指导教师： 杨涛 提交时间： 2016.3.19

CVPR2015 Paper

No: 01

姓名： 张琦

学号： 2013302563

班号： 10011304

CNN 匹配遇上 KNN:准参数化人型分析

摘要：
两个参数和非参数方法已经证明在对人解析任务中有令人鼓励的表现，即分割一个人物图像分成几个语义区域（例如，帽子，手提袋，左臂，脸）。 在这项工作中，我们的目标是开发一个新的解决方法，即从注释数据监测使用新注释（可能是罕见的）图像，并提出一个准参数解析人体模型的灵活性的优势，新的解决方案。 根据经典的 K 近邻（KNN）为基础的非参数框架，参数匹配卷积神经网络（M-CNN）提出来预测在一个匹配的信心，并在测试图像最佳匹配区域的位移为特定的语义区域 KNN 形象。 给定一个测试的形象，我们首先从注释/手动解析人类图像库检索其 KNN 图像。 然后在每个 KNN 图像每个语义区域与信心使用 M CNN 的探伤图像匹配的，并从所有的 KNN 图像匹配的区被进一步稠合，随后一个超像素平滑过程以获得最终的人解析结果。 在 M-CNN 从经典 CNN 的不同点在于，定制交叉图像匹配滤波器引入以表征测试图像和 KNN 图像的语义区域之间的匹配。 匹配滤波器的交叉图像是在不同的卷积层限定，各个目标捕获位移的特定范围。 在一个大的数据集进行全面评估与 7700 批注人的形象展示以及来自国有的最艺术，为人类解析任务准参数模型的显著的性能增益。

1 介绍：
对人的分析，即分割成多个语义区域（例如，帽子，左/右腿，眼镜的人的身体和上半身衣服），已引起在重多的关注百分之年[29，30，4,3]和用作许多基础高层次的应用程序，如服装分类[1]和检索[20]。几个参数和非参数的人类解析方法，提出并证明非常好的效果对人体解析任务。一方面，对位度量方法[30，3，19，4]学习的知识，如不同的语义标签和的区域的外观不同的标签之间的结构关系，从有

注释的图像。这些方法通常依赖于使用制造同盟设计结构模型[4]，这可能不适合具体的数据良好，从而达到最理想的效果。此外，对于另一组新的训练数据和语义标签，新车型必须设计/再培训，这使得这些参数化模型不切实际的，因为新的服装款式可能来自于经常。在另一方面，非参数方法可以灵活地使用上飞新注释的图像和处理是参数模型的代表起诉，这对于更有吸引力实际应用中[17]。这种方法通常首先构建像素级[17]，超像素级[29]或假设级别[27，13，14]测试之间的匹配图像和语料库中注释的图像，然后转移从手动注释的图像的标签所述检测于匹配输出 ING 图像，最后的保险丝转移的标签启发式聚集方案（典型的多数表决）。然而，匹配的质量是通常由缺乏明确的语义限制自下而上超像素或假设。上述的参数和非参数人类解析方法依赖于手工设计的管道多个顺序元件组成，例如，手动精雕细琢特征提取，自下而上的过度分割，人类的姿态估计，手动设计复杂的模型结构体。因此，每一个可能的糟糕表现组分可成为性能的瓶颈整个管道。例如，人的姿势的估计，在上述管道的重要组成部分，它本身是相当有挑战性的任务。这样一个顺序处理商业策略埃及通常使整个管道大多不理想。
而不是多个顺序步骤的组合时，卷积神经网络（CNN）为基础的方法提出了一个图像分析[9，28，22，5]。然而，这些深层次的模型不能轻易更新，以便纳入新的语义标签。为了解决这些问题，我们提出了一个准参数人类解析框架，这既继承的优点的参数和非参数化模型。建议框架能够承担监督

充分利用从注释训练数据的信息，同时易于扩展新添加的标签。核心部分拟议的框架是一个专门设计的配套 CON-卷积神经网络（M-CNN）匹配任何一个 KNN 图像的抽动区（也表示为 KNN 区域中本文），以测试图像。如示于图 1，我们首先应用人类探测器化[6]到探伤图像，将获得的人的中心的 IM-年龄。然后，K 最近邻居（KNN）的图像测试图像从注解/手动解析检索图像库。每个 KNN 图像导出几个抽动区域，这是由掩蔽出背景生成与图像语料库的平均图像地面区域。显示在帽子，裙子和裤子三大 KNN 地区图。然后，在一对测试图像的每个的 KNN 区域被送入提出的 M-CNN 来估计与其相配套的信心和位移。 该匹配的措施如何 KNN 区域相匹配的输入图像而位移描述在 KNN 地区和匹配的内特翻译区域中的测试图像。 匹配的是置信度然后平均超过所有 KNN 地区和阈值处理，以预快译通特定标签是否存在。对于标签预测为存在，如在图帽子和裙子。图 1 中，正确响应标签地图可以从 KNN 重新传送到基于所估计的位移探伤图像求。为标签预测为不可见，诸如裤在图 1，不产生转移标签地图。然后，所有的特定标签的匹配区域相结合，以产生一个概率映射图。最后，对于所有的概率图贝尔由超像素平滑化步骤，以获得精最终的结果分析。输入图像和 KNN 之间的可靠匹配区域是具有挑战性的，因为匹配需要处理语义区域的大型空间变异。例如，该袋可放置在左，右或中前面男人的身体。 所提出的 M-CNN 能够实现准确多不等的匹配。如示于图 2，三个路径，即两个单个图像卷积路径和交叉图像卷积路径。*单个图像卷积路径*接收输入图像或特定 KNN 区域，并产生其判别的分层一层功能层交涉。*交叉图像 CON-卷积路*交叉嵌入图像的过滤器到每一个卷积层表征多不等的匹配。交叉图像过滤器适用于所有前特征图卷积各层，包括单个图像地图和交叉图像特征图。由于规模当跟踪功能的地图增加的感受野了 M-CNN，交叉图像匹配滤波器捕捉从近至远的范围的位移。那里-从跨图像卷积脱颖而出特征图路径可以很好地代表了位移。因为如果没有这特性 TURE 地图由两个单个图像卷积生成路径是出色的功能表现，它们的绝对差分地图被计算为的另一测量的位移。所不同的地图相结合，与十字图像特征映射，然后链接到后面的完全连接层。最后，匹配置信信心和位移倒退。由于 M-CNN 在输入图像和任何 KNN 区域匹配目标任何语义标签，它可以即使新的语义标签工作都包括在内。而不是培养了 M-CNN 为每个标签，我们培养一个统一的 M-CNN 的所有标签全部 KNN 地区。在一个大的数据集进行全面评估 *7700* 人的注释以及图片展示了我们的准参数框架。总结如下：

•我们建立了一个新的深准参数解析人框架。它可以从注释数据，还学习灵活运用新的注释（可能是不常见）。

•我们提出了一个匹配的卷积神经网络工作（M-CNN）来匹配 KNN 的语义区域图像到探伤图像。小说交叉图像过滤器嵌入到不同的卷积层，每旨在捕获位移的特定范围。

•我们整合所有的一步一步的组件（过分割，姿态估计，特征提取，镧贝尔建模等）在传统的管道成一体统一深 CNN 框架。

2. 相关工作

在本节中，我们依次查看参数人类的解析方法，非参数方法和深基于学习的方法。对于人体参数解析，山口等人。[30]提出要加强与像素级人解析评定类别依靠人的姿态估计。 捕捉更复杂的情境信息，Dong 等。[4]签署

了与或图结构的相关性模型一组 parselets，他们的推广工作[3]单向的田间人体解析，在一个框 图像共同分割和地区共同的标签用于人类的解析也被用来捕捉相关不同人的图像之间[31]。 此外，刘等人。[19]利用用户生成的类别标记构建男人解析器。 对于一般的图像解析，泰伊等。[26]提出的检测方法分割。首先，的对象的包围盒由估计 PLAR SVM [26]，在此基础上分割口罩从图像语料库到输入图像传输。在一般情况下，现有的参数方法的功率在很大程度上许多手的次优性能的限制德签署中间组件，如姿态估计，并且也不能容易地扩展到解析新的标签。

在非参数人类解析，像素[17]，ELS [25，7，14]和对象提议[27，13，14，23]分别用于促进非参数图像解析。凯莉，山口等人的模型。[29]传送 pars-从检索到的例子来查询图像 ING 口罩。其标签转印是基于超像素，这是基于低电平 AP-过分割产生线索，因此缺乏语义。[17]利用 SIFT 流构建像素像素的'正确响应研究和 IM-之间的密集变形场年龄。然而，用于求出优化问题 SIFT 流是相当复杂的，并要解决昂贵。回覆，龙等。[21]证明了更好的性能卷积激活功能比传统特色，如 SIFT，为任务要求对应。总体，非参数方法是由不准确的限制该标签转印时导致的噪音/异常值。我们的准参数模型集成了参数化模型和非参数模型所提议的 M-CNN 的优点。存在与 CNN 的架构语义分割部分作品。[6]和它的扩展工作[9]提出了对候选区域由 CNN 的语义分割分类。 Wang 等人。[28]提出了一个共同的任务学习框架，其中，所述对象定位任务和对象分割任务经由 CNN 协作处理。 Farabet 等。[5]训练有素的多尺度 CNN 从原始像素提取深特征用于分配标签到每个像素。复发 CNN [22]提出了加快现场分析和实现国家的最先进的性能。我们的 M-CNN 继承现有 CNN 解析模型的优点在我们的单个图像卷积路径。它在现有的所有基于 CNN 分析模型不同，我们处理一对图像，而不是一个单一的形象，我们把跨图像滤池以具体来说表征多不等的匹配。最后但并非最不重要的，M-CNN 可以毫不费力地处理新的语义标签。

3. 准参数解析人

对于每个人的形象，我们第一个从注释图像库（秒 3.1）检索它的 KNN 图像。然后，M- CNN 预测输入图像和从一个 KNN 图像语义区域，在此基础上的标签图产生（秒 3.2）之间的匹配置信度和位移。最后，所有的标签映射被送入一个后处理程序，以产生分析结果（秒 3.3）。

3.1. K-近邻检索

对于每一个输入图像的，我们使用的人体检测算法[6]，以检测人体。所得人为中心的图像 I，然后重新缩放至 227×227×3 然后，我们从基于所述 Krizhevsky 架构上训练 ILSVRC 2012 分类音响阳离子数据集的预训练 CNN 的模型中的倒数第二个完全连接层提取全局 4,096 维特征[12]。其 KNN 图像 G = {G1，G2，...，GK}从基于所述深特征的图像语料库检索。

3.2. 匹配卷积神经网络

输入，输出和损失函数：给定一个标签 L∈{1，...，L}，其中 L 是总数量的标签中，输入图像 I 和每个 KNN 区域从 KNN 图像 GKL GK 形成一对，并且馈成的 M-CNN。注意，GK 是从图像语料库，因而它的标签图是已知的。 GKL 通过保持标签 L 的区域中的 GK，并且通过从图像库算出的平均图像掩蔽出其它区域产生的。如果 GK 不包含标签 L，GKL 是完全平均图像。

由于图像对{I，GKL}，M-CNN 学会估计它们之间的匹配置信度和位移一个回

归。 1-暗淡匹配置信 CKL 表示 GKL 如何可以匹配 I.它是指示 KNN 区域是否匹配于输入图像的二进制索引。GKL 被认为是与我当且仅当一载有标签 L 相匹配。在测试阶段，CKL 的较高值表明更好的匹配被找到。我们表示在 GKL 作为 Ug 的= [GX1 KL，GY1 KL]和的 Wg = [GX2 KL，曲线 gy2 KL]的 KNN 区域的左上角和右下角的坐标。类似地，匹配区域的我的坐标被表示为用户界面=[IX1，Iy1]和 WI = [Ix2 的，IY2]。坐标是由高度和图像的宽度为范围[0,1]归一化。4 维位移漳州灿坤代表 UG 和用户界面，工作组和 WI 之间的差异。而是采用了分类科幻阳离子损失[12]，我们通过减少地面真相[CKL，漳州灿坤]和预测[∽CKL，漳州灿坤∽]之间ℓ 2 距离训练 M-CNN。

其中 Φ （·）是包含的特定网络连接 C 图像中的标签集。式中的第一个任期。 （ 1 ）可以，第二项对应于位移损失匹配置信度的损失。我们惩罚位移损耗当两个 KNN 图像 GK 和输入图像 I 含有标签 L 。然后，所有的培训对的损失相加和参数通过反向传播教训。

架构：由于 KNN 图片是基于全球外观相似性检索，每个标签的 KNN 区域可定位完全不同的图像。例如，在图 3，袋可放置在左侧或右侧，或在人体的前面。 M- CNN 旨在通过嵌入在不同的卷积层交叉图像匹配滤池估计多不等的匹配。如示于图 2，M- CNN 含有两种路径，即，两个单图像卷积路径和一个交叉图像卷积路径这三个路径的输出被进一步稠合来估计匹配置信和位移。单一映像路径的目的是为分层特征的代表性，而交叉图像卷积路径估计输入对之间的位移。单张图像卷积路径：我们有图的顶部和底部行中的单个图像路径两个实例。 2，其每一个分别处理 I 或 GKL。它们共享相同的架构和提取我还是 GKL 的功能表示。输出处于“conv5”各自的特征图。在此路径中，在下一个卷积层的单个图像滤池被连接到那些特征图在前面的层，作为绿色图虚线示出。2. 非线性加到每个卷积层的输出。特征图的尺寸被逐渐通过使用 2 的步幅所有卷积层降低。 M-CNN 和在[12]的基础设施之间的最重要的区别是，M-CNN 删除池层。虽然池是用于增强翻译不变性的物体识别是有用的，它失去所必需的准确预测的标签[10]的位置的精确的空间信息。有关网络的参数，例如图像/特征图的尺寸，核尺寸的细节/数字示于图 2.单个图像卷积路径的强大表示能力奠定了匹配的置信度和位移的精确估计的基础。十字交叉图像卷积路径：交叉图像卷积路径位于图的中间一行。2.在输出“conv5”层交叉图像特征图。第 j 层 XCJ 在第 m 交叉图像特征图，由卷积相应匹配滤波器（包括三种组分 FIJ，P，M，FRJ，Q，M 和 FCJ，T，M）与两个烧毛生成米图像（XI J-1，p，xR 的 J-1，q）的和交叉图像特征图中（XC J-1，t）的 - 1 个层。该 FIJ，P，M 组件连接到第 p 次（所有的 P）输入图像特征图曦 J-1，P 在 J - 1 个层 XCJ，男。类似地，FRJ，Q，M 组件链接第 q（所有的 Q）KNN 区域特征映射 xR 的 J-1，Q 到 XCJ，男。此外，FCJ，T，M 组件链接第 t（出所有 T）的交叉图像特征映射 XC J-1，T 为 XCJ，男。从一个层中的匹配滤波器到下一个的操作被示出为在图紫色虚线。 2.在数学上，横特征图 XCJ，m 被计算如下：

其中，*表示卷积和 BJ，m 是第 m 个输出地图的偏差。最大值（0，·）是非线性激活 FUNC 化，并且被操作元件明智。从式。（2)我们可以看到，在接下来的层交叉图像特征图通过考虑在前面的层单和交叉图像特征图，并且因此 I 和 KNN 区域 GKL 能够有效地估计输入图像之间的位移计算。注意，与 M-CNN 的沿，所述单个图像和交叉图像卷积路径的不同层的接受网络连

接的视场逐渐增加[32]。以这种方式，多范围的匹配可以实现的。
最后，我们融合来自两个单图像路径和一个交叉图像路径的特征图。更具体来说，在输入图像和 KNN 区域（从单个图像卷积路径）的特征图的绝对差异是第一个计算出的，然后被堆叠与交图像卷积路径的输出。我们的融合施加在特征图。它是由“连体”[2]架构，其计算表示的绝对差不同。实验表明，在我们的融合策略通过使更多的空间信息优于“连体”。

3.3. 后期处理

定由 M-CNN 推定的匹配 CON 组和位移，输入图像的解析结果可以计算如下。首先，含有 1 个标签的我置信通过平均计算匹配置信度为 CKL 满足升∈ϕ（GK）所有 KNN 地区。如果该置信度超过阈值ξ1 越大，标签 L 被预测为在输入图像中可见，否则预测为不可见的。其次，我们预计在可见标签的位置。更具体来说，匹配区域的输入图像中的坐标我基于匹配位移 TKL 和 GKL 的坐标被计算。然后，我们变形 GKL 的相关地面真标签掩模到匹配区域在一以这种方式，我们得到每个标签 L∈[1，L]的概率地图 M1 的我。我们像素明智的最大值都毫升的所有标签，并得到前台概率。比ξ2 被视为粗糙前景的阈值的概率较大的象素，其余的是粗糙的背景。粗糙的前景和背景由滤波器尺寸 10 进一步削弱生产网络最终前景和背景的种子。基于种子，可以得到由该算法[8]背景概率。得到的背景概率与概率前景映射表 M1，L∈[1，L]，在此基础上结合起来，我们可以得到的逐像素的最大后验概率（MAP）分配一个人类最初的解析结果。最后，我们再过段我用熵率基于分割算法[18]尊重实际语义标签的边界，并受到了广大的覆盖像素的最初的解析结果的分配超像素的标签。

4. 实验

4.1. 实验设置

数据集：我们使用的数据集[15]像素明智地由 18 个类别去网络由每日照片集定义标记为[4]。该数据集包含 7700 图像（6,000 培训，700 验证和 1000 进行测试）。我们采用 4 个评价指标，即准确度，平均准确率，平均召回和平均 F-1 的分数超过像素[30]。训练图像生成双：要减少过度的模型训练拟合和部分解决检测误差，我们用 1 和 1.2 倍放大裁剪人为中心的图像区域。我们也水平镜像图像。总之，每一个图像具有 4 变型和训练数据可被大大增强。对于每 6000 个训练图像，50 KNN 图像从图像库中检索。每个训练图像，每个区域 KNN 形成训练对。不均匀采样训练对平衡不同的标签后，我们应受科幻有 5 个万双，其中甚至租税的 ILSVRC2012 [12]。我们舒夫 F　　训练对上，以增加每个时代的多样性。实施细则：我们的来自框架[11]下实现了 M-CNN 和使用随机梯度下降的有 128 例子批量大小，0.9 的动力和重量衰减 0.0005 训练它。我们使用的所有图层平等的学习速度。学习速率由当验证错误率停止与当前学习率降低 10 分手动调节。学习率在 0.0005 初始化。我们培养的 M-CNN 的大约 50 时代，这需要 11 至 12 天在一个 NVIDIA GTX TITAN。在训练阶段，我们第一个计算出意味着和匹配置信和整个图像语料的位移和方差元素明智地正常化由均值和方差的输出训练。测试时，我们的项目匹配的置信度和位移由 M-CNN 估计的均值和方差的绝对值。在后处理步骤中，ξ1 和ξ2 被设定的阈值在 0.8 和 0.5。KNN 区域的每个输入图像的数量被设置为 9。

4.2. 结果与分析

随着国家的最艺术的比较：我们比较我们的 M-CNN 基于准参数人解析框架有两个国家的的艺术：等。 [30]和空

间 [29]。我们用自己的公开可用的代码和相同的6000训练图像作为我们的公平比较的方法培养他们的模型。我们不与Dong等进行比较。 [4]，因为它们的代码是不公开的和他们的方法据报道，比我们慢。对于所有标签的

平均结果示于表1等人的方法。[30]和空间 [29]上训练同一6000训练图像和测试同1000图像作为M-CNN，他们的平均得分F1达到41.80％和44.76％。我们的“M-CNN”显着超过21.01％，山口等优于这两种基准。[30]和用于空间 [29] 18.05％。“M-CNN”也给出了在前台的准确性巨大的推动作用：两个基线达到55.59％，山口等人。 [30]和空间62.18％[29]，而“M-CNN”获得73.98％。 “M-CNN”还获得高得多的精确度（64.56％对37.54％为[30]和用于52.75％[29]），以及更高的召回（65.17％对51.05％为[30]和用于49.43％[29] ）。此VERI音响ES提供的M-CNN基于准参数框架的有效性。我们还提出了F1-分数在表2一般而言每个标签，“M-CNN”显示出比基线高得多的性能。在小型语义地区，如帽子，腰带，包包和围巾预测标签而言，我们的方法实现大增益，例如43.38％和11.43％[30]和2.95％[29]的围巾，57.87％和24.53％[30]，30.52％[29]的袋子，38.45％和14.68％[30]和16.94％[29]皮带。这表明，我们的准参数网络能有效地捕捉标签之间的内在联系，并有力地预测各种服装款式和姿势标签口罩。

我们的网络的消融：我们还广泛探索不同的CNN的架构，以更透明地展示在M-CNN各成分的有效性。 M-CNN的体系结构示于图。 2和其它变体是通过逐渐加入/消除在不同的层的交图像滤池构成。 M-CNN包含4层交叉的图像匹配滤池CONV2到。“M-CNN（5,4,3,2,1）”是由“CONV1”层到“M-CNN”加入11×11×6交叉图像滤池获得。所添加的匹配滤池被施加输入图像和KNN区域的RGB通道构成的层叠图像。此外，我们将继续以消除层交叉图像匹配滤池层，生产“M-CNN（跨5,4,3）”，“M-CNN（跨5,4）”，“M-CNN（5跨）“和”M-CNN（W / O）“。请注意，没有匹配滤池是在使用“M-CNN（W / O）”架构，其中“W / O”代表没有。对于公平的比较，我们保持每一层不变不同的M-CNN变化特征图的数量。因此，除去跨图像滤池的数目被均匀地加入到相应的两个单图像层。例如，要素的数量映射为单和交叉图像卷积路径是“CONV2”所有30个。之后，我们去掉交叉配滤池得出“M-CNN（跨5,4,3）”，“M-CNN（跨5,4）”和“M-CNN（跨5）”，功能映射数所述单个图像中的卷积路径被设置为45.另外，与名为“连体”的经典基于CNN 音响阳离子架构比较[2]，这是两个相同的子网络的一个复合结构。两个子网络的输出是完全连接的线性

层的反应，其绝对差异计算功能于F T中的网络连接最终匹配置信度和位移。最后，我们与比较“M-CNN W / O SS”，这与M-CNN一样除了超像素平滑处理步骤被跳过。平均成绩超过以下意见表1中提供的所有标签中。逐步增加跨图像匹配滤池进更多的卷积层产生4的变化“M-CNN”，其中“M-CNN（W / O）”，“M-CNN（跨5）”，“M-CNN（5 ，4）“，”MCNN（5,4,3）“和”M-CNN“。他们的F1评分从56.99％，58.07％，60.36％提高到62.81％。最高的F1得分是由“M-CNN”达到62.81％。逐步完善的性能验证，在插入更多的交叉图像匹配成多个卷积层可以帮助实现更好的匹配。在“CONV1”图层添加一个交叉图像匹配内核下降，从62.81％的F1比分61.53％。究其原因，相对较低的结果是，对应于第一个交叉图像匹配内核的接受场很小，只涉及一个语义标签的某些部

分。然而这项工作，即匹配置信度和位移，的目标是定义在语义标签层次去连接，从而超出插入“CONV1”层交叉图像匹配内核的接受场。当在 M-CNN 的生长越深，接受网络连接的视场变得更大，具有较高
两者均以覆盖语义标签，其可以推定语义标签级的位移。此外，“MCNN（W / O）”的性能比“连体”更好，因为我们的最终任务是估计位移。“M-CNN（W / O）”计算两个“conv5”层的特征图之间的差异，而“连体”计算两个完全连接的功能，其中，在 2 暗淡的图像的空间结构被部分丢失的差异。下部 F1 分数“M-CNN W / O SS”与“M-CNN”相比，证明了采用超像素平滑技术可以更好地保护边界信息虽然是像素“标签的简单和快速的投票。性能优越的“M-CNN W / O SS”比国家的最艺[30，29]表明，我们的 M-CNN 有直接预测更可靠的标签口罩，即使没有后处理步骤的能力。
可以归因于该可靠标签从 KNN 区域转移。例如，在第一个栏三个图像的包，围巾和帽子都成功 MCNN 位于而是由空间 P 完全错过。另一个例子是 M- CNN 的成功科幻 NDS 的第一个行中的小太阳镜，这是由空间错过。此外，我们可以观察到的 M - CNN 的结果是稳健姿势变化。作为底行所示，M- CNN 可以准确估计左右臂的位置，而空间不能。优越的性能，因为 MCNN 是一个模型，而空间依赖于一个独立的姿态估计的预处理。另一个观察是，我们的分割区域是更完整而空间区域是分散的，如右下结果。这是因为空间基于缺乏明确的语义传输的标签。

五，结论和未来工作

在这项工作中，我们将处理由一个新的准参数模型人解析。我们的单向网络版准参数框架继承了参数和非参数方法分析的优点。表征多不等的匹配，我们提出了一个匹配卷积神经网络，它包含了更好的特征表示并在交叉图像匹配滤池被嵌入到卷积层的交叉图像卷积路径中的两个单个图像卷积路径。大量的实验结果清楚地表明，从对国家的最艺术的准参数模型显着的性能提升。在未来，我们将扩展框架以其它基于典范的任务，如面部解析。此外，我们计划使用其它更多的电网结构，例如，GoogLeNet [24]。 确认这项工作由中国国家自然科学基金项目（No.61422213，61332012，61328205）和 100 人才，中国中国科学院计划的支持。它也被从 Adobe 礼物资金的支持下，中国的院士，批准号 XDA06010701 的战略重点研究项目，中国的高技术研究发展计划（no.2013AA013801），广东省自然科学基金（no.S2013050014548 ），以及科学技术的广州珠江之星（no.2013J2200067）的计划。

References

[1] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In ECCV, 2012.
[2] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In CVPR, 2005.
[3] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan. Towards unified human parsing and pose estimation. In CVPR, 2014.
[4] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan. A deformable mixture parsing model with parselets. In ICCV, 2013.

[5] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. TPAMI, 2013.
[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv preprint arXiv:1311.2524, 2013.
[7] S. Gould, J. Zhao, X. He, and Y. Zhang. Superpixel graph label transfer with learned distance metric. In ECCV, 2014.
[8] V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman. Geodesic star convexity for interactive image segmentation. In CVPR, 2010.
[9] B. Hariharan, P. Arbel´aez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In ECCV, 2014.
[10] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler. Learning human pose estimation features with convolutional networks. arXiv preprint arXiv:1312.7302, 2013.
[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093, 2014.
[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classi fication with deep convolutional neural networks. In NIPS, 2012.
[13] D. Kuettel and V. Ferrari. Figure-ground segmentation by transferring window masks. In CVPR, 2012.
[14] D. Kuettel, M. Guillaumin, and V. Ferrari. Segmentation propagation in imagenet. In ECCV, 2012.
[15] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, L. Lin, and S. Yan. Deep human parsing with active template regression. 2015.
[16] M. Lin, Q. Cheng, and S. Yan. Network in network. In ICLR, 2014.
[17] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. TPAMI, 2011.
[18] M.-Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa. Entropy rate superpixel segmentation. In CVPR, 2011.
[19] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan. Fashion parsing with weak color-category labels. TMM, 2014.
[20] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In CVPR, 2012.
[21] J. Long, N. Zhang, and T. Darrell. Do convnets learn correspondence. In NIPS, 2014.
[22] P. H. O. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In ICML, 2014.
[23] J. A. R. Serrano and D. Larlus. Predicting an object location using a global image representation. In ICCV, 2013.
[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. arXiv preprint arXiv:1409.4842, 2014.

[25] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In ECCV, 2010.
[26] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In CVPR, 2013.
[27] F. Tung and J. J. Little. Collageparsing: Nonparametric scene parsing by adaptive overlapping windows. In ECCV, 2014.
[28] X. Wang, L. Zhang, L. Lin, Z. Liang, and W. Zuo. Deep joint task learning for generic object extraction. In NIPS, 2014.
[29] K. Yamaguchi, M. H. Kiapour, and T. L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In ICCV, 2013.
[30] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In CVPR, 2012.
[31] W. Yang, P. Luo, and L. Lin. Clothing co-parsing by joint image segmentation and labeling. In CVPR, 2014. [32] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. arXiv preprint arXiv:1311.2901, 2013.