综述报告：CVPR 2018 Best Paper——Taskonomy: Disentangling Task Transfer Learning

严凯　学号：21721299

推荐理由

去年的这篇CVPR最佳论文研究了一个非常新颖的课题，它既不是经典的CV问题如**分类**、**识别**、**定位**、**检测**、**跟踪**、**分割**等，也不是近年来较为新颖的如**风格迁移**，**VQA**，**基于文本的图像生成**、**由图像生成场景知识图**等视觉问题，而是 视觉任务之间的关系研究 ——根据得出的关系可以帮助在不同任务之间做**迁移学习**。文章的终极目标，是通过计算任务相似性，进一步计算选取（不是手动设计）针对目标任务的multi-task组合进行训练，并实现以少量数据尽可能接近 `fully supervised learning` 性能。这相比于之前已经看惯了的针对各种单个单个任务的研究，这篇论文可谓是计算机视觉领域的一股春风。

关键词

任务空间 、 相关性 、 避免冗余 、 迁移学习

创新性求解思路

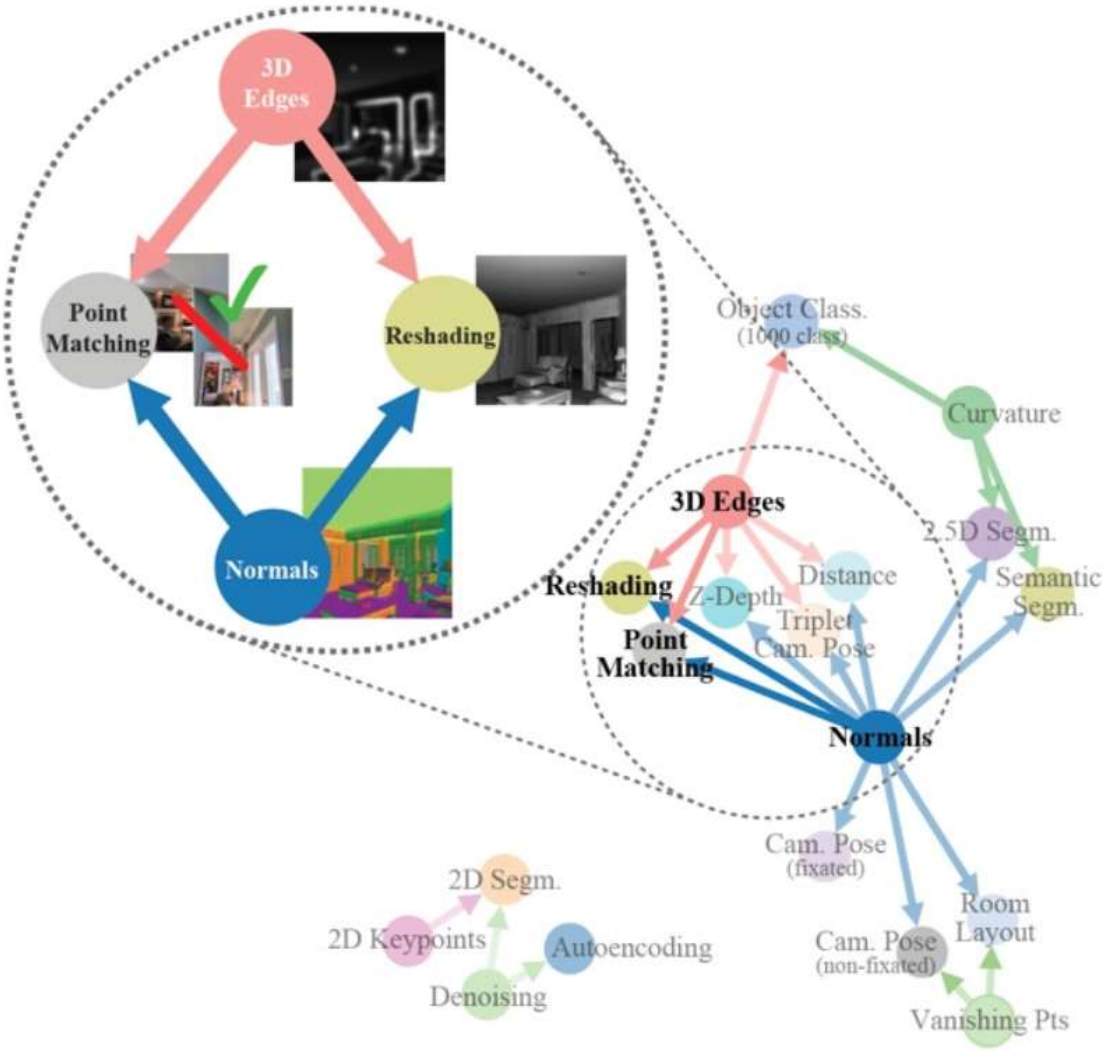
- 在谈及本文的创新性求解思路之前，不妨先看一下贯穿本文始终的也是作者提出的最基本、根本的设想：**视觉任务之间有什么关系吗，还是说它们都是各自独立的?**
  根据以往的经验，其实不难发现：答案是**肯定**的——比如深度估计和表面法线预测之间的关系（法线 (Surface Normals) 是由深度 (Depth) 求导得来），语义分割 (Semantic Segmentation) 又似乎和遮挡边缘测试 (Occlusion edge detection) 有着千丝万缕的关联。不管是从我们的直觉上还是借助一些知识，比如我们知道表面法线预测模型、深度预测模型或者室内布局模型都可以为物体识别带来不小的帮助：所以任务之间肯定是有一些关系的。
- 长远来看，计算机视觉着眼于解决大多数甚至所有视觉任务，但现有方法大多尝试将视觉任务逐一击破。这种方法造成了两个问题：第一， 逐一击破需要为每一项任务收集大量数据，随着任务数量的增多，这将会是不可行的；第二，逐一击破会带来不同任务之间的冗余计算和重复学习。总的来说，逐一击破的策略忽略了视觉任务之间的**关联性**。
- 这篇论文尝试达到的目的：希望能有效测量并利用视觉任务之间的关联来**避免重复学**

习，从而用更少的数据学习我们感兴趣的一组任务。也即，这篇文章并非是针对某个具体的视觉问题（任务）进行探讨，而是想找出并 量化 此前计算机视觉领域中存在的大量学习任务之间的各种关联，并尝试通过这些关联来使得在执行一组具体的学习任务时尽可能减少训练所需的数据量，大概就是这个意思。

- 相比于上面我自己的概括，论文中更规范的表述为：

> 如果两个视觉任务A、B具有关联性，那么在任务A中习得的representations理应可为解决任务B提供有效的统计信息 。

比如，假定有预测法线的网络和预测遮挡边缘测试的网络，我们可以通过结合两个网络的representations来快速通过少量数据解决Reshading和点匹配 (Point matching)。基于这些关联，本文利用BIP (Binary Integer Programming) 求得对于一组我们感兴趣的任务，如何去最优分配训练数据量。



左上方的圆圈中即展示了“通过将曲面法线估计器和遮挡边缘检测器学习到的特征结合起来，用少量标记数据就能快速训练用于Reshading和点匹配的优质网络”。

- 更进一步，作者论述了这些关系的影响以及它们具有的重要作用。
  四个要点：
  - 任务之间的关系是存在的（上面也已经说过了）
  - 这些关系可以通过计算性的方式得到，不需要我们人类的知识参与
  - 各种任务属于 一个有结构的空间 ，而不是一些各自独立的概念
  - 它可以为我们提供一个用于迁移学习的统一化的模型

其中第三点不仅是对上面**关联性**的补充说明，也强调了这些所有的关联性存在于一个统一的底层结构中。这也说明了任选一些任务出来我们都可以问这样的问题：它们之间有没有关系、有多大关系。为了回答这些问题，就要对任务之间的关系、任务之间的冗余有一个全局的认识，需要把任务作为一个集体来看待，而不是作为单个单个的任务。这样我们才能利用它们之间的关系和冗余度达到更高的效率。

- 迁移学习 ——实际上，迁移学习之所以可行就是因为任务间的这些关系。从高抽象层次上讲，如果能够迁移或者翻译一个模型学到的内部状态，这就有可能会对学习解决别的任务起到帮助 —— 如果这两个任务之间存在某种关系的话。

本文的方法

简单概括，方法分为**两个大阶段**，**四个小步**。第一大阶段涉及前三小步，作者量化了不同视觉任务之间的关联，并将任务关联表达成一个 `affinity matrix` （关联矩阵）。第二大阶段，也就是最后一步，作者对求得的affinity matrix进行**最优化**，求得如何最高效地去学习一组任务。这个最高效的策略会由一个指向图 (directed graph) 来表示，在这篇文章中，它被称为 **`Taskonomy`** 。

要解决的问题及相关定义

- 要解决的问题：在有限的监督预算 γ 下最大化我们在一组目标任务 (target tasks) $\mathcal{T} = \{t_1, \ldots, t_n\}$ 上的表现。
- 相关定义：
  - 一组**起始任务** (source tasks) $\mathcal{S}$，其定义为我们可从零学习的任务。
  - **监督预算** γ ，定义为多少起始任务我们愿意从零开始学习。由于从零开始学习需要收集大量数据，所以监督预算表达了我们所面对的金钱、计算力和时间上的限制。
  - $\mathcal{T} \setminus \mathcal{S}$ 代表了我们感兴趣但不能从零学习的任务，比如一个只含少量训练数据的任务。
  - $\mathcal{S} \setminus \mathcal{T}$ 代表了我们不感兴趣但可以从零学习的任务，如 `jigsaw` 、 `colorization` 等自我监督的视觉任务。从零开始学习可以帮助我们更好的学习 $\mathcal{T}$ 。
  - $\mathcal{T} \cap \mathcal{S}$ 代表了我们感兴趣也能从零学习的任务，但考虑到从零学习会消耗监督预算，所以我们希望从中选择出符合预算的一组任务从零学习，余下的通过少量数据的迁移学习来实现。
  - $\mathcal{V} = \mathcal{T} \cup \mathcal{S}$ 称为任务词典 (task dictionary)。
  - 视觉任务 t 定义为一个基于图片的方程 f_t 。

数据集选取

这篇文章收集了一个有四百万张图片的数据集，每张图片均有26个不同视觉任务的标注 (ground truth)。这26个任务涵盖了2D的、3D的和语义的任务，构成了该项research的任务词典。因为这26个任务均有标签， $\mathcal{S}$ 也就是这26个任务。

具体步骤

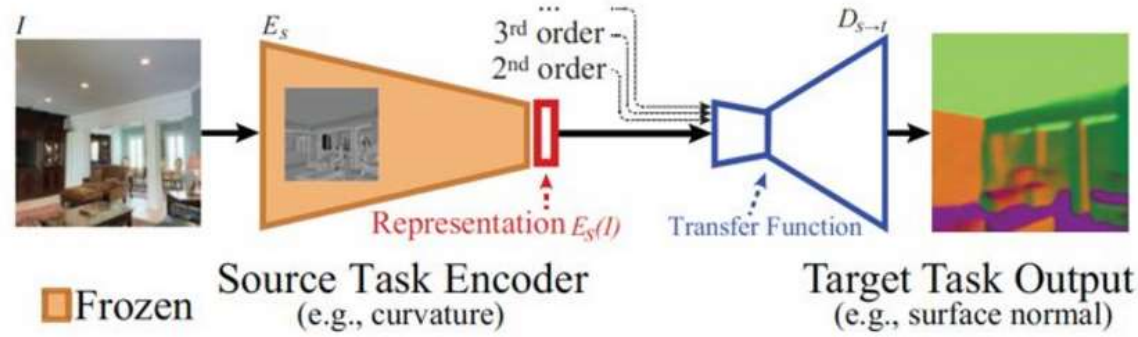
- step1：从零学习
  对于每个起始任务，我们为其从零开始学习一个神经网络。
  为了能更好地控制变量从而比较任务关联，每个任务的神经网络具有相似的encoder和decoder结构。所有的encoder都是相同的类 `ResNet 50` 结构。因为每个任务的输出维度各不相同，所以decoder的结构对不同的任务各不相同，但都只有几层，远小于encoder的大小。值得一提的是，作者指出：

  > decoder泛指read out functions，比如classification的FC Layers也算为decoder。

- step2：迁移学习
  对于一个起始任务 $s \in \mathcal{S}$ 和一个目标任务 $t \in \mathcal{T}$，**以 s 的representation作为输入来学习 t**。我们将freeze任务 s 的encoder参数，并基于encoder的输出 (representations) 学习一个浅层神经网络read out function。严谨来讲，如果我们用 E_s 表示 s 的encoder，f_t 表示 t 的标注，L_t 表示 t 的loss函数，$I \in \mathcal{D}$ 来表示图片和迁移训练集，D 表示要迁移学习的浅层神经网络，学习目标为：

  $$D_{s\to t} := \arg\min_\theta \mathbb{E}_{I\in\mathcal{D}}\left[L_t\left(D_\theta\left(E_s(I)\right), f_t(I)\right)\right]$$

  对于所有 s 和 t 组合，作者都训练了一个 $D_{s\to t}$。对于 t，不同的 $E_s(I)$ 会对 $D_{s\to t}$ 的表现造成不同的影响。更具关联的 s 会为 t 提供更有效的统计信息，从而仅用1/60的训练数据（相较于从零学习）就能取得不错的结果；相反不具备关联的 s 则并不能有此表现。因此，作者得出结论：$D_{s\to t}$ 在 t 任务中的表现可以很好地代表了 s 之于 t 的关联性。
  此外，针对**多对一**的高阶关联，作者在文章中根据一阶的表现（所谓一阶就是上面的**一对一**关联），对于小于五阶的高阶将前五的所有组合作为输入，对于5阶以上的高阶选择结合一阶表现前n的起始任务作为输入。

  论文中的大量插图有更直观的体现，我这里就不再多放了。
- step3：基于序数的规范化
  这一步的目标是用一个 `affinity matrix` 量化任务之间的关联。其中还用到了运筹学中的层次分析法（Analytic Hierarchy Process），我只是略微了解了一下，这里就不提了，反正也不是本文提出的创新思路。
  从不同任务迁移到目标任务 t 所产生的loss因为 t 的不同而不在一个量级上，需要进行归

一化。用论文原话来说：

> 虽然从上步习得的迁移网络中我们获得了许多的loss值 $L_{s\to t}$ ，但因这些loss值来自于不同的loss 函数，它们的值域有很大差别。如果我们把这些loss值直接放入一个矩阵，那么这个矩阵内的值及其不均匀，并不能有效反应任务之间的关联。同时，简单的线性规范化也并不能解决问题，因为任务的loss值和表现并不构成线性关系（0.01的 l_2 loss并不代表其表现两倍好于0.02的loss）。

作者采用了上面提到的运筹学中的层次分析法（AHP）进行量化。
首先对于每一个目标任务构建pairwise tournament矩阵W_t，其纵轴和横轴均对应所有的起始任务及我们计算过的高阶组合。给定一个测试集D_{test}，每一个元素$w_{i,j}$表示在测试集中，有多大几率s_i到t比s_j到t误差更小（有几成 $I \in D_{test}$ 会使 $L_t(D_{s_i\to t}(I)) < L_t(D_{s_j\to t}(I))$）。然后计算$W'_t = W_t/W_t^T$，现在$W'_t$的第$(i,j)$项表示$s_i$的表现比$s_j$好几倍。于是：

$$w'_{i,j} = \frac{\mathop{\mathbb{E}}_{I\in D_{test}}[D_{s_i\to t}(I) > D_{s_j\to t}(I)]}{\mathop{\mathbb{E}}_{I\in D_{test}}[D_{s_i\to t}(I) < D_{s_j\to t}(I)]}$$

在把 W'_t 规范化成数值和为1的矩阵后，我们将 s_i 相对于 t 的关联性（抑或可迁移性）定义为 W'_t 的第 i 项principal eigenvector。将所有目标任务的 W'_t 合并起来，我们获得最终的affinity matrix P。

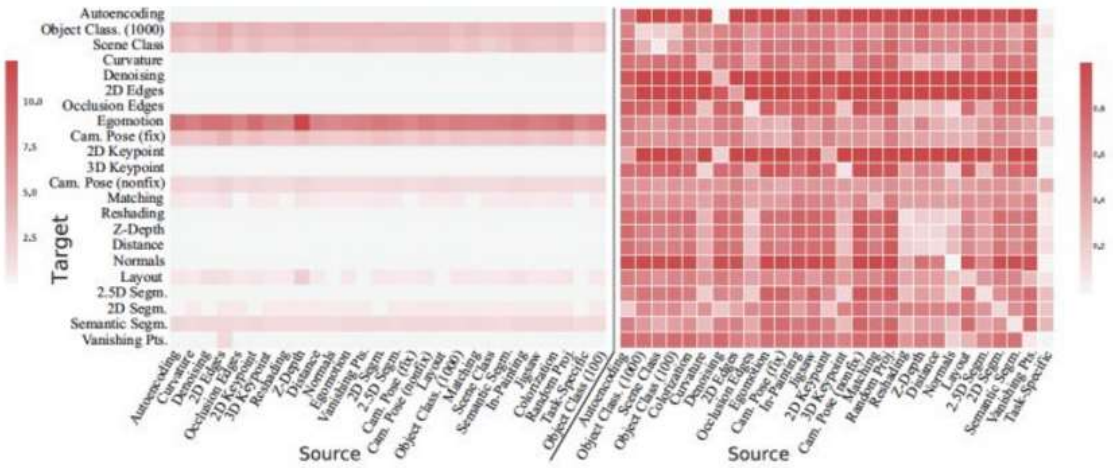


Figure 7: **First-order task affinity matrix** before (left) and after (right) Analytic Hierarchy Process (AHP) normalization. Lower means better transfered. For visualization, we use standard affinity-distance method $dist = e^{-\beta\cdot P}$ (where $\beta = 20$ and e is element-wise matrix exponential). See supplementary material for the full matrix with higher-order transfers.

至此作者通过上述affinity matrix量化了任务之间的关联性。

- step4：BIP (Binary Integer Programming) 最优化
  利用affinity matrix，从任务字典里抽取一组任务和迁移方式，**等价于一个子图选取问题。**
  子图选取问题可以抽象成一个BIP问题求解。文章针对这个最优化问题给了三个约束：
  - 如果选了一种迁移，那么该迁移的源任务和目标任务都要出现在子图中；
  - 每个目标任务有且只有一个迁移（将从零学习在图中定义为从自己到自己的迁移，即一条自己到自己的edge）；

- 监督信息加和不超过预算。

这三个限制条件的具体数学表达式我就不在这里列出来了。
至此，本文已经通过解最优subgraph selection获得了最有效迁移学习策略，如下图所示：

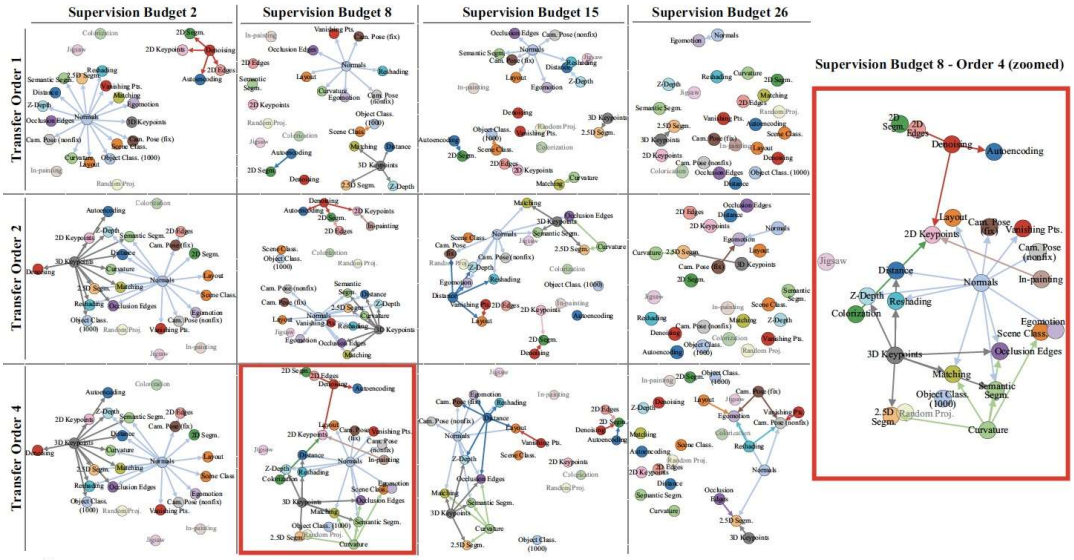


Figure 8: **Computed taxonomies** for solving 22 tasks given various supervision budgets (x-axes), and maximum allowed transfer orders (y-axes). One is magnified for better visibility. Nodes with incoming edges are target tasks, and the number of their incoming edges is the order of their chosen transfer function. Still transferring to some targets when tge budget is 26 (full budget) means certain transfers started performing better than their fully supervised task-specific counterpart. See the interactive solver website for color coding of the nodes by *Gain* and *Quality* metrics. Dimmed nodes are the source-only tasks, and thus, only participate in the taxonomy if found worthwhile by the BIP optimization to be one of the sources.

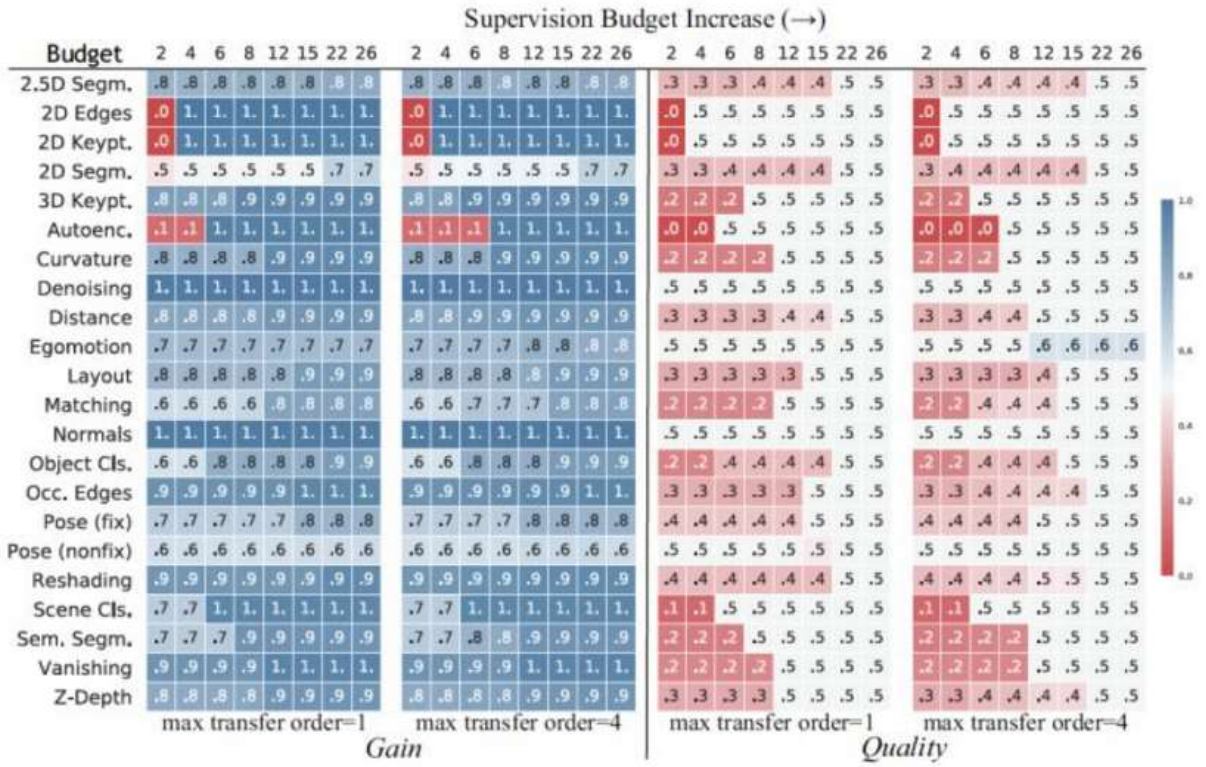
实验结果

先列一下作者在论文中的结果：
Taskonomy项目训练了3000+个神经网络，总耗时～50000小时的GPU。从零学习消耗120k张图片，迁移学习为16k张图片。

评判标准

- **迁移获利** (Gain) : 如果我们不进行迁移学习，我们只能基于少量的数据从零学习。迁移获利是指迁移学习相较于从零学习的胜率（见Ordinal Normalization部分）。
- **迁移质量** (Quality) : 用少量数据迁移学习相较于用大量数据从零学习的胜率。

图表



Taskonomy的迁移获利 (左) 和质量 (右) 的图表。两图的纵轴为所有目标任务，横轴为监督预算，胜率在0-1之间。

结论

可见，对于一个26个任务的目标集，在只有一半甚至1 / 3的监督预算时，Taskonomy 计算出的监督分配会使整体表现远远打败从零学习（迁移获利），并近似于（胜率超过40%）大量数据完全监督学习（迁移质量）。

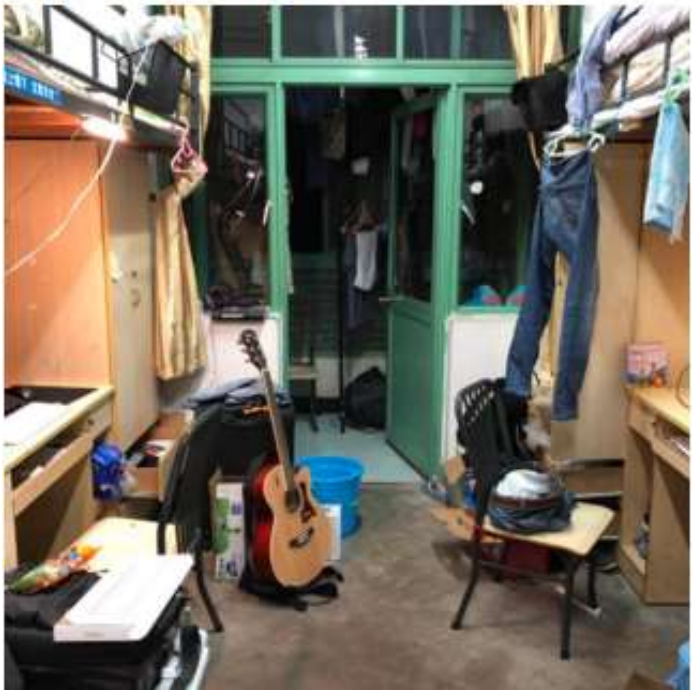
我的总结

最后总结一下：
本文向着认识视觉任务空间的目标迈出了积极的一步，这点比起以往把视觉问题分成单个单个的任务有了本质性的拓展——作者把任务作为一个结构化空间中的群体，而不是作为各自独立存在的概念来看待；同时这是一个完全可计算化的框架，它可以帮助我们做迁移学习，也可以帮助我们寻找通用化的感知模型。

我自己的实验测试

测试图片：（宿舍实时取景）

Input Image

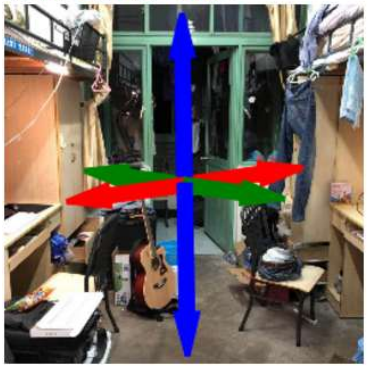


- Surface Normals

Image Reshading

2D Texture Edges

Vanishing Points

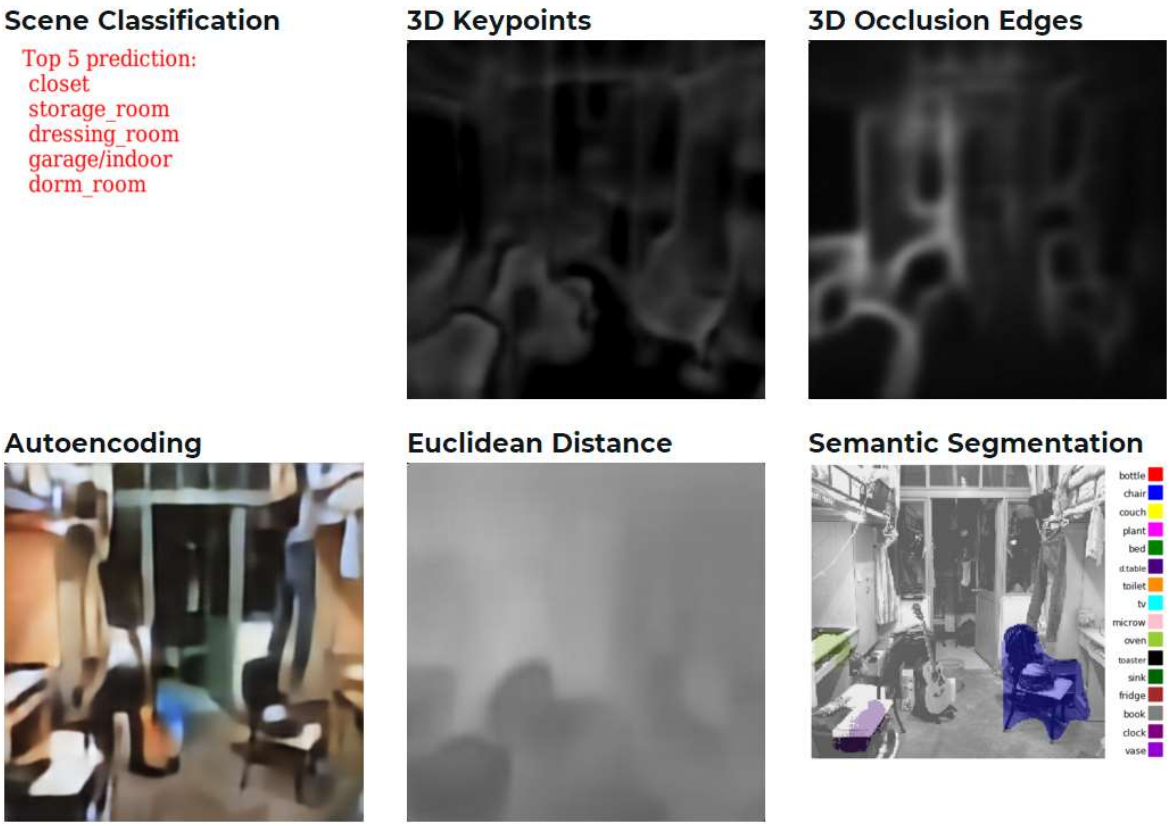


Unsupervised 2.5D Segm.

Room Layout

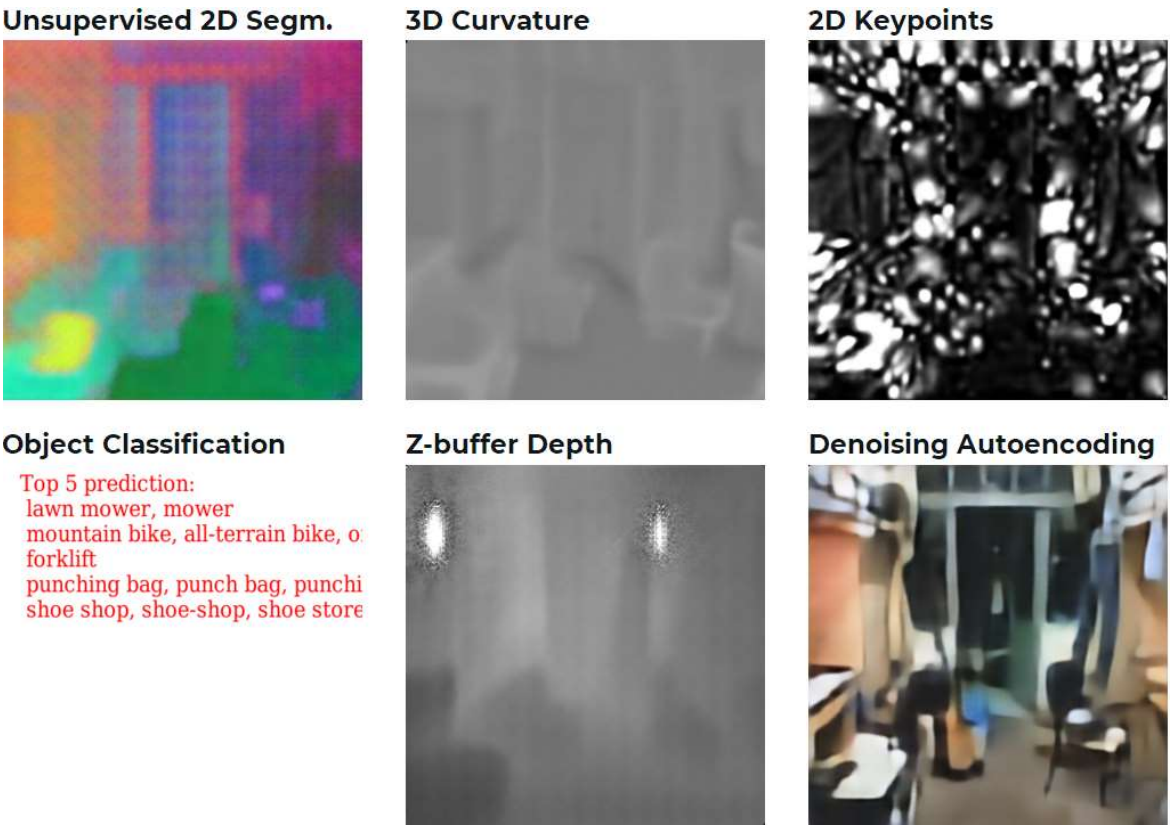
`Unsupervised 2.5D Segm.` 几乎不能看。。。。。。

`Room Layout` 也明显有偏差。



关于 `Scene Classification` 的 `Top 5 prediction`，预测出来的5个条目分别是：壁橱、储藏室、更衣室、室内车库和宿舍，其中只有第一条和最后一条正确，考虑到我在的宿舍在取景时并没有整理过房间，一是物件比较多，二是比较乱，所以精度不高也能接受。

至于 `2D Texture Edges`，识别度还是相当高的。



Colorization

Image In-painting