(19)中华人民共和国国家知识产权局

(12)实用新型专利

(10)授权公告号 CN 210428520 U
(45)授权公告日 2020.04.28

(21)申请号 201920242515.0
(22)申请日 2019.02.26
(30)优先权数据
62/636,018 2018.02.27 US
16/280,963 2019.02.20 US
(73)专利权人 意法半导体国际有限公司
地址 荷兰阿姆斯特丹
专利权人 意法半导体股份有限公司
(72)发明人 S•P•辛格 T•勃伊施 G•德索利
(74)专利代理机构 北京市金杜律师事务所 11256
代理人 王茂华
(51)Int.Cl.
G06N 3/063(2006.01)
G06N 3/08(2006.01)
(ESM)同样的发明创造已同日申请发明专利

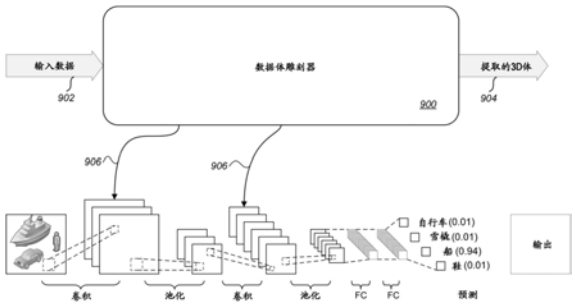
权利要求书2页 说明书35页 附图23页

(54)实用新型名称

用于深度学习加速的集成电路

(57)摘要

本公开涉及用于深度学习加速的集成电路。实施例包括板载存储器、应用处理器、数字信号处理器(DSP)集群、可配置的加速器框架(CAF)和至少一个通信总线架构。通信总线将应用处理器、DSP集群和CAF通信地耦合到板载存储器。CAF包括可重配置的流交换器和数据体雕刻单元,其具有耦合到可重配置的流交换器的输入和输出。数据体雕刻单元具有计数器、比较器和控制器。数据体雕刻单元被布置为接收形成三维(3D)特征图的特征图数据的流。3D特征图被形成为多个二维(2D)数据平面。数据体雕刻单元还被布置为标识3D特征图内的3D体、并且从3D特征图隔离在3D体内的数据以用于在深度学习算法中进行处理。实施例能够高效地标识和隔离三维特征图内的三维体。



CN 210428520 U

1.一种用于深度学习加速的集成电路,其特征在于,包括:

板载存储器;

应用处理器;

数字信号处理器集群;

可配置的加速器框架;以及

至少一个通信总线架构,将所述应用处理器、所述数字信号处理器集群和所述可配置的加速器框架通信地耦合到所述板载存储器,其中所述可配置的加速器框架包括:

可重配置的流交换器;以及

数据体雕刻单元,所述数据体雕刻单元具有耦合到所述可重配置的流交换器的至少一个输入和耦合到所述可重配置的流交换器的输出,所述数据体雕刻单元具有计数器、比较器和控制器,所述数据体雕刻单元被布置为:

经由所述至少一个输入接收特征图数据的流,所述特征图数据的流形成三维特征图,所述三维特征图被形成为多个二维数据平面;

标识所述三维特征图内的三维体,所述三维体在尺寸上小于所述三维特征图;

从所述三维特征图隔离在所述三维体内的数据,以用于在深度学习算法中进行处理;以及

经由所述输出提供隔离的数据。

2.根据权利要求1所述的集成电路,其特征在于,所述数据体雕刻单元还被布置为:

经由所述至少一个输入接收限定第一二维数据平面中的感兴趣区域的输入信息,所述输入信息包括所述感兴趣区域的至少一个第一坐标和足以形成所述第一二维数据平面中的封闭二维体的另外信息;

加载并且按顺序排好所述计数器,使得所述第一二维数据平面中的每个数据以选择的顺序被分析;以及

使用所述比较器确定被分析的每个数据是否在所述封闭二维体内,其中提供所述隔离的数据输出包括提供被确定为在所述封闭二维体内的每个数据。

3.根据权利要求1所述的集成电路,其特征在于,所述数据体雕刻单元还被布置为:

经由所述至少一个输入接收限定第一二维数据平面中的感兴趣区域的输入信息,所述输入信息包括所述感兴趣区域的至少一个第一坐标和足以形成所述第一二维数据平面中的封闭二维体的另外信息;

加载并且按顺序排好所述计数器,使得所述第一二维数据平面中的每个数据以选择的顺序被分析;

使用所述比较器确定被分析的每个数据是否在所述封闭二维体内;以及

生成包括被确定为在所述封闭二维体内的每个数据的有序数据结构。

4.根据权利要求3所述的集成电路,其特征在于,所述数据体雕刻单元还被布置为:

将在所述三维特征图的多个封闭二维体内的数据包括在所述有序数据结构中,其中所述多个二维数据平面中的每个二维数据平面具有相应的封闭二维体,并且其中每个相应的封闭二维体与在相邻二维数据平面中被限定的至少一个其他封闭二维体相关联。

5.根据权利要求3所述的集成电路,其特征在于,所述多个二维数据平面中的每个二维数据平面具有限定于其中的多个封闭二维体。

6.根据权利要求5所述的集成电路，其特征在于，在所选择的二维数据平面上的所述多个封闭二维体中的个体封闭二维体是非重叠的。

7.根据权利要求1所述的集成电路，其中所述集成电路被形成为片上系统。

用于深度学习加速的集成电路

技术领域

[0001] 本公开总体上涉及改善例如卷积神经网络(CNN)中的深度机器学习系统的灵活性、数据局部性和更快执行。更具体地但非排他性地,本公开涉及用于深度学习加速引擎的数据体(data volume)雕刻器(sculptor)。

背景技术

[0002] 已知的计算机视觉、语音识别和信号处理应用受益于对学习机器的使用。本公开中所讨论的学习机器可以落入机器学习、人工智能、神经网络、概率推理引擎、加速器等的技术题目下。这样的机器被布置为快速地执行数百、数千以及数百万并发操作。常规学习机器可以递送数百万亿次浮点运算(即,每秒一万亿(10^{12})次浮点运算)的计算能力。

[0003] 在一些情况下,学习机器被组织为深度卷积神经网络(DCNN)。DCNN领域中的开创性著作是Y.LeCun等人的“Gradient-Based Learning Applied To Document Recognition”(Proceedings of the IEEE,vol.86,no.11,第2278-2324页,1998),其最终利用“AlexNet”赢得2012ImageNet大规模视觉识别挑战。如在Krizhevsky,A.Sutskever,I.和Hinton,G.的“ImageNet Classification With Deep Convolutional Neural Networks”(NIPS,第1-9页,Lake Tahoe,NV(2012))中描述的AlexNet是第一次表现明显好于经典方法的DCNN。

[0004] DCNN是处理大量数据并且通过合并数据内的最接近地相关的特征、做出关于数据的宽泛预测、并且基于可靠的结论和新的合并完善预测来自适应地“学习”的基于计算机的工具。DCNN被布置在多个“层”中并且在每个层处做出不同类型的预测。

[0005] 例如,如果面部的多幅二维图片作为输入被提供给DCNN,则DCNN将学习各种面部特性,诸如边缘、曲线、角度、点、颜色对比度、亮点、暗点等。这些一个或多个特征在DCNN的一个或多个第一层处被学习。然后,在一个或多个第二层中,DCNN将学习各种可识别的面部特征,诸如眼睛、眉毛、前额、头发、鼻子、嘴巴、脸颊等;其中的每一个可与所有其他特征区分开。即,DCNN学习识别眼睛并将眼睛与眉毛或任何其他面部特征区分开。在一个或多个第三和之后的后续层中,DCNN学习整个面部和更高阶特性,诸如种族、性别、年龄、情绪状态等。DCNN甚至在一些情况被教导识别个人的特定身份。例如,随机图像可以被识别为面部,并且面部可以被识别为奥兰多·布鲁姆、安德烈·波伽利或某个其他身份。

[0006] 在其他的一些示例中,DCNN可以被提供有动物的多幅图片,并且DCNN可以被教导标识狮子、老虎和熊;DCNN可以被提供有汽车的多幅图片,并且DCNN可以被教导标识并区分不同类型的车辆;并且许多其他DCNN也可以被形成。DCNN可以用于学习句子中的词语模式、识别音乐、分析个体购物模式、打视频游戏、创建交通路线,并且DCNN也可以用于许多其他基于学习的任务。

[0007] 图1A-图1J可以在本文中被统称为图1。

[0008] 图1A是卷积神经网络(CNN)系统10的简化图示。在CNN系统中,像素的二维阵列由CNN处理。CNN分析10×10输入对象平面以确定“1”是否被表示在该平面中,“0”是否被表示

在该平面中，或者“1”或“0”是否都没有被实施在该平面中。

[0009] 在10×10输入对象平面中，每个像素是被照亮的或未被照亮的。为图示的简单起见，被照亮的像素被填充(例如，暗色)并且未被照亮的像素不被填充(例如，亮色)。

[0010] 图1B图示了图1A的CNN系统10确定第一像素图案图示“1”并且第二像素图案图示“0”。然而，在现实世界中，图像并不总是如图1B中图示的那样整洁地对齐。

[0011] 在图1C中，示出了不同形式的一和零的若干变型。在这些图像中，普通人类观察者将容易意识到具体数字被平移或缩放，但是观察者还将正确地确定图像表示“1”还是“0”。按照这些原则，不用思考，人类观察者展望图像旋转、数字的各种加权、数字的大小调整、移位、倒转、重叠、破碎、相同图像中的多个数字、以及其他这样的特性。然而，以编程方式，在传统计算系统中，这样的分析非常困难。各种图像匹配技术是已知的，但是甚至关于非常小的图像大小，这种类型的分析也快速地压倒可用的计算资源。然而，对比之下，CNN系统10可以以可接受的准确度正确地识别每幅处理的图像中的一、零、一和零两者、或者既没有一也没有零，即使CNN系统10先前从未“看到”过确切的图像。

[0012] 图1D表示分析(例如，数学上组合)未知图像的部分以及已知图像的对应部分的CNN操作。例如，左侧未知图像的3像素部分B5-C6-D7可以被识别为匹配右侧已知图像的对应3像素部分C7-D8-E9。在这些和其他情况下，各种其他对应的像素布置也可以被识别。一些其他对应关系被图示在表1中。

[0013] 表1-对应的已知图像分段和未知图像分段

[0014]

图 1D 左侧未知图像	图 1D 右侧已知图像
C3-B4-B5	D3-C4-C5
C6-D7-E7-F7-G6	D8-E9-F9-G9-H8
E1-F2	G2-H3
G2-H3-H4-H5	H3-I4-I5-I6

[0015] 在识别到已知图像的分段或部分可以被匹配到未知图像的对应分段或部分的情况下，进一步识别到，通过统一部分匹配操作，整个图像可以以完全相同的方式被处理同时得到先前未计算出的结果。换句话说，特定部分的大小可以被选择，并且已知图像可以然后逐部分被分析。当已知图像的任何给定部分内的图案在数学上与未知图像的类似大小的部分组合时，生成表示这些部分之间的相似度的信息。

[0016] 图1E图示了图1D的右侧已知图像的六个部分。每个部分(也称为“核”)被布置为3个像素乘3个像素的阵列。在计算上，被照亮的像素在数学上被表示为正“1”(即，+1)；并且未被照亮的像素在数学上被表示为负“1”(即，-1)。为简化图1E中的图示起见，每个图示的核也被示出具有图1D的列和行引用。

[0017] 图1E中示出的六个核是代表性的并且被选择以便于理解CNN系统10的操作。清楚的是，已知图像可以利用重叠或非重叠核的有限集合来表示。例如，考虑3个像素乘3个像素的核大小以及具有一(1)步幅的重叠核的系统，每个10×10像素图像可以具有64个对应的核。

[0018] 第一核跨越列A、B、C和行1、2、3中的9个像素。

[0019] 第二核跨越列B、C、D和行1、2、3中的9个像素。

[0020] 第三核跨越列C、D、E和行1、2、3中的9个像素。

[0021] 核的这种图案继续直到第八核跨越列H、I、J和行1、2、3中的9个像素。

[0022] 核对齐以这种方式继续直到第57核跨越列A、B、C和行8、9、10，并且以此类推直到第64核跨越列H、I、J和行8、9、10。

[0023] 在其他的一些CNN系统中，核可以是重叠的或非重叠的，并且核可以具有2、3、或某个其他数目的步幅。用于选择核大小、步幅、位置等的不同策略由CNN系统设计者基于过去的结果、分析研究或者以某种其他方式来选择。

[0024] 返回到图1D和图1E的示例，总共64个核使用已知图像中的信息来形成。第一核以最上面、最左边的3×3阵列中的9个像素开始。接下来的七个核各自被顺序地向右移位一列。第九核返回到头三列并且向下下降到第二行，类似于基于文本的文档的回车操作，其概念是从二十世纪的手动打字机得到的。遵循这种图案，图1E示出了图1D(b)中的10×10图像的第7、第18、第24、第32、第60和第62核。

[0025] 顺序地，或者以某种其他已知图案，已知图像的每个核与处于分析中的图像的对应大小的像素集合对齐。在完全分析的系统中，例如，已知图像的第一核在概念上在核位置中的每个核位置中叠加在未知图像上。考虑图1D和图1E，第一核在概念上在第1号核的位置(图像的最左边、最上面的部分)中叠加在未知图像上，然后第一核在概念上在第2号核的位置中叠加在未知图像上，以此类推，直到第一核在概念上在第64号核的位置(图像的最下面、最右边的部分)中叠加在未知图像上。针对64个核中的每个核重复该流程，并且执行总共4096个操作(即，64个核在64个位置中的每个位置中)。以这种方式，还示出了当其他CNN系统选择概念叠加的不同核大小、不同步幅、以及不同图案时，则操作的数目将改变。

[0026] 在CNN系统10中，每个核在分析中的未知图像的每个部分上的概念叠加被执行为被称为卷积的数学过程。核中的九个像素中的每个像素基于该像素是被照亮还是未被照亮而被给定正“1”(+1)或负“1”(-1)的值，并且当核被叠加在分析中的图像的部分上时，核中的每个像素的值乘以图像中的对应像素的值。由于每个像素具有值+1(即，被照亮)或-1(即，未被照亮)，则乘法将总是得到+1或-1。附加地，由于4096个核操作中的每个核操作使用9像素核来处理，所以在非常简单的CNN中在单个未知图像分析的第一阶段执行总共36,864个数学操作(即，9×4096)。很清楚，即使简单的CNN系统也要求巨大的计算资源，并且针对更复杂的CNN系统的计算要求以指数方式增长。

[0027] 如刚刚描述的，核中的9个像素中的每个像素乘以分析中的图像中的对应像素。核中的未被照亮像素(-1)当乘以主体未知图像中的未被照亮像素(-1)时将得到指示在该像素位置处的“匹配”的+1(即，核和图像两者均具有未被照亮像素)。类似地，核中的被照亮像素(+1)乘以未知图像中的被照亮像素(+1)时也得到匹配(+1)。另一方面，当核中的未被照亮像素(-1)乘以图像中的被照亮像素(+1)时，结果指示在该像素位置处的不匹配(-1)。并且当核中的被照亮像素(+1)乘以图像中的未被照亮像素(-1)时，结果也指示在该像素位置处的不匹配(-1)。

[0028] 在执行了单个核的九个乘法运算之后，乘积结果将包括九个值；九个值中的每一个是正一(+1)或负一(-1)。如果核中的每个像素与未知图像的对应部分中的每个像素相匹配，则乘积结果将包括九个正一(+1)值。备选地，如果核中的一个或多个像素与分析中的未

知图像的部分中的对应像素不匹配，则乘积结果将具有至少一些负一(-1)值。如果核中的每个像素未能与分析中的未知图像的对应部分中的对应像素相匹配，则乘积结果将包括九个负一(-1)值。

[0029] 考虑像素的数学组合(即，乘法运算)，应意识到乘积结果中的正一(+1)值的数目和负一(-1)值的数目表示核中的特征与图像中该核在概念上被叠加的部分匹配的程度。因此，通过对所有乘积求和(例如，对九个值求和)并且除以像素的数目(例如，九)，单个“品质值”被确定。品质值表示核与分析中的未知图像的部分之间的匹配的程度。品质值的范围可以从没有核像素相匹配时的负一(-1)到核中的每个像素具有与未知图像中的其对应像素相同的被照亮/未被照亮状态时的正一(+1)。

[0030] 本文中参考图1E描述的动作还可以被统称为被称为“滤波”的操作中的第一卷积过程。在滤波操作中，在未知图像中搜索已知图像中的特定感兴趣部分。滤波的目的是用可能性的对应预测来标识在未知图像中是否找到感兴趣特征以及在未知图像中哪里找到感兴趣特征。

[0031] 图1F图示了滤波过程中的卷积的十二个动作。图1G示出了图1F的十二个卷积动作的结果。在每个动作中，未知图像的不同部分利用选择的核来处理。选择的核可以被识别为图1B的代表性数字一(“1”)中的第十二个核。代表性“1”在图1B中被形成为10个像素乘10个像素的图像中的一组被照亮像素。从最上面、最左边的角落开始，第一核覆盖3个像素乘3个像素的部分。第二至第八核顺序地向右移动一列。以回车的方式，第九核在第二行最左边的列中开始。核10-16对于每个核顺序地向右移动一列。核17-64可以被类似地形成，使得图1B中的数字“1”的每个特征被表示在至少一个核中。

[0032] 在图1F(a)中，3个像素乘3个像素的选择的核在概念上被叠加在未知图像的最左边、最上面的区段上。在这种情况下选择的核是图1B的数字“1”的第十二核。图1F(a)中的未知图像对人类观察者而言看起来像是移位的形状不佳的数字一(即，“1”)。在卷积过程中，选择的核中的每个像素的值(其对于被照亮像素而言为“+1”并且对于未被照亮像素而言为“-1”)乘以未知图像中的每个对应像素。在图1F(a)中，五个核像素被照亮，并且四个核像素未被照亮。未知图像中的每个像素未被照亮。相应地，当所有九个乘法被执行时，五个乘积被计算为“-1”，并且四个乘积被计算为“+1”。九个乘积被求和，并且得到的值“-1”除以九。出于这个原因，图1G(a)的对应图像示出了对于未知图像的最左边、最上面的区段中的核得到的核值“-0.11”。

[0033] 在图1F(b)、1F(c)和1F(d)中，核像素跨图像的列被顺序地向右移动。由于头六列和跨越头六列的头三行的区域中的每个像素也未被照亮，所以图1G(b)、1G(c)和1G(d)均示出了计算的核值“-0.11”。

[0034] 图1F(e)和1G(e)示出了与早前计算的核值“-0.11”不同的计算的核值。在图1F(e)中，被照亮核像素中的一个匹配未知图像中的被照亮像素中的一个。该匹配通过图1F(e)中的变暗像素示出。由于图1F(e)现在具有匹配的/不匹配的特性的不同集合，并且另外，由于核像素中的另一个匹配未知图像中的对应像素，所以预料得到的核值将增大。的确，如图1G(e)中所示，当执行了九个乘法运算时，核中的四个未被照亮像素匹配未知图像中的四个未被照亮像素，核中的一个被照亮像素匹配未知图像中的一个被照亮像素，并且核中的四个其他被照亮像素不匹配未知图像中的四个未被照亮像素。当九个乘积被求和时，结果“+1”

除以九以得到第五个核位置中的计算的核值“+0.11”。

[0035] 当核在图1F(f)中被进一步向右移动时，被照亮核像素中的不同的一个匹配未知图像中的对应的被照亮像素。图1G(f)将该组匹配的和不匹配的像素表示为核值“+0.11”。

[0036] 在图1F(g)中，核又被向右移动一列，并且该位置中，核中的每个像素匹配未知图像中的每个像素。由于执行了九个乘法，所以当核的每个像素乘以未知图像中的其对应像素时得到“+1.0”，九个乘积的和被计算为“+9.0”，并且针对该特定位置的最终核值被计算为(“即，9.0/9”)+1.0，其表示完美匹配。

[0037] 在图1F(h)中，核被再次向右移动，其得到如图1G(h)中图示的单个被照亮像素匹配、四个未被照亮像素匹配、以及核值“+0.11”。

[0038] 核继续如图1F(i)、1F(j)、1F(k)和1F(l)中示出的那样被移动，并且在每个位置中，核值在数学上被计算。由于在图1F(i)至图1F(l)中没有核的被照亮像素被叠加在未知图像的被照亮像素上，所以针对这些位置中的每个位置的计算的核值为“-0.11”。核值在图1G(i)、1G(j)、1G(k)和1G(l)中在相应的四个核位置中被示出为“-0.11”。

[0039] 图1H图示了核值的图(map)的堆叠。当图1B中的数字“1”的第十二核被移动到未知图像的每个位置中时，图1H中的最上面的核图被形成。第十二核将被识别为在图1F(a)至图1F(l)和图1G(a)至图1G(l)中的每一个中使用的核。对于选择的核在概念上被叠加在未知图像上的每个位置，核值被计算，并且核值被存储在核图上的其相应位置中。

[0040] 另外在图1H中，其他滤波器(即，核)也被应用到未知图像。为了讨论的简单，图1B中的数字“1”的第29核被选择，并且图1B中的数字“1”的第61核被选择。对于每个核，创建独特的核图。多个创建的核图可以被设想为具有与被应用的滤波器(即，核)的数目相等的深度的核图的堆叠。核图的堆叠也可以被称为滤波的图像的堆叠。

[0041] 在CNN系统10的卷积过程中，单个未知图像被卷积以创建滤波的图像的堆叠。堆叠的深度与被应用到未知图像的滤波器(即，核)的数目相同或者以其他方式基于被应用到未知图像的滤波器(即，核)的数目。其中对图像应用滤波器的卷积过程也被称为“层”，因为它们可以被堆叠在一起。

[0042] 如图1H中显而易见的，在卷积分层过程期间生成大量数据。另外，每个核图(即，每个滤波的图像)具有与其原始图像几乎一样多的值。在图1H中呈现的示例中，原始未知输入图像由100个像素(10×10)形成，并且所生成的滤波图具有64个值(8×8)。仅仅核图的大小的简单减小被实现，因为所应用的9像素核值(3×3)不能完全处理在图像的边缘处的最外面的像素。

[0043] 图1I示出了显著减少由卷积过程产生的数据量的池化特征。池化过程可以对滤波的图像中的一幅、一些或全部执行。图1I中的核图被识别为图1H的最上面的滤波图，其利用图1B中的数字“1”的第12核被形成。

[0044] 池化过程引入“窗口大小”和“步幅”的概念。窗口大小是窗口的尺寸，使得窗口内的单个最大值将在池化过程中被选择。窗口可以被形成为具有m个像素乘n个像素的尺寸，其中“m”和“n”是整数，但是在大多数情况下，“m”和“n”是相等的。在图1I中示出的池化操作中，每个窗口被形成为2个像素乘2个像素的窗口。在池化操作中，4像素窗口在概念上被叠加到核图的选择部分上，并且在窗口内，最高值被选择。

[0045] 在池化操作中，以类似于将核在概念上叠加在未知图像上的方式，池化窗口在概

念上被叠加到核图的每个部分上。“步幅”表示池化窗口在每个池化动作之后被移动多少。如果步幅被设置为“二”,则池化窗口在每个池化动作之后被移动两个像素。如果步幅被设置为“三”,则池化窗口在每个池化动作之后被移动三个像素。

[0046] 在图1I的池化操作中,池化窗口大小被设置为2×2,并且步幅也被设置为二。第一池化操作通过选择核图的最上面、最左边的角落中的四个像素来执行。由于窗口中的每个核值已经被计算为“-0.11”,所以来自池化计算的值也为“-0.11”。值“-0.11”被放置在图1I中的池化输出图中的最上面、最左边的角落中。

[0047] 池化窗口然后被向右移动两个像素的选择步幅,并且第二池化操作被执行。再次地,由于第二池化窗口中的每个核值被计算为“-0.11”,所以来自池化计算的值也为“-0.11”。值“-0.11”被放置在图1I中的池化输出图中的最上面行的第二条目中。

[0048] 池化窗口然后被向右移动两个像素的步幅,并且窗口中的四个值被评估。第三池化动作中的四个值为“+0.11”、“+0.11”、“+0.11”和“+0.33”。这里,在该组四个核值中,“+0.33”是最高值。因此,值“+0.33”被放置在图1I中的池化输出图中的最上面行的第三条目中。池化操作不关心在窗口中哪里找到最高值,池化操作简单地选择落入窗口的边界内的最高(即,最大)值。

[0049] 剩余的13个池化操作也以类似的方式被执行,以便填充图1I中的池化输出图的剩余部分。类似的池化操作也可以针对其他生成的核图(即,滤波的图像)中的一些或全部执行。进一步考虑图1I的池化输出,并且进一步考虑选择的核(即,图1B中的数字“1”的第十二核)和未知图像,应意识到在池化输出的右上角找到最高值。这是因为当核特征被应用到未知图像时,所选择的感兴趣特征的像素(即,核)与未知图像中的类似布置的像素之间的最高相关也是在右上角找到的。还应意识到池化输出具有在其中捕获的、松散地表示未池化的较大大小的核图中的值的值。如果未知图像中的特定图案被搜索,则图案的大致位置可以从池化输出图中学习。即使特征的实际位置不是肯定已知的,观察者也可以意识到在池化输出中检测到特征。实际特征可以在未知图像中被向左或向右移动一点,或者实际特征可以被旋转或以其他方式与核特征不相同,但是,特征的出现及其大体位置可以被识别。

[0050] 还在图1I中图示了可选的归一化(normalization)操作。归一化操作通常由修正线性单元(ReLU)执行。ReLU标识池化输出图中的每个负数并且在归一化的输出图中用值零(即,“0”)来代替负数。由一个或多个ReLU电路进行的可选的归一化过程帮助减少原本可能由关于负数执行的计算所要求的计算资源工作负荷。

[0051] 在ReLU层中的处理之后,归一化的输出图中的数据可以被平均以便预测在未知的图像中是否找到由核表征的感兴趣特征。以这种方式,归一化的输出图中的每个值被用作指示在图像中是否存在该特征的加权的“投票”。在一些情况下,若干特征(即,核)被卷积,并且预测被进一步组合以更宽泛地表征图像。例如,如图1H中图示的,从数字“1”的已知图像得到的三个感兴趣核与未知图像进行卷积。在通过各个层处理每个核之后,做出关于未知图像是否包括示出数字“1”的一个或多个像素图案的预测。

[0052] 总结图1A-图1I,从已知图像选择核。不是已知图像的每个核都需要被CNN使用。相反,被确定为“重要”特征的核可以被选择。在卷积过程产生核图(即,特征图)之后,核图被传递通过池化层、以及归一化(即,ReLU)层。输出图中的所有值被平均(即,求和和相除),并且来自求平均的输出值被用作未知图像是否包含在已知图像中找到的特定特征的预测。在

示例性情况下，输出值被用于预测未知图像是否包含数字“1”。在一些情况下，“投票的列表”还可以被用作到后续堆叠的层的输入。这种处理方式强烈地加强标识的特征并减少微弱标识(或未标识)的特征的影响。考虑整个CNN，二维图像被输入到CNN并且产生一组投票作为其输出。在输出处的该组投票被用于预测输入图像是否包含由特征表征的感兴趣对象。

[0053] 图1A的CNN系统10可以被实现为一系列操作层。一个或多个卷积层可以跟随有一个或多个池化层，并且一个或多个池化层可以可选地跟随有一个或多个归一化层。卷积层从单个未知图像创建多个核图，其原本被称为滤波的图像。多个滤波的图像中的大量数据利用一个或多个池化层而被减少，并且数据量进一步由通过移除所有负数来归一化数据的一个或多个ReLU层减少。

[0054] 图1J更详细地示出了图1A的CNN系统10。在图1J(a)中，CNN系统10将10个像素乘10个像素的输入图像接受到CNN中。CNN包括卷积层、池化层、修正线性单元(ReLU)层以及投票层。一个或多个核值与未知的10×10图像配合被卷积，并且来自卷积层的输出被传递到池化层。一个或多个最大池化操作对由卷积层提供的每个核图执行。来自池化层的池化输出图被用作到产生归一化的输出图的ReLU层的输入，并且包含于归一化的输出图中的数据被求和并且相除以确定关于输入图像是否包括数字“1”或数字“0”的预测。

[0055] 在图1J(b)中，图示了另一CNN系统10a。CNN系统10a中的CNN包括多个层，其可以包括卷积层、池化层、归一化层、以及投票层。来自一层的输出被用作到下一层的输入。在通过卷积层的每个通道中，数据被滤波。相应地，图像数据和其他类型的数据两者都可以被卷积以搜索(即，滤波)任何特定特征。当通过池化层时，输入数据大体保持其预测信息，但是数据量被减少。由于图1J(b)的CNN系统10a包括许多层，所以CNN被布置为预测输入图像包含许多不同特征中的任何一个特征。

[0056] CNN的一个其他特性是使用反向传播来减少误差并改进神经网络的质量以识别巨大量的输入数据之中的特定特征。例如，如果CNN到达小于1.0的预测，并且预测稍后被确定为是准确的，则预测的值与1.0之间的差被认为是误差率。由于神经网络的目标是准确地预测特定特征是否被包括在输入数据集中，所以CNN可以被进一步引导以自动地调节在投票层中应用的加权值。

[0057] 反向传播机制被布置为实现梯度下降的特征。梯度下降可以被应用在二维图上，其中图的一个轴线表示“误差率”，并且图的另一轴线表示“权重”。以这种方式，这样的梯度下降图将优选呈现抛物线形状，使得如果误差率高，则该得到的值的权重将是低的。当误差率下降时，则得到的值的权重将增大。因此，当实现反向传播的CNN继续操作时，CNN的准确性具有继续自动改进其本身的潜力。

[0058] 使用机器学习方法的已知目标识别技术的性能通过将更强大的模型应用到更大的数据集并且实现更好的技术来防止过拟合而被改进。两个已知的大数据集包括LabelMe和ImageNet。LabelMe包括成百上千的完全分割的图像，并且超过22,000类别中的多于1500万的高分辨率的带标签图像被包括在ImageNet中。

[0059] 为了从数百万图像中学习数千个对象，被应用到图像的模型要求大的学习能力。具有足够学习能力的一种类型的模型是卷积神经网络(CNN)模型。为了补偿关于巨大数据池的特定信息的缺乏，CNN模型利用数据集合的至少一些先验知识(例如，统计平稳性/非平

稳性、空间性、时间性、像素相关性的局部性等)来布置。CNN模型还利用设计者可选择的特征集合(诸如能力、深度、广度、层数等)来布置。

[0060] 早先的CNN利用大型专用超级计算机来实现。常规CNN利用定制的强大的图形处理单元(GPU)来实现。如由Krizhevsky描述的,“与2D卷积的高度优化实现配对的当前GPU足够强大以促进对感兴趣的大型CNN的训练,并且诸如ImageNet的最近的数据集包含足够的带标签示例以在没有严重过拟合的情况下训练这样的模型”。

[0061] 图2A-图2C在本文中可以被统称为图2。

[0062] 图2A是已知AlexNet DCNN架构的图示。如由Krizhevsky描述的,图1示出了“两个GPU之间的职责的描绘。一个GPU运行图的顶部处的层部分而同时另一个运行底部处的层部分。GPU仅在某些层处进行通信。网络的输入为150,528维,并且网络的剩余层中的神经元的数目由253,440-186,624-64,896-64,896-43,264-4096-4096-1000给出。”

[0063] Krizhevsky的两个GPU实现高度优化的二维(2D)卷积框架。最终网络包含具有权重的八个学习的层。八个层包括:五个卷积层CL1-CL5,其中的一些跟随有最大池化层;以及具有最终1000路softmax的三个完全连接层FC,最终1000路softmax产生1000个类标签上的分布。

[0064] 在图2A中,卷积层CL2、CL4、CL5的核仅被连接到先前层中在相同GPU上处理的核图。对比之下,卷积层CL3的核被连接到卷积层CL2中的所有核图。完全连接层FC中的神经元被连接到先前层中的所有神经元。

[0065] 响应归一化层在卷积层CL1、CL2之后。最大池化层在响应归一化层以及卷积层CL5两者之后。最大池化层对相同核图中的神经元的相邻组的输出求和。修正线性单元(ReLU)的非线性被应用到每个卷积和完全连接层的输出。

[0066] 图1A的AlexNet架构中的第一卷积层CL1利用大小为11×11×3的96个核以及4个像素的步幅对224×224×3输入图像进行滤波。该步幅是核图中的相邻神经元的感受野中心之间的距离。第二卷积层CL2将第一卷积层CL1的响应归一化的且池化的输出作为输入并且利用大小为5×5×48的256个核对第一卷积层的输出进行滤波。第三、第四和第五卷积层CL3、CL4、CL5被连接到彼此而没有任何中介池化或归一化层。第三卷积层CL3具有被连接到第二卷积层CL2的归一化的、池化的输出的大小为3×3×256的384个核。第四卷积层CL4具有大小为3×3×192的384个核,并且第五卷积层CL5具有大小为3×3×192的256个核。完全连接层各自具有4096个神经元。

[0067] AlexNet架构的八层深度似乎重要,因为特定测试披露了移除任何卷积层导致不可接受地削弱的性能。网络的大小受实现的GPU上可用的存储器的量限制并且受被认为可容忍的训练时间量限制。图1A的AlexNet DCNN架构花费五至六天来在两个NVIDIA GEFORCE GTX580 3GB GPU上训练。

[0068] 图2B是诸如NVIDIA GEFORCE GTX 580GPU的已知GPU的框图。GPU是包含采用灵活标量架构的32个统一设备架构处理器的流式多处理器。GPU被布置用于纹理处理、阴影图处理、以及其他以图形为中心的处理。GPU中的32个处理器中的每个处理器包括完全流水线的整数算术逻辑单元(ALU)和浮点单元(FPU)。FPU符合针对浮点算术的IEEE 754-2008工业标准。GPU在这种情况下被特别地配置用于台式机应用。

[0069] GPU中的处理以被称为曲数(warp)的32个线程的组来调度。32个线程中的每个线

程同时执行相同指令。GPU包括两个曲数调度器和两个指令分发单元。在该布置中，两个独立的曲数可以同时被发出并执行。

[0070] 本背景部分中讨论的所有技术方案不一定是现有技术并且不应当仅仅由于其在背景部分中的讨论而被假定为现有技术。按照这些原则，在背景部分中讨论的或与这样的技术方案相关联讨论的现有技术的问题的任何识别不应当被当作现有技术，除非明确被陈述为现有技术。相反，对背景部分中的任何技术方案的讨论都应当被当作本发明人对特定问题的方案的部分，其本身也可以是具有创造性的。

实用新型内容

[0071] 为了提高标识和隔离三维特征图内的三维体的效率，提出了一种集成电路。

[0072] 一种集成电路可以被概述为包括：板载存储器(例如，随机存取存储器(RAM))；应用处理器；数字信号处理器(DSP)集群；可配置的加速器框架(CAF)；以及至少一个通信总线架构，其将应用处理器、DSP集群、以及CAF通信地耦合到板载存储器，其中CAF包括：可重配置的流交换器(switch)；以及数据体雕刻单元，其具有耦合到可重配置的流交换器的至少一个输入和耦合到可重配置的流交换器的输出，数据体雕刻单元具有计数器、比较器和控制器，数据体雕刻单元被布置为：经由至少一个输入接收特征图数据的流，特征图数据的流形成三维(3D)特征图，3D特征图被形成为多个二维(2D)数据平面；标识3D特征图内的3D体，3D体在尺寸上小于3D特征图；从3D特征图隔离在3D体内的数据以用于在深度学习算法中进行处理；以及经由输入提供所隔离的数据。

[0073] 数据体雕刻单元还可以被布置为：经由至少一个输入接收限定第一2D数据平面中的感兴趣区域的输入信息，输入信息包括感兴趣区域的至少一个第一坐标和足以形成第一2D数据平面中的封闭2D体的另外信息；加载并按顺序排好计数器，使得第一2D数据平面中的每个数据以选择的顺序被分析；使用比较器来确定被分析的每个数据是否在封闭2D体内，其中提供隔离的数据输出包括提供被确定为在封闭2D体内的每个数据。

[0074] 数据体雕刻单元还可以被布置为：经由至少一个输入接收限定第一2D数据平面中的感兴趣区域的输入信息，输入信息包括感兴趣区域的至少一个第一坐标和足以形成第一2D数据平面中的封闭2D体的另外信息；加载并按顺序排好计数器，使得第一2D数据平面中的每个数据以选择的顺序被分析；使用比较器来确定被分析的每个数据是否在封闭2D体内；以及生成包括被确定为在封闭2D体内的每个数据的有序数据结构。

[0075] 数据体雕刻单元还可以被布置为将在3D特征图的多个封闭2D体内的数据包括在有序数据结构中，其中多个2D数据平面中的每个2D数据平面具有相应的封闭2D体，并且其中每个相应的封闭2D体与在相邻2D数据平面中限定的至少一个其他封闭2D体相关联。多个2D数据平面中的的每个2D数据平面可以具有限定于其中的多个封闭2D体。在所选择的2D数据平面上的多个封闭2D体中的个体封闭2D体可以是非重叠的。集成电路可以被形成为片上系统。

[0076] 本实用新型的实施例能够高效地标识和隔离三维特征图内的三维体。

[0077] 已经提供本实用新型内容从而以简化的形式介绍下面在具体实施方式中进一步描述的某些构思。除非另行明确陈述，否则本实用新型内容不标识要求保护的技术方案的关键特征或必要特征，其也不旨在限制要求保护的技术方案的范围。

附图说明

[0078] 参考以下附图描述非限制性和非穷尽性实施例，其中除非另行说明，否则类似的标记在各个视图中指代类似的部分。附图中的元件的大小和相对位置不一定按比例绘制。例如，各个元件的形状被选择、放大并且定位以改进绘图易读性。如所绘制的元件的特定形状已经为了便于绘图中的识别而被选择。下文参考附图描述一个或多个实施例，在附图中：

[0079] 图1A是卷积神经网络(CNN)系统的简化图示；

[0080] 图1B图示了图1A的CNN系统确定第一像素图案图示“1”并且第二像素图案图示“0”；

[0081] 图1C示出了一和零的不同形式的若干变型；

[0082] 图1D表示利用已知图像的对应部分来分析(例如，数学上组合)未知图像的部分的CNN操作；

[0083] 图1E图示图1D的右侧已知图像的六个部分；

[0084] 图1F图示滤波过程中的卷积的12个动作；

[0085] 图1G示出图1F的12个卷积动作的结果；

[0086] 图1H图示核值的图的堆叠；

[0087] 图1I示出显著减少由卷积过程产生的数据量的池化特征；

[0088] 图1J更详细地示出了图1A的CNN系统；

[0089] 图2A是已知AlexNet DCNN架构的图示；

[0090] 图2B是已知GPU的框图；

[0091] 图2C是来自使用网络流来链接多个视频剪辑中的管方案的T-CNN论文的示例；

[0092] 图3是具有集成于其中的、被图示为框图的DCNN处理器实施例的示例性移动设备；

[0093] 图4是描绘可配置加速器框架(CAF)的实施例，诸如图3的图像和深度卷积神经网络(DCNN)协同处理器子系统；

[0094] 图5是更详细的流交换器实施例；

[0095] 图6是卷积加速器(CA)实施例；

[0096] 图7是图示由卷积神经网络算法内的数据体雕刻器900单元支持的数据路径的高级别框图；

[0097] 图8A-图8C图示并呈现感兴趣区域内的各种雕刻的三维(3D)体，其被用在机器学习算法中，该机器学习算法诸如预测或分类视频流中的选择动作或场景的机器学习算法；

[0098] 图9是与图3-图6的硬件加速的DCNN处理器集成的数据体雕刻器单元的一个实施例；以及

[0099] 图10是图示至少一种数据体雕刻方法的数据流程图。

具体实施方式

[0100] 本实用新型可以通过参考本实用新型的优选实施例的以下具体实施方式而更容易地理解。应理解，本文中使用的术语仅仅是为了描述特定实施例的目的并且不旨在为限制性的。还应理解，除非在本文中特别限定，否则本文中使用的术语应被给予如在相关领域中已知的其传统含义。

[0101] 已知神经网络中的深度卷积处理在执行如图形中的对象分类的动作时产生优良

的结果。然而，欠发展的是高效地检测并分类视频流中的对象、场景、动作或其他感兴趣点的过程。因为视频数据是复杂的，并且因为视频缺乏如此容易地被附着到图像数据的标注，所以用于检测视频内的感兴趣点的手段尚未受到如此多的关注。在已经做出了解决问题的尝试的情况下，主要方法必须在两个主要阶段中应用卷积神经网络技术。第一个阶段试图标识单个帧中的“动作”，并且然后第二个阶段试图将疑似动作跨若干帧相关联。这些方法在卷积神经网络中创建一个流以在空间上标识特征并在网络中创建第二个分离的流以在时间上标识特征。

[0102] 在来自中佛罗里达大学(UCF)的计算机视觉研究中心(CRCV)的Rui Hou和其他人的论文“Tube Convolutional Neural Network(T-CNN)for Action Detection in Videos”中提出了推进视频分类技术的一种建议。在该论文(在本文中被称为T-CNN论文)中，Hou建议创建被称为管卷积神经网络(T-CNN)的端到端深度网络以用于视频中的动作检测。T-CNN的本质包括能够基于三维(3D)卷积特征来识别并定位动作的统一深度网络。视频被划分成相等长度的剪辑，并且对于每个剪辑，基于3D卷积网络特征来生成一组管方案。随后，不同剪辑的管方案被链接在一起作为数据的网络流，并且使用链接的视频方案来执行时空动作检测。

[0103] T-CNN论文描述了其中视频剪辑被馈送到管方案网络(TPN)中以获得一组管方案的方法。来自每个视频剪辑的管方案根据它们的“动作度分数”被链接，并且相邻方案之间的重叠被分析以形成针对视频中的时空动作定位的完整管方案。然后，感兴趣管的池化被应用到管方案以生成用于动作标签预测的固定长度特征向量。对T-CNN论文而言重要的是创建具有跨每个帧链接的特别地隔离的空间信息(即，“管方案”)的时间上相邻帧的TPN。

[0104] 图2C是来自使用网络流来链接多个视频剪辑中的管方案的T-CNN论文的示例。在该图中，在三个单独的视频剪辑中标识两个管方案。管方案由上标标识符和下标标识符标识。第一管方案具有上标“1”(管^{1})，第二管方案具有上标“2”(管^{2})。下标信息(C1,C2,C3)用于标识从其提取相应管方案的剪辑(例如，管$^{1}_{C1}$，管$^{2}_{C3}$，等等)。

[0105] 在本公开中不讨论T-CNN论文中提出的基于管的处理的益处和成功。相反，本公开说明并描述用于标识和隔离3D特征图内的三维(3D)体以用于深度学习算法中的处理的改进的机制。因此，本公开是经由将基于硬件的数据体雕刻器结构与其他深度卷积神经网络结构集成的系统、设备和方法超越T-CNN论文和类似的工作的改进。除此之外，本公开因此教导了用于产生在卷积神经网络中处理的“管方案”的新的、更快的、更高效的、且更低功率的设备和方法。

[0106] 根本地，发明人已经意识到学习机器可以在附加专用硬件结构与实现学习机器的架构集成或者以其他方式对实现学习机器的架构可用的情况下得到改进。可以做出的一个这样的改进包括实现如本文所描述的一个或多个数据体雕刻器900单元的结构和动作。本实用新型的数据体雕刻器900单元是本公开中描述的特定类型的基于硬件的数据解析器，其可以被是实现在各种各样的学习机器上。然而，为简洁起见，本公开包括以DESOLI等人的题为DEEP CONVOLUTIONAL NETWORK HETEROGENEOUS ARCHITECTURE的美国专利申请No.15/423,272中公开的特定深度卷积神经网络中对本实用新型的数据体雕刻器900单元的实现，该申请通过引用被并入本申请中。该特定深度卷积网络异构架构学习机器公开了片上系统(SoC)，其具有系统总线、耦合到系统总线的多个可寻址的存储器阵列、耦合到系统总线的

至少一个应用处理器核心、以及耦合到系统总线的可配置的加速器框架。可配置的加速器框架是图像和深度卷积神经网络(DCNN)协同处理系统。SoC还包括耦合到系统总线的多个数字信号处理器(DSP),其中多个DSP将功能与可配置的加速器框架协同以执行DCNN。

[0107] 图3-图6及其伴随的详细描述说明并呈现可配置为高性能的、能量高效的硬件加速DCNN处理器的示例性片上系统(SoC)110的元件。图7是图示由卷积神经网络算法内的数据体雕刻器900单元支持的数据路径的高级别框图。图8A-图8C及其伴随的详细描述说明并呈现感兴趣区域内的各种雕刻的三维(3D)体,其被用在机器学习算法中,该机器学习算法诸如预测或分类视频流中的选择动作或场景的机器学习算法。图9是与图3-图6的硬件加速的DCNN处理器集成的数据体雕刻器900单元的一个实施例,并且图10是图示被布置为生成限定3D特征图内的3D体的有序数据结构的至少一种数据体雕刻方法1000的数据流程图。

[0108] 图3的示例性SoC 110(其特别有用于机器学习应用)实现:图像和DCNN协同处理器子系统400(图4),其可以可互换地被称为可配置的加速器框架;架构上高效的流交换器500(图5),其以前所未有的水平创建数据局部性;一组卷积加速器600(图6),其执行输入特征数据与从对神经网络的训练得到的核数据的卷积;以及一组数据体雕刻器单元900,其被特别地布置用于深度学习引擎(图9)。

[0109] 图3是具有集成于其中的、被图示为框图的DCNN处理器实施例的示例性移动设备100。移动DCNN处理器被布置为片上系统(SoC)110,然而也预见到其他布置(例如,多个芯片,单个集成电路中的若干芯片裸片等)。图示的SoC 110包括多个SoC控制器120、可配置的加速器框架(CAF)400(例如,图像和DCNN协同处理器子系统)、SoC全局存储器126、应用(例如,主机)处理器128、以及多个DSP 138,其中的每一个被直接地或间接地通信耦合到主(例如,系统)通信总线132和次通信(例如,DSP)总线166。

[0110] 可配置的加速器框架(CAF)400被通信耦合到系统总线166,其提供用于使CAF 400的卷积加速器根据需要访问SoC全局存储器126并且根据需要与DSP 138通信的机制。在图4中更详细地图示了CAF 400。

[0111] SoC 110包括各种SoC控制器120,其中的一些控制SoC 110,并且其中的其他控制一个或多个外围设备。SoC控制器120包括应用(例如,主机)处理器128(例如,ARM处理器或某种其他主机处理器)、时钟发生器168(例如,时钟管理器)、重置控制器170、以及功率管理器172,以提供对SoC 110和其他部件的各种计时、功率消耗和其他方面的附加支持、控制和管理。控制外设的其他SoC控制器120包括低速外围I/O接口130和外部存储器控制器174以与SoC 110被嵌入到其中的示例性设备100的外部芯片、部件或存储器通信或以其他方式访问SoC 110被嵌入到其中的示例性设备100的外部芯片、部件或存储器。

[0112] 应用处理器128可以用作中间模块或用作到SoC 110与之集成的示例性电子设备100的其他程序或部件的接口。在一些实施例中,应用处理器128可以被称为应用处理器核心。在各种实施例中,应用处理器128在启动时加载SoC配置文件并且根据配置文件来配置DSP 138和CAF 400。当SoC 110处理一批或多批输入数据(例如,图像)时,应用处理器128可以基于配置文件来协同CAF 400或DSP 138的重新配置,配置文件本身可以基于DCNN层和拓扑。

[0113] SoC 110还包括支持SoC控制器120与DSP 138之间以及SoC控制器120与CAF 400之间的通信的主通信总线132(例如,AXI-先进可扩展接口)。例如,DSP 138或CAF 400可以经

由主通信总线132与应用处理器128、一个或多个外围控制器/外围通信接口(低速外围I/O)130、经由外部存储器控制器174的外部存储器(未示出)或其他部件通信。SoC控制器120还可以包括其他支持和合作设备，诸如时钟管理器(例如，时钟发生器)168、重置控制器170、功率管理器172，以提供对SoC 110和其他部件的附加的定时和功率管理。

[0114] 在一些实施例中，并且如图3中图示的，多个DSP 138以多个DSP集群被布置，诸如第一DSP集群122、第二DSP集群140以及为简化图示未提及的若干其他DSP集群。

[0115] 每个DSP集群122、140包括多个(例如，两个)DSP 142、152、多个(例如，两个)局部DSP纵横开关144、154以及DSP集群纵横开关145、155。特定集群中的每个DSP 142、152能够经由DSP集群纵横开关145、155与其他DSP 142、152通信。每个DSP 142、152能够经由其对应的局部DSP纵横开关144、154访问对应的指令高速缓存146、156和局部DSP存储器148。在一个非限制性实施例中，每个指令高速缓存146、156是4路16kB指令高速缓存，并且每个局部DSP存储器148、158是针对其对应的DSP的64kB的局部RAM存储。每个DSP集群122、140还包括共享的DSP集群存储器160、159和用于访问SoC全局存储器160、159的集群DMA 162、164。

[0116] 每个DSP集群122、140经由DSP集群纵横开关145、155被通信耦合到全局DSP集群纵横开关150以使得每个DSP集群122、140中的每个DSP 142、152能够与彼此通信并且与SoC 110上的其他部件通信。全局DSP集群纵横开关150使得每个DSP能够与多个DSP集群138中的其他DSP通信。

[0117] 附加地，全局DSP集群纵横开关150被通信耦合到系统总线166(例如，次通信总线、xbar-SoC纵横开关等)，其使得每个DSP能够与SoC 110的其他部件通信。例如，每个DSP 142、152可以与CAF 400的一个或多个部件(例如，一个或多个卷积加速器)通信或者经由系统总线166访问SoC全局存储器126。在一些实施例中，每个DSP 142、152可以经由其对应的DSP集群122、140的DMA 162、164与SoC存储器126通信。此外，DSP 142、152可以经由系统总线166根据需要与控制器120或SoC 110的其他模块通信。每个DSP经由其局部DSP纵横开关144、154、其DSP集群纵横开关145、155、以及全局DSP集群纵横开关150访问系统总线166。

[0118] 多个DSP 138可以被分配或指派为执行特定指令以加速DCNN的其他操作。这些其他操作可以包括在DCNN过程期间被执行的非卷积操作，其在一些情况下主要由CAF 400执行。这些非卷积操作的示例包括但不限于最大或平均池化、非线性激活、跨通道响应归一化、表示总DCNN计算的小部分但是更适合于未来算法演变的分类、或者其他操作，例如Min、Max、Sqrt、Mac、Butterfly、Average、2-4SIMD ALU。在一些情况下，先前已经使用DSP 138中的一个或多个执行的操作现在使用本文参考图7描述的用于深度学习加速结构的算术单元来执行。因此，本文描述的处理器及其相关联的计算设备的改进的操作可以由本文描述的算术单元结构实现。

[0119] DSP 138可以与CAF 400中的CA的操作并发地(例如，并行地)操作并且与数据传输并发地(例如，并行地)操作，其可以通过中断、邮箱(mailbox)或用于并发执行的某种其他同步机制来同步。

[0120] 在各种实施例中，SoC存储器126包括用于存储CAF 400或DSP 138的部件可访问的数据的多个存储器部件。在至少一个实施例中，SoC存储器126以层级型存储器结构被配置。在一个非限制性示例中，SoC存储器126包括各自具有1M字节的存储空间的四个SRAM库。

[0121] 在至少一个实施例中，可配置的加速器框架(CAF)400可以被组织为SoC 110的图

像和DCNN协同处理器子系统。如本文所描述的，CAF 400包括可重配置的数据流加速器结构，其将高速相机接口与用于深度学习加速的算术单元(图6)中的算术单元、传感器处理流水线、剪裁器、颜色转换器、特征检测器、视频编码器、八通道数字麦克风接口、流式DMA、以及多个卷积加速器中的任何一个或多个连接。

[0122] 结合图4描述关于CAF 400的附加细节。简单来说，CAF 400诸如从相机接口或其他传感器接收传入信息(例如，图4中的图像数据，但是在不同的实施例中为其他类型的流式数据)，并且将传入数据分布到CAF 400的各种部件(例如，结合图6更详细地描述的卷积加速器，结合图7-图10更详细地描述的数据体雕刻器900单元，等等)和/或多个DSP 138中的一个或多个以采用DCNN并识别传入图像中的对象。

[0123] CAF 400经由到或来自不同种类的源或槽(sink)部件的可配置的完全连接交换器利用单向链路来传送数据流。例如，可配置的完全连接交换器(其结合图5更详细地描述)可以经由直接存储器访问(DMA)将数据传送到SoC全局存储器126、I/O接口(例如，相机)、以及各种类型的加速器(例如，卷积加速器(CA)600、数据体雕刻器900单元等)。在一些情况下，CAF 400在启动时基于从特定SoC配置工具接收的信息被配置，并且CAF 400在运行时期间基于限定的DCNN层和拓扑或从一个或多个DSP 138、应用处理器128等等接收的信息被重新配置。

[0124] CAF 400允许限定在运行时可选择数目的并发的虚拟处理链。CAF 400还包括全功能背压机制以控制到框架的各种部件的数据流。CAF 400被布置用于流多播操作，其使得数据流能够在多个块实例处重新使用。链接的列表控制整个卷积层的完全自主处理。分组的或链接到一起的多个加速器并行地处置针对特征图数据和多个核的变化的大小。将卷积加速器(CA)600进行分组以实现更大的计算实体使得能够选择可用数据带宽、预算功率和可用处理资源的可接受的优化的平衡。每个CA 600包括行缓冲区以与单个存储器访问并行地提取高达预定数目(例如，12)的特征图数据字。进一步支持CA 600结构的是数据体雕刻器900单元，其隔离并递送表示特征图中的3D感兴趣区域的三维(3D)体。不是如将由其他设备完成的从CAF 400的外部确定一个或多个建议的3D体，而是将数据保留在CAF 400架构内，由此实现显著的速度和数据吞吐量增益。

[0125] 在每个CA(600)中，基于寄存器的核缓冲区提供多个读取端口(例如，36个)，同时多个定点乘积累加(MAC)单元(例如，36个16位MAC单元)每个时钟周期执行多个MAC操作(例如，每个时钟周期高达36个操作)。加法器树将针对每个核列的MAC结果累加。对MAC操作的重叠的基于列的计算允许对特征图数据针对多个MAC的可接受的优化的重新使用，由此减少与冗余存储器访问相关联的功耗。

[0126] 核集合被按批划分顺序处理并且中间结果可以被存储在SoC全局存储器126中。各种核大小(例如，高达12×12)、各种批大小(例如，高达16)、以及并行核(例如，高达4)可以由单个CA 600实例处置但是任何大小的核可以利用累加器输入来适应。

[0127] 可配置的批大小和并行核的可变数目实现针对跨不同单元共享的可用的输入和输出带宽和可用的计算逻辑资源的可接受的优化的权衡。

[0128] CAF 400中的CA 600的不同的可接受的优化配置针对每个DCNN层被确定。这些配置可以使用以DCNN描述格式开始的整体工具(诸如Caffe'或TensorFlow)来确定或调节。CA 600当核用每权重8位或更少位进行非线性量化时支持运行中核解压缩和凑整，其中第1误

差率对于8位增加高达0.3%。

[0129] 图4是描绘可配置的加速器框架(CAF)400的一个实施例，诸如图3的图像和深度卷积神经网络(DCNN)协同处理器子系统400。CAF 400可以被配置用于图像处理、音频处理、预测分析(例如，技巧游戏、营销数据、人群行为预测、天气分析和预报、遗传图谱、疾病诊断以及其他科学、商业、消费和这样的处理)或者某种其他类型的处理；特别是包括卷积操作的处理。

[0130] CAF 400也被布置有许多可配置的模块。一些模块是可选的，并且一些模式是必需的。许多可选的模块通常被包括在CAF 400的实施例中。CAF 400的一个必需的模块例如是流交换器500。流交换器500提供设计时间参数化的、运行时可重配置的加速器互连框架以支持基于数据流的处理链。另一必需的模块例如是一组CAF控制寄存器402。其他模块也可以是必需的。CAF 400的可选模块包括系统总线接口模块404、所选择数目的DMA引擎406(例如，DMA控制器)、所选择数目的外部设备接口408、所选择数目的处理模块410、所选择数目的卷积加速器(CA)600、以及所选择数目的数据体雕刻器900单元(例如，1个、2个、4个、8个或其他数目)。

[0131] 流交换器500是利用多个单向“流链路”形成的可重配置的单向互连结构。流链路被布置为将多位数据流从加速器、接口和其他逻辑模块传送到流交换器500并且从流交换器500传送到加速器、接口和其他逻辑模块。

[0132] 除了流交换器500，CAF 400还可以包括系统总线接口模块404。系统总线接口模块404提供到SoC 110的其他模块的接口。如图3的示例性实施例中示出的，CAF 400被耦合到次通信总线166。在其他情况下，CAF 400可以被耦合到主通信总线132或某个其他通信机制。控制信息可以通过CAF 400的系统总线接口模块404被单向地或双向地传递。这样的接口用于向主机处理器(例如，DSP集群130的DSP、应用处理器128或另一处理器)提供对用于控制、操作或以其他方式引导框架的特定特征的所有CAF控制寄存器402的访问。在一些实施例中，每个DMA引擎406、外部设备接口408、处理模块410、卷积加速器600和数据体雕刻器900具有到具有限定集合的配置寄存器(例如，以CAF控制寄存器402形成)的配置网络的接口。

[0133] CAF 400包括多个DMA引擎406。在图4中，图示了十六个DMA引擎406a至406p，但是一些其他数目的DMA引擎可以根据由半导体从业者在设计时做出的一个或多个选择而被包括在SoC 110的其他一些实施例中。DMA引擎406被布置为提供针对输入数据流、输出数据流、或输入和输出数据流的双向通道。在这些情况下，大量数据被传递到CAF 400中，从CAF 400被传递出去，或者被传递到CAF 400中并且从CAF 400被传递出去。例如，在一些情况下，一个或多个DMA引擎406用于从存储器或从产生大量视频数据的数据源设备(例如，高清(HD)视频相机)传递流式视频数据。视频中的一些或全部可以从源设备被传入，从SoC全局存储器126被传入或传出到SoC全局存储器126，等等。

[0134] 在一个示例性实施例中，一个或多个DMA引擎406被连接到具有一个输入端口504(图5)和一个输出流端口516(图5)的流交换器500。DMA引擎406可以被配置成输入或输出模式。DMA引擎406可以被配置为打包数据并将数据发送到可在主通信总线132、次通信总线166上访问的任何地址位置、或者某个其他地址位置。DMA引擎406还可以附加地或备选地被配置为对提取的数据拆包并将拆包的数据转化为数据流。

[0135] 图4的CAF 400包括设计时间可选择的、运行时可配置的多个外部设备接口408。外部设备接口408提供与产生(即,源设备)或消耗(即,槽设备)数据的外部设备的连接。在一些情况下,传递通过外部设备接口408的数据包括流式数据。传递通过外部设备接口408的流式数据的量可以在一些情况下是预定的。备选地,传递通过外部设备接口408的流式数据的量可以是不确定的,并且在这样的情况下,外部设备可以简单地产生或消耗数据,无论何时特定外部设备被启用并且被如此引导。通过外部设备接口408耦合的外部设备可以包括图像传感器、数字麦克风、显示监视器、或其他源设备和槽设备。在图4中,外部设备接口408包括数字视频接口(DVI)外部设备接口408a、第一图像传感器接口和图像信号处理器(ISP)外部设备接口408b以及第二图像传感器接口和ISP外部设备接口408c。也预见到其他接口,但是为图示的简单,仅示出了三个外部设备接口408。

[0136] 多个处理模块410被集成在CAF 400中。为简单图示了三个处理模块410,但是另一所选择数目(例如,两个、四个、八个、十六个)的处理模块410还可以在设计时由半导体从业者集成在CAF 400中。第一处理模块410是被布置为执行某个视频(即,MPEG)处理和某个图像(即,JPEG)处理的MPEG/JPEG处理模块410a。第二处理模块410是H264处理模块410b,其被布置为执行特定视频编码/解码操作。第三处理模块410是颜色转换器处理模块410n,其被布置为对某个多媒体数据执行基于颜色的操作。

[0137] 在许多情况下,DMA控制器406、外部设备接口408、处理模块410、卷积加速器600、数据体雕刻器900单元以及集成在CAF 400中的其他模块是由半导体从业者在设计时从库中选择的IP模块。半导体从业者可以指定模块的数目、特定模块的特征、总线宽度、功率参数、布局、存储器可用性、总线访问、以及许多其他参数。

[0138] 表2是库中的IP模块的非穷举的示例性列表,其中的任一项可以由半导体从业者并入到CAF 400中。在许多情况下,当新模块被设计时,以及当现有模块被修改时,新IP将被添加到诸如表2的库的库。

[0139] 表2-IP模块的CAF库

[0140]

功能单元	应用
RGB/YUV 传感器接口	接口
拜耳传感器接口	接口
视频输出接口（DVI）	接口
增强 I/O（传感器接口、视频输出、叠加）	接口
ISP（图像信号处理器）	信号处理

[0141]

功能单元	应用
Mini ISP（图像信号处理器）	信号处理（拜耳-> RGB）
GP 颜色转换器单元	通用
图像剪裁器和大小调整器单元	通用
变形滤波器单元	通用
背景移除单元（+阴影移除）	背景/前景分割
参考帧更新单元	背景/前景分割
JPEG 编码器	编码器
JPEG 解码器	解码器
H264 编码器	编码器
H264 编码器	编码器（基线，仅内部）
修正和镜头失真校正	立体视觉
Census 变换单元（BRIEF）	立体视觉
立体视觉深度图生成器	立体视觉
特征点检测器（FAST）	特征检测
特征检测（Viola Jones）	面部检测（例如，积分图像、ISA 扩展）
特征检测（光流）	面部跟踪
特征点提取器（DoG+SIFT）	特征检测–高斯差分加尺度不变特征变换
特征提取	边缘提取（Sobel，Canny）
时钟和中断管理器	系统控制
调试支持单元	调试
GP IO 单元	通用
用于神经网络的 3D 卷积加速器	处理
数据体雕刻器	隔离所选择的 3D 特征体

[0142] 在图4的可配置的加速器框架(CAF)400中，表示了八个卷积加速器600，CA0至CA7。在其他的一些CAF 400实施例中，形成了不同数目的卷积加速器。卷积加速器600的数目和在每个卷积加速器600中可用的特定特征在一些情况下基于由半导体从业者在设计时选择

的参数值。

[0143] 卷积加速器(CA)600是具有所选择数目(例如,一个、两个、四个、八个)的输入和输出流链路端口的数据处理单元。一个或多个配置寄存器(例如,一组配置寄存器)被布置为控制CA 600的操作。在一些情况下,配置寄存器被包括在CAF控制寄存器402中,并且在这些或其他情况下,某些配置寄存器被形成为CA 600的一部分。

[0144] 一个或多个卷积加速器模板模块可以被包括在IP模块库(诸如参考表2描述的库)中。在这些情况下,存储于IP模块库中的数据包括减少构建实现加速器的核心功能的新加速器所需的工作的相关构建块。预定义的一组配置寄存器可以被扩展。在流链路端口处形成或者以其他方式定位的可配置的FIFO可以用于吸收数据率波动并提供放松处理链中的某些流控制约束所需的一些缓冲余量。

[0145] 通常,每个CA 600或消耗数据、生成数据,或既消耗数据又生成数据。消耗的数据传递通过可重配置的流交换器500的第一流链路,并且流式传输的数据传递通过流交换器500的第二流链路。在至少一些实施例中,CA不能够直接访问可由主通信总线132(图3)、次通信总线166(图3)或其他总线地址访问的存储器地址空间。然而,如果需要对在系统总线上传递的数据的随机存储器访问,则CA 600还可以使用可选的总线端口接口,其可以沿着图4的系统总线接口模块404的线路,其用于包括准许DMA引擎访问系统总线上的存储器位置的若干事情。如以上所讨论的,一些CA 600实现是库的一部分,其可以在其他CAF 400实施例中以用于简单地以全局系统定义文件来实例化CA 600。

[0146] 一个或多个数据体雕刻器模板模块还可以被包括在IP模块库(诸如参考表2描述的库)中。这里,预定义的一组配置寄存器还可以被扩展以提供用于所包括的数据体雕刻器单元的配置的参数存储。参数与任何期望数目的计数器、比较器、控制单元、计算单元、数据储存库、复用器电路、临时存储电路以及其他电路的配置相关联。

[0147] 每个数据体雕刻器900在输入流接口处接收包括一系列帧的信息;每个帧被形成为二维(2D)数据结构。数据体雕刻器900将确定帧中的每个帧的第一维和第二维,并且基于第一维和第二维,数据体雕刻器900将进一步针对每个帧确定要从相应帧提取的感兴趣区域的位置和大小。被传递到数据体雕刻器900单元中的数据可以源自于可重配置的流交换器500、CAF 400框架内部或外部的存储器、传感器或特定接口、或者来自某个其他源。按照这些原则,这些类型的数据源中的每一个可以在一些情况下消耗由数据体雕刻器900生成的数据。如本文所讨论的,一些数据体雕刻器900实现是库的部分,其可以在其他CAF 400实施例中用于简单地以全局系统定义文件来实例化数据体雕刻器。

[0148] 机器学习系统的系统级程序员期望有为它们的特定实现选择期望的编程模型的灵活性。为了支持这种高级灵活性,CAF 400被布置有可重配置的流交换器500。如本公开中所描述的,流交换器500用作数据传输结构以改进逻辑块(IP)重新使用、数据重新使用以及对其他部件和逻辑的重新使用,其进而允许减少片上和片外存储器流量,并且其提供大得多的灵活性来在不同的应用使用情况下利用相同逻辑块。集成于流交换器500中的是多个单向链路,多个单向链路被布置为经由可配置的完全连接交换器将数据流传送到不同种类的数据源、数据槽以及数据源和数据槽、从不同种类的数据源、数据槽以及数据源和数据槽传送数据流、以及将数据流传送到不同种类的数据源、数据槽以及数据源和数据槽并从不同种类的数据源、数据槽以及数据源和数据槽传送数据流,数据源和数据槽诸如直接存储

器访问(DMA)控制器、I/O接口(例如,相机)、以及各种类型的加速器。

[0149] 传送的数据可以采取任何期望的格式,诸如光栅扫描图像帧的流、面向宏块的图像的流、音频流、原始数据块、输入或数据体雕刻器值的流、或任何其他格式。流交换器500还可以沿着由每个单元转发到一个或多个或多个目标单元(在其中处理控制信息)的处理链传送消息、命令、或其他类似的控制信息。控制信息可以用于信号通知事件,重新配置处理链本身,或者引导其他操作。

[0150] 图5是更详细的流交换器实施例500。流交换器500包括用户可选择的、设计时可配置的第一数目的流链路输入端口504和用户可选择的、设计时可配置的第二数目的流链路输出端口516。在一些情况下,存在与所存在的输出端口一样多的输入端口。在其他的一些情况下,存在比输出端口多的输入端口,并且在另外的其他情况下,存在比输入端口多的输出端口。输入端口的数目和输出端口的数目在设计时被限定。

[0151] 在图5的流交换器500实施例中,详细地示出了一个流链路502实施例。还图示了其他流链路502a、502b,但是为图示的简单没有细节。流链路502a、502b大体沿着流链路502的线路被布置,并且为清楚起见,在本公开中,图示的流链路中的任何可以被标识为流链路502。

[0152] 在运行时,流交换器500根据写入到CAF控制寄存器402(图4)中的某些CAF控制寄存器的配置数据来将输入流链路端口通过流链路502通信地耦合到输出流链路端口。在实施例中,输入流链路端口504中的一个或多个可以被期望地布置为在相同时钟周期上将接收的数据流同时转发到一个或多个(多播)输出端口516。因此,一个输入流链路端口可以被通信地耦合(例如,电连接到数据的通路)到一个或多个输出流链路接口,其导致输入数据流的物理复制。流链路502提供用于传送数据流以及与数据流相关联的控制信息的直接单向接口。在这样的实施例中,单个控制信号(其可以在一些情况下在单个专用或共享数据路径上被传播)提供流控制。

[0153] 流链路的一些导体用于传递数据;一些其他导体可以包括数据有效性指示器、第一像素指示器、最后一个像素指示器、行类型定义、以及暂停信号。暂停信号被用作背压(例如,流控制)机制。在流链路的一些实施例中,图像数据、命令数据、控制信息、消息等以基于帧的协议沿着通过流交换器500的处理链被传递。

[0154] 在流交换器500中,每个输出端口516与特定流链路502相关联。在图5中,例如,输出端口X与流链路502相关联。另外,一个或多个输入端口504与每个流链路相关联。在一些情况下,例如,每一个输入端口504与每一个流链路502相关联。以这种方式,每个输入端口504可以将数据同时或在不同的时间传递到任何和所有输出端口516。

[0155] 流链路的个体通信路径管道都是单向的。即,每个通信路径管道上的信号仅在一个方向上流动。在一些情况下,多个通信路径管道单向地接受从输入端口接收的数据并将数据传递到一个或多个输出端口。在这些情况下,以及在其他情况下,单个通信路径管道从输出端口单向地接收命令信息(例如,流控制信息)并将命令信息传递到一个或多个输入端口。在一些其他情况下,从输出端口接收的并且被传递到一个或多个输入端口的命令信息在两个或更多个通信路径管道上被传递。

[0156] 如图5的详细的流交换器502中示出的,来自多个输入端口504的一组单向通信路径管道被传递到数据交换器506中。在一些情况下,来自每个输入端口504的一组单向通信

路径管道被传递到数据交换器506中。在其他的一些情况下，一个或多个但少于全部的输入端口504的单向通信路径管道被传递到特定流链路502的数据交换器506中。数据交换器506可以包括复用器逻辑、解复用器逻辑、或者某种其他形式的交换逻辑。

[0157] 如图5中示出的，从多个输入端口504传递到流链路502中的数据可以同时存在于数据交换器506的输入节点处。选择机制508被布置为确定哪个输入数据被传递通过数据交换器506。即，基于选择机制508，来自输入端口A、B、C、D中的一个的输入数据通过数据交换器506被传递到数据交换器506的输出。输出数据将在$N_{A...D}$单向通信路径管道上被传递，$N_{A...D}$单向通信路径管道将匹配所选择的输入端口的单向通信路径管道的数目。

[0158] 选择机制508根据流交换器配置逻辑510被引导。流交换器配置逻辑510在运行时确定哪个输入端口504应向相关联的输出端口供应数据，并且基于该确定，流交换器配置逻辑510形成被传递到数据交换器506的合适的选择信号。流交换器配置逻辑510在运行时并且实时操作。流交换器510可以从CAF控制寄存器、从DSP集群122的DSP(图3)、从应用处理器128、或者从某个其他控制设备获取引导。另外，流交换器配置逻辑510还可以从消息/命令逻辑512获取引导。

[0159] 在一些实施例中，数据均一地被传递通过每个特定流链路502。即，在一些情况下，一个流链路502被配置(例如，流交换器配置逻辑510、CAF控制寄存器等)为协作地传递任何数目N的第一数据(例如，位、字节、字、半字节、元组或一些其他数据样本等)，并且一个或多个其他流链路502类似地配置为传递对应的第二数据。在该配置中，对于被传递通过第一流链路502的每个数据，存在被传递通过其他一个或多个流链路502中的每一个的对应数据。

[0160] 在其他的一些实施例中，数据不是均一地被传递通过每个特定流链路502。例如，数据可以被交织，或者以另一种非均一方式被传递。在被交织的实施例中，各种流链路502可以被配置为交织数据。在一个这样的被交织示例中，第一流链路502可以被布置为从第一源(例如，输入端口504)传递“M”个数据，然后第一流链路502可以被布置为从第二源(例如，不同的输入端口504)传递“N”个数据。

[0161] 备选地，在又一个交错实施例中，两个流链路502可以被布置成以非均一方式传递不同数目的数据。即，在第一流链路502正在传递“M”个数据的同时，第二流链路502同时或并发传递“N”个数据。在本文描述的示例中，“M”和“N”是整数。在某些情况下，“M”和“N”是不同的整数。

[0162] 在一些流交换器500实施例中，例如由接口或加速器传递通过输入端口504的某些特定信息由流交换器500的一个或多个流链路502中的命令逻辑512识别并且用于对一个或多个流链路502实现再编程。在这些或其他的一些实施例中，流交换器500被配置为根据固定模式来合并数据流。例如，在至少一个情况下，流交换器500可以被布置为通过在两个或更多个输入端口504上传递的输入流之间切换来选择数据并将数据传递到输出端口516。例如，在每个行、每个帧、每N个事务之后或者通过某种其他措施，流交换器500可以被配置为将来自不同输入端口504的数据传递到所选择的输出端口516。

[0163] 从数据交换器506传递的数据可以在一些情况下传递通过一个或多个可选的输出同步逻辑级514。输出同步逻辑级514可以用于存储或以其他方式缓冲从耦合到输入端口504的数据源朝向耦合到输出端口516的数据槽设备传递的所选择量(例如，一位或多位，几个字节或许多字节等)的数据。这样的缓冲、同步以及其他这样的操作可以当数据源设备和

数据槽设备以不同速率、在不同阶段、使用不同时钟源或以可以彼此异步的其他方式来操作时被实现。

[0164] 流交换器500包括背压暂停信号机制，其用于将流控制信息从槽设备传递到源设备。流控制信息从槽设备被传递以通知数据流源设备降低其数据速率。降低数据速率将帮助避免槽设备中的数据溢出。

[0165] 背压暂停信号机制的一个部分包括包含于每个输入端口中的背压暂停信号路径。背压暂停信号路径被布置为背压单向通信路径管道。在图5中，图示了四个背压输入端口机制BP_A、BP_B、BP_C、BP_D；针对图示的输入端口中的每个输入端口一个背压输入端口机制。在其他的一些实施例中，每个输入端口的背压机制可以包括一个或多个单向通信路径管道。在一些实施例中，每个输入端口的背压机制具有相同数目的单向通信路径管道，其可以例如为单个管道。在这些情况下，例如，当耦合到特定输入端口的数据源设备检测到背压机制上的信号被断言时，特定数据源设备将减缓或停止被传递到相关联的输入端口的数据量。

[0166] 每个输出端口516包括背压机制的另一部分。针对图5的三个图示的输出端口X、Y、Z中的每一个图示了一个输出端口背压机制，BP_X、BP_Y、BP_Z。在一些情况下，每个输出端口背压机制包括相同数目的单向通信路径管道(例如，一个)。在其他的一些情况下，至少一个输出端口具有含有与另一输出端口的另一背压机制不同数目的单向通信路径管道的背压机制。

[0167] 输出端口背压机制管道被传递到每个流链路502中的组合背压逻辑518。在图5中，背压逻辑518接收背压控制信号BP_X、BP_Y、BP_Z。组合背压逻辑518还从流交换器配置逻辑510接收控制信息。组合背压逻辑518被布置为通过输入端口504的输入端口背压机制将相关的流控制信息传递到特定数据源设备。

[0168] 图6是卷积加速器(CA)实施例600。CA 600可以被实现为图4的卷积加速器600中的任何一个或多个。

[0169] CA 600包括三个输入数据接口和一个输出数据接口，其各自被布置用于耦合到流交换器500(图5)。第一CA输入数据接口602被布置用于耦合到第一流交换器输出端口516，第二CA输入数据接口604被布置用于耦合到第二流交换器输出端口516，并且第三CA输入数据接口606被布置用于耦合到第三流交换器输出端口516。CA输出数据接口608被布置用于耦合到所选择的流交换器输出端口504。每个CA输入数据接口602、604、606和输出数据接口608被耦合到的特定流交换器500端口可以默认地、在启动时或在运行时被确定，并且特定耦合可以在运行时以编程方式被改变。

[0170] 在一个示例性实施例中，第一CA输入数据端口602被布置为将批数据的流传递到CA 600中，第二CA输入数据端口604被布置为将核数据的流传递到CA 600中，并且第三CA输入数据端口606被布置为将特征数据的流传递到CA 600中。输出数据端口608被布置为传递来自CA 600的输出数据流。

[0171] CA 600包括若干内部存储器缓冲区。内部存储器缓冲区可以在一些实施例中共享共同的存储器空间。在其他的一些实施例中，内部存储器缓冲区中的一些或全部可以是与彼此分离且不同的。内部存储器缓冲区可以被形成为寄存器、触发器、静态或动态随机访问存储器(SRAM或DRAM)或者以某种其他结构配置来形成。在一些情况下，内部存储器缓冲区可以使用多端口架构来形成，多端口架构例如一个设备执行存储器中的数据“存储”操作同

时另一设备执行存储器中的数据“读取”操作。

[0172] 第一CA内部缓冲区610被物理地或虚拟地布置为与第一CA输入数据接口602一致。以这种方式，流式传输到CA 600中的批数据可以被自动存储于第一CA内部缓冲区610中，直到数据被传递到CA 600中的特定数学单元，诸如加法器树622。第一CA内部缓冲区610可以固定具有在设计时确定的大小。备选地，第一CA内部缓冲区610可以被限定有在启动时或运行时以编程方式确定的可变大小。第一CA内部缓冲区610可以是64字节、128字节、256字节或某个其他大小。

[0173] 第二CA内部缓冲区612和第三CA内部缓冲区614沿着第一CA内部缓冲区610的线路被形成。即，第二CA内部缓冲区612和第三CA内部缓冲区614可以各自具有在设计时确定的它们自己的固定大小。备选地，第二CA内部缓冲区612和第三CA内部缓冲区614可以具有在启动时或运行时以编程方式确定的可变大小。第二CA内部缓冲区612和第三CA内部缓冲区614可以是64字节、128字节、256字节或某个其他大小。第二CA内部缓冲区612被物理地或虚拟地布置为与第二CA输入数据接口604一致以自动地存储流式传输的核数据，直到核数据被传递到专用于存储核缓冲数据的专用第四CA内部缓冲区616。第三CA内部缓冲区614被物理地或虚拟地布置为与加法器树622一致以自动地存储求和的数据，直到其可以被传递通过CA输出接口604。

[0174] 第四CA内部缓冲区616是被布置为期望地存储核数据并将存储的核数据应用到多个CA乘积累加(MAC)单元620的专用缓冲区。

[0175] 第五CA内部缓冲区618是被布置为接收通过第三CA输入接口606传递的流式传输的特征数据的特征行缓冲区。一旦被存储于特征行缓冲区中，特征数据就被应用到多个CA MAC单元620。应用到CA MAC单元620的特征和核缓冲区数据在数学上根据本文描述的卷积操作进行组合，并且得到的来自CA MAC单元620的输出乘积被传递到CA加法器树622。CA加法器树622在数学上组合(例如，求和)传入的MAC单元数据和通过第一CA输入端口被传递的批数据。

[0176] 在一些情况下，CA 600还包括可选的CA总线端口接口624。CA总线端口接口624当其被包括时可以用于将数据从SoC全局存储器126或某个其他位置传递到CA 600中或者从CA 600中传递出来。在一些情况下，应用处理器128、DSP集群122的DSP或者某个其他处理器引导对数据、命令或其他信息到或自CA 600的传递。在这些情况下，数据可以通过CA总线端口接口624被传递，CA总线端口接口624可以本身被耦合到主通信总线132、次通信总线166或者某个其他通信结构。

[0177] 在一些情况下，CA 600还可以包括CA配置逻辑626。CA配置逻辑626可以完全驻留在CA 600中，部分驻留在CA 600中，或者在CA 600远程。配置逻辑600可以例如被完全或部分地实施在CAF控制寄存器402、SoC控制器120或SoC 110的某种其他结构中。

[0178] 图7是图示由卷积神经网络算法内的数据体雕刻器900单元支持的数据路径的高级别框图。如该图中图示的，具有各种特征和复杂性的图像数据被流式传输到卷积过程中，卷积过程可以利用形成于移动设备100(图3)的SoC 110(图3)中的可配置的加速器框架(CAF)400(图4)中的卷积加速器600(图6)来执行。输入数据902(其可以例如为特征图数据或者包括特征图数据)被流式传输到数据体雕刻器900中。根据数据体雕刻器900单元生成的雕刻的输出数据904从数据体雕刻器900被流式传输出去。

[0179] 在卷积过程中，输入图像的流被传递到神经网络中。为了预测、检测、或以其他方式标识特定特征(例如，动作、场景、对象、或者某个其他特征)，在每幅图像中隔离感兴趣区域并且将感兴趣区域通过多幅图像的深度链接以产生三维(3D)体。核通过每幅图像的宽度和高速、通过多幅图像的深度并通过3D体被卷积，以产生特征图的集合或“堆叠”。池化层执行子采样操作并将在一层处的特征簇的输出组合成后续层中的单个特征。一个或多个附加的雕刻、卷积和池化层进一步对输入数据进行滤波，并且一个或多个完全连接操作贯穿层被执行以将一层中的特征与其他层中的对应特征相关。在完全连接操作之后，分类/预测从数据显露。

[0180] 通过形成基于硬件的数据体雕刻器900单元，并且将这些雕刻器耦合到流交换器500，卷积神经网络的卷积过程或跨数据平面的序列处理感兴趣区域中的数据的其他机器学习设备可以以改进的灵活性、数据局部性和速度来执行。

[0181] 图8A-图8C可以在本文中被统称为图8。这些图图示并呈现感兴趣区域内的各种雕刻的三维(3D)体，其被使用在诸如预测或分类视频流中的所选择动作或场景的机器学习算法中。

[0182] 在图8A中，从多个二维(2D)数据平面908A、908B、908N来形成特征图。示出了三个数据平面，其也可以被称为“帧”，然而，特征图可以包括任何数目的两个或更多个数据平面。在许多但不是所有的情况下，连续数据帧表示数据帧的时间序列或系列。例如，如果每个数据平面都表示以每秒30帧捕获的视频数据的单幅图像，则从75帧的系列形成的特征图表示两个半(2.5)秒的视频数据的连续流。在其他的一些情况下，数据平面根本不是图像数据，并且相邻的帧可以在时间上相关或者基于一个或多个不同的特性。

[0183] 在第一2D数据帧908A中，期望提取(例如，隔离，分开，区分，标记等)2D数据平面中的来自封闭2D体内的信息。限定感兴趣区域的信息可以与2D数据平面的数据一起被流式传输，从机器算法或另一源被传入，被存储在数据存储库(例如，图4中的控制寄存器402)中，或者信息可以来自某个其他地方。信息在许多实施例中将包括感兴趣区域的至少一个第一坐标和足以形成2D数据平面中的封闭2D体的进一步信息。在图8A中，第一坐标可以包括限定矩形2D感兴趣区域910A的左上坐标910TL的2D(例如，x-y)信息。第二坐标可以包括限定矩形2D感兴趣区域910A的右下坐标910BR的2D(例如，x-y)信息。使用这两个坐标，整个封闭2D体910A可以被确定。对于包括在2D数据平面908A中的每个数据，可以确定具有其自己的特定坐标(例如，x-y)的数据是落入封闭2D体910A内部还是封闭2D体910A外部。因为封闭2D体910A是矩形的，所以2D体910A的外周边界利用线性数学运算来快速地且高效地确定。

[0184] 在第二2D数据帧908B中，期望提取第二2D体910B。第二2D体910B对应于第一2D体910A，但是第二2D体910B可以具有一个或多个不同的尺寸、不同的旋转取向、不同的数据或其他差异。然而，在3D体的雕刻中，期望将第一2D体910A与第二2D体910B链接在一起。如在第一数据平面908A中的第一感兴趣区域的情况下，限定第二感兴趣区域的信息可以与第二2D数据平面908B的数据一起被流式传输，从机器算法或另一源被传入，存储于数据存储库(例如，图4中的控制寄存器402)中，或者信息可以来自不同源。为了简化附图，未在第二2D体910B中示出左上坐标和右下坐标，但是这些坐标可以被确定，或者第二2D体910B的边界可以以不同方式被确定。

[0185] 根据第一2D体910A和第二2D体910B的这些原则，还可以隔离任何数目的附加的2D

体。“第N个”2D体910N被示出在图8A的第N个2D数据平面908N中。从图8A的表示中，应清楚的是，可以隔离2D数据平面的序列中的每个2D数据平面中的2D体。从一个2D数据平面到另一个的信息也可以被链接。在至少一个情况下，隔离感兴趣区域中的多个2D体或2D体的“堆叠”包括生成有序数据结构。有序数据结构以许多方式被布置。例如，有序数据结构可以存储2D数据平面的被确定为在封闭2D体内的每个数据或者足以表示数据在感兴趣区域内的信息。在一些情况下，形成元组，元组包括感兴趣区域标识符和对应于2D感兴趣区域的至少一个链表或其部分。如在链表中，多个元组可以指向相邻2D数据平面的前向元组和后向元组。以这种方式，单个有序数据结构可以呈现跨特征图而封闭感兴趣区域的3D体。更具体地，为了说明而非限制本原理，有序数据结构可以表示该组2D数据帧908A、908B、908N中的矩形2D感兴趣区域910A、910B、910N的组合。

[0186] 一个或多个感兴趣区域可以在每个2D数据平面中被隔离。在图8A中，示出了两个感兴趣区域，但是在一些情况下，几十、几百或甚至更多的感兴趣区域被选择以用于隔离和进一步的卷积神经网络处理。

[0187] 第二感兴趣区域912A、912B、912N被表示在图8A的2D数据平面(例如，帧)中。第二感兴趣区域被示出为椭圆，然而，可以期望具有曲线部分的任何感兴趣区域，并且数据体雕刻器900的操作的原理对于本领域技术人员而言将是显而易见的。

[0188] 图8A的“细节A”部分呈现了用于在数学上确定具有曲线部分的2D感兴趣区域的边界的一种技术。例如，椭圆可以作为平面上的围绕两个特定焦点的曲线被分析。焦点是可选择的，或者可以以其他方式被解读，使得当从曲线上的任何点和每个点到两个焦点的距离被求和时，结果将是恒定值。因此，线性或其他更复杂的数学计算可以用于确定所选择的2D感兴趣区域中的任何曲线区的边界。

[0189] 总结图8A中示出的内容中的一些，在数据体雕刻器900单元中接收特征图数据的流。特征图数据的流将三维(3D)特征图形成为多个二维(2D)数据平面。三个2D数据平面908A、908B、908N被示出，然而，特征图可以包括任何数目的数据平面。在3D特征图内标识两个3D体；3D体中的每一个在尺寸上小于3D特征图。第一3D体大体是矩形的并且包括特征图的第一2D数据平面中的矩形2D感兴趣区域910A、特征图的第二2D数据平面中的矩形2D感兴趣区域910B、以及特征图的第N个2D数据平面中的矩形2D感兴趣区域910N。第二3D体是曲线的(即，大体椭圆形)并且包括特征图的第一2D数据平面中的具有曲线部分的2D感兴趣区域912A、特征图的第二2D数据平面中的具有曲线部分的2D感兴趣区域912B、以及特征图的第N个2D数据平面中的具有曲线部分的2D感兴趣区域912N。3D特征图的落入感兴趣区域中的3D体中的任一个3D体内的隔离的数据被隔离以用于在深度学习算法中的处理。

[0190] 图8B按照图8A的原则。可以限定3D特征图内的任何数目的3D体，然而，为了简化该图，仅仅示出了两个不同的3D体。

[0191] 在图8B中，第一3D体由特征图的第一2D数据平面中的非对称四边形2D感兴趣区域914A、特征图的第二2D数据平面中的非对称四边形2D感兴趣区域914B、以及特征图的第N个2D数据平面中的非对称四边形2D感兴趣区域914N限定。

[0192] 在图8B中，第二3D体由特征图的第一2D数据平面中的多边形2D感兴趣区域916A、特征图的第二2D数据平面中的多边形2D感兴趣区域916B、以及特征图的第N个2D数据平面中的多边形2D感兴趣区域916N限定。

[0193] 图8C是图8B的第一2D数据平面908A的不同视图。附加的信息被添加在图8C中。具体地，示出了一组数据点。这些数据点是非限制性的，并且表示可以以许多方式执行包括限定封闭2D感兴趣区域的信息。在至少一些实施例中，2D感兴趣区域可以使用感兴趣区域的一个或多个第一坐标连同足以形成2D数据平面中的封闭2D体的进一步信息来限定。

[0194] 在图8C中，2D数据帧908A中的左上原点(0,0)被示出为如同指示两个轴线X和Y。其他坐标系当然是可限定的，并且图8A-图8C的X-Y坐标系不是限制性的。然而，使用这样的坐标系准许任何数目的点的高效标识以限定2D数据平面中的封闭2D体。

[0195] 针对第一2D体限定的数据点包括非对称四边形2D感兴趣区域914A的左上坐标914TL、非对称四边形2D感兴趣区域914A的右上坐标914TR、非对称四边形2D感兴趣区域914A的左下坐标914BL、以及非对称四边形2D感兴趣区域914A的右上坐标914TBR。

[0196] 针对另一2D体限定的数据点包括多边形2D感兴趣区域916A的第一坐标916P1、多边形2D感兴趣区域916A的第二坐标916P2、多边形2D感兴趣区域916A的第三坐标916P3、多边形2D感兴趣区域916A的第四坐标916P4、多边形2D感兴趣区域916A的第五坐标916P5、以及多边形2D感兴趣区域916A的第六坐标916P6。

[0197] 图9是与图3-图6的硬件加速的DCNN处理器集成的数据体雕刻器900单元的一个实施例。该实施例以虚线表示数据体雕刻器900单元，以指示数据体雕刻器900的部分可以可选地独立于流交换器500或者与流交换器集成，并且为此，数据体雕刻器900单元还可以与可配置的加速框架400(图4)的其他结构共享特征。数据体雕刻器900的任何特定特征的位置或者一个结构相对于任何其他结构的位置可以适当地由系统半导体从业者布置。

[0198] 数据体雕刻器900单元包括计数器库918、比较器单元920、计算单元922和控制单元924。数据体雕刻器900的其他结构未示出以避免不必要地使数据体雕刻器900的某些特征模糊不清。

[0199] 在图9中特别图示以辅助对数据体雕刻器900单元的讨论的是流交换器500，其可以进一步参考图5和图5的相关联的讨论来理解。输入数据902经由至少一个输入接口被传递到流交换器中。输入数据可以包括向量数据、标量数据、或者某种其他格式的数据。例如，输入数据902可以包括汇总起来形成如本文公开中所描述的特征图的图像数据的流。

[0200] 输出数据904A、904B、904C、904N从流交换器被传递出去。输出数据可以包括特征图数据、三维(3D)特征体数据、被确定为落入3D特征体内的数据、表示3D特征体内的数据的有序数据结构数据、空数据、或者某种其他数据。

[0201] 可选的控制信息通信路径926被包括在图9中。控制信息通信路径926可以用于将控制信息传递到流交换器中。控制信息可以例如是限定2D数据平面中的感兴趣区域的控制信息、与特征图、2D或3D感兴趣区域相关联的尺寸信息、或者用于某种其他目的的信息。

[0202] 计数器库918可以包括任何期望数目的计数器。计数器可以是整数计数器、向上计数计数器、向下计数计数器、移位计数器等。计数器可以包括预定的初始化参数、自动重置功能、自动加载功能、警报功能、中断触发功能等。在一些情况下，一个或多个计数器可以被级联，使得当第一计数器达到阈值时，第二计数器执行计数并且第一计数器重置。以这种方式，例如，一个或多个计数器可以用于产生索引值以访问2D数据平面中的每个数据。此外，在这样的系统中，2D数据平面、3D数据体、或者某种其他结构中的每个个体数据可以基于来自计数器库918的一个或多个计数器的值而被唯一地标识。

[0203] 比较器单元920被布置为将一个数据(例如,数据值)与另一数据或多个其他数据值进行比较的单元。比较器单元920可以包括任何数目的比较器。每个比较器可以被布置为接受向量数据、标量数据或者其他形式的数据。被比较的一些数据可以在易失性或非易失性数据存储库中被存储为恒定数据、可再编程的数据或者某种其他类型的信息。比较器单元的比较器可以被布置为输出单个信号作为比较的结果。单个值可以被断言为高或低、正或负、或者采用某种其他方式。单个值可以基于任何期望的比较结果被断言,诸如,例如,大于、小于、等于等。在一些情况下,不是单个值,而是一个或多个比较器可以被布置为输出表示被比较的第一值与被比较的第二值之间的差的差值。

[0204] 计算单元922可以被布置为处理器、控制器、状态机、或者某种其他这样的计算器。计算单元922可以被布置为执行简单的线性数学运算,诸如加法和减法。另外或者在备选方案中,计算可以被布置为执行更复杂的数学运算,诸如乘法、三角函数、浮点运算等。在一些情况下,计算单元可以被用于计算如参考图8A-图8C的封闭2D体的边界。在一些情况下,计算单元922可以被布置为快速地且高效地确定特征体中的2D数据平面的特定数据是否落入感兴趣区域内。

[0205] 控制单元924可以被布置为确定针对封闭2D数据体上的一个或多个点的坐标。控制单元可以从数据存储库(例如,图4的CAF控制寄存器402)、从输入数据902、从反馈数据、从机器学习算法、或者从某个其他源汲取信息。控制单元还可以被布置为产生在可选的控制信息通信路径926上传递的信息或者产生其他信令信息。

[0206] 如从图9的描述和本文中呈现的结构显而易见的,本文中公开的数据体雕刻器900单元是非常灵活的且可以以许多方式来配置。例如,在一些情况下,数据体雕刻器900单元用于2D感兴趣区域提取,并且在其他的一些情况下,数据体雕刻器单元900用于3D感兴趣区域提取。由于已知在一些情况下,目标检测机器学习算法生成每帧200至300个感兴趣区域,并且由于这些感兴趣区域在一些情况下是最大池化的、平均池化的、并且已经执行了其他操作,因此具有快速灵活的基于硬件的单元提供巨大价值。因为2D和3D特征体提取与使用流交换器的其他数据操作协同地执行,所以这些数据提取/隔离特征提供在任何其他已知方式中不可获得的速度增大、数据处理效率以及功率减少益处。此外,各种配置可以通过将参数保存在诸如CAF控制寄存器402(图4)的数据存储库中、通过机器学习算法的控制、通过嵌入或以其他方式包含于输入数据流中的信息、或者通过其他手段来建立或预先建立。

[0207] 由本文中描述的数据体雕刻器900单元提取的封闭体是从任何数目的源中选择的。在一些情况下,封闭的2D或3D体的特性(例如,大小、形状、取向和其他特征)经由寄存器(例如,图4的CAF控制寄存器402)中的值来编程。在一些情况下,特性被嵌入输入数据中。在另外的其他情况下,特性通过机器学习算法或者通过某种其他手段来确定。在示例性但非限制性的情况下,特性包括特征图的几何形状(例如,高度H、宽度W、深度D)、特征或感兴趣区域的几何形状(例如,高度h、宽度w、深度d)、从其提取特征的输入数据的标识符、是否要从特征图的每个数据平面提取感兴趣区域(即,连续体)或者是否要从特征图的所选择集合的数据平面提取感兴趣区域(即,不连续体)的指示符、以及其他特性。在这些或另外的其他情况下,特性还可以引导用于输出数据的格式。

[0208] 在一些情况下,数据体雕刻器900单元将输出被确定为在所限定的感兴趣区域内的实际特征图数据。在其他的一些情况下,数据体雕刻器900将输出简单地标识被确定为在

所限定的感兴趣区域内或外部的特征图数据的有序数据结构。在一些不同的情况下，数据体雕刻器900将输出特征图的所有数据，但是一些数据将被标记或以其他方式被指示为在所限定的感兴趣区域内，并且剩余的数据将被标记或以其他方式被指示为在所限定的感兴趣区域的外部。在另外的其他情况下，数据体雕刻器900将一对一地输出表示整个特征图的数据，但是在所限定的感兴趣区域内的数据将是实际数据，并且在所限定的感兴趣区域的外部的数据将是空数据(例如，零、常数、确定的空值、或者某种其他所选择的指示符)。

[0209] 现在描述数据体雕刻器900的一些示例性实施例和使用这些实施例的示例性方法。实施例不是限制性的。相反，实施例被描述以向本领域技术人员阐明本文中公开的数据体雕刻器900的灵活性和能力以推进卷积神经网络技术，特别是在分类、预测、或者以其他方式识别视频数据流中的特征(例如，场景、对象、动作等)的实现中。

[0210] 图9中实施的数据体雕刻器900具有至少一个流输入接口和一个或多个流输出接口。在一些实施例中，输入流接口接收一系列2D数据结构作为输入数据902。数据结构可以例如是包括单像素值的图像帧。在卷积神经网络的上下文中，输入帧还可以是网络内的特征数据结构。这些特征数据结构可以包括图像数据，但是它们不一定必需是图像数据或者以任何方式与图像数据相关。

[0211] 在一些情况下，在数据体雕刻器900的输入流接口处接收的数据是“原始”数据流。在一些情况下，原始数据流具有开始标签、停止标签、或者开始标签和停止标签两者。在另外的一些其他情况下，输入数据流是光栅扫描结构，并且输入数据902被布置为被展现有开始指示符、停止指示符、以及类型标识符的个体“行”的序列。在这种上下文中，数据体雕刻器900被布置为针对帧序列中的每个帧“剪裁掉”个体感兴趣区域。这些感兴趣区域可以重叠，但是它们不必重叠。另外，一个或多个数据体雕刻器900单元可以被布置为从每个帧隔离、提取、或以其他方式剪裁掉一个、两个或几十个感兴趣区域。

[0212] 为了执行本文中描述的功能，使数据体雕刻器900单元清楚每个帧的几何形状(例如，尺寸)并且清楚待提取的感兴趣区域的几何形状(例如，位置、大小、取向等)。每个输入帧的几何形状可以在输入数据流902嵌入或以其他方式(例如，以光栅扫描方式)包括这样的数据的情况下被自动提取。备选地，例如，如果输入数据流90仅2包括原始数据帧，则几何形状可以被预编程到配置寄存器(例如，控制单元924，图4中的CAF控制寄存器402等)，被编程为机器学习算法的部分，或者以其他方式被散播到数据体雕刻器900单元。

[0213] 关于数据体雕刻器900单元的灵活性，应意识到，提取或以其他方式处理2D和3D数据体的许多卷积神经网络操作将提取在一个帧中具有一种几何形状并且在另一帧中具有不同几何形状的体。因此，数据体雕刻器900足够灵活以准许被提取的感兴趣区域的几何形状针对每个输入帧而变化或者针对比所有输入帧少的多个输入帧而变化。

[0214] 在矩形感兴趣区域(例如，图8A，矩形感兴趣区域910A、910B、910N)的示例性情况下，提供至少两个二维(2D)坐标。尽管本文中描述的数据体雕刻器900单元可以从任何形状的感兴趣区域中提取数据，但是现在描述矩形感兴趣区域以简化本公开。

[0215] 在这种情况下，两个坐标对表示特征图的第一2D数据平面中的矩形2D感兴趣区域的左上坐标910TL和右下坐标910BR。对于每个新的数据帧(例如，特征图的第二和第N个2D数据平面中的矩形2D感兴趣区域910B、910N)，左上坐标和右下坐标的更新的对必须被接收、加载给数据体雕刻器900、或以其他方式对数据体雕刻器900已知。这些更新的坐标对区

分标准“剪裁”功能，在标准“剪裁”功能中提取的区域不能在不同帧之间变化并且这样的不同的感兴趣区域不被链接在一起以形成3D体。

[0216] 在一些情况下，机器学习算法将实例化多个输出数据流904A、904B、904C、904N。在这些情况中的一些中，具有提取的一个感兴趣区域或多个感兴趣区域的相同数据流可以被传递通过多个输出接口。在这些情况中的其他一些情况中，具有不同几何形状的感兴趣区域被传递通过多个输出接口。

[0217] 考虑矩形2D感兴趣区域情况的当前示例性情况，计数器库918的两个计数器可以用于跟踪在输入接口处的输入数据流902中接收的当前输入像素的实际位置。为了当前示例性情况，两个计数器中的第一计数器被称为“x_cnt”并且两个计数器中的第二计数器被称为“y_cnt”。对于处理的每个输入像素，x_cnt值可以递增直到达到相关联的“行”的大小。相关联的行的大小可以从内部配置寄存器(例如，控制单元924、图4中的CAF控制寄存器402等)、通过检索输入数据中的行标签、如由机器学习算法引导的、或者通过某种其他手段得到。在处理每个输入像素时，当确定达到行的结尾时，则x_cnt计数器被重置为零，并且y_cnt计数器递增一。在处理了每个行的每个像素之后，当确定达到帧的结尾时，则x_cnt计数器和y_cnt计数器两者均被重置为零，并且可以开始针对特征图中的下一2D帧的处理。

[0218] 当处理每一个像素时，考虑中的像素的坐标被分析。本示例中的分析可以包括将x_cnt计数器值和y_cnt计数器值与针对当前帧的所确定的感兴趣区域的“左上TLx,TLy”和“右下BRx,BRy”角坐标进行比较。如果方程1证明为真，则像素被确定为在感兴趣区域内。否则，如果方程1证明为假，则像素被确定为在感兴趣区域外部。

[0219] $TLx(N) \geq x\_cnt \geq BRx(N) \text{ AND } TLy(N) \geq y\_cnt \geq BRy(N)$ (1)

[0220] 在一些情况下，当像素被确定为在感兴趣区域内时，像素被转发到输出接口作为感兴趣区域内的有效数据。在这些情况下，像素可以被标记为或以某种其他方式被标识为在感兴趣区域内，或者可以不被标记为或不以某种其他方式被标识为在感兴趣区域内。在其他的一些情况下，有序结构被创建或更新以包括指示像素在感兴趣区域内的信息。

[0221] 在一些情况下，当像素被确定为在感兴趣区域外部时，像素被简单地丢弃。在其他的一些情况下，像素被转发到输出接口，但是被标记为或以某种其他方式被标识为在感兴趣区域外部。在其他的一些情况下，空数据(其可以为零值、常数或空数据的任何其他表示)被转发到输出接口。在另外的其他情况下，相同的有序结构被更新，或者不同有序结构被创建，以包括指示像素在感兴趣区域外部的信息。

[0222] 鉴于方程1和刚刚描述的处理，进一步意识到数据体雕刻器900单元的灵活性。在这些情况下，可选的控制信息通信路径926可以被断言、加载、或以其他方式用于传递关于像素的状态(如在感兴趣区域内或外部)的信息。

[0223] 在本当前示例中，数据体雕刻器900针对每个帧分析感兴趣区域数据，并且感兴趣区域几何形状可以随着每个帧改变或者可以不改变。感兴趣区域几何形状(其可以包括至少一个第一坐标和足以形成帧内的封闭2D体的附加信息)针对每个帧被更新或是可更新的。感兴趣区域几何形状信息可以被本地存储在控制单元924中、计数器库918的初始化参数中、或者与数据体雕刻器900相关联的某个其他区中。备选地或另外，这样的几何参数或者相关联的数据可以在输入数据902中被传递，并且经由计算单元922、控制单元924或某种其他手段被检索或被以其他方式计算(例如，线性数学、射线投射算法、积分方程等)。另外，

可以限定的是机器学习算法的一些外部单元或部分与帧几何形状一起或者与帧几何形状分离地提供坐标信息(例如,角坐标、区域中心坐标、多个顺序点坐标等)。这些参数可以在运行中(即,在运行时)被提供、提前(例如,在构建时、在初始化时等)被提供、作为预编程的信息被提供、作为硬件确定的信被提供息、或者通过某种其他手段被提供。

[0224] 在本示例的一些情况下,无论感兴趣区域是矩形、某种其他多边形、曲线特征、或者某种其他复杂的感兴趣区域,输出帧都可以再次变成矩形区域。如果是这种情况下,例如,则输出特征的尺寸可以针对被确定为在多边形感兴趣区域中的每个像素根据方程2来设置。

[0225] $(MAX(x_{coord})-MIN(x_{coord}))\times(MAX(y_{coord})-MIN(y_{coord}))$ (2)

[0226] 在除了本公开中描述的那些之外的任何其他已知的设备、系统或方法中未提供灵活的数据体雕刻器900的优点。至多,限定特征图中的2D或3D体的功能松散地利用经由源地址、目的地地址、源步幅、行数和按字节的行宽提供必要的信息的一个或多个直接存储器访问(DMA)引擎来完成。这些解决方案,除了非常不同于本公开的集成的数据体雕刻器900单元之外,要求在存储器中加载并存储特征体和提取的感兴趣区域以及由主机微控制器的介入以生成链接的寻址信息。因此,任何先前的变通方案不能够在不暂时将大量数据存储到存储器并从存储器中检索大量数据的情况下从可以产生感兴趣区域信息和特征体的在前单元直接与特征体数据同时地流式传输感兴趣区域信息。自然,对于本领域技术人员清楚的是,这些变通方案是更慢的、更低效的且功耗很高。

[0227] 图10是图示至少一种数据体雕刻方法的数据流程图。在一些情况下,方法被布置为生成限定3D特征图内的三维(3D)体的有序数据结构。在其他的一些情况下,方法被布置为处理特征图的二维(2D)数据平面中的每个数据并输出被确定为处于所确定的感兴趣区域内的每个数据。为了示例起见,本文中在处理被称为特征图的3D数据块/体的卷积神经网络的上下文中描述图10的方法,这包括被堆叠在一起以形成3D“特征图”的被称为“特征”的个体2D数据平面。这里,如图9中图示的且本文描述的数据体雕刻器900单元被用于从输入特征图内隔离、“雕刻”或者以其他方式提取所确定的“体”。

[0228] 一些已知的卷积神经网络应用要求访问任意的感兴趣区域,其可以甚至限制于单个平面(即,2D)感兴趣区域。其他已知的应用,例如T-CNN论文中公开的应用,要求访问任意的3D感兴趣区域。本文描述的数据体雕刻器900单元以没有其他结构或方法可以的方式提供这些功能。本文描述的数据体雕刻器900单元被布置为一个或多个硬件块,其可以从存储于存储器中、从图像或其他传感器直接流式传输的、从卷积加速器600(图6)传递的或者从某个其他源被传递到数据体雕刻器900中的现有特征图提取确定的3D体感兴趣区域。

[0229] 图10的示例性方法利用如图9中描绘的示例性数据体雕刻器900来执行。

[0230] 在至少一个情况下,图10的方法在移动设备100(图3)的集成电路中被执行。集成电路可以被形成为片上系统110(图3)或者以一些其他封装、裸片、芯片、控制器、计算设备等被形成。集成电路包括板载随机存取存储器(RAM)126(图3)、应用处理器128(图3)、数字信号处理器(DSP)集群138(图3)、可配置的加速器框架(CAF)400(图3、图4)、以及将应用处理器128、DSP集群138、和CAF 400通信地耦合到RAM 126的至少一个通信总线架构166、132(图3)。CAF 400在至少一个情况下包括可重新配置的流交换器500(图4、图5)以及具有耦合到可重配置的流交换器500的输入和耦合到可重配置的流交换器500的输出(图9)的数据体

雕刻单元900(图4、图9)。

[0231] 方法中的数据体雕刻器900单元除其他结构外具有计数器库918、比较器单元920、和控制单元924。数据体雕刻器900被形成为硬件块,硬件块执行并加速从处于卷积神经网络深度机器学习算法中的处理下的给定特征图中对具有确定的几何形状的3D体的提取。数据体雕刻器900在其输入处接收输入特征图,并且数据体雕刻器900在其输出处产生表示一个或多个提取的3D体的信息。也被提供到数据体雕刻器900单元的是针对特征图中的每个2D数据平面的一组配置参数。参数包括以下中的任何一个或多个:特征图的几何形状(例如,高度、宽度和深度(H,W,D)),限定感兴趣区域的几何形状的一系列值(例如,与被指定为包括感兴趣区域标识符元组的特征内的每个感兴趣区域的左上和右下坐标相对应的链表、对应于感兴趣区域的特定点的一个或多个坐标等),视情况而可能标识特征或特征图的开始和结束的参数(例如,形成连续3D体的独立开始和结束索引,或者实现对不连续3D体的提取的开始和结束索引的链表),以及任何其他参数。在3D体的情况下,数据体雕刻器900的输出可以包括感兴趣区域标识符,以将个体2D感兴趣区域“串在一起”、连结、或者以其他方式关联以形成期望的提取的输出3D体。

[0232] 图10的数据体雕刻方法1000包括数据体雕刻器900的动作,并且处理在1002处开始。

[0233] 在1004处,接收特征图数据的流。在一些情况下,特征图数据的流被形成为三维(3D)特征图,并且3D特征图被形成为多个二维(2D)数据平面。

[0234] 处理继续到1006,其中接收限定感兴趣区域的输入信息。输入信息可以被限定在2D数据平面中。在一些情况下,输入信息包括感兴趣区域的至少一个第一坐标和足以形成2D数据平面中的封闭2D体的另外信息。在接收感兴趣区域信息后,一个或多个计数器可以被加载并按顺序排好,使得2D数据平面中的每个数据以所选择的顺序被分析。在一些情况下,多个封闭2D体被限定在3D特征的每个2D数据平面中。在一些情况下,所选择的2D数据平面中的多个封闭2D体中的某些封闭2D体是非重叠的,并且在其他的一些情况下,所选择的2D数据平面中的2D体中的某些2D体确实重叠。

[0235] 在1008处,标识3D特征图内的3D体。3D体在尺寸上小于3D特征图。

[0236] 并且在1010处,处理继续。这里,来自3D特征图的在3D体内的数据被隔离以用于在深度学习算法中进行处理。隔离数据的动作可以采用比较器单元920,其被布置为确定所分析的每个数据是否在封闭2D体内。

[0237] 在隔离期间或之后,在1012处,或者在某个其他时间,被确定为在封闭2D体内的每个数据或与之相关联的信息可以从数据体雕刻器900被输出。在其他的一些情况下,不是输出数据,数据体雕刻器900单元可以代替地生成并输出包括被确定为在封闭2D体内的每个数据的有序数据结构。这里,在3D特征图的多个封闭2D体内的数据可以被包括在有序数据结构中。每个2D数据平面可以由其相应的封闭2D体限定,并且每个相应的封闭2D体可以在有序结构中与在相邻2D数据平面中限定的至少一个其他封闭2D体相关联。

[0238] 图10的方法中的处理在1014处结束。

[0239] 考虑图10的方法的另一实现,在1004处的处理包括接收在数据体雕刻900单元的输入流接口处的信息。信息包括一系列帧,并且每个帧被形成为二维(2D)数据结构。在一些情况下,该系列的2D数据结构包括由单像素值组成的图像帧。备选地在一些情况下或另外,

该系列的二维数据结构包括卷积神经网络内的非图像特征数据结构。一系列的帧可以被接收作为具有开始标签和停止标签的原始数据流。备选地或另外，该系列的帧可以被接收作为光栅扫描结构，其中光栅扫描结构的每个个体行被展现有开始标签、停止标签和类型标识符。

[0240] 在1006处，确定帧中的每个帧的第一维和第二维，并且基于第一维和第二维，针对每个帧确定要从相应帧提取的感兴趣区域的位置和大小。有时，确定要从每个帧提取的多个感兴趣区域。并且在这些情况下，要从每个帧提取的多个感兴趣区域中的感兴趣区域可以是重叠的或非重叠的。数据体雕刻器900可以在一些情况下使用在输入流接口处接收的信息来自动地从每个帧提取感兴趣区域的位置和大小。备选地或另外，数据体雕刻器900可以从参数存储库检索感兴趣区域的位置和大小。要从第一帧提取的感兴趣区域的位置和大小中的至少一项可以与要从第二帧提取的感兴趣区域的对应的位置或大小不同。

[0241] 在一些情况下，在1008处，分析二维坐标的对，以确定要从相应帧提取的感兴趣区域的位置和大小。这里，有时，二维坐标的对包括要从相应帧提取的感兴趣区域的左上坐标和右下坐标。

[0242] 在一些情况下，在1008处，分析单个点和围绕该单个点的半径，以确定要从相应帧提取的感兴趣区域的位置和大小。

[0243] 在另外的其他情况下，在1008处，分析限定多边形的多个点，以确定要从相应帧提取的感兴趣区域的位置和大小。这里，或者在其他的一些情况下，确定要从相应帧提取的感兴趣区域的位置和大小包括分析多个点和多个点中的至少两个点之间的距离。

[0244] 在1010和1012处的处理包括从每个帧中提取帧中的在感兴趣区域内的数据。为了实现该动作，提取包括：1)针对每个帧中的在要从相应帧提取的相应感兴趣区域外部的每个数据，将空数据传递通过数据体雕刻单元的输出接口；以及2)针对每个帧中的在要从相应帧提取的相应感兴趣区域内的每个数据，将数据传递通过数据体雕刻单元的输出接口。

[0245] 在一些情况下，在1010处，数据的隔离(即，提取)包括初始化第一计数器和第二计数器。第一计数器和第二计数器被布置为跟踪在输入流接口处接收的帧的每个数据的位置。跟踪每个数据的位置在这种情况下包括使用来自第一计数器和第二计数器的计数值作为帧内的数据的坐标。隔离还包括将数据的坐标与限定感兴趣区域的界限值进行比较并且从该比较来确定数据是在感兴趣区域外部还是在感兴趣区域内。

[0246] 在另外的其他情况下，在1010和1012处，例如在确定要从相应帧提取的多个感兴趣区域的情况下，数据的隔离和输出可以包括针对多个感兴趣区域中的每个感兴趣区域将分离的且不同的空数据或帧数据同时从数据体雕刻单元传递出去。在这些情况中的一些中，将空数据传递通过数据体雕刻单元的输出接口通过传递来自帧的数据并断言指示数据在要从相应帧提取的相应感兴趣区域外部的信号来执行。

[0247] 考虑图10的方法的又一实现，在1010处的处理包括将从被形成于集成电路中的可重配置流交换器500传递的流式数据接收到数据体雕刻器900单元中。集成电路可以特别地被布置用于卷积神经网络操作。流式数据限定被形成为一系列二维(2D)数据平面的三维(3D)特征图。在一些情况下，3D特征图包括卷积神经网络中处于分析的图像数据。3D特征图的几何形状可以由高度、宽度和深度(H,W,D)限定。

[0248] 在1010处，数据体雕刻器900单元生成限定3D特征图内的3D体的有序数据结构。3D

体在尺寸上小于3D特征图。所雕刻的3D体的几何形状可以由高度、宽度和深度(h,w,d)限定。生成有序数据结构在一些情况下可以包括形成具有与二维(2D)感兴趣区域的坐标相对应的一系列值的至少一个链表。在其他的一些情况下,生成有序数据结构可以包括形成至少一个元组。至少一个元组可以包括感兴趣区域标识符和至少一个链表或其对应于2D感兴趣区域的部分。在另外的其他情况下,生成有序数据结构可以包括:选择与特征图的第一2D数据平面相对应的开始索引;选择与特征图的最后2D数据平面相对应的结束索引;以及将所选择的开始索引和结束索引与至少一个链表一起包括在有序数据结构中,使得3D体被限定在特征图的第个2D数据平面与最后2D数据平面之间。并且在这些情况中的一些情况下,生成有序数据结构可以包括基于共同的感兴趣区域标识符,将开始索引与结束索引之间的一系列2D数据平面关联在一起。

[0249] 有时,在1010处,在数据体雕刻器900单元通过形成具有与二维(2D)感兴趣区域的坐标相对应的一系列值的至少一个链表来生成有序数据结构的情况下,坐标包括2D感兴趣区域的左上坐标和右下坐标。在2D感兴趣区域是圆形的情况下,2D感兴趣区域的坐标可以包括对应于单个点的坐标和围绕该单个点的半径。在2D感兴趣区域可以是多边形的情况下,2D感兴趣区域的坐标可以包括与限定多边形的多个点相对应的点坐标。并且在2D感兴趣区域包括至少一个曲线的情况下,2D感兴趣区域的坐标可以包括对应于多个点的坐标和多个点中的至少两个点之间的距离。

[0250] 在1012处,有序数据结构一旦被生成就被传递通过可重配置流交换器500。

[0251] 本公开涉及“半导体从业者”。半导体从业者一般是半导体设计和制造领域中的普通技术人员。半导体从业者可以是学位工程师或具有这样的技术以便指导并平衡半导体制造项目的特定特征(诸如几何形状、布局、功率使用、包括的知识产权(IP)模块等)的另一技术人员或系统。半导体从业者可以理解或者可以不理解被执行以形成裸片、集成电路或其他这样的设备的制造过程的每个细节。

[0252] 图10是可以由移动计算设备100的实施例使用的多个非限制性过程。在此方面,每个描述的过程可以表示包括用于实现(一个或多个)指定的逻辑功能的一个或多个可执行指令的软件代码的模块、片段或部分。还应当指出,在一些实现中,过程中指出的功能可以以不同的顺序进行,可以包括附加的功能,可以并发地进行,和/或可以被省略。

[0253] 本公开中的图图示一个或多个非限制性计算设备实施例(诸如移动设备100)的部分。计算设备可以包括在常规计算设备装置中发现的操作性硬件,诸如一个或多个处理器、易失性和非易失性存储器、遵从各种标准和协议的串行和并行输入/输出(I/O)电路、有线和/或无线联网电路(例如,通信收发器)、一个或多个用户界面(UI)模块、逻辑、以及其他电子电路。

[0254] 除其他外,本公开的示例性移动设备(例如,图3的移动设备100)可以以任何类型的移动计算设备来配置,诸如智能电话、平板计算机、膝上型计算机、可穿戴设备(例如,眼镜、外套、衬衫、裤子、袜子、鞋、其他衣服、帽子、头盔、其他头戴件、腕表、项链、吊坠、其他首饰)、车辆安装的设备(例如,火车、飞机、直升机、无人飞行器、无人潜水器、无人地面车辆、汽车、摩托车、自行车、小型摩托车、悬滑板、其他个人或商用交通工具)、工业设备等。因此,移动设备包括未图示的其他部件和电路,诸如,例如,显示器、网络接口、存储器、一个或多个中央处理器、相机接口、音频接口、以及其他输入/输出接口。在一些情况下,示例性移动

设备还可以以不同类型的低功率设备来配置，诸如头戴式视频相机、物联网(IoT)设备、多媒体设备、移动检测设备、入侵者检测设备、安全设备、人群监控设备、或者某种其他设备。

[0255] 如本文所描述的处理器包括中央处理单元(CPU)、微处理器、微控制器(MCU)、数字信号处理器(DSP)、专用集成电路(ASIC)、状态机等。因此，如本文所描述的处理器包括控制至少一个操作的任何设备、系统或其部分，并且这样的设备可以以硬件、固件或软件、或者这些中的至少两个的某种组合来实现。与任何特定处理器相关联的功能可以是集中或分布的，无论是本地还是远程。处理器可以可互换地指代被配置为执行编程的软件指令的任何类型的电子控制电路。编程的指令可以是高级软件指令、编译的软件指令、汇编语言软件指令、目标代码、二进制代码、微代码等。编程的指令可以驻留在内部或外部存储器中或者可以被硬编码为状态机或控制信号的集合。根据本文提及的方法和设备，一个或多个实施例描述可由处理器执行的软件，其当被执行时执行方法动作中的一个或多个。

[0256] 在一些情况下，本公开中描述的一个处理器或多个处理器以及附加地本公开中描述的示例性移动设备的更多或更少的电路可以被提供在集成电路中。在一些实施例中，在本文附图的处理器中示出的元件中的全部(例如，SoC 110)可以被提供在集成电路中。在备选实施例中，在本文附图中描绘的布置中的一个或多个(例如，SoC 110)可以由两个或更多个集成电路提供。一些实施例可以由一个或多个裸片实现。一个或多个裸片可以被封装在相同或不同的封装中。所描绘的部件中的一些可以被提供在集成电路或裸片的外部。

[0257] 本文附图中示出的和本文描述的处理器可以在设计时在拓扑、最大可用带宽、每单位时间的最大可用操作、最大并行执行单元、以及其他这样的参数中的一项或多项方面被固定。处理器的一些实施例可以提供在运行时的可再编程功能(例如，用于实现DCNN的SoC模块和特征的重新配置)。可再编程功能中的一些或全部可以在一个或多个初始化阶段被配置。可再编程功能中的一些或全部可以被配置为在运行中没有延时、具有可屏蔽的延时、或者可接受的延时水平。

[0258] 如本领域技术人员已知的，如本公开中描述的计算设备以及是这样的计算设备的移动设备100具有一个或多个存储器，并且每个存储器包括用于读和写的易失性和非易失性计算机可读介质的任何组合。易失性计算机可读介质包括例如随机存取存储器(RAM)。非易失性计算机可读介质包括例如只读存储器(ROM)、诸如硬盘的磁性介质、光盘、闪存设备等。在一些情况下，特定存储器被虚拟地或物理地分离成单独的区，诸如第一存储器、第二存储器、第三存储器等。在这些情况下，应理解，存储器的不同划分可以处于不同设备中或者被实施在单个存储器中。存储器在一些情况下是非暂态计算机介质，其被配置为存储被布置为由处理器执行的软件指令。

[0259] 在本公开中，存储器可以以一种配置或另一种配置来使用。存储器可以被配置为存储数据。在备选方案中或者另外，存储器可以是非暂态计算机可读介质(CRM)，其中CRM被配置为存储可由处理器执行的指令。指令可以被单独地存储或者被存储为文件中的成组的指令。文件可以包括函数、服务、库等。文件可以包括一个或多个计算机程序或者可以为较大计算机程序的部分。备选地或另外，每个文件可以包括对执行本公开中描述的系统、方法和装置的计算功能有用的数据或者其他计算支持材料。

[0260] 本文中图示和描述的计算设备(计算设备100是其中的一个示例)还可以包括在常规计算设备中找到的操作性软件，诸如操作系统或任务循环、用于引导通过I/O电路的操作

的软件驱动器、联网电路、以及其他外围部件电路。另外，计算设备可以包括操作性应用软件，诸如用于与其他计算设备进行通信的网络软件、用于构建和维护数据库的数据库软件、以及在合适的情况下用于将通信和/或操作工作负荷分布在各种处理器间的任务管理软件。在一些情况下，计算设备是具有本文列出的硬件和软件中的至少一些的单个硬件机器，并且在其他的一些情况下，计算设备是在服务器场中一起工作以执行本文描述的一个或多个实施例的功能的硬件和软件机器的联网汇集。计算设备的常规硬件和软件的一些方面为简单而未在图中示出，但是技术实践人员容易理解。

[0261] 当如本文中所描述的那样被布置时，每个计算设备可以从通用的且非特定的计算设备转变为包括被配置用于特定且具体的目的的硬件和软件的组合设备。按照这些原则，组合设备的特征带来对技术计算领域至今未见的且未知的改进。

[0262] 如果在本文描述的移动设备或支持网络设备中存在任何数据结构，则数据结构可以以单个数据库或多个数据库被形成。在一些情况下，硬件或软件存储库在它们与之相关联的一个或多个特定系统的各种功能之间被共享。数据库可以被形成为局部系统或局域网的部分。备选地或另外，数据库可以被远程地形成，例如形成在“云”计算系统内，其将能够经由广域网或某种其他网络访问。

[0263] 在至少一个实施例中，本文描述的移动设备可以经由网络上的通信与其他设备通信。网络可以涉及互联网连接或某种其他类型的局域网(LAN)或广域网(WAN)。实现或形成网络的部分的结构的非限制性示例包括但不限于以太网、双绞线以太网、数据用户环路(DSL)设备、无线LAN、WiFi、基于蜂窝的网络等。

[0264] 按钮、小键盘、计算机鼠标、存储器卡、串行端口、生物传感器读取器、触摸屏等可以单独地或协同地有用于移动设备或如本文描述的其他这样的设备的操作者。设备可以例如将控制信息输入到系统中。显示器、打印机、存储器卡、LED指示器、温度传感器、音频设备(例如，扬声器、压电设备等)、振动器等全部有用于将输出信息呈现给这些移动设备的操作者。在一些情况下，输入和输出设备被直接耦合到本文描述的控制系统并且被电子地耦合到处理器或其他操作性电路。在其他的一些情况下，输入和输出设备经由一个或多个通信端口(例如，RS-232、RS-485、红外、USB等)传递信息。

[0265] 除非另行限定，否则本文使用的技术和科学术语具有与如本实用新型所属领域的普通技术人员通常理解的相同的意义。尽管与本文中描述的方法和材料相似或等价的任何方法和材料也可以在实践或测试本实用新型中被使用，但是本文描述了有限数目的示例性方法和材料。

[0266] 在前面的描述中，阐述了某些具体细节以提供对各种所公开的实施例的透彻理解。然而，相关领域技术人员将意识到，实施例可以在没有这些具体细节中的一个或多个的情况下或者利用其他方法、部件、材料等来实践。在其他的一些实例中，未详细示出或描述与包括客户端和服务器计算系统的电子和计算系统以及网络相关联的公知结构，以避免不必要地使实施例的描述模糊不清。

[0267] 除非上下文另行要求，否则在说明书和随附的权利要求书中，词语“包括”及其变型，例如“包含”和“具有”应在开放式的包含性的意义上来理解，例如，“包括但不限于”。

[0268] 贯穿本说明书对“一个实施例”或“实施例”及其变型的引用意味着结合实施例描述的特定特征、结构或特性被包含在至少一个实施例中。因此，贯穿本说明书在各个地方中

出现的短语“在一个实施例中”或“在实施例中”不一定全部指代同一实施例。另外，在一个或多个实施例中可以以任何适当的方式来组合特定特征、结构和特性。

[0269] 如在本说明书和随附权利要求书中所使用的，单数形式的“一”、“一个”和“所述”包括复数指代，除非内容和上下文另行清楚指示。还应当指出，连接术语“和”和“或”一般在最宽泛的意义上被用于包括“和/或”，除非内容和上下文清楚地指示包含性或排他性(视情况而定)。另外，“和”和“或”的组成当在本文中被记载为“和/或”时旨在涵盖包括所有相关联的项或构思的实施例以及包括少于所有相关联的项或构思的一个或多个其他备选实施例。

[0270] 在本公开中，连接列表利用逗号，其可以被已知为牛津逗号、哈佛逗号、连续逗号、或者另一类似术语。这样的列表旨在连接词语、子句或句子，使得逗号之后的内容也被包括在列表中。

[0271] 本文中提供的公开内容的标题和摘要仅是为了方便起见并且不限制或解释实施例的范围或含义。

[0272] 以上描述的各种实施例可以被组合以提供另外的实施例。实施例的方面可以在必要时被修改，以采用各种专利、申请和公布的构思来提供更进一步的实施例。

[0273] 可以鉴于上述具体实施方式对实施例进行这些和其他改变。总体上，在随附权利要求书中，使用的术语不应当被理解为将权利要求限制于本说明书和权利要求书中公开的具体实施例，而是应当被理解为包括所有可能的实施例连同这样的权利要求被授予的等价方案的完整范围。因此，权利要求不受本公开内容限制。

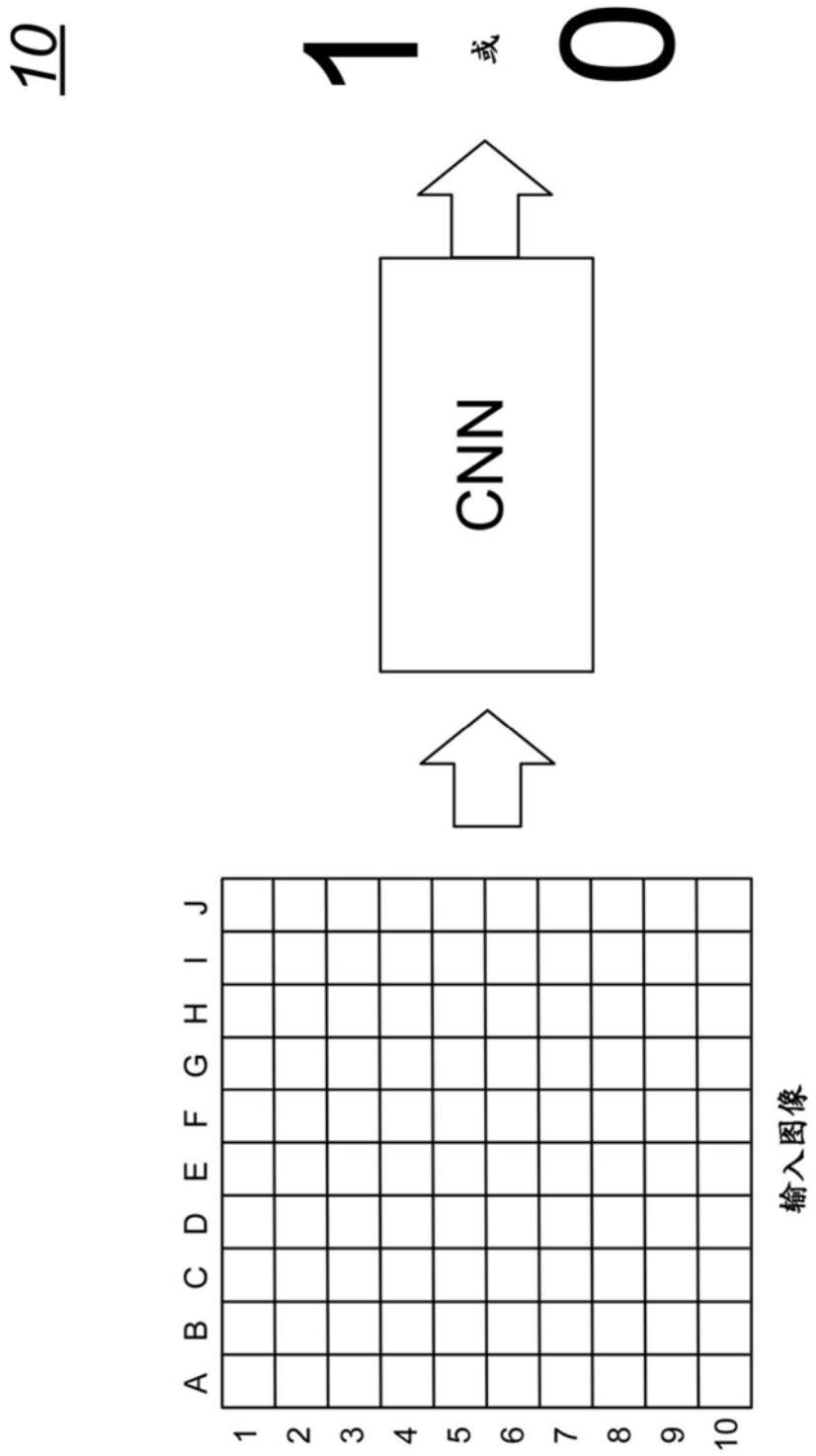


图1A

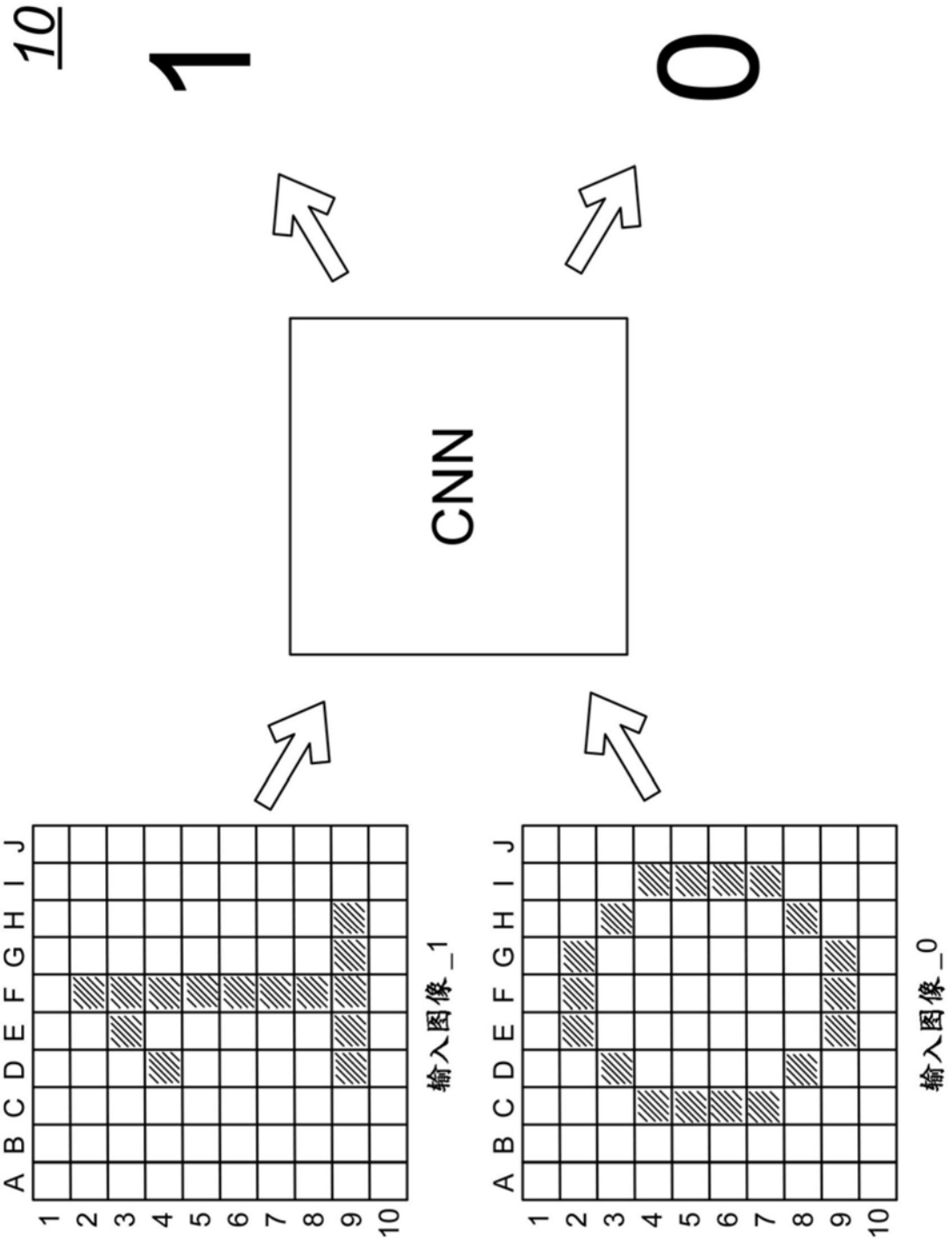


图1B

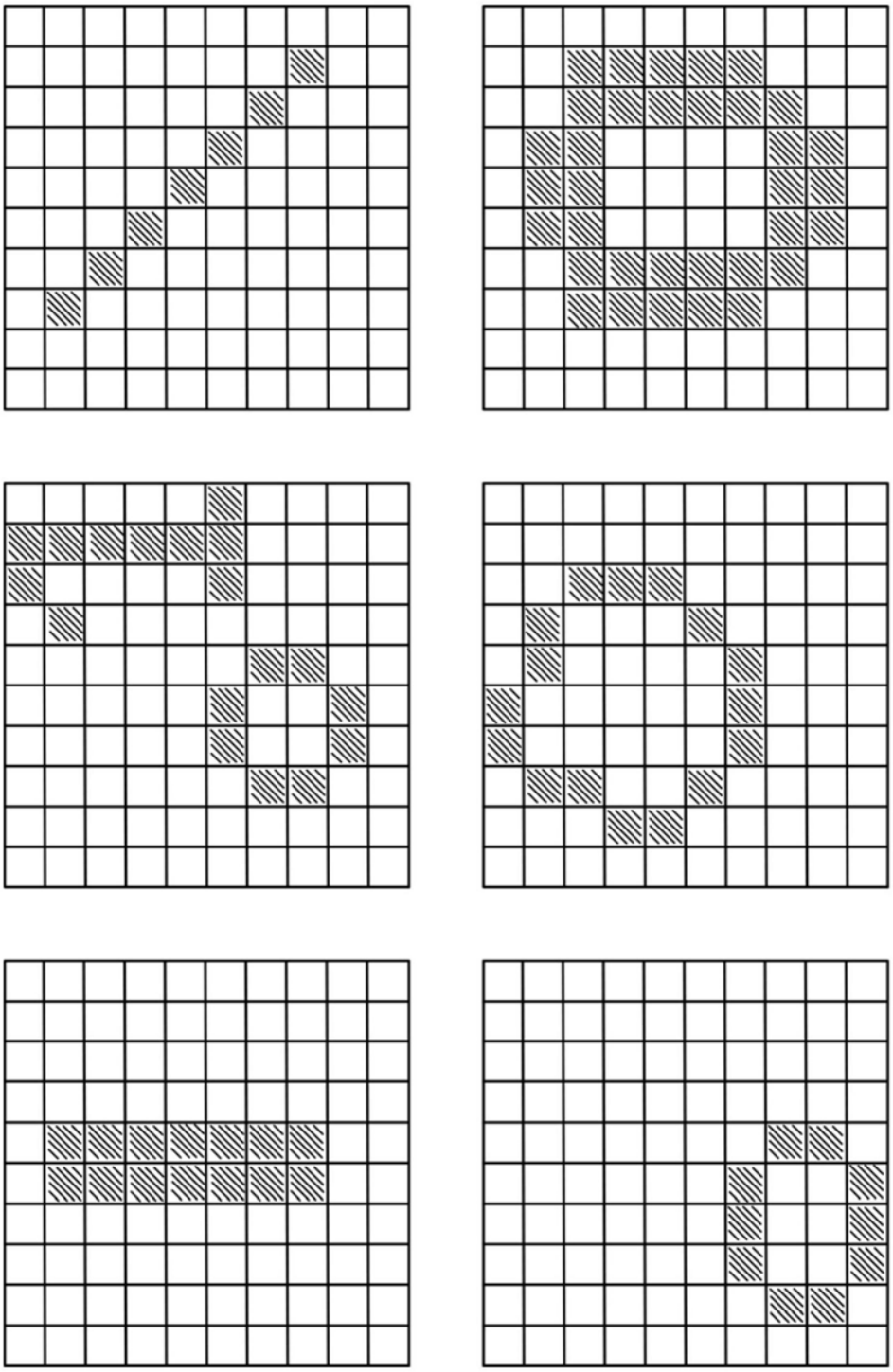


图1C

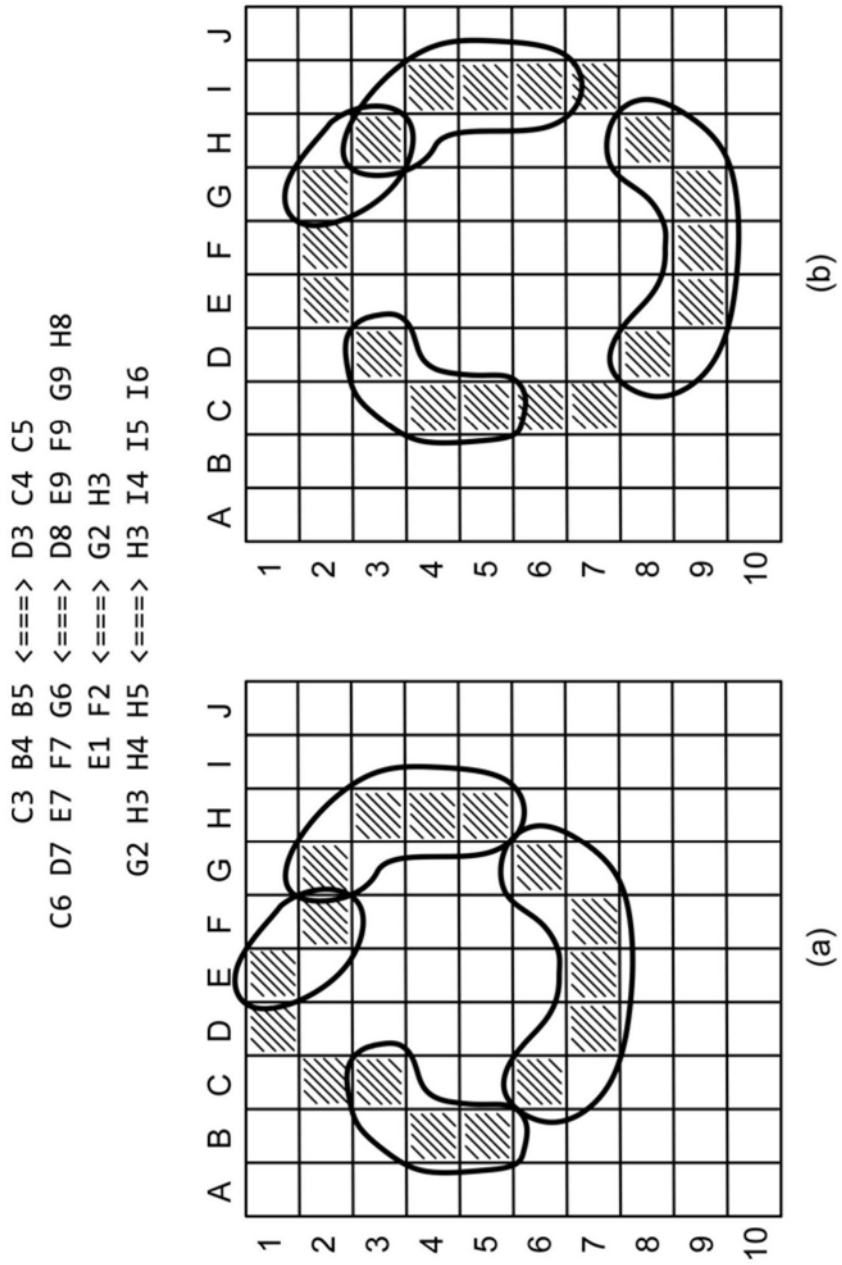

C3 B4 B5 <===> D3 C4 C5
C6 D7 E7 F7 G6 <===> D8 E9 F9 G9 H8
E1 F2 <===> G2 H3
G2 H3 H4 H5 <===> H3 I4 I5 I6
A B C D E F G H I J
1 2 3 4 5 6 7 8 9 10
(a)
(b)

图1D

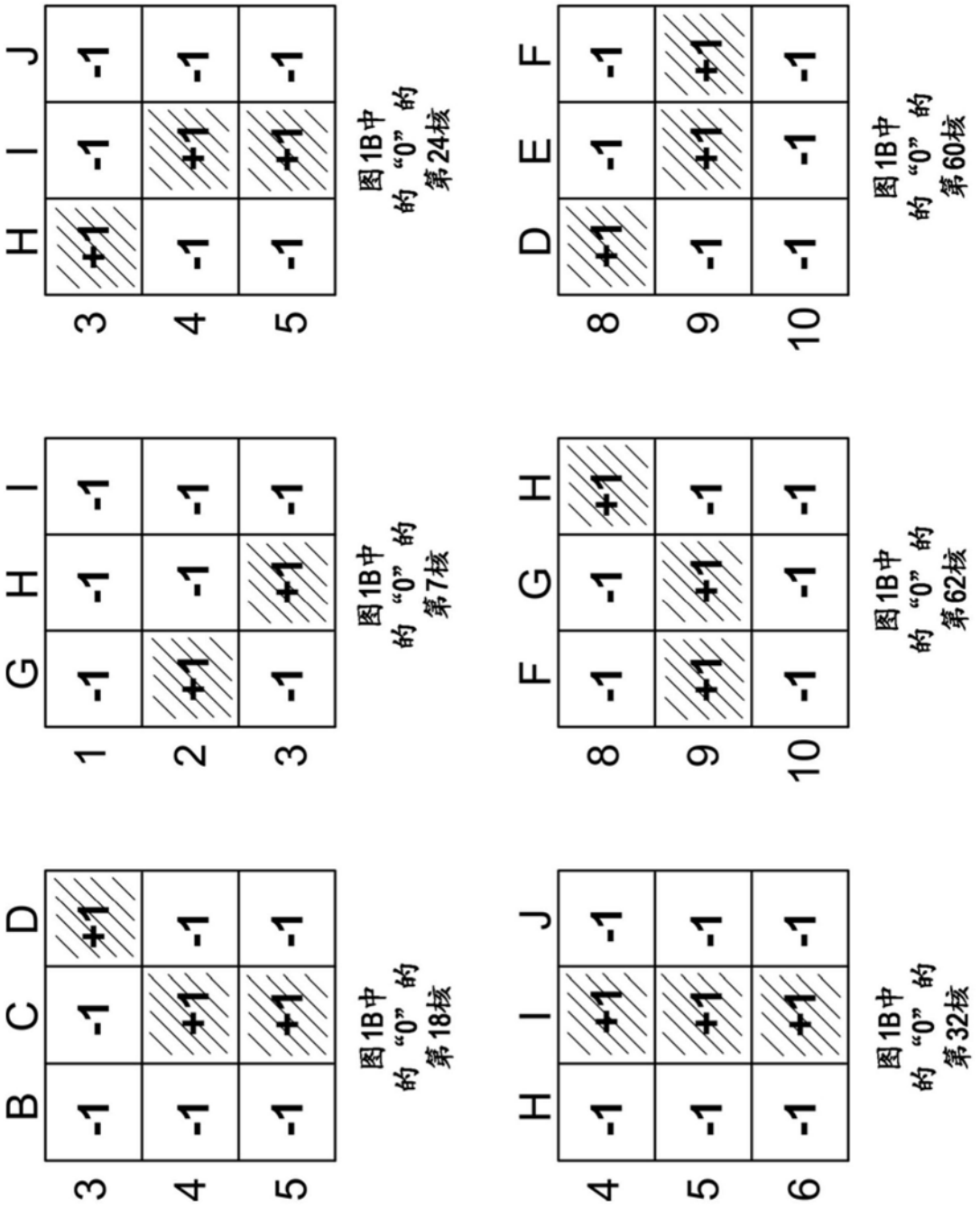

H I J
3 +1 -1 -1
4 -1 +1 -1
5 -1 +1 -1
图1B中的“0”的第24核
D E F
8 +1 -1 -1
9 -1 +1 +1
10 -1 -1 -1
图1B中的“0”的第60核
G H I
1 -1 -1 -1
2 +1 -1 -1
3 -1 +1 -1
图1B中的“0”的第7核
F G H
8 -1 -1 +1
9 +1 +1 -1
10 -1 -1 -1
图1B中的“0”的第62核
B C D
3 -1 -1 +1
4 -1 +1 -1
5 -1 +1 -1
图1B中的“0”的第18核
H I J
4 -1 +1 -1
5 -1 +1 -1
6 -1 +1 -1
图1B中的“0”的第32核

图1E

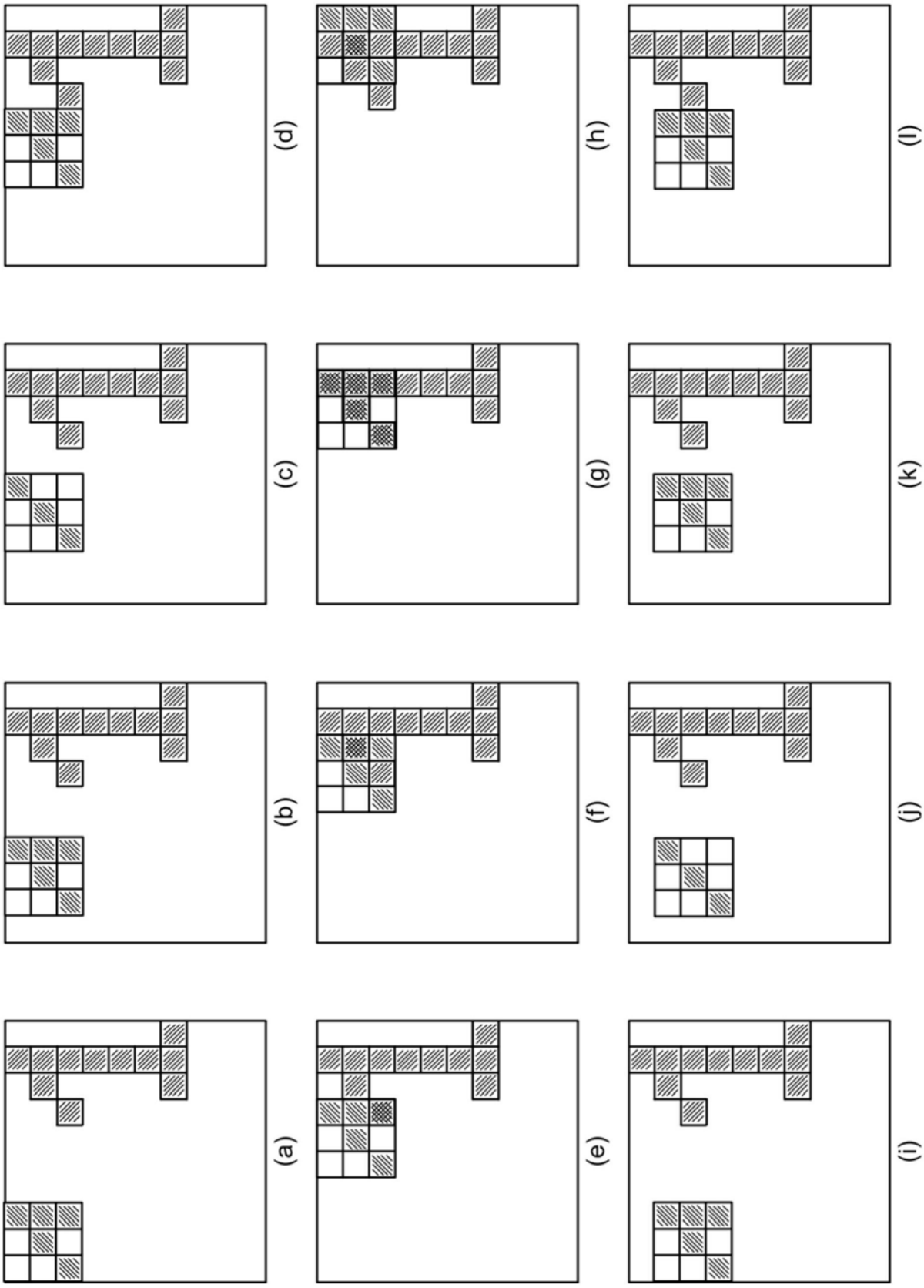


图1F

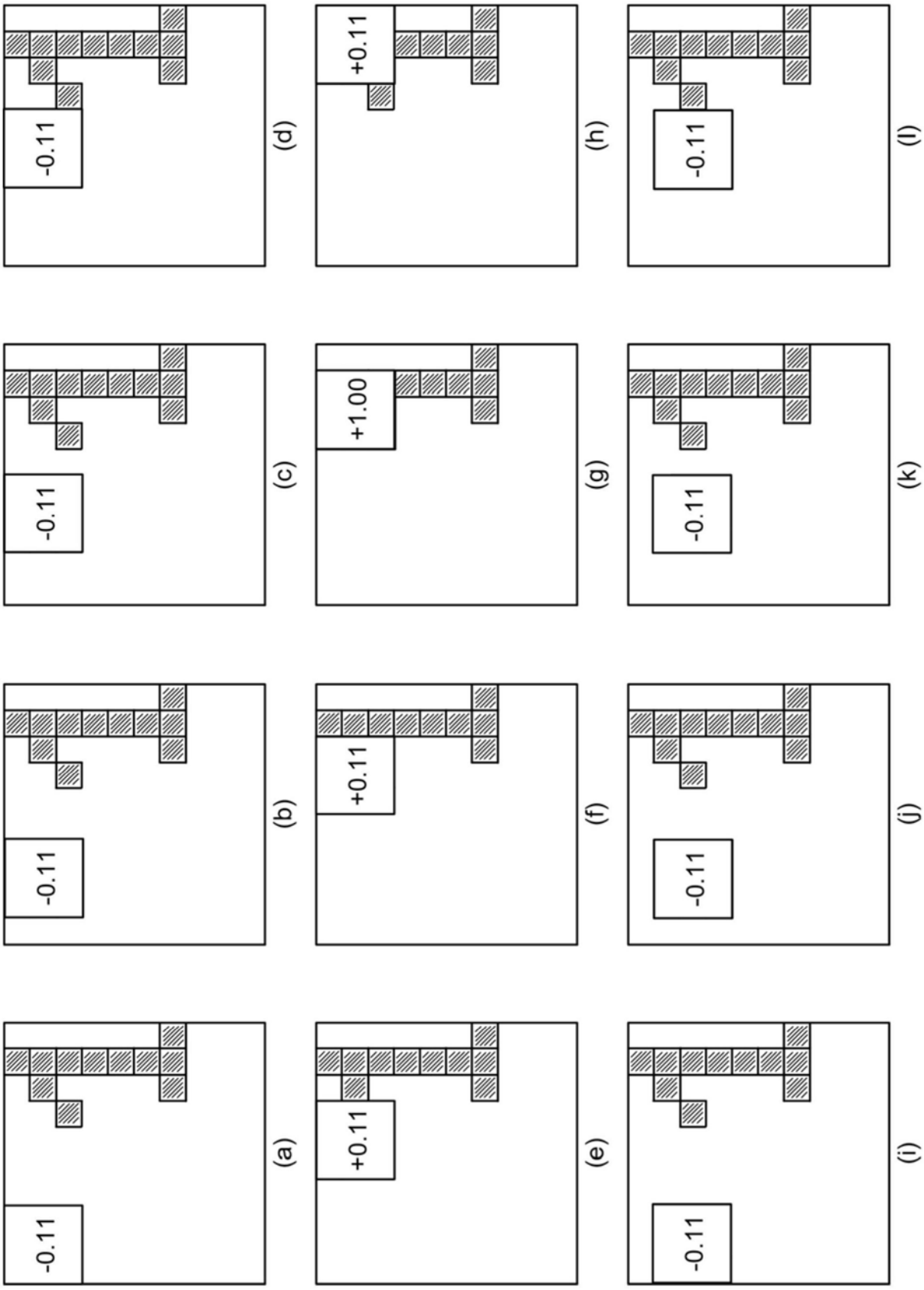


图1G

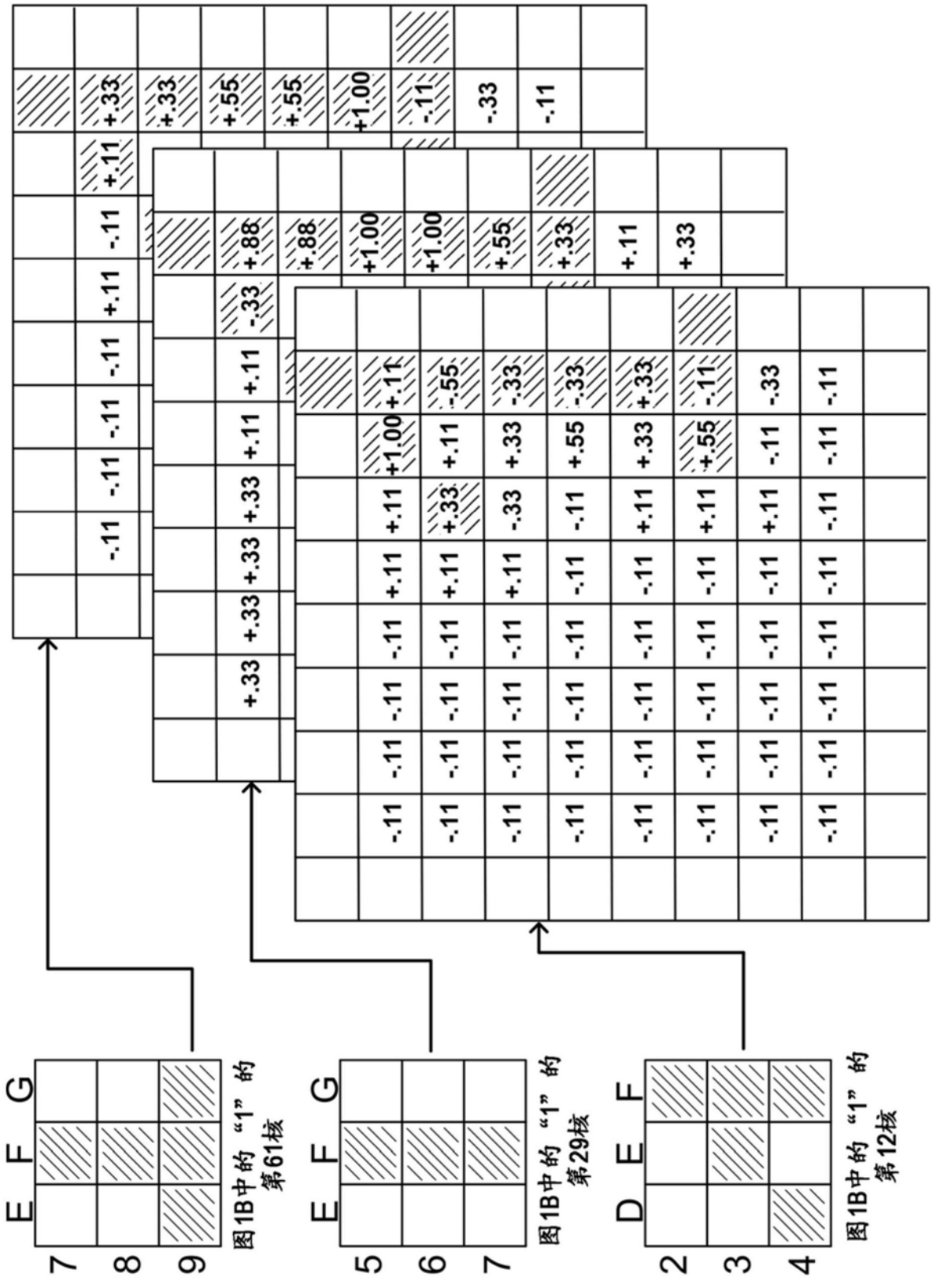

D E F
2 3 4
图1B中的“1”的第12核
E F G
5 6 7
图1B中的“1”的第29核
E F G
7 8 9
图1B中的“1”的第61核

图1H

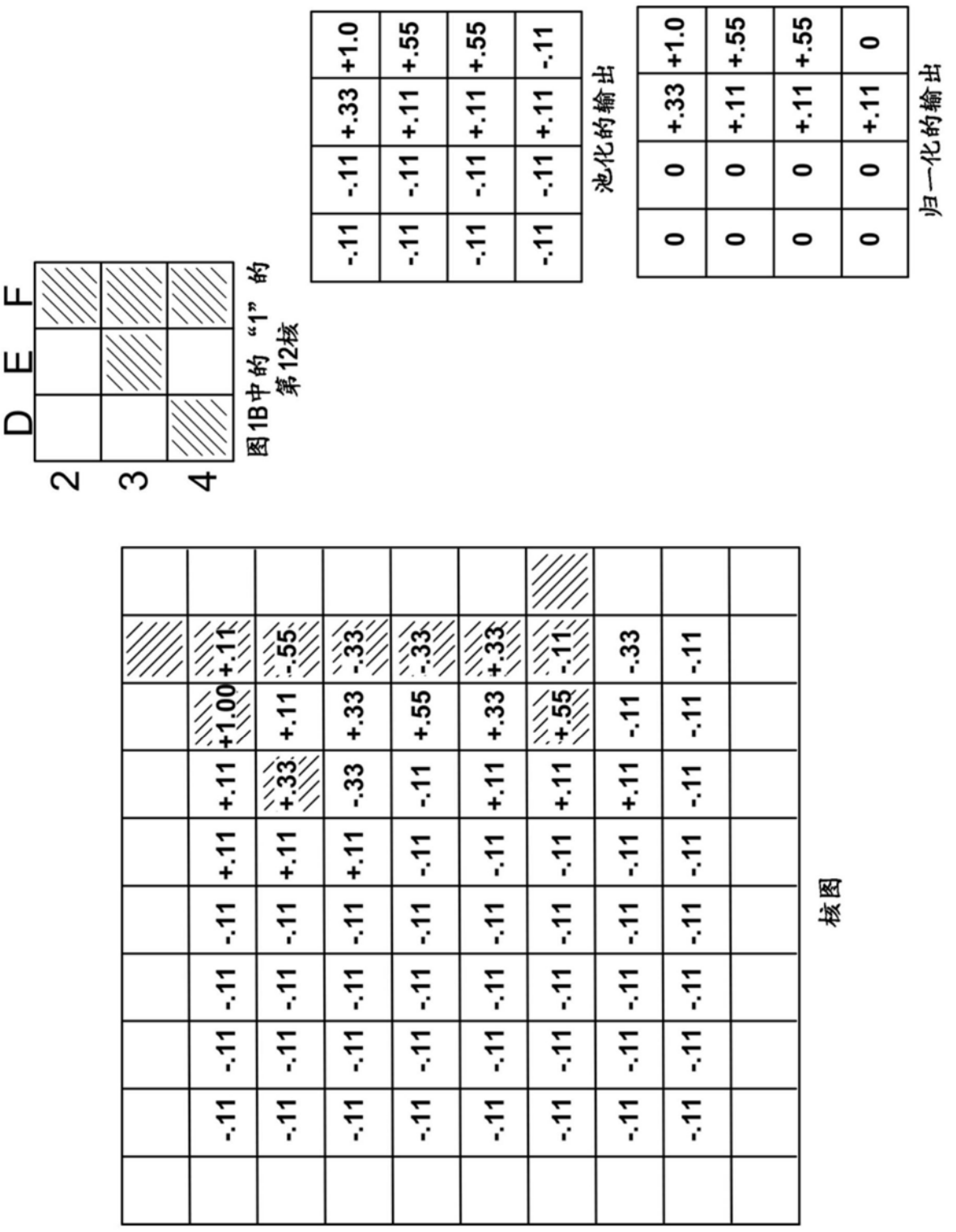

D E F
2 3 4
图1B中的“1”的第12核
-.11 -.11 +.33 +1.0
-.11 -.11 +.11 +.55
-.11 -.11 +.11 +.55
-.11 -.11 +.11 -.11
池化的输出
0 0 +.33 +1.0
0 0 +.11 +.55
0 0 +.11 +.55
0 0 +.11 0
归一化的输出
核图
+1.00 +.11 -.55 -.33 -.33 +.33 -.11 -.33 -.11
+.11 +.11 +.33 -.33 -.11 +.11 +.11 +.11 -.11 -.11
-.11 -.11 -.11 +.11 -.11 -.11 -.11 -.11 -.11 -.11
-.11 -.11 -.11 +.11 -.11 -.11 -.11 -.11 -.11 -.11
-.11 -.11 -.11 -.11 -.11 -.11 -.11 -.11 -.11 -.11
-.11 -.11 -.11 -.11 -.11 -.11 -.11 -.11 -.11 -.11
-.11 -.11 -.11 -.11 -.11 -.11 -.11 -.11 -.11 -.11
-.11 -.11 -.11 -.11 -.11 -.11 -.11 -.11 -.11 -.11

图1I

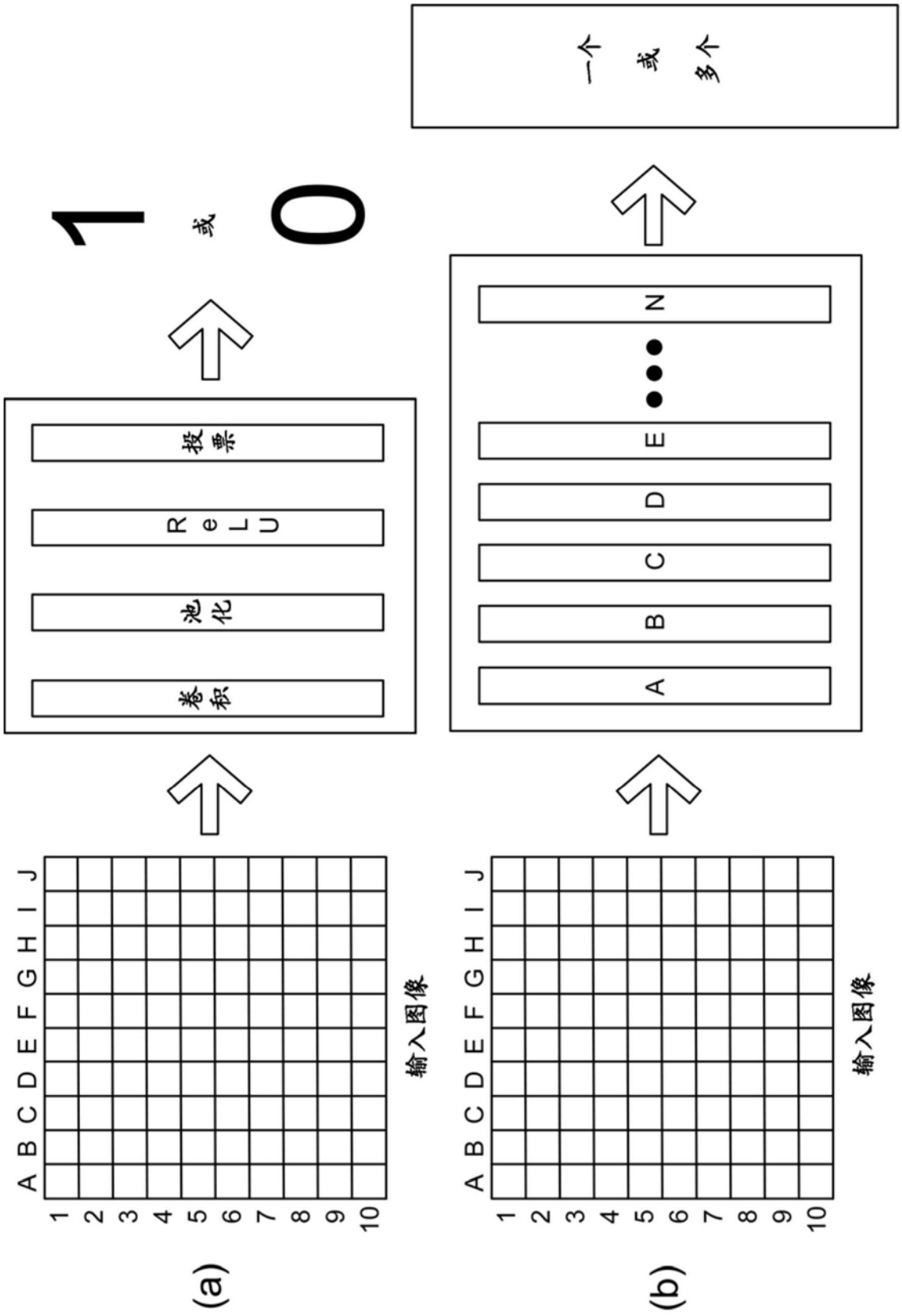

一个 或 多个
1 或 0
投票
ReLU
池化
卷积
A B C D E ••• N
输入图像
A B C D E F G H I J
1 2 3 4 5 6 7 8 9 10
(a)
(b)

图1J

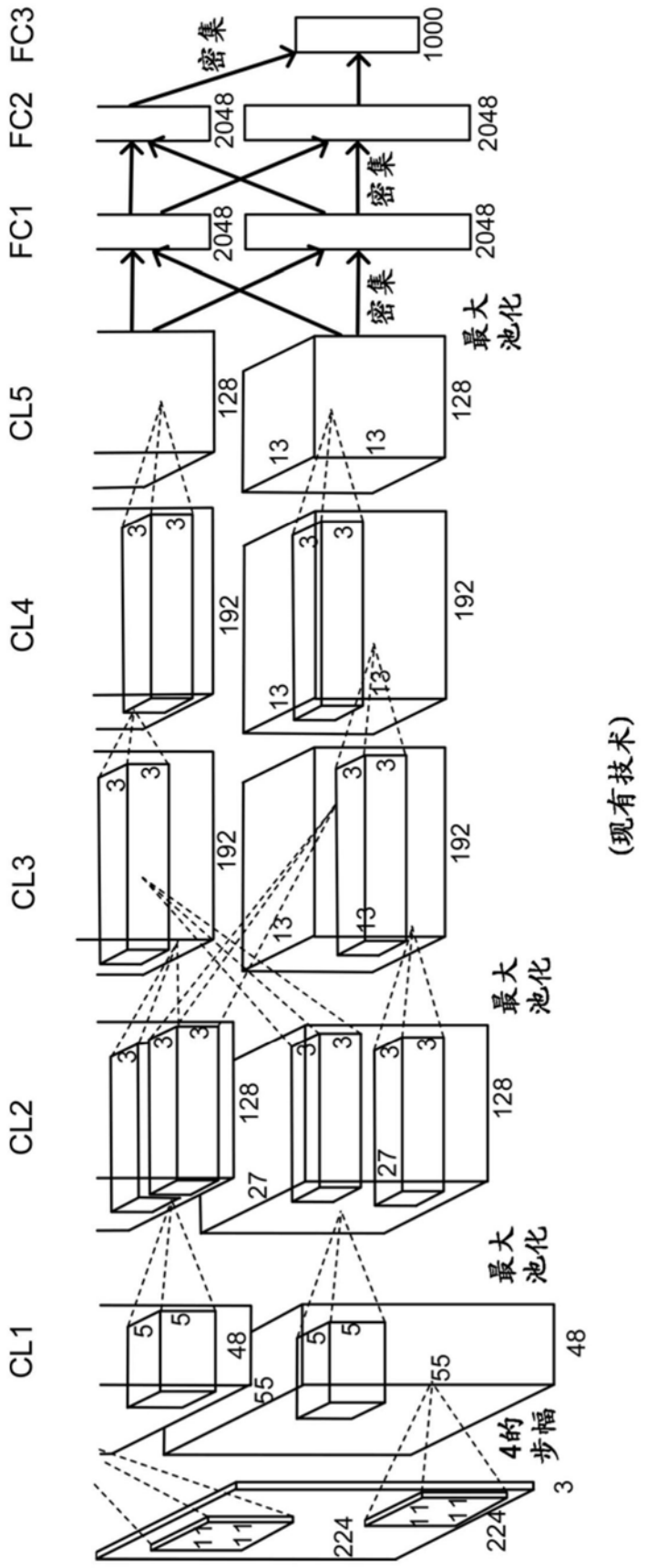
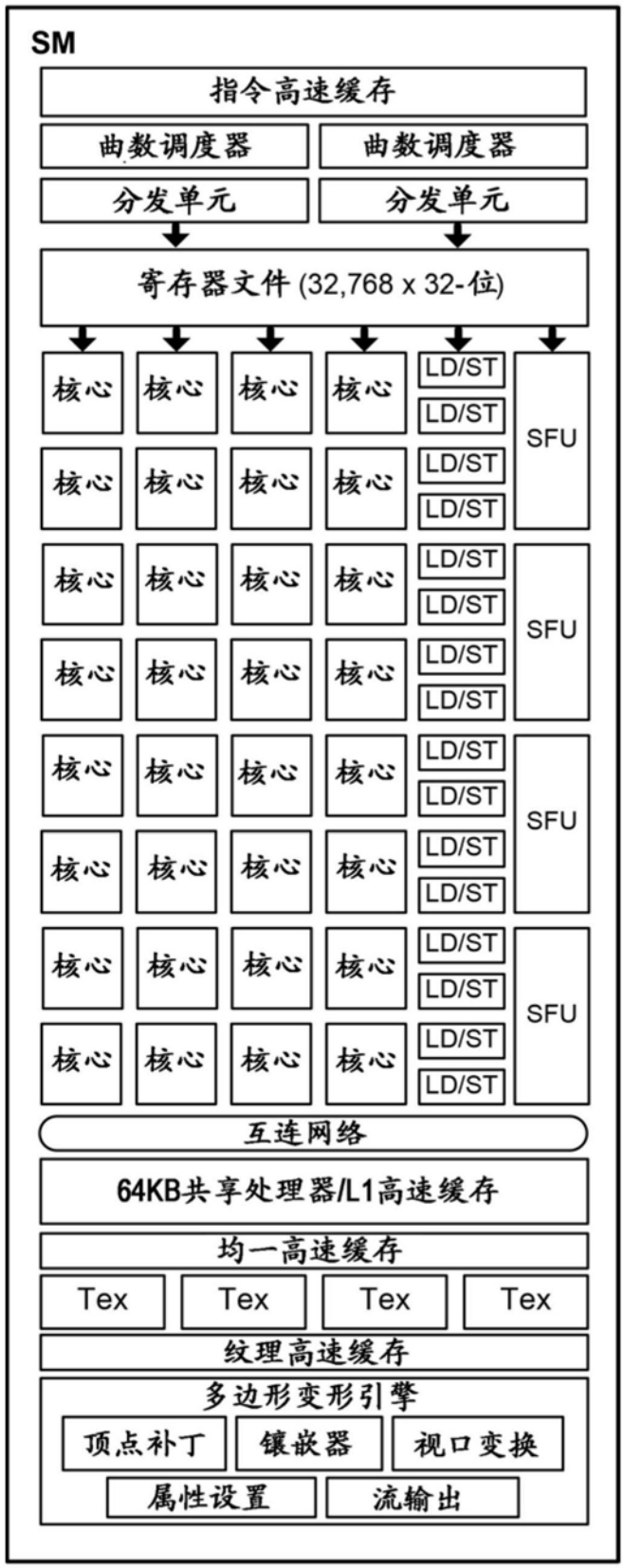
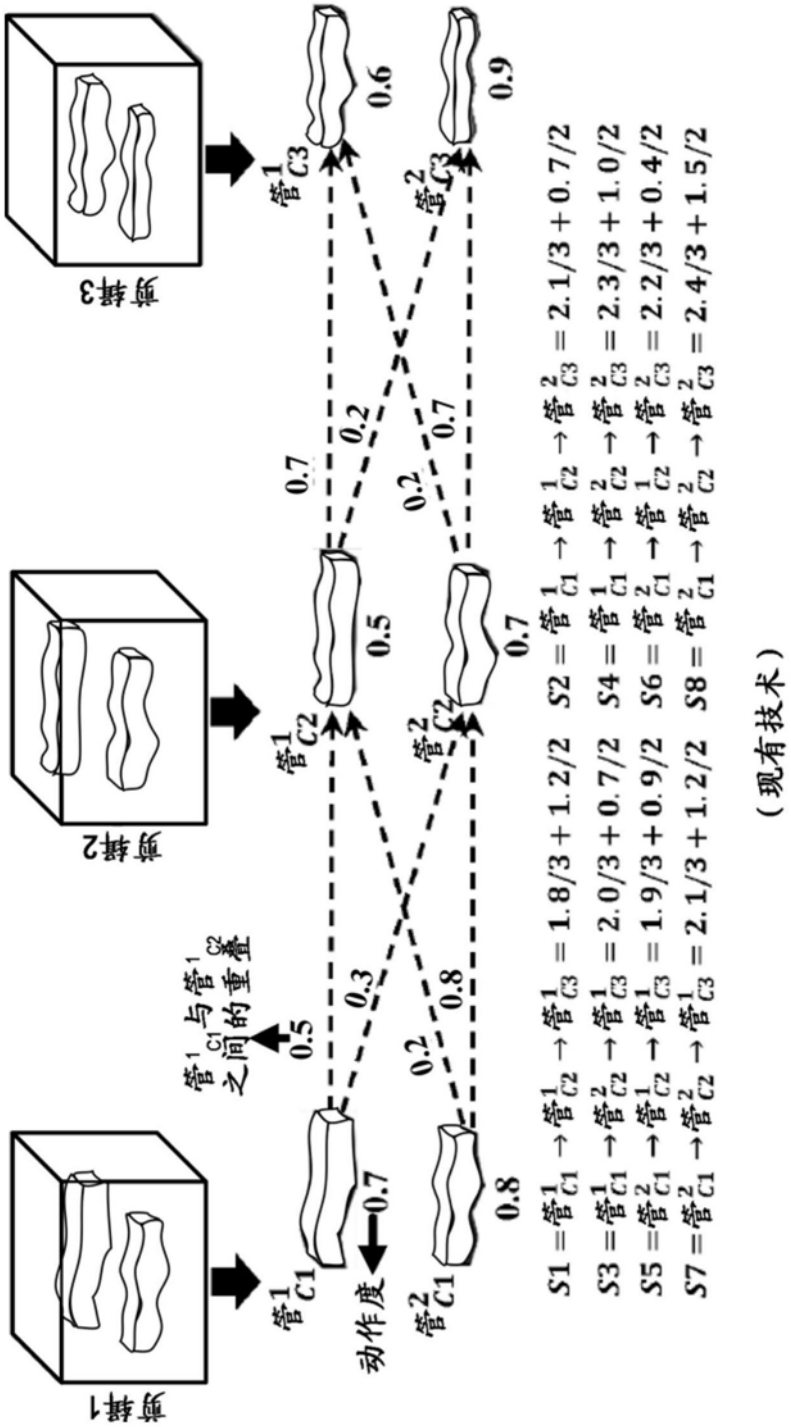


图2A


SM
指令高速缓存
曲数调度器
曲数调度器
分发单元
分发单元
寄存器文件 (32,768 x 32-位)
核心
LD/ST
SFU
互连网络
64KB共享处理器/L1高速缓存
均一高速缓存
Tex
纹理高速缓存
多边形变形引擎
顶点补丁
镶嵌器
视口变换
属性设置
流输出

（现有技术）

图2B


剪辑1
剪辑2
剪辑3
动作度
0.7
0.8
0.6
0.9
0.5
0.7
0.3
0.8
0.2
0.7
0.2
0.2
0.5
$S1 = 1.8/3 + 1.2/2$
$S2 = 2.1/3 + 0.7/2$
$S3 = 2.0/3 + 0.7/2$
$S4 = 2.3/3 + 1.0/2$
$S5 = 1.9/3 + 0.9/2$
$S6 = 2.2/3 + 0.4/2$
$S7 = 2.1/3 + 1.2/2$
$S8 = 2.4/3 + 1.5/2$

（现有技术）

图2C

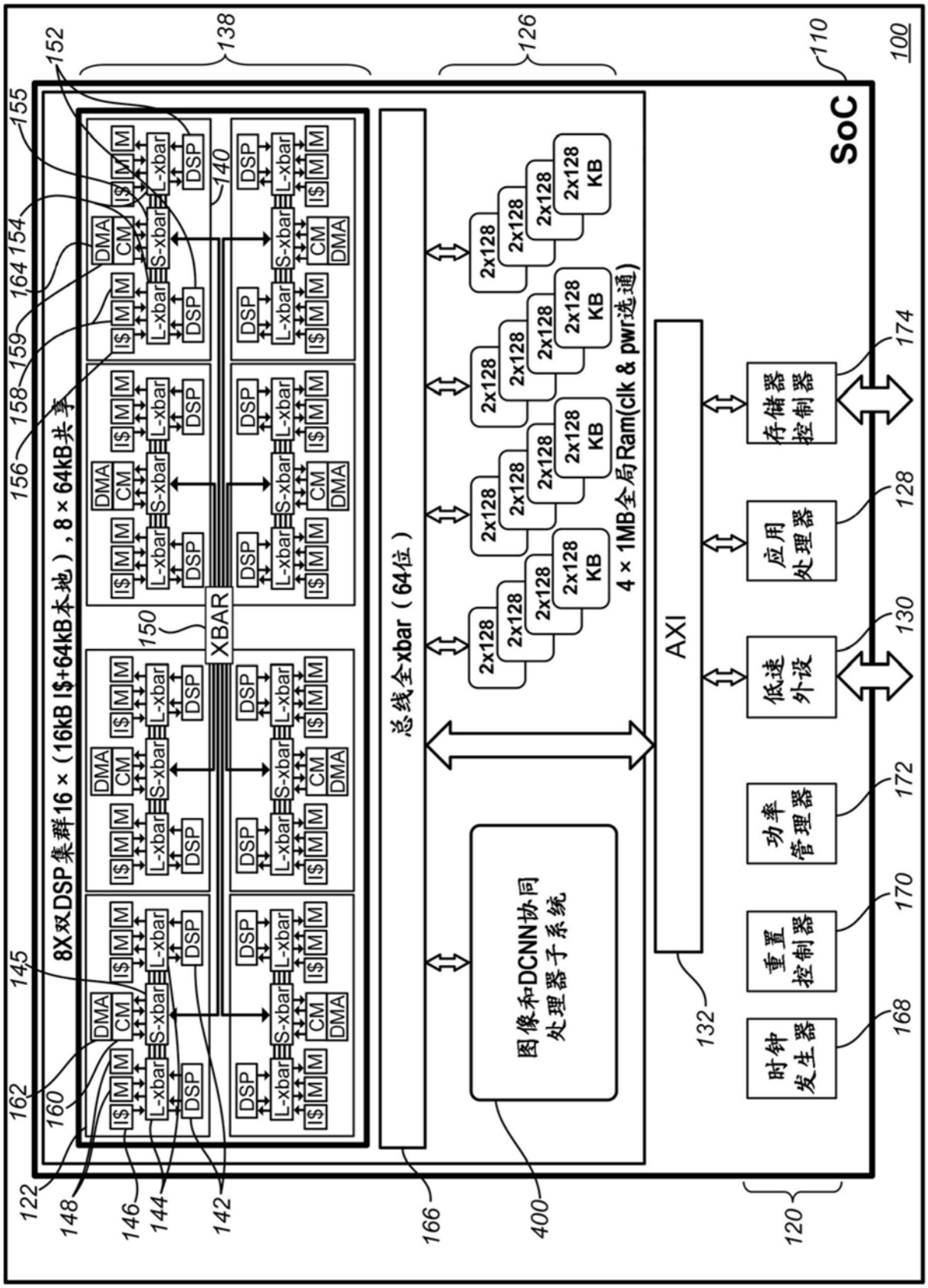

SoC
100
110
8X双DSP集群16×（16kB I$+64kB本地），8×64kB共享
XBAR
总线全xbar（64位）
4×1MB全局Ram(clk & pwr选通)
2x128 KB
AXI
图像和DCNN协同处理器子系统
低速外设
应用处理器
存储器控制器
功率管理器
重置控制器
时钟发生器
DSP
I$
M
L-xbar
S-xbar
DMA
CM

图3

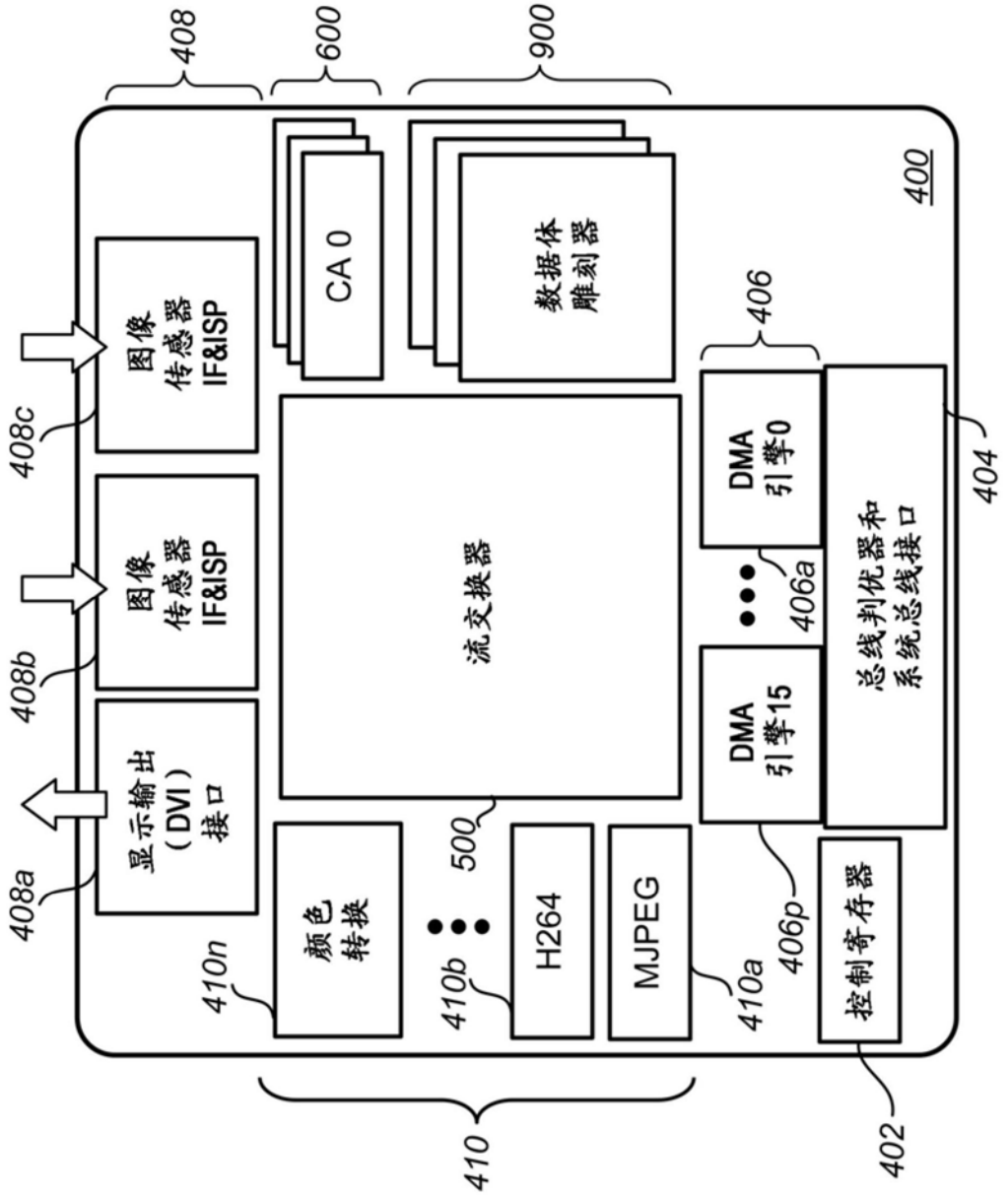

408
600
900
400
图像传感器IF&ISP
图像传感器IF&ISP
显示输出(DVI)接口
CA 0
数据/体雕刻器
流交换器
DMA引擎0
DMA引擎15
总线判优器和系统总线接口
颜色转换
H264
MJPEG
控制寄存器
408c
408b
408a
406
406a
406p
404
500
410n
410b
410a
410
402

图4

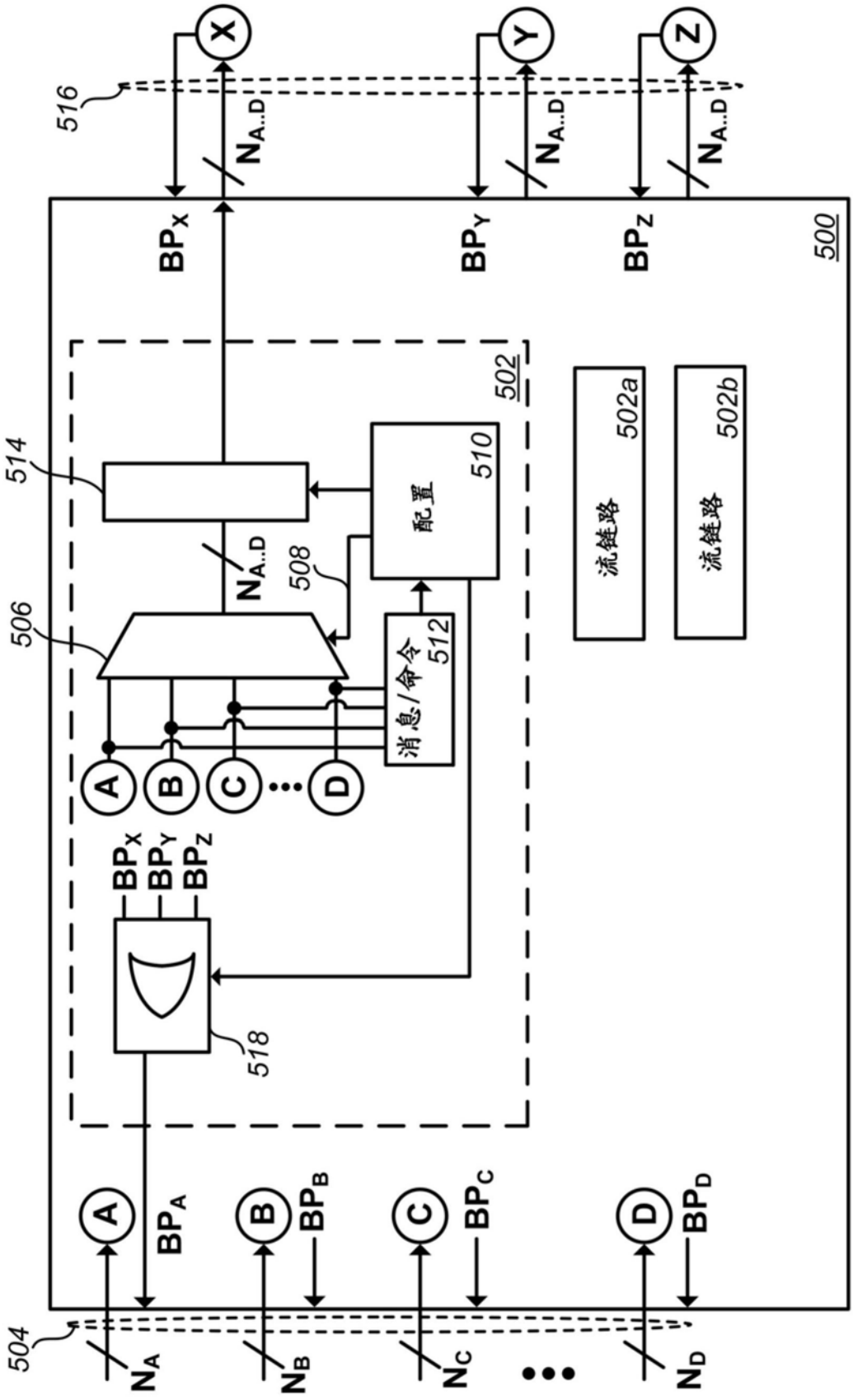

X
Y
Z
516
$N_{A..D}$
BP_X
BP_Y
BP_Z
500
502
514
配置
510
508
506
消息/命令
512
流链路
502a
流链路
502b
A
B
C
D
518
BP_A
BP_B
BP_C
BP_D
504
N_A
N_B
N_C
N_D

图5

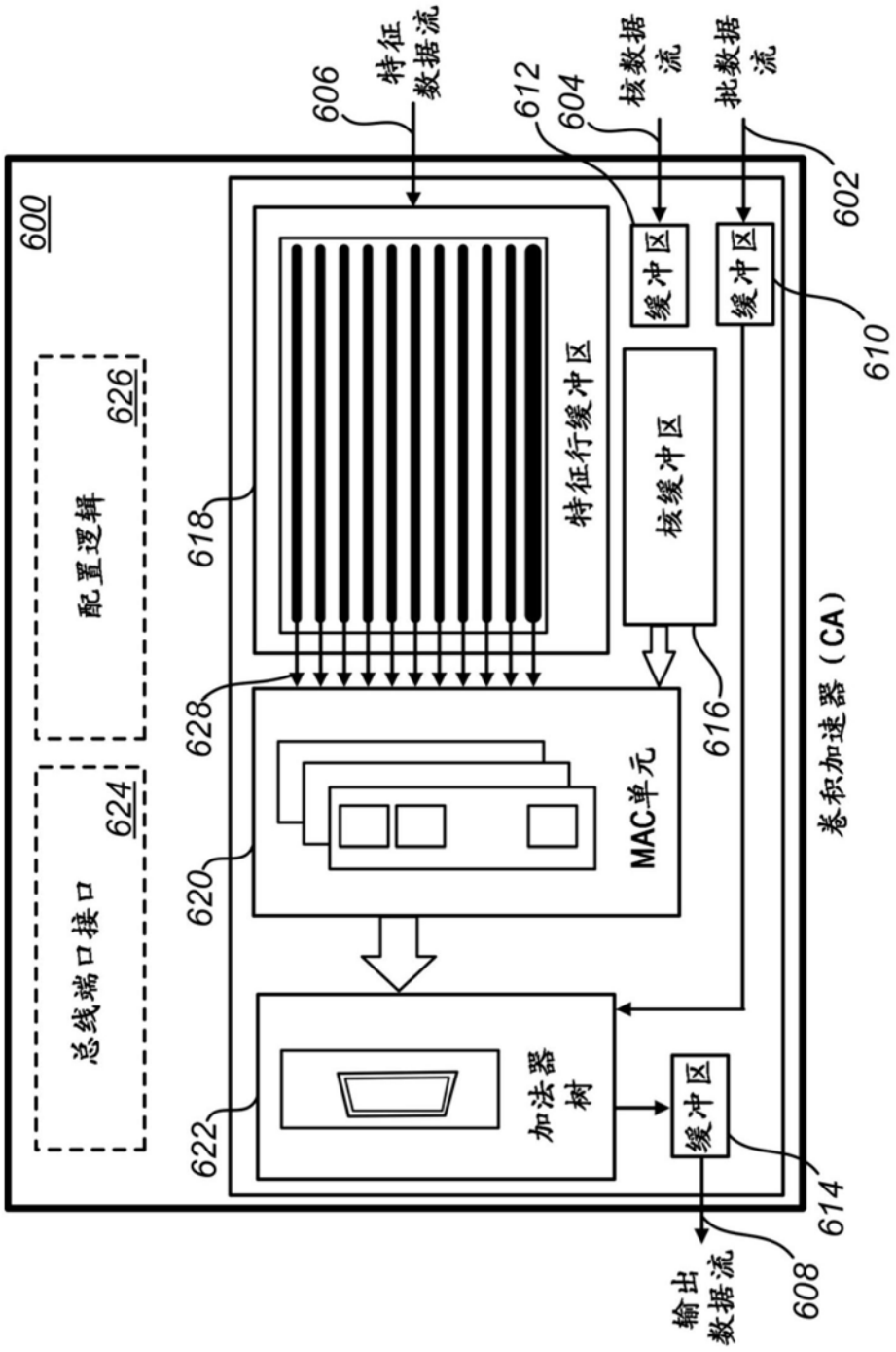

600
606
特征数据流
612
604
核数据流
批数据流
602
缓冲区
缓冲区
610
配置逻辑
626
特征行缓冲区
618
核缓冲区
628
MAC单元
616
总线端口接口
624
620
卷积加速器（CA）
加法器树
622
缓冲区
614
输出数据流
608

图6

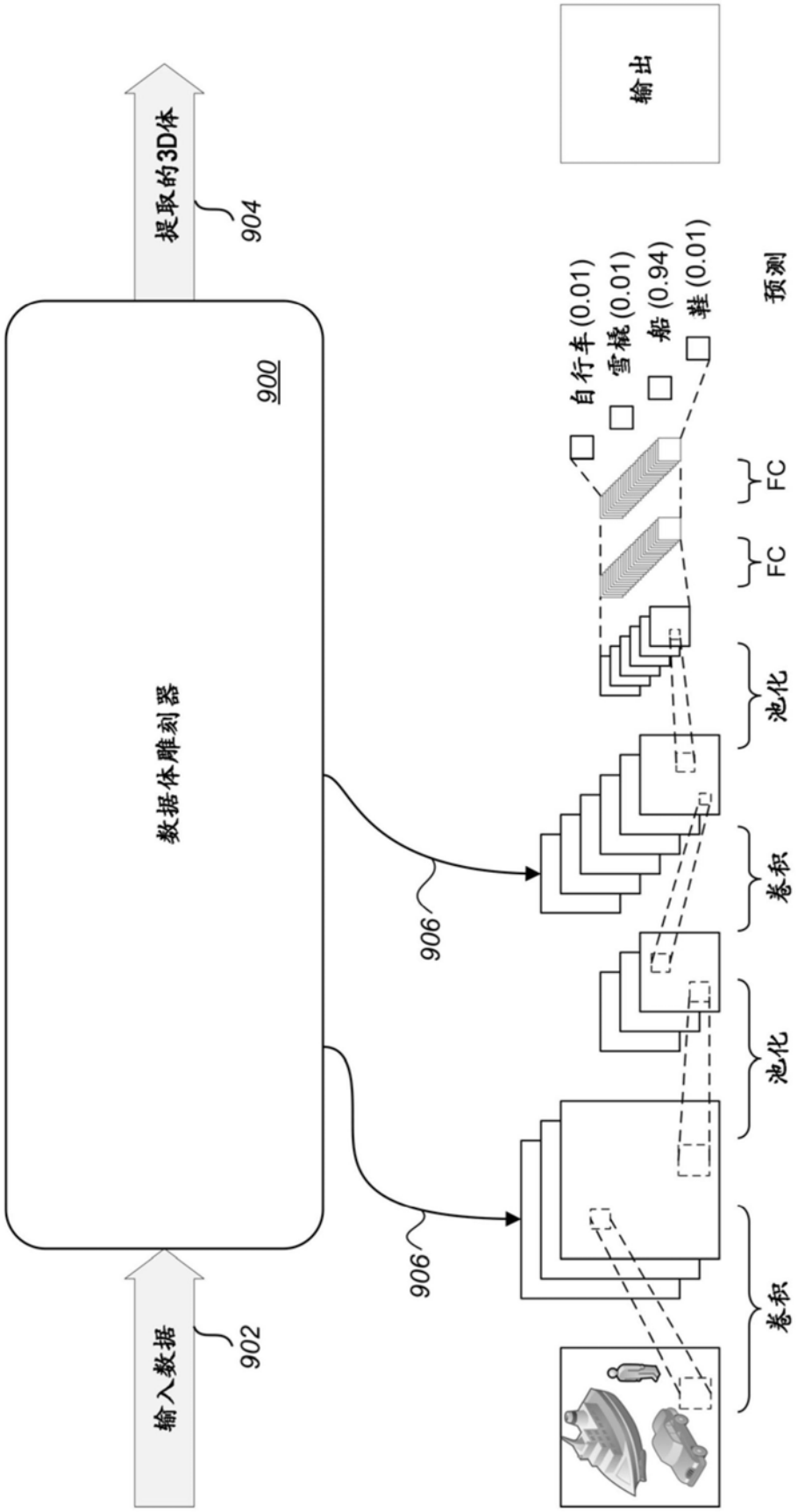

输入数据
902
数据体雕刻器
900
提取的3D体
904
906
906
卷积
池化
卷积
池化
FC
FC
预测
自行车 (0.01)
雪橇 (0.01)
船 (0.94)
鞋 (0.01)
输出

图7

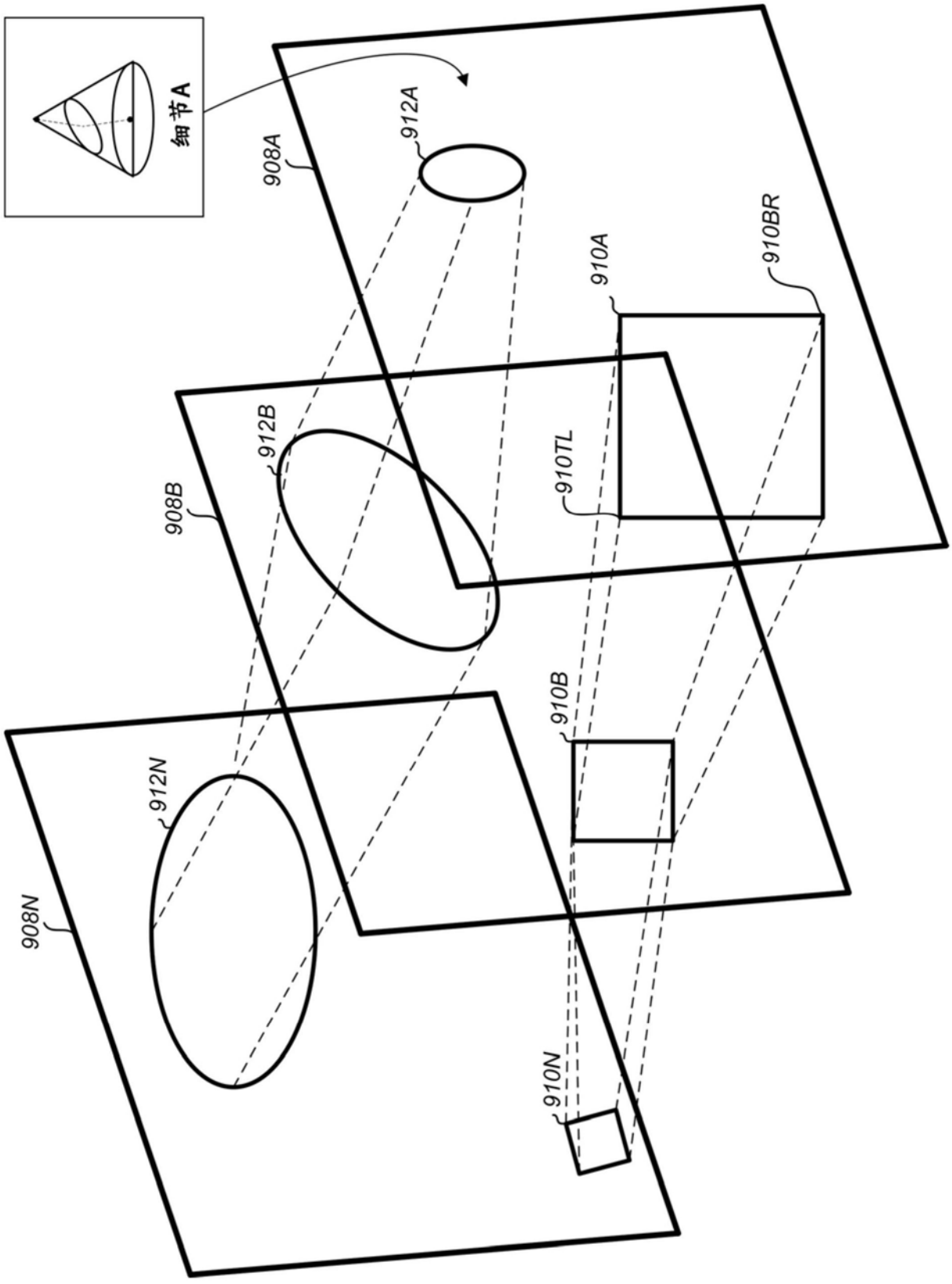


图8A

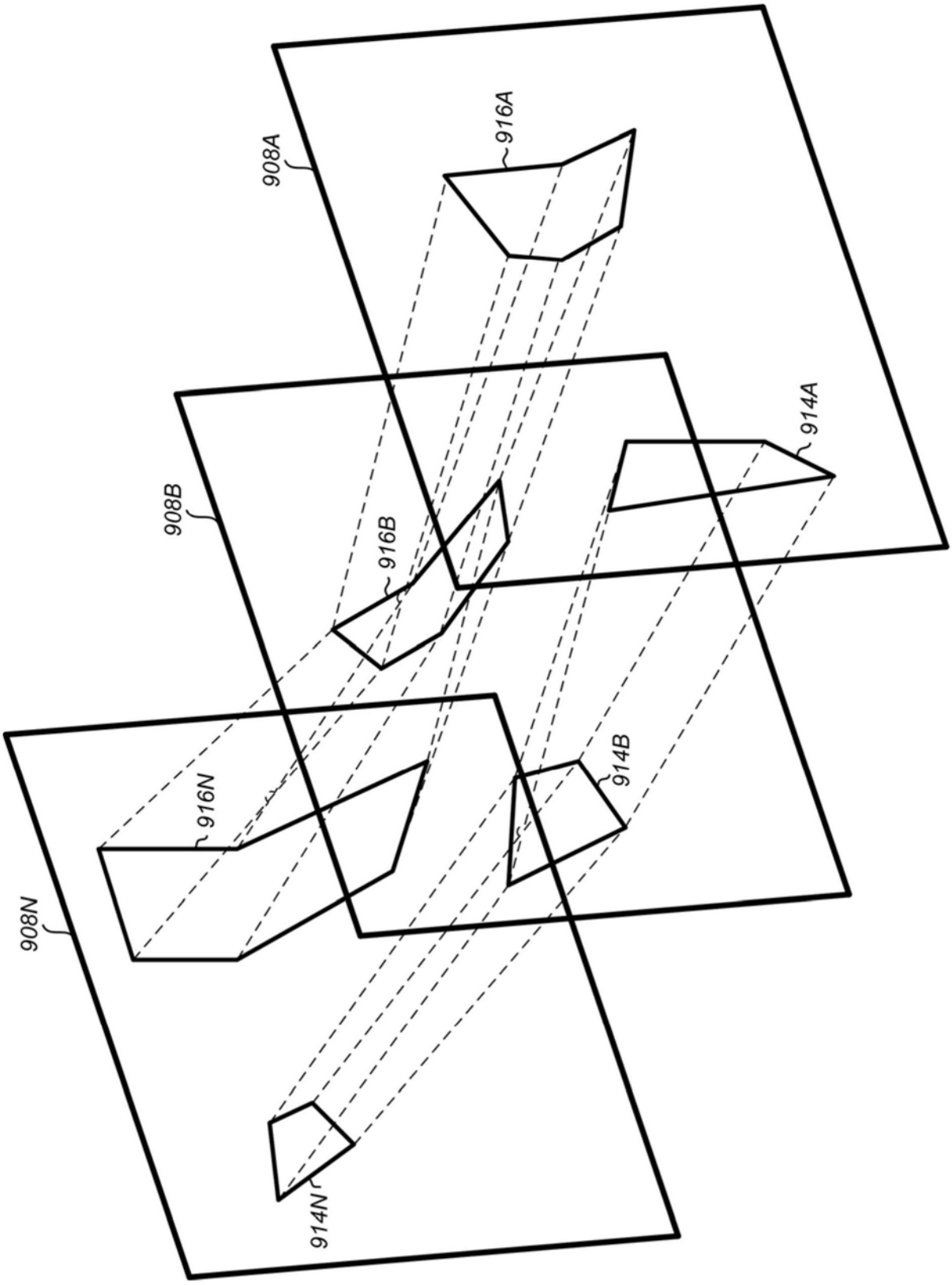


图8B

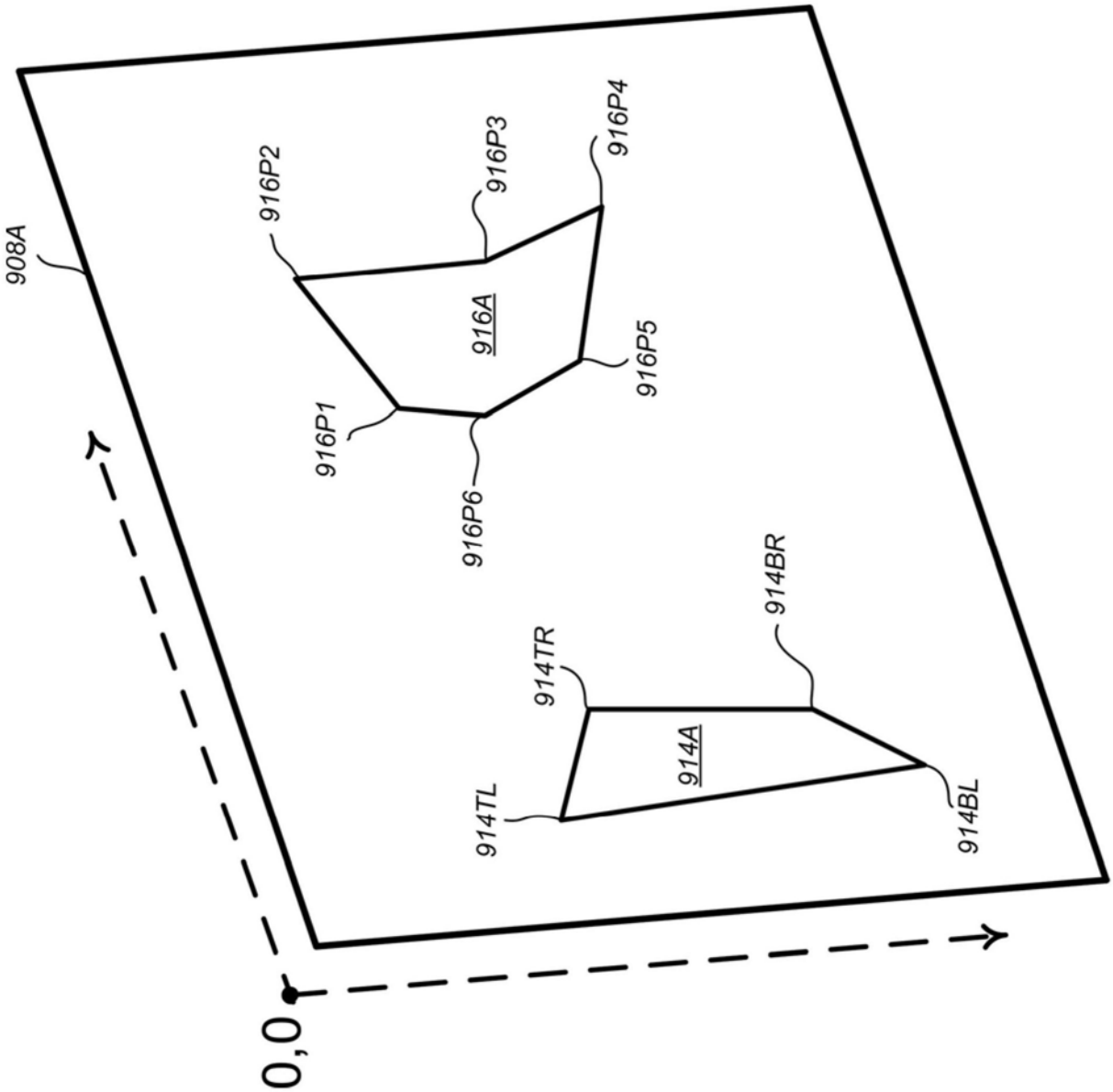

908A
916P2
916P3
916P4
916A
916P5
916P1
916P6
914BR
914TR
914A
914TL
914BL
0,0

图8C

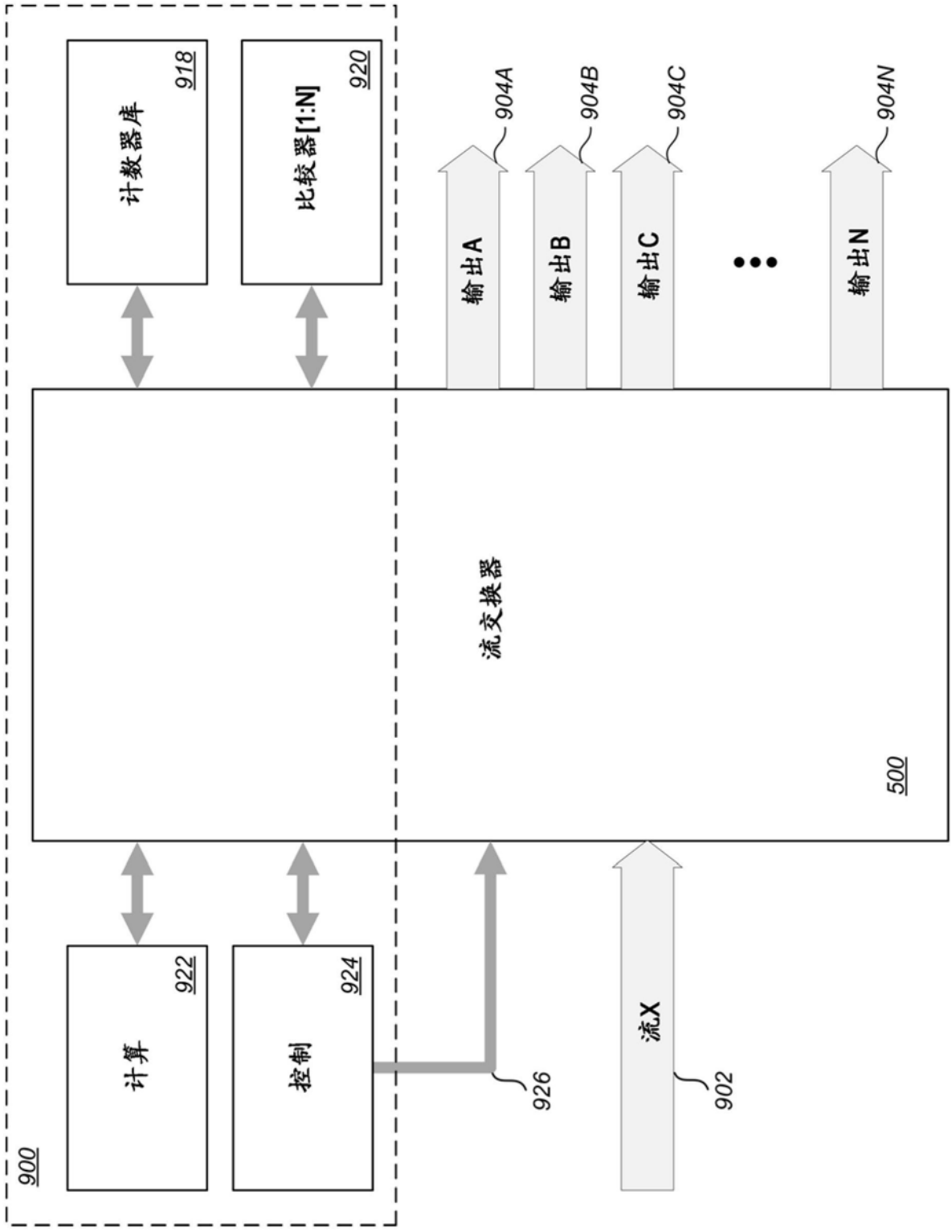

计数器库
918
比较器[1:N]
920
904A
904B
904C
904N
输出A
输出B
输出C
输出N
流交换器
500
计算
922
控制
924
926
流X
902
900

图9

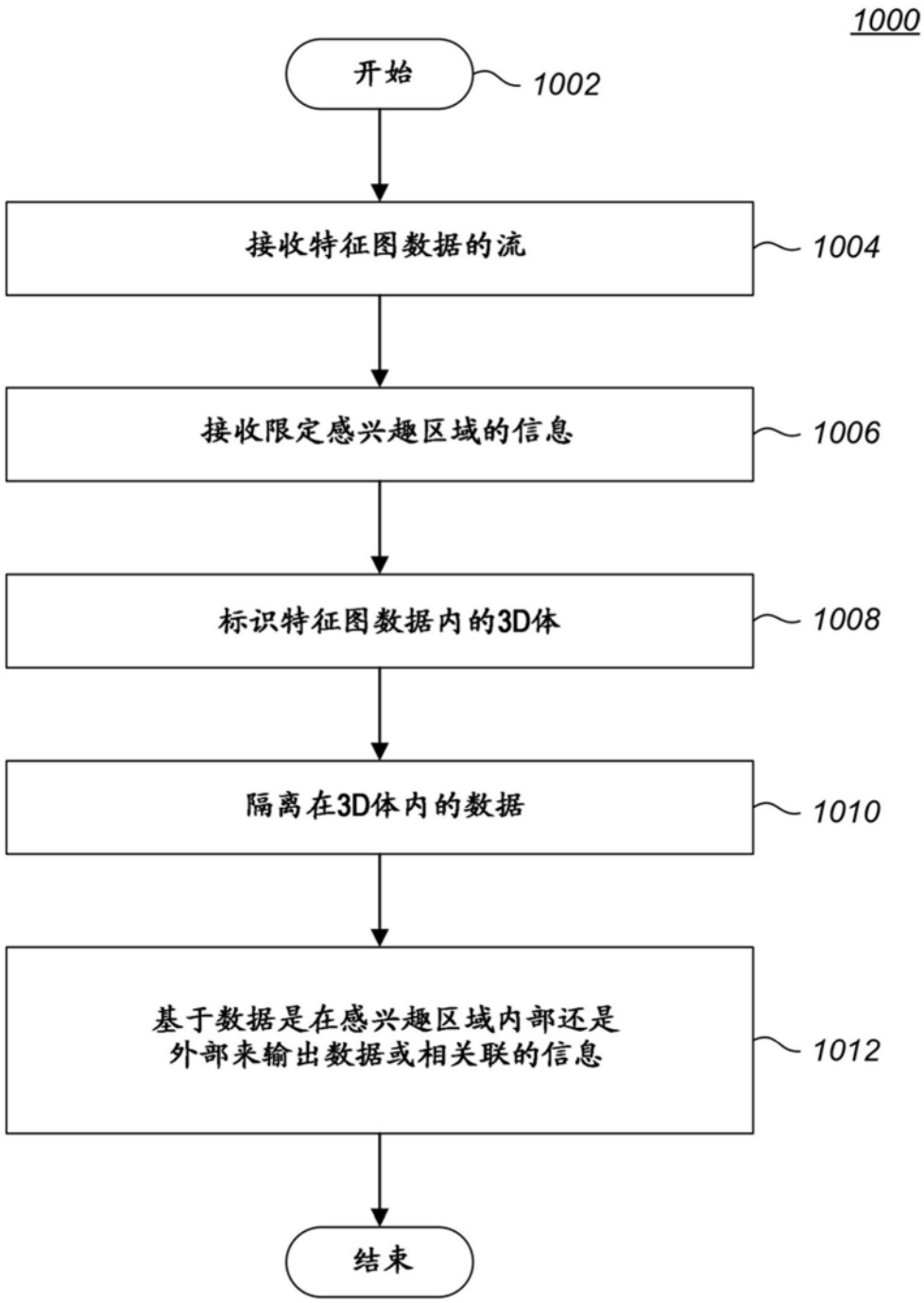

1000
开始
1002
接收特征图数据的流
1004
接收限定感兴趣区域的信息
1006
标识特征图数据内的3D体
1008
隔离在3D体内的数据
1010
基于数据是在感兴趣区域内部还是
外部来输出数据或相关联的信息
1012
结束

图10