IVD-Net: Intervertebral Disc Localization and Segmentation in MRI with a Multi-modal UNet

Jose Dolz[✉], Christian Desrosiers, and Ismail Ben Ayed

ETS Montreal, Montreal, Canada
{jose.dolz,christian.desrosiers,ismail.benayed}@etsmtl.ca

Accurate localization and segmentation of intervertebral disc (IVD) is crucial for the assessment of spine disease diagnosis. Despite the technological advances in medical imaging, IVD localization and segmentation are still manually performed, which is time-consuming and prone to errors. If, in addition, multi-modal imaging is considered, the burden imposed on disease assessments increases substantially. In this paper, we propose an architecture for IVD localization and segmentation in multi-modal magnetic resonance images (MRI), which extends the well-known UNet. Compared to single images, multi-modal data brings complementary information, contributing to better data representation and discriminative power. Our contributions are three-fold. First, how to effectively integrate and fully leverage multi-modal data remains almost unexplored. In this work, each MRI modality is processed in a different path to better exploit their unique information. Second, inspired by HyperDenseNet [11], the network is densely-connected both within each path and across different paths, granting the model the freedom to learn *where* and *how* the different modalities should be processed and combined. Third, we improved standard U-Net modules by extending inception modules [22] with two convolutional blocks with dilated convolutions of different scale, which helps handling multi-scale context. We report experiments over the data set of the public MICCAI 2018 Challenge on Automatic Intervertebral Disc Localization and Segmentation, with 13 multi-modal MRI images used for training and 3 for validation. We trained IVD-Net on an NVidia TITAN XP GPU with 16 GBs RAM, using ADAM as optimizer and a learning rate of 1×10^{-5} during 200 epochs. Training took about 5 h, and segmentation of a whole volume about 2–3 s, on average. Several baselines, with different multi-modal fusion strategies, were used to demonstrate the effectiveness of the proposed architecture.

1 Introduction

Intervertebral disc (IVD) degeneration [1] is one of the main causes for chronic low back pain (LBP), which has become a major public health problem in our society and a leading cause of function incapacity [24]. Magnetic resonance imaging (MRI) is the preferred modality to evaluate lumbar degenerative disc disease because it offers a good soft tissue contrast without ionizing radiation [12].

Advances in multi-modal MRI have increased the quality of diagnosis, treatment and follow-up in many diseases. However this comes at the cost of an increased amount of data, imposing a burden on disease assessments. Visual inspections of such an enormous amount of medical images are prohibitively time-consuming, prone to errors and unsuitable for large-scale studies. Developing robust methods for automatic IVD localization and segmentation from multi-modal MRI is thus essential for the diagnosis and treatment of spine pathologies. Having such methods could also reduce the manual work required by clinicians, and provide a faster and more consistent diagnosis.

Over the years, various semi-automated and automated techniques have been proposed for IVD localization and segmentation [2,4]. Recently, deep convolutional neural networks (CNNs) have shown outstanding performance for this task, outperforming previous segmentation approaches [5,14,16,27,31]. For example, Ji et al. [14] proposed a standard CNN for IVD segmentation, where the inference was performed pixel-wise by extracting a patch around each pixel. In addition, the authors evaluated different patch strategies, such as 2D or 2.5D patches, as well as the impact of vicinity size. More recently, a deeply supervised multi-scale fully CNN was proposed in [27] for the segmentation of IVDs in MR-T2 weighted images. An interesting feature of this work is its use of multi-scale deep supervision in the architecture, which alleviates the risk of vanishing gradient during training. Despite achieving satisfactory results, these works have mostly focused on single-modality scenarios.

Integrating multi-modal images in deep learning segmentation methods has also gained growing attention recently. Multi-modal segmentation in CNNs is typically addressed with an *early fusion* strategy, where multiple modalities are merged from the original input space of low-level features [10,15,18,23,29] (See Fig. 1, *left*). By concatenating image modalities at the input of the network, we explicitly assume that the relation between different modalities is simple (e.g., linear), which may not correspond to the characteristics of the multi-modal data at hand [21]. To better account for the complexity of multi-modal data, other studies investigated *late fusion* strategies [19], where each modality is processed by an independent CNN and the multi-modal outputs are merged in a deep layer, as in the architecture depicted in Fig. 1, *middle*. This *late fusion* strategy was demonstrated to outperform *early fusion* on infant brain segmentation [19]. More recently, Aygün et al. explored different ways of combining multiple modalities [3]. In this work, all modalities are considered as separate inputs to different CNNs, which are later fused at an 'early', 'middle' or 'late' point. Although it was found that 'late' fusion provides better performance, as in [19], this method relies on a single-layer fusion to model the relation between all modalities. Nevertheless, as demonstrated in several works [21], relations between different modalities may be highly complex and cannot easily be modeled by a single layer. To account for the non-linearity in multi-modal data modeling, we recently proposed a CNN that incorporates dense connections not only between pairs of layers within the same path, but also between layers across different

paths [9,11]. This architecture, known as *HyperDenseNet*, obtained very competitive performance in the context of infant and adult brain tissue segmentation with multi-modal MRI data.
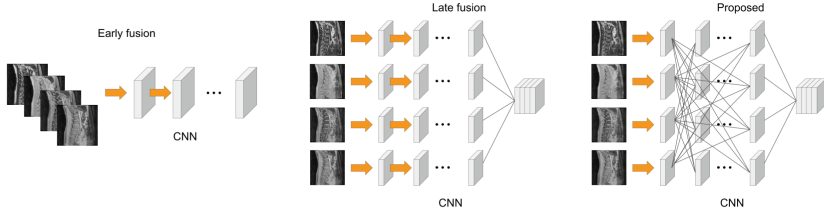
Fig. 1. Typical feature-fusion strategies (*left* and *middle*) and proposed fusion technique (*right*).

In the context of IVD localization and segmentation, Li et al. [17] have also considered multi-modal images. Specifically, they proposed a multi-scale and modality dropout learning framework, which employed four MRI modalities. To capture multi-scale context and handle the scale variations of IVDs, three different paths process regions extracted from the same location but at different scales. In addition, a random modality voxel dropout strategy is used to reduce feature co-adaptation between multiple modalities, and encourage each single modality to learn discriminative information independently.

Nevertheless, the combination of multi-modal data at various levels of abstraction has not been fully exploited for IVD localization and segmentation. In this work, we adopt the strategy presented in [9,11] and propose a multi-path architecture [8] called IVD-Net, where each modality is employed as input of one pathway, with dense connectivity used between the layers, within and across paths (Fig. 1, *right*). Furthermore, we extend the standard convolutional module of InceptionNet [22] by including two additional dilated convolutional blocks, which can help to learn larger context. In our previous work on multi-modal ischemic stroke lesion segmentation [8], we showed this model to outperform architectures based on early and late fusion, as well as several state-of-art segmentation networks.

2   Methodology

The proposed IVD-Net architecture follow the structure of UNet [20]. This well-known model is composed of two paths: one contracting and one expanding. While the former collapses the input image into a set of high level features forming a compact intermediate representation of the input, the latter employs these features to generate a pixel-wise segmentation mask. Furthermore, it includes skip-connections, which connect the outputs from shallow layers to the input of subsequent layers, with the goal of transferring information that may have been lost in the encoding path during the compression process.

2.1   Processing Multiple Modalities Separately

In order to fully exploit multi-modal data, we adopt the hyper-dense connectivity approach of [11] in the current work. To achieve this dense connectivity pattern, we first create an encoding path composed of multiple streams, each of them processing a different image modality. The main goal of employing separate streams for different modalities is to disentangle information that otherwise would be fused from an early stage, limiting the learning capabilities of the network to capture complex relationships between modalities. The structure of the proposed IVD-Net architecture is depicted in Fig. 2.
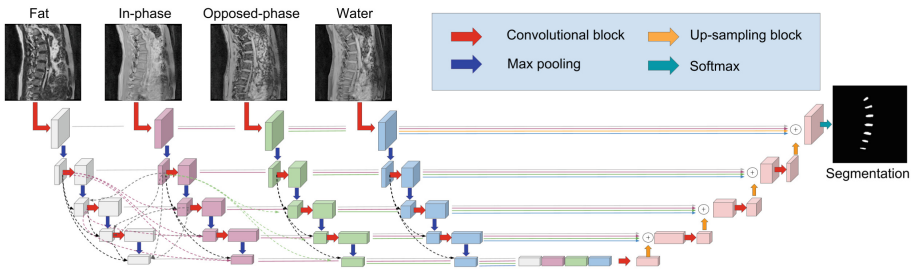

Fig. 2. Proposed IVD-Net architecture for IVD segmentation in multi-modal images, which extends the traditional UNet. Dotted lines represent some of the dense connectivity patterns adopted in this extended version of UNet.

2.2   Extended Inception Module

Meaningful areas in an image may undergo extremely large variation in size. In our particular case, as 3D segmentation is assessed in a 2D slice-wise manner, the region occupied by the IVD varies from one image to another. For instance, when the 2D sagittal slice corresponds to the center of the vertebral column, every IVD will appear in the image, whereas only one or two IVDs will be present in the image when the sagittal plane is located at extremes. This makes the selection of an accurate and general kernel size difficult. While a smaller kernel is better for local information, a larger kernel can capture information that is distributed globally. This idea is exploited in InceptionNet [22], where convolutions with multiple kernel sizes operate on the same level. Furthermore, in more recent versions, $n \times n$ convolutions are factorized to a combination of $1 \times n$ and $n \times 1$ convolutions, resulting in a 33% memory reduction.

To facilitate the learning of multiple contexts, we included two dilated convolutional blocks in parallel to the existing blocks in an inception module. Dilation rates of these blocks are different, which helps learning from different receptive fields, thereby increasing the context of the original inception modules. In addition, we removed max-pooling from the proposed architecture, as dilated convolutions were shown to be a better alternative, which captures more effectively the global context [25]. Our extended inception modules are depicted in Fig. 3.
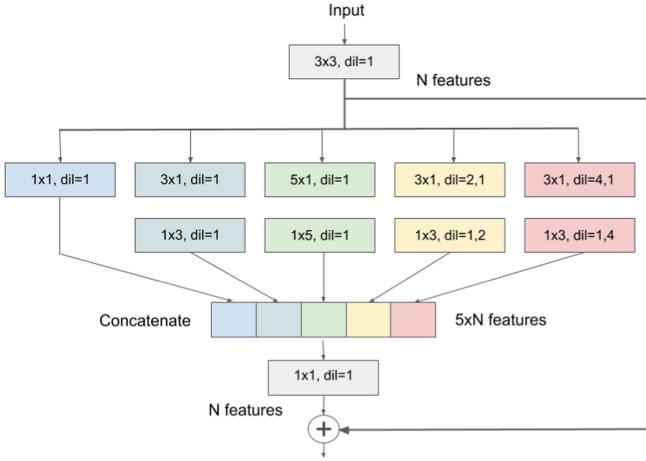
Fig. 3. Proposed extended inception modules. The module on the left employs standard convolutions while the module on the right adopts the idea of asymmetric convolutions [22].

2.3   Hyper-dense Connectivity

Inspired by the recent success of densely connected architectures for medical image segmentation [6, 11, 26], we adopted hyper-dense connections in the proposed model. The benefits of employing dense connections in the network are four-fold [11, 13]. First, as demonstrated in [11], dense connections between multiple streams can better model relationships between different modalities. Second, flow of information and gradients through the entire network is facilitated by the use of direct connections between all layers, which alleviates the problem of vanishing gradient. Third, including short paths to all feature maps in the network introduces an implicit deep supervision. Fourth, dense connections have a regularizing effect, reducing the risk of over-fitting on tasks with smaller training sets.

Formulation. Let x_l denote the output of the lth layer, and H_l be a mapping function, which corresponds to a convolution layer followed by a non-linear activation. In standard CNNs, the output of the lth layer is typically obtained from the output of the previous layer x_{l-1} as

$$x_l \; = \; H_l\big(x_{l-1}\big). \tag{1}$$

In a densely-connected network, nevertheless, all feature outputs are concatenated in a feed-forward manner, i.e.,

$$x_l \; = \; H_l\big([x_{l-1}, x_{l-2}, \ldots, x_0]\big), \tag{2}$$

where [. . .] denotes a concatenation operation.

In the present work, as in HyperDenseNet [9,11], the outputs from previous layers in different streams are also concatenated to form the input of subsequent layers. This connectivity yields a much more powerful feature representation than early or late fusion strategies in a multi-modal context, as the network is capable of learning more complex relationships between the different modalities within and in-between all levels of abstraction. For simplicity, let us consider the scenario with only two modalities. Let $\boldsymbol{x}_l^1$ and $\boldsymbol{x}_l^2$ denote the outputs of the lth layer in streams 1 and 2, respectively. Then, the output of the lth layer in a given stream s can be defined as

$$\boldsymbol{x}_l^s \;=\; H_l^s\big([\boldsymbol{x}_{l-1}^1, \boldsymbol{x}_{l-1}^2, \boldsymbol{x}_{l-2}^1, \boldsymbol{x}_{l-2}^2, \ldots, \boldsymbol{x}_0^1, \boldsymbol{x}_0^2]\big). \tag{3}$$

Furthermore, recent works have found that shuffling and interleaving complete feature maps (or single feature maps elements) in a CNN can improve its performance, as it serves as a strong regularizer [7,28,30]. Inspired by this, we concatenate feature maps in a different order for each branch and layer, where the output of the lth layer now becomes

$$\boldsymbol{x}_l^s \;=\; H_l^s\big(\pi_l^s([\boldsymbol{x}_{l-1}^1, \boldsymbol{x}_{l-1}^2, \boldsymbol{x}_{l-2}^1, \boldsymbol{x}_{l-2}^2, \ldots, \boldsymbol{x}_0^1, \boldsymbol{x}_0^2])\big), \tag{4}$$

with π_l^s being a function that permutes the feature maps given as input. Thus, in the case of two image modalities, the outputs of the lth layers in both streams can be defined as

$$\boldsymbol{x}_l^1 \;=\; H_l^1\big([\boldsymbol{x}_{l-1}^1, \boldsymbol{x}_{l-1}^2, \boldsymbol{x}_{l-2}^1, \boldsymbol{x}_{l-2}^2, \ldots, \boldsymbol{x}_0^1, \boldsymbol{x}_0^2]\big)$$
$$\boldsymbol{x}_l^2 \;=\; H_l^2\big([\boldsymbol{x}_{l-1}^2, \boldsymbol{x}_{l-1}^1, \boldsymbol{x}_{l-2}^2, \boldsymbol{x}_{l-2}^1, \ldots, \boldsymbol{x}_0^2, \boldsymbol{x}_0^1]\big)$$

A detailed example of the adopted hyper-dense connectivity for the case of two image modalities is depicted in Fig. 4. This figure shows a section (only three levels) of a deep CNN where the two image modalities are processed in separated paths and modules are linked in a hyper-dense fashion.

3   Materials

3.1   Dataset

The provided IVD dataset is composed of 16 3D multi-modal MRI data sets of at least 7 IVDs of the lower spine, collected from 8 subjects in two different stages. Each MRI data set contains four aligned high-resolution 3D volumes: in-phase, opposed-phase, fat and water images. In addition to the MRI images, corresponding reference manual segmentations were provided. More detailed information about the dataset can be found at the IVD website[1].

---

[1] https://ivdm3seg.weebly.com.
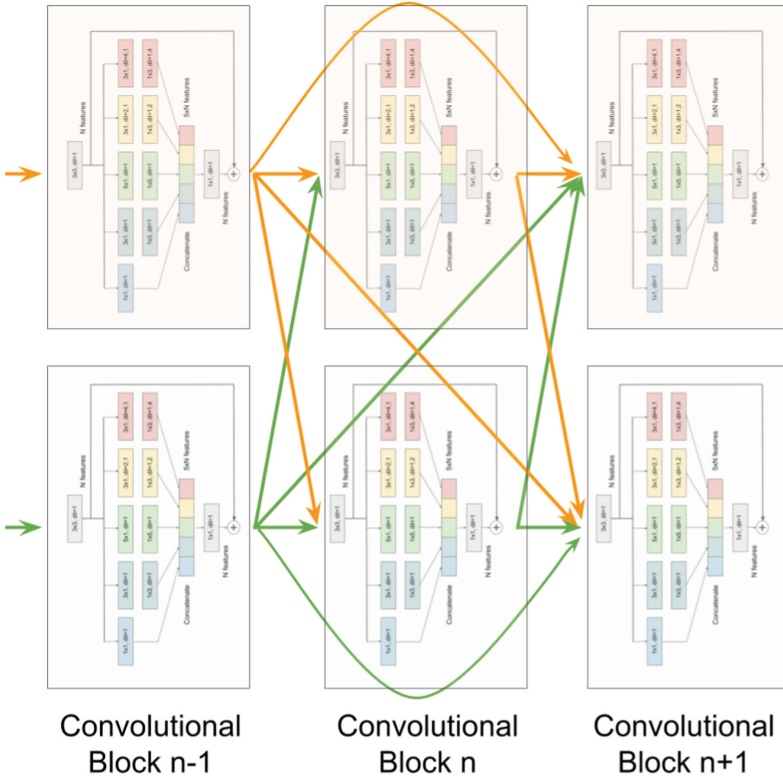
Fig. 4. Detailed version of a section of the proposed dense connectivity in multi-modal scenarios. For simplicity, two image modalities (in orange and in green) are considered in this example. While boxes represent a complete convolutional block of the proposed type, arrows indicate the connectivity pattern between modules. (Color figure online)

3.2   Evaluation Metrics

Even though segmentation is performed in a 2D-slice fashion, once all the 2D sagittal slices for a given patient have been segmented, they are stacked to reconstruct the original 3D volume. The metrics introduced below are therefore employed to evaluate performance on the whole 3D image. While the first metric is used to evaluate the segmentation accuracy, the second one serves as a measure of localization error.

Dice Similarity Coefficient (DSC). We first evaluate performance using Dice similarity coefficient (DSC), which compares volumes based on their overlap. Let V_{ref} and V_{auto} be the reference and automatic segmentations of a given tissue class and for a given subject, respectively. The DSC for this subject is defined as

$$\text{DSC}\big(V_{\text{ref}}, V_{\text{auto}}\big) \;=\; \frac{2\,|\,V_{\text{ref}} \cap V_{\text{auto}}\,|}{|\,V_{\text{ref}}\,| + |\,V_{\text{auto}}\,|} \tag{5}$$

Localization Distance. To evaluate the localization error, we compute the 3D barycenters of ground-truth and predicted IVDs, and measure their Euclidean distance. Results are given in voxels.

3.3   Implementation Details

Baselines. Several architectures are used to demonstrate the effectiveness of the proposed network. As baselines, we consider two UNet versions, the first one with early fusion and the other with late fusion. In early fusion, following the procedure employed in most works, all MRI image modalities are merged into a single input which is processed through a unique path. In contrast, for late fusion, each MRI modality is processed in a separate stream, and learned features of different modalities are fused in a later stage. In both early and late fusion, the extended inception module of Fig. 3 is employed, however asymmetric convolutions are replaced by standard $n \times n$ convolutions in these baselines. Another difference with respect to standard UNet is that feature maps from skip connections are summed before being fed into convolutional modules of the decoding path, instead of being concatenated.

Proposed Network. In terms of architecture, the proposed IVD-Net network and the one employed with late fusion strategy are very similar. As introduced in Sect. 2.3, the main difference is that feature maps from previous layers and different paths are concatenated and fed into the subsequent layers, following Eq. (4). Details of the resulting architecture are provided in Table 1. The first version of the proposed network employs the same convolutional module as the two baselines, whereas the second version adopts asymmetric convolutions instead (Fig. 3).

Training. We employed Adam optimizer to train the proposed architectures, with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. Training converged after 200 epochs with an initial learning rate of 1×10^{-4}, reduced by half after 100 epochs. Four images were used in each mini-batch. The same values for all hyper-parameters were employed across all architectures. Implementation of the analyzed architectures was done in PyTorch and experiments were performed on an NVidia TITAN XP GPU with 16 GBs RAM. While training was done in around 5 h, inference on a whole 3D volume took in 2–3 s on average. Images were normalized between 0 and 1 and no other pre- or post-processing steps were used. Furthermore, no data augmentation was employed to boost the performance of the networks. For all architectures, we used the four MRI modalities provided by the organizers as input. While 13 scans were employed for training 3 scans were used for validation.

4   Results

Quantitative results obtained with the different architectures are reported in Table 2. First, we observe that by simply fusing all image modalities at the input of the network provides the lowest mean DSC value. Adopting a late

**Table 1.** Layer placement of the proposed hyper-dense connected UNet.

	Name	HyperDense connectivity	
		Feat maps (input)	Feat maps (output)
Encoding Path (each modality)	Conv Layer 1	$1 \times 256 \times 256$	$32 \times 256 \times 256$
	Max-pooling 1	$32 \times 256 \times 256$	$32 \times 128 \times 128$
	Layer 2	$128 \times 128 \times 128$	$64 \times 128 \times 128$
	Max-pooling 2	$64 \times 128 \times 128$	$64 \times 64 \times 64$
	Layer 3	$384 \times 64 \times 64$	$128 \times 64 \times 64$
	Max-pooling 3	$128 \times 64 \times 64$	$128 \times 32 \times 32$
	Layer 4	$896 \times 32 \times 32$	$256 \times 32 \times 32$
	Max-pooling 4	$256 \times 32 \times 32$	$256 \times 16 \times 16$
	Bridge	$1920 \times 16 \times 16$	$512 \times 16 \times 16$
Decoding Path	Up-sample 1	$512 \times 16 \times 16$	$256 \times 32 \times 32$
	Layer 5	$256 \times 32 \times 32$	$256 \times 32 \times 32$
	Up-sample 2	$256 \times 32 \times 32$	$128 \times 64 \times 64$
	Layer 6	$128 \times 64 \times 64$	$128 \times 64 \times 64$
	Up-sample 3	$128 \times 64 \times 64$	$64 \times 128 \times 128$
	Layer 7	$64 \times 128 \times 128$	$64 \times 128 \times 128$
	Up-sample 4	$64 \times 128 \times 128$	$32 \times 256 \times 256$
	Layer 8	$32 \times 256 \times 256$	$32 \times 256 \times 256$
	Softmax layer	$32 \times 256 \times 256$	$2 \times 256 \times 256$

fusion strategy instead of early fusion achieves a mean DSC of 0.9086. Moreover, we see that our hyper-densely connected IVD-Net architecture brings a boost in performance compared to the more 'naive' early or late fusion strategies. When employing the extended module with standard convolutions (Fig. 3), we obtained a mean DSC of 0.9162, whereas the use of asymmetric convolutions in the proposed module provided the best performance in terms of mean DSC. These results are in line with values of localization distance, where the proposed architecture outperforms simpler fusion strategies. Nevertheless, in this case, the proposed network integrating standard convolutions slightly outperforms the architecture with asymmetric convolutions.

Table 2. Results on validation subjects obtained by the different architectures.

| Architecture	DSC	Localization distance (voxels)
Baseline_EarlyFusion	0.8981 ± 0.0293	0.7701 ± 1.5872
Baseline_LateFusion	0.9086 ± 0.0339	0.7400 ± 1.6009
IVD-Net	0.9162 ± 0.0192	$\mathbf{0.4145 \pm 0.2698}$
IVD-Net (asym)	$\mathbf{0.9191 \pm 0.0179}$	0.4470 ± 0.2641

Qualitative evaluation of the proposed IVD-Net architecture is assessed in Figs. 5 and 6. First, ground truth and automatic contours obtained with IVD-Net are depicted on the sagittal plane in Fig. 5 for two validation subjects. Then, 3D rendered volumes for the ground truth and CNN segmentation are compared in Fig. 6. In both figures, we can see that the segmentation obtained by our architecture is very close to the manual annotated data, which aligns with the quantitative results in Table 2.
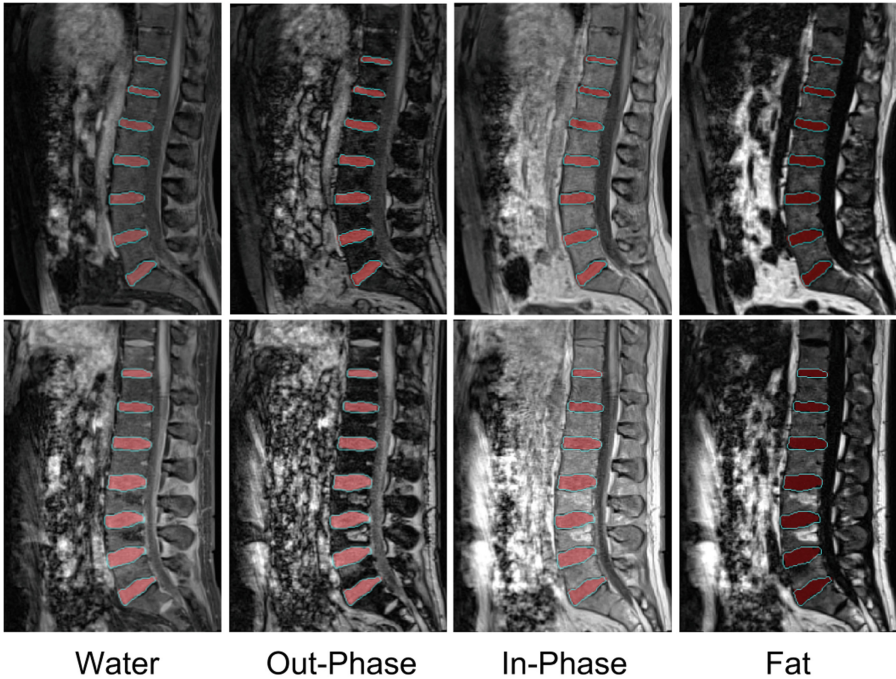
Fig. 5. Visual results for two subjects of the validation set. While the area in red represents the ground truth, bluish contours depict the automatic contours by our IVD-Net (asym) method in the different image modalities. (Color figure online)

5   Discussion

We have presented an architecture called IVD-Net that can efficiently leverage information from multiple image modalities for inter-vertebral disc segmentation. Following recent research on multi-modal image segmentation [8,11], our architecture adopts dense connectivity between multiple paths in the encoding section, each of them processing single modalities. Specifically, convolutional layers in any stream receive as input the features maps of all previous layers in the same stream as well as from other streams.

Ground Truth   Proposed CNN   Ground Truth + Proposed CNN
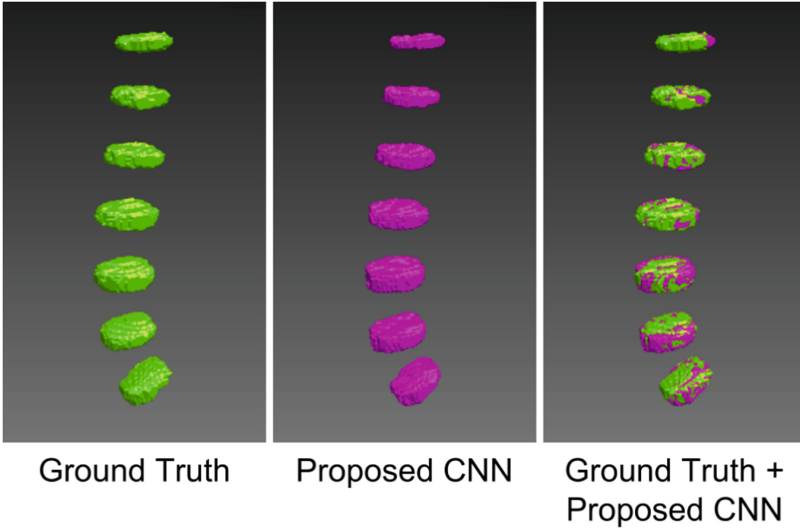
Fig. 6. 3D visualization of the ground truth, segmentation achieved by the proposed network and the combination of both for a subject on the validation set.

We have demonstrated that naive feature fusion strategies, such as simply merging information at an early or late stage, may be insufficient to fully exploit information in multi-modal scenarios. By allowing the network to learn how to combine learned features from separate modalities, it can capture more complex relationships between multiple sources. This improves its representation power, which ultimately results in a boost on performance. These findings are in line with recent works on multi-modal image segmentation [9,11,19]. For example, high-level features were combined at a late stage in [19], outperforming an early fusion strategy in the context of infant brain segmentation. In a recent work, we demonstrated that adopting more complex fusion techniques, referred to as hyper-dense connectivity, surpasses the performance of other features fusion strategies in the challenging tasks of infant and adult brain tissue segmentation [9,11].

Even though considering 3D context typically helps improve performance, we treated each volume as a stack of 2D sagittal slices (see Fig. 7). The main reason for this is that manual segmentations provided in this challenge were performed slice-wise in the sagittal plane. Thus, when looking at these annotations in the axial plane, a sharp contour is observed. As CNNs will generally provide a smooth contour, we assumed that tackling this problem as a 3D task would have led to lower values during evaluation. Furthermore, IVD localization is assessed after volumetric segmentation is done. This means that the process of localization itself is not optimized during training. A possible solution to overcome this limitation in the future might be to investigate multi-task architectures that can be trained end-to-end, so that both localization and segmentation tasks can be jointly optimized.
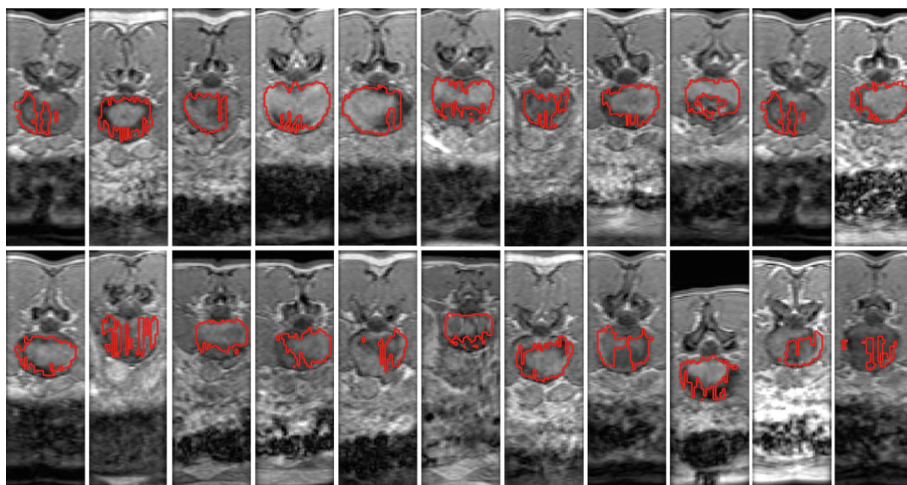
Fig. 7. Examples of manual annotations from the training set seen on axial slices.

References

1. An, H.S., et al.: Introduction: disc degeneration: summary. Spine **29**(23), 2677–2678 (2004)
2. Ben Ayed, I., Punithakumar, K., Garvin, G., Romano, W., Li, S.: Graph cuts with invariant object-interaction priors: application to intervertebral disc segmentation. In: Székely, G., Hahn, H.K. (eds.) IPMI 2011. LNCS, vol. 6801, pp. 221–232. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22092-0_19
3. Aygün, M., Şahin, Y.H., Ünal, G.: Multi modal convolutional neural networks forbrain tumor segmentation. arXiv preprint arXiv:1809.06191 (2018)
4. Chen, C., et al.: Localization and segmentation of 3D intervertebral discs in MR images by data driven estimation. IEEE Trans. Med. Imaging **34**(8), 1719–1729 (2015)
5. Chen, H., Dou, Q., Wang, X., Qin, J., Cheng, J.C.Y., Heng, P.-A.: 3D fully convolutional networks for intervertebral disc localization and segmentation. In: Zheng, G., Liao, H., Jannin, P., Cattin, P., Lee, S.-L. (eds.) MIAR 2016. LNCS, vol. 9805, pp. 375–382. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43775-0_34
6. Chen, L., Wu, Y., DSouza, A.M., Abidin, A.Z., Wismüller, A., Xu, C.: MRI tumor segmentation with densely connected 3D CNN. In: Medical Imaging 2018: Image Processing. International Society for Optics and Photonics (2018)
7. Chen, Y., Wang, H., Long, Y.: Regularization of convolutional neural networks using shufflenode. In: 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 355–360. IEEE (2017)

8. Dolz, J., Ben Ayed, I., Desrosiers, C.: Dense multi-path U-Net for ischemic stroke lesion segmentation in multiple image modalities. arXiv preprint arXiv:1810.07003 (2018)

9. Dolz, J., Ben Ayed, I., Yuan, J., Desrosiers, C.: Isointense infant brain segmentation with a hyper-dense connected convolutional neural network. In: 2018 IEEE 15th International Symposium on Biomedical Imaging, ISBI 2018, pp. 616–620. IEEE (2018)

10. Dolz, J., Desrosiers, C., Wang, L., Yuan, J., Shen, D., Ben Ayed, I.: Deep CNN ensembles and suggestive annotations for infant brain MRI segmentation. arXiv preprint arXiv:1712.05319, 2017

11. Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., Ben Ayed, I.: HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation. In: IEEE Transactions on Medical Imaging (2018, in press)

12. Hamanishi, C., Matukura, N., Fujita, M., Tomihara, M., Tanaka, S.: Cross-sectional area of the stenotic lumbar dural tube measured from the transverse views of magnetic resonance imaging. J. Spinal Disord. **7**(5), 388–393 (1994)

13. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, vol. 1, p. 3 (2017)

14. Ji, X., Zheng, G., Belavy, D., Ni, D.: Automated intervertebral disc segmentation using deep convolutional neural networks. In: Yao, J., Vrtovec, T., Zheng, G., Frangi, A., Glocker, B., Li, S. (eds.) CSI 2016. LNCS, vol. 10182, pp. 38–48. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-55050-3_4

15. Kamnitsas, K., et al.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. **36**, 61–78 (2017)

16. Kim, S., Bae, W., Masuda, K., Chung, C., Hwang, D.: Fine-grain segmentation of the intervertebral discs from MR spine images using deep convolutional neural networks: BSU-Net. Appl. Sci. **8**(9), 1656 (2018)

17. Li, X., et al.: 3D multi-scale FCN with random modality voxel dropout learning for intervertebral disc localization and segmentation from multi-modality MR images. Med. Image Anal. **45**, 41–54 (2018)

18. Moeskops, P., Viergever, M.A., Mendrik, A.M., de Vries, L.S., Benders, M.J., Išgum, I.: Automatic segmentation of MR brain images with a convolutional neural network. IEEE Trans. Med. Imaging **35**(5), 1252–1261 (2016)

19. Nie, D., Wang, L., Gao, Y., Sken, D.: Fully convolutional networks for multi-modality isointense infant brain image segmentation. In: 13th International Symposium on Biomedical Imaging (ISBI), pp. 1342–1345. IEEE (2016)

20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

21. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep boltzmann machines. J. Mach. Learn. Res. **15**, 2949–2980 (2014)

22. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR, pp. 2818–2826 (2016)

23. Valverde, S., et al.: Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. NeuroImage **155**, 159–168 (2017)

24. Wieser, S., et al.: Cost of low back pain in switzerland in 2005. Eur. J. Health Econ. **12**(5), 455–467 (2011)

25. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)

26. Yu, L., et al.: Automatic 3D cardiovascular MR segmentation with densely-connected volumetric ConvNets. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 287–295. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_33

27. Zeng, G., Zheng, G.: DSMS-FCN: a deeply supervised multi-scale fully convolutional network for automatic segmentation of intervertebral disc in 3D MR images. In: Glocker, B., Yao, J., Vrtovec, T., Frangi, A., Zheng, G. (eds.) MSKI 2017. LNCS, vol. 10734, pp. 148–159. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-74113-0_13

28. Zhang, T., Qi, G.-J., Xiao, B., Wang, J.: Interleaved group convolutions. In: CVPR, pp. 4373–4382 (2017)

29. Zhang, W., et al.: Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage **108**, 214–224 (2015)

30. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: an extremely efficient convolutional neural network for mobile devices. arXiv preprint arXiv:1707.01083 (2017)

31. Zheng, G., et al.: Evaluation and comparison of 3D intervertebral disc localization and segmentation methods for 3D T2 MR data: a grand challenge. Med. Image Anal. **35**, 327–344 (2017)