2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-7358-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISBI53787.2023.10230365

MULTI-TASK CURRICULUM LEARNING FOR PARTIALLY LABELED DATA

Won-Dong Jang[1⋆] *Stanislav Lukyanenko*[4⋆] *Donglai Wei*[5] *Jiancheng Yang*[6]
Brian Leahy[1,2] *Helen Yang*[3] *Dalit Ben-Yosef*[7,8] *Daniel Needleman*[1,2] *Hanspeter Pfister*[1]

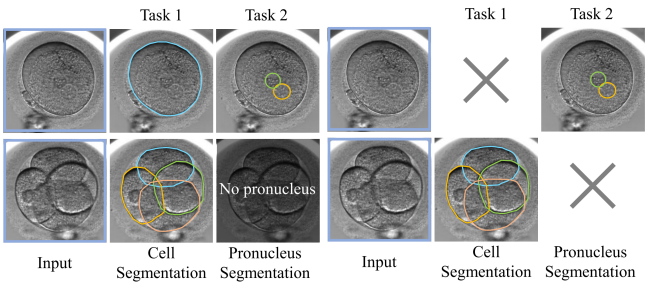
[1] School of Engineering and Applied Sciences, [2] Department of Molecular and Cellular Biology,
[3] Harvard Graduate Program in Biophysics, Harvard University, MA, USA
[4] Technical University of Munich, Germany
[5] Boston College, MA, USA
[6] EPFL, Switzerland
[7] Fertility & IVF Institute, Tel Aviv Sourasky Medical Center, Israel
[8] Dept. of Cell and Developmental Biology; Sagol School of Neuroscience - Tel Aviv University, Israel

ABSTRACT

Incomplete labels are common in multi-task learning for biomedical applications due to several practical difficulties, *e.g.*, expensive annotation efforts by experts, limit of data collection, different sources of data. A naive approach to enable joint learning for partially labeled data is adding self-supervised learning for tasks without ground truths by augmenting an input image and forcing the multi-task model to return the same outputs for both the input and augmented images. However, the partially labeled setting can result in imbalanced learning of tasks since not all tasks are trainable with ground truth supervisions for each data sample. In this work, we propose a multi-task curriculum learning method tailored for partially labeled data. For balanced learning of tasks, our multi-task curriculum prioritizes less performing tasks during training by setting different supervised learning frequencies for each task. We demonstrate that our method outperforms standard approaches on one biomedical and two natural image datasets. Furthermore, our learning method with partially labeled data performs better than the standard multi-task learning methods with fully labeled data for the same number of annotations.

1. INTRODUCTION

In multi-task learning, a single model learns to perform multiple tasks at the same time. Current multi-task learning algorithms mainly use fully-labeled data [1, 2], where all tasks' ground truths are available for each data. However, in practice (*e.g.*, biomedical applications), it is usual to have incomplete labels [3, 4]. Each input data may only have labels for particular tasks. For example, an embryo microscopy data might have a ground truth for cell segmentation, but not for pronucleus segmentation as exemplified in the lower row of Figure 1 (b). The labels are missing mainly because the annotation by clinicians or experts is costly. Besides, each task requires a different degree of effort. When humans label ground truths, there is a labor budget for annotation. For a limited labeling budget, there are two choices in the annotation. One is to label each image with full annotations, and the other is to label more images with partial annotations. If partially labeled data enables better multi-task learning than fully labeled data, we can reduce number of required annotations to achieve the same multi-task performance. However, researches on incomplete labels have been out of focus in the multi-task learning literature, *e.g.*, balancing between tasks and joint training of tasks. Figure 1 compares fully labeled and partially labeled datasets.



(a) Fully labeled dataset (b) Partially labeled dataset

Fig. 1. Comparison of different settings for multi-task learning of cell segmentation and pronucleus segmentation. (a) A fully labeled dataset, in which every image has both annotations for cell segmentation and pronucleus segmentation. (b) A partially labeled dataset, in which every image has annotations only for partial tasks.

In this paper, we develop a multi-task learning method tailored for partially labeled data. For each partially labeled data sample, not all tasks are simultaneously trainable with supervisions due to missing labels, which can induce different learning speeds between tasks. In multi-task learning, different learning speeds between tasks can lead to inferior performance [5]. Increasing model size or sharing fewer parameters has been known as a general solution, but either is yet another design choice. Instead, we propose a curriculum learning, which is applicable to existing multi-task models developed for fully-labeled data. We compose a sequential learning schedule for curriculum learning that makes a model to train less performing tasks more frequently with ground truth supervisions. It can prevent early convergence of quickly learned tasks. In our experiments, we conduct experiments for a partially labeled embryo dataset [4] in which each image has partial tasks' labels. We implement our curriculum learning method upon a baseline multi-task learning model, MTAN [6], to showcase our learning method's applicability to existing multi-task learning models. We demonstrate that our method improves the baseline on the biomedical dataset. To compare differ-

⋆equal contributions. This work has been completed while Won-Dong Jang was in Harvard University.

ent labeling strategies, we simulate the partially labeled setting on the two natural image datasets, NYU v2 [1] and Cityscapes [2], by using partial annotations for each image.

We have three major contributions. First, we introduce a tractable method for multi-task learning on partially labeled data. Our method can adapt existing models designed for the fully-labeled setting. Thus, our method will directly benefit from the improvement in the actively researched fully-labeled setting. Second, we address the problem of training tasks at different rates using the curriculum learning by taking into account each task's learning progress. Third, we investigate various possible labeling strategies to optimally assign limited labeling budgets for future dataset annotation. We show that partial annotations are more effective than full annotations.

2. RELATED WORKS

Biomedical Multi-task Learning. Many biomedical applications use multi-task learning. For skin cancer diagnosis, Coppola *et al.* [7] and Chen *et al.* [8] jointly trained classifiers and a segmentation model. Multi-task learning is beneficial for CT images. Bragman *et al.* [9] made a model to yield segmentation of organs and synthesize MRI. He *et al.* [10] developed a multi-task model that segment out organs while checking the existence of organs in CT images. Shi *et al.* [11] had collected CT images with multiple organs and performed multi-organ segmentation. However, most of them use fully labeled data, which is expensive to obtain. Shi *et al.* [11] utilized partially labeled data, but they still require fully labeled data during training.

Multi-task Learning for Partially Labeled Data. To the best of our knowledge, there is no multi-task learning work that assumes the identical setting to our partially labeled data. There are many multi-label learning algorithms with missing labels [12, 13, 14], but they focus on complementing missing labels. Semi-supervised multi-task learning is closest to our problem setting [15, 16, 17]. Chen *et al.* [16] developed a semi-supervised method that uses fully labeled data and totally unlabeled data, whereas we assume missing labels in each sample. He *et al.* [17] addressed partial input data instead of partial labels, where the model knows the ground truth labels in all tasks, but only partial input features. Liu *et al.* [15] predicted only one type of label, *e.g.,* like or dislike of artwork, and treated classification on each dataset, *e.g.,* each user, as a single task. Therefore, it is hard to apply these methods to multi-modal prediction tasks (*e.g.*, segmentation and classification) or add upon existing multi-task methods (*e.g.*, MTAN [6]).

Curriculum Learning. Curriculum learning [18] optimizes the learning order of data in training. It started from the observation that human learns faster when the learning difficulty gradually increases compared to learning randomly. Pentina *et al.*'s method [19] is the closest to our multi-task curriculum learning. They determine the learning order of tasks to learn highly related tasks sequentially. However, curriculum learning focuses on which sample to learn first while our method optimizes which task to learn with priority. Also, no curriculum learning method takes into account the learning frequencies of each task.

3. OUR METHOD

We first introduce a multi-task model's learning objective modified for partially labeled data and then propose a curriculum learning method that enables balanced learning of tasks.

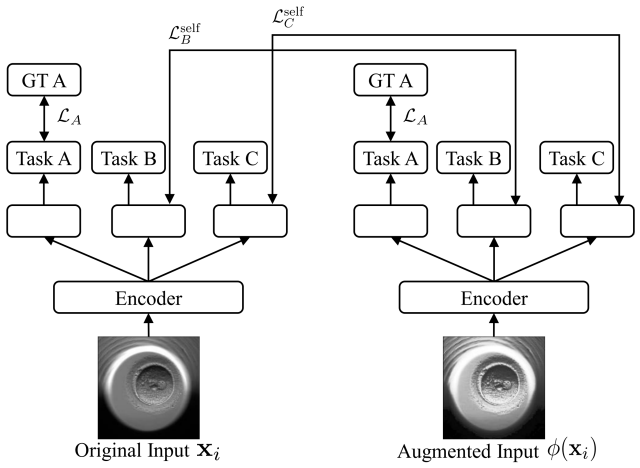


Fig. 2. Self-supervised learning. We first apply augmentation to an input image $\mathbf{x}_i$ to generate an augmented input image $\phi(\mathbf{x}_i)$. We feed both images into a model. We compute a loss for a labeled task L_A as well as absolute difference losses between outputs from unlabeled tasks L_B^{self} and L_C^{self}. We can calculate a loss L_A using either an original input or an augmented one for a labeled task.

3.1. Learning Objective for Partially Labeled Data

For a given set of tasks, a multi-task model $\mathcal{M}$ optimizes its parameters θ by minimizing all tasks' losses $\mathcal{L}_1, \ldots, \mathcal{L}_T$, where T indicates the number of tasks. For a fully labeled data sample, the learning objective is the sum of all tasks' losses. However, in a partially labeled dataset, each input sample is available to optimize only tasks with ground truths. We can formulate its learning objective as

$$\min_{\theta} \sum_{t \in \mathcal{T}_i} w_t \mathcal{L}_t(\mathcal{M}(\mathbf{x}_i, \theta)_t, \mathbf{y}_{i,t}), \tag{1}$$

where w_t denotes a weight for task t, $\mathbf{y}_{i,t}$ is a task t's ground truth for input $\mathbf{x}_i$, and $\mathcal{T}_i$ is a set of tasks with available ground truths for the sample $\mathbf{x}_i$. Based on the loss function (1), we update the model's parameters by

$$\theta^* = \theta - \alpha \sum_{t \in \mathcal{T}_i} w_t \nabla \mathcal{L}_t(\mathcal{M}(\mathbf{x}_i, \theta)_t, \mathbf{y}_{i,t}), \tag{2}$$

where α is a tunable step size parameter. Thus, gradients through tasks with no ground truths are not available, which can induce unstable learning, *e.g.*, imbalanced learning speeds or negative transfers between tasks.

Self-supervised learning can be a natural solution to address this issue, which makes unlabeled tasks trainable. Recently, there have been intensive researches in unsupervised and semi-supervised learning [20, 21, 22]. The key idea is to apply an augmentation to an input image and make a model to yield the same outputs for the original and augmented inputs. Inspired by this, we augment an input image and make a model return the same outputs for the input and augmented images, as shown in Figure 2. In this way, we can train both labeled and unlabeled tasks.

We first apply augmentation $\phi(\cdot)$ to an input image $\mathbf{x}_i$ to obtain an augmented image $\phi(\mathbf{x}_i)$ and feed them into a model. For the original and augmented images, we represent their intermediate features fed into the prediction layers for task t as $\hat{\mathcal{M}}(\mathbf{x}_i)_t$ and $\hat{\mathcal{M}}(\phi(\mathbf{x}_i))_t$, respectively. Then, we define a loss for self-supervised learning as

$$\mathcal{L}_t^{\text{self}}(\mathbf{x}_i) = \left| \hat{\mathcal{M}}(\mathbf{x}_i)_t - \hat{\mathcal{M}}\left(\phi(\mathbf{x}_i)\right)_t \right| \tag{3}$$

By adding it to the supervised loss, $\mathcal{L}_t$, the final loss function becomes

$$\sum_{t \in \mathcal{T}_i} \mathcal{L}_t(\mathbf{x}_i) + \lambda \sum_{t \in \mathcal{T}_i^{\complement}} \mathcal{L}_t^{\text{self}}(\mathbf{x}_i), \tag{4}$$

where λ is a parameter controlling the weights to the self-supervised loss. The final loss enables gradients to flow from all tasks even they are partially labeled.

3.2. Multi-task Curriculum Learning

Based on the loss function (4), a multi-task model for partially labeled data learns tasks in $\mathcal{T}_i$ with ground truth supervisions and the rest with self-supervisions. As such, determining which tasks to learn with supervisions at each iteration can differ the performance of the multi-task model. In this work, we define curriculum learning as a problem of determining what tasks to learn with supervisions at each iteration in every epoch. In other words, curriculum learning determines each task's learning frequency with supervisions, called *supervised learning frequency*, within each epoch. An adequate curriculum can maximize positive transfers between tasks by balancing their learning speeds.

As a baseline learning curriculum, one may train the model using the uniform curriculum, in which tasks take turn to get a supervision. At iteration k, the uniform curriculum determines a task with supervision $\tau(k)$ as $\mod(k, T) + 1$, where $\mod(\cdot, \cdot)$ is a modulo operator. However, the uniform curriculum can induce negative transfers between tasks, where sharing features leads to performance degradation due to task imbalance. If the model learns a task notably faster than other tasks, the model's multi-task performance can be underachieving at the end of training [5]. Widely used solutions to address the rate imbalance are sharing fewer parameters across tasks and training a bigger network to allow more flexibility in parameter sharing. However, these solutions are yet another design choices.

Instead, we propose a novel curriculum learning method, which is applicable to existing multi-task learning algorithms. We assign different supervised learning frequencies for each task based on the learning progress of each task. To estimate each task's learning progress, we measure relative performance gains after every epoch on a validation set by comparing validation scores to ones from pretrained single-task models. We define a score gain as

$$\Delta s_t = s_t / \hat{s}_t, \tag{5}$$

where s_t and $\hat{s}_t$ are performances on a validation set for task t of a multi-task model and the single-task model, respectively. Note that the single-task model is pretrained, and its final performance is used as $\hat{s}_t$. If Δs_t is lower than 1, it means a multi-task model is inferior to a single-task model. For metrics that lower values indicate better performance, we use the inverse of this score instead.

As different learning speeds between tasks can result in the underachievement of multi-task learning, we prioritize worse-performing tasks over better-performing ones by training the earlier with supervisions more frequently. To compose a curriculum, we define a frequency of learning task t with a supervision at each epoch as

$$f(t) = N \cdot \frac{\exp(-\Delta s_t / \sigma)}{\sum_i \exp(-\Delta s_i / \sigma)}, \tag{6}$$

where σ is a temperature and N is the number of iterations per epoch. The temperature controls the influence of the performance gains in determining the supervised learning frequency of each task. If the temperature is huge, our curriculum becomes a uniform sequential learning, which has the same supervised learning frequencies for all tasks.

Table 1. Comparison of our method with different settings to MTAN on the embryo test dataset [4]. We add the self-supervised learning (SL) and multi-task curriculum learning (CL) to MTAN one at a time. All scores are higher better. The best result is boldfaced.

Method	Stage Accuracy	Cell mAP	Pronucleus mAP	Overall *mean* Δs_t
Single Task	80.14	69.95	67.37	1
MTAN	80.21	68.55	65.51	0.984
MTAN + SL (Ours)	79.89	69.74	66.14	0.992
MTAN + CL (Ours)	82.37	70.32	66.49	1.007
MTAN + SL + CL (Ours)	81.84	70.59	66.86	**1.008**

Based on the supervised learning frequencies, we schedule a training curriculum for the next epoch, *i.e.*, we determine a task to sample at each iteration, $\tau(k)$, using the tasks' supervised learning frequencies, $f(1), \ldots, f(T)$. To prevent excessive imbalances between tasks in a single epoch, we maximize the number of alternations of tasks in the learning curriculum. We update the learning frequencies every epoch and compose a learning curriculum accordingly.

4. EXPERIMENTS

As baseline multi-task learning methods, we chose Cross-Stitch [23] and MTAN [6] to showcase our method's efficacy, as they are well-acknowledged in the field [24, 25] and highly reproducible. For the baseline methods, we exploit a uniform curriculum learning, in which all tasks have the same frequency for supervised learning. When possible, we compose a batch with samples from different tasks to train the baselines to cover tasks maximally. We benchmark our method on the embryo dataset [4]. We also evaluate our method on NYU v2 [1] and Cityscapes [2], which are the standard benchmark datasets for multi-task learning, to show our method's generalizability and simulate different experimental settings. We will make our code publicly available, including architectural details, augmentation policies, training information, and detailed benchmark results.

4.1. Embryo Light Microscopy Dataset

Visual analysis of embryos is necessary for *in vitro fertilization*. We define a multi-task learning problem that aims at addressing the three tasks introduced in the recent embryo analysis work [4]. The three tasks are stage classification, cell instance segmentation, and pronucleus instance segmentation.

The embryo dataset [4] consists of partially labeled data. Stage classification aims at classifying the developmental status of an embryo, which is one of 13 classes. The purpose of cell instance segmentation is to segment cell instances. A goal of pronucleus instance segmentation is to detect and segment pronuclei, which only appear in embryos at 1 cell stage. Stage classification performance is accessed using classification accuracy. We measure mean average precision (mAP) for the cell and pronucleus instance segmentation tasks. Finally, we report a *mean* Δs_t, which is an average of score gains across tasks, to assess multi-task models' overall performance. We believe the *mean* Δs_t is more important than the task-wise metrics, as it measures overall performance.

To build a multi-task model, we incorporate MTAN [6] to Mask R-CNN [26]. Due to the large memory requirement from Mask R-CNN, Cross-Stitch [23], a soft parameter sharing model, was not usable. For stage classification, we apply a global average pool-

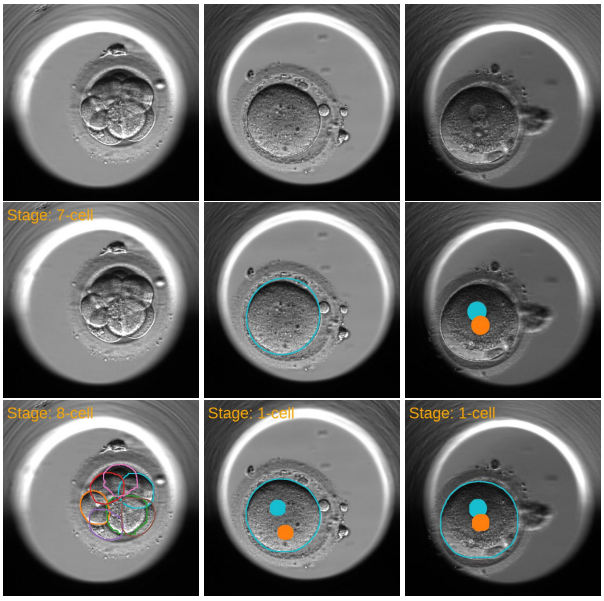


Fig. 3. Multi-task prediction results on the embryo dataset [4]. From the top to the bottom, each row presents input images, ground-truths (partially labeled), and our prediction results, respectively.

ing and a fully-connected layer to the Mask R-CNN's backbone feature. For cell and pronucleus instance segmentation, we use Mask R-CNN's prediction heads, such as RPN, classification and bounding box heads, and segmentation head. The model has 117M parameters. We set σ and λ as 0.02 and 0.1, respectively.

Table 1 compares our multi-task curriculum learning with the MTAN baseline. Our method (MTAN + SL + CL) improves the baseline (MTAN) by 2.4%. Moreover, our method outperforms the baseline across all tasks. It demonstrates that applying our learning strategy to an existing multi-task learning method can lead to a performance increase. Figure 3 shows the qualitative results of our multi-task curriculum learning.

Cumulative Effects. To show the efficacy of our multi-task curriculum learning, we conduct ablation studies. We add self-supervised learning (SL) and multi-task curriculum learning (CL) one at a time to the baselines. Table 1 includes the results with the different model settings. While the self-supervised learning (MTAN + SL) marginally increases the performance by 0.8%. Our curriculum learning (MTAN + CL) improves the baseline by 2.3%. By adding the curriculum learning to the self-supervised learning (MTAN + SL + CL), we achieve the best performance across tasks.

4.2. Applications for Natural Images

Our subsequent interest is the comparison between various labeling budget assignment strategies. The majority of multi-task datasets have their data either fully labeled or labeled separately without any overlap. We are interested in how creating some subset of fully labeled data influences the performance of our method.

Since our biomedical dataset has limited annotation overlap between tasks, we use two standard datasets for multi-task learning instead: NYU v2 [1] and Cityscapes [2]. Both datasets consist of fully labeled data, *i.e.*, each image has annotations for all tasks. Hence, we can partially omit ground truths to simulate datasets with different task overlaps. We define three settings, 0%, 50%, and 100%

Table 2. Results on the NYU v2 [1] and Cityscapes [2] evaluation sets. We set three types of task overlaps. Each entry shows *mean*±*std*. The better results are boldfaced for each comparison.

Task Overlap	Method	*mean* Δs_t NYU v2	*mean* Δs_t Cityscapes
100%	Single Task	1	1
	Cross-Stitch	1.051±0.026	0.686±0.02
	MTAN	1.080±0.015	0.962±0.04
50%	Single Task	1	1
	Cross-Stitch	1.105±0.029	0.651±0.01
	Cross-Stitch + Ours	**1.130**±0.045	**0.697**±0.02
	MTAN	1.094±0.015	0.924±0.05
	MTAN + Ours	**1.142**±0.014	**1.050**±0.09
0%	Single Task	1	1
	Cross-Stitch	1.041±0.048	0.981±0.25
	Cross-Stitch + Ours	**1.072**±0.066	**0.991**±0.09
	MTAN	1.066±0.016	1.372±0.36
	MTAN + Ours	**1.130**±0.007	**1.419**±0.39

task overlaps. If the overlap ratio is 0%, task labels are mutually exclusive; each image has a label only for one of the tasks. In contrast, if the overlap ratio is 100%, a training set becomes a fully labeled dataset; every image has labels for all the tasks. We set the same labeling budget, a total number of labels, for all overlap ratios. NYU v2 [1] consists of indoor images and their corresponding annotations for semantic segmentation, depth estimation, and surface normal estimation. Cityscapes [2] contains street-view images and their corresponding annotations for semantic segmentation and inverse depth estimation. For evaluation, we follow the benchmarking method used in MTAN [6]. As an overall metric, we use a *mean* Δs_t. We run the multi-task learning methods five times and report their mean and standard deviation.

Table 2 compares ours to the baselines on NYU v2 and Cityscapes. We report results with 100%, 50%, and 0% task overlaps. Our method improves the two baseline methods and the single-task models across all three tasks. We observe that our method with 50% and 0% task overlaps yield better performance than the baseline methods do with 100% task overlap.

Performance Comparison. While our method provides a moderate performance increase in embryo analysis, it significantly improves the baselines on the natural image datasets. We conjecture this comes from the differences between the datasets. While the natural image datasets are diverse and complex, the embryo images have limited diversity. It may imply that our method performs better when labeling is expensive, and data varies a lot. Our future works include follow-up research to analyze this phenomenon.

Labeling Budget. We note that our performance on 50% overlap is better than both on 0% and 100%. Hence, we believe it is possible to achieve a higher multi-task learning performance by partially labeling the data with an overlap instead of full annotations.

5. CONCLUSIONS

We have introduced a multi-task learning method for partially labeled data. Our problem setting is realistic, as collecting all labels for every image may not be possible. This is especially serious when labeling is extremely expensive, *e.g.*, biomedical labels. Since our multi-task curriculum learning is a simple add-on for any multi-task framework, it could be a basic technique for partially labeled data.

Acknowledgment

The authors thank the dedicated team of embryologists and medical professionals at the Institution of Reproduction and IVF, Lis Maternity Hospital, Tel Aviv Sourasky Medical Center. This work was funded in part by NIH grants 5U54CA225088 and R01HD104969, NSF Grant NCS-FO 1835231, the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard (award number 1764269), the Harvard Quantitative Biology Initiative, and Sagol fund for studying embryos and stem cells; Perelson Fund.

Compliance with Ethical Standards

This research study was conducted retrospectively using human subject data made available in restricted access by Tel-Aviv Medical Center, Israel. Ethical approval was not required as we reused previously introduced data.

6. REFERENCES

[1] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*. 2012, pp. 746–760, Springer.

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.

[3] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, and Bjoern Menze, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019.

[4] B. D. Leahy, W.-D. Jang, H. Y. Yang, R. Struyven, D. Wei, Z. Sun, K. R. Lee, C. Royston, L. Cam, and Y. Kalma, "Automated Measurements of Key Morphological Features of Human Embryos for IVF," *arXiv preprint arXiv:2006.00067*, 2020.

[5] Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei, "Dynamic task prioritization for multitask learning," in *ECCV*, 2018, pp. 270–287.

[6] Shikun Liu, Edward Johns, and Andrew J. Davison, "End-to-end multi-task learning with attention," in *CVPR*, 2019.

[7] Davide Coppola, Hwee Kuan Lee, and Cuntai Guan, "Interpreting Mechanisms of Prediction for Skin Cancer Diagnosis Using Multi-Task Learning," in *CVPR Workshops*, 2020.

[8] S. Chen, Z. Wang, J. Shi, B. Liu, and N. Yu, "A multi-task framework with feature passing module for skin lesion classification and segmentation," in *ISBI*, Apr. 2018.

[9] Felix JS Bragman, Ryutaro Tanno, Sebastien Ourselin, Daniel C. Alexander, and Jorge Cardoso, "Stochastic filter groups for multi-task cnns: Learning specialist and generalist convolution kernels," in *ICCV*, 2019.

[10] Tao He, Junjie Hu, Ying Song, Jixiang Guo, and Zhang Yi, "Multi-task learning for the segmentation of organs at risk with label dependence," *Medical Image Analysis*, vol. 61, pp. 101666, Apr. 2020.

[11] Gonglei Shi, Li Xiao, Yang Chen, and S. Kevin Zhou, "Marginal loss and exclusion loss for partially supervised multi-organ segmentation," *arXiv:2007.03868 [cs]*, July 2020, arXiv: 2007.03868.

[12] Baoyuan Wu, Zhilei Liu, Shangfei Wang, Bao-Gang Hu, and Qiang Ji, "Multi-label learning with missing labels," in *ICPR*. 2014, pp. 1964–1968, IEEE.

[13] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon, "Large-scale multi-label learning with missing labels," in *ICML*. 2014, pp. 593–601, PMLR.

[14] Wei Bi and James Kwok, "Multilabel classification with label correlations and missing labels," in *AAAI*, 2014, vol. 28.

[15] Qiuhua Liu, Xuejun Liao, Hui Li Carin, Jason R. Stack, and Lawrence Carin, "Semi-supervised multitask learning," in *NeurIPS*, 2007.

[16] Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng, "A Multi-task Mean Teacher for Semi-supervised Shadow Detection," in *CVPR*, 2020.

[17] Yi He, Baijun Wu, Di Wu, and Xindong Wu, "On Partial Multi-Task Learning," in *European Conference on Artificial Intelligence*, 2020, p. 8.

[18] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston, "Curriculum learning," in *ICML*, 2009, pp. 41–48.

[19] Anastasia Pentina, Viktoriia Sharmanska, and Christoph H. Lampert, "Curriculum learning of multiple tasks," in *CVPR*, 2015.

[20] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *arXiv preprint arXiv:2001.07685*, 2020.

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020, pp. 9729–9738.

[22] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[23] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert, "Cross-stitch networks for multi-task learning," in *CVPR*, 2016, pp. 3994–4003.

[24] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn, "Gradient Surgery for Multi-Task Learning," in *NeurIPS*, 2020.

[25] Ximeng Sun, Rameswar Panda, Rogerio Feris, and Kate Saenko, "AdaShare: Learning What To Share For Efficient Deep Multi-Task Learning," in *NeurIPS*, 2020.

[26] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *ICCV*, 2017.