Deep Learning-Based Image Segmentation on Multimodal Medical Imaging

Zhe Guo, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li

***Abstract*—Multimodality medical imaging techniques have been increasingly applied in clinical practice and research studies. Corresponding multimodal image analysis and ensemble learning schemes have seen rapid growth and bring unique value to medical applications. Motivated by the recent success of applying deep learning methods to medical image processing, we first propose an algorithmic architecture for supervised multimodal image analysis with cross-modality fusion at the feature learning level, classifier level, and decision-making level. We then design and implement an image segmentation system based on deep convolutional neural networks to contour the lesions of soft tissue sarcomas using multimodal images, including those from magnetic resonance imaging, computed tomography, and positron emission tomography. The network trained with multimodal images shows superior performance compared to networks trained with single-modal images. For the task of tumor segmentation, performing image fusion within the network (i.e., fusing at convolutional or fully connected layers) is generally better than fusing images at the network output (i.e., voting). This paper provides empirical guidance for the design and application of multimodal image analysis.**

***Index Terms*—Computed tomography (CT), convolutional neural network (CNN), magnetic resonance imaging (MRI), multimodal image, positron emission tomography (PET).**

I. Introduction

In THE field of biomedical imaging, use of more than one modality (i.e., multimodal) on the same target has become a growing field as more advanced techniques and devices have become available. For example, simultaneous acquisition of positron emission tomography (PET) and computed tomography (CT) [1] has become a standard clinical practice for a number of applications. Functional imaging techniques such as PET which lacks anatomical characterization while providing quantitative metabolic and functional information about diseases can work together with CT and magnetic resonance imaging (MRI) which provide details on anatomic structures via high contrast and spatial resolution to better characterize lesions [2]. Another widely used multimodal imaging technique in neuroscience studies is the simultaneous recording of functional MRI (fMRI) and electroencephalography (EEG) [3], which offers both high spatial resolution (through fMRI) and temporal resolution (through EEG) on brain dynamics.

Correspondingly, various analyses using multimodal biomedical imaging and computer-aided detection systems have been developed. The premise is that various imaging modalities encompass abundant information which is different and complementary to each other. For example, in one deep-learning-based framework [4], automated detection of solitary pulmonary nodules were implemented by first identifying suspect regions from CT images, followed by merging them with high-uptake regions detected on PET images. As described in a multimodal imaging project for brain tumor segmentation [5], each modality reveals a unique type of biological/biochemical information for tumor-induced tissue changes and poses "somewhat different information processing tasks." Similar concepts have been proposed in the field of ensemble learning [6], where decisions made by different methods are fused by a "meta-learner" to obtain the final result, based on the premise that the different priors used by these methods characterize different portions or views of the data.

There is a growing amount of data available from multimodal medical imaging and a variety of strategies for the corresponding data analysis. In this paper, we investigate the differences among multimodal fusion schemes for medical image analysis, based on empirical studies in a segmentation task. In their review, James and Dasarathy provide a perspective on multimodal image analysis [7], noting that any classical image fusion method is composed of "registration and fusion of features from the registered images." It is also noted in the survey work of [8] that networks representing multiple sources of information "can be taken further and channels can be merged at any point in the network."

Motivated by this perspective, we advance one step further from the abstraction of image fusion methods in [7] and propose an algorithmic architecture for image fusion strategies that can cover most supervised multimodal biomedical image analysis methods. This architecture also addresses the need for a unified framework to guide the design of methodologies for multimodal image processing. Based on the main

Manuscript received March 16, 2018; revised July 13, 2018 and November 5, 2018; accepted December 9, 2018. Date of publication January 1, 2019; date of current version March 1, 2019. This work was supported by the National Institutes of Health under Grant 1RF1AG052653-01A1, Grant 1P41EB022544-01A1, and Grant C06 CA059267. The work of Z. Guo was supported by the China Scholarship Council. *(Zhe Guo and Xiang Li contributed equally to this work.) (Corresponding author: Quanzheng Li.)*

Z. Guo is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: guo_zion@bit.edu.cn).

X. Li, N. Guo, and Q. Li are with the Department of Radiology, Massachusetts General Hospital, Boston, MA 02114 USA (e-mail: xli60@mgh.harvard.edu; guo.ning@mgh.harvard.edu; li.quanzheng@mgh.harvard.edu).

H. Huang is with the Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15260 USA (e-mail: heng.huang@pitt.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TRPMS.2018.2890359

stages of machine learning models, our design includes fusing at the feature level, fusing at the classifier level, and fusing at the decision-making level. We further propose that optimizing a multimodal image analysis method for a specific application should consider the possibility of all the three strategies and select the most suitable one for the given use case.

Successes in applying deep convolutional neural networks (CNNs) for natural image [9] and medical image [10], [11] processing have been recently reported. Further, for the task of automatic tumor segmentation, CNNs have been applied to segmentation of tumors in brain [5], [12], [13], liver [14], breast [15], lung [16], [17], and other regions [18]. These deep learning-based methods have achieved superior performance compared to traditional methods (such as level set or region growing) with good robustness toward common challenges in medical image analysis, including noise and subject-wise heterogeneity. Deep learning on multimodal images (which are also referred to as multisource/multiview images) is an important topic with growing interest in the computer vision and machine learning community. To name a few, works in [19] proposed the cross-modality feature learning scheme for shared representation learning. Work in [20] developed a multiview deep learning model with deep canonical correlated autoencoder and shared representation to fuse two views of data. Similar multisource modeling has also been applied for image retrieval [21] by incorporating view-specific and view-shared nodes in the network. In addition to the correlation analysis, consistency evaluation across different information sources is used by multisource deep learning framework in [22] to estimate trustiness of information sources. When image views/sources are unknown, the multiview perceptron model introduced in [23] explicitly perform classification on views of the input images as an added route in the network. Various methods have also been developed for deep learning-based works for multimodal/multiview medical analysis. For example, work in [24] used shared image features from unregistered views of the same region to improve classification performance. Framework proposed in [25] fuses imaging data with non-image modalities by using a CNN to extract image features and jointly learn their nonlinear correlations using another deep learning model. The multimodal feature representation framework introduced in [26] fuses information from MRI and PET in a hierarchical deep learning approach. The unsupervised multimodal deep belief network [27] encoded relationships across data from different modalities with data fusion through a joint latent model.

However, there has been little investigation from a systematic perspective about how multimodal imaging should be used. There are few empirical studies on how different fusing strategies can affect segmentation performance. In this paper, we address this problem by testing different fusion strategies through different implementations of CNN architecture.

A typical CNN for supervised image classification consists of: 1) convolutional layers for feature/representation learning, which utilize local connections and shared weights of the convolutional kernels followed by pooling operators, resulting in translation invariant features and 2) fully connected layers for classification, which use high-level image features extracted from the convolutional layers as input to learn the complex mapping between image features and labels. CNN is a suitable platform to test and compare the different fusion strategies as proposed above in a practical setting, as we can customize the fusion location in the network structure: either at the convolutional layers, fully connected layers, or network output.

II. Materials and Methods

A. Algorithmic Architecture for Multimodal Image Fusion Strategies and Summary of Related Works

As any supervised learning-based method consists of three stages: 1) feature extraction/learning; 2) classification (based on features); and 3) decision making (usually a global classification problem but varies), we summarize the three strategies for fusing information from different image modalities, as shown below.

1) *Fusing at Feature Level:* Multimodality images are used together to learn a unified image feature set, which shall contain the intrinsic multimodal representation of the data. The learned features are then used to support the learning of a classifier.
2) *Fusing at the Classifier Level:* Images of each modality are used as separate inputs to learn individual feature sets. These single-modality feature sets will be then used to learn a multimodal classifier. Learning the single-modality features and learning the classifier can be conducted in an integrated framework or separately (e.g., using unsupervised methods for learning the single-modality features then train a multimodality classifier).
3) *Fusing at the Decision-Making Level:* Images of each modality are used independently to learn a single-modality classifier (and the corresponding feature set). The final decision of the multimodality scheme is obtained by fusing the output from all the classifiers, which is commonly referred to as "voting" in [6], although the exact scheme of decision making varies across methods.

Any practical scenario using supervised learning on multimodal medical images belongs to one of these fusion strategies, and most of the current literature reports can be grouped accordingly. Works in [28] (co-analysis of fMRI and EEG using CCA), [29] (co-analysis of MRI and PET using PLSR), and [30] (co-learning features through pulse-coupled neural network) perform feature-level fusion of the images. Works in [31] (using features of contourlets), [32] (using feature of wavelet), [33] (using features of wavelets), and [34] (using features learned by linear discriminant analysis) perform the classifier-level fusion. Several other works in image segmentation, such as [35] (fusing the results from different atlases by majority voting) and [36] (fusing the support vector machine results from different modalities by majority voting), as well as the multimodal brain tumor segmentation framework [5] (using majority vote for fusing results from different algorithms, rather than modalities) belong to decision-level fusion.

B. Data Acquisition and Preprocessing

In this paper, we use the publicly available soft-tissue sarcoma (STS) dataset [37] from the cancer imaging archive [38] for model development and validation. MRI is mainly used for the diagnosis of STS, while other options including CT or ultrasound [39], [40]. As STS poses high risk of metastasis (especially to lung) leading to low survival rates, a comprehensive characterization of STSs including imaging-based biomarker identification is a crucial task for better adapted treatment. Accurate segmentation of the tumor region plays an important role for image interpretation, analysis, and measurement. The STS dataset contains a total of four imaging modalities: FDG-PET/CT and two anatomical MR imaging sequences (T1-weighted and T2-weighted fat-saturated). Images from all those four modalities have been preregistered to the same space. It should be noted that throughout this paper we regard T1- and T2-weighted imaging as two "modalities" because they portray different tissue characteristics. The STS dataset encompasses 50 patients with histologically proven STSs of the extremities. The FDG-PET scans were performed on a PET/CT scanner (Discovery ST, GE Healthcare, Waukesha, WI, USA) at the McGill University Health Centre. PET attenuation corrected images were reconstructed (axial plane) using an ordered subset expectation maximization (OSEM) iterative algorithm. PET slice thickness resolution was 3.27 mm and the median in-plane resolution was 5.47×5.47 mm^2. For MRI imaging, T1 sequences were acquired in the axial plane for all patients while T2 (or short tau inversion recover) sequences were scanned in different planes. The median in-plane resolution for T1-weighted MR imaging was 0.74×0.74 mm^2 and T2-weighted MR was 0.63×0.63 mm^2. The median slice thickness was 5.5 mm and 5.0 mm for T1 and T2, respectively.

The gross tumor volume (GTV) was manually annotated based on the T2-weighted MR images by expert radiologists with access to the other modalities. After drawing the GTV on T2 images, corresponding contours of these annotations for the other modalities were then obtained using rigid registration with the commercial software MIM (MIM software Inc., Cleveland, OH, USA). As the PET/CT images have a much larger fields of view than the MR images, they were truncated to the regions with MR images. In addition, the PET images were first converted to standardized uptake values (SUVs) and linearly up-sampled to the same resolution of other modalities. Pixel values for all three modalities are linearly normalized to the value interval of 0–255 according to the original pixel value.

The final data used as input in this analysis has four modalities of imaging (PET, CT, T1, and T2 MR), all in the same image size for each subject while the size varies across different subjects. A sample multimodal image set is illustrated in Fig. 1. In the analysis, image patches with the size of 28×28 are extracted from all images. A patch is labeled as "positive" if its center pixel is within the annotation (i.e., tumor-positive) region and labeled as "negative" otherwise. On average, around 1 million patches were extracted from each subject, with around 0.1 million positive patches. During the training phase, to balance the number of positive and negative patches, we randomly selected negative patches to the same number of positive patches. During the testing phase, we used all the patches for segmentation.

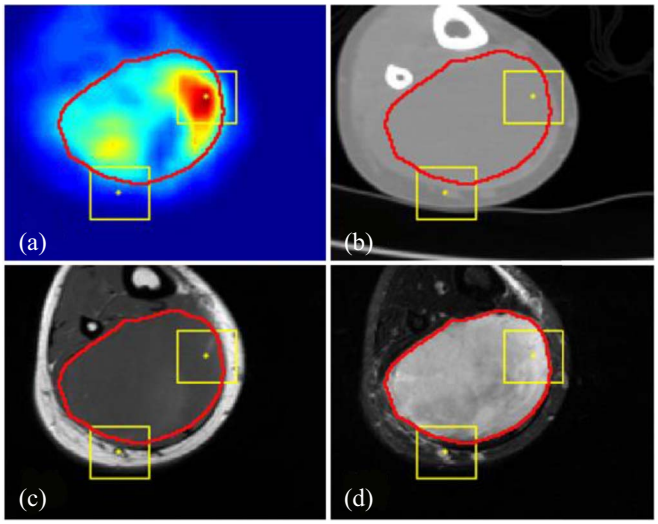

Fig. 1. Multimodal images on the same position from a randomly selected subject. (a) PET; (b) CT; (c) T1; and (d) T2. The image size of this subject is 133×148. Red line is the contour of ground truth from manual annotation. Two yellow boxes illustrate the size of patches (28×28) used as the input for CNN. The center pixel of one patch is within the tumor region and another patch outside the tumor region.

C. Multimodal Image Classification Using CNN

We implemented and tested three fusion strategies in three different patch-based CNNs with corresponding variations in network structures, as illustrated in Fig. 2: 1) the Type-I fusion network represents feature-level fusing; 2) the Type-II fusion network represents classifier-level fusing; and 3) the Type-III fusion network represents decision-level fusion. All the networks use same set of image patches as input. The network outputs, which are the labels of the given patches, are aggregated by assigning the corresponding label to the pixel in the patch center in the output label maps. All the single and multimodal networks were implemented in TensorFlow and run on a single NVIDIA 1080Ti GPU. Training time for a single-modal network on the current dataset was around 3 h. For multimodal networks of all types, the training time was around 10 h. Testing time (i.e., segmentation on new images) of any single or multimodal network was negligible (<2 min).

For the Type-I fusion network, patches from different modalities are transformed into a 3-D tensor ($28 \times 28 \times k$, where k is the number of modalities) and convoluted by a $2 \times 2 \times k$ kernel as shown in Fig. 2(a), to fuse the high-dimensional features to the 2-D space thus performing the feature-level fusion. Outputs from the k-dimensional kernel are then convoluted by typical 2×2 kernels. For the Type-II fusion network, the features are learned separately through each modality's own convolutional layers. Outputs of the last convolutional layers from each modality, which are considered the high-level representation of the corresponding images, are used to train a single fully connected network (i.e., classifier), as in Fig. 2(b).

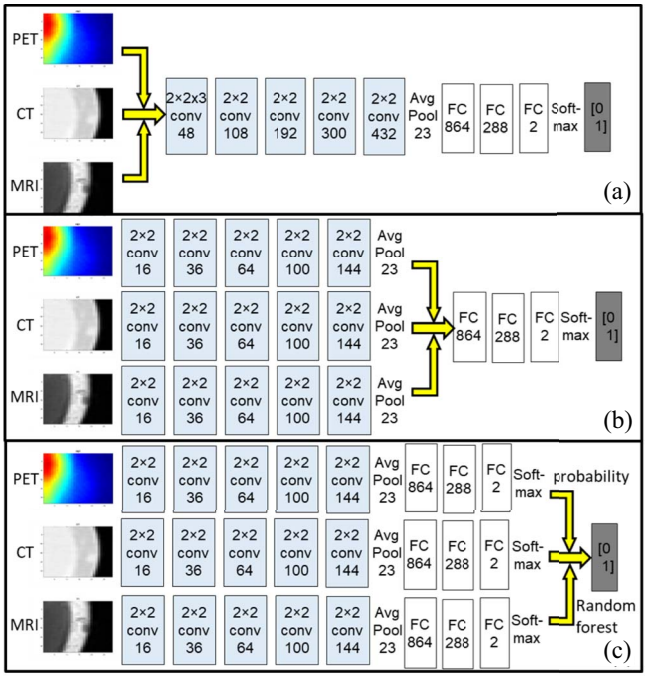

Fig. 2. Illustration of the structure for (a) Type-I fusion networks, (b) Type-II fusion network and (c) Type-III fusion network. The yellow arrows indicate the fusion location.

For the Type-III fusion network, for each modality we train a single-modal 2-D CNN to map its own image to the annotation. The prediction results (i.e., patch-wise probability) of these single-modality networks are then ensembled together to obtain the final decision (i.e., patch-wise label). The ensemble can be done in many ways: the simplest form is majority voting (i.e., label of a patch is set to the majority label from classifiers). In this paper, we utilized the random forest algorithm [41] to train a series of decision trees for the patch-wise label classification, as random forest has been shown to be capable of achieving better generalizability and avoid overfitting in many applications. The random forest algorithm uses bootstrap sampling of the data to learn a set of decision trees, where a random subset of data is used at each decision split. Details of implementation can be found in [42]. The random forest algorithm in this paper uses an ensemble of 10 bagged decision trees, each tree with maximum depth of 5. These hyper parameters were determined through grid search.

D. Experiments on Synthetic Low-Quality Images

While it is expected that multimodal imaging should offer additional information for lesion classification resulting in better performance compared with single-modality methods, it is interesting to investigate the extent of such a benefit. To answer this question, and at the same time simulate a practical scenario of low-dose imaging, we generated synthetic low-quality images by adding random Gaussian noise into the original images and used them for training and testing in both the single-modality and multimodality networks, following the same tenfold cross-validation scheme. Images after adding Gaussian noise were normalized to the same value

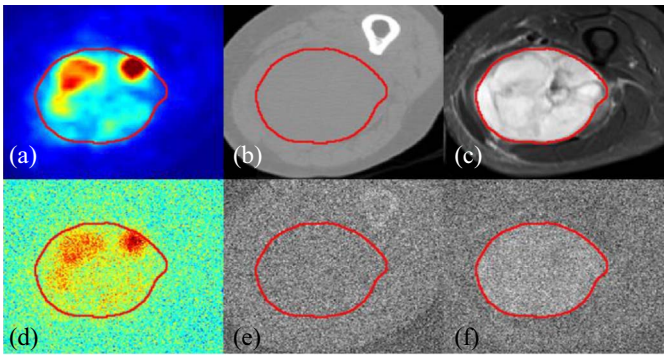

Fig. 3. Sample multimodal image before and after adding Gaussian noise. (a) Ground truth shown as red contour overlaid on PET image, (b) CT image, and (c) T2 image. After adding noise, (d) PET image, (e) CT image, and (f) T2 image. The magnitude factor k equals to 1.

interval as the original images to ensure similar settings in the imaging parameters. A sample multimodal image before and after adding noise is visualized in Fig. 3.

As seen in Fig. 3, when the noise magnitude is 1 (standard deviation equal to the 90% of the cumulative histogram distribution value of the image), low-quality PET images maintain good contrast of the tumor region with blurred boundaries. Similarly, tumor regions can be visually identified from the low-quality T2 image, but the contrast is very low. On the other hand, the contrast of CT image after adding noise became so low that tumors cannot be distinguished from background. Apparently performing segmentation on these synthetic images will be challenging for certain modalities, which is similar to the case of low-dose image analysis.

E. Segmentation and Performance Evaluation

The whole image set containing PET, CT, and MR T1-weighted and T2-weighted images from 50 patients were divided into training (including validation) and testing sets, based on the tenfold cross-validation scheme. In each run of the cross-validation experiment for the single-modality and Type-I/II networks, PET + CT + T1 or PET + CT + T2 images from 45 patients were used for training the three-modality network, while the remaining 5 patients were used for testing and performance evaluation. With a total of ten runs, images from every patient were tested. In each run, the same number (around 5 million) of positive and negative samples were used as training input. Model performances were evaluated based on pixel-wise accuracy by comparing the predicted label with the ground truth from human annotation, as well as Sørensen–Dice coefficient [43] (DICE coefficient) which equals to twice the number of voxels within both regions divided by the summed number of voxels in each region to measure the similarity between predicted region and annotation region. It should be noted while labels of all the patches from the five test patients in each run were predicted during the testing phase, we calculated the prediction result based on equal numbers of positive and negative patches, in order to overcome the problem of unbalanced samples. For the Type-III network with random forest, the prediction was based on the single-modality networks in the tenfold cross validation.

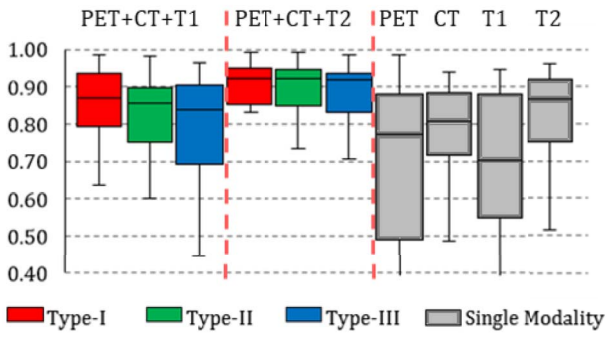

Fig. 4. Box chart for the statistics (median, first/third quartile and the min/max) of the DICE coefficient across 50 subjects. Each box corresponds to one specific type of network trained and tested on one specific combination of modalities. For example, the first box from the left shows the prediction statistics of Type-I fusion network trained and tested on images from PET, CT, and T1-weighted MR imaging modalities.

Patch-wise probabilities of each patch being within the tumor region from the single-modality networks are then combined to train a random forest (training labels of the patches are ground truth) in the same tenfold cross-validation approach.

We also performed comparison between the model performance using three modalities (PET, CT, and MRI T1 or T2) and two modalities (PET + CT, CT + T2 and PET + T2). Hyper parameters remain similar for these networks with alteration of network structure for the number of modalities. For example, multimodal PET + CT fusion Type-I network has two input channels, the images will go through $2 \times 2 \times 2$ convolutional kernel followed by 2-D 2×2 kernels. All fused networks were implemented with Type-I, Type-II, and Type-III strategies. Raw outputs of the networks, which are of patch-wise classification results, were transformed to the "label map" by assigning each pixel in the input image the label of the patch centered at it.

III. Results

A. Performances Comparison Between Single-Modality Networks and Multimodality Fusion Networks

DICE coefficient of single and multimodality networks are summarized in the box charts of Fig. 4: average DICE of Type-I, II, and III fusion networks on PET + CT + T1 is 82%, 80%, and 77%, respectively. Average DICE of Type-I, II, and III fusion networks on PET + CT + T2 is 85%, 85%, and 84%, respectively. Average performance of a single-modality network is 76%, 68%, 66%, and 80% for PET, CT, T1, and T2 images. From the statistics, it can be seen that the DICE of single-modality networks are all lower than the multimodality fusion networks, while no network achieved result higher than 80%. The networks trained and tested on the T2-weighted MR had the best performance. The reason is that: 1) annotation is performed on T2 images and 2) T2 relaxation is more sensitive to STS, as illustrated in Fig. 1(d). It is also interesting to observe that the performance of PET-based network is the worst in average while PET is designed to detect the tumor presence. This is mainly caused by the necrosis in the center of large tumor which barely show uptake in FDG-PET images.

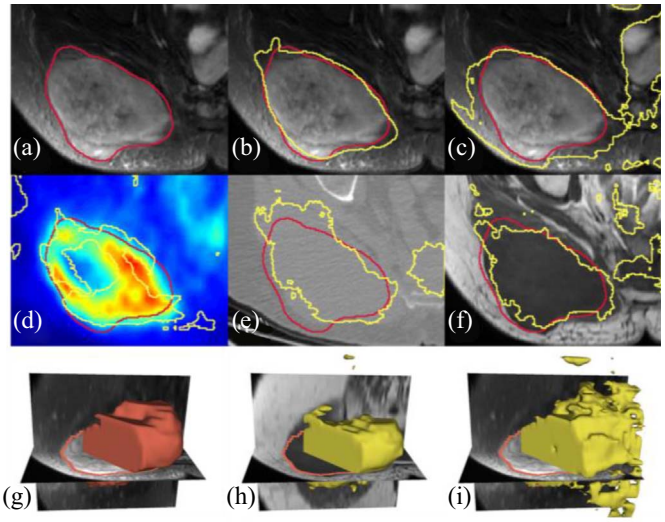

Fig. 5. (a) Ground truth shown as red contour line overlaid on T2-weighted MR image. (b) Result from Type-II fusion network based on PET + CT + T1. (c) Result from single-modality network based on T2. (d)–(f) Results from single-modality network based on PET, CT, and T1, respectively. (g) 3-D surface visualization of the ground truth. (h) 3-D surface visualization of the result from Type-III fusion network based on PET + CT + T1. (i) 3-D surface visualization of the result from single-modality network based on T2.

Although the annotation was mainly performed on the T2-weighted images, the fusion network trained and tested on the combination of PET, CT, and T1 (without T2) achieved better result, on average compared to the single-modality network based on T2 images (by around 2% improvement). Such result shows that while modalities other than T2 might be inaccurate and/or insufficient to capture the tumor region in single, the fusion network (using any of the fusion scheme) can automatically take advantage of the combined information. An illustrative example is shown in Fig. 5, where the multimodality fusion network Fig. 5(b) can obtain the better result comparing with T2-based single-modality network [Fig. 5(c)]. A closer examination of the single-modality networks based on PET, CT, and T1 shows that neither of these three modalities can lead to a good prediction: PET [Fig. 5(d)] suffers from the necrosis in the center issue as discuss above, while a large region of false positive is presented in CT, T1, and T2 results [Fig. 5(c), (e), and (f)].

B. Performance Comparison on Synthetic Low-Quality Image

By training and testing the single-modality and multimodality networks on the synthetic low-quality images with Gaussian noise, we obtained label maps and corresponding prediction accuracies. Model performance on both original images and low-quality images are summarized in Fig. 6, under the noise magnitude k of 1. From the statistics, three important observations can be made.

First, when the image quality degrades, the segmentation performance decreased for all the networks. However, the level of decrease for single-modality networks was far higher than the multimodality networks. For example, the result of segmentation on single-modality low-quality CT images decreases to random guessing, which is in correspondence with what has been observed in Fig. 3. On the other hand, performance of

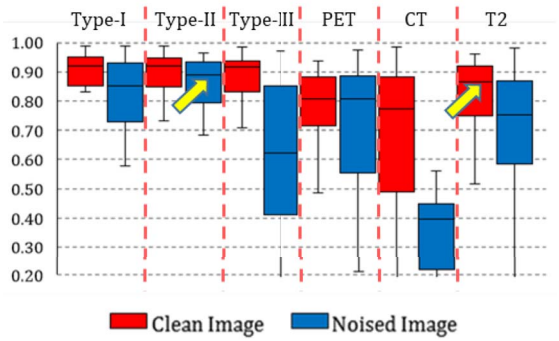

Fig. 6. Box chart for the statistics (median, first/third quartile and the min/max) of the DICE coefficient across 50 subjects. Red box stands for networks trained and tested on original clean images and blue box stands for networks based on synthetic noised image.

all Type-I and II networks only slightly decrease: their DICE measurements are all above 80%.

Second, it is interesting to find that the performance of multimodality networks based on low-quality images is on the same level or even higher than the performance of single-modality networks based on original images, as indicated by the arrows in Fig. 6. The observation indicates that multimodal imaging can be useful in low-image quality settings (such as low-dose scans), as its analytic performance is far less impacted by the degraded image quality. Fig. 7 shows an example consisting of the results from three single-modality networks (on original image) and one Type-II fusion network (on low-quality image). Networks based on PET as a single modality cannot define correct tumor boundaries while at the same time they generate false positives outside the tumor region. Networks based on single CT and T2 MRI can delineate the rough tumor boundaries but with either false positive outliers [Fig. 5(c), from T2]) or incorrect boundary definition [Fig. 7(b), from CT]. On the other hand, the performance of a multimodal fusion network on the same subject is clearly superior [Fig. 7(d)], although it was trained and tested on noised images (as visualized in the background). Further examination of multimodal fusion network performance on low-quality images with different noise magnitudes shows that on low-to-mid noise magnitudes ($k = 0.5/1$), the performance of multimodal fusion networks is similar to performance on original clean images. Specifically, for $k = 0.5$ (i.e., standard deviation of Gaussian noise is almost half of the image intensity), there is no significant difference ($p < 0.05$) between the segmentation result on original and noised images for each subject. At higher noise magnitudes ($k = 2$), the model performance deteriorates to below 80% (0.62% for Type-I, 75% for Type-II, and 52% for Type-III fusion networks), which is worse than the single-modality performance on T2 images.

Third, among different fusion strategies, fusion networks of Type-I and II perform largely the same per the statistics, both on original images and on low-quality noised images. While Type-III networks with random forest have consistently worse performance. It is important to find it performs the worst among the three strategies, as it is commonly applied in multimodal image studies. Finally, with regard to the

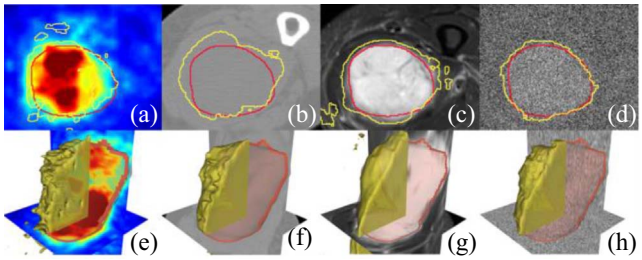

Fig. 7. Network result on different modalities. Contour line of the ground truth annotation (red line) and network performance (yellow line). (a) Single PET network on PET image. (b) Single CT network on CT image. (c) Single T2 network on T2 image. (d) Fused noisy PET/CT T2 image on noisy T2. (e)–(h) 3-D surface visualization of the segmentation results in (a)–(d).

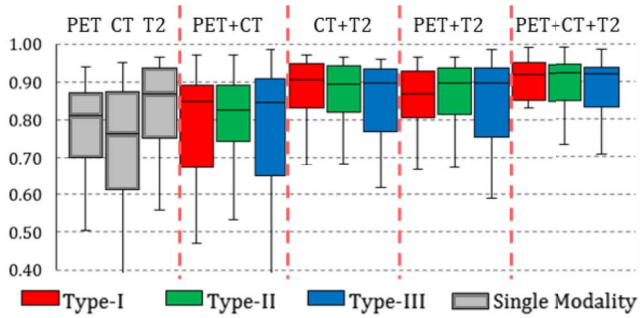

Fig. 8. Box chart for the statistics (median, first/third quartile and the min/max) of the DICE coefficient across 50 subjects. Red box stands for network train and test on Type-I network, blue box stands for Type-II network, and green stands for Type-III. Performances of single-modality network are shown as gray boxes to the left for reference.

computational complexity which affects training and testing time as well as hardware cost, and the ease of implementation, networks with earlier fusion (Type-I) are superior for their simplicity in model structure.

C. Performance Comparison Using Different Modality Combinations

Based on the observations on model performance difference between multimodal and single-modal networks and the detailed investigation of the label maps from network results, we have found that additional imaging modalities can offer new information to the segmentation task even with lowered image quality. Yet it is still unclear how (and whether) different modalities contribute to the multimodality network. In other words, if little or no performance increase is consistently observed from a certain combination of imaging modalities compared with its single-modality counterpart, then we can conclude that the extra modality is not contributing to the segmentation task. To this end, we trained and tested the multimodal fusion networks on additional combinations of imaging modalities as introduced in the methods section. Statistics of the performance of these networks are summarized in Fig. 8.

From Fig. 8, it can be observed that a fusion network based on PET + T2 has similar but lowered performance compared to a fusion network based on PET + CT + T2, showing CT has a limited contribution to the segmentation. More importantly, a fusion network based on PET + CT has significantly ($p < 0.05$) higher performance than single-modality networks

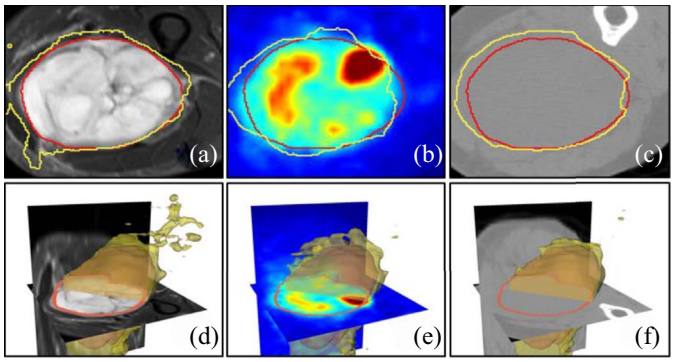

Fig. 9. Contour line of the ground truth annotation (red line) and segmentation result (yellow line). (a) Single-modality network on T2. (b) Multimodality network on T2 + PET (Type-I). (c) Multimodalities network on T2 + PET + CT. (d)–(f) 3-D surface visualization of the segmentation results in (a)–(c).

on PET or CT for Type-I and Type-II fusion strategies, indicating that a low-contrast imaging modality (such as CT) can significantly improve the segmentation accuracy for functional imaging (PET). To further illustrate this, Fig. 9 is an example case of segmentation from a single-modality network based on T2 images, a multimodality network based on T2 + PET images and a multimodality network based on T2 + PET + CT images. By gradually adding extra modalities, the resulting tumor region segmentation is shown corresponding improvements: a single-modality T2 network can delineate a rough boundary of the tumor but also generates false positives in the bottom left corner due to the confusing boundaries of the anatomical structures. This error is then corrected by utilizing functional information from PET images (where such anatomical deviations show little contrast) to form a multimodal fusion network [Fig. 9(b)]. By incorporating CT images, the segmentation boundary is further smoothed, achieving the best possible performance.

IV. Conclusion

Based on the network performance comparison, we empirically demonstrate several findings. First, comparison results between multimodality and single-modality networks in Sections III-A and III-B show that multimodal fusion networks perform better than single-modal networks. More interestingly, fusion networks based on synthetic low-quality images perform better than single-modality networks on high-quality images, at certain noise levels. This finding brings in new evidence for the benefit of multimodal imaging in medical applications in which one of the modalities can only provide images with limited quality, such as screening or low-dose scans. It is then a better option to utilize more than one modality for better analytics.

Second, comparison results of fusion strategies in Section III-A shows that for the task of tumor region segmentation using CNN, performing fusion within the network (at the convolutional layer or fully connected layer) is better than outside the network (at network output through voting), even when the voting weights are learned by using sophisticated classification algorithms such as random forest. As voting is commonly used by multimodal analytics, this conclusion could provide empirical guidance for the corresponding model design (e.g., consider an integrated multimodal framework through registration rather than voting).

Third, modality combination results in Section III-C show that multimodal fusion networks can take advantage of the additional anatomic or physiological characterizations provided by different modalities, even if the extra modality can only provide limited contrast in the target region. This conclusion is in accordance with "weak learnability" in the field of ensemble learning [44], indicating that as long as a learner (or source of information, as the imaging modality in this context) can perform slightly better than random guessing, it can be added into a learning system to improve its performance.

Although we have only tested the framework on a single dataset using one set of simple network structures, most of the current conclusions we draw from the empirical results are not dependent upon the exact data used. We are aiming to test more network structures including end-to-end semantic segmentation networks, on datasets with more types of modalities in future work.

In addition, as fully CNN, such as U-Net [45], has been widely used in medical image analysis especially semantic segmentation, we performed the same segmentation task using U-Net based on Type-I fusion scheme. Structure of U-Net used in this paper consists of four convolution layers for encoding and four deconvolution layers for decoding, in accordance with input image size (128×128). Other model parameters and implementation details can be found in our previous work [46]. Comparison between the segmentation result from U-Net-based and CNN-based fusion networks (all Type-I) shows that these two methods achieved very similar performance, with relative difference $<0.5\%$. This result shows that with the same fusion scheme, actual performance is similar for different segmentation methods (e.g., between patch-based and encoder-decoder-based methods). Further, it shows that fusion schemes introduced in this paper is not dependent on the implementation of segmentation, thus it can serve as a general design rule for multimodal image segmentation.

Our algorithmic architecture (three fusion strategies) only covers supervised, classification-purposed methods. Yet we also note that there exist unsupervised methods in medical image analysis such as gradient flow-based methods for image segmentation [47], as well as well-established deformable image registration algorithms [48]. These unsupervised methods can also be applied to multimodal images, while their fusion schemes can be studied by an extension of the current framework.

While the empirical study is performed on a well-registered image dataset, we recognize that registration across different imaging modalities is a vital part of any fusion model. All three types multimodal fusion networks used in this paper assumes good voxel-level correspondence, while erroneous registration across different modalities in an incoming patient can lead to dramatically decreased prediction performance within the misaligned region, depending on the number of modalities affected by the misalignment and its severity. This limitation has inspired us for a plan to develop an integrated

framework consisting of iterative segmentation and registration through alternative optimization, with shared multimodal image features.

ACKNOWLEDGMENT

The authors would like to thank MGH and BWH Center for Clinical Data Science for computational resources and Dr. J. Thrall for proofreading this paper.

REFERENCES

[1] T. Beyer *et al.*, "A combined PET/CT scanner for clinical oncology," *J. Nucl. Med.*, vol. 41, no. 8, pp. 1369–1379, 2000.

[2] U. Bagci *et al.*, "Joint segmentation of anatomical and functional images: Applications in quantification of lesions from PET, PET-CT, MRI-PET, and MRI-PET-CT images," *Med. Image Anal.*, vol. 17, no. 8, pp. 929–945, 2013.

[3] M. Czisch *et al.*, "Altered processing of acoustic stimuli during sleep: Reduced auditory activation and visual deactivation detected by a combined fMRI/EEG study," *NeuroImage*, vol. 16, no. 1, pp. 251–258, 2002.

[4] A. Teramoto, H. Fujita, O. Yamamuro, and T. Tamaki, "Automated detection of pulmonary nodules in PET/CT images: Ensemble false-positive reduction using a convolutional neural network technique," *Med. Phys.*, vol. 43, no. 6, pp. 2821–2827, 2016.

[5] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[6] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, nos. 1–2, pp. 1–39, 2010.

[7] A. P. James and B. V. Dasarathy, "Medical image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 19, pp. 4–19, Sep. 2014.

[8] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[9] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[10] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2016.

[11] J. H. Thrall *et al.*, "Artificial intelligence and machine learning in radiology: Opportunities, challenges, pitfalls, and criteria for success," *J. Amer. College Radiol.*, vol. 15, no. 3, pp. 504–508, 2018.

[12] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, Jan. 2017.

[13] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1240–1251, May 2016.

[14] P. F. Christ *et al.*, "Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2016, pp. 415–423.

[15] R. Rouhi, M. Jafari, S. Kasaei, and P. Keshavarzian, "Benign and malignant breast tumors classification based on region growing and CNN segmentation," *Exp. Syst. Appl.*, vol. 42, no. 3, pp. 990–1002, 2015.

[16] S. Wang *et al.*, "Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation," *Med. Image Anal.*, vol. 40, pp. 172–183, Aug. 2017.

[17] X. Feng, J. Yang, A. F. Laine, and E. D. Angelini, "Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2017, pp. 568–576.

[18] J. Wang *et al.*, "Technical note: A deep learning-based autosegmentation of rectal tumors in MR images," *Med. Phys.*, vol. 45, no. 6, pp. 2560–2564, 2018.

[19] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 1–8.

[20] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37. Lille, France, 2015, pp. 1083–1092.

[21] Y. Kang, S. Kim, and S. Choi, "Deep learning to hash with multiple representations," in *Proc. IEEE 12th Int. Conf. Data Min.*, 2012, pp. 930–935.

[22] L. Ge, J. Gao, X. Li, and A. Zhang, "Multi-source deep learning for information trustworthiness estimation," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, Chicago, IL, USA, 2013, pp. 766–774.

[23] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Multi-view perceptron: A deep model for learning face identity and view representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–9.

[24] G. Carneiro, J. Nascimento, and A. P. Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 652–660.

[25] T. Xu, H. Zhang, X. Huang, S. Zhang, and D. N. Metaxas, "Multimodal deep learning for cervical dysplasia diagnosis," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2016, pp. 115–123.

[26] H.-I. Suk, S.-W. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, Nov. 2014.

[27] M. Liang, Z. Li, T. Chen, and J. Zeng, "Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 12, no. 4, pp. 928–937, Jul./Aug. 2015.

[28] N. M. Correa, T. Eichele, T. Adali, Y.-O. Li, and V. D. Calhoun, "Multi-set canonical correlation analysis for the fusion of concurrent single trial ERP and functional MRI," *NeuroImage*, vol. 50, no. 4, pp. 1438–1445, 2010.

[29] M. Lorenzi *et al.*, "Multimodal image analysis in Alzheimer's disease via statistical modelling of non-local intensity correlations," *Sci. Rep.*, vol. 6, Apr. 2016, Art. no. 22161.

[30] X. Xu, D. Shan, G. Wang, and X. Jiang, "Multimodal medical image fusion using PCNN optimized by the QPSO algorithm," *Appl. Soft Comput.*, vol. 46, pp. 588–595, Sep. 2016.

[31] G. Bhatnagar, Q. M. J. Wu, and Z. Liu, "Directive contrast based multimodal medical image fusion in NSCT domain," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1014–1024, Aug. 2013.

[32] R. Singh and A. Khare, "Fusion of multimodal medical images using Daubechies complex wavelet transform—A multiresolution approach," *Inf. Fusion*, vol. 19, pp. 49–60, Sep. 2014.

[33] Y. Yang, "Multimodal medical image fusion through a new DWT based technique," in *Proc. 4th Int. Conf. Bioinformat. Biomed. Eng.*, 2010, pp. 1–4.

[34] X. Zhu, H.-I. Suk, S.-W. Lee, D. Shen, and D. Shen, "Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 607–618, Mar. 2016.

[35] S. Klein *et al.*, "Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information," *Med. Phys.*, vol. 35, no. 4, pp. 1407–1417, 2008.

[36] H. Cai *et al.*, "Probabilistic segmentation of brain tumors based on multi-modality magnetic resonance images," in *Proc. 4th IEEE Int. Symp. Biomed. Imag. Nano Macro*, 2007, pp. 600–603.

[37] M. Vallières, C. R. Freeman, S. R. Skamene, and I. E. Naqa, "A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities," *Phys. Med. Biol.*, vol. 60, no. 14, pp. 5471–5496, 2015.

[38] K. Clark *et al.*, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, 2013.

[39] T. E. E. S. N. W. Group, "Soft tissue and visceral sarcomas: ESMO Clinical practice guidelines for diagnosis, treatment and follow-up," *Ann. Oncol.*, vol. 25, no. S3, pp. 102–112, 2014.

[40] R. Grimer, I. Judson, D. Peake, and B. Seddon, "Guidelines for the management of soft tissue sarcomas," *Sarcoma*, vol. 2010, May 2010, Art. no. 506182.

[41] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[42] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[43] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[44] R. E. Schapire, "The strength of weak learnability," *Mach. Learn.*, vol. 5, no. 2, pp. 197–227, 1990.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[46] M. Zhang, X. Li, M. Xu, and Q. Li, "RBC semantic segmentation for sickle cell disease based on deformable U-Net," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2018, pp. 695–702.

[47] G. Li *et al.*, "3D cell nuclei segmentation based on gradient flow tracking," *BMC Cell Biol.*, vol. 8, no. 1, p. 40, 2007.

[48] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Trans. Med. Imag.*, vol. 32, no. 7, pp. 1153–1190, Jul. 2013.