Enhancing Multi-Task Learning for Image Segmentation using soft-attention blocks and self-supervised auxiliary tasks

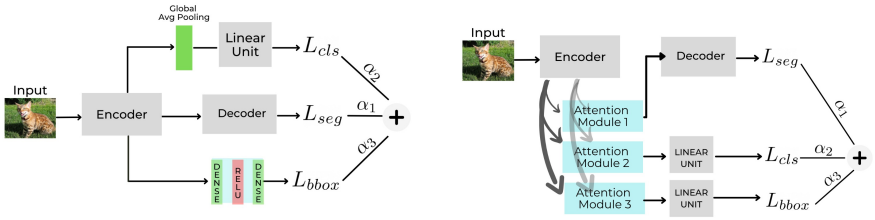
Arnab Pushilal, Andrea Achilleos, Tabish Ahmed, Sulakshana Chakraborty, Hamze Muse, Michael Pawlik,, Jesus Solano, and Gishean Thayaparan

University College London
London, UK

1 Introduction

Multi-Task Learning (MTL) became widely used since it was proved that generalisation in Machine Learning (ML) is improved through parallel training of tasks on a shared representation [3,24]. Using this approach, a lot of new MTL architectures appeared, where shared representations between related tasks, led to better performance on a target task [32]. This was applied in the related tasks of detection and classification [10, 23], as well as to detection and segmentation [8], and it was shown that related tasks act as "regularizers" for each other [33]. This method of performance improvement, has been widely utilised in various domains like computer vision [13], linguistic evaluation [34] and multi-lingual translations [6]. The sharing of the tasks has been explored through two different approaches: the originally noted hard-parameter sharing [3] and soft-parameter sharing [7]. Recently, MTL framework has evolved more as an "umbrella term" used widely [5,9,14] and often finds parallels in continual learning [31] and transfer learning [17, 19].

In this study, the effectiveness of different MTL architectures was explored, by first only evaluating a baseline inspired on `SegNet` architecture [1]; then modifying the baseline to implement MTL framework with two additional tasks (image classfication and bounding-boxes prediction); and finally by fine-tuning the MTL network using pre-trained weights from VGG-16 [27] trained on ImageNet [4]. Furthermore, inspired by the success of attention approaches in computer vision [16], a more recent and novel MTL network that includes task specific attention modules was investigated. This network is called Multi-Task Attention Network (`MTAN`) [16]. Finally, auxiliary tasks of image denoising, edge detection and image colourisation were implemented. For all of them, self-supervision strategies were used to create the ground-truth labels. The outlook of this research was to investigate the effect of parameter sharing on the effectiveness of learning, the effect of self-supervision on the model's performance and the effectiveness of fine-tuning in the context of simultaneous MTL.



(a) `SegNet` (b) `MTAN`

Figure 1: Illustration of network architectures.

2 Method

Learning of multiple tasks is typically explored either simultaneously [3] or sequentially [21, 22, 26, 29, 30]. For this study, simultaneous learning of tasks was chosen [3]. It not only avoids an unwanted bias from defining a "training sequence" (i.e. order in which tasks are learned to benefit each other), but it also enables the tasks to mutually benefit each other because all details are available to all the tasks at the same time. For example, if classification was learned before segmentation, then classification will not benefit from learning segmentation, and it will not help the the segmentation task. Though we do opt for simultaneous training, we experimented and also inculcated sequential transfer in the vein of fine-tuning by using pre-trained weights trained on ImageNet [4].

2.1 Image Segmentation: `SegNet`

In this empirical study, a `SegNet` model [1] was implemented as the baseline for image segmentation. This model is an encoder-decoder network for image segmentation, topologically inherited from a VGG-16 [27] architecture. Initially, `SegNet` was used only for the single target task of segmentation. Then, an MTL architecture was implemented by adding two layers at the end of the `SegNet` encoder, in order to perform binary classification and bounding box regression. Figure 2 depicts an illustration of `SegNet` The encoder has been chosen to have shared parameters for all of the 3 tasks and in effect, choosing a hard-parameter sharing scheme. This choice was inspired from the segmentation and classification scheme for cancer detection developed by [15].

2.2 Multi-Task Learning with Attention

Multi-Task Learning with Attention (`MTAN`) [16] based on `SegNet` was implemented in order to enhance the relative importance given to each task in training. The parameter sharing scheme involved a shared network and task specific attention modules. The original implementation was modified by applying the attention module only to the encoder. The soft attention masks were applied to every one of the five blocks of the encoder, that learns the relative importance of the shared parameters for each task. As there were shared layers along with task specific attention modules, both shared and task specific features were learnt, in a self-supervised manner. Compared to other alternatives such as Cross-Stitch

networks [18] and Progressive Networks [25], `MTAN` has shown to be more memory efficient both in terms of the number of parameters required for each task, but also in terms of data. Additionally, the number of parameters was found to be close enough to the baseline model to be comparable as in Figure 4.

2.3 Losses & Metrics

Cross-entropy was used for the loss of both segmentation & classification and mean-squared loss was used for the bounding-box. The metrics used for segmentation were F1, and Jaccard Index (IOU) and accuracy was used for classification. The IOU weighted by the classes was also monitored. Existing methods for loss tuning like SoftAdapt [11], DWA [16], GM [12] losses were investigated but did not improving the target task. Therefore, manually fixed weights were initialised for each of the tasks as follows:

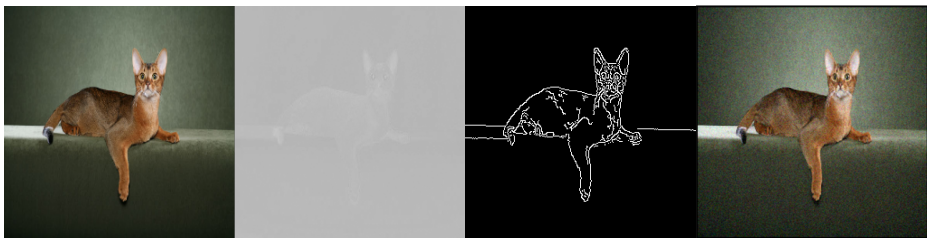
$$L_{total} = \alpha_1 L_{seg} + \alpha_2 L_{cls} + \alpha_3 L_{bbox}, \tag{1}$$

where $\alpha_1, \alpha_2, \alpha_3$ were chosen as hyper-parameters instead of being determined based on the rate of change of the losses. This was done with respect the scale of the losses and with keeping in mind the improvement of the target task. As seen in Figure3a, the balancing of losses was crucial in achieving a good performance for MTL.

2.4 Self-Supervised Auxiliary Tasks

For our Open-Ended Question (EOQ), Carls Doersch and Andrew Zisserman [5] proposed that an interesting extension of their paper was to use similar self-supervised tasks on deeper networks, for example VGG-16. To that end, as `SegNet` is based on the VGG-16 architecture, three auxiliary tasks were implemented on the above-mentioned `MTAN` architecture. The three auxiliary tasks tackled were the colourisation of grayscale images, the detection of edges in colour images (Canny filter) and the removal of normally distributed noise from an image.

2.4.1 Denoising The restoration from a noisy image to a clean image has been of great interest since it was introduced [36] from the perspective of neural networks. For this experiment, denoising was achieved by using as training data noisy images, constructed by adding small Gaussian noise with std 2 to clean, original, images, which were used as targets. A noisy image is shown in Figure 2d. The training loss used was the Mean Squared Error Loss (MSE).



(a) Original image. (b) Color (c) Canny (d) Noisy

Figure 2: Visualisation of inputs/outputs for self-supervised auxiliary tasks.

2.4.2 Edge Filter Detection This auxiliary task was inspired by the findings of [28], who recently showed that adding a shape stream that processes the information of boundaries, boosts the performance of image segmentation. In this experiment, a canny edge detector was implemented [2], as the proxy to find the boundaries of objects in the input image. Canny filter available in OpenCV was used to find the boundaries in images. A new task branch was added after the encoder in `MTAN`, that predicts the boundaries in the images.

2.4.3 Colourisation In order to implement colourisation, the baseline `SegNet` network was modified firstly by changing the number of input channels of the first layer from 3 to 1 color channel. RGB images were converted to LAB [20]. A single L channel was taken as an input and used to predict the A and B channels which were outputted in the final layer of the network. Pre-trained weights were added by skipping the first layer, where there was a mismatch between VGG our network for this task [35].

3 Evaluation

3.1 Experimental Setup

3.1.1 Data Our data consisted of pre-processed images from the Oxford-IIIT dataset, consisting of two class labels namely - dogs and cats. Our validation and test sets consisted of only dogs images.

3.1.2 Experiments The tests performed as part of our study include: (1) `SegNet` **1 task**, (2) `SegNet` **pre-trained 3 task**, (3) `SegNet` **pre-trained 1 task**, (4) `SegNet` **pre-trained with attention** and both models with additional auxiliary tasks. For the ablative study, each of the two original tasks were removed and they were individually trained with the target task. Cross-validation was deemed too expensive, since fine-tuning was performed for the entire encoder-decoder architecture for all models. The validation IOU was monitored on *Tensorboard* and the hold-out test set was used to test the models.

3.1.3 Hardware: The models for `SegNet` were trained on a single NVDIA GTX 1650 with 4GB capacity. The `MTAN` model did not fit on the above-mentioned GPU, thus it was trained on a Telsa T4 with 12GB capacity on Colab Pro. All

	`SegNet`			`MTAN`					
	1 Task	**1 Task**	**3 Task**	**3 Task**	**No BBox**	**No Class**	**Canny**	**Noisy**	**Color**
Pre-trained Weights	False	True	True	True	True	True	True	True	True
Average IOU	0.775	0.869	0.866	0.874	0.864	**0.876**	0.865	0.824	0.847
Weighted IOU	0.782	0.869	0.870	0.879	0.868	**0.880**	0.871	0.829	0.852
Segmentation F1	0.853	0.921	0.919	0.924	0.918	**0.926**	0.918	0.889	0.906
Classification Accuracy	-	-	0.423	0.796	0.786	-	0.728	0.327	0.746
BBOX L2 Loss	-	-	463.4	353.9	-	**347.3**	376.1	1345	444.1

Table 1: Model evaluation on the test set.

models were trained for 30 epochs with a *batchsize* of 5 due to hardware constraints.

3.1.4 Losses: Naively we could initialise equal weights to all 3 tasks as in Figure 3(a). To achieve comparable accuracy with 1-task `SegNet`, in the multi-task networks $\alpha_1, \alpha_2, \alpha_3$, were tuned while inspecting the validation IOU. The obtained weights for each task were $\alpha_1 = 0.7, \alpha_2 = 0.1$ and $\alpha_3 = 0.2$. It was found that if α_1 was increased further, the training for the other two tasks was influenced negatively, as shown in Figure 3a. Additionally, the bounding box loss was adjusted by a constant in order to be at the same scale as the other two losses. This was done for both `MTAN` and `SegNet`.

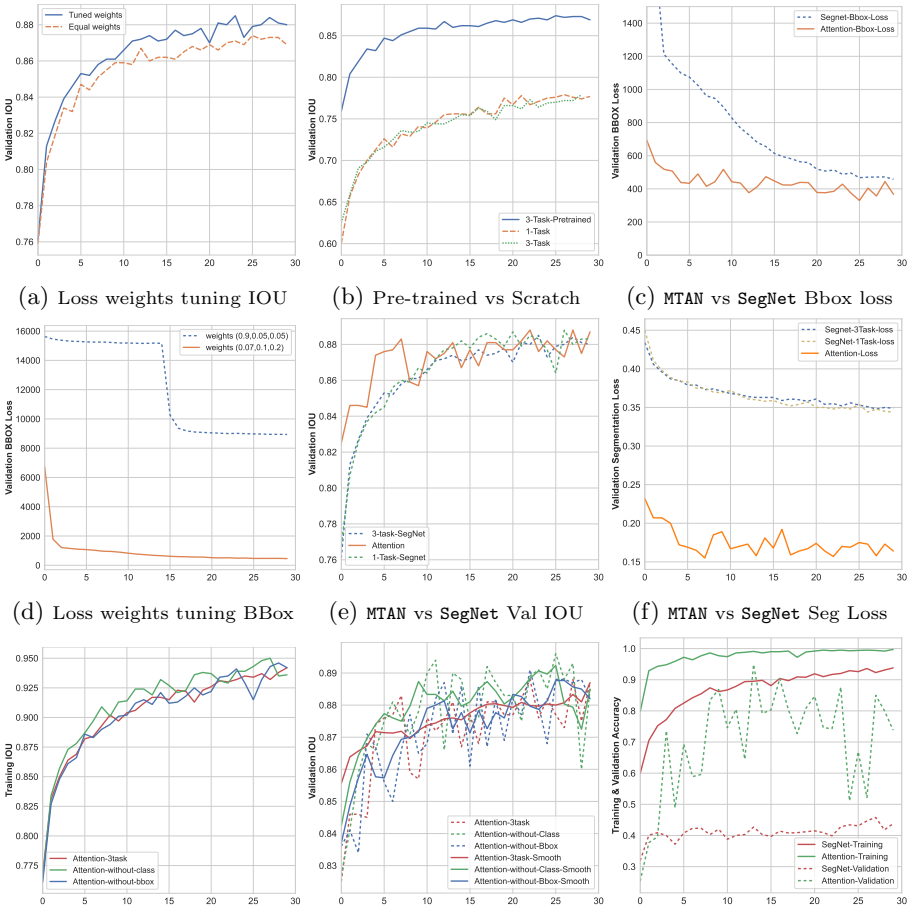
3.2 Results

As seen in Figure 3f, the `MTAN` model had the lowest segmentation loss, which means its better confidence about predictions is higher. However, the validation IOU was comparable for `SegNet` and `MTAN`, as depicted in Figure 3e. Moreover, after using `MTAN` the performance for the bounding box and classification task was improved (See Figure 3a). Additionally, faster convergence was observed for the auxiliary tasks using the attention blocks. This was attributed to the fact that both of these tasks had an individual network instead of a shared encoder as in `SegNet`. Figure 3b shows the impact of fine-tuning using pre-trained VGG weights was quick convergence and a significant performance improvement for all 3 tasks. Additionally, it was found that `SegNet` was over-fitting on the training set with respect to classification (Figure 3i), possibly because of the difference of the distribution of the labels in the training and validation sets. It was also observed that the different versions of `MTAN` were less prone to over-fitting. Finally, regarding the ablation study, Figures 3h and 3g show that the training of the model with the bounding-box, resulted in better performance than the model with classification, which hypothesises that the classification task induces a negative transfer.

All models were tested on a hold-out test set and the results are depicted in Figure 3. The `MTAN` architecture with the bounding box (but no classification) task, negligibly outperformed all other models, as far as the target task is concerned. With respect to the self-supervised auxiliary tasks, the simultaneous learning with the Canny filter (i.e. image gradients as outputs) outperformed the other task for all chosen metrics. The model that trained the fastest was the `MTAN` colourisation ($\approx$ 10 minutes per epoch). Interestingly, it was found that adding denoising to `MTAN` resulted in negative transfer. This finding was reflected in the relatively low testing IOU and the poor performance in other auxiliary tasks - especially the classification accuracy. Another riveting finding, is that the pre-trained weights of VGG-16 handled the LAB inputs equally well, when it in fact was trained on RGB images.

4 Discussion & Conclusion

Two MTL architectures were implemented, namely `SegNet` and `MTAN`. It was found that using pre-trained weights for `MTAN` did not significantly impact the performance of the target segmentation task. Additionally, `MTAN` was found to



(a) Loss weights tuning IOU (b) Pre-trained vs Scratch (c) `MTAN` vs `SegNet` Bbox loss

(d) Loss weights tuning BBox (e) `MTAN` vs `SegNet` Val IOU (f) `MTAN` vs `SegNet` Seg Loss

(g) Ablation Study Train IOU (h) Ablation Study Val IOU (i) Classification Acc.

Figure 3: Evaluation of different models on training and validation sets. Horizontal axis is the number of epochs for all of the sub-figures.

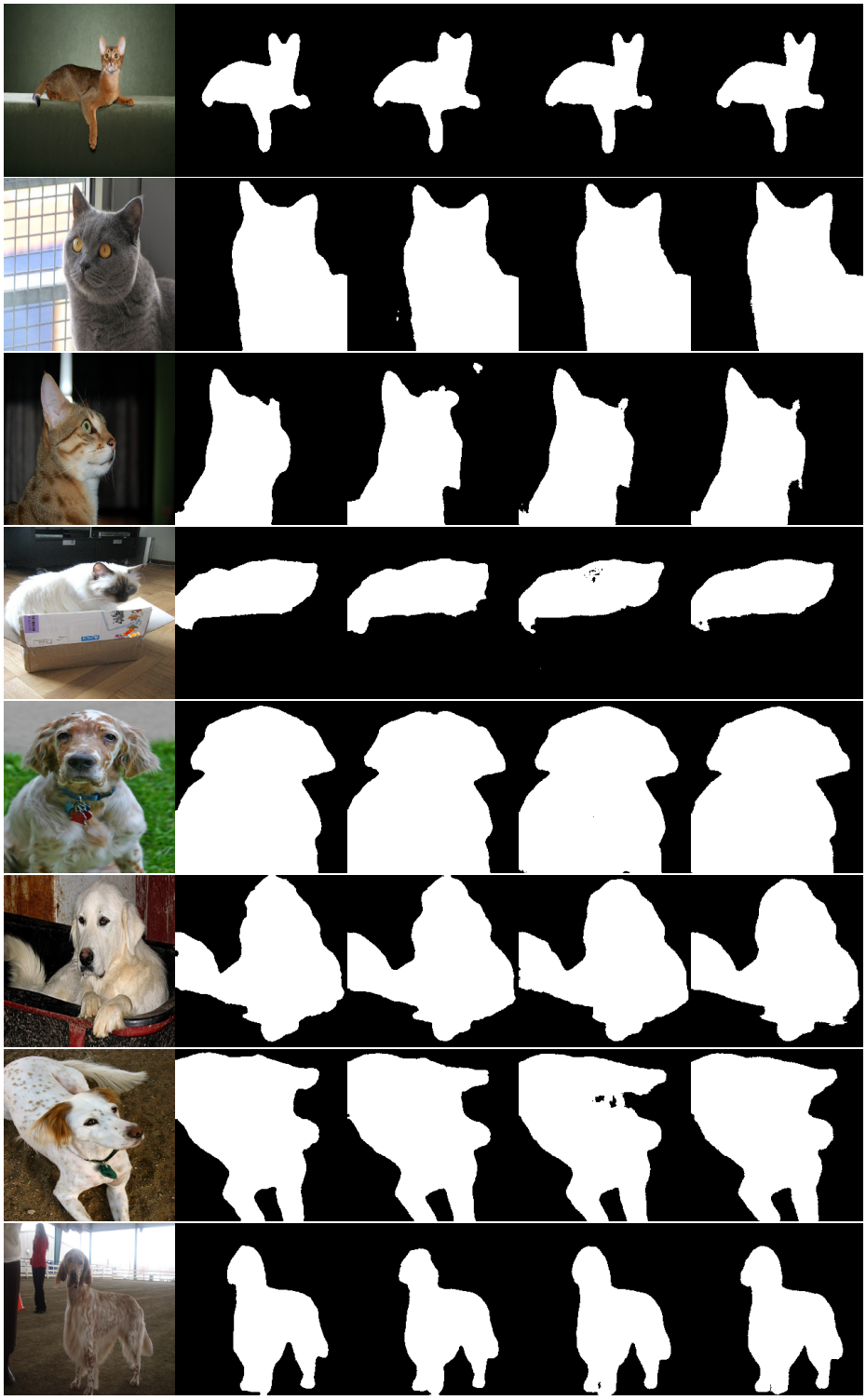
improve the performance of auxiliary tasks (i.e classification and bounding box regression) as well, hence exhibiting flexibility in learning other auxiliary tasks, when compared to the hard-parameter sharing model. Differences in the target task performance metric between models were small and this phenomenon was attributed to the fact that the model size was large and when using pre-trained weights, it is hard to outperform the `SegNet` baseline. Figure 4 depicts a visual representation of predicted segmentation masks for the best performing models.

From the ablation studies, it would be interesting to look into the attention masks for bounding box versus classification, to establish the reason that `MTAN` with bounding box was marginally better. It was found that the auxiliary self-supervised tasks did not improve the segmentation task. Whether a combination of them would lead to a performance improvement, remains an open question. An extension to our study that would potentially result in better test performance is to use the validation set for training after hyper-parameter tuning.

References

1. Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.
2. John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.
3. Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.
4. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
5. Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2051–2060, 2017.
6. Yajie Dong, Jean-Michel Caruge, Zhaoqun Zhou, Charles Hamilton, Zoran Popovic, John Ho, Matthew Stevenson, Guo Liu, Vladimir Bulovic, Moungi Bawendi, et al. 20.2: Ultra-bright, highly efficient, low roll-off inverted quantum-dot light emitting devices (qleds). In *SID Symposium Digest of Technical Papers*, volume 46, pages 270–273. Wiley Online Library, 2015.
7. Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers)*, pages 845–850, 2015.
8. Nikita Dvornik, Konstantin Shmelkov, Julien Mairal, and Cordelia Schmid. Blitznet: A real-time deep network for scene understanding. In *Proceedings of the IEEE international conference on computer vision*, pages 4154–4162, 2017.
9. Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117, 2004.
10. Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
11. A Ali Heydari, Craig A Thompson, and Asif Mehmood. Softadapt: Techniques for adaptive loss weighting of neural networks with multi-part loss functions. *arXiv preprint arXiv:1912.12355*, 2019.
12. Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *2016 IEEE international conference on Robotics and Automation (ICRA)*, pages 4762–4769. IEEE, 2016.
13. Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.
14. Abhishek Kumar and Hal Daume III. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*, 2012.
15. Thi-Lam-Thuy Le, Nicolas Thome, Sylvain Bernard, Vincent Bismuth, and Fanny Patoureaux. Multitask classification and segmentation for cancer diagnosis in mammography. *arXiv preprint arXiv:1909.05397*, 2019.
16. Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1871–1880, 2019.

17. Mingsheng Long, Jianmin Wang, Guiguang Ding, Sinno Jialin Pan, and S Yu Philip. Adaptation regularization: A general framework for transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(5):1076–1089, 2013.
18. Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3994–4003, 2016.
19. Weike Pan, Evan Xiang, Nathan Liu, and Qiang Yang. Transfer learning in collaborative filtering for sparsity reduction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, 2010.
20. Abhishek Thakur Sumit Pandey and Priyanka Dusane Nidhi Sharma. Image colorization using deep learning. *International Journal for Scientific Research and Engineering Trends*, 2019.
21. Lorien Y Pratt. Non-literal transfer among neural network learners. *Colorado School of Mines: MCS-92-04*, 1992.
22. Lorien Y Pratt, Jack Mostow, Candace A Kamm, Ace A Kamm, et al. Direct transfer of learned information among neural networks. In *Aaai*, volume 91, pages 584–589, 1991.
23. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.
24. Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.
25. Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.
26. Noel E Sharkey and Amanda JC Sharkey. Adaptive generalisation. *Artificial Intelligence Review*, 7(5):313–328, 1993.
27. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
28. Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5229–5238, 2019.
29. Sebastian Thrun. Lifelong learning: A case study. Technical report, 1995.
30. Sebastian Thrun and Tom M Mitchell. Learning one more thing. Technical report, 1994.
31. Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012.
32. Partoo Vafaeikia, Khashayar Namdar, and Farzad Khalvati. A brief review of deep multi-task learning and auxiliary task learning. *arXiv preprint arXiv:2007.01126*, 2020.
33. Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European Conference on Computer Vision*, pages 268–285. Springer, 2020.
34. Jichuan Wang and Xiaoqian Wang. *Structural equation modeling: Applications using Mplus*. John Wiley & Sons, 2019.
35. Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016.
36. Y-T Zhou, Rama Chellappa, Aseem Vaid, and B Keith Jenkins. Image restoration using a neural network. *IEEE transactions on acoustics, speech, and signal processing*, 36(7):1141–1151, 1988.



(a) Original (b) SegNet-1T (c) SegNet-3T (d) MTAN-3T (e) MTAN-NoCl

Figure 4: Illustration of predicted mask for different models