landmark selection $(x_{ref}, y_{ref}, z_{ref})$

Bounding box

$$(x_{ref} - \frac{\Delta_x}{2}, y_{ref} - \frac{\Delta_y}{2}, z_{ref} + -\frac{\Delta_z}{2})$$

Basic idea: Get refined values from ϕ_t conditioned on z_t

$$\hat{\phi}^{ROI} = Concat(S_x \times \phi_x^{ROI}, S_y \times \phi_y^{ROI}, S_z \times \phi_z^{ROI})$$

forecasted embedding → carries semantic information on the respiratory phase

Dense deformation predicted by the model

$$S_i = \sigma_2(\sigma_1(W_c c + W_\phi \phi_i^{ROI})W_s)$$

Deformation forecasting | Results

4D prediction

Geometrical errors (in mm) and image similarity measures obtained with the motion model and different temporal predictors. Values are mean ± std (95th percentile).

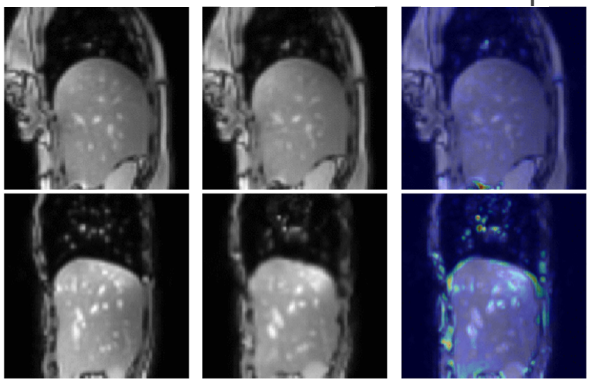
Predictive module	TRE (mm)	NCC	MSE	Time (ms)
ConvGRU [37]	$1.60 \pm 1.09\ (3.17)$	$0.71 \pm 0.14\ (0.89)$	$0.16 \pm 0.09\ (0.32)$	11.1 ± 1.2
ConvLSTM [36]	$1.37 \pm 0.92\ (2.60)$	$0.76 \pm 0.13\ (0.91)$	$0.13 \pm 0.09\ (0.22)$	11.9 ± 1.4
Transformer	$1.34 \pm 0.87\ (2.51)$	$0.76 \pm 0.13\ (0.91)$	$0.14 \pm 0.09\ (0.23)$	27.2 ± 1.2
Transformer+prior	$\mathbf{1.25 \pm 0.74 (2.13)}$	$\mathbf{0.81 \pm 0.11 (0.95)}$	$\mathbf{0.10 \pm 0.07 (0.18)}$	29.5 ± 1.6

2D + t prediction

Comparison with related approaches. Values are mean (std.)

Method	PSNR	SSIM	MSE
SVG [10]	22.6(3.1)	0.71(0.17)	0.09(0.06)
R-Unet [38]	25.4(5.5)	0.77(0.13)	0.08(0.04)
LMC [25]	23.5(2.3)	0.71(0.13)	0.10(0.04)
Proposed	**26.3(4.5)**	**0.78(0.11)**	**0.07(0.05)**

Ground-truth Prediction Diff map

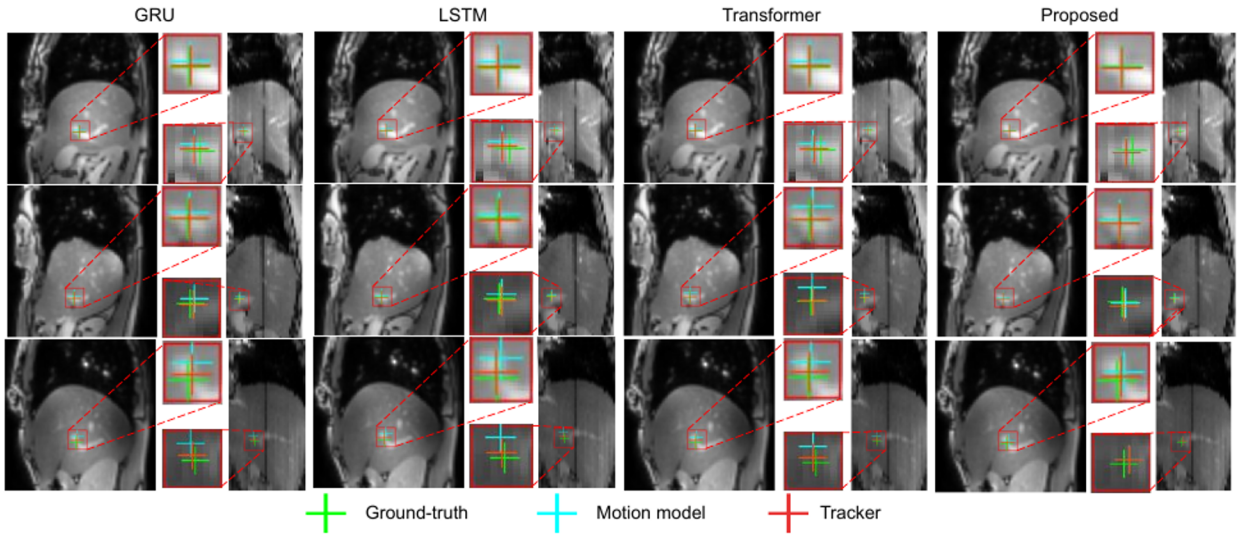


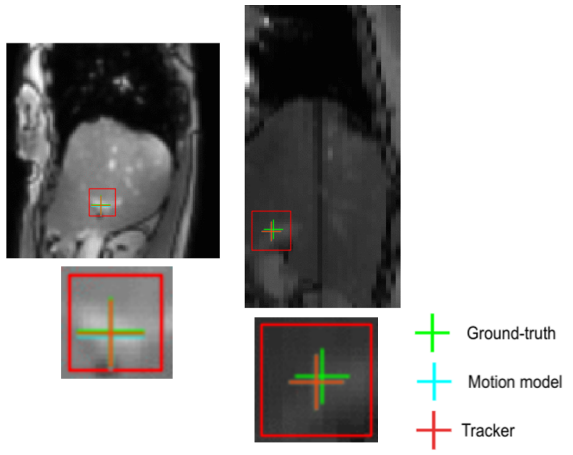
4D Tracking | Results

Tracker module reduces the error by **63%** compared with the 4D motion model

Tracking errors (in mm) using the motion model. Values are mean ± std [P_{90}]

Method	TRE (horizon = 450 ms)	TRE (horizon = 900 ms)	TRE (horizon = 1350 ms)
Initial motion	6.52 ± 3.41 [8.19]	6.35 ± 3.11 [8.0]	6.42 ± 3.40 [8.23]
MM + GRU	2.65 ± 1.93 [5.47]	2.72 ± 1.89 [5.45]	2.66 ± 1.86 [5.38]
MM + LSTM	2.68 ± 1.73 [4.90]	2.66 ± 1.70 [4.81]	2.59 ± 1.66 [4.68]
MM + Transf.	2.61 ± 1.58 [4.80]	2.56 ± 1.57 [4.73]	2.54 ± 1.55 [4.67]
MM + Transf. + prior	2.55 ± 2.11 [6.22]	2.56 ± 1.45 [4.70]	2.60 ± 2.08 [6.15]
MM + GRU + tracker	1.75 ± 1.19 [3.17]	1.78 ± 1.19 [3.19]	1.77 ± 1.17 [3.13]
MM + LSTM + tracker	1.66 ± 1.21 [3.25]	1.61 ± 1.16 [3.13]	1.57 ± 1.13 [3.03]
MM + Transf. + tracker	1.65 ± 1.17 [3.21]	1.63 ± 1.16 [3.16]	1.61 ± 1.15 [3.11]
MM + Transf. + prior + tracker	**1.56 ± 1.13 [3.09]**	**1.53 ± 1.11 [3.04]**	**1.52 ± 1.10 [2.98]**





Outlook

- Self-supervised models to forecast future deformations from 2D images

  1) RNN-based: leveraged spatial transformations to implicitly regress motion fields

  2) Attention-based: feasibility of Transformer to learn future representations as learnable queries conditioned on prior knowledge

- Clinically relevant accuracy and real-time application (inferences within a few ms)

- Deployed in unseen cases without prior steps

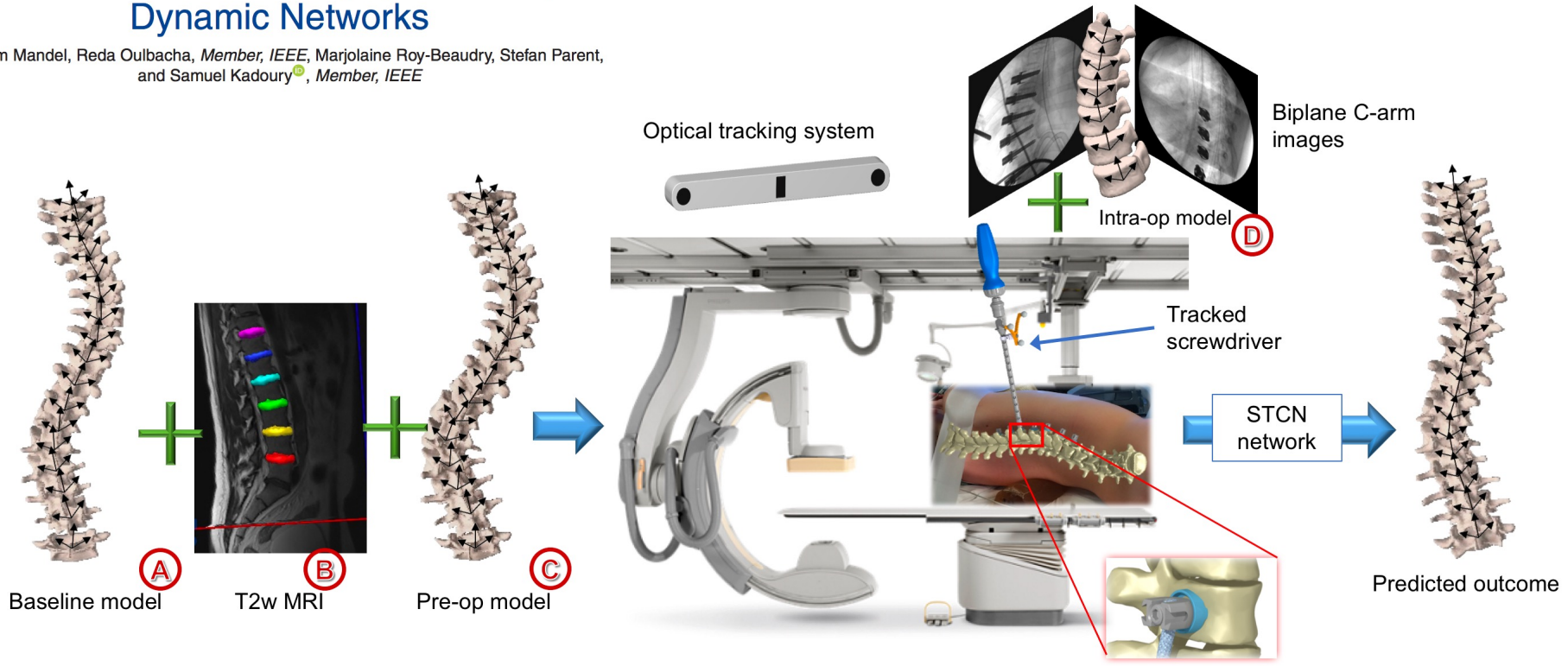
- Tracker: viable alternative for real-time image-guided interventions where pairs of up-to-date volumes are not available

Spatio-temporal network for spine surgery outcome

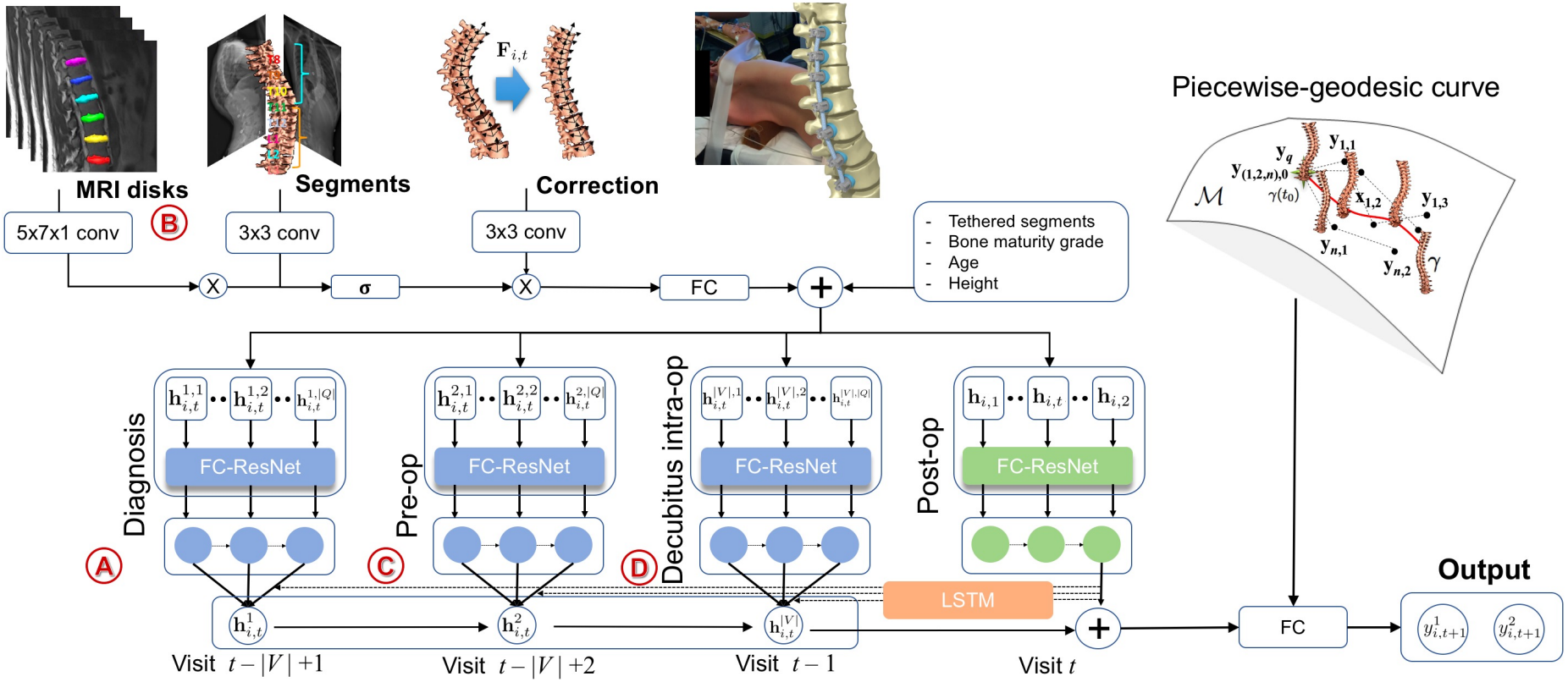
IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 40, NO. 2, FEBRUARY 2021 491

Image-Guided Tethering Spine Surgery With Outcome Prediction Using Spatio-Temporal Dynamic Networks

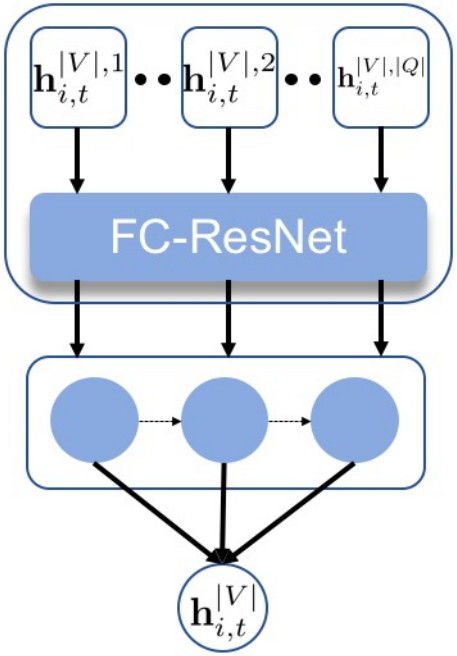
William Mandel, Reda Oulbacha, *Member, IEEE*, Marjolaine Roy-Beaudry, Stefan Parent, and Samuel Kadoury, *Member, IEEE*



Spatio-temporal shape prediction model



Geometric gating mechanism



- FC-ResNet is used to retrain the geometric features for each serial visit V
- Includes long-term and short-term geometric temporal changes:

$$\mathbf{h}_{i,t}^{v,q} = \text{FC-ResNet}([\mathbf{y}_{i,t}^{v,q}; \mathbf{e}_{i,t}^{v,q}], \mathbf{h}_{i,t}^{v,q-1})$$

- $\mathbf{h}_{i,t}^{v,q}$ representing the spine region i at time q from a previous visit v for the predicted time t ;
- $\mathbf{e}_{i,t}^{v,q}$ are the features related to the additional external parameters :
  - # of tethered segments;
  - Age;
  - Height;
  - Bone maturity status (Risser grade).

Time shifting mechanism

- The representation $\mathbf{h}_{i,t}^{v}$ for each previous visit is found by the sum of all visits within the time frame v close to the input case:

$$\mathbf{h}_{i,t}^{v} = \sum_{q \in Q} \alpha_{i,t}^{v,q} \mathbf{h}_{i,t}^{v,q}$$

Assigs the importance of the time gap q in a visit v

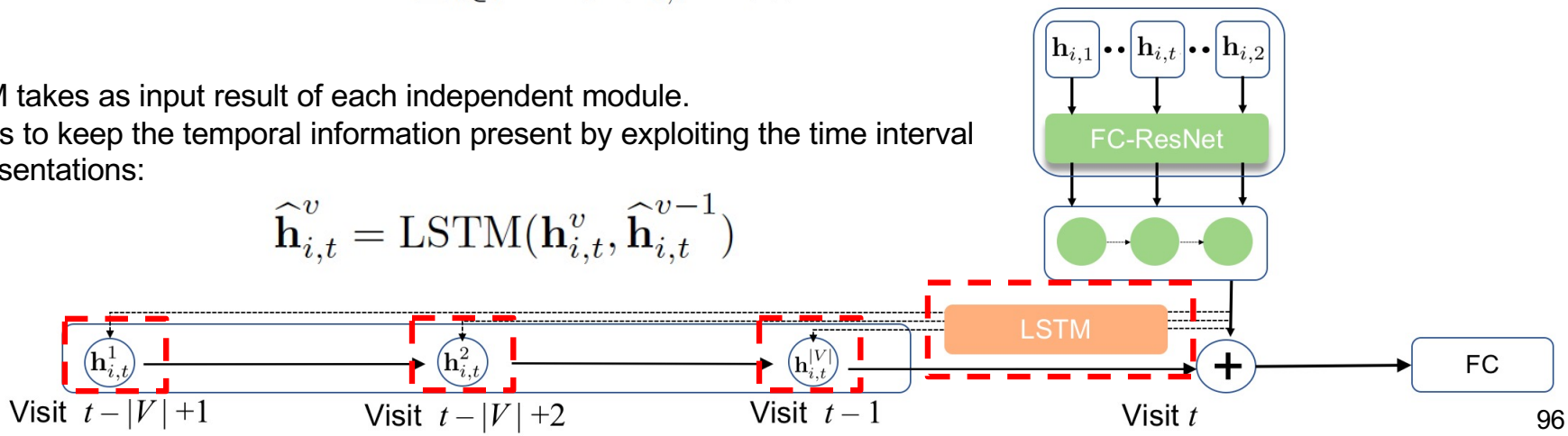
- Weight parameters compare the learned spatio-temporal representation:

$$\alpha_{i,t}^{v,q} = \frac{\exp(\Phi(\mathbf{h}_{i,t}^{v,q}, \mathbf{h}_{i,t}))}{\sum_{v \in V} \exp(\Phi(\mathbf{h}_{i,t}^{v,q}, \mathbf{h}_{i,t}))}$$

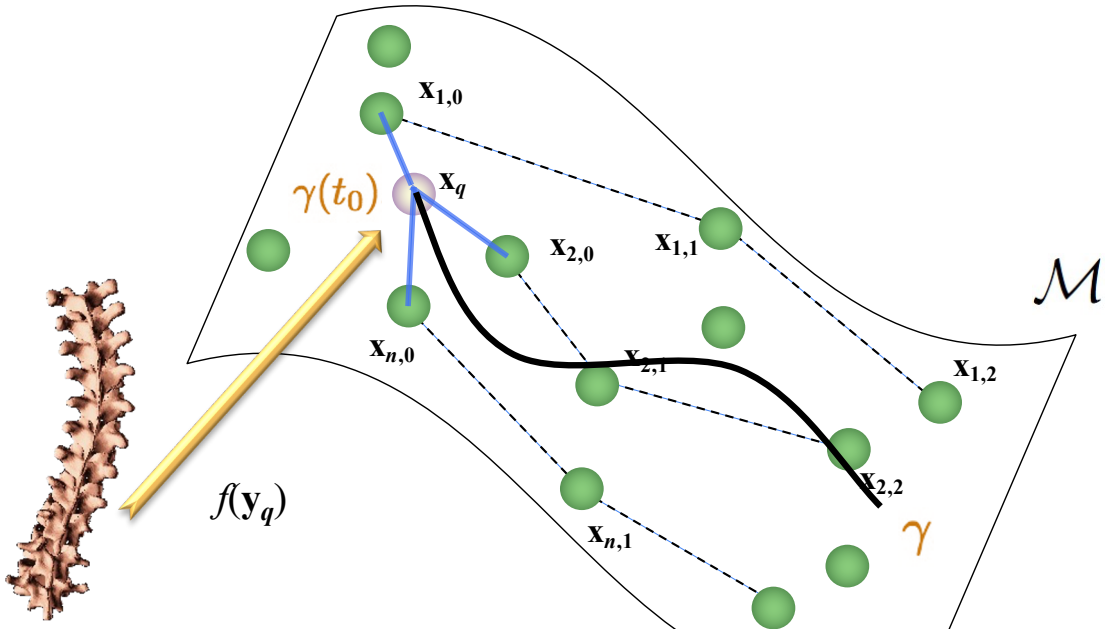
Φ measures the similarity in correction flows between pairs of spines.

- LSTM takes as input result of each independent module.
- Allows to keep the temporal information present by exploiting the time interval representations:

$$\widehat{\mathbf{h}}_{i,t}^{v} = \mathrm{LSTM}(\mathbf{h}_{i,t}^{v}, \widehat{\mathbf{h}}_{i,t}^{v-1})$$



Piecewise-geodesic group-average trajectory



Input 3D model

1. Projection of input onto $\mathcal{M}$ using kernel function f
2. Neighborhood selection K_d at t_0
3. Shifting of time measurements to common baseline
4. Regression of smooth piecewise-geodesic curve $\gamma : [t_1, t_N]$
5. Minimization of discretized energy term $E(\gamma)$ with quadratic optimization problem (Boumal *et al.*, IFAC, 2011):

$$E(\gamma) = \frac{1}{K_d}\sum_{i=1}^{K_d}\sum_{j=0}^{t_N} w_i \|\gamma(t_{i,j}) - (\mathbf{x}_{i,j} - (\mathbf{x}_{i,0} - \mathbf{x}_q))\|^2 + \frac{\lambda}{2}\sum_{i=1}^{K_d}\alpha_i\|v_i\|^2 + \frac{\mu}{2}\sum_{i=1}^{K_d}\beta_i\|a_i\|^2$$

Minimize geodesic dis… — Minimize velocity of $\dot{\gamma}(t_i)\frac{\mathrm{D}^2\gamma}{\mathrm{d}t^2}(t_i)$ — Acceleration penalty factor… — …exhibit discontinuities or rapid transi…

Training with piecewise-geodesic trajectory

1. Projection of $\mathbf{y}_{i,t}$ onto the piecewise-geodesic curve γ
2. The output predictions are given by:

$$[y^1_{i,t+1}, y^2_{i,t+1}] = \tanh(\mathbf{W}_{fa}\mathbf{h}^c_{i,t} + \mathbf{b}_{fa})$$

3. Triplet loss function integrating the network output combined with the piecewise-geodesic curve is defined as:

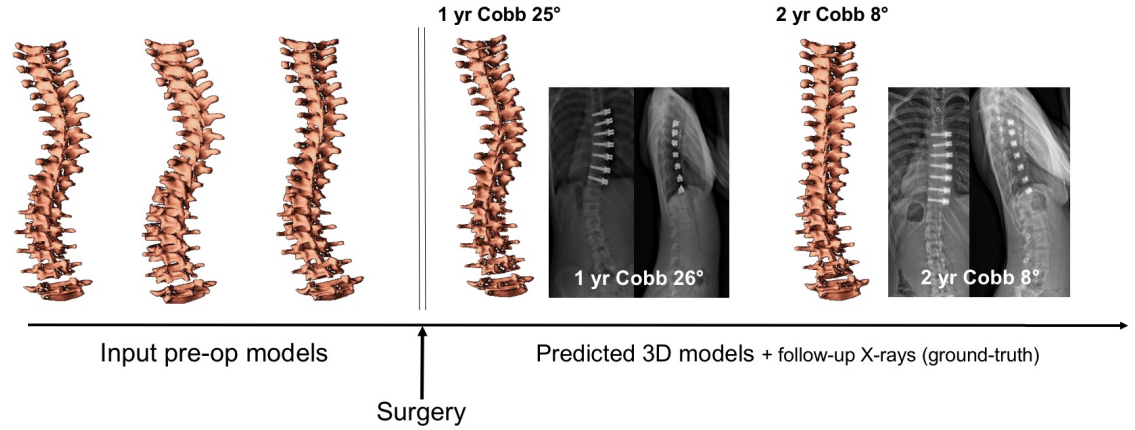
$$\mathcal{L} = \sum_{i=1}^{n} \beta(y^1_{i,t+1} - \widehat{y}^1_{i,t+1})^2 + (1-\beta)(y^2_{i,t+1} - \widehat{y}^2_{i,t+1})^2 + \lambda((y^1_{i,t+1} - \gamma(1)) + (y^2_{i,t+1} - \gamma(2)))$$

with β and λ modulating actual outcomes $(\widehat{y}^1_{i,t+1}, \widehat{y}^2_{i,t+1})$ and trajectory regularization, respectively.

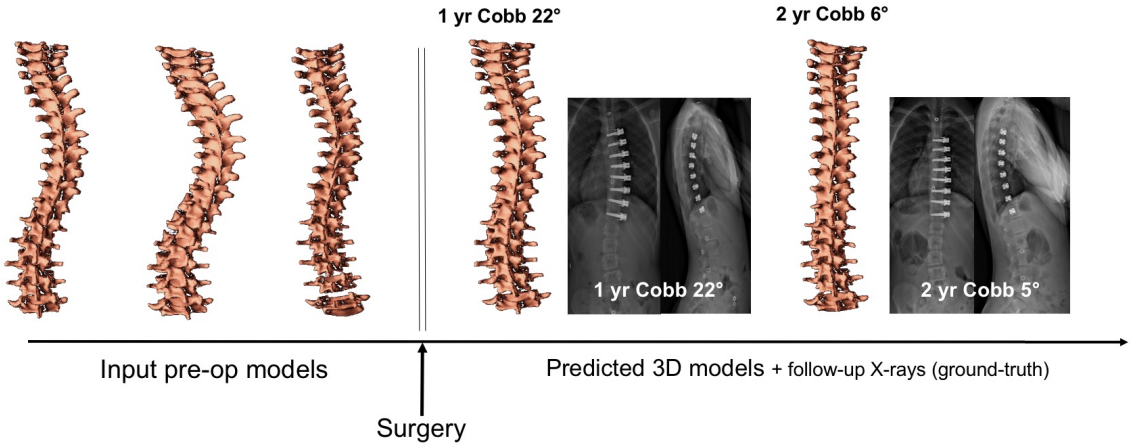
4. Repeating this process ensure regularization along γ.

Shape prediction results of 3D geometry

12-yr F Risser 0 Instrumented from T4 to T10



13-yr F Risser 1 Instrumented from T3 to T10



Prediction accuracy results

3D RMS errors (mm), Dice (%) and Cobb angles (^{0}), compared with biomechanical model (Cobetto et al. 2018) , ST-ResNet (Zheng et al. AAAI 2017), deep auto-encoders (AE; Thong et al. 2016), and ST-Manifold (MICCAI 2018).

	1-year visit		
	3D RMS (mm)	Dice	Cobb
Biomec. Simulation	3.8 ± 1.3	83 ± 3.7	3.5 ± 1.0
Deep AE	4.9 ± 3.5	80 ± 4.3	5.7 ± 2.7
ST-ResNet	3.4 ± 1.7	85 ± 3.6	4.0 ± 2.4
ST-Manifold (Mandel et al.)	3.0 ± 1.0	90 ± 2.7	2.1 ± 0.8
STCN (no reg. + gating)	2.5 ± 0.8	91 ± 2.6	2.0 ± 0.6
STCN (no regularization)	2.2 ± 0.7	92 ± 2.5	1.9 ± 0.6
Proposed	1.9 ± 0.8	93 ± 2.5	1.7 ± 0.5

Questions