

Multi-Modal Learning from Unpaired Images: Application to Multi-Organ Segmentation in CT and MRI

Vanya V. Valindria¹, Nick Pawlowski¹, Martin Rajchl¹, Ioannis Lavdas¹, Eric O. Aboagye¹,
Andrea G. Rockall², Daniel Rueckert¹, and Ben Glocker¹

¹Imperial College London

²The Royal Marsden NHS Foundation Trust

v.valindria15@imperial.ac.uk

Abstract

Convolutional neural networks have been widely used in medical image segmentation. The amount of training data strongly determines the overall performance. Most approaches are applied for a single imaging modality, e.g., brain MRI. In practice, it is often difficult to acquire sufficient training data of a certain imaging modality. The same anatomical structures, however, may be visible in different modalities such as major organs on abdominal CT and MRI. In this work, we investigate the effectiveness of learning from multiple modalities to improve the segmentation accuracy on each individual modality. We study the feasibility of using a dual-stream encoder-decoder architecture to learn modality-independent, and thus, generalisable and robust features. All of our MRI and CT data are unpaired, which means they are obtained from different subjects and not registered to each other. Experiments show that multi-modal learning can improve overall accuracy over modality-specific training. Results demonstrate that information across modalities can in particular improve performance on varying structures such as the spleen.

1. Introduction

In clinical practice, multiple imaging modalities are used to capture anatomical structures such as major abdominal organs. For example, CT and MRI both give clear images of organs such as the liver, the heart, or the kidneys, however, the visual appearance is completely different due to the very different underlying physical principles of each imaging technique. For the human eye, however, it is quite easy to identify the same structures in different modalities. For example, after showing an abdominal CT scan and pointing out where the liver is, even a non-expert will be able to find

the liver in an MRI scan. For algorithms trained to segment livers in CT, however, it is quasi impossible to segment the livers in MRI due to the different nature of the imaging features that are learned in a modality-specific setting. In this paper, we explore the power of multi-modal learning to extract abstract representations that are modality-independent and have the potential to increase accuracy and robustness when employed as features in a supervised setting.

Deep-learning based approaches such as convolutional neural networks (CNNs) have become popular in medical imaging but they require large amounts of data to achieve good results, near human-level capability [3, 22]. However, in practice, it is often difficult to acquire sufficiently many manually annotated datasets from one modality. Hence, there is a strong desire to utilise all available data for the same or similar tasks even if the images are from different modalities. The main question is whether the information from one modality can actually improve the performance of a task on another modality [4]. We aim to answer this question through an extensive set of experiments with novel CNN architectures for multi-modal learning.

In computer vision, previous works have shown that multi modal learning often provides better performance [31, 30, 27]. In general, multi-modal learning is a more natural way to learn, as humans perceive information through various modalities, such as vision and sound [26]. However, capturing the correspondence between modalities and inferring meaningful information that can be shared between tasks are the main challenges in a multi-modal setting.

In medical imaging, most previous work explores multi-contrast segmentation using a single imaging modality: MRI - with different imaging sequences which can be seen as different modalities [13, 9, 7, 16]. All previous works on multi-modal segmentation use *paired* data, meaning that the images are acquired from the same subject and co-

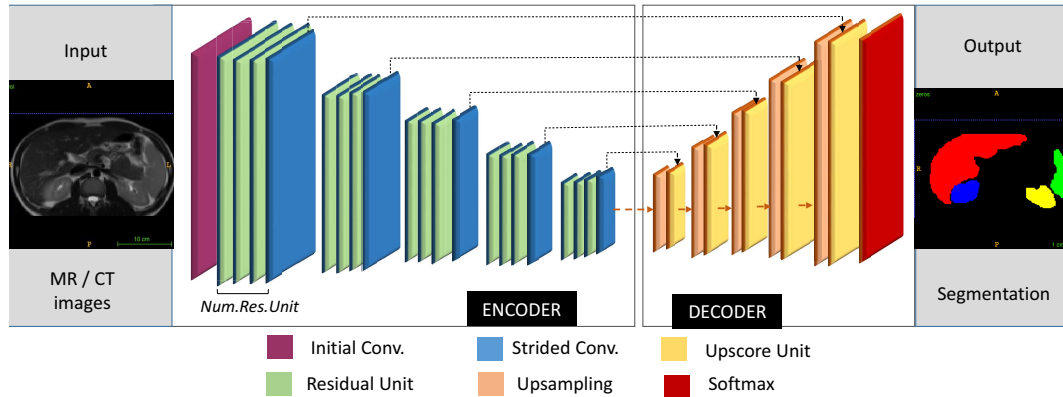


Figure 1. Baseline network architecture for multi-organ segmentation: FCN based network with encoder-decoder structure

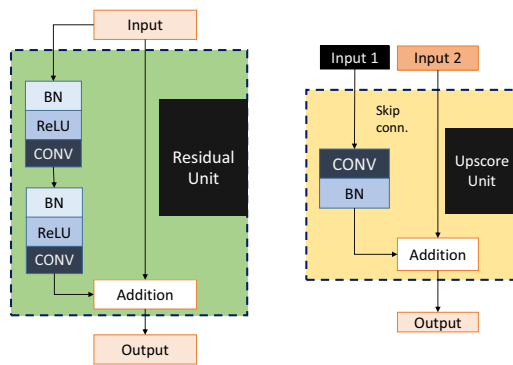


Figure 2. Left: Residual unit with pre-activation. Right: Upscore Unit. Input 1: Higher resolution features from the encoder (same scale) to add. Input 2: Input features from decoder to be upscored.

registered across modalities. In our case, all of the data are *unpaired*, the acquisitions are from different subjects and images are not registered. Hence, to the best of our knowledge, learning across unpaired data from different imaging modalities has not been explored yet in medical imaging.

We investigate multi-modal learning for multi-organ segmentation in CT and MRI. We set up different network architectures and multi-modal training schemes to evaluate the effect on segmentation accuracy for each modality.

The contributions of this paper are as follows:

- We learn modality-independent features from CT and MRI using an encoder-decoder network architecture
- We explore various dual-stream networks with different strategies for sharing information between tasks
- We show improved segmentation accuracy of varying organs when using multi-modal learning

2. Related work

Multi-modal learning has been widely studied for various applications in machine learning and computer vision. One of the challenges is the need to find a common shared representation from different modalities. Previous work on multi-modal data has shown the benefit of shared latent representations for generative tasks [31, 27, 20], with most approaches being based on Canonical Correlation Analysis (CCA) [2] and Autoencoders (AEs) [20].

One of the early works in multi-modal learning allows a bimodal deep AE to reconstruct both modalities (audio and video data of speech), even if only one modality is present [20]. Although AE-based models are good for self-and-cross reconstruction, they do not guarantee correlated common representations [19]. In attempt to address that issue, [5] proposes Correlational Neural Networks that combine the advantage of CCA to encourage correlation in the common learned representations.

Most works in medical image segmentation refer to the term *multi-modality* when using different sequences in MRI, such as T1-weighted, T2-weighted, diffusion, or functional MRI. The contrast and information in each sequence is different. There are varying approaches to these multi-modal segmentation problems, even though the most common approach is by concatenating the inputs from different modalities as different input channels to the networks [8, 13, 16]. Another option is to train one CNN for each modality and fusing multi-modality features from high-layers of each network [21]. High-level representations from different modalities are complementary, and might yield improvement in performance. In multi-modal brain MRI segmentation, cross-modality convolution with a deep encoder-decoder architecture has been explored by [29] with convolutional LSTM to model the sequential correlations between slices. They show that cross-modality convolution can effectively aggregate the information between modalities to produce better results [29]. These works as-

sume that all modalities are always available for each individual subject.

The Hetero-modal network architecture (HeMIS) [9] and scalable multi-modal CNN (ScaleNet) [7] are built for multi-modal brain MRI segmentation that can deal with missing modalities, however, they assume modalities are from the same subjects and co-registered. In our case, CT and MRI are unpaired, i.e., one modality per subject.

Transfer learning is one way to share knowledge between different domains. Fine-tuning a pre-trained model from natural images is shown to be useful for medical image analysis, even though the source and target domains are very different [28]. Transfer learning between CT and MRI has been explored in [32]. They showed that the accuracy of MRI segmentation can be improved by using shape priors from CT. In deep learning based segmentation, learning different modalities needs different parameters in order to work well on a particular network architecture [17]. In general, it appears that training each modality separately has so far led to best results, which might be due to the use of non-optimal architectures.

In medical imaging, training multi-task on one network had been conducted by [18]. They showed that the performance of learning multi-tasks and multi-modalities on a single CNN is *equivalent* to a single network trained specifically for one-task and one-modality. Instead of using a single network for multi-task learning, [14] applied a multi-modal encoder-decoder network with shared latent space and shared skip connections. Their results show a potential of shared representations from different modalities to improve the multi-task performance.

Multi-modal learning can also be applied to image synthesis, where a joint representation in a common latent space preserves the local consistency of the images [11]. Using this representation, cross-modality synthesis in MRI (T2 to T1, FLAIR to T2 and vice versa) can be achieved in high-resolution. Multi-modal image synthesis of MRI is explored in [12, 6] by composing the networks of encoders (for each modality), fusing latent representations, and a decoder to produce an output image.

Our work exploits the shared representation in multi-modal learning, particularly in unpaired CT and MR images, which has not been investigated before.

3. Multi-modal learning from unpaired images

Previous multi-modal segmentation approaches [9, 7] learn to fuse the information from modalities and to handle missing modalities. Thus, pairs of n -modalities have to be available at least once in training time. This approach is not feasible when there are only unpaired images from different modalities. Our challenge is to build a model which can take *unpaired* inputs from n -image modalities to produce an accurate segmentation for each modality.

3.1. Network architectures

As a baseline, we use a fully convolutional network (FCN) [25] with residual layers [10] for multi-organ image segmentation (see Figure 1). The network can be seen as an encoder-decoder architecture.

Encoder. The network consists of a number of residual feature encoding blocks with pre-activation [10], as shown in Figure 2. We specify the number of residual units in each scale, which is chosen experimentally. The residual unit block consist of batch normalisation (*BN*), activation function (*ReLU*), followed by 3D convolutional layers (*Conv*) for extracting image features. To handle the stride convolution, we add pooling to the input before the addition in residual unit. Then, we use padding if the number of filters are different between input and output. Before moving on to the next scale, the features are downsampled via strided convolutions. We use kernel size 3^3 with stride 1 and padding size 1 for all convolutional layer, and kernel size twice the stride for strided convolution.

Shared representation. Multi-modal learning focuses on gaining shared representations from multi-modal data. Since our data is unpaired, straightforward concatenation of extracted features from different modalities is not reasonable. Moreover, at test time we want to be able to perform segmentation on a single modality acquired for a single subject. We aim to improve the shared representation by using an encoder-decoder architecture, to investigate which part of encoder or decoder should be shared across modalities.

Decoder. The decoding stage uses fully convolutional layers to map a target output modality. Feature maps learnt at different scales are upsampled to the original resolution. The decoder consists of upscore units as depicted in Figure 2 with kernel 1^3 . The input is upsampled before going into the upscore unit. Inside the upscore unit, the features from the encoder are learnt to produce a sparse feature map. We apply skip connections, with the 3D convolution of the encoder features followed by batch normalisation [10]. This is added to the input features to be upscored.

Finally, the rescaled output is fed into a softmax layer to produce the probabilistic label map. The highest softmax probability for each class is computed to yield the final segmentation result.

3.2. Dual-stream

To effectively infer the multi-modal features from CT and MRI, we propose a dual-stream network architecture. By using individual streams for each modality, we can investigate when it is best to merge or split the streams. Dual-stream also allows to process *unpaired* images in one network, which is not possible with standard multi-channel architectures [13, 9] for *paired* data. The images from unpaired multi-modal data have no correspondences which makes it difficult to find correlations across modalities.

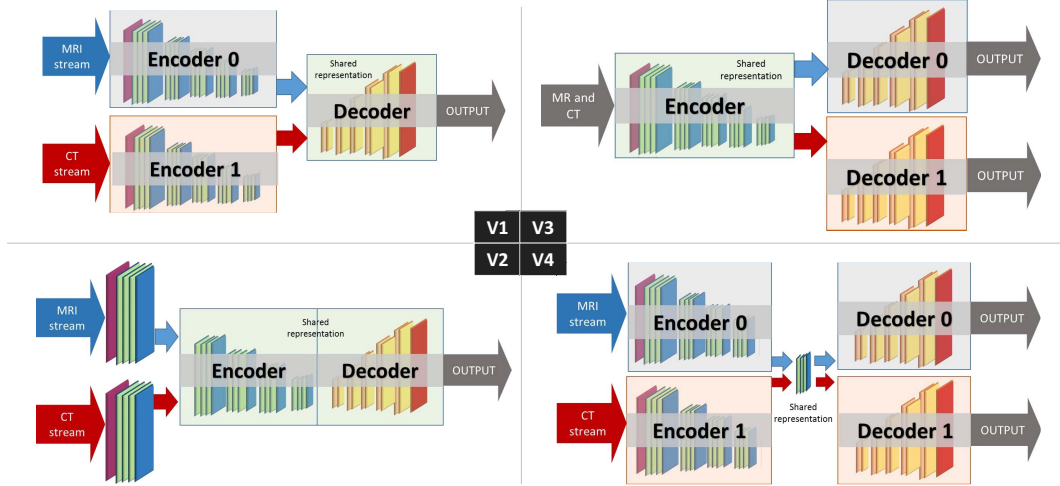


Figure 3. Dual-stream models. Version 1 - modality-specific encoder and shared decoder. Version 2 - shared part-of-encoder and decoder. Version 3 - shared encoder and modality-specific decoder. Version 4 - different streams in encoder, shared last layers of encoder, and modality-specific decoder.

To exploit the commonality among different modalities, streams of encoder/decoder are connected with each other via the shared latent representation.

For the dual-stream model, we also implement an architecture with skip connections to propagate features from encoders to decoders. The FCN-based architecture is the most suitable choice for dual-streams as the features can be merged from different scales in the encoder which vary in semantic information. Implementing the dual-stream model is slightly more complex in U-Net [24] and DeepMedic architectures [13]. For U-Net, the encoder feature maps need to be *concatenated* with the upsampled feature maps from the decoder at every scale. For DeepMedic, the multi-scale nature of the network needs to be taken into account. Hence, for this paper we opt to use an FCN-based model in all our experiments.

4. Experimental Setup

In this section, we evaluate the multi-modal encoder-decoder networks for multi-organ segmentation using two datasets: MRI and CT data. The implementation details are below.

4.1. Data and pre-processing

We use two datasets of 3D abdominal images from MRI and CT [15] to evaluate the models. We train the model with a roughly balanced dataset: 34 subjects for MRI and 30 subjects for CT to keep it agnostic to the modality intensity distribution. For both CT and MRI data we split the data into 2-folds for cross-validation. From the MRI database we used T2-weighted images. Both CT and MRI scans have the same four organs manually annotated: liver, spleen, right

kidney, and left kidney delineated by clinical experts. All of the CT and MRI data are unpaired, which means that the acquisitions are from different subjects. All experiments were carried out using NVIDIA GPUs.

We perform all experiments on resampled data with 2 mm voxel spacing. The original abdominal scans of MRI and CT capture the area from neck to knee, with varying image sizes according to patient body size. In order to have similar FOV (Field of View) between all of the images, we crop the volumes so that they only cover all four organs to be segmented - removing uninformative areas. We take specific preprocessing measures for CT images to account for the differences in intensity distribution. We clip all values lower than -1000, before zero-mean unit-variance intensity normalisation. We add Gaussian noise and intensity offsets to the images for data augmentation, as a common approach in training the deep networks. Then, we use weighted class sampling during training to compensate for class-imbalance.

4.2. Training Details

We implemented the model using DLTK (Deep Learning Toolkit) [23], with a Tensorflow backend [1]. We use the DLTK FCN segmentation model as a baseline network, which has been optimised for abdominal multi-organ segmentation. We experiment with both cross-entropy and Dice-loss, and choose the Dice-loss for the better training stability. Let $o_i \in [0, 1]$ be the i the output of the network and let $y_i \in [0, 1]$ be the corresponding label. The Dice-loss is then defined as follows:

$$L_{Dice} = \frac{2 \sum_i o_i y_i}{\sum_i o_i + \sum_i y_i}$$

The model is trained using Adam optimisation with a learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, $\epsilon = 10^{-5}$. We use mini-batch training of 16 training examples with size 64^3 , to provide enough context at 2 mm resolution. The number of filters for each scale is 16, 64, 128, 256, and 512. Training was run for 10k steps. The training examples were shuffled from all training images. However, for multi-modal learning, training examples are drawn in an alternating way: one iteration contains only MRI data, and the subsequent iteration contains only CT data. Since we use alternating batches from modalities in training, during testing procedure we can simply feed single modality data for inference.

Evaluation of segmentation performance.

We repeated the training and testing procedures for all of the models three times, to average out random variations. The results presented will be the total of all experiments and the average accuracies from all scores. Results are evaluated using the Dice score for the four individual abdominal organs: liver, spleen, right kidney, and left kidney.

4.3. Benchmark

To demonstrate the benefit of multi-modal learning, we extensively compare several different approaches and baselines as detailed below.

4.3.1 Baseline multi-organ segmentation model

The segmentation model in Figure 1 was used for modality-specific segmentation. Here, we trained the network on MRI data and only tested on MRI data, and the same procedure was done for CT data. The number of residual units is 3 per scale and the learning rate is 0.001 - chosen via experimentation. As a baseline, one network was trained separately for each imaging modality.

4.3.2 Joint learning

The same model as in Figure 1 with the same parameters, network capacity, and optimisation was trained for a combination of CT and MR data. The only difference in this joint learning is the data, we trained a single network with input from *both* CT and MRI data to perform the segmentation task. Additionally, we use alternating batches in training phase, one iteration for MRI batch and the next iteration fetched the CT batch. Hence, the network can learn from balanced examples from both modalities.

4.3.3 Dual-stream

Dual-stream models are built to handle multi-modal segmentation in one network, with one stream per modality. If the input is from MRI data then stream 0 will be chosen, while stream 1 will be activated if the network has CT input

Table 1. Multi-modal learning in MR images

Organ	Individual	Joint	V1	V2	V3	V4
Liver	0.913	0.903	0.877	0.862	0.905	0.914
Spleen	0.772	0.772	0.728	0.732	0.736	0.790
Right Kidney	0.835	0.842	0.804	0.792	0.819	0.871
Left Kidney	0.821	0.771	0.748	0.741	0.810	0.833

Table 2. Multi-modal learning in CT images

Organ	Individual	Joint	V1	V2	V3	V4
Liver	0.914	0.875	0.900	0.902	0.895	0.919
Spleen	0.824	0.834	0.822	0.824	0.817	0.859
Right Kidney	0.820	0.806	0.824	0.815	0.793	0.838
Left Kidney	0.808	0.822	0.833	0.818	0.819	0.851

data. In the training phase, we also apply alternating batches as in joint learning, and the loss function is back-propagated per modality. The optimisation in all dual-stream networks are the same to provide the lowest error and a fair comparison. We still use the same hyper-parameter settings in the baseline model, with a lower learning rate 0.0001. Different dual-stream versions are illustrated in Figure 3.

Version 1 (V1). This model has a separate encoder for each modality. MRI data will go to encoder 0 and CT data will go into encoder 1, and share the same decoder. As in [9] and [7], this type of independent, modality-specific in the back-end and fusing in the front-end model is shown to work with multi-modal segmentation. The decoding task in this model is expected to benefit from high-resolution representations, as the decoder is connected from different encoders.

Version 2 (V2). Both modalities share most part of encoder, and decoder. A separate stream for CT and MRI is only applied in the input and residual units at the first scale. Then, they share the same encoder and decoder part.

Version 3 (V3). CT and MRI share the same encoder but have different streams in decoding part. By having both modalities share the same encoder, they expect to generate a single shared representation to discover correlation across the modalities, which is then modality-specific decoded. This model is similar to [20] when the model only has one modality to reconstruct both modalities. In our case, the segmentation output is for each modality.

Version 4 (V4). Inspired by one of the architectures in bi-modal deep AE [20] and multi-modal encoder-decoder networks in [14], we use dual-stream in both encoder and decoder. Both MR and CT flows into a different stream in the encoding part, but only share weights at the last scale of encoder. These layers can be seen as a shared latent representation which is then fed into a modality-specific decoder. In other words, all encoder/decoder streams are connected with the shared latent representation.

Ground truth	Individual MR	CTMR-joint	Dualstream V.1	Dualstream V.2	Dualstream V.3	Dualstream V.4
Liver	0.905	0.906	0.902	0.892	0.915	0.926
Spleen	0.856	0.864	0.845	0.852	0.866	0.893

Ground truth	Individual MR	CTMR-joint	Dualstream V.1	Dualstream V.2	Dualstream V.3	Dualstream V.4
Right Kidney	0.584	0.338	0.568	0.326	0.664	0.762
Left Kidney	0.713	0.329	0.481	0.383	0.758	0.761

Ground truth	Individual MR	CTMR-joint	Dualstream V.1	Dualstream V.2	Dualstream V.3	Dualstream V.4
Liver	0.899	0.854	0.811	0.836	0.864	0.907
Spleen	0.750	0.734	0.810	0.786	0.704	0.840
Right Kidney	0.729	0.900	0.891	0.887	0.865	0.921
Left Kidney	0.822	0.878	0.908	0.857	0.839	0.923

Figure 4. Multi-modal learning on MRI. Liver (red), spleen (green), right kidney (blue), left kidney (yellow).

Table 3. Overview on multi-modal learning on all organs

All organs	Individual	Joint	V1	V2	V3	V4
Average	0.838	0.828	0.817	0.811	0.824	0.860
Std.Deviation	0.019	0.021	0.059	0.065	0.034	0.011

5. Results and Discussion

Multi-modal learning can leverage the shared information in both modalities in one pass, unlike the traditional transfer learning with the sequential training (initial training and fine-tuning the pretrained networks). Additionally, from our experiments, we found that fine-tuning did not improve over models trained jointly and individually.

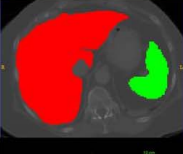
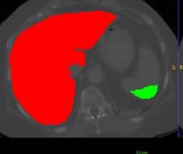
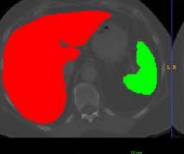
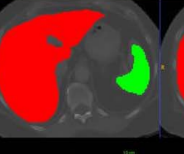
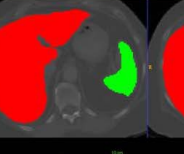
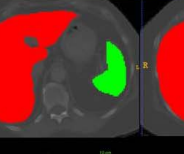
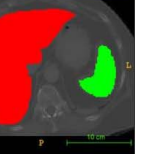
We compare the performance of multi-modal learning (joint and dual-stream versions) to individual learning in MR and CT data. Because we ran the benchmark procedures (training and testing in 2-fold cross-validation for each modality) in three repetitions for each model, we present the results in Table 1, 2, 3 as the average scores from all experiments.


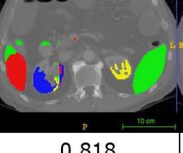
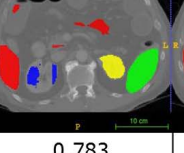
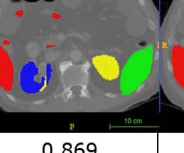
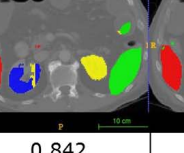
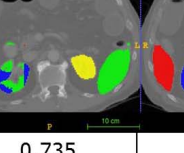
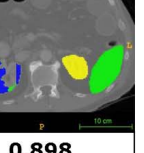
We report how the multi-modal learning affects the Dice scores of each organ segmentation for different modalities. For MRI, in Table 1, we can observe that multi-modal learn-

ing by dual-stream V4 generally improves the performance for all organs, compared to individual segmentation. From Table 1, a better accuracy is achieved in spleen and left kidney segmentation. Note that there is a significant improvement in right kidney segmentation, by four Dice scores. Joint CT-MR also gives a slight increase in the right kidney accuracy.

Table 2 shows the benefit of multi-modal learning in CT data with dual-stream V4. It is interesting to note that the multi-modal learning is more helpful to segment organs in CT (than in MR), as shown for left kidney where all multi-modal models (joint and dual-stream models) improve the individual performance. A high rise of Dice scores is shown in small organs, especially spleen and left kidney. Joint CT-MR also gives slightly better accuracy in spleen segmentation. In both modalities, the accuracy of liver segmentation with multi-modal learning is consistent, as individually trained segmentation already gives a high accuracy.

In terms of overall performance, we report the results for all organ segmentation of benchmark models in Table 3. Dual-stream V4 outperforms the individual training with almost three Dice scores increment. However, other multi-modal learning models fail to improve the segmentation of

Ground truth	Individual CT	CTMR-joint	Dualstream V.1	Dualstream V.2	Dualstream V.3	Dualstream V.4
						
Spleen	0.384	0.891	0.846	0.836	0.904	0.907

Ground truth	Individual CT	CTMR-joint	Dualstream V.1	Dualstream V.2	Dualstream V.3	Dualstream V.4
						
Right Kidney	0.818	0.783	0.869	0.842	0.735	0.898
Left Kidney	0.723	0.867	0.912	0.855	0.893	0.937

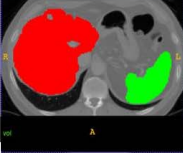
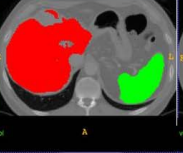
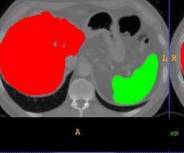

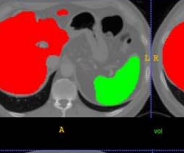
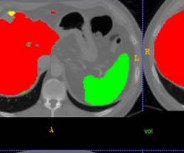

Ground truth	Individual CT	CTMR-joint	Dualstream V.1	Dualstream V.2	Dualstream V.3	Dualstream V.4
						
Liver	0.931	0.918	0.931	0.934	0.936	0.943
Spleen	0.942	0.908	0.940	0.934	0.930	0.951
Right Kidney	0.935	0.921	0.928	0.936	0.932	0.939
Left Kidney	0.743	0.902	0.919	0.916	0.918	0.935

Figure 5. Multi-modal learning on CT. Liver (red), spleen (green), right kidney (blue), left kidney (yellow).

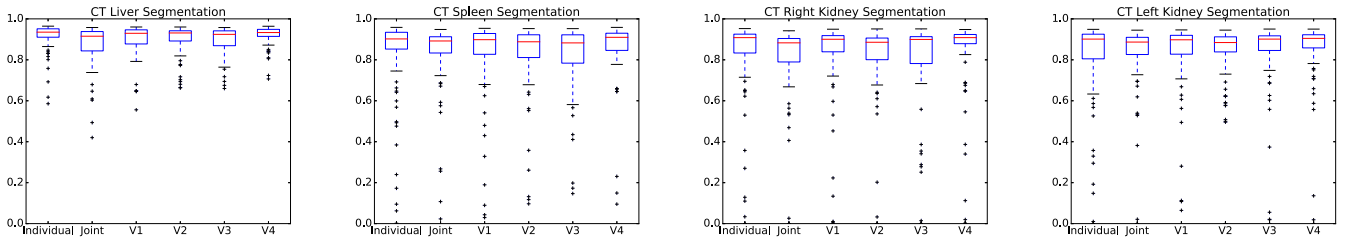


Figure 6. Boxplot of different organ segmentations on CT data: liver, spleen, right kidney, and left kidney segmentation (left-right) Different models on x-axis and Dice scores on y-axis.

individual learning. Despite of its lower accuracy than dual-stream V4, joint CT-MR can still train a single neural network to do quite well on both modalities. It is because the network effectively compresses the information of both databases into a model that usually fits one modality and

improves the performance. We did further experiments to make the network's capacity larger, but we did not obtain better results in joint CT-MR.

From Table 3, a shared decoder (as in dual-stream V1 and V2) cannot effectively leverage the multi-modal fea-

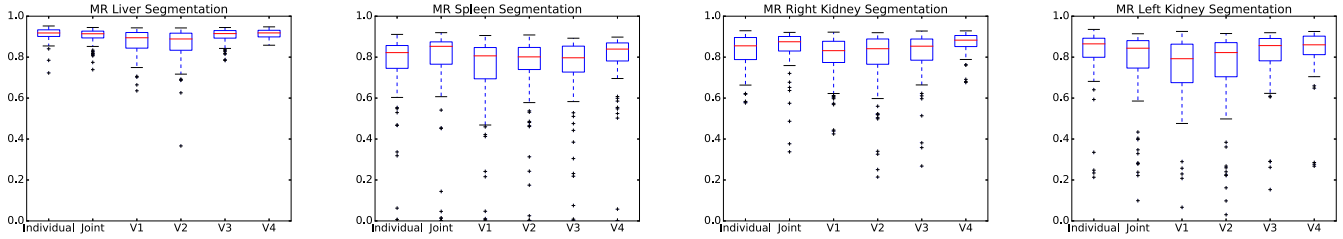


Figure 7. Boxplot of different organ segmentations on MRI data: liver, spleen, right kidney, and left kidney segmentation (left-right). Different models on x-axis and Dice scores on y-axis.

tures for segmentation. In dual-stream V1, we tried to combine the high-level features from both modalities in a shared decoding phase. In dual-stream V2, we only extract low-level features in different streams, to be encoded and decoded together. In contrast to shared-decoder in *paired* multi-modal data [7, 9], this model cannot learn effectively in *unpaired* data. Dual-stream V1 and V2 are forced to have the same decoding phase. Meanwhile, dual-stream V3 and V4 give better performance than V1 and V2 because of their modality-specific decoders. It is interesting to note that different streams in upscoring could increase the multi-modal learning capability, as shown in [20, 14]. However, in dual-stream V3, the encoders are tied between modalities, making it harder to leverage the multi-modal features. As a result, dual-stream V4 with split streams of encoder, shared latent representation (last layers on encoder), and modality-specific decoder gives overall best performance.

By visual inspection, we can see how multi-modal learning can improve multi-organ segmentation. For example in MRI data in Figure 4 (top), multi-modal learning can reduce the false positive of liver segmentation and increase spleen accuracy. In Figure 4 (middle), both kidneys give better accuracies in dual-stream V4. Overall organ segmentation in Figure 4 (bottom) is improved by dual-stream V4.

We also visualise how multi-modal learning can help the segmentation in CT. In Figure 5 (top), we can see how all multi-modal learning (CT-MR joint, dual-stream V1, V2, V3 and V4) can help spleen segmentation. Dual-stream V4 segmentation of left kidney has a significant increase in Dice scores (Figure 5-middle), compared to individually trained segmentation. Multi-organ segmentation is improved by dual-stream V4 in Figure 5 (bottom).

Box-plots in Figure 6 and 7 from all experiments illustrate the benefit of multi-modal learning. Dual-stream V4 gives an improved segmentation performance compared to individual learning, particularly in kidneys and spleen. In dual-stream V4, the outliers are reduced as well as the variance on Dice scores, for both modalities.

From [18] we learn that multi-task segmentation on a single network performs similarly on multi-modal data. From our experiments, we believe that multi-modal learning

in a single task (multi-organ segmentation) can benefit from sharing information, even though the data are *unpaired*.

6. Conclusion

In this paper, we investigate the benefit of multi-modal learning on unpaired multi-modal CT and MRI segmentation. We present a comparative study of different multi-modal training schemes to better exploit the sharing of information. A novel dual-stream network architecture was introduced for that purpose. By multi-modal learning, shared representation on both modalities can help the network solving the same task with limited training data. Experimental results on both MR and CT demonstrate improved state-of-the-art segmentation accuracies, especially on varying organs such as spleen and kidneys. The power of learning shared representations from different datasets appears as a promising direction for future work.

Acknowledgments

Vanya Valindria is supported by the Indonesia Endowment for Education (LPDP) - Indonesia Presidential PhD Scholarship programme. Nick Pawlowski is supported by a Microsoft Research PhD Scholarship and the EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, Grant Reference EP/L016796/1). Martin Rajchl is supported by an Imperial College Research Fellowship. The authors thank NVIDIA for the donation of two Titan X GPUs for our research.

The MRI data has been collected as part of the MALIBO project funded by the Efficacy and Mechanism Evaluation (EME) Programme, an MRC and NIHR partnership (EME project 13/122/01). The views expressed in this publication are those of the author(s) and not necessarily those of the MRC, NHS, NIHR or the Department of Health.

References

- [1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia,

- R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [2] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.
 - [3] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vailant, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi, et al. Human-level cmr image analysis with deep fully convolutional networks. *arXiv preprint arXiv:1710.09289*, 2017.
 - [4] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *CVPR 2016 Tutorial*, 2017.
 - [5] S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran. Correlational neural networks. *Neural computation*, 2016.
 - [6] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsaftaris. Multimodal mr synthesis via modality-invariant latent representation. *IEEE Transactions on Medical Imaging*, 2017.
 - [7] L. Fidon, W. Li, L. C. Garcia-Peraza-Herrera, J. Ekanayake, N. Kitchen, S. Ourselin, and T. Vercauteren. Scalable multimodal convolutional networks for brain tumour segmentation. In *MICCAI*, pages 285–293. Springer, 2017.
 - [8] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.
 - [9] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio. Hemis: Hetero-modal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 469–477. Springer, 2016.
 - [10] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016.
 - [11] Y. Huang, L. Shao, and A. F. Frangi. Simultaneous super-resolution and cross-modality synthesis of 3d medical images using weakly-supervised joint convolutional sparse coding. *CVPR*, 2017.
 - [12] T. Joyce, A. Chartsias, and S. A. Tsaftaris. Robust multimodal mr image synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 347–355. Springer, 2017.
 - [13] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.
 - [14] R. Kuga, A. Kanezaki, M. Samejima, Y. Sugano, and Y. Matsushita. Multi-task learning using multi-modal encoder-decoder networks with shared skip connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–411, 2017.
 - [15] B. Landman, Z. Xu, J. E. Iglesias, M. Styner, T. R. Langerak, and A. Klein. Multi-atlas labeling beyond the cranial vault. <https://www.synapse.org/Synapse:syn3193805/wiki/217789>. doi:10.7303/syn3193805. Accessed June 2017.
 - [16] I. Lavdas, B. Glocker, K. Kamnitsas, D. Rueckert, H. Mair, A. Sandhu, S. A. Taylor, E. O. Aboagye, and A. G. Rockall. Fully automatic, multi-organ segmentation in normal whole body magnetic resonance imaging (mri), using classification forests (cfs), convolutional neural networks (cnns) and a multi-atlas (ma) approach. *Medical Physics*, 2017.
 - [17] F. Milletari, S.-A. Ahmadi, C. Kroll, A. Plate, V. Rozanski, J. Maiostre, J. Levin, O. Dietrich, B. Ertl-Wagner, K. Bötzel, et al. Hough-cnn: Deep learning for segmentation of deep brain regions in mri and ultrasound. *Computer Vision and Image Understanding*, 2017.
 - [18] P. Moeskops, J. M. Wolterink, B. H. van der Velden, K. G. Gilhuijs, T. Leiner, M. A. Viergever, and I. Išgum. Deep learning for multi-task medical image segmentation in multiple modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 478–486. Springer, 2016.
 - [19] T. Mukherjee, M. Yamada, and T. M. Hospedales. Deep matching autoencoders. *arXiv preprint arXiv:1711.06047*, 2017.
 - [20] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.
 - [21] D. Nie, L. Wang, Y. Gao, and D. Sken. Fully convolutional networks for multi-modality isointense infant brain image segmentation. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pages 1342–1345. IEEE, 2016.
 - [22] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, R. Guerrero, S. Cook, A. de Marvao, D. O’Regan, et al. Anatomically constrained neural networks (acnn): Application to cardiac image enhancement and segmentation. *IEEE Transaction of Medical Imaging*, 2017.
 - [23] N. Pawłowski, S. I. Ktena, M. C. Lee, B. Kainz, D. Rueckert, B. Glocker, and M. Rajchl. Dltk: State of the art reference implementations for deep learning on medical images. In *Medical Imaging meet NIPS Workshop*, 2017.
 - [24] 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*, pages 234–241. Springer, 2015.
 - [25] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):640–651, 2017.
 - [26] R. Socher and L. Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 966–973. IEEE, 2010.
 - [27] M. Suzuki, K. Nakayama, and Y. Matsuo. Joint multimodal learning with deep generative models. *ICLR 2017 workshop*, 2017.

- [28] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.
- [29] K.-L. Tseng, Y.-L. Lin, W. Hsu, and C.-Y. Huang. Joint sequence learning and cross-modality convolution for 3d biomedical segmentation. *CVPR*, 2017.
- [30] X. Wang, G. Oxholm, D. Zhang, and Y.-F. Wang. Multi-modal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. *CVPR*, 2017.
- [31] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. A. Bernal, and J. Luo. Deep multimodal representation learning from temporal data. *CVPR*, 2017.
- [32] Y. Zheng. Cross-modality medical image detection and segmentation by transfer learning of shapel priors. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 424–427. IEEE, 2015.