基于图的半监督学习理论及其应用

苏 航，董胤蓬，胡文波

清华大学

0 引言

随着计算机、通信、传感等信息技术的飞速发展，以及互联网应用的日益普及，人们能够以更加快速、方便、廉价的方式来获取和存储数据资源，使得数字化信息以指数方式迅速增长。面对极度膨胀的数据量，传统的以标注数据为基础的机器学习模型已经远远不能满足数据分析和处理的要求，使人们处于数据到处“泛滥”而所获知识和价值甚少的困境[1]。半监督学习在一定程度上避免了数据资源的浪费，通过充分地利用少量的有监督数据和大量的无监督数据来改善算法性能，从而解决了监督学习的模型泛化能力不强和无监督学习的模型不精确等问题。因此，半监督学习可以最大限度地发挥数据的价值，使机器学习模型从体量巨大、结构繁多的数据中挖掘出隐藏在背后的规律，成为近年来机器学习领域比较活跃的研究方向，被广泛应用于社交网络分析、文本分类、计算机视觉和生物医学信息处理等诸多的领域[2-3]。

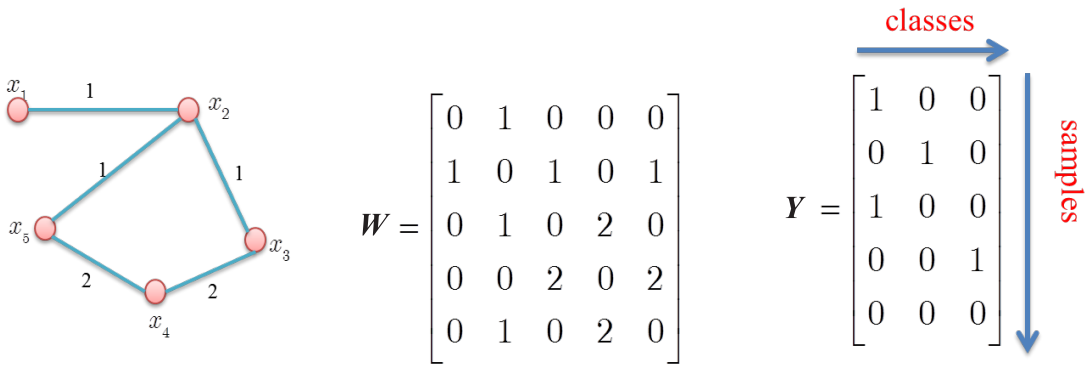
基于图的半监督学习方法是半监督学习研究中被广泛采用的一种方法，近年来有大量的工作专注在此领域，也产生了诸多卓有成效的进展[4]。该方法将数据样本间的关系映射为一个相似度图，其中图的节点表示数据点(包括标记数据和无标记数据)；图的边被赋予相应权重，代表数据点之间的相似度，通常来说相似度越大，权重越大。对无标记样本的识别可以通过图上标记信息传播的方法实现，节点之间的相似度越大，标签越容易传播；反之，传播概率越低。基于图的半监督学习算法简单有效，符合人类对于数据样本相似度的直观认知，又可以针对实际问题来灵活地定义数据之间的相似性，具有很强的灵活性。尤其需

要指出的是，基于图的半监督学习具有坚实的数学基础作为保障，通常可以得到闭式最优解，因此具有广泛的适用范围。该方法的代表性论文也因此获得了 2013 年国际机器学习大会“十年最佳论文奖”[5]，可看出该范式的影响力和重要性。

本文将在总结基于图的半监督学习的基本理论基础上，结合近年来一些最新进展，对基于图的半监督学习的学习方法、主要进展、相关应用，以及未来发展做综述简介。

1 基于图的半监督学习一般性框架

在众多的半监督学习算法中，基于图的半监督学习是近年来最为活跃的研究领域之一（见图 1）。假设有标记样本集为 $L=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}$ 以及未标记的样本集合 $U=\{x_{l+1},x_{l+2},\cdots x_n\}$，其中 x_i 为样本点、y_i 为相应的标签。该类方法利用有标注和无标注的数据构建数据图 $G=(V,E)$，其中顶点 V 表示样本点，包括了有标记和未标记的样本；E 表示连接两个顶点的相似程度。半监督学习可以通过基于图的邻接关系，将标签从有标签的数据点传播到无标签数据点，从而实现相应样本点的分类。



(a) 相似度图　　(b) 邻接矩阵　　(c) 标签矩阵

图 1 基于图的半监督学习

在标签传播过程中，保持已标注数据的标签不变，使其像一个源头把标签传向未标注节点。每个节点根据相邻节点的标签来更新自己的标签，与该节点相似度越大，其标签就越容易传播到相邻节点，相似节点的标签越趋于一致。当迭代过程结束时，相似节点的概率分布也趋于相似，可以划分到同一个类别中，从而完成标签传播过程。标签传播在数学上可以通过多种方式来实现，其中比较经典的是高斯场和谐波函数方法[5]，该方法将图上每个样本点的标签由离散值松弛到了连续域，并将随机场上的能量函数优化问题转换成了高斯场上的能量函数优化问题，可以通过求解目标函数

$$\boldsymbol{Y}_u^*=\underset{\boldsymbol{Y}_u\in\{0,1\}}{\arg\min}\ \mathrm{tr}\left([\boldsymbol{Y}_l;\boldsymbol{Y}_u]^{\mathrm{T}}\begin{bmatrix}\boldsymbol{L}_{ll}&\boldsymbol{L}_{lu}\\\boldsymbol{L}_{ul}&\boldsymbol{L}_{uu}\end{bmatrix}[\boldsymbol{Y}_l;\boldsymbol{Y}_u]\right) \tag{1}$$

来实现。其中，$\boldsymbol{X}_l$ 和 $\boldsymbol{Y}_u$ 分别是标记样本和无标记样本的标签矩阵；$\boldsymbol{L}=\begin{bmatrix}\boldsymbol{L}_{ll}&\boldsymbol{L}_{lu}\\\boldsymbol{L}_{ul}&\boldsymbol{L}_{uu}\end{bmatrix}$ 是相似度图的拉普拉斯矩阵。对方程（1）中

的 $\boldsymbol{Y}_u$ 微分，可以得到相应的闭式解为

$$\boldsymbol{Y}_u^* = -\boldsymbol{L}_{uu}^{-1}\boldsymbol{L}_{ul}\boldsymbol{Y}_l \tag{2}$$

此方法通过对数值域进行松弛，使得顶点的边缘概率密度具有封闭形式解，并且后续的很多方法都采用了相似的松弛，包括局部和全局一致性学习 [7]、基于流形正则化 [8] 等方法。

1.1 核心问题

基于图的半监督学习需要利用图来模拟低维流形，因此图的构造在模型学习当中起到的是核心作用，包括如何构建边，以及如何给边赋权重 [9]。边的构建就是解决如何保持样本空间的局部几何结构的问题，目前广泛采用的方法是构建 ε 近邻图或者 k-最近邻图的方法来实现 [6]。另一个问题是如何给边赋权重，由于样本的低维流形分布未知，因此很难通过计算任意两点沿着流形的距离来定义边的权重，这也成为影响模型性能的最重要因素。目前普遍采用的是基于欧式距离及其变体的距离定义，如高斯核函数、余弦核函数等。但这也导致了目前图的结构严重依赖于用户手动设定的相关距离参数，当参数设定不合理或者参数的设定和任务不匹配时，就会严重地影响半监督学习算法的性能，甚至导致其比监督学习或者无监督学习更差的性能，从而成为制约半监督学习算法应用的最重要因素之一。目前，如何根据问题选取合适的模型参数，仍然是一个亟待解决的问题。

1.2 基本假设

需要指出的是，同其他半监督学习的方法类似，以上基于图的半监督学习的成立依赖于如下的 2 个基本假设 [6]。

(1) **聚类假设**：当两个样本位于同一聚类簇时，它们在很大的概率下有相同的标签。此假设表明，不同标签样本的分界面不应该出现在样本密度较大的区域，因此大量未标记样本能够帮助标明样本空间中数据分布的稠密和稀疏区域。

(2) **流形假设**：位于一个低维流形的局部极小邻域的样例，在很大的概率下具有相同的标签。这一假设主要考虑模型的局部特性，利用大量未标记样本使得数据空间更加稠密，从而有助于更加准确地刻画数据的局部特性。

通常情况下，聚类假设和流形假设是不矛盾的，分别着眼于数据的整体特征和局部特征。基于图的半监督学习算法依赖于这些基本的假设。大量的研究表明，当不满足这些基本假设时，无监督的样本不仅不能改进模型性能，反而起到恶化的作用 [6]。因此，如何使模型满足这些条件，或者在部分条件不满足的情况下更加“安全”地利用无监督样本仍然是一个活跃的研究领域。

2 近期若干进展

2.1 半监督主动学习

同半监督学习类似，主动学习是利用未标记样本的另一大类技术 [9]，在学习的过程中自行挑选出一些未标记样本并要求用户进行标记，然后将这些样本加入训练集进行常规的监督学习，其技术难点在于如何使用尽可能少的查询来提升泛化性能。显然，若能将半监督学习与主动学习相结合，极大地促进对未标记样本进行利用。2003 年，Zhu 等 [10]

首次提出将主动学习和半监督学习相结合的技术，在深入研究标签传播理论的基础上，通过贪心算法，寻找使模型性能获得最大提升的样本点进行标记。近年来，Su 等[11]进一步利用了拉德马赫复杂度（Rademacher Complexity）理论，通过最小化的预测误差期望，来批量化的选取最有信息量的样本点进行人工标记，并成功地应用于图像分割中。到目前为止，主动学习和半监督学习的融合仍然是一个极有价值的研究方向，有很多问题尚待解决。

2.2 在线修正传播理论

在实际应用中，由于问题的复杂性、时间演进等原因，不可避免地存在大量的分类错误。在很多应用中，用户往往会重新介入到分类的任务中，对其中的错误进行重新修订。可以想像的是在样本空间中，往往存在大量的类似错误，理想的情况是人工修正也能够同时对相似的错误进行修正。但是很显然，将这些用户修正的样本当作标定样本进行重新训练需要消耗大量的时间和计算量，这就需要采用增量学习的模式，在既有模型的基础上有针对性地进行微调，从而实现模型的动态演进。

为了解决这一问题，Su 等[12]提出了基于修正传播算法的模型动态更新算法，通过引入一个修正传播矩阵，可以有效地将人工修正传播到其他的未标记样本中，从而极大地提高了人工修正的利用效率，降低修正传播的计算量。作者通过在原有的相似度图中引入了虚监督节点，并将其关联到被修正的节点上，将关联边的权值设为 $+\infty$，从而有效地将人工的修正等效传导到整个样本空间中。如图（2）所示，其中深蓝圆圈代表了虚监督，其关联的圆圈代表了被人工修订的样本（记所有和虚监督关联的样本点集合为 $\boldsymbol{K}$），推导得到修正传播可以通过如下的方程实现：

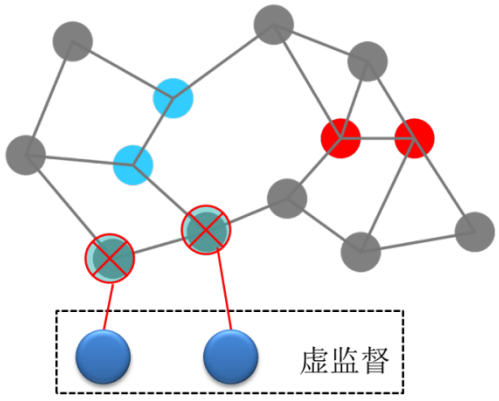


图 2 基于增广图的修正传播

$$\boldsymbol{Y}_u^+ = \boldsymbol{Y}_u + \boldsymbol{\Gamma}_{uu}(:,\boldsymbol{K})\left(\boldsymbol{\Gamma}_{uu}(\boldsymbol{K},\boldsymbol{K})\right)^{-1}(\boldsymbol{Y}_s - \boldsymbol{Y}_u(\boldsymbol{K})) \tag{3}$$

其中，$\boldsymbol{\Gamma}_{uu}=\boldsymbol{L}_{uu}^{-1}$，是无标记样本所对应的拉普拉斯矩阵的逆；$\boldsymbol{\Gamma}_{uu}(:,\boldsymbol{K})\left(\boldsymbol{\Gamma}_{uu}(\boldsymbol{K},\boldsymbol{K})\right)^{-1}$是对应的修正传播矩阵，对应的是$\boldsymbol{\Gamma}_{uu}$中被修订的样本。这样，修正传播算法只需要计算一个很小的矩阵的逆（$\left(\boldsymbol{\Gamma}_{uu}(\boldsymbol{K},\boldsymbol{K})\right)^{-1}$），算法的效率将得到极大提高。相关研究者应用修正传播算法已经在图像分类[12]、生物图像分割[13]等诸多任务中取得了很好结果。类似的方法也可以应用于其他的半监督学习框架中。

2.3 半监督学习的高效算法

计算效率始终是半监督学习中所面临的重要瓶颈之一，由于基于图的半监督学习往往要进行拉普拉斯矩阵的逆运算，其时间复杂度是 $O(N^3)$，甚至出现内存溢出的现象，难以应用于大规模数据的运算。因此如何提高半监督学习算法的效率，一直是一个研究

的热点。Zhang[14] 在 2009 年利用 Nystrom 近似解决相似性矩阵规模较大问题，但这个算法存在不能保证相似性矩阵半正定的问题。Fergus 等 [15] 提出在标记预测函数使用图拉普拉斯的光滑特征向量，但此算法过分依赖维度可分离的数据密度假设。针对这些问题，Liu 等 [17] 提出基于锚点的建图方式，以及相应的半监督分类算法框架，有效地解决了对大数据集的分类问题。但是这一方法采用无监督聚类的方法得到锚点，对噪声较为敏感；同时将节点关联为固定数量的锚点，自适应性较差。针对这一问题，Su 等 [18] 在锚点的构建中引入了稀疏性的约束，可以通过字典学习的方式，自适应学习得到相应的锚点，以及节点和锚点之间的关联关系，从而有效提升了算法的性能。

3 典型应用

基于图的半监督学习技术已在众多领域中得到广泛应用，下面仅撷几例说明。

3.1 生物图像分析

在生物图像的分析中，每天大量的生物实验往往很容易得到海量的无类标签的样例，但往往需要使用特殊设备或经过昂贵且用时非常长的实验过程，进行人工标记才能得到有类标签的样本 [6]。因此，如何将半监督学习应用到生物图像分析中改进系统性能，是一个非常有价值的研究方向。相关的研究者在此方面已经进行了大量的探索，并成功将半监督学习算法应用于相差显微镜图像下的细胞分割问题 [19]（如图 3 所示），以及双光子显微镜的脑细胞的分割 [20] 中。

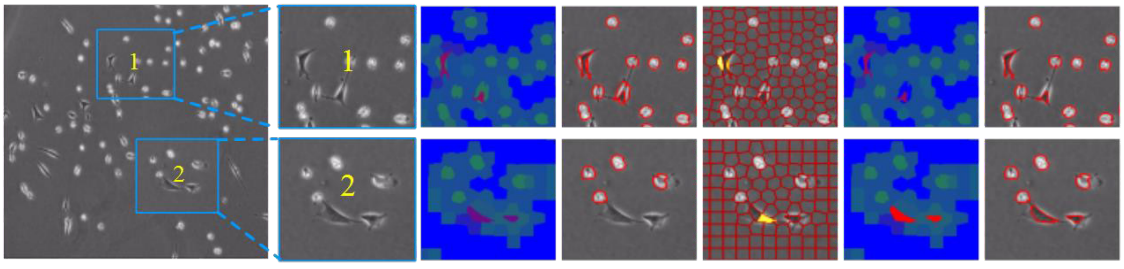


图 3 半监督学习显微图像下的细胞分割 [19]

3.2 文本分析

半监督学习比较典型的应用就是在自然语言处理领域的应用。由于互联网的日益发达，指数级增长的网络资源能进行人工标记的网页微乎其微，给半监督学习提供了广泛的应用场景。He 等 [21] 提出将基于图的半监督学习和生成模型结合的方法，利用图的传播来估计条件概率，用线性回归估计类先验信息，为基于图的半监督学习和主题模型等生成式模型的结合提供了可能。近年来，Hu 等 [22] 提出将拉普拉斯平滑约束作为正则项，引入到 LDA 等贝叶斯主题模型的框架中 [23]，从而实现了问题的主题提取和文档分类。

3.3 哈希算法

哈希学习通过将数据表示成二进制码的形式，可以显著减少数据的存储和通信开销，成为大数据学习中的一个研究热点。基于半监督学习的哈希学习算法可以充分利用样本的监督学习和海量的无标记数据，近年来受到了广泛的关注 [24]。基于图半监督学习的方法由于其可以很好表征样本之

间的相似度关系，可以和度量学习进行深度的融合，从而可以得到更加紧致的哈希编码，并且广泛地应用于图像和视频的检索问题中[25]。

4 结束语

经过多年的发展，基于图的半监督学习获得了快速的发展，在很多方面的技术已经趋于成熟，并逐渐应用于很多实际问题中。但是需要指出的是仍然存在很多有待研究的问题。

首先，目前的半监督学习方法鲁棒性不足。一方面表现在目前的方法在某些情况下无法保证实现“安全”的半监督学习，即在利用了未标记样本后，性能不但没有提升，反而还会有所下降。因此，如何更加安全地利用无标记样本，从理论和技术上仍然有很多问题需要探索[26]。另一方面，随着互联网和众包技术的发展，人们可以获得相对大量的标记，但是这类标记通常质量较差，还存在大量的噪声，甚至恶意的误导，如何更加有效地利用这些低质量的标注数据是半监督学习领域又一个重要的研究方向[27]。

其次，近年来以深度学习为代表的新一代人工智能方法取得一系列重要突破，受到了前所未有的关注，但是这些方法往往依赖大量的高质量训练数据。在训练数据量有限的情况下，深度神经网络的性能往往受到很大局限，一些规模巨大的深度神经网络也容易出现过拟合，使得测试性能远低于训练性能。因此，如何结合深度学习和半监督学习的优势，充分利用人工标定和海量的无标记数据将是未来非常有价值的研究方向。一些初步的半监督深度学习方法的尝试在较为简单的情形中取得了较好的效果[28]。随着深度学习理论的深入尤其是深度生成模型的发展[29]，相关的研究者在这方面进行了大量的探索，使得深度学习在半监督学习领域的在图像的生成和分类等诸多应用中，都取得了显著的进展[30]。

参考文献

[1] C. Lynch.Big data: How do your data grow. Nature, no. 7209 (2008): 28-29.

[2] 周志华 . 基于分歧的半监督学习 . 自动化学报 39, no. 11 (2013): 1871-1878.

[3] 刘建伟，刘媛，罗雄麟 . 半监督学习方法 . 计算机学报 38, no. 8 (2015): 1592-1617.

[4] Liu, Wei, Jun Wang, and Shih-Fu Chang. Robust and scalable graph-based semisupervised learning.Proceedings of the IEEE100, no. 9 (2012): 2624-2638.

[5] Zhu, Xiaojin, Zoubin Ghahramani, and John D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In Proceedings of the 20th International conference on Machine learning (ICML-03), pp. 912-919. 2003.

[6] Subramanya, Amarnag, and Partha Pratim Talukdar. Graph-based semi-supervised learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 8, no. 4 (2014): 1-125.

[7] Zhou, Denny, Olivier Bousquet, Thomas N. Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In Advances in neural information processing systems, pp. 321-328. 2004.

[8] Belkin, Mikhail, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of machine learning research 7, no. Nov (2006): 2399-2434.

[9] Liu, Wei, and Shih-Fu Chang. Robust multi-class transductive learning with graphs.In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 381-388. IEEE, 2009.

[10] Zhu, Xiaojin, John Lafferty, and Zoubin Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining, vol. 3. 2003.

[11] Hang Su, Hua Yang, Shibao Zheng, Sha Wei, Shuang Wu.Towards Active Annotation for Detection of Numerous and Scattered Objects.Multimedia and Expo, IEEE International Conference on (ICME), 2015.

[12] Hang Su, Zhaozheng Yin, Takeo Kanade, Seungil Huh.Active Sample Selection and Correction Propagation on a Gradually-Augmented Graph. Computer Vision and Pattern Recognition, IEEE Conference on (CVPR), 2015.

[13] Su, Hang, Zhaozheng Yin, Seungil Huh, Takeo Kanade, and Jun Zhu. Interactive cell segmentation based on active and semi-supervised learning. IEEE transactions on medical imaging 35, no. 3 (2016): 762-777.

[14] K. Zhang, J. T. Kwok, and B. Parvin.Prototype vector machine for large scale semi-supervised learning. in Proc. Int. Conf. Mach. Learn., 2009, pp. 1233-1240.

[15] Fergus, Rob, Yair Weiss, and Antonio Torralba. Semi-supervised learning in gigantic image collections.In Advances in neural information processing systems, pp. 522-530. 2009.

[16] Wang, M., Fu, W., Hao, S., Tao, D. and Wu, X., 2016. Scalable semi-supervised learning by efficient anchor graph regularization. IEEE Transactions on Knowledge and Data Engineering, 28(7), pp.1864-1877.

[17] Liu, Wei, Junfeng He, and Shih-Fu Chang. Large graph construction for scalable semi-supervised learning. In Proceedings of the 27th international conference on machine learning (ICML-10), pp. 679-686. 2010.

[18] Su, Hang, Jun Zhu, Zhaozheng Yin, Yinpeng Dong, and Bo Zhang. Efficient and Robust Semi-supervised Learning Over a Sparse-Regularized Graph. In European Conference on Computer Vision, pp. 583-598. Springer International Publishing, 2016.

[19] Hang Su, Zhaozheng Yin, Seungil Huh, Takeo Kanade.Cell segmentation in phase contrast microscopy images via semi-supervised classification over optics-related features. Medical Image Analysis, vol. 17, pp. 746-765, Oct. 2013.

[20] Kun Xu, Hang Su, Jun Zhu, Ji-Song Guan, and Bo Zhang. “Neuron Segmentation Based on CNN With Semi-Supervised Regularization.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 20-28. 2016.

[21] He, Jingrui, Jaime G. Carbonell, and Yan Liu. Graph-Based Semi-Supervised Learning as a Generative Model.In International Joint Conference on Artificial Intelligence (IJCAI), vol. 7, pp. 2492-2497. 2007.

[22] Wenbo Hu, Jun Zhu, Hang Su, Jingwei Zhuo, and Bo Zhang. Semi-supervised Max-margin Topic Models with Manifold Posterior Regularization.In International Joint Conference on Artificial Intelligence (IJCAI), Melbourne, Australia.

[23] Jun Zhu, Ning Chen, Eric P. Xing. Bayesian inference with posterior regularization and applications to infinite latent SVMs. Journal of Machine Learning Research, 2014, 15: 1799.

[24] Liu W, Wang J, Kumar S, et al.Hashing with graphs. In: Proceedings of the 28th International Conference on Machine Learning (ICML), Washington, 2011, 1-8.

[25] Zhang, S., Li, J., Jiang, M., & Zhang, B. (2017). Scalable Discrete Supervised Multimedia Hash Learning with Clustering.IEEE Transactions on Circuits and Systems for Video Technology (IEEE TCSVT).

[26] Li, Yu-Feng, and Zhi-Hua Zhou. Towards making unlabeled data never hurt.IEEE Transactions on Pattern

Analysis and Machine Intelligence 37, no. 1 (2015): 175-188.

[27] Li, Yu-Feng, James T. Kwok, and Zhi-Hua Zhou. Towards Safe Semi-Supervised Learning for Multivariate Performance Measures.In The AAAI Conference on Artificial Intelligence, pp. 1816-1822. 2016.

[28] Weston J, Ratle F, Mobahi H, et al. Deep learning via semi-supervised embedding. Neural Networks: Tricks of the Trade. Springer Berlin Heidelberg, 2012: 639-655.

[29] Kumar, Abhishek, Prasanna Sattigeri, and P. Thomas Fletcher. Improved Semi-supervised Learning with GANs using Manifold Invariances. In Advances in Neural Information Processing Systems (NIPS 2017).

[30] Li, Chongxuan, Kun Xu, Jun Zhu, and Bo Zhang. Triple Generative Adversarial Nets.In Advances in Neural Information Processing Systems (NIPS 2017).

苏航

清华大学计算机科学与技术系助理研究员。CCF 计算机视觉专家委员会委员，CCAI 机器学习专业委员会通讯委员。在 CVPR、IJCAI 等国际会议和期刊发表论文 50 余篇，多次受邀担任 TPAMI、TIP 等国际期刊的审稿人，以及 ICML、CVPR 等国际会议的技术委员会委员或审稿人。承担自然科学基金、“973”等多个国家级项目。曾获得 MICCAI“青年学者”、AVSS 的“最佳论文”等重要奖项。主要研究方向为计算机视觉、机器学习。

董胤蓬

清华大学计算机科学与技术系在读博士研究生。在 CVPR、IJCAI 等国际会议和期刊发表论文近 10 篇，多次率队获得国际人工智能相关比赛冠军，曾获 CCF 优秀大学生、北京市优秀毕业生和清华大学“钟士模”奖学金。

胡文波

清华大学计算机科学与技术系在读博士。在 IJCAI、NSR 等国际会议和期刊发表多篇论文，曾获得清华大学“斯伦贝谢”奖学金。