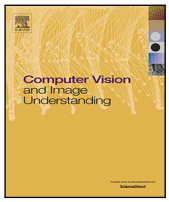
Computer Vision and Image Understanding 202 (2021) 103085

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Multi-modal semantic image segmentation

Akila Pemasiri *, Kien Nguyen, Sridha Sridharan, Clinton Fookes

Image and Video Research Lab, Queensland University of Technology, 2 George Street, GPO Box 2434, Brisbane, QLD 4001, Australia

ARTICLE INFO

Communicated by Smith John

Keywords:
Segmentation
X-ray
Mask R-CNN
Neural networks

ABSTRACT

We propose a modality invariant method to obtain high quality semantic object segmentation of human body parts, for four imaging modalities which consist of visible images, X-ray images, thermal images (heatmaps) and infrared radiation (IR) images. We first consider two modalities (i.e. visible and X-ray images) to develop an architecture suitable for multi-modal semantic segmentation. Due to the intrinsic difference between images from the two modalities, state-of-the-art approaches such as Mask R-CNN do not perform satisfactorily. Insights from analysing how the intermediate layers within Mask R-CNN work on both visible and X-ray modalities have led us to propose a new and efficient network architecture which yields highly accurate semantic segmentation results across both visible and X-ray domains. We design multi-task losses to train the network across different modalities. By conducting multiple experiments across visible and X-ray images of the human upper extremity, we validate the proposed approach, which outperforms the traditional Mask R-CNN method through better exploiting the output features of CNNs. Based on the insights gained on these images from visible and X-ray domains, we extend the proposed multi-modal semantic segmentation method to two additional modalities; (viz. heatmap and IR images). Experiments conducted on these two modalities, further confirm our architecture's capacity to improve the segmentation by exploiting the complementary information in the different modalities of the images. Our method can also be applied to include other modalities and can be effectively utilized for several tasks including medical image analysis tasks such as image registration and 3D reconstruction across modalities.

1. Introduction

Semantic segmentation underpins many computer vision applications across several domains since region of interest identification uplifts the accuracy of the end results (Pemasiri et al., 2019b). Many of these applications utilize in addition to visible images, other imaging modalities such as X-ray images and IR images. Algorithms for computer vision applications, including human action recognition and patient analysis (Pemasiri et al., 2019a; Ji et al., 2018; Khowaja and Lee, 2019), 3D reconstruction (Elanattil et al., 2018; Kundu et al., 2014) and object detection (Salscheider, 2019) based on visible images use semantic segmentation as a subroutine of the process. When considering the X-ray images, establishing semantic segmentation is vital for processes such as joint space width assessment, bone age assessment and preoperative planning (Bullock et al., 2019; Wu and Mahfouz, 2016; Hržić et al., 2019). Thermal images (heatmaps) and infrared (IR) images are also employed in computer vision applications across many domains including industrial inspection, medical image analysis and surveillance where these applications also require semantic segmentation as part of the process (Duarte et al., 2014; Qiao et al., 2017).

Recent advances in neuro-computing have motivated the application of neural networks for semantic segmentation tasks, resulting in robust and impressive baseline frameworks including Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2015), and Mask R-CNN (He et al., 2017). However, the existing work on semantic segmentation has been aimed only on a single selected image modality (Bates et al., 2019; Hu et al., 2019; Li et al., 2018a; Wollmann et al., 2019; Bullock et al., 2019).

Multi-modal information processing is beneficial in many computer vision applications as the information available in different modalities complement each other (Pemasiri et al., 0000; Jian et al., 2017; Nakamura et al., 2013). For an example, in medical image analysis, visible images capture the external features of an organ (e.g., skin information) where as X-ray images capture the information about internal parts (e.g., bones and muscles) and the combined information is useful for medical analysis. In addition to these modalities, thermal images present temperature distribution of the surface which can capture pathophysiologic information related to physical conditions (Duarte et al., 2014; Sonkusare et al., 2019). However, the development of these cross modality applications demand robust lower-level computer vision algorithms which can be applied on multi-modal images. With

* Corresponding author.
E-mail address: akila.10@cse.mrt.ac.lk (A. Pemasiri).

https://doi.org/10.1016/j.cviu.2020.103085
Received 5 November 2019; Received in revised form 24 June 2020; Accepted 25 August 2020
Available online 5 September 2020

this requirement the question to be addressed is how to combine the complementary information from multiple sources considering the intrinsic differences in their nature. In the case of semantic segmentation, the open question is how to distill information from multiple imaging modalities in a single and unified architecture.

Compared with single-modality image processing, multi-modal image processing exhibits some inherent challenges, arising from potential dissimilarity where the regions in two images can be differently textured or one image may be textured while the other is homogeneous (Torabi and Bilodeau, 2013). In addition, the level of information presented in each image modality is different. For instance, considering images of a hand from visible, X-ray, heatmap and IR modalities, visible images contain details of skin, nails, clothes and accessories, whereas X-ray images contain details of bones and muscles, while heatmap images contain the temperature distribution of the surface and IR images contains the details of the degree of infrared wave penetration through the objects in the image (Fig. 1).

In this paper, we present an architecture which uses a single model to establish pixel wise semantic object segmentation of human body parts (i.e. multi-class), regardless of the modality, (i.e. a visible or an X-ray image). In the context of this paper we use the term "Semantic segmentation" to associate the pixels in the image, which belong to the object of interest (e.g., identifying components of human upper extremity with the class labels of "elbow", "forearm", "hand", "humerus", "shoulder" and "wrist") with the correct class labels. Using a single model for semantic segmentation of multiple modalities, not only seamlessly incorporates all sub-tasks into one, but also reduces the number of parameters and improves the accuracy by enabling end-to-end training and inference. Training a single network on multi-modal images will integrate complementary information that is presented in each modality, and thus will result in robust kernels which are capable of capturing information from all the modalities under consideration. In developing our multi-modal semantic segmentation framework, we exploit the convolution neural networks' capability of learning patterns at early stages (Li et al., 2018b; Long et al., 2014), including the patterns in different modalities. Our proposed approach has the advantage that the models trained on modalities where the samples are abundant (e.g., visible images) can be used on other image modalities (e.g., X-ray) which are hard to acquire. In our approach, we utilize the capabilities of Mask R-CNN by exploiting the kernels which generate the most informative features to achieve effective multi-modal and robust object segmentation.

We also demonstrate the effectiveness of our model in exploiting complementary information across other imaging modalities (i.e. heatmaps and IR images), for single-class image segmentation. The reason to use single-class semantic segmentation for this evaluation as opposed to multi-class, multi-modal segmentation which we demonstrate with visible and X-ray images, is the limited availability of datasets, which we will discuss in detail in Section "Experiments" of this paper. Our proposed architectural innovation however, is generic and can be used to perform semantic multi-class image segmentation with any combination of multi-modal data, using a single deep learning model.

Medical images of human upper extremity are used in detecting many pathological conditions including Rheumatoid arthritis, upper limb fractures (e.g., proximal humeral fracture, humeral shaft fracture, supracondylar fracture), dislocations (e.g., shoulder dislocation, radial head dislocation and carpal dislocation) and deformities (e.g., radial club hand, ulnar dysplasia and radiohumeral synostosis). Due to these wide range of applications of human upper extremity for which we have large amounts of X-ray, heatmap and IR images available, we have chosen this body part to demonstrate the efficacy of our method; the method we propose however, can easily be applied to the other body parts as well.

For the X-ray images, we use the **mu**sculoskeletal **ra**diograph (MURA) dataset (Rajpurkar et al., 2017). For visible images, we have collected our own dataset, which contains the corresponding visible human body parts (i.e. elbow, forearm, hand, humerus, shoulder and wrist). To ensure that our database is representative of a typical healthcare environment, we have collected visible images under challenging conditions including varying lighting conditions, different resolutions and different backgrounds. This database will be made publicly available to enable researchers to replicate our results in this paper. In addition to releasing the database, code for the robust object segmentation method introduced in this paper, as well as the groundtruth region annotation for the MURA X-ray image dataset for the upper extremity of the human body which we have manually annotated will also be available for researchers to follow-up on our work. In addition using the HandNet dataset (Wetzler et al., 2015), we obtained the heatmap images and IR images of human hands. As this dataset does not contain other components of the upper extremity (i.e. elbow, forearm, humerus, shoulder and wrist), we use the HandNet dataset (Wetzler et al., 2015) to validate our method for other modalities and to demonstrate its applicability in other modalities.

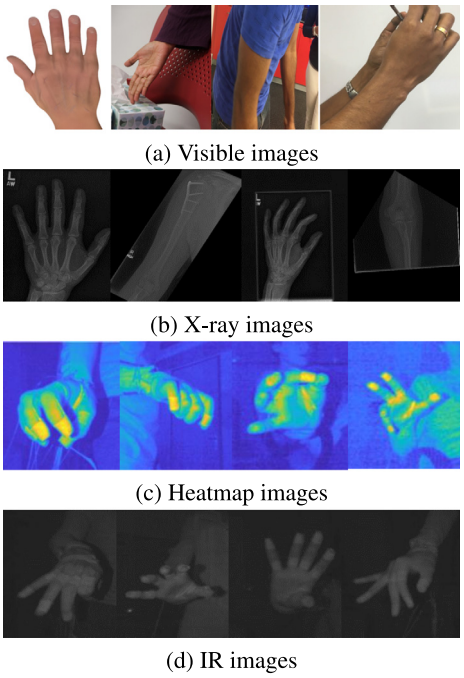


Fig. 1. Sample images from visible (first row), X-ray (second row), heatmap (third row) and IR (fourth row) images, illustrate the difference in visual texture and level of information.

The major contributions of our paper are summarized below:

1. We have developed a CNN based modality invariant semantic image segmentation framework, where the most informative layers of CNNs are effectively utilized as the features for class prediction, region identification and mask identification across multi-modal images.
2. Our multimodal semantic segmentation implementation has outperformed the widely used segmentation framework "Mask R-CNN" for the visible, X-ray, thermal and IR images, and this method can be extended to include other image modalities such as CT scan and MRI.
3. Our code and the dataset we used in this work are released so that the researchers are able to replicate our results and further advance the area of multi-modal semantic image segmentation of medical images.[1]

[1] https://github.com/PemasiriAkila/Multi-modal-Semantic-Image-Segmentation.

(a) Input image. (b) Traditional segmentation (c) Semantic segmentation

Fig. 2. Results of traditional segmentation methods and semantic segmentation for an input X-ray image.

The remainder of the paper is organized as follows. In Section 2 we analyse the recent literature on object detection and segmentation. Section 3 describes the proposed system and its subsections elaborate on each component of the system. It should be noted that in the method development we used only the visible and X-ray images, which was then validated for other imaging modalities. In Section 4, we present our experimental results for multi-class multi-modal semantic segmentation using visible and X-ray images as well single-class multi-modal semantic segmentation using visible, X-ray, heatmap and IR images. In Section 5, we conclude the paper with a discussion of the results on the presented method.

2. Related work

Object detection and segmentation literature has progressed through the decades. Early stages of image segmentation include approaches that use the low level image features such as brightness and texture gradient (Martin et al., 2003). However, when considering the object segmentation in contrast to the image segmentation, the former requires more details on the structure and the shape of the object category. Many of the recent approaches for object segmentation are preceded by object detection, where the object boundary identification is used as a prior for the detection task (Elgammal et al., 2000; Gonfaus et al., 2010; Mittal and Paragios, 2004). Therefore efforts on improving the detection process have a paramount effect on the improvement of the segmentation process.

It is notable that neural network based approaches have produced significant improvement for object detection and segmentation tasks (Rowley et al., 1998; Sermanet et al., 2013; Vaillant et al., 1994). However, early approaches based on CNNs dealt with specific object categories in constrained applications such as faces (Rowley et al., 1998; Vaillant et al., 1994) and pedestrians (Sermanet et al., 2013). In the more recent literature, rich feature hierarchies for accurate object detection and semantic segmentation (Girshick et al., 2014), have been introduced such as the concept of Regions with CNN features coined as "R-CNN" with the aim to accomplish higher accuracy values for object detection. They have first extracted the region proposals for the whole image, following which CNN based features are extracted on each of these regions which are then subjected to SVM based classification.

One major drawback with R-CNN is that to extract features of each candidate region proposal, it requires a forward pass of the CNN. To overcome this a number of approaches known as "Fast R-CNN" (Girshick, 2015) has been suggested, where a Region of Interest (RoI) pooling layer is used to extract feature vectors for object proposals. Feature vectors use the feature map generated by the forward pass of the input image over the CNN. To alleviate the bottleneck caused by having to obtain the candidate region proposals using a separate selective search process in Fast R-CNN, "Faster R-CNN" (Ren et al., 2015) has been introduced. In Faster R-CNN, a single CNN is used to estimate the candidate proposals as well as to extract features for the classification.

The main objective in developing the R-CNN, Fast R-CNN and Faster R-CNN architectures are to perform effective object detection and classification. Mask R-CNN (He et al., 2017) has extended the Faster R-CNN by adding another branch for object mask detection resulting in state-of-art accuracy for object segmentation. However, all these approaches suggested in the literature for detection and segmentation have mainly focused on visible images.

When considering the other imaging modalities such as X-ray images, the existing methods for segmentation are limited to those which have targeted on identifying the components of the musculoskeletal system (e.g., identifying bone regions of a given X-ray image or extracting the whole body part that has been captured in the X-ray image) (Ali et al., 2006; Feng, 2006; Isard and Blake, 1998; Manos et al., 1994; Stolojescu-CriŞan and Holban, 2013) (See for example Fig. 2(b), where different colours indicate the different regions resulting from segmentation). They do not segment all the components of a particular body part (e.g., identifying all the components of the wrist) (Fig. 2), which limits the semantic nature of the segmentation results. The existing IR and heatmap segmentation methods are based on hand engineered features (Duarte et al., 2014; Qiao et al., 2017) such as pixel intensities which make the algorithms vulnerable to noise while generating less robust results. In addition, all these methods are targeted on modality specific features which limit their ability to be used with other imaging modalities.

In contrast to these previous approaches, we develop an architecture to detect the objects and carry out the segmentation while being invariant to the modality of the image, whether they are visible, X-ray, thermal or IR. Our method enables a deeper semantic segmentation which can provide better localization of regions of interest, as illustrated in Fig. 2c for the X-ray image. To the best of our knowledge, this is the first work to consider model invariant semantic segmentation for multi-modal images.

3. Methodology

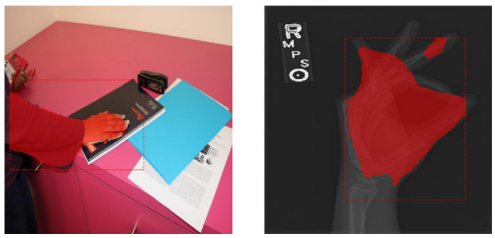
In this section we provide an overview of Mask R-CNN and then elaborate the insight of deriving our method based on the observations on visible and X-ray image segmentation.

3.1. Mask R-CNN

As described in Section 2, state-of-the-art object detection and segmentation are driven by CNNs and Mask R-CNN has depicted robust results for both these tasks. In this section, we will elaborate strengths and weaknesses of Mask R-CNN and how we can exploit the Mask R-CNN's strengths to make it invariant for object segmentation for visible and X-ray images.

Mask R-CNN has (He et al., 2017) been extended from Faster R-CNN (Ren et al., 2015), where the objective of the Faster R-CNN is to generate accurate object bounding boxes, Mask R-CNN provides a refined segmentation of the objects. In addition to the added branch for segmentation, Mask R-CNN introduces the RoIAlign layer which aligns the regions of the feature map along with the regions of the original image.

The foundation of our approach is based on the observations that we encountered while attempting to use Mask R-CNN for object segmentation. When the original Mask R-CNN was trained on a dataset which consisted of human body parts in the visible image modality, it was identified that though it can generalize well for visible images, it does not generate masks of adequate quality for X-ray images (Fig. 3). In most of the cases, it was not detecting all the components of the X-ray object. Conversely, when a model trained on X-ray images was tested on a set of visible images, one major observation that was made is that it was erroneously detecting other objects in the environment as human body parts (Fig. 4).



(a) Tested on a visible image. (b) Tested on an X-ray image.

Fig. 3. Results that were obtained from the model trained on visible images illustrating its poor performance in mask identification in X-ray images.

(a) Tested on an X-ray image. (b) Tested on a visible image.

Fig. 4. Results that were obtained from the model trained on X-ray images illustrating its poor performance in mask identification in visible images.

Fig. 5. Results on images which were tested on the model trained on multi-modal image dataset.

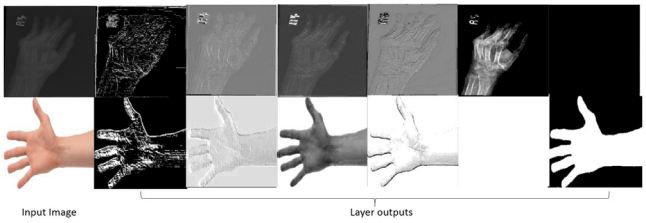


Fig. 6. Visualization of layer outputs of Mask R-CNN for a visible image and for an X-ray image, the first column depicts the input images. The 2nd, 3rd, 4th and 5th columns show the layer outputs where informative features have been generated for both X-ray and visible images, 6*th* column shows an example for a layer output where for the X-ray image the generated feature is informative but for the visible image it does not, and the 7th column shows an example for the opposite scenario.

An intuitively obvious solution to overcome the above problem is to use a dataset, which contains the images from both the modalities. However, though this approach results in the increased quality of the output compared to the models that were trained on a single modality (Figs. 3 and 4), the newly generated masks still suffer from the problems of not recognizing the complete object or mistakenly identifying the parts in the image that do not belong to the object (Fig. 5).

Our next step in finding the solution to this problem was to gain a greater insight into the architecture of Mask R-CNN. When the output of the layers of Mask R-CNN were inspected, a simple yet crucial observation was made. That is, only some of the kernels in Mask R-CNN generates discerning features for both the image modalities. A subset of feature maps, extracted from some random kernels of the Mask R-CNN is depicted in Fig. 6. The first column of Fig. 6 shows the input image, where the first row of it is the input X-ray image and the second row shows the input visible image. The rest of the columns depict the feature output from some random kernels. It can be seen that some kernels have generated features for both modalities, whereas other kernels generated sensible features for only one of the two modalities. The kernels that were generating sensible features for both image modalities were consistent for all the body parts under our consideration. Furthermore, to validate our observation, Mask R-CNN was trained 3 times separately, with the pretrained weights on the COCO dataset (Lin et al., 2014) for 100 epochs, we observed that the layers which generated features for the modalities remained the same.

3.2. Modality invariant segmentation

Motivated by the observations described in Section 3.1, we introduce a new architecture, which generates robust results for visible images as well as X-ray images.

As described above, for a set of initial images we identified the kernels in the network that created discerning features for both visible and X-ray images. Based on these kernel numbers, our architecture incorporates a new feature concatenation layer where the resized feature outputs of the layers that generate features for both the modalities from Mask R-CNN are concatenated. The input to the our model (Fig. 7) is images (either visible or X-ray) and the input images are passed through the network, where the features are generated at every level. Then using bilinear interpolation, the features are resized to the size of 224×224 which is the input image size, and the feature outputs that are related to the kernels which generated discerning features are concatenated together before feeding to the RoI pooling layer.

Depending on the backbone architecture used in the Mask R-CNN, the dimension of the feature concatenation layer vary. The architecture that is depicted in Fig. 7 corresponds to the ResNet101-C4 (He et al., 2016) architecture. Following the previous work on Faster R-CNN we employed ResNet architecture and ResNeXt architecture (Xie et al., 2017) with depth of 50 or 101 layers. Following the commonly adopted notation (Dai et al., 2016; He et al., 2016; Huang et al., 2017b; Shrivastava et al., 2016) we denote the ResNet architecture with 50 layers where the features are extracted from 4th stage convolution layer as "ResNet-50-C4". In addition, we analyse Feature Pyramid Network (FPN) (Lin et al., 2017) in combination with ResNet architecture.

When the features generated by Resnet-50-C4, Resnet-101-C4, ResNet-50-FPN, ResNet-101-FPN, ResNeXt-50-C4, ResNeXt-101-C4, ResNeXt-50-FPN and ResNeXt-101-FPN were investigated it was identified that there are 2100, 3435, 2421, 3606, 2241, 3527, 2627, and 3677 feature outputs that create features for both visible and X-ray images. These layers were automatically selected by thresholding the pixel values of output features. In the current implementation, to select the kernels which generate the discerning features, first we crop the images based on the bounding boxes and then pass through the network. From the features generated by each kernel we eliminate the kernels which have generated features with either 0s or 1s. The concatenation layer is followed by a convolutional layer, which is then subjected to RoI align layer, which then get segregated into mask prediction branch and to class and bounding box prediction branch.

We use a multi-task loss function to train the network,

$$L = L_{class} + \lambda L_{box} + \mu L_{mask}, \tag{1}$$

where L_{class} denotes the classification loss, L_{box} denotes the loss defined on bounding box identification and L_{mask} denotes the loss defined on mask identification. The combined loss is calculated as in Eq. (1), where λ and μ are the hyper-parameters that are used for maintaining

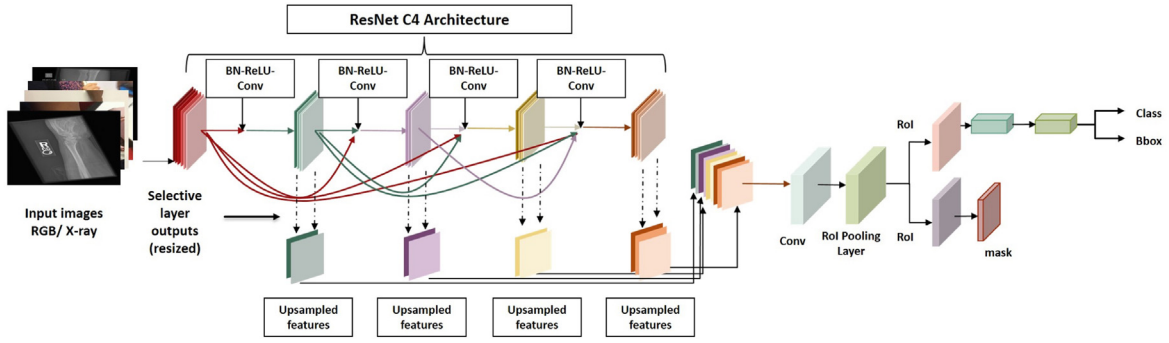


Fig. 7. Our architecture, where the selected layer outputs are resized and fused to generate the new layer, which is then subjected for RoI extraction, mask extraction and class label extraction.

the balance between the three losses that are aggregating for the final loss value.

The L_{class} is computed based on the softmax over 7 outputs of the fully connected layer in the classification branch. The last layer of the classification branch contains 7 outputs as the dataset we have used 6 classes and the default background layer is also taken into consideration. For each given Region of Interest (RoI), a probability distribution is associated using a softmax over the 7 outputs of the fully connected layer. For a given image, for a given Region of Interest (RoI) with groundtruth class label u, the L_{class} is the negative log softmax value for the groundtruth class label Eq. (2),

$$L_{class} = -log\left(p_u\right). \tag{2}$$

A bounding box is defined using four parameters t_x, t_y, t_w and t_h which denote the x coordinate of the right corner of the bounding box, y coordinate of the right corner of the bounding box, width of the bounding box and height of the bounding box respectively. Let $t^u = \left(t^u_x, t^u_y, t^u_w, t^h_x\right)$ be the bounding box parameter tuple of instance u of groundtruth annotation, and t^v be the predicted tuple the L_{box} is defined as,

$$L_{box} = \sum_{i\epsilon x,y,w,h} smooth_{L1}\left(t^v_i - t^u_i\right). \tag{3}$$

In Eq. (3), the term $smooth_{L1}$ is defined as,

$$smooth_{L_1}(r) = \begin{cases} 0.5q^2 & if\ |q| < 1 \\ |q| - 0.5 & otherwise. \end{cases} \tag{4}$$

It should be noted that the L_{box} is not calculated for the *background* class.

In this work, the mask is defined as a binary array where the foreground pixels of the objects are marked. The mask branch output consists of K masks in $m\times m$ resolutions, where for each of the entries in $m \times m$ array a per-pixel sigmoid is applied and when an object belongs to the class k, the L_{mask} is defined as in Eq. (5),

$$L_{mask} = -\frac{1}{m^2} \sum_{1\le i,j\le m} \left[y_{ij} log(\hat{y}^k_{ij}) + \left(1 - y_{ij}\right) log\left(1 - \hat{y}^k_{ij}\right)\right]. \tag{5}$$

In Eq. (5), y_{ij} is the groundtruth label of the cell (i, j) where as the $\hat{y}_{ij}$ is the predicted mask associated with the same cell.

4. Experiments

The experiments that we have conducted are divided into three main sections: (1) Multi-class semantic segmentation for visible and X-ray images, (2) Single-class segmentation for multi-modal images (i.e. visible, X-ray, heatmap and IR images) and (3) Segmentation on CT images. The first set of experiments (multi-class semantic segmentation on visible and X-ray images) are conducted to demonstrate our model's capacity to exploit features presented in multimodal images and the second set of experiments are conducted to validate our method on other imaging modalities (i.e. heatmap images and IR images). However, it should be noted that the second set of experiments were limited to the single class segmentation due to limited availability of datasets. Thirdly, we conducted the experiments to evaluate our method's ability in handling CT images (Yang et al., 2018), using the AAPM Thoracic Auto-segmentation datasets which contains 5 classes (i.e. Esophagus, Heart, Left lung, Right lung and Spinal cord) and their train/test split was used for the evaluations.

For the training of all the networks, we use the same learning rates and the weight decay values that have been used in the preceding approaches including Fast R-CNN (Girshick, 2015; Ren et al., 2015). We use the pretrained backbone architectures where the training has been performed on COCO (Lin et al., 2014) dataset and the convolution layer was initiated with random values. The learning rate, weight decay and the momentum were set to 0.02, 0.00001 and 0.9 accordingly.

To benchmark our method, we used Mask R-CNN (He et al., 2017).

Evaluation Metrics

For the first and second evaluations as the evaluation metrics, we used average precision (AP), which is calculated using Intersection over Union (IoU),

$$IoU = \frac{Area\left(b_p \cap b_t\right)}{Area\left(b_p \cup b_t\right)}, \tag{6}$$

based on a defined threshold.

In Eq. (6), b_p is the prediction, which is the bounding box prediction or the mask prediction, and b_t is the corresponding groundtruth value. The precision is calculated as,

$$Precision = \frac{TP}{TP + FP}, \tag{7}$$

where a detection is defined as a true positive (TP), if $IOU \ge threshold$ and as a false positive (FP), if $IOU < threshold$. In this evaluation we used the thresholds of 0.5 and 0.75 and AP_{50} denotes the calculated average precision when the $threshold$ is set to 50%.

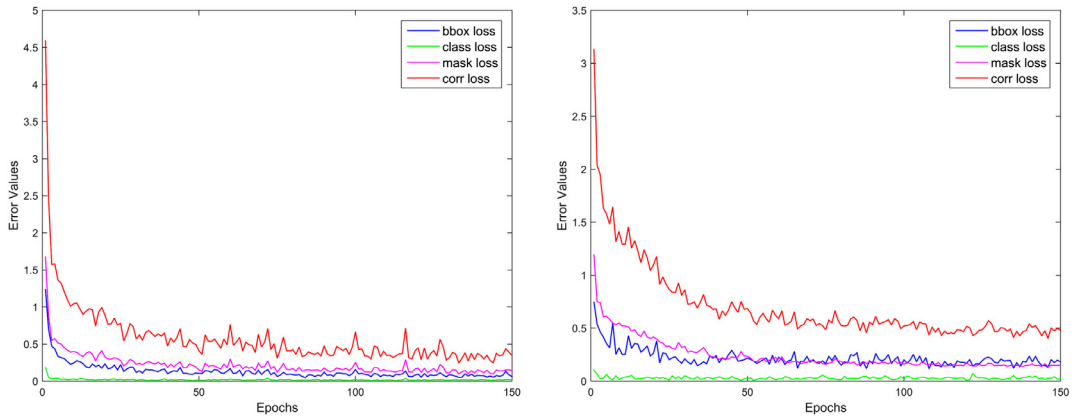
For the experiments on the CT image dataset we used Dice score Eq. (8) which is the most common evaluation protocol used for the CT segmentation accuracy (Yang et al., 2018).

$$Dice = \frac{2 * Area\left(b_p \cap b_t\right)}{Area\left(b_p\right) + Area\left(b_t\right)}, \tag{8}$$

In Eq. (8), the numerator estimates the area of overlap between the groundtruth segmentation (b_t) and the predicted segmentation (b_p), while the denominator estimates the total number of pixels in both the segmentations.

4.1. Multi-class segmentation on X-ray and visible images

The evaluations that we describe in this section were performed with two objectives. Firstly, selecting the optimum backbone for feature extraction is a critical design decision that affects the accuracy level.



(a) Training losses. (b) Validation losses.

Fig. 8. Training and validation losses for the our method when it was trained on a dataset which contained both visible and X-ray images.

Therefore, in this section we employee a number of backbone architectures (Resnet-50, Resnet-101, ResNeXt-50 and ResNeXt-101) and estimate the corresponding precision values, to select which backbone architecture yields the best accuracy values. Secondly, we assess our model's capability in robustly segmenting different parts in human upper extremity, and evaluate the benefits of using multi-modal images in a single network modal. In this section we also evaluate the effectiveness of using different values for the parameters in the loss function Eq. (1).

Datasets

For the X-ray image dataset we used the MURA dataset (Rajpurkar et al., 2017) which contains 8 bit X-ray images of the human upper extremity. The MURA dataset contains X-ray images of elbow, forearm, hand, humerus, shoulder and wrist. For this dataset we generated manually annotated mask groundtruth. For the visible images, we generated a dataset which contains visible images of the same body parts that are included in MURA dataset and the corresponding groundtruth masks. The visible image dataset contains the images captured under challenging conditions including different backgrounds, resolutions, occlusions and illumination conditions as well as the subjects wearing different accessories such as wrist watches, jackets and bangles. The total number of images that were captured is 1400, where there were 817 annotations of the hand, 908 annotations for the wrist, 1023 annotations for the forearm, 986 annotations for the elbow, 807 annotations for the humerus and 722 annotations for the shoulder. However, it should be noted that the visible images and X-ray images are not captured on the same subjects as these two sets were constructed independently, and thus the training on multi-modal images is not affected by the subject related features. The training that we have carried out can easily be adapted to the simpler case, where the images captured are on the same subjects. From the annotated X-ray images, we extracted a subset such that it contains a comparable number of annotations to the number of annotations for each class in the visible image dataset. For each of the classes we randomly divided our dataset into 60%, 20% and 20%, which we used as the training set, validation test and the test set respectively.

Training and Testing on Single Modalities: First, we trained Mask R-CNN and our model separately on the visible image dataset and the X-ray image dataset, and were tested on a dataset of the same modality. It should be noted that the λ and μ which are the hyper-parameters of the loss function Eq. (1) were set to 1 following (He et al., 2017). The obtained results are recorded in Table 1, and it can be observed that our method surpasses the Mask R-CNN in small margins, for visible images, and for the X-ray images our method outperforms Mask R-CNN by large margins.

When the results and the intermediate layers which generated discerning features were analysed, it was identified that for X-ray images there is a large percentage of kernels which do not generate discerning features in Mask R-CNN. However, for the visible images Mask R-CNN has a high percentage of layers which generate discerning features. However, Mask R-CNN utilizes all the layers for the classification and mask identification tasks. In contrast our method uses the features which encode distinctive features. The difference in the two implementations has given a high marginal improved accuracy for X-ray images in our method, whereas the improvement of the accuracy for the visible images is limited.

Training and Testing on Multi-modal Datasets: We then trained both the modals (Mask R-CNN and our method) on the multi-modal dataset, which contains both X-ray and Visible images. Similar to the single modality training, we have divided the combined dataset for each body part into 60%, 20% and 20%, for training, validation and testing. The training losses and the validation losses that we obtained for when training our architecture on the multi-modal dataset are depicted in Fig. 8.

Some qualitative results that we obtained for both visible and X-ray images are depicted in Fig. 9. This illustrates the effect of selecting the discerning features in contrast to using all the features for successive operations. Moreover in work by Pemasiri et al. (2019a), the benefits of cross-dataset learning are demonstrated with possible applications while elaborating on how the architecture of the network can easily be fine-tuned according to the application under the consideration, specifically for static and dynamic analysis in a clinical context.

When the original Mask R-CNN and the our method were tested on the same datasets for mask identification, the obtained average precision values are indicated in Table 2. By analysing the table it can be understood that the we outperform the Mask R-CNN for both the X-ray and visible images. In addition, it can be identified that the features extracted from deeper and advanced networks can perform better comparatively. From Table 2 it can be understood that ResNeXt-101-FPN has yielded the highest average precision for both the modalities and for both architectures, except for the AP_{50} in visible modality on which the Mask R-CNN with backbone ResNeXt-101-C4 has performed at the best. However, when comparing the results of Mask R-CNN and our architecture it was identified that the later with all the backbone architectures have performed better than the best performing setting of Mask R-CNN.

The average precision for bounding box identification using the original Mask R-CNN and our-method, which have been trained on the same datasets is recorded in Table 3. It can be identified that the same

Table 1
Average precision on single image modality segmentation on Mask R-CNN and our method for mask identification.

Method	Backbone	Trained & tested on visible images		Trained & tested on X-ray images	
		AP_{50}	AP_{75}	AP_{50}	AP_{75}
Mask R-CNN	Resnet-50-C4	58.48	34.49	38.08	22.08
Mask R-CNN	Resnet-101-C4	61.18	34.78	44.73	24.20
Mask R-CNN	ResNet-50-FPN	59.83	32.75	40.81	23.03
Mask R-CNN	ResNet-101-FPN	64.51	32.36	46.74	25.92
Mask R-CNN	ResNeXt-50-C4	65.59	35.90	42.75	22.99
Mask R-CNN	ResNeXt-101-C4	67.39	35.56	45.47	23.93
Mask R-CNN	ResNeXt-50-FPN	68.61	39.23	46.49	24.72
Mask R-CNN	ResNeXt-101-FPN	70.44	40.85	47.13	25.87
Our method	Resnet-50-C4	59.32	35.46	54.45	29.65
Our method	Resnet-101-C4	61.81	35.52	56.98	35.26
Our method	ResNet-50-FPN	60.44	33.26	57.15	33.50
Our method	ResNet-101-FPN	64.94	32.58	59.50	33.04
Our method	ResNeXt-50-C4	66.02	36.05	61.41	35.90
Our method	ResNeXt-101-C4	67.78	35.66	62.52	31.79
Our method	ResNeXt-50-FPN	68.98	39.30	64.25	39.00
Our method	ResNeXt-101-FPN	**70.73**	**40.86**	**65.60**	**33.73**

Table 2
Average precision on visible images and X-ray images for image segmentation, where the models have been trained on multi-modal dataset.

	Backbone	Trained on multi-modal tested on visible images		Trained on multi-modal tested on X-ray images		Average	
		AP_{50}	AP_{75}	AP_{50}	AP_{75}	AP_{50}	AP_{75}
Mask R-CNN	Resnet-50-C4	39.1	27.4	37.7	21.8	38.4	24.6
Mask R-CNN	Resnet-101-C4	43	29.1	44.3	23.3	43.7	26.2
Mask R-CNN	ResNet-50-FPN	41.6	28.3	40.3	22.4	41.0	25.4
Mask R-CNN	ResNet-101-FPN	43.8	29.5	46.1	25	45.0	27.3
Mask R-CNN	ResNeXt-50-C4	41.9	28.7	42.1	22.8	42.0	25.8
Mask R-CNN	ResNeXt-101-C4	**45.6**	29.3	44.8	23.9	45.2	26.6
Mask R-CNN	ResNeXt-50-FPN	44.1	29.4	45.7	24.3	44.9	26.9
Mask R-CNN	ResNeXt-101-FPN	44.7	**29.6**	**46.2**	**25.3**	**45.5**	**27.5**
Our method	Resnet-50-C4	57.3	31.3	53.4	27.7	55.4	29.5
Our method	Resnet-101-C4	59.7	33.1	54.4	28.8	57.1	31.0
Our method	ResNet-50-FPN	58.2	32.4	53.1	28.2	55.7	30.3
Our method	ResNet-101-FPN	62.6	34.6	54.9	29.3	58.8	32.0
Our method	ResNeXt-50-C4	62.9	34.9	55.8	30	59.4	32.5
Our method	ResNeXt-101-C4	64.1	35.3	56.9	31.7	60.5	33.5
Our method	ResNeXt-50-FPN	65.1	35.7	58.2	32.4	61.7	34.1
Our method	ResNeXt-101-FPN	**66.8**	**37.1**	**58.9**	**33.5**	**62.9**	**35.3**

Table 3
Average precision on visible images and X-ray images for bounding box identification, where the models have been trained on multi-modal dataset.

	Backbone	Trained on multi-modal tested on visible images		Trained on multi-modal tested on X-ray images		Average	
		AP_{50}	AP_{75}	AP_{50}	AP_{75}	AP_{50}	AP_{75}
Mask R-CNN	Resnet-50-C4	42.10	29.05	42.05	25.46	42.07	27.26
Mask R-CNN	Resnet-101-C4	46.92	34.98	49.34	28.36	48.13	31.67
Mask R-CNN	ResNet-50-FPN	42.61	29.94	44.06	27.54	43.34	28.74
Mask R-CNN	ResNet-101-FPN	48.90	36.62	51.80	33.03	50.35	34.83
Mask R-CNN	ResNeXt-50-C4	44.62	34.50	48.63	27.83	46.63	31.17
Mask R-CNN	ResNeXt-101-C4	**50.48**	35.48	50.34	31.13	50.41	33.30
Mask R-CNN	ResNeXt-50-FPN	49.25	36.07	50.40	33.02	49.83	34.54
Mask R-CNN	ResNeXt-101-FPN	49.54	**39.33**	**52.95**	**33.62**	**51.24**	**36.48**
Our method	Resnet-50-C4	63.29	39.75	57.28	30.09	60.28	34.92
Our method	Resnet-101-C4	65.56	38.99	61.70	33.01	63.63	36.00
Our method	ResNet-50-FPN	63.34	33.13	56.44	30.81	59.89	31.97
Our method	ResNet-101-FPN	65.81	38.70	61.75	33.86	63.78	36.28
Our method	ResNeXt-50-C4	68.40	44.18	62.49	34.26	65.44	39.22
Our method	ResNeXt-101-C4	69.13	42.84	63.58	36.39	66.35	39.62
Our method	ResNeXt-50-FPN	69.27	41.36	63.95	36.91	66.61	39.13
Our method	ResNeXt-101-FPN	**69.80**	**46.31**	**64.16**	**36.94**	**66.98**	**41.63**

observations that were made for the mask prediction are presented in bounding box identification as well.

Hyper Parameter Selection for the Loss Function: In the loss function Eq. (1), λ and μ were set to 1, which is the parameter value that has been used in the original work of Mask R-CNN (He et al., 2017). Furthermore from the evaluations that were conducted it was identified that the deeper backbone architectures favourably affect the accuracy values. Then, to evaluate the effect of λ and μ values we conducted an ablative study by changing the parameter values. Since the bounding box identification and mask identification are interrelated, in this study we assigned the same value for λ and μ. The obtained results are

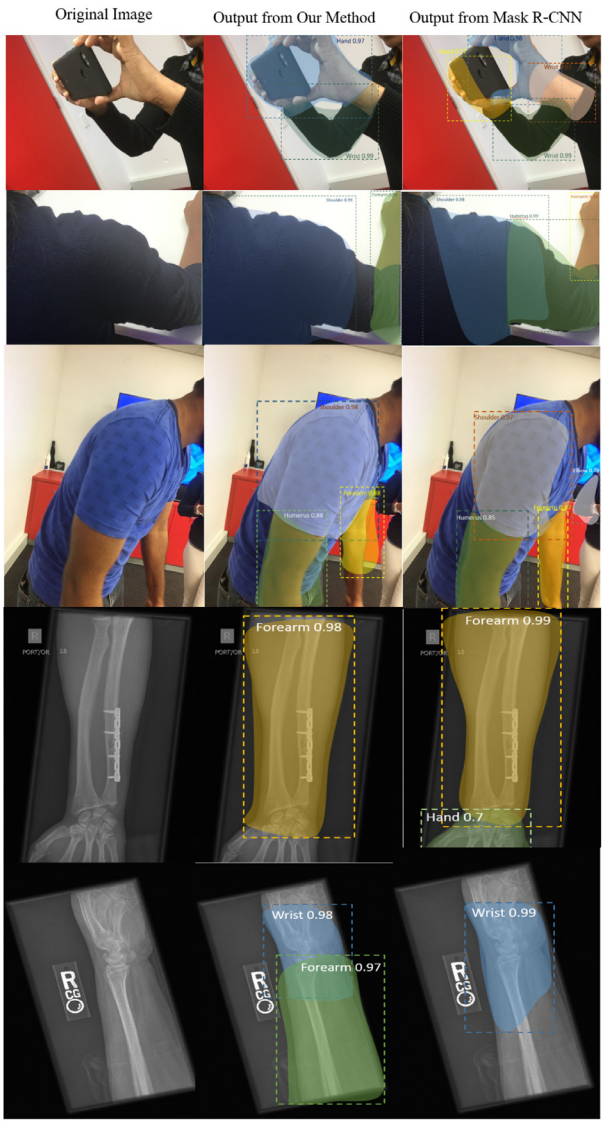


Fig. 9. Qualitative results on visible images and X-ray images on object identification, mask identification and bounding box identification. The first column indicates the input images, second column indicates the results from Mask R-CNN and the third column indicates the results from our method.

recorded in Table 4. It can be seen that assigning higher values to the λ and μ has favourably affected the results.

4.2. Single-class segmentation on multi-modal images

The evaluations that we describe in this section were carried out to validate our method on other imaging modalities, which includes heatmap images and IR images.

Datasets

Using the HandNet (Wetzler et al., 2015) dataset, we obtained IR images and heatmap images along with their groundtruth annotation of the mask. HandNet dataset contains 214 971 images in total. However as the groundtruth annotations for hand masks in X-ray images were limited to 4200 images, to eliminate the class imbalance we randomly selected 8400 images from HandNet dataset, and for 4200 of them we obtained the IR images and for the rest we obtained the heatmap images. For the visible images, the dataset we collected has 918 hand annotations. Therefore we used images from Rendered Hand Pose Dataset (RHD dataset Zimmermann and Brox, 2017) along with our own dataset to make a dataset with 4200 visible images.

Architecture Setting

By the ablative studies that we discussed in the previous section, it was identified that feature extraction network of ResNeXt-101-FPN produced the best accuracy values for majority of experiments. Therefore, in this evaluation we have used ResNeXt-101-FPN as the backbone architecture. Similar to the layer identification that we conducted for visible and X-ray images, when the features generated for some heatmap images and IR images from Mask R-CNN were analysed, it could be identified that only some kernels generates features for these two modalities, yet these kernel numbers were consistent among the images from a single modality. For the ResNeXt-101-FPN architecture, it was identified that there are 3011 feature outputs which generated discerning features for all the modalities under consideration, which are visible, X-ray, heatmap and IR images. Therefore, in the architecture that was used for this evaluation 3011 feature outputs were concatenated.

First we trained both the model (the Mask R-CNN and our model) using heatmap image dataset and IR image dataset separately and evaluated using a test set from the same imaging modality. The results that we obtained for mask identification and bounding box identification are recorded in Table 5. It can be seen that our model is capable of obtaining better results for both imaging modalities, when compared with Mask R-CNN.

Then we trained both Mask R-CNN and our method on multi-modal image dataset, which includes visible, X-ray, visible, heatmap and IR images. The obtained average precision values for this setting, are depicted in Table 6. From the results it can be observed that our method outperforms Mask R-CNN by larger margins. Compared to the single modality training Mask R-CNN demonstrates a significant decrease in the accuracy values, which demonstrates the adverse effect of incorporating the feature outputs which are generating discerning features only for some imaging modalities. Qualitative results for the IR images and heatmap images are depicted in Fig. 10 and the observations that were made in quantitative results are reflected by the qualitative results as well.

It should be noted that the traditional segmentation methods (Ali et al., 2006; Feng, 2006; Isard and Blake, 1998; Manos et al., 1994) were not evaluated in this paper, as the objective of this work is to obtain accurate semantic segmentation (Fig. 2c) in contrast to the traditional segmentation process (Fig. 2b).

4.3. Multi-class segmentation on CT images

Dataset

Using the CT image dataset (Yang et al., 2018), we obtained CT images along with their groundtruth annotation of the mask. This dataset contains free-breathing (FB) CT images from 60 patients and annotations for five organs (i.e. Esophagus, Heart, Left lung (Lung_L), Right lung (Lung_R) and Spinal cord).

Results

We used the standard train/test split of the dataset and the evaluations were conducted using the Dice score Eq. (8). For the benchmarking, we compared the best network configuration of Yang, Jinzhong, et al. (Yang et al., 2018) with ours. The obtained results are recorded in Table 7 and some qualitative results are depicted in Fig. 11.

4.4. Discussion on the generalization of the proposed approach

Generalization capacity in terms of imaging modalities: The proposed method has been validated on four imaging modalities (i.e. visible, X-ray, heatmap and IR images), and extending the method into other imaging modalities does not demand much effort. Kernel selection needs to be carried out by passing a sample of images across the

Table 4
Average precision on visible images and X-ray images for different λ and μ values in Eq. (1) for image segmentation, where the models have been trained on multi-modal dataset.

Method	Backbone	λ	μ	Trained on multi-modal tested on visible images		Trained on multi-modal tested on X-ray images		Average	
				AP_{50}	AP_{75}	AP_{50}	AP_{75}	AP_{50}	AP_{75}
Mask R-CNN	ResNeXt-101-C4	1	1	50.48	35.48	50.34	31.13	50.41	33.30
Mask R-CNN	ResNeXt-101-C4	2	2	52.36	37.73	52.80	33.38	52.58	35.56
Mask R-CNN	ResNeXt-101-C4	3	3	52.81	39.27	53.90	35.05	53.36	37.16
Mask R-CNN	ResNeXt-101-FPN	1	1	49.54	39.33	52.95	33.62	51.25	36.48
Mask R-CNN	ResNeXt-101-FPN	2	2	51.42	41.58	55.41	35.87	53.42	38.73
Mask R-CNN	ResNeXt-101-FPN	3	3	51.87	43.12	56.51	37.54	54.19	40.33
Our method	ResNeXt-101-C4	1	1	69.13	42.84	63.58	36.39	66.36	39.62
Our method	ResNeXt-101-C4	2	2	71.01	45.09	66.04	38.64	68.53	41.87
Our method	ResNeXt-101-C4	3	3	71.46	46.63	67.14	40.31	69.30	43.47
Our method	ResNeXt-101-FPN	1	1	69.80	46.31	64.16	36.94	66.98	41.63
Our method	ResNeXt-101-FPN	2	2	72.05	48.77	66.41	39.11	69.23	43.94
Our method	ResNeXt-101-FPN	**3**	**3**	**73.59**	**49.87**	**68.08**	**39.89**	**70.84**	**44.88**

Table 5
Average precision on IR images and heatmap images for bounding box identification and segmentation, where the models have been trained on individual image modality.

Task	Method	λ	μ	Trained & tested on IR images		Trained & tested on Heatmap images	
				AP_{50}	AP_{75}	AP_{50}	AP_{75}
Segmentation	Mask R-CNN	1	1	49.21	28.32	44.13	25.68
		2	2	51.38	30.42	46.21	27.41
		3	3	51.88	30.86	46.57	27.75
	Our method	1	1	58.53	31.19	61.33	43.47
		2	2	60.44	33.23	63.49	45.12
		3	3	**60.70**	**33.57**	**63.83**	**45.36**
Bounding box identification	Mask R-CNN	1	1	52.77	31.29	48.19	27.36
		2	2	54.49	33.35	50.26	28.87
		3	3	54.83	33.63	50.50	29.09
	Our method	1	1	61.18	34.52	64.56	47.08
		2	2	62.90	36.35	66.51	48.47
		3	3	**63.21**	**36.60**	**66.73**	**48.65**

Table 6
Average precision on IR images and heatmap images for bounding box identification and segmentation, where the models have been trained on multi-modal images.

Task	Method	λ	μ	Trained on multimodal & tested on IR images		Trained on multimodal & tested on Heatmap images	
				AP_{50}	AP_{75}	AP_{50}	AP_{75}
Segmentation	Mask R-CNN	1	1	39.33	21.83	32.62	18.66
		2	2	41.43	23.71	34.51	20.38
		3	3	41.8	24.08	34.83	20.7
	Our method	1	1	56.12	43.38	59.87	47.34
		2	2	58.16	45.47	62.04	49.04
		3	3	60.71	48.06	64.78	51.16
Bounding box Identification	Mask R-CNN	1	1	43.54	31.04	37.09	23.42
		2	2	45.36	33.06	39.34	25.2
		3	3	46.65	34.61	41.06	26.61
	Our method	1	1	60.11	45.9	64.01	51.51
		2	2	62.36	48.46	66.73	53.76
		3	**3**	**63.94**	**50.36**	**68.84**	**55.52**

network, which is followed by updating the model to extract features from the selected kernels, and then the training can be performed using the loss functions mentioned in Section in this paper.

Generalization capacity in terms of architectures: In the current implementation variations of ResNet architecture (Resnet-50, Resnet-101, ResNeXt-50 and ResNeXt-101) have been used. The methodology can be used with any other backbone (e.g., DenseNet Huang et al., 2017a and MobileNet Howard et al., 2017) architectures by replacing the feature extraction component of the architecture depicted in Fig. 7, with the corresponding feature extraction framework. Since the features generated by the kernels depend on the backbone architecture in the process of adopting the proposed method to a different architecture, the kernel selection needs to be performed by passing sample images across the network.

Generalization capacity in terms of classes: By the performed experiments, it is evident that the proposed method has the ability to capture the semantic segments on human upper extremity. Extending the model to other body parts does not require much effort. By the empirical studies, we observed that for a given modality the kernels which generated discerning features for all the body parts under the consideration of this paper (those body parts are elbow, forearm, hand, humerus, shoulder and wrist) remained the same. Therefore the existing models can be used directly for other body parts which are from the same modality.

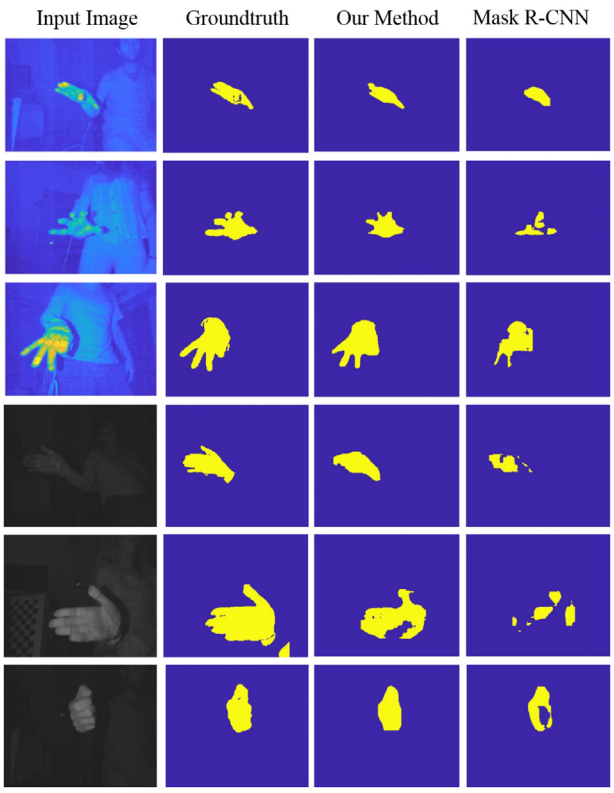


Fig. 10. Qualitative results on IR and heatmap images on segmentation. The first column indicates the input images, second column indicates the groundtruth masks, third column indicates the results obtained by our method and the fourth column indicates the results that were obtained from Mask R-CNN.

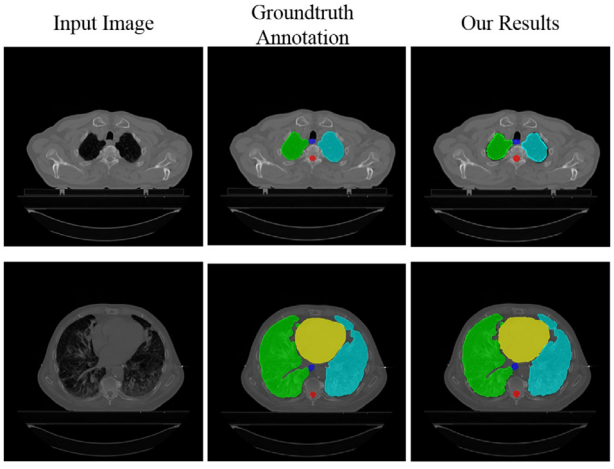


Fig. 11. Qualitative results on CT image dataset. The first column indicates the input images, second column indicates the groundtruth masks and third column indicates the results obtained by our method. Spinal cord, Esophagus, Heart, Lung_L and Lung_R are denoted by red, blue, green, yellow and cyan colours respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Conclusion

In this paper we have presented a modality invariant segmentation framework, where the most informative layers of CNNs are effectively utilized as the features for class prediction, region identification and mask identification across multi-modal images. The different textures of imaging modalities and the features that are generated by the convolutional neural networks for these different textures are taken into consideration when devising our architecture. In contrast to the previous approaches where the successive layers of the network use all the outputs from the previous layers, and the features of the last convolutional layer are subjected for the mask extraction, we filter the most informative outputs of all the layers in the network. The current implementation has outperformed the widely used segmentation framework " Mask R-CNN" for the visible, X-ray, heatmap and IR images, and this method can be extended to other image modalities such as CT scan and MRI by observing the convolutional neural network's behaviour on the modalities under consideration. Using a single model for semantic segmentation of multiple modalities, not only seamlessly incorporates all sub-tasks into one, but also reduces the number of parameters and improves the accuracy by enabling end-to-end training and inference. This is meaningful in the real world, since many modalities (e.g., X-ray) are quite hard to acquire and by using a single model, we can train the network with the other modalities (e.g., visible) where samples are abundant. Our code and the dataset we used in this work are released so that the broader research community is able to replicate our results and further advance the area of multi-modal image segmentation of medical images.

Table 7
Dice score values obtained for CT image dataset.

	Left lung	Right lung	Heart	Esophagus	Spinal cord
Yang et al. (2018)	0.97	0.97	0.93	0.72	0.88
Ours	**0.97**	**0.97**	**0.96**	**0.78**	**0.91**

CRediT authorship contribution statement

Akila Pemasiri: Conception of ideas, Methodology development, Data preparation, Implementation, Validation, Analysis of results, Writing - original draft. **Kien Nguyen:** Conception of ideas, Supervision, Methodology development, Analysis of results, Writing, Editing. **Sridha Sridharan:** Conception of ideas, Supervision, Methodology development, Analysis of results, Writing, Editing. **Clinton Fookes:** Conception of ideas, Supervision, Methodology development, Analysis of results, Writing, Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The research presented in this paper was supported by an Australian Research Council (ARC) grant DP170100632.

References

Ali, M.A., Dooley, L.S., Karmakar, G.C., 2006. Object based image segmentation using fuzzy clustering. In: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, Vol. 2. IEEE, p. II.

Bates, R., Irving, B., Markelc, B., Kaeppler, J., Brown, G., Muschel, R.J., Brady, M., Grau, V., Schnabel, J.A., 2019. Segmentation of vasculature from fluorescently labeled endothelial cells in multi-photon microscopy images. In: IEEE Transactions on Medical Imaging, Vol. 38. IEEE, pp. 1–10.

Bullock, J., Cuesta-Lázaro, C., Quera-Bofarull, A., 2019. XNet: a convolutional neural network (CNN) implementation for medical x-ray image segmentation suitable for small datasets. In: Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging, Vol. 10953. International Society for Optics and Photonics, 109531Z.

Dai, J., He, K., Sun, J., 2016. Instance-aware semantic segmentation via multi-task network cascades. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3150–3158.

Duarte, A., Carrão, L., Espanha, M., Viana, T., Freitas, D., Bártolo, P., Faria, P., Almeida, H., 2014. Segmentation algorithms for thermal images. Proc. Technol. 16, 1560–1569.

Elanattil, S., Moghadam, P., Sridharan, S., Fookes, C., Cox, M., 2018. Non-rigid reconstruction with a single moving RGB-D camera. In: 2018 24th International Conference on Pattern Recognition. ICPR, IEEE, pp. 1049–1055.

Elgammal, A., Harwood, D., Davis, L., 2000. Non-parametric model for background subtraction. In: European Conference on Computer Vision. Springer, pp. 751–767.

Feng, D., 2006. Segmentation of Bone Structures in X-ray Images (Ph.D. thesis). School of Computing National University of Singapore, Superviser Dr. Leow Wee Kheng.

Girshick, R., 2015. Fast r-cnn. arXiv preprint arXiv:1504.08083.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 580–587.

Gonfaus, J.M., Boix, X., Van de Weijer, J., Bagdanov, A.D., Serrat, J., Gonzalez, J., 2010. Harmony potentials for joint classification and segmentation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, pp. 3280–3287.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, pp. 2980–2988.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Hržić, F., Štajduhar, I., Tschauner, S., Sorantin, E., Lerga, J., 2019. Local-entropy based approach for X-ray image segmentation and fracture detection. Entropy 21 (4), 338.

Hu, J., Chen, Y., Yi, Z., 2019. Automated segmentation of macular edema in OCT using deep neural networks. Med. Image Anal. 55, 216–227.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017a. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al., 2017b. Speed/accuracy trade-offs for modern convolutional object detectors. In: IEEE CVPR.

Isard, M., Blake, A., 1998. Condensation—conditional density propagation for visual tracking. Int. J. Comput. Vis. 29 (1), 5–28.

Ji, J., Buch, S., Soto, A., Carlos Niebles, J., 2018. End-to-end joint semantic segmentation of actors and actions in video. In: Proceedings of the European Conference on Computer Vision, ECCV. pp. 702–717.

Jian, B.-L., Chen, C.-L., Chu, W.-L., Huang, M.-W., 2017. The facial expression of schizophrenic patients applied with infrared thermal facial image sequence. BMC Psychiatry 17 (1), 229.

Khowaja, S.A., Lee, S.-L., 2019. Semantic image networks for human action recognition. arXiv preprint arXiv:1901.06792.

Kundu, A., Li, Y., Dellaert, F., Li, F., Rehg, J.M., 2014. Joint semantic segmentation and 3d reconstruction from monocular video. In: European Conference on Computer Vision. Springer, pp. 703–718.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., Heng, P.A., 2018a. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. In: IEEE Transactions on Medical Imaging, Vol. 37. IEEE, pp. 2663–2674.

Li, H., Ellis, J.G., Zhang, L., Chang, S.-F., 2018b. PatternNet: Visual pattern mining with deep neural network. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. ACM, pp. 291–299.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: CVPR, Vol. 1, No. 2. p. 4.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.

Long, J.L., Zhang, N., Darrell, T., 2014. Do convnets learn correspondence? In: Advances in Neural Information Processing Systems. pp. 1601–1609.

Manos, G., Cairns, A., Rickets, I., Sinclair, D., 1994. Segmenting radiographs of the hand and wrist. Comput. Methods Programs Biomed. 43 (3–4), 227–237.

Martin, D.R., Fowlkes, C.C., Malik, J., 2003. Learning to detect natural image boundaries using brightness and texture. In: Advances in Neural Information Processing Systems. pp. 1279–1286.

Mittal, A., Paragios, N., 2004. Motion-based background subtraction using adaptive kernel density estimation. In: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, Vol. 2. IEEE, p. II.

Nakamura, T., Sato, M., Kajimoto, H., 2013. Semi-automatic scoring method for torticollis by using kinect: 328. Mov. Disorders 28.

Pemasiri, A., Ahmedt-Aristizabal, D., Nguyen, K., Sridharan, S., Dionisio, S., Fookes, C., 2019a. Semantic segmentation of hands in multimodal images: A region new-based CNN approach. In: 2019 IEEE 16th International Symposium on Biomedical Imaging. ISBI 2019, IEEE, pp. 819–823.

Pemasiri, A., Nguyen, K., Sridharan, S., Fookes, C., 0000. Unified 2D and 3D hand pose estimation from a single visible or X-ray image.

Pemasiri, A., Thanh, K.N., Sridharan, S., Fookes, C., 2019b. Semantic correspondence in the wild. In: 2019 IEEE Winter Conference on Applications of Computer Vision. WACV, IEEE, pp. 1137–1146.

Qiao, Y., Wei, Z., Zhao, Y., 2017. Thermal infrared pedestrian image segmentation using level set method. Sensors 17 (8), 1811.

Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R.L., et al., 2017. MURA dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:1712.06957.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. pp. 91–99.

Rowley, H.A., Baluja, S., Kanade, T., 1998. Neural network-based face detection. IEEE Trans. Pattern Anal. Mach. Intell. 20 (1), 23–38.

Salscheider, N.O., 2019. Simultaneous object detection and semantic segmentation. arXiv preprint arXiv:1905.02285.

Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y., 2013. Pedestrian detection with unsupervised multi-stage feature learning. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, pp. 3626–3633.

Shrivastava, A., Sukthankar, R., Malik, J., Gupta, A., 2016. Beyond skip connections: Top-down modulation for object detection. arXiv preprint arXiv:1612.06851.

Sonkusare, S., Ahmedt-Aristizabal, D., Aburn, M.J., Nguyen, V.T., Pang, T., Frydman, S., Denman, S., Fookes, C., Breakspear, M., Guo, C.C., 2019. Detecting changes in facial temperature induced by a sudden auditory stimulus based on deep learning-assisted face tracking. Sci. Rep. 9 (1), 4729.

Stolojescu-CriŞan, C., Holban, Ş., 2013. A comparison of X-ray image segmentation techniques. Adv. Electr. Comput. Eng. 13 (3).

Torabi, A., Bilodeau, G.-A., 2013. A LSS-based registration of stereo thermal–visible videos of multiple people using belief propagation. Comput. Vis. Image Underst. 117 (12), 1736–1747.

Vaillant, R., Monrocq, C., Le Cun, Y., 1994. Original approach for the localisation of objects in images. IEE Proc.-Vis. Image Signal Process. 141 (4), 245–250.

Wetzler, A., Slossberg, R., Kimmel, R., 2015. Rule of thumb: Deep derotation for improved fingertip detection. In: Xie, X., Jones, M.W., Tam, G.K.L. (Eds.), Proceedings of the British Machine Vision Conference. BMVC, BMVA Press, pp. 33.1–33.12.

Wollmann, T., Gunkel, M., Chung, I., Erfle, H., Rippe, K., Rohr, K., 2019. GRUU-Net: Integrated convolutional and gated recurrent neural network for cell segmentation. Med. Image Anal.

Wu, J., Mahfouz, M.R., 2016. Robust x-ray image segmentation by spectral clustering and active shape model. J. Med. Imaging 3 (3), 034005.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, pp. 5987–5995.

Yang, J., Veeraraghavan, H., Armato, III, S.G., Farahani, K., Kirby, J.S., Kalpathy-Kramer, J., van Elmpt, W., Dekker, A., Han, X., Feng, X., et al., 2018. Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. Med. Phys. 45 (10), 4568–4581.

Zimmermann, C., Brox, T., 2017. Learning to estimate 3d hand pose from single rgb images. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4903–4911.