DIGITOUR：房地产物业的自动数字导览

Prateek Chhikara
Housing.com 古尔冈，印度
prateek.chhikara@housing.com

Harshul Kuhar
Housing.com 古尔冈，印度
harshul.kuhar@housing.com

Anil Goyal
Housing.com 古尔冈，印度
anil.goyal@housing.com

Chirag Sharma
Housing.com 古尔冈，印度
chirag.shar
ma@housing.com

arXiv:2301.06680v1 [cs.CV] 17 Jan 2023

摘要

虚拟导览或数字导览是一种虚拟现实技术，允许用户远程体验特定地点。目前，这类虚拟导览采用两步策略制作。首先，摄影师拍摄一张 360◦ 等距圆柱投影图像；随后，标注人员团队手动关联这些图像，以实现用户的“漫游”体验。虚拟导览大规模普及面临的主要挑战，是人工标注和关联图像所耗费的时间与成本。因此，本文提出了一套端到端流水线，用于自动生成房地产物业的三维虚拟导览，所用图像为等距圆柱投影图像。我们提出了一种新颖的基于 HSV 的配色方案，用于纸质标签；在使用 360◦ 相机拍摄等距圆柱投影图像之前，需将这些标签放置在不同位置。这些标签具有两个特点：*i)* 它们带有编号，便于摄影师按顺序放置标签；*ii)* 它们采用双色设计，有助于在图像中更好地完成标签检测（使用 YOLOv5 架构）和数字识别（使用定制的 MobileNet 架构）任务。最后，我们根据检测到的标签关联所有等距圆柱投影图像。我们通过在 Housing.com 数据库中采集的真实世界等距柱状投影图像数据集，展示了所提流程的效率。

CCS 概念

• **计算方法 → 目标检测；目标识别。**

关键词

计算机视觉、房地产、虚拟导览、数字识别。

ACM 参考格式：

Prateek Chhikara、Harshul Kuhar、Anil Goyal 和 Chirag Sharma。2023年。DIGITOUR：面向房地产物业的自动**数字导览**。载于第六届数据科学与数据管理联合国际会议（第*10*届 *ACM IKDD CODS* 和第*28*届 *COMAD*）（*CODS-COMAD 2023*），*2023*年*1*月*4*日至*7*日，印度孟买。 ACM，纽约，美国纽约州，5页。https://doi.org/10.1145/ 3570991.3571060

允许免费制作本作品的全部或部分内容的数字版或纸质副本，供个人或课堂使用，但不得为营利或商业利益而制作或分发副本，且副本必须附有本声明和首页完整引文。对于本作品中由 ACM 以外的其他方拥有版权的组件，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、重新出版、发布到服务器或重新分发至列表，则须事先取得具体许可和/或支付费用。请向 permissions@acm.org 申请许可。*CODS-COMAD 2023*，*2023*年*1*月*4*日至*7*日，印度孟买© 2023年计算机协会。ACM ISBN 978-1-4503-9797-1/23/01…$15.00 https://doi.org/10.1145/3570991.3571060

1 引言

过去几年，由于互联网的使用更加便捷，在线房地产工具的需求大幅增长，尤其是在印度等国家 [13]。许多在线房地产平台（例如 Housing.com、Proptiger.com、Makaan.com 等）供业主、开发商和房地产经纪人发布待售和出租房产。每天，这些平台都会收到 8，000 至 9，000 条新房源。迄今为止，用户在这些平台上浏览图像、快照或视频，而这些内容可能无法建立足够的置信度，帮助用户做出决定并完成交易。为应对这一挑战并提升用户体验，虚拟导览是一种潜在的解决方案。虚拟导览由相互关联的图像构成，让观看者能够远程体验特定地点（一个常见的例子是谷歌街景 [1]）。近来，虚拟导览的需求不断增长，因为它们能更好地与用户/顾客互动，尤其适用于房地产 [11], 酒店与餐馆 [6], 大学与学校等行业。总体而言，虚拟导览分为三类：*i)* 二维视频导览、*ii)* 360◦视频虚拟导览，以及*iii)* 360◦ 静态图像虚拟导览。与二维视频导览和 360◦ 视频导览相比，基于静态等距柱状图像的虚拟导览能带来更强的沉浸感和交互性，从而有助于做出更好的决策，并避免不必要的看房。



图 1：在房屋入口处拍摄的等距圆柱投影图像示例，展示了客厅。

通常，创建虚拟导览的流程由以下组件构成：

i) **等距柱状图像采集：**等距圆柱投影图像是一种将球形物体表示为二维图像的方式（如图 1 所示）。基本上，它是一幅球形全景图，垂直视角为 180°，水平视角为 360°。一个简单的例子是将地球（球形物体）投影到二维地图上。这些图像是

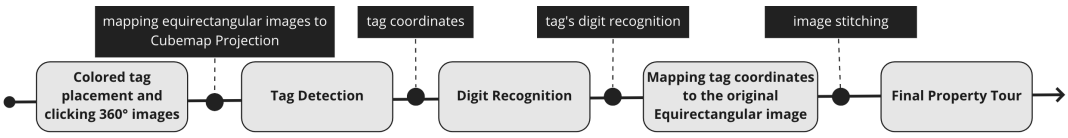

图2：所提出方法的端到端流程DIGITOUR

使用 Ricoh-Theta1、Insta3602 等 360。相机拍摄的。

ii) 连接等距圆柱投影图像：对于任何地点，我们都会有多张等距柱状投影图像。例如，在房地产中，我们通常会有卧室、客厅、厨房、餐厅等房间的等距图像。要获得完整的“漫游”体验，必须在图像之间建立导航连接。此外，从一个位置到其他位置可能有多条路线。比如，我们可以从客厅前往厨房、卧室、阳台等。因此，连接所有等距圆柱投影图像至关重要（示例请参见图3）。通常，这项工作需要手动完成，既成本高昂又耗时 [10]。

iii) 发布虚拟导览：在建立等距圆柱投影图像之间的连接后，我们可以将最终的虚拟导览发布到云端或应用程序中。

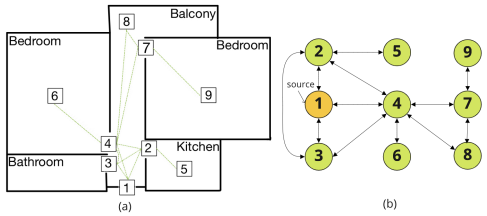

图 3：物业平面图示例（a）及其用于构建数字导览的连接关系（b）。数字表示标签和等距圆柱投影图像的位置。

在实现上述流程自动化时，主要挑战之一是需要人工标注等距圆柱投影图像之间的连接关系。通常，创建一次虚拟导览需要 20 分钟以上。为解决自动创建数字导览的难题，我们提出了一种面向房地产等距圆柱投影图像的端到端流水线（称为 DIGITOUR ）。DIGITOUR 流程由以下组件构成：

i) 彩色标签放置并拍摄 360° 图像：我们提出了一种新型纸质标签，它们采用双色设计并带有编号，可帮助下游计算机视觉任务（即标签识别和数字识别）更好地学习，并实现等距圆柱投影图像的自动拼接。

ii) 将等距柱状投影图像映射为立方体贴图投影：我们使用公开可用的 Python 库 vrProjector[3] 将等距柱状投影图像映射为对应的立方体贴图投影（对应六个立方体面）。

[1]https://theta360.com/en/etc/technology.html [2]
https://www.insta360 2 3 360 .com/product/insta-onex
https://github.com/bhautikj/vrProjector

iii) 标签检测：对于每个立方体面，我们提出使用 YOLOv5 [5] 架构检测图像中的彩色标签。**iv) 数字识别：**我们提出使用轻量级定制 MobileNet 模型进行数字识别。 [8]。

最后，我们利用检测到的标签连接所有等距柱状投影图像。我们在两个数据集（“绿色标签”和“双色标签”）上，对各项任务（标签检测和数字识别）以及端到端流水线进行了实验验证。实验结果表明，端到端流水线在所有类别上的加权平均 mAP 和 F1 分数分别为 88.12 和 95.81，IoU 阈值为 0.5。下一节将介绍所提出的算法 DIGITOUR，随后第 3 节将呈现实验结果。



数字	颜色	HSV 值
0	绿色	(112°, 59%, 76%)
1	Red	(3°, 84%, 84%)
2	紫罗兰色	(266°, 73%, 85%)
3	棕色	(25°, 82%, 58%)
4	Pink	(316°, 59%, 87%)
5	Grey	(0°, 0%, 64%)
6	黄色	(43°, 72%, 89%)
7	浅蓝色	(194°, 62%, 87%)
8	深蓝色	(224°, 97%, 93%)
9	橙色	(25°, 79%, 94%)

图4：为每个数字设计的双色标签格式和配色方案，以及对应的 HSV 值。

2 DIGITOUR

本节介绍生成数字导览的完整流程（简称 DIGITOUR，如图2所示）。

2.1 标签放置与图像采集

为任何房地产物业创建数字导览时，必须从房产的不同位置（如卧室、客厅、厨房等）拍摄 360。图像，然后将这些图像自动拼接在一起，从而无需亲临现场即可获得“漫游”体验。因此，为了连接多张等距柱状投影图像，我们建议在房产各处的地面上放置纸质标签，并将相机（在我们的案例中使用的是 Ricoh-Theta）放在场景中央，以拍摄整个场地（前、后、左、右和下方）。此外，我们确保场景中没有昏暗照明等干扰因素以及“不需要的”伪影，以便更好地进行模型训练和推理。如图4所示，我们将标签的尺寸标准化为

6”×6”，并赋予其两项特性：*i)*标签带有编号，便于摄影师按顺序放置标签；*ii)* 标签采用双色设计，从而将数字识别问题构建为分类任务，并促进下游计算机视觉任务（即标签检测和数字识别）的学习。请注意，数字 0 到 9 分别采用 HSV 色彩方案中的不同颜色，并且如图4所示，标签的首位数字带有黑色圆圈，以便与末位数字区分开来。将纸质标签标准化的原因在于，这样可以训练出不受失真、标签放置角度、光源反射、模糊状况和相机质量影响的标签检测与数字识别模型。

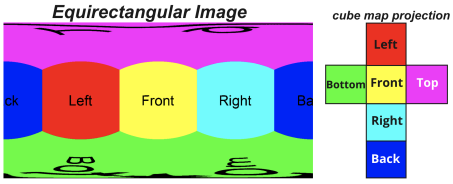

图 5：将等距圆柱投影图像转换为对应的六个立方体贴图投影面。

2.2 将等距圆柱投影图像映射到立方体贴图投影

等距圆柱投影图像由单张图像构成，其宽高比为 2:1（如图 1 所示）。在我们的案例中，图像使用尺寸为 $4096 \times 2048 \times 3$ 的 Ricoh-Theta 相机拍摄。通常，等距圆柱投影图像中的每个点都对应球体上的一个点，并且图像会沿“纬度”方向拉伸。由于等距圆柱投影图像中的内容存在失真，因此很难直接从中检测标签并识别数字。例如，图 1 中的标签在图像底部中央被拉伸。因此，有必要将图像映射到失真较小的投影形式，并切换回原始等距圆柱投影图像，以构建数字导览。在这项工作中，我们提出采用立方体贴图投影，它由表示立方体六个面的六张图像组成。在此，球面坐标空间中的每个点都对应立方体某个面上的一个点。如图 5 所示，我们使用 vrProjector Python库，将等距圆柱投影图像映射到尺寸为 $1024 \times 1024 \times 3$ 的立方体的六个面（左、右、前、后、顶和底）。

2.3 标签检测

获得对应立方体各个面的六张图像后，我们会检测每张图像中所放置标签的位置。对于标签检测，我们使用了最先进的 YOLOv 5 [5] 模型。我们先使用 COCO 权重初始化网络，然后在自己的数据集上进行训练。如图 6 所示，该模型以一张图像作为输入，并返回检测到的标签，以及边界框的坐标和预测置信度。模型在我们的数据集上训练了 100 个训练轮次，批次大小为 32。

2.4 数字识别

对于检测到的标签，我们需要识别标签上的数字。在真实环境中，检测到的标签可能存在方向错误

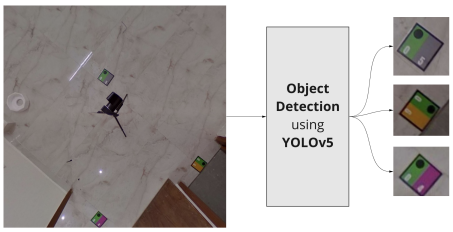

图 6：使用 YOLOv5 进行标签检测。

、光线不足、房间灯泡反光等问题。由于这些原因，使用光学字符识别（OCR）引擎很难获得良好的数字识别效果。因此，我们使用了一个基于 ImageNet 权重初始化的定制 MobileNet [8] 模型，该模型利用标签中的颜色信息进行数字识别。在所提出的架构中，我们将原始 MobileNet 的最终分类模块替换为丢弃层和全连接层，其中包含 20 个节点，分别代表从 1 到 20 的标签。图 7 展示了所提出的架构。训练模型时，我们使用 Adam 作为优化器，学习率为 0.001，折扣因子（ρ）为 0.1。我们使用分类交叉熵作为损失函数，并将批次大小设为 64，训练轮次设为 50。

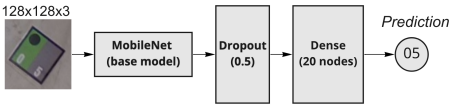

图 7：使用自定义 MobileNet 模型进行数字识别。

2.5 将标签坐标映射到原始360° 图像并创建虚拟导览

检测到标签并识别出数字后，我们使用 Python 库 vrProjector 将立方体贴图坐标映射回原始等距圆柱投影图像。图 8 展示了一个输出示例。对于每张等距圆柱投影图像，检测到的标签构成图中的节点，节点之间由边连接。在同一房产的后续等距圆柱投影图像中，随着检测到更多标签，图中会加入更多节点。最后，我们根据图像上标注的已识别数字，依次连接多张等距圆柱投影图像，由此得到的图就是图 3(b) 所示的虚拟导览。

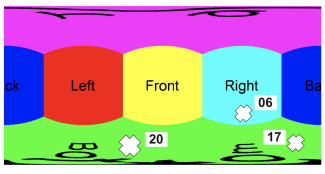

图 8：将标签映射到原始等距圆柱投影图像。

3 实验

本节将通过一项实证研究，展示我们的 DIGITOUR 流程的性能。

3.1 数据集

我们在印度古尔冈（一个一线城市）的多处住宅房产中放置标签，并使用 Ricoh-Theta 相机拍摄等距柱状投影图像，以此收集数据。拍摄图像时，我们确保满足了若干条件，例如所有房门均已打开、灯光均已开启、移除了“不需要”的物品，并放置标签以覆盖房产的每个区域。遵循这些要求后，每处住宅房产平均拍摄了 7 或 8 张等距柱状投影图像。最后，我们在以下生成的数据集上验证了我们的方法（根据标签的背景颜色划分）。

• **绿色标签** （）：我们将些标签（编号为 1 至 20）的背景颜色设为绿色。我们从 212 处房产中收集了 1572 张等距柱状投影图像。将这些等距圆柱投影图像转换为立方体贴图投影后，我们得到 9432 张图像（对应于立方体面）。由于并非所有立方体面都有标签（例如顶面），因此我们得到 1503 张至少带有一个标签的图像。

• **所提出的双色标签** （）：对这些标签，我们从 350 处房产中收集了 2654 张等距柱状投影图像。最终，我们得到 2896 张至少带有一个标签的图像（对应于立方体面）。

最后，我们使用 LabelImg[4] 标注立方体贴图投影图像中的标签。LabelImg 是一款开源工具，可用于以 Pascal VOC 和 YOLO 等多种格式标注图像。所有实验中，我们保留 20% 的数据用于测试，其余数据用于训练。

3.2 结果

3.2.1 评估指标。 按照文献中的做法 [7], ，我们采用准确率（A）、精确率（P）、召回率（R）、F1 分数（f1）和平均精度均值（mAP@k），在不同交并比（IoU）阈值（以 k 表示）下评估标签检测任务。我们将数字识别视为分类任务，因此采用宏平均和加权平均准确率（A）、精确率（P）、召回率（R）和 F1 分数（f1）作为所有类别的评估指标 [3]。

表 2：标签检测性能（%）

标签类型		P	R	mAP@0.5	mAP @0.95
绿色	96.5	98.2		99.2	67.7
双色		98.5	98.4	99.3	71.4

3.2.2 标签检测。我们针对两个数据集训练了 YOLOv5 模型，并采用了与第 2.3 节中讨论的相同实验设置。从表 2 可以看出，YOLOv5 架构在两种彩色标签上的性能相近。改变标签的背景颜色并未显著提升性能。

3.2.3 数字识别。我们以两种方式定义数字识别任务：*i)* 作为 OCR（光学字符识别）任务；*ii)* 作为分类问题。在“绿色标签”数据集上，我们对 PaddleOCR [2] 模型和定制训练的 InceptionV3 [12] 模型进行了实验验证。然而，PaddleOCR 和 InceptionV3 模型获得的 F1 分数分别为 33% 和 24%。最后，我们在“双色标签”数据集上对用于数字识别任务的各种最先进模型进行了实验验证（如表 3 所示）。从表 3 可以看出，与 MobileNet [8], VGG19 [9] 和 ResNet50 [4]相比，InceptionV3 [12]模型表现最佳。此外，InceptionV3 和 MobileNet 架构在双色标签数据集上的性能相近。不过，InceptionV3 和 MobileNet 模型的单张图像平均响应时间分别为 8.26 毫秒和 2.74 毫秒。因此，我们决定选择 MobileNet 作为数字识别任务的最终模型。

表 3：不同最先进模型在双色标签数据集上的数字识别任务比较。

		宏平均			加权平均		
模型	**A**	**P**	**R**	**f1**	**P**	**R**	**f1**
InceptionV3	99.2	98.1	98.6	98.2	99.2	99.3	99.3
MobileNet	98.9	99.0	98.6	98.8	99.0	98.9	98.9
VGG19	96.1	96.1	94.5	95.2	96.2	96.1	96.1
ResNet50	93.4	90.8	92.8	90.5	94.3	93.4	93.5

3.2.4 端到端评估。本节介绍所提出的端到端流水线的实验结果。对于任何输入图像，我们首先检测标签，最后识别标签上写的数字。由此，我们能够确定真正例（检测并正确读取的标签）、假正例（检测到但读取错误的标签）和假负例（未检测到的标签）。在 0.5 IoU 阈值下，得到的平均精度均值、精确率、召回率和 F1 分数分别为 88。12、93。83、97。89 和 95。81。请注意，所有指标均在 20 个类别上进行加权平均。如果某房产所有等距柱状投影图像中的标签都被检测并正确读取，我们就能获得准确率为 100% 的虚拟导览，因为图中的所有节点都被检测到，并与相应的边连接。在实验中，我们为 94。55% 的房产准确生成了准确率为 100% 的虚拟导览。出现误差的原因是一些彩色伪影被误检测为标签，以及光照条件不佳。

4 结论

我们提出了一种端到端流程（DIGITOUR），用于自动为房地产物业生成数字导览。对于此类物业，我们首先在物业的每个区域放置所提出的双色纸质标签。然后拍摄等距柱状投影图像，并将这些图像映射为失真较小的立方体贴图图像。获得对应于立方体面的六张图像后，我们使用 YOLOv5 模型检测标签的位置，接着使用 MobileNet 模型识别数字。下一步是将检测到的坐标以及识别出的数字映射回原始等距柱状投影图像。最后，我们将所有等距柱状投影图像拼接起来，构建虚拟导览。我们在真实世界数据集上验证了该流程，并表明端到端流程在所有类别上的加权平均 mAP 和 F1 分数（IoU 阈值为 0.5）分别达到 88.12 和 95.81。

[4]https://github.com/tzutalin/labelImg

参考文献

[1] Dragomir Anguelov、Carole Dulong、Daniel Filip、Christian Frueh、Stéphane Lafon、Richard Lyon、Abhijit Ogale、Luc Vincent 和 Josh Weaver。2010。谷歌街景：捕捉街道层面的世界。《计算机》43，6（2010），32–38。[2] Yuning Du、Chenxia Li、Ruoyu Guo、Cheng Cui、Weiwei Liu、Jun Zhou、Bin Lu、Yehua Yang、Qiwen Liu、Xiaoguang Hu 等。2021。PP-OCRv2：超轻量级 OCR 系统的技巧集。*arXiv* 预印本 *arXiv:2109.03144*（2021）。[3] Margherita Grandini、Enrico Bagli 和 Giorgio Visani。2020。多类别分类的度量指标：综述。*arXiv* 预印本 *arXiv:2008.05756*（2020）。[4] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。2016。用于图像识别的深度残差学习。载于电气电子工程师学会计算机视觉与模式识别会议论文集。770–778。[5] Glenn Jocher、Ayush Chaurasia、Alex Stoken、Jirka Borovec、NanoCode012、Yonghye Kwon、TaoXie、Jiacong Fang、imyhxy、Kalen Michael、Lorna、Abhi-ram V、Diego Montes、Jebastin Nadar、Laughing、tkianai、yxNONG、Piotr Skalski、Zhiqiang Wang、Adam Hogan、Cristi Fati、Lorenzo Mammana、AlexWang1900、Deep Patel、DingYiwei、Felix You、Jan Hajek、Laurentiu Diaconu 和 Mai Thanh Minh。2022。*ultralytics/yolov5：v6.1——TensorRT、TensorFlow*、边缘张量处理器和 *Open-VINO* 导出与推理。https://doi.org/10 5281 6222936 。/zenodo。AONAW 2009 [6] Mohammad 和 Hafiz Ismail。。面向旅游目的地的交互式 360度虚拟导览的开发与评估。*J. Inform. Technol. Impact* 9（2009），137–182。[7] Rafael Padilla、Sergio L Netto 和 Eduardo AB Da Silva。2020年。目标检测算法性能指标综述。载于*2020*年国际系统、信号与图像处理会议（*IWSSIP*）。电气电子工程师学会，237–242。[8] Mark Sandler、Andrew Howard、Menglong Zhu、Andrey Zhmoginov 和 Liang-Chieh Chen。2018年。MobileNetV2：倒置残差与线性瓶颈。载于电气电子工程师学会计算机视觉与模式识别会议论文集。4510–4520。[9] Karen Simonyan 和 Andrew Zisserman。2014年。用于大规模图像识别的超深卷积网络。*arXiv*预印本 *arXiv:1409.1556*（2014）。[10] Chinu Subudhi。2021年。前沿360度虚拟导览。https://www.mindtree.com/insights/resources/cutting-edge-360-degree-virtual-tours [11] Mohamad Zaidi Sulaiman、Mohd Nasiruddin Abdul Aziz、Mohd Haidar Abu Bakar、Nur Akma Halili 和 Muhammad Asri Azuddin。2020年。Matterport：疫情COVID-19期间房地产行业的一种全新营销方式——虚拟导览。载于媒体与视觉设计创新国际会议（*IMDES 2020*）。Atlantis Press，221–226。[12] Christian Szegedy、Vincent Vanhoucke、Sergey Ioffe、Jon Shlens 和 Zbigniew Wojna。2016年。重新思考面向计算机视觉的Inception架构。载于电气电子工程师学会计算机视觉与模式识别会议论文集。2818–2826。[13] Fahim Ullah、Samad ME Sepasgozar、Sara Shirowzhan 和 Steven Davis。2021年。在数字颠覆环境下对在线房地产平台用户感知进行建模：一种集成KANO-SISQual方法。《远程信息处理与信息学》63（2021），101660。