VX-Mask R-CNN: Multimodal Semantic Object Segmentation for Human Upper Extremity

Anonymous ACCV 2018 submission

Paper ID 1193

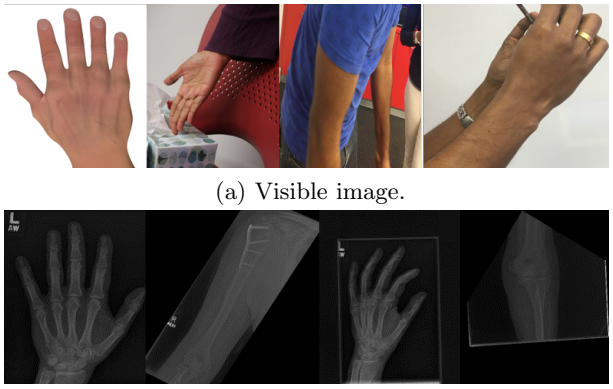
Abstract. In this paper we propose a modality invariant method to obtain high quality semantic object segmentation of human body parts, where the considered modalities are X-ray images and visible images. Due to the intrinsic difference between images from two modalities, state-of-the-art approaches such models as Mask R-CNN do not perform satisfactorily. We analyse performance of intermediate layers within the Mask R-CNN on both X-ray and visible modalities. The insights from this analysis has led us to propose a new and efficient network architecture which yields highly accurate semantic segmentation results across both X-ray and visible domains. In the proposed network, dubbed as VX-Mask R-CNN, multi-task losses are combined to train the network. 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. Our method can be applied to other modalities and can be effectively utilized for medical image analysis tasks such as image registration and 3D reconstruction across modalities.

1 Introduction

Leveraging the performance of medical imaging systems with computer vision techniques has progressed over many years with the aim to improve both diagnosis and therapy outcomes [1–3]. X-ray images are widely used in the medical domain to observe the structures of human anatomy including bones, tissues and muscles [4, 5]. Some computer vision tasks such as 3D reconstruction, may require more than one X-ray image of the patient. However, exposing patients to many X-ray images has adverse health affects and can result in a higher cancer risk [6]. In such cases the visible images (RGB images) of the same body part can be used in combination with X-ray images to mitigate this problem. The advantages of using visible images in this task are that they can be captured conveniently, with very low cost while incurring minimal harm to patients. However, this requires innovative cross-modal processing in low level image processing tasks such as image registration, object identification, segmentation and correspondence.

Compared with single-modality image processing, cross-modal image processing exhibits some inherent challenges, arising from potential visual dissimilarity where the regions in two images can be differently textured or one image may be

textured while the other is homogeneous [7]. In addition, the level of information presented in each image modality is different. For instance, considering images of a hand from visible and X-ray modalities, visible images contain details of skin, nails, clothes and accessories whereas X-ray images contain details of bones and muscles (Figure 1).



(a) Visible image.

(b) X-ray image.

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

In this paper we present a method to establish object boundary and pixel wise semantic object segmentation of human body parts regardless of the modality, whether it is an X-ray or a visible image. Object boundary identification and semantic segmentation underpin many applications including human pose estimation [8], semantic understanding [9], 3D object reconstruction [10], human tracking [11] and action recognition [12]. Moreover, recent advancements in neural networks have motivated their applicability for semantic segmentation tasks, resulting in robust and impressive baseline frameworks including Fast R-CNN [13], Faster R-CNN [14], and Mask R-CNN [15]. Inspired by this progression, we develop a new method coined as "VX Mask R-CNN" where we exploit the layers which generate the most informative features the capabilities of Mask-RCNN to achieve effective multi-modal and robust object segmentation. For the X-ray images, we use the **mu**sculoskeletal **ra**diograph (MURA) dataset [16]. For visible images, we have collected our own dataset, which contains human body parts (i.e. elbow, forearm, hand, humerus, shoulder and wrist) images under challenging environments such as different lighting conditions, different resolutions and different backgrounds which we will make publicly available to enable researchers to replicate our results in this paper. In addition to the robust object segmentation method introduced in this paper, we provide the groundtruth region marking by manual annotation for the MURA X-ray image dataset for the upper extremity of the human body which will also be available for researchers

to follow-up on our work. This work has targeted on human upper extremity due to the unavailability of large X-ray image datasets for other body parts; the method we propose however can easily be applied to the other body parts as well.

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 each component of the system. In Section 4, we present our experimental results and 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 decades. Early stages of image segmentation include approaches that uses the low level image features such as brightness and texture gradient [17]. However, when considering the object segmentation in contrast to the image segmentation, it 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 can operate as a prior for the detection task [18–20]. Therefore efforts on improving the detection process have paramount effect on the improvement of segmentation process.

It is notable that neural network based approaches have granted significant enhancement for the object detection and segmentation task [21–23]. However, in the early approaches based on CNNs they dealt with specific object categories in constrained applications such as faces [21, 23] and pedestrians [22]. In the most recent literature, in rich feature hierarchies for accurate object detection and semantic segmentation [24], has introduced the concept of Regions with CNN features coined as "R-CNN" with the target of accomplishing higher accuracies for object detection. They have first extracted the region proposals for the whole image, where then the 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 extensive number of forward passes "Fast R-CNN" [13] 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 froward 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" [14] 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 objectives of R-CNN, Fast R-CNN and Faster R-CNN are object detection and classification. Mask R-CNN [15] has extended the Faster R-CNN by adding another branch for object mask detection resulting in state-of-art accuracies for object segmentation. However, the approaches suggested in the liter-

ature for detection and segmentation have focused on visible images. In contrast to the previous approaches we develop an architecture to detect the objects and carryout the segmentation while being invariant to the image modality, whether they are X-ray or visible. To the best of our knowledge, this is the first paper to consider model invariant semantic segmentation for visible and X-ray images.

3 VX-Mask R-CNN

In this section we provide an overview of Mask R-CNN and then elaborate our method on deriving VX-Mask R-CNN.

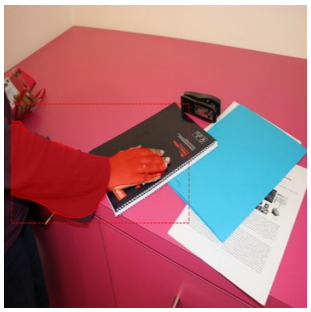
3.1 Mask R-CNN

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.

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 (Figure 2). 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 (Figure 3).

An intuitive solution to overcome the problems that are mentioned above is to use a dataset, which contains the images from both the modalities. Besides the increased quality of the output compared to the models that were trained on a single modality (Figure 2 and Figure 3), the newly generated masks also suffered from the problems of not recognizing the complete object or mistakenly identifying the parts in the image that do not belong to the object (Figure 4).

With the aim of gaining an insight for Mask R-CNN, when the output of the layers were inspected, a simple yet crucial observation was made. That is, only some of the layers in Mask R-CNN generates discerning features for both the image modalities. A subset of feature maps, extracted from some random layers of the Mask R-CNN is depicted in Figure 5. The first column of Figure 5 shows the input image, where the first row of it is the input visible image and the second row shows the input X-ray image. The rest of the columns depict the feature output from some random layers. It can be seen that some layers have generated features for both the input images, where as some layers only have generated sensible features for either of input images. The layers that were generating sensible features for both image modalities were consistent for all the body parts under our consideration.

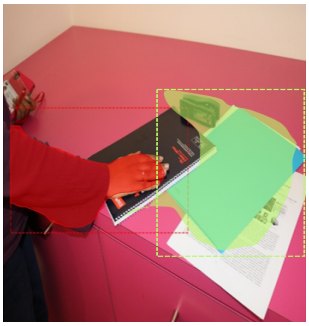


(a) Tested on a visible image.

(b) Tested on an X-ray image.

Fig. 2: 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. 3: Results that were obtained from the model trained on X-ray images illustrating its poor performance in mask identification in visible images.

Fig. 4: Results on images which were tested on the model trained on multimodal image dataset.

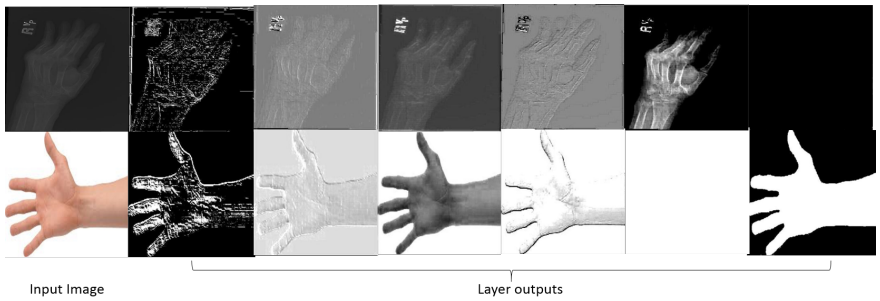


Fig. 5: Visualization of layer outputs of Mask R-CNN for a visible image and for an X-ray image, the first column depicts the input images. The 2nd, 3rd, 4th and 5th columns show the layer outputs where informative features have been generated for both X-ray and visible images, 6*th* column shows an 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.

3.2 VX-Mask R-CNN

Motivated by the observations described in Section 3.1, in this paper we introduce the architecture of VX-Mask R-CNN, which generates robust results for visible images as well as X-ray images.

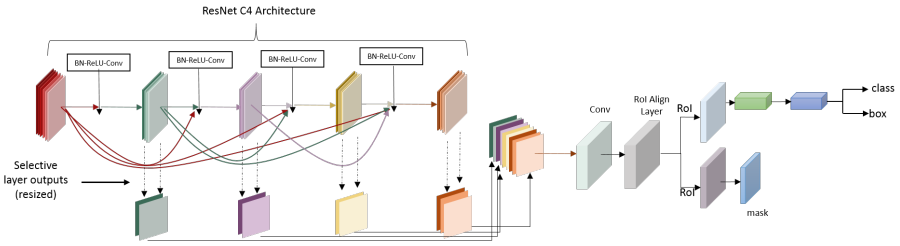


Fig. 6: VX-Mask architecture, where the selected layer outputs are resized and fused to generate the new layer, which is then subjected for RoI extraction, mask extraction and class label extraction.

VX-Mask introduces a new feature concatenation layer where the feature output of the layers that generate features for both the modalities from Mask R-CNN are concatenated. Depending on the backbone architecture used in the Mask R-CNN the dimension of feature concatenation layer vary. The architecture that is depicted in Figure 6 corresponds to the ResNet101-C4 [25] architecture. Following the previous work on Faster R-CNN we employed ResNet architecture and ResNeXt architecture [26] with depth of 50 or 101 layers. Following the

commonly adopted notation [27, 25, 28, 29] 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) [30] 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. 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 Equation 1, where λ and μ are the hyper-parameters that are used for maintaining 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 the consideration. 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 (Equation 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 the equation 3, the term $smooth_{L1}$ is defined as,

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

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 by the average binary cross entropy,

$$Binary\,Cross\,Entropy = -\left(ylog\left(p\right)\right) + \left(1-y\right)log\left(1-p\right), \tag{5}$$

of the k^{th} mask. In this process K is the number of classes under consideration. In Equation 5, y is the label for the mask that can either be 1 or 0, depending on whether that point correspondence to a foreground pixel or a background pixel and p stands for the output of the neuron.

4 Experiments

4.1 Dataset

For the X-ray image dataset we used the MURA dataset[16] which contains X-ray images of the human body parts. 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.

4.2 Experiment Setting

For the VX-Mask training we use the same learning rates and the weight decay values that have been used in the preceding approaches including Fast R-CNN [13, 14]. We use the pretrained backbone architectures where the training has been performed on COCO [31] dataset and the convolution layer was initiated with random values. The λ and μ were set to 1, and the learning rate, weight decay and the momentum was set to 0.02, 0.00001 and 0.9 accordingly.

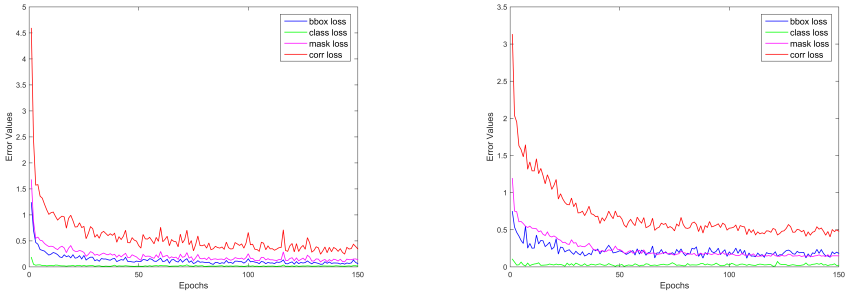
The training accuracies and the validation for the combined dataset are depicted in Figure 7.

4.3 Evaluation Matrices

The evaluation matrix we used the average precision (AP), which is been 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 Equation 6, b_p is the prediction, which is the bounding box prediction or the mask prediction, and b_t is the corresponding groundtruth value. In this evaluation we used the thresholds of 0.5 and 0.75.



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

Fig. 7: Training and validation losses.

4.4 Results

To compare our method with original Mask R-CNN we conducted a set of experiments where previously unseen images were fed into our architecture as well as to the Mask R-CNN. When the original Mask R-CNN and the VX-Mask were tested on the same datasets for mask identification, the obtained average precision values are indicated in Table 1. By analysing the table it can be understood that the VX-Mask R-CNN outperforms 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 1 it can be understood that ResNeXt-101-FPN has yielded the highest average precision for both the modalities and for both Mask R-CNN and VX-Mask R-CNN, 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 VX-Mask R-CNN 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 VX-Mask R-CNN, which have been trained on the same datasets is recorded in Table 2. It can be identified that the same observations that were made for the mask prediction are presented in bounding box identification as well. Some qualitative results that we obtained for both visible and X-ray images are depicted in Figure 8. From the qualitative results it can be identified that the when the results gained by VX-Mask R-CNN are compared with the original Mask R-CNN results, that the VX-Mask R-CNN is capable of generating more precise results, while identifying more components and regions of upper extremity of human body. However, it should be noted that even when using VX-Mask R-CNN, there are circumstances that where all the components that are presented in the image are not properly identified. For an instance, in the 3rd row of Figure 8 the elbow of the person has not been identified by both the

approaches, instead the combination of forearm and the elbow has been detected as the forearm.

Table 1: Average precision on visible images and X-ray images for image segmentation

	Backbone	On Visible Images		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**
VX-Mask R-CNN	Resnet-50-C4	57.3	31.3	53.4	27.7	55.4	29.5
VX-Mask R-CNN	Resnet-101-C4	59.7	33.1	54.4	28.8	57.1	31.0
VX-Mask R-CNN	ResNet-50-FPN	58.2	32.4	53.1	28.2	55.7	30.3
VX-Mask R-CNN	ResNet-101-FPN	62.6	34.6	54.9	29.3	58.8	32.0
VX-Mask R-CNN	ResNeXt-50-C4	62.9	34.9	55.8	30	59.4	32.5
VX-Mask R-CNN	ResNeXt-101-C4	64.1	35.3	56.9	31.7	60.5	33.5
VX-Mask R-CNN	ResNeXt-50-FPN	65.1	35.7	58.2	32.4	61.7	34.1
VX-Mask R-CNN	ResNeXt-101-FPN	**66.8**	**37.1**	**58.9**	**33.5**	**62.9**	**35.3**

5 Conclusion

In this paper we have presented modality invariant Mask R-CNN called "VX Mask R-CNN", where the most informative layers of CNNs are effectively utilized as the features for class prediction, region identification and mask identification across multimodal images. The different textures of the two image modalities (X-ray and visible) and the features that are generated by the convolutional neural networks for these different textures are taken into 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 is subjected for the mask extraction, we filter the most informative outputs of all the layers in the network. The current implementation has outperformed the original Mask R-CNN for the visible and X-ray images, and this method can be extended to other image modalities by observing the convolutional neural network's behaviour on the modalities under consideration.

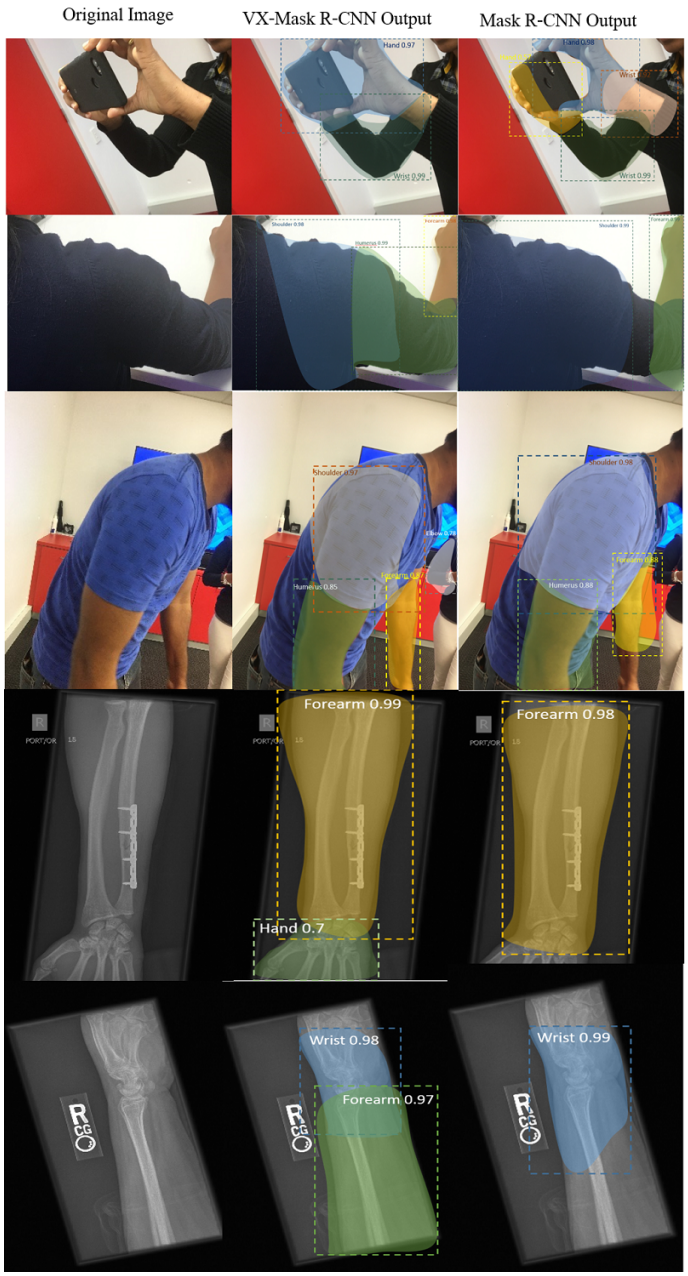


Fig. 8: Qualitative results on visible images and X-ray images on object identification, mask identification and bounding box identification.

Table 2: Average precision on visible images and X-ray images for bounding box identification

	Backbone	On Visible Images		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**
VX-Mask R-CNN	Resnet-50-C4	63.29	39.75	57.28	30.09	60.28	34.92
VX-Mask R-CNN	Resnet-101-C4	65.56	38.99	61.70	33.01	63.63	36.00
VX-Mask R-CNN	ResNet-50-FPN	63.34	33.13	56.44	30.81	59.89	31.97
VX-Mask R-CNN	ResNet-101-FPN	65.81	38.70	61.75	33.86	63.78	36.28
VX-Mask R-CNN	ResNeXt-50-C4	68.40	44.18	62.49	34.26	65.44	39.22
VX-Mask R-CNN	ResNeXt-101-C4	69.13	42.84	63.58	36.39	66.35	39.62
VX-Mask R-CNN	ResNeXt-50-FPN	69.27	41.36	63.95	36.91	66.61	39.13
VX-Mask R-CNN	ResNeXt-101-FPN	**69.80**	**46.31**	**64.16**	**36.94**	**66.98**	**41.63**

References

1. Chen, C.h.: Computer vision in medical imaging. Volume 2. World scientific (2014)
2. Cootes, T.F., Taylor, C.J.: Statistical models of appearance for medical image analysis and computer vision. In: Medical Imaging 2001: Image Processing. Volume 4322., International Society for Optics and Photonics (2001) 236–249
3. Florack, L.: Image structure. computational imaging and vision (1997)
4. Farber, L., Tardos, G., Michaels, J.N.: Use of x-ray tomography to study the porosity and morphology of granules. Powder Technology **132** (2003) 57–63
5. Hu, S., Hoffman, E.A., Reinhardt, J.M.: Automatic lung segmentation for accurate quantitation of volumetric x-ray ct images. IEEE transactions on medical imaging **20** (2001) 490–498
6. Rao, V.M., Levin, D.C.: The overuse of diagnostic imaging and the choosing wisely initiative. Annals of internal medicine **157** (2012) 574–576
7. Torabi, A., Bilodeau, G.A.: A lss-based registration of stereo thermal–visible videos of multiple people using belief propagation. Computer Vision and Image Understanding **117** (2013) 1736–1747
8. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. arXiv preprint arXiv:1804.06208 (2018)
9. Neuhold, G., Ollmann, T., Bulo, S.R., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy. (2017) 22–29
10. Yang, B., Rosa, S., Markham, A., Trigoni, N., Wen, H.: 3d object dense reconstruction from a single depth view. arXiv preprint arXiv:1802.00411 (2018)
11. Li, J., Zhao, J., Wei, Y., Lang, C., Li, Y., Sim, T., Yan, S., Feng, J.: Multi-human parsing in the wild. arXiv preprint arXiv:1705.07206 (2017)

12. Liu, Y., Wang, Q.: A new simple human abnormal action detection based on static images. In: Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on. Volume 1., IEEE (2011) 578–581
13. Girshick, R.: Fast r-cnn. arXiv preprint arXiv:1504.08083 (2015)
14. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. (2015) 91–99
15. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE (2017) 2980–2988
16. Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R.L., et al.: Mura dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:1712.06957 (2017)
17. Martin, D.R., Fowlkes, C.C., Malik, J.: Learning to detect natural image boundaries using brightness and texture. In: Advances in Neural Information Processing Systems. (2003) 1279–1286
18. Elgammal, A., Harwood, D., Davis, L.: Non-parametric model for background subtraction. In: European conference on computer vision, Springer (2000) 751–767
19. Mittal, A., Paragios, N.: 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. Volume 2., Ieee (2004) II–II
20. Gonfaus, J.M., Boix, X., Van de Weijer, J., Bagdanov, A.D., Serrat, J., Gonzalez, J.: Harmony potentials for joint classification and segmentation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 3280–3287
21. Rowley, H.A., Baluja, S., Kanade, T.: Neural network-based face detection. IEEE Transactions on pattern analysis and machine intelligence **20** (1998) 23–38
22. Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y.: Pedestrian detection with unsupervised multi-stage feature learning. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 3626–3633
23. Vaillant, R., Monrocq, C., Le Cun, Y.: Original approach for the localisation of objects in images. IEE Proceedings-Vision, Image and Signal Processing **141** (1994) 245–250
24. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 580–587
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
26. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE (2017) 5987–5995
27. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 3150–3158
28. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al.: Speed/accuracy trade-offs for modern convolutional object detectors. In: IEEE CVPR. (2017)
29. Shrivastava, A., Sukthankar, R., Malik, J., Gupta, A.: Beyond skip connections: Top-down modulation for object detection. arXiv preprint arXiv:1612.06851 (2016)

30. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. Volume 1. (2017) 4
31. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, Springer (2014) 740–755