计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(7):239−244 [doi: 10.15888/j.cnki.csa.007499]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于可变形卷积的手绘图像检索①

王文超

(中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
通讯作者: 王文超, E-mail: wwchao1227@163.com

摘　要: 手绘图像仅包含简单线条轮廓, 与色彩、细节信息丰富的自然图像有着截然不同的特点. 然而目前的神经网络大多针对自然图像设计, 不能适应手绘图像稀疏性的特性. 针对此问题, 本文提出一种基于可变形卷积的手绘检索方法. 首先通过 Berkerly 边缘检测算法将自然图转化为边缘图, 消除域差异. 然后将卷积神经网络中的部分标准卷积替换为可变形卷积, 使网络能够充分关注手绘图轮廓信息. 最后分别将手绘图与边缘图输入网络并提取全连接层特征作为特征描述子进行检索. 在基准数据集 Flickr15k 上的实验结果表明, 本文方法与现有方法相比能够有效提高手绘图像检索精度.

关键词: 手绘图; 图像检索; 可变形卷积; 边缘检测; 神经网络

引用格式: 王文超.基于可变形卷积的手绘图像检索.计算机系统应用,2020,29(7):239–244. http://www.c-s-a.org.cn/1003-3254/7499.html

Sketch-Based Image Retrieval with Deformable Convolution

WANG Wen-Chao

(College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: Sketches contain only simple lines and contours, which have completely different characteristics from natural images with rich colors and details. However, the current neural networks are mostly designed for natural images and cannot adapt to the sparseness of sketches. Aiming at this problem, this study proposes a sketch-based image retrieval method based on deformable convolution. First, the Berkeley edge detection algorithm is used to transform the natural image into edge map to eliminate domain differences. Then replace part of the standard convolution in the convolutional neural networks with deformable convolution, so that the network can fully focus on the outlines of the sketches. Finally, sketches and edge maps are sent to the network separately, and extract the fully connected layer features as feature descriptors for retrieval. Experimental results on the benchmark dataset Flickr15k show that the proposed method can effectively improve the accuracy of sketch-based image retrieval compared with existing methods.

Key words: sketch; image retrieval; deformable convolution; edge detection; neural networks

手绘对于人类而言是非常直观且通用的工具, 是人类的本能, 从原始时期就被用来描述人类所看到的现实世界. 近年来, 随着智能手机、平板电脑、手绘板等移动设备的普及, 手绘图像的获取更加容易, 使得手绘图像检索的研究变得日益繁盛和重要起来. 手绘图像检索是一种通过手绘图像检索自然彩图的图像检索技术, 是以图搜图技术的一种. 与文本标签相比, 手绘图像所包含的信息更加丰富, 表达更加生动形象; 与自然图像相比, 手绘图像钩玄提要, 只保留了物体最基本的骨架轮廓信息, 而且手绘允许人们随心所欲地表达想要描述的物体. 由于手绘检索独特的优势, 目前已在图片检索、在线商城、安防等领域得到了足够的重视

① 基金项目: 中央高校基本科研业务费专项资金 (18CX06048A)
Foundation item: The Fundamental Research Funds for the Central Universities of China (18CX06048A)
收稿时间: 2019-12-16; 修改时间: 2020-01-07; 采用时间: 2020-01-14; csa 在线出版时间: 2020-07-03

与应用. 比如在线商城根据用户通过手绘图像推荐相似商品, 公安人员通过素描画像进行嫌疑犯定位等等. 近期流行的“你画我猜”等应用背后也是手绘识别技术的体现.

1 相关工作

手绘图像检索相关研究可追溯至自上世纪 90 年代, 早期工作如 GF-HOG[1]、HELO[2]、RST-HELO[3]等多通过设计手工特征对线条轮廓进行特征表达, 但由于手绘图像具有抽象性、随意性等特点并未取得良好效果. 随着深度学习的兴起, 基于神经网络的方法开始逐渐奏效. 2014 年, Yu Qian 针对手绘图像稀疏性的特点, 设计了第一个适用于手绘图像的卷积神经网络 Sketch-a-Net[4], 作者利用手绘图像绘制过程中线条的时序顺序, 将手绘图分为 5 种不同层次的表达并据此构建了 5 分支网络, 最后通过贝叶斯融合得到最终特征描述. 作者将 Sketch-a-Net 用于手绘图像分类并取得了良好效果. 2016 年, 新加坡南洋理工大学[5]针对自然图像背静杂乱的特点, 引入目标检测中的 RPN (Region Proposal Network) 网络对于图像中可能存在的目标予以定位, 并将手绘图像与 RPN 网络生成的每个 Region Proposal 进行相似度比较以此实现实例检索. 其网络结构延续了 Sketch-a-Net 的设计思路, 然而此方法由于操作繁琐检索效率稍显不足. 同年, 北京邮电大学提出一种基于孪生网络 (Siamese Networks) 的手绘检索方法[6], 通过选取正负样本对的形式使网络对图像相似性进行建模. 基础网络结构则为模仿 Sketch-a-Net 搭建的小型卷积网络, 然而由于网络深度较浅并未取得良好效果. 2018 年, Bui Tu 在 Siamese Networks 基础上采用了 Triplet Networks 进行手绘图像检索[7], 该方法每次选取三个样本: 一张作为参照的手绘图像, 一张与手绘图像同类别的自然图像作为正样本, 另一张不同类别的作为负样本, 通过 Triplet loss 进行网络训练. 作者通过实验发现网络结构采用 AlexNet[8]、GoogLeNet[9]等在 ImageNet 上训练过的网络比采用 Sketch-a-Net 能够取得更好的效果. 类似的, Huang F 和 Seddati O 等人的工作[10,11]也得出了类似的结论. 文献[10]采用 Alexnet 作为基础网络, 而文献[11]提出的 Quadruplet Networks 则采用了 Resnet-18[12].

通过以上介绍可知, Sketch-a-Net 虽然针对手绘图像特点设计, 但是一方面其 5 分支结构过于复杂不易操作, 另一方面由于网络深度较浅, 特征表达能力不足, 直接将其用于手绘检索并未取得理想效果. 另一方面, 迁移学习在图像分类、目标检测、目标跟踪等诸多计算机视觉任务中均体现出明显优势, 采用在自然图像数据集 ImageNet 上训练得到的 VGG[13]、GoogLeNet 等作为基础网络结构并针对特定任务特定数据进行微调, 可以有效减少网络训练难度并带来性能提升. 因此手绘图像检索近期工作也逐渐倾向于迁移学习而忽略了对手绘图像特点的探索.

手绘图像与自然图像有着截然不同的特性. 自然图像颜色、背景以及纹理细节信息丰富, 而手绘图像仅由简单的线条轮廓组成, 因此, 设计适合手绘图像稀疏性特点的卷积神经网络结构仍然是有必要的. 为此, 本文提出一种基于可变形卷积的手绘检索方法, 打破标准卷积只能在矩形感受野内均匀采样的特点, 通过学习卷积核的位置偏移量使得网络关注到手绘图像轮廓区域, 以获得更加鲁棒的特征表达.

2 基于可变形卷积的手绘图像检索

该部分将从消除域差异、网络结构与训练、特征表达与相似度度量等方面对本文提出的基于可变形卷积的手绘图像检索方法进行详细描述.

2.1 消除域差异

实现手绘图像的跨域检索首先要做的就是消除手绘域与自然图像域之间的域差异. 本文沿用现有工作最常采用的思路, 将自然图像通过边缘检测转化为类手绘图, 即边缘图, 以此减小域间差异实现跨域检索. 与现有工作最常采用的 Canny 边缘检测相比, 本文采用的 Berkerly 边缘检测算法[14]通过训练分类器得到每一个像素属于边缘的概率, 通过设置恰当的阈值, 能够最大程度上保留目标的主体轮廓而消除不必要的细节信息干扰. 图 1 为 Canny 算子与 Berkerly 算法效果对比, 第 1 行为原始图像, 第 2 行为 Canny 边缘检测效果图, 第 3 行为本文采用 Berkerly 边缘检测效果图.

2.2 可变形卷积

卷积操作是图像处理最基础也是最常用的操作之一. 以卷积作为主要操作的卷积神经网络同样在计算机视觉领域大放异彩. 每个卷积核都是一个滤波模板, 通过卷积核与图像在特定邻域内做卷积操作可以检测图像是否具备某些特征. 以 3 × 3 尺寸的卷积核为例, 对于输入图像 x, 卷积操作在中心位置 p_{center} 处的响应

$y(p_{\text{center}})$ 可定义为:

$$y(p_{\text{center}}) = \sum_{p_i \in F} w(p_i) \cdot x(p_{\text{center}} + p_i)$$
$$F = \{(-1,-1),(-1,0),\cdots,(0,1),(1,1)\} \quad (1)$$

其中, F 表示 3 × 3 卷积核定义的感受野位置, w 表示卷积核采样权重.

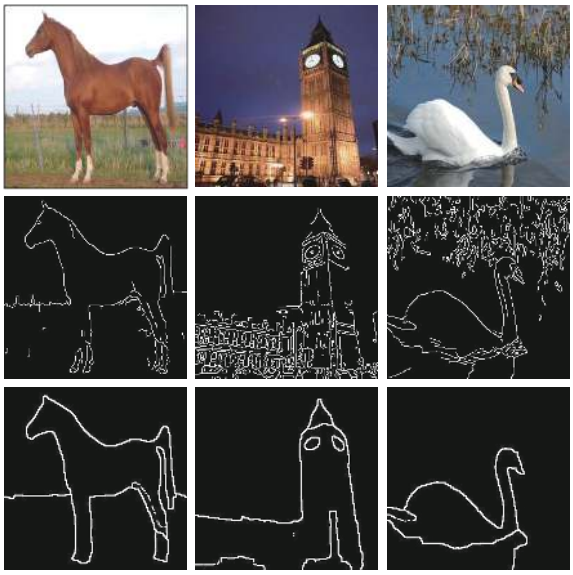

图 1 边检检测效果对比

对于自然图像而言, 由于其细节信息丰富, 任意位置处的像素信息均有可能对图片内容理解做出贡献, 因此在使用标准卷积进行均匀采样的情况下能够取得良好效果. 而使用标准卷积处理手绘图像时, 卷积核感受野内的背景像素几乎对手绘图像识别没有帮助, 因此如果能够打破规格化的矩形感受野, 使得卷积操作能够更偏重于提供有用信息的线条像素而忽略部分背景像素, 将更有助于手绘图像的特征学习. 为此, 本文将可变形卷积引入手绘图像检索.

在可变形卷积中, 卷积核采样位置不再由卷积核尺寸限制, 而是通过学习得到, 即式 (2) 中的 *Offset*. 引入可变形卷积后, 卷积操作可定义为:

$$y(p_{\text{center}}) = \sum w(p_i) \cdot x(p_{\text{center}} + Offset_i) \quad (2)$$

可变形卷积由 Dai Jifeng 首次提出并用于目标检测任务[15], 通过引入可变形卷积可以使网络适应图片中不同尺寸的目标. 文献[15]仅对输入特征图的每一个位置学习一对偏移量, 而各通道之间共享学习到的偏移量. 然而, 不同通道代表的特征一般并不相同, 因此本文充分考虑了通道之间的差异性, 设计了更为灵活多样的可变形卷积操作, 具体如图 2.

如图 2 所示, 给定输入图像或卷积网络中间层的特征图, 记其尺寸大小为 $W \times H \times C$, 其中 W 表示宽度, H 表示高度, C 表示通道数. 本文方法需要对输入特征图每个通道的每个位置学习 x 和 y 两个方向的偏移量, 因此需要学习 $2WHC$ 个参数. 该过程可通过 $2C$ 个 3 × 3 卷积核对输入特征图进行卷积操作实现, 得到的偏移量特征图尺寸为 $W \times H \times 2C$. 值得注意的是, 网络学习到的偏移量 (即采样位置) 并不要求是整数, 而且极有可能是浮点数, 因此采样位置的像素值需要通过对输入特征图进行双线性插值得到.

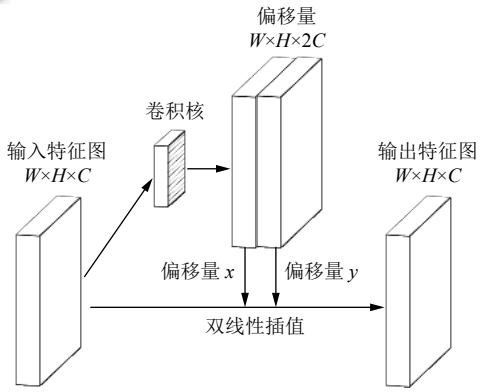

图 2 可变形卷积示意图

本文采用的可变形卷积模块可以替换标准卷积网络中的任意卷积层而不影响网络的整体结构, 具有简单灵活且高效的特点.

2.3 网络结构

如图 3 所示, 本文采用 VGG-16 网络结构作为基准网络结构, VGG 网络由于其优异的效果与可拓展性现已成为图像检索任务最常选用的网络结构之一. 需要说明的是, 本文引入的可变形卷积模块可与任意卷积神经网络结合, 并不依赖于某个具体网络结构. 如图所示, 本文将 VGG-16 网络每个 block 的第一层卷积层由原来的标准卷积替换为可变形卷积, 在实验部分将会对该替换选择进行分析.

2.4 网络训练

由于手绘图像数量较少, 本文采用类手绘图 (边缘图) 进行网络训练. 将数据集按照 1:1 的比例划分训练集与测试集, 以 ImageNet 上预训练权重为初始化网络参数, 通过类别交叉熵损失进行迁移学习. 一方面, 与

自然图像相比边缘图的数量仍然较少, 通过迁移学习而不是从头训练会取得更好的效果; 另一方面, 本文引入的可变形卷积模块在初始状态时偏置为零, 并不会改变网络状态, 因此可以通过预训练权重进行微调.

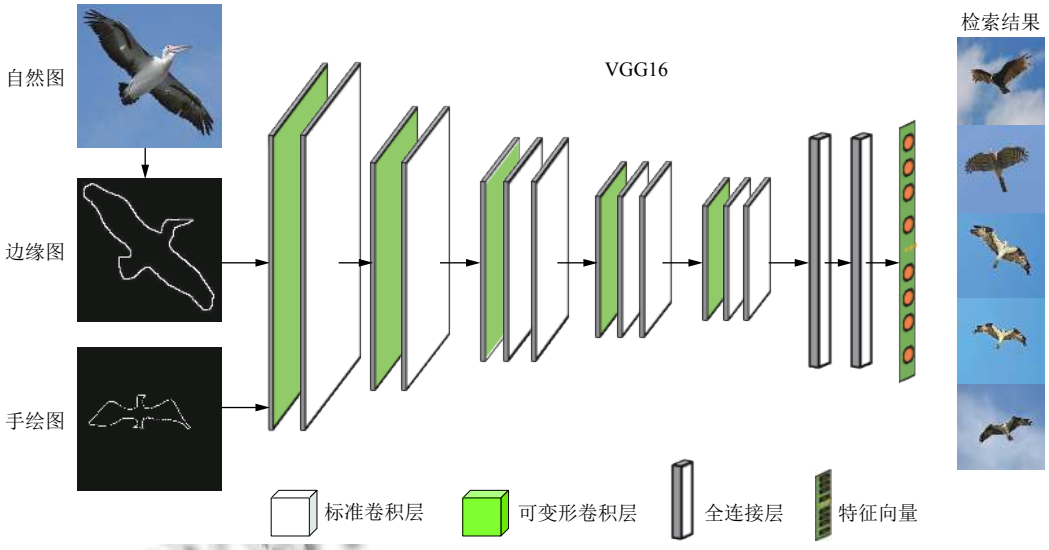

图 3 整体网络框架图

2.5 特征表达与相似度度量

与卷积层相比, 全连接层往往包含更多的语义信息, 因此本文提取加入可变形卷积的 VGG-16 网络的第二个全连接层特征作为输入图像的特征向量. 将自然图像通过边缘检测算法转化为边缘图后, 手绘图像与边缘图可以共享一个网络进行特征提取.

本文采用欧式距离衡量手绘图像特征 s_i 与自然图像特征 n_i 之间的相似度. 基于特征向量距离的远近, 相似度公式定义为:

$$S(s_i,n_i)=1-\frac{d(s_i,n_i)}{\sum_{j=1}^{k}d(s_i,n_i)} \tag{3}$$

其中, $d(,)$ 表示两个向量的欧式距离, k 表示检索结果总数, $S(,)$ 越大表示图像相似度越高.

2.6 检索算法流程

本文基于可变形卷积的手绘检索算法描述如算法 1.

算法 1. 基于可变形卷积的手绘图像检索算法

1) 通过 Berkerly 边缘检测算法将数据库中的自然图像转化为二值化边缘图
2) 将所有边缘图输入基于可变形卷积的 VGG 网络并提取全连接层特征作为特征描述子
3) 将给定的手绘图像输入基于可变形卷积的 VGG 网络并提取全连接层特征作为特征描述子
4) 通过欧氏距离进行特征相似度度量
5) 返回检索结果

3 实验分析

3.1 基准数据集

本文选用手绘图像检索常用的数据集 Flickr15k 为基准数据集进行实验验证与对比. Flickr15k 同时包含手绘图像和与之对应的自然彩图. 其中手绘图像 329 幅, 分别属于 33 个类别, 由 10 名非专业手绘创作者绘制而成. 自然图像共 14 460 幅, 分属 60 个类别.

3.2 评价标准

本文采用图像检索任务最常用的 mAP (mean Average Precision) 指标作为主要评价标准, mAP 值越高代表检索效果越好.

3.3 实验环境

本文所有实验均在以下环境配置中进行: Intel Xeon CPU E5 处理器, 一块 GeForce GTX Titan X 显卡, 以 TensorFlow 为后端的 Keras 深度学习框架.

3.4 可变形卷积替换选择

为验证可变形卷积对检索精度的影响, 本文通过将原始 VGG-16 网络不同卷积层由原来的标准卷积替换为可变形卷积并进行多次对比试验, 实验结果记录如表 1 所示. 其中 BxCy 代表第 x 个 block 中的第 y 个卷积层. 对勾表示该层采用可变形卷积.

实验表明, 将 VGG-16 中每个 block 的第一个卷积层替换为可变形卷积, 或者将第二个卷积层替换为可变形卷积, 均会对检索精度带来不同程度的提升, 说

明本文引入的可变形卷积是有效的. 但是将每个 block 前两层均替换为可变形卷积时, 效果反而不如仅替换一层, 原因可能是连续堆叠多层可变形卷积使得引入的偏移量参数相互影响, 不易优化所致. 因此本文在 VGG-16 每个 block 的第一个卷积层使用可变形卷积.

表 1 可变形卷积替换选择

mAP	B1C1	B1C2	B2C1	B2C2	B3C1	B3C2	B3C3	B4C1	B4C2	B4C3	B5C1	B5C2	B5C3
45.2													
49.5	√		√		√			√			√		
48.4		√		√		√			√			√	
47.2	√	√	√	√	√	√		√	√		√	√	

3.5 实验对比

本文选取经典手工特征描述子 HOG、GF-HOG、RST-HELO 等方法, 以及采用深度特征的 Siamese CNN、Triplet CNN、Quadruplet CNN 等方法与本文提出的算法进行对比, 不同方法在 Flickr15k 上的 mAP 表 2 所示.

通过表 2 结果可知, 本文提出的基于可变形卷积的手绘检索方法是有效的. 其得到的特征描述子平均检索精度远超手工特征描述子 30%~40%, 而且与同为深度特征的其他方法相比平均检索精度也能有较大幅的提升.

图 4 为采用本文基于可变形卷积的手绘检索方法对于 Flickr15k 数据集的部分检索结果. 左侧为输入的手绘图像, 右侧为从自然图像数据库中检索到的相似度 top-8 排名的图像. 可以看出本文方法能够取得较为理想的检索效果.

表 2 Flickr15k 上各方法 mAP 对比

分类	方法	mAP
手工特征	HOG[16]	10.93
	GF-HOG[1]	12.22
	HELO[2]	12.36
	RST-HELO[3]	20.22
	PerceptualEdge[17]	15.13
深度特征	3Dshape[18]	18.31
	Siamese CNN[6]	19.54
	Triplet CNN[7]	24.45
	Quadruplet Network[11]	32.16
	Query-adaptive CNN[10]	32.3
	本文	49.5

图 4 本文方法检索结果

4 总结

本文分析了针对手绘图像稀疏性等特点设计神经网络结构的必要性, 提出基于可变形卷积的手绘检索方法, 使得卷积神经网络更加关注手绘图像轮廓信息以获取更鲁棒的特征表达. 本文以 VGG-16 为基础网络, 通过在 Flickr15k 数据集上的实验验证了引入可变形卷积的效果增益, 并对可变形卷积的添加位置进行了讨论. 另外, 本文方法具有良好的拓展性和迁移性, 可变形卷积模块可以添加到任意卷积神经网络, 该手绘检索流程也同样适用于其他手绘图像数据库的检索.

参考文献

1 Hu R, Collomosse J. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. Computer Vision and Image Understanding, 2013, 117(7): 790–806. [doi: 10.1016/j.cviu.2013.02.005]

2 Saavedra JM, Bustos B. An improved histogram of edge local orientations for sketch-based image retrieval. Proceedings of the 32nd DAGM Symposium Joint Pattern Recognition Symposium. Darmstadt, Germany. 2010. 432–441.

3 Saavedra JM. RST-SHELO: Sketch-based image retrieval using sketch tokens and square root normalization. Multimedia Tools and Applications, 2017, 76(1): 931–951. [doi: 10.1007/s11042-015-3076-5]

4 Yu Q, Yang YX, Song YZ, *et al*. Sketch-a-net that beats humans.Proceedings of the British Machine Vision Conference. 2015. 7–10.

5 Bhattacharjee SD, Yuan JS, Hong WX, *et al*. Query adaptive instance search using object sketches. Proceedings of the 24th ACM international conference on Multimedia. Amsterdam, the Netherlands. 2016. 1306–1315.

6 Qi YG, Song YZ, Zhang HG, *et al*. Sketch-based image retrieval via siamese convolutional neural network. Proceedings of 2016 IEEE International Conference on Image Processing. Phoenix, AZ, USA. 2016. 2460–2464.

7 Bui T, Ribeiro L, Ponti M, *et al*. Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression. Computers & Graphics, 2018, 71: 77–87.

8 Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, CA, USA. 2012. 1097–1105.

9 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 1–9.

10 Huang F, Jin C, Zhang YJ, *et al*. Sketch-based image retrieval with deep visual semantic descriptor. Pattern Recognition, 2018, 76: 537–548. [doi: 10.1016/j.patcog.2017.11.032]

11 Seddati O, Dupont S, Mahmoudi S. Quadruplet networks for sketch-based image retrieval. Proceedings of 2017 ACM on International Conference on Multimedia Retrieval. Bucharest, Romania. 2017. 184–191.

12 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–778.

13 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations. 2015. 1–14.

14 Arbeláez P, Maire M, Fowlkes C, *et al*. Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5): 898–916. [doi: 10.1109/TPAMI.2010.161]

15 Dai JF, Qi HZ, Xiong YW, *et al*. Deformable convolutional networks. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy. 2017. 764–773.

16 Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA. 2005. 886–893.

17 Qi YG, Song YZ, Xiang T, *et al*. Making better use of edges via perceptual grouping. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 1856–1865.

18 Wang F, Kang L, Li Y. Sketch-based 3d shape retrieval using convolutional neural networks. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 1875–1883.