# 3DCV Final Projection

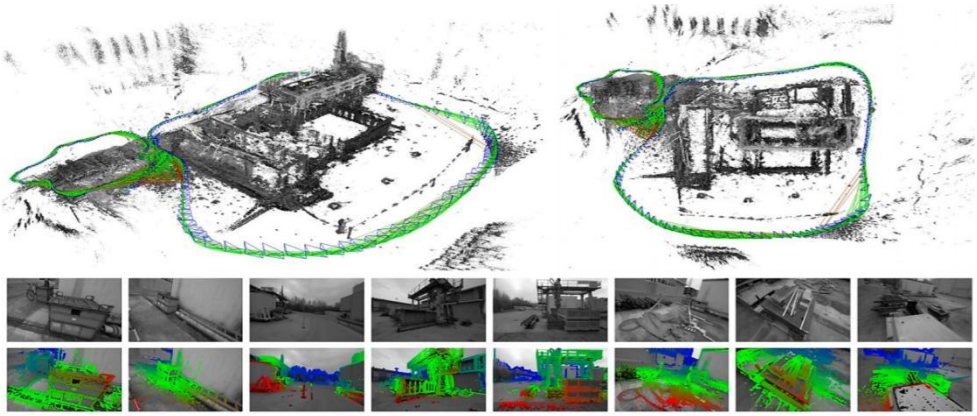
## Camara pose regression using MapNet with RGB-D dual-stream

網媒所碩一 r10944059 俞正楠
網媒所碩一 r10944060 陳政翰

# OUTLINE

1. Introduction
2. Related work
3. Proposed method
4. Experiment
5. Conclusion
6. Reference

# 1. Introduction



相機位姿可以利用在機器人導航，SLAM（Simultaneous localization and mapping），AR（Augmented Reality）等多個應用場景中，因此如何快速準確地獲得一個相機位姿就成爲了一個關鍵問題。傳統方法往往需要較大的計算空間與時間，再加之，深度學習在現如今發展迅速，我們因此希望通過深度學習的方法通過 sensor 獲得的 rgb 和 depth 圖像相對快速地去預測得到一個比較精準的相機位姿信息。那這個問題從本質上來説其實就是要解決一個回歸問題。

在此 project 中使用了一種基於 RGB-D 圖像的絕對相機姿態回歸方法，此方法融合顏色和深度資訊，實現更精確的定位性能。使用雙流網路架構分別處理彩色圖像和深度圖像，再通過此模型去對比單 RGB 與單深度資訊 model。

# 2. Related work

## PoseNet

PoseNet[1]是第一個提出了一種能夠預測出 6 個 DOF 的相機位姿回歸的方法，

它使用了 GoogLeNet 來提取特徵之後通過兩個 FC 層來預測出一個 Pose。以下

是它的 Loss Function：

$$L(I_t) = \|\mathbf{c}_t - \hat{\mathbf{c}}_t\|_2 + \beta \cdot \left\| \mathbf{r}_t - \frac{\hat{\mathbf{r}}_t}{\|\hat{\mathbf{r}}_t\|} \right\|_2$$

其中 β 是一個人爲設定的平衡因子。

# Mapnet

在此 project 中我們有參考 Mapnet[2]網路架構。而 MapNet 除了使用圖像外還利用了例如視覺里程計和 GPS 等低成本且十分普遍的感測器的輸入，並且將它們融合到一起進行相機定位。這些輸入表示傳統上已被用於集束調整或位姿圖優化的幾何約束，在 MapNet 訓練中被設定為損失項，也用於推理過程。除了直接提高定位精度之外，這種方法允許使用來自場景附加的未標記的視頻序列以自監督的方式更新 MapNet。

Mapnet 發現使用單位四元數的對數來替換四元數有更好的表現

$$\mathbf{q} = (u, \mathbf{v}), \mathbf{r} = \log \mathbf{q} = \begin{cases} \frac{\mathbf{v}}{\|\mathbf{v}\|} \cos^{-1} u & \text{if } \mathbf{v} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

在 loss function 中 Mapnet 不僅最小化了單張圖片絕對位姿的損失，同時也計算了相鄰兩張圖片相對位姿的損失。這使得 Mapnet 能夠更好的利用視覺里程計的信息。

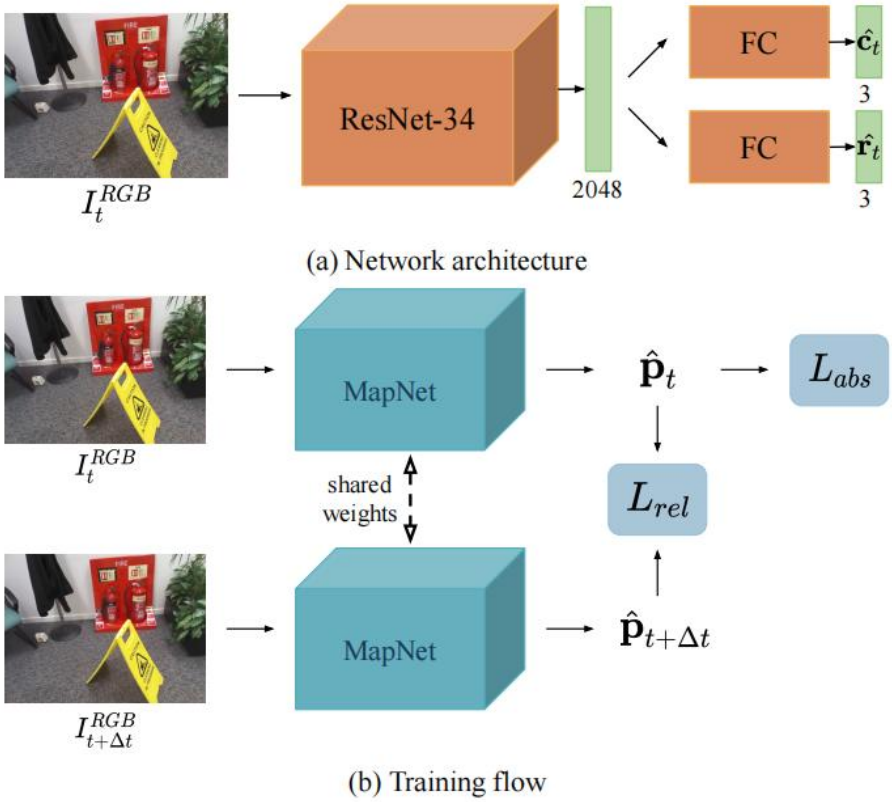
$$L_{abs}(I_t) + L_{rel}(I_t, I_{t+\Delta t}) = h(\hat{\mathbf{p}}_t, \mathbf{p}_t) + h(\hat{\mathbf{v}}_t^{t+\Delta t}, \mathbf{v}_t^{t+\Delta t})$$

$$\mathbf{v}_t^{t+\Delta t} = (\mathbf{c}_t - \mathbf{c}_{t+\Delta t}, \mathbf{r}_t - \mathbf{r}_{t+\Delta t})$$

h(·)為以下函數:

$$L(I_t) = e^{-\hat{s}_c} \cdot \|\mathbf{c}_t - \hat{\mathbf{c}}_t\|_1 + e^{-\hat{s}_r} \cdot \|\mathbf{r}_t - \hat{\mathbf{r}}_t\|_1 + \hat{s}_c + \hat{s}_r$$

下圖為 Mapnet 的架構和訓練機制圖:




(a) Network architecture

(b) Training flow

## Depth Completion

由於 sensor 獲取到的深度圖像往往會有一些無效區域，這些區域沒有一個有效的深度值，因此我們希望通過一種類似顏色填補的方法去填補一個深度值給

予這些無效區域。

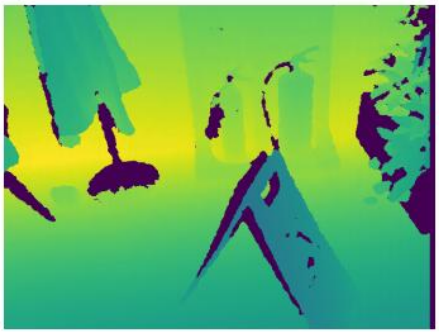
在此方法中我們使用 colorization method[3],此方法最初是針對只需對給定灰度圖像的幾個區域進行顏色注釋即可獲得全彩色圖像的情況而設計的。原始深度圖像類似於要著色的灰度圖像，並且原始深度圖像中的有效值類似於顏色注釋。因此，我們可以將這種著色方法修改為一種簡單的深度補全方法。這種深度完成優化的假設是，具有相似強度的相鄰圖元將具有相似的深度。通過最小化圖元 r 處的深度值與其相鄰 N（r）處深度值的加權和之間的差值來插補缺失值。

$$\min_F \sum_{\mathbf{r}} (F(\mathbf{r}) - \sum_{\mathbf{s}\in N(\mathbf{r})} w_{\mathbf{rs}} F(\mathbf{s}))^2,$$

$$\text{subject to} F(\mathbf{v}_i) = D(\mathbf{v}_i), \forall \mathbf{v}_i \in \{\mathbf{v} | D(\mathbf{v}) \text{ is valid}\}$$

完成深度填補后，我們將單一深度通道擴展到三個通道作爲輸入到特徵提取器中的輸入。

以下圖片為原深度圖像和填補過後的深度圖像的對比效果。



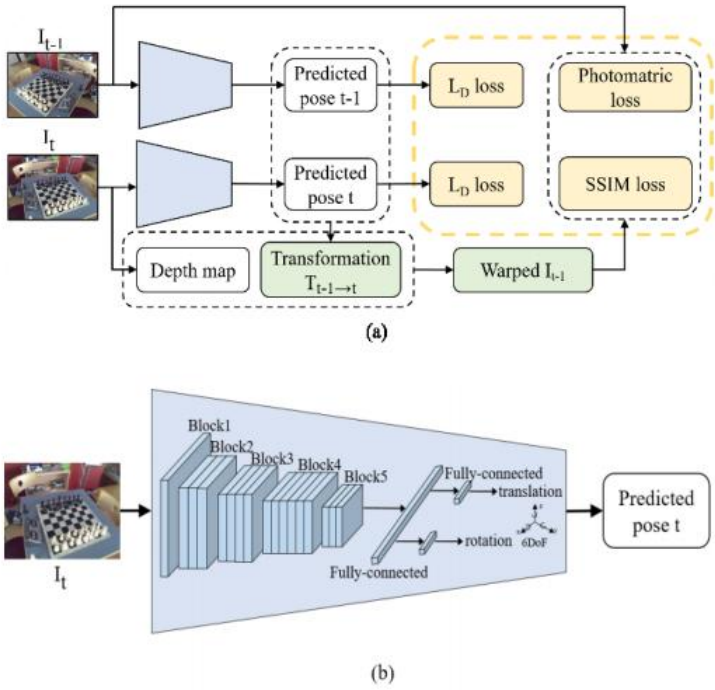
(a) Raw depth image

(b) Full depth image

## 3D Scene Geometry-Aware Constraint

在 loss function 的部分，我們嘗試使用[4]中提到的 loss function，他采用 rgb

圖像作爲預測輸入的同時通過 Depth map 和兩張相鄰序列的圖像得到一個 Transformation，然後把 t-1 時刻的圖像進行一個 warping，得到一張 Warped $I_{t-1}$ 以此去計算一個 SSIM loss，與此同時他還計算了每張圖片的 Photomatric loss。但是在最終實驗中，我們發現該方法實施在我們的架構中表現并沒有特別好，因此最終還是使用了 Mapnet 的 loss function。




以下為該方法的 loss function：

$$L_D = L_D(I_{t-1}) + L_D(I_t)$$

with

$$L_D(I_i) = \|x_i - \hat{x}_i\|_2 + \beta\|q_i - \hat{q}_i\|_2 \quad for\ i \in \{t-1, t\}$$

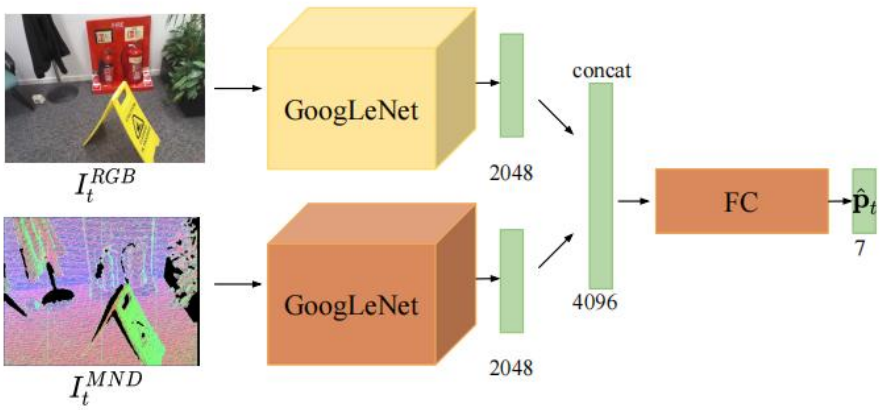
$$L_P = \sum_{i,j} M\left(u_{t-1}^{i,j}\right) \|I_t(i,j) - warped_{t-1}(i,j)\|_1$$

$$L_S = \frac{1 - SSIM(I_t,\ warped_{t-1})}{2}$$

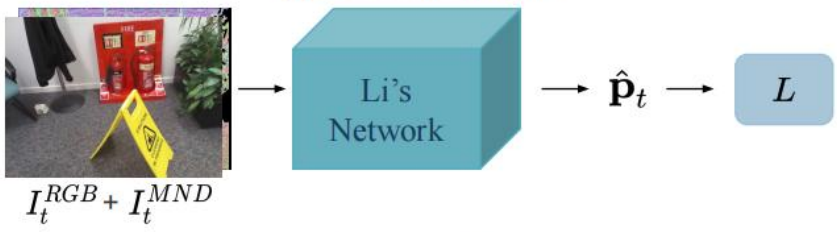
$$L = \lambda_D L_D + \lambda_P L_P + \lambda_S L_S$$

## dual-stream

在這個部分，我們參考了 Li's method[5]，他的方法以彩色圖像和深度圖像作為輸入。為了在將原始深度圖像擴展為三個通道圖像時提供更多的信息，研究人員提出了一種稱為 minimized normal+depth(MND)的編碼方法。他們首先利用深度圖像計算每個像素的單位法向量，然後將歸一化深度圖像和單位法向量的前兩個通道圖像堆疊為三通道 MND 圖像。將彩色圖像和 MND 圖像的特征連接起來，然後輸入到最後一個 FC 層，以回歸相機的姿態。




(a) Network architecture



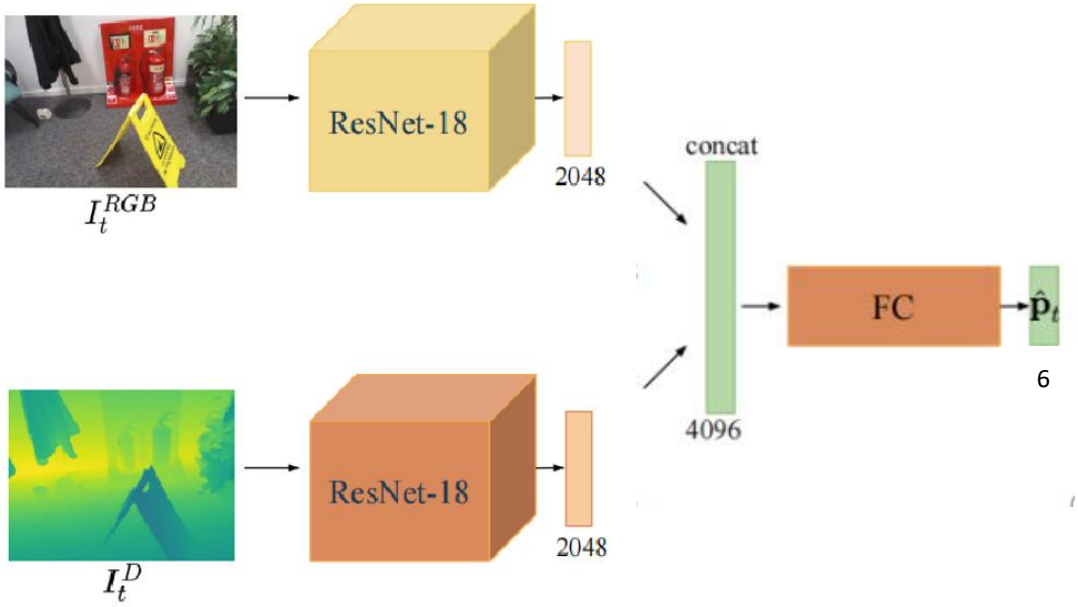

(b) Training flow

# *3. Proposed Method*

## Network Architecture

我們首先使用 Resnet18 作爲特徵提取器來對輸入圖像做一個全局特徵提取，然後對所提出的 RGB-D 雙流網路採用了兩階段訓練機制。在第一階段，顏色流和深度流是獨立的網路，獲得了兩個預測，其中一個僅與顏色資訊相關，另一個

僅與深度資訊相關。第一階段的目標是允許兩條流分別彙聚，以便為下一階段做好準備。第二階段是整個 RGB-D 雙流網路的端到端 train。通過加入第一階段單流模型的結果作爲預訓練模型，雙流網路可以更好地集成顏色特徵和深度特徵，以實現更精確的攝像頭定位。在處理深度圖像時我們還使用一些方法來實現補全深度圖像信息。






## Loss Function

由於 SSIM loss 和 Photomatric loss 在 7scences 數據集中使用效果不佳，在這一部分我們最終采取了 MapNet 和 PoseNet 結合的 loss function。

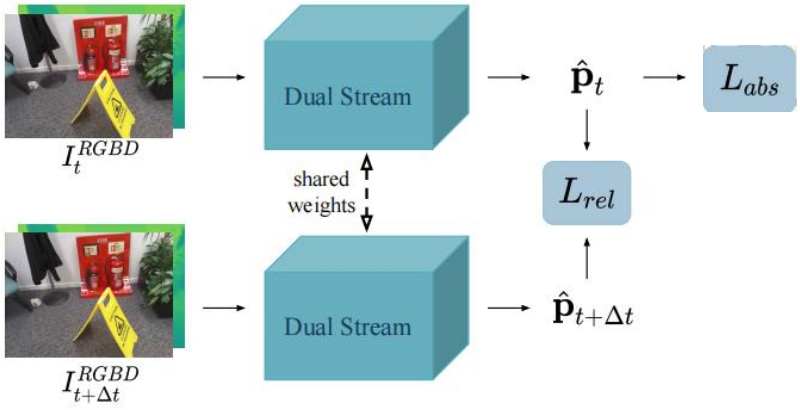
$$L_{abs} + L_{rel} = \sum_{t=1}^{|\mathcal{D}|} h(\hat{\mathbf{p}}_t, \mathbf{p}_t) + \gamma \sum_{t=1}^{|\mathcal{D}|} h(\hat{\mathbf{v}}_t^{t+\Delta t}, \mathbf{v}_t^{t+\Delta t})$$

$$\mathbf{v}_t^{t+\Delta t} = (\mathbf{c}_t - \mathbf{c}_{t+\Delta t}, \mathbf{r}_t - \mathbf{r}_{t+\Delta t})$$

$$h(\hat{\mathbf{p}}, \mathbf{p}) = \|\hat{\mathbf{c}} - \mathbf{c}\|_1 + \beta \cdot \|\hat{\mathbf{r}} - \mathbf{r}\|_1$$

在上述 loss function 中γ和β均爲平衡因子。其中γ設定爲 10，β設定为 3，Δt 設定为 10。

下圖為整個訓練流程的示意圖:




# *4. Experiment*

## Datasets

为了和其他 APR 方法做对比，我们选择了 *7 Scences* 数据集来进行实验。该数据集是由 Kinect 感测器获取的，其中包括了 RGB 和 Depth 图像，下表为该数据集的一些细节。


Figure 4.1: Seven scenes in *7 Scenes* dataset.

Table 4.1: Details of *7 Scenes* dataset.

| Scene | #Frames | | Volume | Ratio of valid depth | | Mean of valid depth | |
|---|---|---|---|---|---|---|---|
| | Train | Test | | Train | Test | Train | Test |
| Chess | 4000 | 2000 | 6 $m^3$ | 80.7% | 86.0% | 181 cm | 170 cm |
| Fire | 2000 | 2000 | 2.5 $m^3$ | 89.5% | 86.6% | 160 cm | 147 cm |
| Heads | 1000 | 1000 | 1 $m^3$ | 84.7% | 82.5% | 84 cm | 82 cm |
| Office | 6000 | 4000 | 7.5 $m^3$ | 86.8% | 86.8% | 188 cm | 192 cm |
| Pumpkin | 4000 | 2000 | 5 $m^3$ | 83.4% | 84.8% | 220 cm | 214 cm |
| RedKitchen | 7000 | 5000 | 18 $m^3$ | 88.7% | 85.9% | 188 cm | 195 cm |
| Stairs | 2000 | 1000 | 7.5 $m^3$ | 86.2% | 86.0% | 189 cm | 177 cm |

# Comparison with Prior Methods

| Scene | PoseNet-17 | Li’s | MapNet | Tian’s | Ours’ |
|---|---|---|---|---|---|
| Chess | 13cm,4.48° | 28cm,7.05° | **8cm,3.25°** | 9cm,4.39° | 11.2cm,4.39° |
| Fire | 27cm,11.3° | 43cm,12.52° | 27cm,11.69° | **25cm**,10.79° | 31cm,**10.21°** |
| Heads | 17cm,13.0° | 25cm,12.72° | 18cm,13.25° | **14cm,12.56°** | 16.9cm,12.66° |
| Office | 19cm,5.55° | 30cm,8.92° | **17cm,5.15°** | 17cm,6.46° | 18cm,5.72° |
| Pumpkin | 26cm,4.75° | 36cm,7.53° | 22cm,**4.02°** | **19cm**,5.91° | 25.5cm,4.73° |
| RedKitchen | 23cm,5.35° | 45cm,9.80° | 23cm,**4.93°** | **21cm**,6.71° | 22.8cm,5.62° |
| Stairs | 35cm,12.4° | 42cm,13.06° | 30cm,12.08° | **26cm**,11.51° | 35cm,**11.46°** |
| Average | 22.9cm,8.12° | 35cm,10.22° | 20.7cm,**7.77°** | **18.7cm**,8.33° | 22.8cm,7.82° |

在實驗部分中，可以看到我們的方法比 Li 和 PoseNet-17 的方法要好，同時

也接近 MapNet 和 Tian 的方法。

# Comparison with dual stream and single stream

| Scene | Only RGB | Only Depth（full） | Dual-stream |
|---|---|---|---|
| Chess | **10.4cm,3.77°** | 15cm,5.53° | 11.2cm,4.39° |
| Fire | **27.4cm**,12.54° | 36.4cm,11.40° | 31cm,**10.21°** |
| Heads | 17.9cm,13.11° | 17.8cm,12.73° | **16.9cm,12.66°** |
| Office | 18.2cm,6.27° | 21.3cm,6.60° | **18cm,5.72°** |

| | | | |
|---|---|---|---|
| Pumpkin | **21.4cm**,5.74° | 33.2cm,6.91° | 25.5cm,**4.73°** |
| RedKitchen | 23.8cm,**5.51°** | 29.6cm,6.59° | **22.8cm**,5.62° |
| Stairs | 29.7cm,11.63° | 45.6cm,13.27° | 35cm,**11.46°** |
| Average | **21.3cm**,8.37° | 28.4cm,9.00° | 22.8cm,**7.82°** |

在與單流模型做對比的過程中我們也發現，我們的雙流模型總體上要表現地更加出色一些，但是可能在個別場景中的表現不如單 RGB-stream 的效果，我認爲這有可能是由於顏色補深度的方法給予了部分錯誤的深度值導致個別場景加入深度信息後的效果不佳。

# *5. Conclusion*

在此次 project 中 我們提出了一種使用 RGB-D 雙流網路來預測相機絕對位姿的方法，并且將其與使用單流(RGB or depth)的方法對比，發現使用了雙流網路之後在實驗結果使用 7scenes datasets 時的表現比使用單流效果總體來的更優。儘管效果沒有能夠超過 Mapnet 和 Tian 的方法，但也能夠與其相比較，未來可以考慮將補充深度值這一步驟加入整個訓練流程作爲一種 end-to-end 的方法，并且思考如何能夠將準確度更加提高。

# *6. Reference*

[1]A. Kendall, M. Grimes, R. Cipolla,"PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization" 2015 IEEE International Conference on Computer Vision.

[2]S. Brahmbhatt,J. Gu,K.Kim,J. Hays, J. Kautz, " Geometry-Aware Learning of Maps for Camera Localization, " in Proceedings of the IEEE international conference on computer vision and Pattern Recognition, pp.2616-2625,2018.

[3]A. Levin,D. Lischinski, and Y.Weiss, " Colorization using optimization, " in ACMSIGGRAPH 2004 Papers, pp.689-694,2004.

[4]M. Tian, Q. Nie, H. Shen, "3D Scene Geometry-Aware Constraint for Camera Localization with Deep Learning" ,ICRA 2020.

[5]R. Li, Q. Liu, J. Gui, D. Gu, H. Hu, " Indoor Relocalization in Challenging Environments With Dual-Stream Convolutional Neural Networks, " IEEE Transactions on Automation Science and Engineering, vol.15, no.2, pp.651-662,2017.