Cas6D : Few shot 6d pose estimation by cascade

出自

題目

Learning to Estimate 6DoF Pose from Limited Data: A Few-Shot, Generalizable Approach using RGB Images (Cas6D)

Introduction

背景

動機

Gen6D：需要太多 (e.g., ≥ 128) reference image

Gen6D：GPU memory issue low feature and voxel resolutions

目的

1. 先取得透過 ViT 處理後的特徵，再去做 Object Detection。

2. 利用 similarity network 找出 top-K reference image
3. cascaded coarse-to-fine refinement module 來偵測出正確 pose
   每一輪都會重建出新的 feature volumne 與 3D voxel，與 Gen6D 不同

相關研究

1. generalizable (model-free) methods,
   a. OnePose [92] and OnePose++ : 先重建出點雲，確認 2D-3D 對應，接著找出 pose

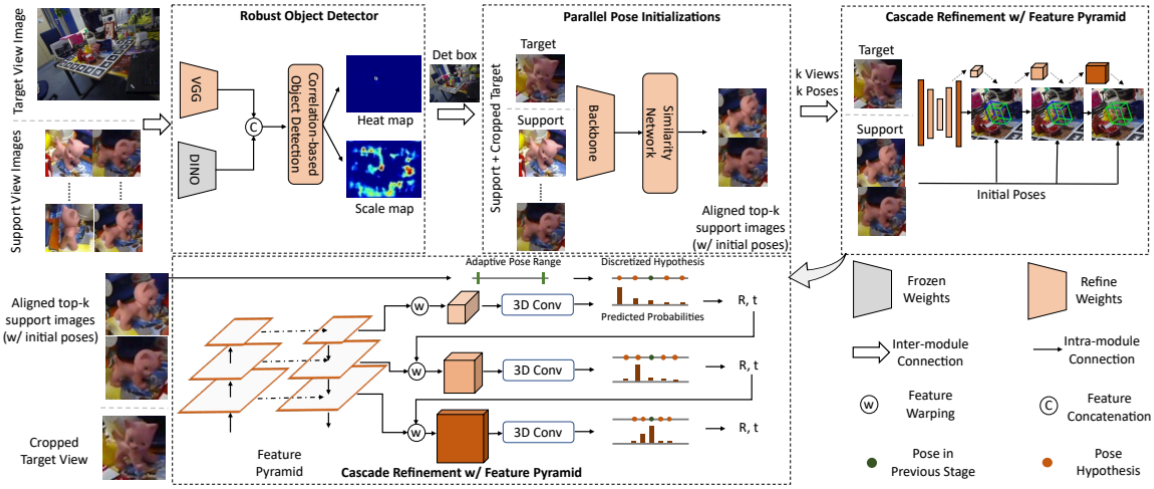
2. few-shot pose estimator 通常有這些做法
   a. local image feature matching
   b. detector-based
   c. detector-free
   d. point cloud registration algorithms : detect 3D keypoints estimate relative transformations

3. coarse to fine architecture :
   a. ZebraPose 是一種 coarse to fine

4. 3D vision Self-supervised methods，通常有這些 tasks
   a. object representation reconstruction
   b. classification
   c. segmentation

方法

1. 找 object box
2. 找 top-K reference image 作為 initial pose
3. discrete pose hypotheses refinement : refine 是從幾個有限的離散 pose 中選擇最佳的出來



robust object detection : 有些 network 會因為 noisy backgrounds or low-textured 導致無法從不同 pose 中找出同一個物體，所以使用 frozen DINO 來抽 feature 找 bounding box

localizing the heat and scale maps

Parallel Pose Initialization : 用 similarity network 從 reference images 中選 top-K 作為 initial pose

cascade feature volume pose refinement

$$\mathcal{L}(\boldsymbol{p}) = \sum_{i=1}^{N_s} \ell(\mathcal{W}(\boldsymbol{I}^{(i)}|\boldsymbol{p}^{(i)}, \boldsymbol{p}), \boldsymbol{I}),$$

p 是當前 predict 的 pose

給定多個 support/reference images I(i) 和它們的姿態 p(i)， 用 warp 把它們轉到 p 並和真實的 target image I 比較差異

W, I 不是唯一的，因為會有 ambiguity ，包含遮擋與對稱; 也會受到背景雜訊影響

Gen6D 直接用一個 neural network 來預測 update 值

volume：32^3 的 voxel ，內涵 6 張 support image 與 target view 做 mean 與 var

最後用 3D ConvNet 來預測 update vector

gen6d : single-scale (top-level) feature representation

對於不同解析度 P2, P3, P4 建立不同解析度的 volumn 16^3, 32^3, 64^3 上面的 feature 長度分別是 64, 32, 16

以 classification 來解 fined 問題，將角度與 transition 量化

然後不斷 focus 在預測的 interval 中 ex: 第一次預測在 50度到 60 度，然後就 focus 在中間，將他們切成 k 格

$$\ell_{pose} = \sum_{k=1}^{N} \lambda_k \cdot \sum_{m=1}^{v_k^3} \|\mathcal{T}_{pr}(p_k \mid k, m) - \mathcal{T}_{gt}(p_k \mid k, m)\|_2 \tag{3}$$

如果某個 stage 猜錯 interval，後面就會很慘

Experiments

可以學習的地方

問題