Semi-supervised Multi-task Learning with Chest X-Ray Images

Abdullah-Al-Zubaer Imran(✉) and Demetri Terzopoulos

Computer Science Department, University of California, Los Angeles, CA, USA
aimran@cs.ucla.edu

Abstract. Discriminative models that require full supervision are inefficacious in the medical imaging domain when large labeled datasets are unavailable. By contrast, generative modeling—i.e., learning data generation and classification—facilitates semi-supervised training with limited labeled data. Moreover, generative modeling can be advantageous in accomplishing multiple objectives for better generalization. We propose a novel multi-task learning model for jointly learning a classifier and a segmentor, from chest X-ray images, through semi-supervised learning. In addition, we propose a new loss function that combines absolute KL divergence with Tversky loss (KLTV) to yield faster convergence and better segmentation performance. Based on our experimental results using a novel segmentation model, an Adversarial Pyramid Progressive Attention U-Net (APPAU-Net), we hypothesize that KLTV can be more effective for generalizing multi-tasking models while being competitive in segmentation-only tasks.

Keywords: Semi-supervised · Multi-tasking · Generative modeling · Classification · Segmentation · KL-Tversky loss · Chest X-ray

1 Introduction

The effective supervised training of deep neural networks normally requires large pools of labeled data. In medical imaging, however, datasets tend to be limited in size due to privacy issues, and labeled data is scarce since manual annotation requires tedious, time-consuming effort by medical experts, making it not only expensive, but also susceptible to subjectivity, human error, and variance across different experts. Although some large labeled datasets are available, they can be seriously imbalanced by over-representation of common problems and under-representation of rare problems.

The success of discriminative models such as regular CNNs for classification or segmentation, depends on large labeled training datasets to make predictions about future unobserved examples. Generative modeling has recently received much attention with the advent of deep generative models, such as GANs. Since they can learn real data distributions, they are becoming increasingly popular given the abundance of unlabeled data.

H.-I. Suk et al. (Eds.): MLMI 2019, LNCS 11861, pp. 151–159, 2019.
https://doi.org/10.1007/978-3-030-32692-0_18

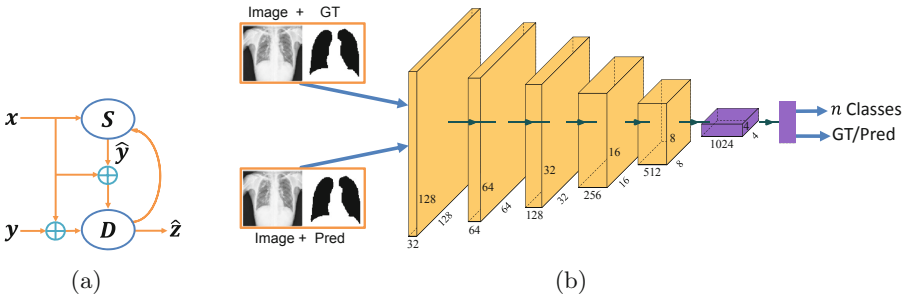
Via generative modeling, we can perform multi-task learning in a semi-supervised manner, without large labeled datasets. In practice, we train a deep learning model to perform a single task (classification, segmentation, detection, etc.) by fine-tuning parameters until its performance no longer improves. The same model can subsequently be enabled to perform better in other tasks. In fact, the domain-specific features from the related tasks are leveraged to improve the generalization of the model through multi-task learning [1]. Hence, one objective regularizes another to accomplish multiple tasks within a common model.

We introduce a novel generative modeling approach to joint segmentation and classification from limited labeled data, in a semi-supervised manner, and apply it to chest X-ray imagery. Our technical contributions are twofold: (1) a novel multi-task learning model for semi-supervised classification and segmentation from small labeled medical image datasets and (2) a new loss function combining absolute KL divergence and Tversky loss (KLTV) for semantic segmentation.

1.1 Related Work

Several single-task classification and segmentation models are available in the chest X-ray literature. Based on the popular segnet architecture, Mittal *et al.* [2] proposed a fully convolutional encoder-decoder with skip connections for lung segmentation in chest X-ray images. Adversarial training of an FCN followed by a CRF has been applied to non-overfitting mammogram segmentation [3]. Adversarial learning has been utilized for segmentation (semantic-aware generative adversarial nets [4], structure correcting adversarial nets [5], etc.) as well as in disease classification from chest X-ray images (semi-supervised domain adaptation [6], attention-guided CNN [7], semi-supervised multi-adversarial encoder [8]).

Unlike the above models, our model jointly performs both classification and segmentation. Several prior efforts address multi-task learning with CNNs and generative modeling. Rezaei *et al.* [9] proposed a GAN model combining a set of auto-encoders with an LSTM unit and an FCN as discriminator for semantic segmentation and disease prediction. Girard *et al.* [10] used a U-Net-like architecture coupled with graph propagation to jointly segment and classify retinal vessels. Mehat *et al.* [11] proposed a Y-Net, with parallel discriminative and convolutional modularity, for the joint segmentation and classification of breast biopsy images. Another multi-tasking model was proposed by Yang *et al.* [12] for skin lesion segmentation and melanoma-seborrheic keratosis classification, using GoogleNet extended to three branches for segmentation and two classification predictions. Khosravan *et al.* [13] used a semi-supervised multi-task model for the joint learning of false positive reduction and nodule segmentation from 3D CT. Ours is the first model to pursue a multi-task learning approach to the analysis of chest X-ray images.



(a) (b)

Fig. 1. (a) Basic structure of the proposed APPAU-Net model. The segmentor S predicts segmentation $\hat{y}$ from a given image x. The discriminator D predicts the class label $\hat{z}$ from image-real label pair (x, y); $z = 0, \ldots, n$ are real disease classes and $z = n + 1$ denotes the predicted class; (b) Detailed architecture of the Discriminator D (as a CNN) network of the APPAU-Net model.

2 Model Description

2.1 Adversarial Pyramid Progressive Attention U-Net

Our proposed APPAU-Net model consists of two major building blocks, a segmentor S and a discriminator D (Fig. 1). S primarily performs segmentation prediction $\hat{y}$ from a given image x. S consists of a pyramid encoder and a progressive attention-gated decoder modifying a U-Net. The S network, which is illustrated in Fig. 2, receives the image input x at different scales in different stages of the encoder [14]. This pyramidal input allows the model to access class details at different scales. Moreover, while lowering resolution, the model can keep track of the ROIs, avoiding the possibility of losing them after the subsequent convolutions. The pyramid input to the encoder network enables the model to learn more locally-aware features crucial to semantic segmentation.

Following [15], with deep-supervision, APPAU-Net generates side-outputs at different resolutions from the decoder. The side-outputs are progressively added to the next side-outputs before reaching the final segmentation at the original image resolution. Combining pyramid inputs and progressive side-outputs helps the model perform better in segmenting small ROIs. The side-output segmentation maps $\hat{y}_i$ are compared to the ground truth mask to calculate the side-losses of varying weights (higher resolutions are usually assigned higher weights). Therefore, the final segmentation loss is calculated as

$$L_{seg_{(x,y)}} = \sum_{i=1}^{4} w_i L_{(y_i, \hat{y}_i)}. \tag{1}$$

However, generating segmentation maps (side-outputs) at different stages of the decoder might lead to loss of spatial detail. In cases with substantial shape variability of the ROIs, this eventually incurs larger false positives. To

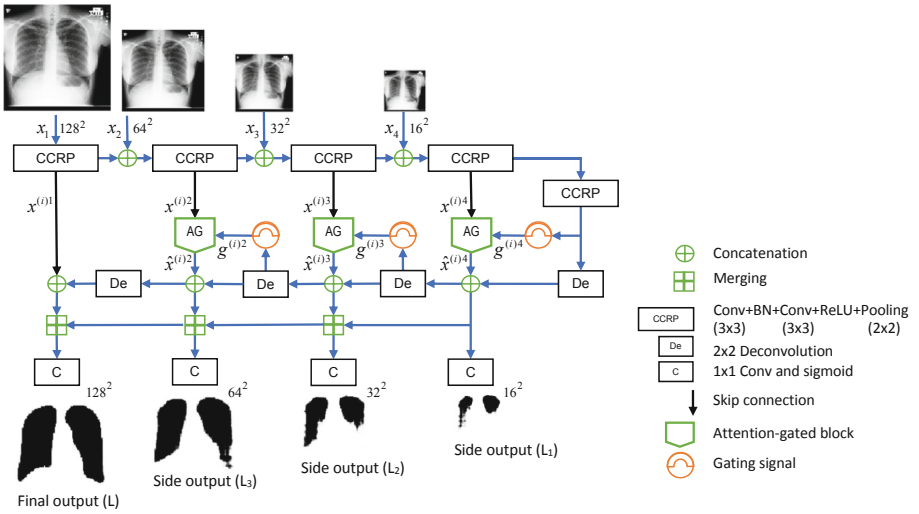


Fig. 2. Architecture of the segmentor or PPAU-Net in our APPAU-Net model. The encoder takes inputs at different scales and progressively adds the side-outputs from the attention-gated decoder. The discriminator takes image-label or image-predicted label pairs and classifies the images.

tackle this problem, we adapt soft-attention gates that help draw relevant spatial features from the low-level feature maps of the pyramid encoder [16]. Feature maps are then propagated to the high-level features to generate side-outputs at different stages of the decoder. Attention-gated (AG) modules produce attention coefficients $\alpha \in [0, 1]$ at each pixel i that scale input feature maps $x^{(i)l}$ at layer l to semantically relevant features $\hat{x}^{(i)l}$. A gating signal from coarser resolution, serves to determine the focus regions through the computation of intermediate maps, as follows:

$$G^l_{\text{attn}} = \psi^T(\sigma(w_x^T x^{(i)l} + w_g^T g^{(i)l} + b_g)) + b_\psi. \tag{2}$$

The linear attention coefficients are computed by element-wise summation and a 1×1 linear transformation. The parameters are w_x, w_g, b_g, and b_ψ. The intermediate maps are then transformed using ReLU σ_1 and sigmoid σ_2 activations. Finally, after element-wise multiplication of the feature map $x^{(i)l}$ (via skip) and nonlinear transformation, $\hat{x}^{(i)l}$ is generated at each decoder stage.

The attention coefficients α_i retain the relevant features by scaling the low level query signal $x^{(i)l}$ through an element-wise product. These pruned features are then concatenated with upsampled output maps at different stages of the decoder. A 1×1 convolution and sigmoid activation is applied on each output map in the decoder to generate the side-outputs at different resolutions. With deep supervision and gating from the pyramid encoder, the model becomes semantically more discriminative.

2.2 Loss Functions

The two building blocks of our APPAU-Net model have different objectives.

Segmentor Loss: As in the semi-supervised learning-scheme, the segmentor's objective is just based on the labeled samples. We employ Tversky loss, a generalization of Dice loss that weighs false negatives higher than false positives in order to balance precision and recall. The segmentor's objective includes a segmentation loss and an adversarial loss, where the segmentor wants the discriminator D to maximize the likelihood for the predicted segmentation generated by the segmentor. We combine an absolute KL divergence with a Tversky loss, proposing the new loss function

$$L_S = L_{S_{seg_{(y,\hat{y})}}} + cL_{S_{adv_{(x,\hat{y})}}}, \tag{3}$$

where $L_{S_{seg_{(y,\hat{y})}}} = aL_{S_{KL}} + bL_{S_{TV}}$, with $L_{S_{KL}} = \sum_i^{m^2} |(y_{pl}(i) - \hat{y}_{pl}^{(i)}) \log(y_{pl}^{(i)}/\hat{y}_{pl}^{(i)})|$, and

$$L_{S_{TV}} = 1 - \frac{\sum_i^{m^2} y_{pl}^{(i)} \hat{y}_{pl}^{(i)} + \epsilon}{\sum_i^{m^2} y_{pl}^{(i)} \hat{y}_{pl}^{(i)} + \alpha \sum_i^{m^2} y_{p\bar{l}}^{(i)} \hat{y}_{pl}^{(i)} + \beta \sum_i^{m^2} y_{pl}^{(i)} \hat{y}_{p\bar{l}^{(i)}} + \epsilon}, \tag{4}$$

where $\hat{y}_{pl}(i)$ is the prediction probability that pixel i is assigned label l (one of the ROI labels) and $\hat{y}_{p\bar{l}}(i)$ is the probability that the pixel i is assigned the non-ROI (background) label. Similarly, $y_{pl}(i)$ and $y_{p\bar{l}}(i)$ denote the pixel-wise mapping labels in the ground-truth masks. Hyper parameters a, b, α, and β can be tuned to weigh the KL-divergence against the Tversky loss (first pair) and weigh FPs against FNs. Small constant ϵ avoids division by zero. The second term in the segmentor's objective is an adversarial loss, where the segmentor wants the discriminator to maximize likelihood for the paired data x and predicted segmentation $\hat{y}$. Therefore, the segmentor's adversarial loss is

$$L_{S_{adv_{(x,\hat{y})}}} = -\mathbb{E}_{x,\hat{y}\sim S} \log[1 - p(z = n+1|(x,\hat{y})]. \tag{5}$$

Since the main objective of the segmentor is to generate the segmentation map, $L_{S_{adv}}$ is usually weighed using a small number c.

Discriminator Loss: The discriminator is trained on multiple objectives—adversary on the segmentor's output and classification of the images into one of the real classes. Since the model is trained on both labeled and unlabeled training data, the loss function of the discriminator D includes both supervised and unsupervised losses. When the model receives image-label pairs (x, y), it is just the standard supervised learning loss

$$L_{D_{sup}} = -\mathbb{E}_{x,y,z\sim p_{data}} \log[p(z = i|x, y; i < n+1)]. \tag{6}$$

Table 1. Segmentation-only performance comparison of different models in four different data setups.

Dataset	Model	DS	JS	SSIM	F1	HD	SN	SP	PR	RC
MCX	U-Net-TV	0.991	0.983	0.950	0.966	2.968	0.965	0.989	0.968	0.965
	U-Net-KLTV	0.990	0.980	0.947	0.962	3.009	0.966	0.985	0.958	0.966
	Attention U-Net-TV	0.984	0.968	0.922	0.937	3.768	0.915	0.987	0.960	0.915
	Attention U-Net-KLTV	0.990	0.980	0.941	0.960	3.063	0.957	0.987	0.962	0.957
	PPAU-Net-TV	0.988	0.978	0.966	0.958	3.143	0.967	0.982	0.949	0.967
	PPAU-Net-KLTV	**0.992**	0.983	0.949	0.989	**2.690**	0.989	0.958	0.989	0.976
SCX	U-Net-TV	0.964	0.931	0.860	0.955	4.181	0.975	0.799	0.936	0.975
	U-Net-KLTV	0.960	0.923	0.850	0.950	4.023	0.963	0.803	0.936	0.963
	Attention U-Net-TV	0.958	0.919	0.842	0.948	4.562	0.983	0.725	0.915	0.983
	Attention U-Net-KLTV	**0.965**	0.933	0.862	0.955	**3.684**	0.946	0.894	0.964	0.946
	PPAU-Net-TV	0.954	0.913	0.838	0.944	4.523	0.983	0.700	0.908	0.983
	PPAU-Net-KLTV	0.964	0.930	0.858	0.954	3.855	0.961	0.836	0.946	0.961
JCX	U-Net-TV	0.989	0.979	0.937	0.985	2.804	0.990	0.956	0.981	0.990
	U-Net-KLTV	**0.990**	0.980	0.939	0.986	**2.553**	0.980	0.988	0.995	0.977
	Attention U-Net-TV	0.988	0.977	0.929	0.983	2.882	0.993	0.940	0.974	0.993
	Attention U-Net-KLTV	0.989	0.977	0.932	0.984	2.781	0.981	0.970	0.986	0.981
	PPAU-Net-TV	**0.990**	0.981	0.941	0.987	2.768	0.992	0.958	0.981	0.992
	PPAU-Net-KLTV	**0.990**	0.979	0.937	0.985	2.751	0.987	0.959	0.982	0.987
CCX	U-Net-TV	**0.978**	0.958	0.907	0.968	**3.322**	0.974	0.928	0.962	0.974
	U-Net-KLTV	0.969	0.939	0.874	0.953	3.502	0.946	0.926	0.960	0.946
	AttnU-Net-TV	0.970	0.941	0.878	0.956	3.643	0.972	0.883	0.940	0.972
	AttnU-Net-KLTV	0.971	0.943	0.877	0.956	3.481	0.944	0.941	0.968	0.944
	PPAU-Net-TV	0.969	0.940	0.875	0.955	3.807	0.978	0.870	0.934	0.978
	PPAU-Net-KLTV	0.967	0.936	0.868	0.951	3.472	0.939	0.932	0.963	0.939

When it receives unlabeled data (x, y) or $(x, \hat{y})$ from two different sources, the unsupervised loss combines the original adversarial losses for image-real label and image-prediction pairs:

$$L_{D_{label}} = -\mathbb{E}_{x,y\sim p_{data}} \log[1 - p(z = n + 1|x, y)] \tag{7}$$

and

$$L_{D_{pred}} = -\mathbb{E}_{(x,\hat{y})\sim S} \log[p(z = n + 1|x, \hat{y})]. \tag{8}$$

3 Experiments and Results

Dataset and Implementation Details: For the supervised segmentation, we used our PPAU-Net model and KLTV as the loss function. We compared against all the preliminary segmentation models and TV loss. Then we performed semi-supervised multi-tasking for semi-supervised disease classification and lung segmentation from chest X-ray images. We used three chest X-ray datasets: the Montgomery County chest X-ray set (MCX) comprising 138 images, the Shenzhen chest X-ray set (SCX) comprising 527 images [17], and the JSRT dataset

(JCX) comprising 247 images [18]. In addition, we created another dataset (CCX) comprising 912 images, by combining prior datasets. Each dataset was split into train and test sets in a 75:25 ratio and 10% of the train set was used for model selection. Except for CCX, all the datasets were used for binary classification (normal/abnormal), while CCX was used for 3-class classification (normal, nodule, tuberculosis). The X-ray images were normalized and resized to 128×128 pixels. For multi-tasking, we used the Adam optimizer with momentum 0.9 and learning rates 1.0^{-5} (S) and 1.0^{-4} (D). Each model was trained using a batch size of 16. All the convolutional layers were followed by batch-normalization, except for the convolutions that generate side-outputs. We performed dropout at a rate of 0.4 in the discriminator. Each model was evaluated after training for 300 epochs. For the classification, along with the overall accuracy, we reported the class-wise F1 scores. For the segmentation, we used the following performance metrics: Dice similarity (DS), Average Hausdorff distance (HD), Jaccard index (JI), Sensitivity (SN), Specificity (SP), F1 score, Structural Similarity Measure (SSIM), Precision (PR), and Recall (RE) scores.

(a) X-Ray (b) Ground Truth (c) APPAU-Net-KLTV (d) APPAU-Net-TV

Fig. 3. Visual comparison of the lung segmentation by the APPAU-Net model with TV loss (d) and KLTV loss (c). The predicted lung mask with TV and KLTV losses are overlaid with the ground truth mask.

Segmentation-Only: At first, we evaluated the performance of our PPAU-Net model for the segmentation-only task, and compared with the baseline models incrementally. Table 1 reports the performance measures of different models with varying choices of loss (TV and KLTV), showing that our model is competitive.

Semi-supervised Multi-task Learning: In the semi-supervised setting, we applied our new APPAU-Net model. Along with TV loss, we used cross-entropy with TV (XETV) loss and the proposed KLTV loss. 10% labeled and 90% unlabeled training data were used for every dataset. Table 2 shows that for all four datasets the APPAU-Net model with the new KLTV loss consistently outperformed the APPAU-Net model with TV and XETV losses in both overlap and distance measures, and suggests that the model with KLTV loss generalizes better in multi-task learning. While both TV and XETV losses tend to lose some accuracy because of the additional classification task, KLTV still achieves good

Table 2. Performance evaluation of the APPAU-Net model for semi-supervised multi-tasking in different data settings.

Dataset	Model	Classification			Segmentation								
		Acc	PR	RE	DS	JI	SSIM	F1	HD	SN	SP	PR	RE
MCX	APPAU-Net-TV	**0.571**	0.690	0.290	0.956	0.916	0.815	0.814	4.514	0.800	0.988	0.953	0.800
	APPAU-Net-XETV	0.514	0.620	0.280	0.929	0.868	0.788	0.778	4.554	0.903	0.856	0.684	0.903
	APPAU-Net-KLTV	0.543	0.680	0.200	**0.974**	0.950	0.880	0.898	**3.914**	0.857	0.944	0.944	0.857
JCX	APPAU-Net-TV	**0.758**	0.000	0.860	0.972	0.945	0.864	0.963	3.755	0.996	0.831	0.929	0.996
	APPAU-Net-XETV	**0.758**	0.000	0.860	0.975	0.952	0.878	0.966	3.489	0.995	0.857	0.939	0.994
	APPAU-Net-KLTV	**0.758**	0.000	0.860	**0.976**	0.953	0.885	0.966	**3.351**	0.975	0.904	0.958	0.975
SCX	APPAU-Net-TV	0.477	0.580	0.300	0.883	0.790	0.713	0.877	6.601	0.999	0.162	0.782	0.992
	APPAU-Net-XETV	**0.553**	0.670	0.290	0.889	0.800	0.720	0.882	6.372	0.997	0.205	0.791	0.997
	APPAU-Net-KLTV	0.508	0.530	0.490	**0.921**	0.853	0.746	0.910	**4.368**	0.992	0.434	0.841	0.992
CCX	APPAU-Net-TV	**0.776**	0.800	0.780	0.874	0.777	0.682	0.845	5.375	0.936	0.576	0.770	0.959
	APPAU-Net-XETV	0.732	0.81	0.70	0.923	0.862	0.768	0.890	4.692	0.974	0.632	0.823	0.954
	APPAU-Net-KLTV	0.750	0.770	0.750	**0.926**	0.863	0.780	0.903	**4.669**	0.979	0.645	0.838	0.953

accuracy, comparable to fully-supervised segmentation models in Table 1 and LF-segnet [2]. Figure 3 shows the segmented lungs by different models, confirming the superior performance of our APPAU-Net with KLTV loss compared to the TV loss.

4 Conclusions

Generative modeling provides unique advantages for learning from small labeled datasets. With adversarial training, we can perform multi-task learning to concurrently accomplish multiple objectives. We proposed and demonstrated in different settings the performance of a novel semi-supervised multi-task learning model for joint classification and segmentation from a limited number of labeled chest X-ray images. Our experimental results confirm that our APPAU-Net model even against the single-task learning of fully supervised models.

References

1. Caruana, R.: Multitask learning: a knowledge-based source of inductive bias. In: International Conference on Machine Learning (1993)
2. Mittal, A., Hooda, R., Sofat, S.: LF-SegNet: a fully convolutional encoder-decoder network for segmenting lung fields from chest radiographs. WPC **101**, 511–529 (2018)
3. Zhu, W., Xiang, X., Tran, T.D., Hager, G.D., Xie, X.: Adversarial deep structured nets for mass segmentation from mammograms. In: ISBI (2018)
4. Chen, C., Dou, Q., et al.: Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest X-ray segment. arXiv:1806.00600 (2018)
5. Dai, W., Dong, N., Wang, Z., Liang, X., Zhang, H., Xing, E.P.: SCAN: structure correcting adversarial network for organ segmentation in chest X-rays. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 263–273. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_30

6. Madani, A., et al.: Semi-supervised learning with GANs for chest X-ray classification with ability of data domain adaptation. In: Proceedings of ISBI (2018)
7. Guan, Q., Huang, Y., et al.: Diagnose like a radiologist: attention guided convolutional net for thorax disease classification. arXiv:1801.09927 (2018)
8. Imran, A.A.Z., Terzopoulos, D.: Multi-adversarial variational autoencoder networks. arXiv preprint arXiv:1906.06430 (2019)
9. Rezaei, M., Yang, H., et al.: Multi-task generative adversarial network for handling imbalanced clinical data. CoRR (2018)
10. Girard, F., Kavalec, C., Cheriet, F.: Joint segmentation and classification of retinal arteries/veins from fundus images. AI Med. **94**, 96–109 (2019)
11. Mehta, S., Mercan, E., Bartlett, J., Weaver, D., Elmore, J.G., Shapiro, L.: Y-Net: joint segmentation and classification for diagnosis of breast biopsy images. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 893–901. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_99
12. Yang, X., Zeng, Z., Yeo, S.Y., et al.: A novel multi-task deep learning model for skin lesion segmentation and classification. arXiv:1703.01025 (2017)
13. Khosravan, N., Bagci, U.: Semi-supervised multi-task learning for lung cancer diagnosis. In: EMBC (2018)
14. Fu, H., Cheng, J., Xu, Y., et al.: Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE TMI **37**, 1597–1605 (2018)
15. Imran, A.-A.-Z., Hatamizadeh, A., Ananth, S.P., Ding, X., Terzopoulos, D., Tajbakhsh, N.: Automatic segmentation of pulmonary lobes using a progressive dense V-network. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS 2018. LNCS, vol. 11045, pp. 282–290. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_32
16. Oktay, O., Schlemper, J., Folgoc, L.L., et al.: Attention U-net: learning where to look for the pancreas. arXiv:1804.03999 (2018)
17. Jaeger, S., Candemir, S., et al.: Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant. Imaging Med. Surg. **4**, 475 (2014)
18. Shiraishi, J., Katsuragawa, S., et al.: Development of a digital image database for chest radiographs with and without a lung nodule. J. Roentgenol. **174**, 71–74 (2000)