

MRI Pulse Sequence Integration for Deep-Learning Based Brain Metastasis Segmentation

Darvin Yi*, Endre Grøvik, Michael Iv, Elizabeth Tong, Kyrre Eeg Emblem, Line Brennhaug Nilsen, Cathrine Saxhaug, Anna Latysheva, Kari Dolven Jacobsen, Åslaug Helland, Greg Zaharchuk, and Daniel Rubin

arXiv:1912.08775v1 [eess.IV] 18 Dec 2019

Abstract—Magnetic resonance (MR) imaging is an essential diagnostic tool in clinical medicine. Recently, a variety of deep learning methods have been applied to segmentation tasks in medical images, with promising results for computer-aided diagnosis. For MR images, effectively integrating different pulse sequences is important to optimize performance. However, the best way to integrate different pulse sequences remains unclear. In this study, we evaluate multiple architectural features and characterize their effects in the task of metastasis segmentation. Specifically, we consider (1) different pulse sequence integration schemas, (2) different modes of weight sharing for parallel network branches, and (3) a new approach for enabling robustness to missing pulse sequences. We find that levels of integration and modes of weight sharing that favor low variance work best in our regime of small data ($n = 100$). By adding an input-level dropout layer, we could preserve the overall performance of these networks while allowing for inference on inputs with missing pulse sequence. We illustrate not only the generalizability of the network but also the utility of this robustness when applying the trained model to data from a different center, which does not use the same pulse sequences. Finally, we apply network visualization methods to better understand which input features are most important for network performance. Together, these results provide a framework for building networks with enhanced robustness to missing data while maintaining comparable performance in medical imaging applications.

Index Terms—Deep Learning, Image Segmentation, Magnetic Resonance, Pulse Sequence, Metastasis, Dropout.

I. INTRODUCTION

Recently, there has been an explosion in applying deep learning methods to magnetic resonance (MR) imaging. Many techniques utilize classification networks to aid diagnosis and detection [1–3]. Other techniques applied the generative capabilities of convolutional neural networks (CNNs) - e.g. through autoencoder networks [4–7] or generative adversarial networks (GANs) [8, 9] - to denoise MR images [10] and enhance signal-to-noise ratio [11, 12]. Finally, deep learning for segmentation has been of core interest to the computational

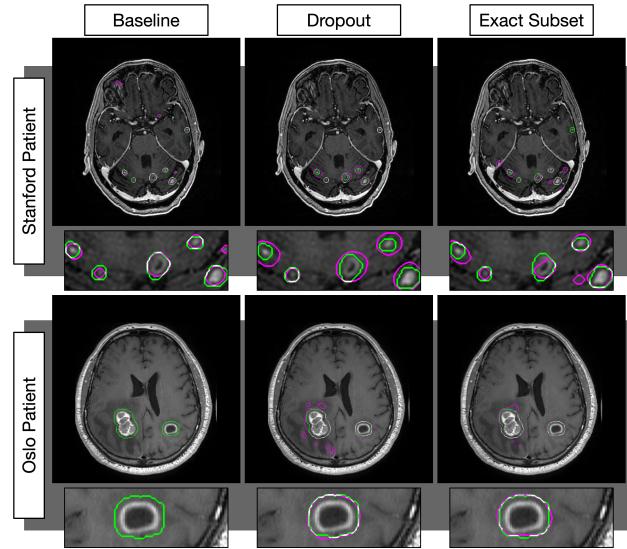


Fig. 1: Visualization of Segmentations on Stanford and Oslo Patients. This figure shows visualizations of the results from our baseline input-level integration network, the dropout trained equivalent, and the baseline network trained (and tested) on the BRAVO-censored data. Our networks were only trained on Stanford patients, and we show examples for a Stanford patient and an Oslo patient, both not used during training. Below each image, we show a magnified zoom on areas of possible interest. Expert Annotation segmentations are outlined in green. Predictions are outlined in purple/magenta. Overlap will be shown as white.

biomedical community since its advent. Biomedical segmentation produced U-Net [13], an architecture used in the broader computer science community to this day. For MR imaging, both 2.5D networks and 3D networks have been commonly used to solve segmentation problems [14–16]. In this work, we will focus on further developing segmentation techniques for MR images.

MR imaging protocols vary from institution to institution and case by case. Typical brain metastasis MRI protocol consists of T2-weighted Fluid attenuation inversion recovery (FLAIR) sequence, pre-contrast and post-contrast T1-weighted sequences. The critical sequence is the post-contrast 3D T1-weighted sequence, which is a high-resolution isotropic sequence acquired by either Inversion Recovery prepended Fast Spoiled Gradient-Echo (IR-FSPGR) or Fast Spin-Echo (FSE) techniques. 3D T1-weighted IR-FSPGR acquire isotropic T1-weighted images with excellent grey-white matter differentiation and are used broadly. It is often difficult to distinguish enhancing metastases

Asterisk indicates corresponding author.

D. Yi* is with the Department of Biomedical Data Science at Stanford University, Stanford, CA 94305 USA e-mail: darvinyi[at]Stanford[dot]EDU
E. Grøvik, K. E. Emblem, and L. B. Nilsen are with the Department for Diagnostic Physics at Oslo University Hospital

M. Iv, E. Tong, and G. Zaharchuk are with the Department of Radiology at Stanford University.

C. Saxhaug and A. Latysheva are with the Department of Radiology and Nuclear Medicine at Oslo University Hospital.

K. D. Jacobsen and Å. Helland is with the Department of Oncology at Oslo University Hospital.

D. Rubin is with the Department of Biomedical Data Science and Department of Radiology at Stanford University.

from background vascular enhancement on post-contrast 3D T1-weighted IR-FSPGR [17]. On the contrary, post-contrast 3D T1-weighted FSE provides inherent background vascular suppression, yielding a higher contrast-to-noise ratio (CNR) than post-contrast FSPGR, making enhancing metastases more conspicuous [18]. No matter the protocol, networks trained with a set of input pulse sequences are notoriously sensitive to missing pulse sequences at run-time [19–22].

In this work, we characterize what we have identified as best practices for pulse sequence integration. We evaluate how to integrate multiple pulse sequences for optimal segmentation and detection performance. We also introduce how using a “dropout” layer on the integration layer for our networks confers robustness to missing pulse sequences with no significant loss of performance given all pulse sequences. Finally, we show that this integration dropout method gives a more consistent training schedule, or behavior of gradients, by analyzing input image saliency maps throughout training.

II. RELATED WORK

Deep Learning and Segmentation. Deep learning-based methods have shown excellent performance for segmentation. Following the first fully convolutional network (FCN) in 2014 [23], semantic segmentation with deep learning has continued to improve with SegNet [24] and U-Net [13] in 2015, PSPNet [25] and DeepLab [26] in 2017, and recently DeepLabv3 [27]. Constant advances in computing power and use of contextual information have provided further benefits.

Segmentation for MR Imaging. The MR community has a strong background in using deep segmentation techniques. Open challenges like the BraTS [14] or the Promise12 [28] challenges have provided many datasets to train segmentation algorithms in the MR domain. Though earlier approaches typically use single-frame 2D FCN approaches [29], current approaches leverage 3D information in either 2.5D or 3D FCNs [15, 30, 31]. Other methods have investigated using deep learning methods from the recurrent network domain [32], a type of network built to handle sequence-like data. These methods represent the variety of deep learning-based segmentation techniques that have proven valuable in the quantitative MR community.

Multi-modal Data Integration. Because of deep network’s generally malleable architecture, many groups have investigated the best modes to combine different data modalities. Our work on integrating different pulse sequences draws from data integration schemes first found in video frame analysis [33], which create parallel shared weights before the layer of integration, after which there is only one shared network. In investigating parallel network branches, we also looked into the work done with siamese networks [7, 34, 35]. Most notably, our study closely models that of Eitel et. al. on training robust RGB-D networks [36]. In the MR space, [22] has been done to create synthetic modalities to replace missing data should the case arise. However, most similar to our own work would be [20]’s 2016 Hetero-Modal Image Segmentation (HeMIS) where each present modality “votes” within a learned latent representation space by means of arithmetic mean. Nonetheless, effectively addressing the problem of missing data remains an important outstanding challenge for MR images.

Dropout Regularization. Dropout, an early addition to the deep learning toolbox, helps decrease learned co-linearities during training [37, 38] by randomly setting activations to zero during training based on a predetermined probability. An equivalent channel-wise dropout is used for training CNNs [39].

Network Performance Visualization. Many techniques, such as the class action map (CAM), have been developed to help visualize network decision making with respect to the original input images [40]. These visualization methods have been quite popular in medical imaging applications as the main form of network explanation [41, 42]. Here, we focus on the idea behind the saliency map, a gradient-based method [43, 44].

Previous Work and Novel Contributions. Our group’s previous work set the baseline performance for our metastasis segmentation task [45]. In comparison, we have updated the segmentation architecture and used detection metrics to better describe our network’s performance. Methodologically, we build on previous works in both medical imaging and real-world computer vision to investigate the best methods of integrating multi-modal MR pulse sequence data for metastasis segmentation. In this paper, we report the following novel contributions:

- 1) An initial investigation into different pulse sequence integration architectures and training techniques for metastasis segmentation in the small data regime.
- 2) Creation and usage of an input-layer dropout that makes the network robust to receiving missing pulse sequences and introduces more consistent network training behavior.
- 3) Evaluation of the performance of the trained network on different pulse sequences to show which input modalities are most and least useful.

III. DATA

This retrospective, multi-center study was approved by our Institutional Review Board. For our experiments, we use brain MR data from two different institutions: Stanford Hospital and Oslo University Hospital. These datasets will hereafter be referred to as the Stanford data and the Oslo data. For both cohorts, all MR image-series were co-registered into one common anatomical space. This was performed using the nordicICE software package (Nordic Neuro Lab, Bergen, Norway). For the Stanford data, all image series were co-registered to a post-contrast 3D T1-weighted inversion recovery fast spoiled gradient echo, whereas for the Oslo data, post-contrast 3D T1-weighted spin echo images was used at the reference series. Furthermore, for the Oslo data, a defacing procedure was applied to anonymize all image-data using an in-house algorithm (MATLAB R2017a version 9.2.0, MathWorks Inc. Natick, MA, USA).

A. Stanford Metastasis Cohort

A total of 156 patients with brain metastases were examined at Stanford Hospital. Inclusion criteria for patient enrollment were presence for known metastatic lesion(s), no prior surgical or radiation treatment, and the availability of the required

MR images. Imaging was performed on both 1.5T (SIGNA Explorer and TwinSpeed, GE Healthcare, Chicago, IL) and 3T (Discovery 750 and 750w and SIGNA Architect, GE Healthcare, Chicago, IL; Skyra, Siemens Healthineers, Erlangen, Germany) clinical scanners. The imaging protocol included pre- and post-contrast T1-weighted 3D fast spin echo, post-Gd T1-weighted 3D axial inversion recovery prepped fast spoiled gradient-echo, and 3D fluid-attenuated inversion recovery. For contrast-enhancement, a dose of 0.1 mmol/kg body weight of gadobenate dimeglumine (MultiHance, Bracco Diagnostics, Princeton, New Jersey) was intravenously administered. Primary cancers from the Stanford dataset consists of lung (99), breast (33), skin/melanoma (7), genitourinary (7), gastrointestinal (5), and other miscellaneous primary sources (5).

B. Oslo Metastasis Cohort

For the Oslo data, a total of 65 patients with brain metastases were examined. To be eligible for inclusion, patients had to receive stereotactic radiosurgery for at least one brain metastasis measured to a minimum of 5 mm in one direction, be untreated or progressive after systemic or local therapy, have confirmed non-small-cell lung cancer or malignant melanoma, be ≥ 18 years of age; have an Eastern Cooperative Oncology Group performance status score of maximum 1, and have a life expectancy of more than 6 weeks. All imaging was performed on a clinical 3T Skyra scanner (Siemens Healthineers, Erlangen, Germany). The imaging protocol included pre- and post-contrast T1-weighted 3D fast spin echo and 3D T2-weighted FLAIR. For contrast-enhancement, a dose of 0.1 mmol/kg body weight of gadoterate (Dotarem, Guerbert, France) was intravenously administered. Note that the Oslo data only had three of the four MR sequences acquired in the Stanford data, and that two patients in the Oslo cohort were missing the pre-contrast T1-weighted 3D fast spin echo. Hence, the data represents a real-world use case of having to create a model that would be robust to missing pulse sequences, a scenario quite common considering that different institutions use differing imaging protocols. The primary cancers of the Oslo data consists of lung (45) and melanoma (20).

IV. METHODS

A. Multiple Levels of Pulse Sequence Integration

As described above, our MR images are 3-D voxel representations of different physical properties captured in different pulse sequences. Because they have been co-registered with each other, the voxel at location $[i, j, k]$ for each pulse sequence captures the same physical location. Thus, pixel-wise operations, such as averaging or channelwise concatenation, are location-preserving and do not have to rely on operations such as flattening.

Common approaches to combine information from different branches of co-localized data are taking some weighted sum (e.g. average) or performing a concatenation in the channel space. In our experiments, we use both of these methods. An outstanding question for integrating information from each of the different pulse sequences is how much to process distinct pulse sequences in isolation and how much to process them after

integration. Different integration schemes have been explored in video frame analysis and other applications, but limited work has been done to characterize their performance for MR images. Although there exists an intractable number of integration schemes, we focus on three levels: (a) input-level integration, (b) mid-level integration, and (c) end-level integration. These three schemes are diagrammed in Figure 2.

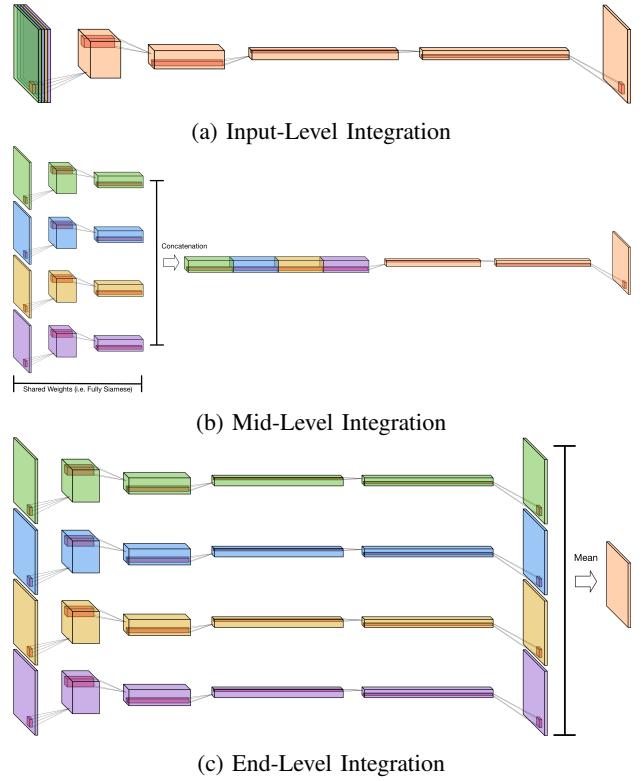


Fig. 2: Pulse Sequence Integration levels. (a) The baseline, as used in previous work [45], integrates the pulse sequence images at the input level. (b) The mid-level integration has both separate and shared layers. We do this integration after DeepLab v3's block 4, before the atrous spatial pooling. (c) End-level integration gives each pulse sequence a separate network and takes the mean of the final predictions. All network diagrams are not representative of the actual convolutional architecture, but merely a placeholder to diagram the larger concept of multi-modal feature integration.

B. Multiple Modes of Weight Sharing

In addition to different levels of integration, we can also have different modes of weight sharing when there are parallel network layers corresponding to different pulse sequences, as in the mid-level and end-level architectures. We considered three major cases: (1) fully independent, (2) fully shared, and (3) L2-tied. *Fully independent weights* have parallel layers that are initialized randomly and receive different gradients. *Fully shared weights* (also referred to as fully siamese), are initialized identically and receive identical gradients. *L2-tied weights* are initialized identically and allowed to receive different gradients, subject to an L2 loss on pairs of parallel weights to penalize diverging weights. These loss functions are represented in a high-level mathematical fashion in equations 1-3. In both the input- and end-level integration schemas, we only have

one basic network that we will call F with corresponding learned parameters θ . However, for the mid-level integration network, we will call the convolution blocks before integration F_e (for early) and the blocks after integration F_l (for late), with corresponding parameters θ_e and θ_l . Cases where weights are allowed to diverge (receive differing gradients) or forced to be identical are denoted by separating a single θ value into $\{\theta_1, \theta_2, \theta_3, \theta_4\}$. Differing colors match those seen in figure 2.

$$\hat{Y} = F(\textcolor{green}{X}_1, \textcolor{blue}{X}_2, \textcolor{orange}{X}_3, \textcolor{purple}{X}_4; \theta) \quad (1)$$

$$\mathcal{L} = \text{CE}(Y, \hat{Y})$$

$$\hat{Y} = F_l(F_e(\textcolor{green}{X}_1; \theta_1), F_e(\textcolor{blue}{X}_2; \theta_2), F_e(\textcolor{orange}{X}_3; \theta_3), F_e(\textcolor{purple}{X}_4; \theta_4); \theta_l) \quad (2a)$$

$$\mathcal{L} = \text{CE}(Y, \hat{Y})$$

$$\hat{Y} = F_l(F_e(\textcolor{green}{X}_1; \theta_e), F_e(\textcolor{blue}{X}_2; \theta_e), F_e(\textcolor{orange}{X}_3; \theta_e), F_e(\textcolor{purple}{X}_4; \theta_e); \theta_l) \quad (2b)$$

$$\mathcal{L} = \text{CE}(Y, \hat{Y})$$

$$\hat{Y} = F_l(F_e(\textcolor{green}{X}_1; \theta_1), F_e(\textcolor{blue}{X}_2; \theta_2), F_e(\textcolor{orange}{X}_3; \theta_3), F_e(\textcolor{purple}{X}_4; \theta_4); \theta_l) \quad (2c)$$

$$\mathcal{L} = \text{CE}(Y, \hat{Y}) + \sum_{i=1}^4 L_2(\theta_i, \bar{\theta})$$

$$\hat{Y} = F(\textcolor{green}{X}_1, \theta_1) + F(\textcolor{blue}{X}_2, \theta_2) + F(\textcolor{orange}{X}_3, \theta_3) + F(\textcolor{purple}{X}_4, \theta_4) \quad (3a)$$

$$\mathcal{L} = \text{CE}(Y, \hat{Y})$$

$$\hat{Y} = F(\textcolor{green}{X}_1, \theta) + F(\textcolor{blue}{X}_2, \theta) + F(\textcolor{orange}{X}_3, \theta) + F(\textcolor{purple}{X}_4, \theta) \quad (3b)$$

$$\mathcal{L} = \text{CE}(Y, \hat{Y})$$

$$\hat{Y} = F(\textcolor{green}{X}_1, \theta_1) + F(\textcolor{blue}{X}_2, \theta_2) + F(\textcolor{orange}{X}_3, \theta_3) + F(\textcolor{purple}{X}_4, \theta_4) \quad (3c)$$

$$\mathcal{L} = \text{CE}(Y, \hat{Y}) + \sum_{i=1}^4 L_2(\theta_i, \bar{\theta})$$

C. Special Case: Combining Pre- and Post-Contrast

We have created a special architecture for integrating pre- and post-contrast images. In addition to testing all combinations of integration levels and weight-sharing modes, we include an additional mid-level integration scheme. Rather than concatenating the pre- and post-contrast layers, we combine them and subtract them in the feature-space. Our reasoning for this is that the features computed from the pre-contrast images and the features from the post-contrast images should only differ locally where the imaging differs, which should correspond to lesions. This choice, which is informed by domain knowledge, may be favorable for the bias-variance tradeoff. Note that this approach makes sense only if the weights in the pre- and post-contrast branches are fully shared.

D. Integration-Level Dropout Layer

A key problem is that a network trained with a given structure (i.e., given preset input channels) is not robust to any missing information during inference. To address this limitation, we propose stochastically zeroing out random pulse sequences during training, an idea inspired by the channel-wise dropout

commonly used when training CNNs. By training on inputs with missing pulse sequences, such networks should be robust to similar inputs during deployment.

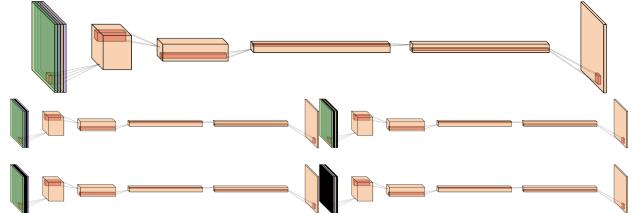


Fig. 3: Integration Level Dropout for Input-Level Integration We take the standard input-level integration architecture (reproduced from Figure 2a, top) and show four potential “dropped out” versions that we might encounter during training (bottom). From our original four input-concatenated pulse sequences, we randomly drop out 0-3 input channels (shown in black). Thus, the network can encounter missing pulse sequences during training.

An important consideration is the weighting function. Because dropping out neurons with probability p during training but having them fully firing during inference creates a difference in the total sum of neuron activations, we can either upweight the dropout layer’s neurons by a factor of $\frac{1}{1-p}$ during training or downweight the dropout layer’s neurons by a factor of $1-p$ during inference. We upweight our input dropout layer by a factor of $\frac{1}{1-p}$ where p is the actual proportion of dropped out pulse sequences, in both training and inference.

An important note is that dropout cannot be done on a purely statistical basis. Normally, each channel has a certain probability of being dropped. However, as we use a 2.5D network structure, we must take care to drop out all of the z-slices of a certain pulse sequence or none at all. We must also make sure to never drop out all four pulse sequences during training. This situation would result in the network receiving an input tensor of all 0s, which would be intractable and lead to unstable training.

E. Visualizing Input Gradient Accumulation

A standard method for visualizing networks is the saliency map, which in its most basic form is the gradient of the loss with respect to the input image:

$$\text{Saliency Image} = \frac{\partial L}{\partial \text{Img}} \quad (4)$$

From equation 4, we construct separate quantities that we term the *saliency aggregate* and the *cumulative saliency aggregate*. The saliency aggregate is the sum of the absolute value of the saliency image, which approximates the sensitivity of the loss with respect to the input image. The cumulative saliency aggregate is the iteration-based cumulative sum of the saliency aggregate during training. These metrics allow us to quantify how the network learns over the course of training.

$$\text{Saliency Aggregate} = \text{Sum} \left(\left| \frac{\partial L}{\partial \text{Img}} \right| \right) \quad (5a)$$

$$\text{Cumulative Saliency Aggregate} = \sum_{\text{iter}=0}^{\text{now}} \text{Sum} \left(\left| \frac{\partial L}{\partial \text{Img}} \right| \right)_{\text{iter}} \quad (5b)$$

F. Implementation Details

In our paper, we use the DeepLabv3 [27] architecture as our segmentation baseline. The architecture is well known for its ability to incorporate an extremely large field of context, most likely due to its atrous (dilated) convolutions. Though a potential weakness of the DeepLabv3 is worse performance on thinner structures (such as the legs of a chair or a light pole), this is not a significant concern for metastasis segmentation, as most lesions are spherical. For mid-level integration architectures, we integrate the modalities after the fourth main residual block, right before the Atrous Spatial Pyramid Pooling (ASPP) layer.

Our input pulse MR's have all been resized to a 256x256 image. The only preprocessing that has been done is adaptive histogram equalization on each slice of the data. This provides two main benefits: (1) each pixel value for every scan is in the same value-space and (2) contrast is enhanced. We believe that at the cost of minimal artifacts, adaptive histogram equalization is a quick and dirty way to fix domain shift errors to a first approximation. For our 2.5D, we use 5 slices. Thus, for our input-level integration networks, we will have an input tensor with 20 channels, 5 slices each for 4 pulse sequences.

To run dropout experiments, we create a custom data loader transform in PyTorch, which takes input data of four pulse sequences and randomly drops out each pulse sequence with a probability of 25%. In the case that all four pulse sequences are dropped out (probability 0.4%), no pulse sequence is dropped. The table below shows the probability distribution for the number of pulse sequences dropped out.

# Seq's Dropped Out	0	1	2	3	4
Probability	32%	42%	21%	4.7%	0%

Furthermore, our main experiments for missing pulse sequences rely on censoring out the BRAVO Post-Contrast sequence. In our training, there is an 11% chance of seeing an input image with only the BRAVO sequence dropped out.

All code was written in PyTorch. Both training and inference were done on two NVIDIA 1080Ti GPUs. During training, we over-sample z-slices with any portion of ground truth annotated lesions by a factor of 10x compared to frames without. Training was done for 10 epochs from random initialization on a dataset of 100 patients, taking about 20 hours. The forward pass inference time is approximately 150ms per z-slice, or about 30 seconds for a single patient. This becomes about 5 seconds per z-slice on a CPU (Intel i7-8700k), which translates to a 30-minute patient inference time.

G. Metrics for Detection and Segmentation

In our work, we treat the segmentation network as both a detection and a segmentation network. To evaluate detection performance, we report the mean average precision (mAP)

value or area under the precision-recall (PR) curve. Generating a PR curve requires (1) a standard to decide when a prediction is a true positive and (2) a probability per predicted lesion. We create predicted lesions by binarizing the complete 3D segmentation probability volume, using an empirically determined probability threshold of 10%. Thus, any 3D connected component of predicted voxels with probability greater than 10% is a single predicted lesion. If the center of mass of the 3D connected component is within 1mm of a ground truth annotation, we call that predicted lesion a true positive. (Given that lesions are small, this metric is conservative.) The predicted probability of the lesion will be the average probability of all of the voxels in that region with respect to the original probability volume. Thus, by definition, the minimum predicted probability for any lesion with this method will be 10% or 0.1. With the constructed PR curve, we report the mAP score and the maximum sensitivity.

To quantify segmentation performance, we report DICE scores for our true positive segmentations at the maximum sensitivity. In other words, we take our binary segmentation map (with probability threshold 10%) and report the DICE scores for all predicted binary lesions with center of mass within 1mm of the expert-annotated lesions.

V. RESULTS

A. Stanford Metastasis Dataset Experiments

TABLE I: Pulse Sequence Integration

Integration Level	Weight Sharing	mAP	Max Sens.	TP DICE
Input	-	46 (44,47)	80	72
Mid	Fully Shared	44 (42,45)	82	<u>74</u>
Mid	Fully Ind.	39 (37,41)	54	70
Mid	L2 Tied	44 (42,45)	<u>83</u>	73
End	Fully Shared	35 (33,37)	52	70
End	Fully Ind.	28 (25,30)	47	62
End	L2 Tied	34 (32,36)	53	71

Below are values for the pre- and post-contrast subtraction.

Mid	Fully Shared	48 (46,49)	86	75
-----	--------------	------------	-----------	-----------

Table I shows results from the different pulse sequence integration and weight sharing architectures. The best values are bolded, while the second-best values are underlined. In general, the best performing network is the pre-/post-contrast subtraction mid-level integration network. The next best networks are the input-level integration network and the fully shared and L2-tied mid-level integration networks.

Table II shows the results of training the networks with the input dropout schema. There is a significant decrease in performance in the two networks with fully independent weight sharing. In addition, we see some network degradation for the L2-tied end-level integration network. However, other networks, including the high-performing networks from Table I (input-level integration, fully shared mid-level, and L2-tied mid-level), show approximately equivalent performance.

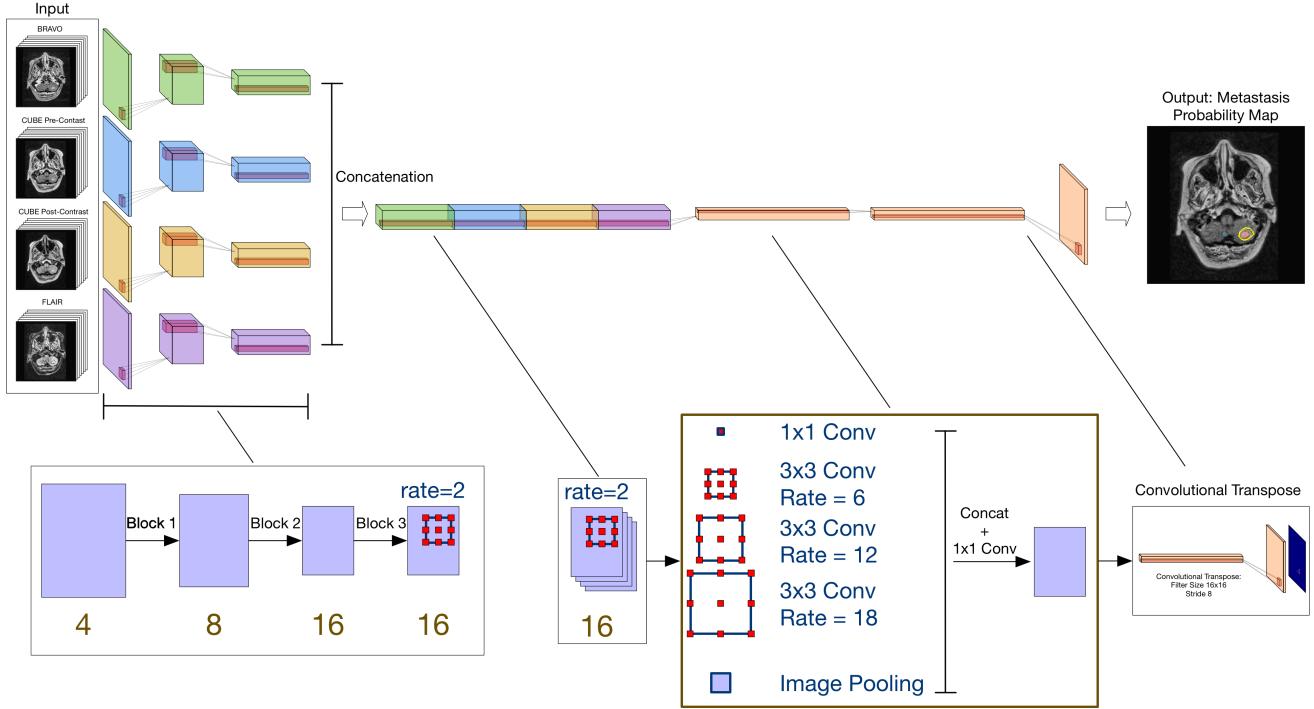


Fig. 4: Deep Learning Pipeline Our main pipeline is simple and is built largely on the DeepLabv3 architecture [27]. For our input-level integration models, we will concatenate all of our input pulse sequence slices together. For our mid-level integration models, we concatenate the features immediately before the ASPP layer, as shown in this figure. Finally, for end-level integration, we will pass through each pulse sequence through an independent DeepLabv3 model and average the output probability maps.

TABLE II: Pulse Sequence Int. with Dropout Training

Integration Level	Weight Sharing	mAP	Max Sens.	TP DICE
Input	-	45 (-2.2%)	80 (0.0%)	72 (0.0%)
Mid	Fully Shared	43 (-2.3%)	83 (+1.2%)	74 (0.0%)
Mid	Fully Ind.	34 (-12.8%)	44 (-18.5%)	71 (+1.4%)
Mid	L2 Tied	45 (+2.3%)	83 (0.0%)	73 (0.0%)
End	Fully Shared	35 (0.0%)	54 (+3.8%)	70 (0.0%)
End	Fully Ind.	24 (-14.3%)	41 (-12.8%)	58 (-6.5%)
End	L2 Tied	34 (0.0%)	50 (-5.7%)	71 (0.0%)

The pre-/post-contrast subtraction network was not trained with input dropout, as dropping out either pre- or post-contrast images alone disrupted the core structure of the network.

B. Input Pulse Sequence Censorship

Table III highlights the need for a network robust to missing pulse sequences. When censoring out the BRAVO pulse sequence during inference on a network trained with all four pulse sequences, the network’s predictive power is lost. One simple solution is to train a network for the set of pulse sequences without the BRAVO, in anticipation of this case. As shown in the last row of Table III, this network has good performance, albeit degraded, in inference on data without the BRAVO pulse sequence.

Table IV shows the results of inference on a BRAVO-censored dataset using networks trained with the input dropout technique, all of which show restored performance. Though

TABLE III: Non-Dropout Model on BRAVO-Censored Data

Integration Level	Weight Sharing	mAP	Max Sens.	TP DICE
Input	-	0 (0,0)	0	-
Mid	Fully Shared	0 (0,0)	0	-
Mid	Fully Ind.	0 (0,0)	0	-
Mid	L2 Tied	0 (0,0)	0	-
End	Fully Shared	0 (0,0)	0	-
End	Fully Ind.	0 (0,0)	0	-
End	L2 Tied	0 (0,0)	0	-

Below is an input-level int. model trained without BRAVO.

Input	-	40 (38,41)	65	71
-------	---	-------------------	-----------	-----------

TABLE IV: Dropout Model on BRAVO-Censored Data

Integration Level	Weight Sharing	mAP	Max Sens.	TP DICE
Input	-	37 (35,39)	62	<u>72</u>
Mid	Fully Shared	40 (38,41)	64	70
Mid	Fully Ind.	34 (32,36)	60	73
Mid	L2 Tied	<u>38</u> (36,40)	65	71
End	Fully Shared	30 (27,32)	55	70
End	Fully Ind.	15 (12,18)	45	63
End	L2 Tied	28 (25,30)	57	<u>72</u>

none of the networks matches the level of the model trained on the exact subset of inference pulse sequences (last row of Table III), performance is approximately equivalent with the

fully shared mid-level integration network trained with dropout. This demonstrates the enhanced robustness to missing pulse sequences conferred by the input dropout layer.

TABLE V: Input-Level Dropout Model on Different Combinations of Input Pulse Sequences

BRAVO Post-C	CUBE Pre-C	CUBE Post-C	FLAIR	mAP	Max Sens.	TP DICE
✓	✓	✓	✓	45 (43,46)	80	72
✗	✓	✓	✓	37 (35,39)	62	72
✓	✗	✓	✓	38 (36,40)	67	71
✓	✓	✗	✓	42 (40,43)	72	70
✓	✓	✓	✗	44 (42,45)	<u>79</u>	72

Having trained networks that can utilize any subset of the four predefined pulse sequences, the relative utility of each modality can be assessed by censoring a particular pulse sequence and using that test set in the input-level dropout model. Table V compares the sensitivity of the network when each pulse sequence is omitted in turn. Unsurprisingly, the network performs best when all of the pulse sequences are given during inference. However, it is most sensitive to the loss of the post-contrast BRAVO and the pre-contrast CUBE, followed by loss of the post-contrast CUBE. The network is minimally sensitive to the absence of the T2-weighted FLAIR.

C. Evaluation on Oslo Dataset

To evaluate the generalizability of the trained network, we evaluated its performance on data from a different center (Oslo). In table VI below, we see a comparison of the dropout-based networks and the exact-subset network. The network trained on all four pulse sequences without the dropout had complete failure, the same as when censoring out BRAVO data on the Stanford test set. All networks below are the same networks shown in previous sections, i.e. trained only on the 100-patient Stanford training set. These values show not only that the networks generalized well to the Oslo data but also that the exact-subset network and the dropout-trained networks showed comparable performance.

It should be noted here that the bottom row represents two different models. Of the 67-patient Oslo cohort, 65 patients had 3 pulse sequences: (1) CUBE pre-contrast, (2) CUBE post-contrast, and (3) T2w FLAIR. However, 2 patients only had 2 pulse sequences: (1) CUBE post-contrast and (2) T2w FLAIR. Thus, we used a network trained on the exact subset of three pulse sequences on the 65 patients and another network trained on the exact subset of two pulse sequences for the remaining two patients. For the dropout model, we only used one model per row but did change the weighting factor between the two groups of patients, as discussed in Section IV-F.

A representative image comparing the baseline input-level network, the same network trained with dropout, and the same network trained without dropout on the exact subset data (BRAVO-censored) has been shown in figure 1. Worth noting in figure 1 is the fact that our baseline model will have complete failure on the Oslo patient, since it did not receive

TABLE VI: Testing on Oslo Data (no BRAVO Scans)

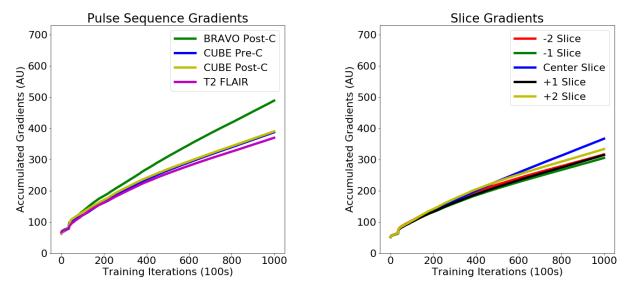
Integration Level	Weight Sharing	mAP	Max Sens.	TP DICE
Input	-	68 (66,69)	92	85
Mid	Fully Shared	67 (65,69)	90	83
Mid	Fully Ind.	61 (58,63)	75	82
Mid	L2 Tied	68 (66,69)	92	84
End	Fully Shared	59 (57,60)	70	80
End	Fully Ind.	36 (33,38)	45	72
End	L2 Tied	55 (53,57)	67	81
Below is an input-level int. model trained on without BRAVO.				
Input	-	67 (65,69)	92	84

all four pulse sequences it was trained with. However, both the dropout model and the model trained without BRAVO perform well on the Oslo Patient.

D. Visualization with Saliency Map Cumulative Sum

Below, we investigate the cumulative aggregate saliency (Equation 5b) for our input-level integration networks trained with and without dropout. In Figures 5 and 6, we look at a single input-level integration network (corresponding to the input-level integration networks shown in Tables 1-6). On the left, we see the cumulative aggregate saliencies separated by input pulse sequences. On the right, we see the cumulative aggregate saliencies separated by z-slice position.

We also trained three additional input-level networks each with and without the dropout method. The networks were trained identically apart from random parameter initializations and different mini-batch orderings. These networks, along with the original network, have their pulse sequence saliencies plotted in Figures 7 and 8. These results demonstrate that the dropout method allows for a more consistent training schedule, learning at approximately equal rates from all available pulse sequences besides FLAIR. Additional quantitative results are not shown for these three networks, but they are comparable to those of the original network.



(a) Pulse Sequence Gradients

(b) Slice Gradients

Fig. 5: Cumulative Aggregate Saliencies for Network Trained without Dropout.

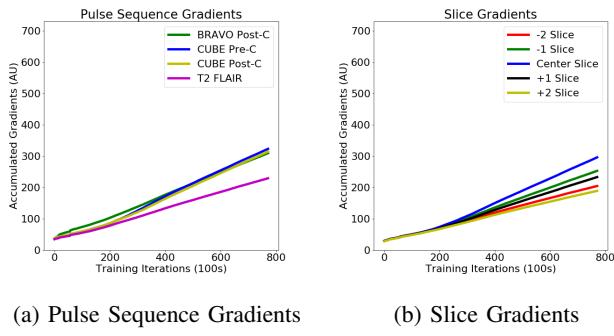


Fig. 6: Cumulative Aggregate Saliencies for Network Trained with Dropout.

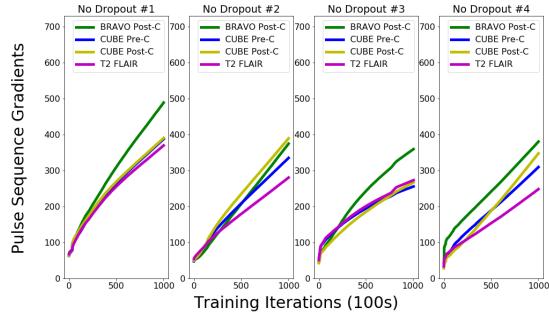


Fig. 7: Four Stochastically Different Training Runs of No-Dropout Networks.

VI. DISCUSSION

A. Pulse Sequence Integration and the Bias-Variance Tradeoff

In general, the best networks have the most shared parameters or most regularization (Table I). This leads us to believe that the small dataset (100 patients) favors constraints on networks and precludes identification of the optimal architecture. It is clear that the total compute of the input-level integration network is the lowest, as it lacks replicated layers. Since mid-level and end-level integration show no clear benefit, we advise using the input-level integration network in the small data regime.

Intriguingly, the best network is a mid-level integration network where the pre- and post-contrast CUBE features are

subtracted at the integration level rather than concatenated. We believe that this result fits with the bias-variance tradeoff in the small data regime. By constructing a high-bias model that only the difference of pre- and post-contrast features should be important, we constrain our model in a way that minimizes variance error. However, we argue that the assumptions behind subtracting the pre- and post-contrast features are well founded, adding minimal bias error to the model.

B. Dealing with Missing Pulse Sequences

As shown in Table III, missing pulse sequences destroyed the performance of our trained models, motivating us to develop a single network robust to missing pulse sequences at inference time. Such a network would be essential for real-world applications, as the types of data taken may vary over time or between institutions. Training with integration-level dropout restored network performance, as shown in Table IV. We compared the performance of these models to that of a network trained on the exact subset of provided data without the dropout method, given in the last row of Table III. Though this exact-subset network performs best overall, the fully-shared and L2-tied mid-level integration dropout models closely match its performance. We can also see from Table I that introducing the dropout method does not degrade performance when we are given all pulse sequences.

One important consideration is the difference in model selection between the exact-subset and dropout networks. Every 1000 iterations, we tested our model on a small validation set ($n=6$). If the model performed better than previously, we saved that state as the new best model, overwriting the previous best. The exact-subset network was validated on a set that had the exact subset, while the dropout models were validated on inputs with all four pulse sequences available. Thus, the dropout models were biased for the best performance given no missing pulse sequences. We additionally trained an input-level integration network with dropout but with the validation set having the BRAVO scan censored out. This created a model that performed slightly better than the exact subset network (mAP=42, max sensitivity=69, TP DICE=72). However, performance was slightly reduced when testing on the test set without any censorship (mAP=44, max sensitivity=75, TP DICE=72).

C. Input Dropout Network to Assay Pulse Sequence Importance

Having trained a flexible network that could use any subset of pulse sequences, we sought to leverage this feature to probe changes in performance given different input pulse sequence combinations, as shown in Table V. By censoring each of the input pulse sequences individually, we found the largest degradations in performance when censoring the post-contrast BRAVO or the pre-contrast CUBE. Since the post-contrast BRAVO was used to create the annotations and all other sequences were registered to the BRAVO, this scan would be free from registration errors and would have the highest spatial correlation to expert annotations. We also think that the pre-contrast CUBE serves as a point of comparison with

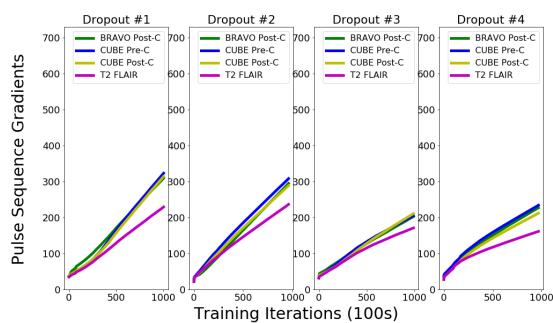


Fig. 8: Four Stochastically Different Training Runs of Dropout Networks.

respect to the post-contrast scans. On the other hand, the post-contrast CUBE is less important, likely because the BRAVO sequence, which is also a post-contrast image, captures similar information. Finally, missing the FLAIR has very little impact on the network's performance. The FLAIR is often used in clinical practice to indicate edema and other fluid as a larger marker for localizing lesions, but our annotations center on the core metastatic lesions, not edema. Thus, the FLAIR signal does not correlate with the primary annotations.

Additionally, we can look at the cumulative aggregate saliency for the input image, as described in Section IV-E. Figure 5 shows the input image cumulative aggregate saliency when separated by (a) pulse sequence and (b) z-slice location. Looking at the z-slice location, we see very predictable behavior, with the aggregate saliency being strongest when centered on the location of the annotation, followed by the $+1/-1$ slices, and then finally the $+2/-2$ slices. Repeating this analysis for the pulse sequences of a no-dropout model shows that the BRAVO post-contrast images have the largest gradients. However, the behavior of cumulative aggregate saliency can change between training runs with otherwise identical hyperparameters (Figure 7), although the BRAVO scan often eventually shows the largest gradients. Comparing Figures 6 and 8 (which visualize the models trained with input-level dropout), we observe that the dropout training creates a regular training pattern over time. When we look at the saliencies for different dropout runs, we notice the same pattern: all sequences other than the FLAIR train at approximately equal rates, while the FLAIR is always found to be the least correlated with the annotations. We believe that since our expert annotations did not segment edema portions of the metastasis, there was minimal correlation between the FLAIR images and our segmentation targets. In addition, adding dropout regularization creates a more consistent training behavior. This could be meaningful both scientifically and practically to create a consistent set of networks. However, this property could be detrimental if stochasticity in the networks is desired, such as in ensembling many networks.

D. Robustness to Inference of Multi-center Data

By testing our model on the Oslo data, we provide a real-world use case for training our models with input-level dropout, as these data do not have the same pulse sequences as that of our training data from Stanford. Standard practice involves training a separate model for every subset of pulse sequences, an exponentially difficult task with respect to number of pulse sequences. (For n pulse sequences, we would have to train $(2n-1)$ models.) However, in table VI, we show comparable performance on the Oslo data between the models trained with and without dropout.

We also notice a substantial boost in performance on the Oslo test set compared to the Stanford test set. We attribute this to the fact that the Oslo data had, on average, much larger metastasis lesions, which should be more easily detected. The Oslo data was also taken at a lower slice thickness, which gave more resolution for prediction.

VII. CONCLUSION

We set out to investigate best practices for integrating different pulse sequences of MR data using metastasis segmentation as a model task. We found that within the small data regime ($n=100$), models that had lower variance error performed better. Our best performing model was one that encoded in a subtraction between the learned-features from the pre- and post-contrast images. This leads us to believe that leveraging the clinical relationship between pulse sequences and mathematically encoding this knowledge in our network architecture could potentially boost network performance without requiring more data.

We also created and tested using an input dropout layer to train a single network that was robust to receiving any subset of the input pulse sequences (excluding the trivial case of an empty subset). With this method, the network not only preserves performance when given the full set of pulse sequences but also rivals the results of a model trained on the exact subset of pulse sequences when tested on that subset of pulse sequences. Using this dropout network, we were able to show that a single network could be trained on data from one institution (Stanford Hospital) and generalize to test data from another institution (Oslo University Hospital).

Finally, we ran network interpretation methods to better understand how our new dropout network performs in different circumstances. By running different subsets of pulse sequences through our network, we revealed which pulse sequences were most important in the decision-making process of our network. By visualizing the accumulated image gradients throughout training, we were able to show that the dropout network not only learned more evenly from the input pulse sequences but also gave more consistent training schedules. These results provide new insight on how CNNs can be most effectively applied to the multi-modal problem of medical images.

ACKNOWLEDGMENT

We'd like to thank both Stanford Hospital and Oslo University Hospital for providing the data needed to complete this study. We acknowledge the T15 LM 007033 NLM Training grant in funding this project. This work was also supported in part by grants from the National Cancer Institute, National Institutes of Health, U01CA142555, 1U01CA190214, 1U01CA187947 and U01CA242879.

REFERENCES

- [1] K. Bäckström, M. Nazari, I. Y.-H. Gu, and A. S. Jakola, "An efficient 3d deep convolutional network for alzheimer's disease diagnosis using mr images," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 149–153.
- [2] D. Lee, J. Yoo, and J. C. Ye, "Deep residual learning for compressed sensing mri," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017, pp. 15–18.
- [3] L.-L. Zeng, H. Wang, P. Hu, B. Yang, W. Pu, H. Shen, X. Chen, Z. Liu, H. Yin, Q. Tan *et al.*, "Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity mri," *EBioMedicine*, vol. 30, pp. 74–85, 2018.
- [4] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.
- [5] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, "Wasserstein auto-encoders," *arXiv preprint arXiv:1711.01558*, 2017.

- [6] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.
- [7] N. Wang and D.-Y. Yeung, "Learning a deep compact image representation for visual tracking," in *Advances in neural information processing systems*, 2013, pp. 809–817.
- [8] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, 2017, pp. 214–223.
- [9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [10] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2016, pp. 241–246.
- [11] E. Gong, J. M. Pauly, M. Wintermark, and G. Zaharchuk, "Deep learning enables reduced gadolinium dose for contrast-enhanced brain mri," *Journal of Magnetic Resonance Imaging*, vol. 48, no. 2, pp. 330–340, 2018.
- [12] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen, "Image reconstruction by domain-transform manifold learning," *Nature*, vol. 555, no. 7697, p. 487, 2018.
- [13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [14] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycski *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.
- [15] K. Kamnitsas, E. Ferrante, S. Parisot, C. Ledig, A. V. Nori, A. Criminisi, D. Rueckert, and B. Glocker, "Deepmedic for brain tumor segmentation," in *International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries*. Springer, 2016, pp. 138–149.
- [16] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 565–571.
- [17] Y. J. Bae, B. S. Choi, K. M. Lee, Y. H. Yoon, L. Sunwoo, C. Jung, and J. H. Kim, "Efficacy of maximum intensity projection of contrast-enhanced 3d turbo-spin echo imaging with improved motion-sensitized driven-equilibrium preparation in the detection of brain metastases," *Korean journal of radiology*, vol. 18, no. 4, pp. 699–709, 2017.
- [18] M. Majigsuren, T. Abe, T. Kageji, K. Matsuzaki, M. Takeuchi, S. Iwamoto, Y. Otomi, N. Uyama, S. Nagahiro, and M. Harada, "Comparison of brain tumor contrast-enhancement on t1-cube and 3d-spgr images," *Magnetic Resonance in Medical Sciences*, pp. 2014–0129, 2015.
- [19] M. Havaei, N. Guizard, H. Larochelle, and P.-M. Jodoin, "Deep learning trends for focal brain pathology segmentation in mri," in *Machine learning for health informatics*. Springer, 2016, pp. 125–148.
- [20] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "Hemis: Heteromodal image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 469–477.
- [21] R. Li, W. Zhang, H.-I. Suk, L. Wang, J. Li, D. Shen, and S. Ji, "Deep learning based imaging data completion for improved brain disease diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2014, pp. 305–312.
- [22] G. van Tulder and M. de Bruijne, "Why does synthesized data improve multi-sequence classification?" in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 531–538.
- [23] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [24] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
- [25] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.
- [26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
- [27] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.
- [28] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang *et al.*, "Evaluation of prostate segmentation algorithms for mri: the promise12 challenge," *Medical image analysis*, vol. 18, no. 2, pp. 359–373, 2014.
- [29] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, 2015.
- [30] K. Fritscher, P. Raudaschl, P. Zaffino, M. F. Spadea, G. C. Sharp, and R. Schubert, "Deep neural networks for fast segmentation of 3d medical images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 158–165.
- [31] Y. Xu, T. Géraud, and I. Bloch, "From neonatal to adult brain mr image segmentation in a few seconds using 3d-like fully convolutional network and transfer learning," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 4417–4421.
- [32] X. Zhao, Y. Wu, G. Song, Z. Li, Y. Zhang, and Y. Fan, "A deep learning model integrating fcnn and crfs for brain tumor segmentation," *Medical image analysis*, vol. 43, pp. 98–111, 2018.
- [33] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.
- [34] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *European conference on computer vision*. Springer, 2016, pp. 850–865.
- [35] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*. Springer, 2015, pp. 84–92.
- [36] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 681–687.
- [37] J. Ba and B. Frey, "Adaptive dropout for training deep neural networks," in *Advances in Neural Information Processing Systems*, 2013, pp. 3084–3092.
- [38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [39] E. A. Smirnov, D. M. Timoshenko, and S. N. Andrianov, "Comparison of regularization methods for imagenet classification with deep convolutional neural networks," *Aasri Procedia*, vol. 6, pp. 89–94, 2014.
- [40] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.
- [41] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*, "CheXnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.
- [42] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball *et al.*, "Mura: Large dataset for abnormality detection in musculoskeletal radiographs," *arXiv preprint arXiv:1712.06957*, 2017.
- [43] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *International conference on machine learning*, 2015, pp. 597–606.
- [44] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.
- [45] E. Grøvik, D. Yi, M. Iv, E. Tong, D. Rubin, and G. Zaharchuk, "Deep learning enables automatic detection and segmentation of brain metastases on multisequence mri," *Journal of Magnetic Resonance Imaging*, 2019.