Introduction to 6D Pose Estimation

Jikai Wang

What is 6D Object Pose Estimation?

- Definition:
  - 6D Object Pose Estimation is a field of computer vision and robotics that determines the **position and orientation** of objects in 3D space using 2D images.
  - Widely used in Robotics, AR or Auto Driving.

Key Concepts

- **Position (3D)**: X, Y, Z coordinates
- **Orientation (3D)**: Roll, Pitch, Yaw (Rotation around X, Y, Z axes)
- **6 Degrees of Freedom (DoF)**: Combination of 3D position and 3D orientation

Single-Frame Pose Estimation vs. Sequence Pose Tracking

- **Single-Frame Pose Estimation**: Determining pose from a single image.
  - Advantages: Simpler, less computationally intensive.
  - Use Cases: Static environments, one-time detection.
- **Sequence Pose Tracking**: Determining pose over a sequence of images.
  - Advantages: More accurate over time, handles motion and changes.
  - Use Cases: Robotics, real-time applications.

Instance-Level vs. Category-Level

- **Instance-Level Pose Estimation**: Recognizing and estimating pose of a specific, known object.
  - Example: Detecting a specific cup on a table.
- **Category-Level Pose Estimation**: Recognizing and estimating pose of objects within a category, not specific instances.
  - Example: Detecting any cup regardless of its specific appearance.

Instance-Level vs. Category-Level

- **Instance-Level Pose Estimation**: Recognizing and estimating pose of a specific, known object.
  - Example: Detecting a specific cup on a table.
- **Category-Level Pose Estimation**: Recognizing and estimating pose of objects within a category, not specific instances.
  - Example: Detecting any cup regardless of its specific appearance.

Model-based vs. Model-free

- **Model-based Estimation**: Uses a pre-defined 3D model of the object.
  - Advantages: High accuracy for known objects.
  - Challenges: Requires detailed models, less flexible.
- **Model-free Estimation**: Does not rely on a specific model, learns from data.
  - Advantages: Flexible, can handle novel objects.
  - Challenges: Requires large amounts of training data.

Seen Objects vs. Novel Objects

- **Seen Objects**: Objects that the system has encountered and learned before.
  - Use Cases: Controlled environments, industrial applications.
- **Novel Objects**: Objects that the system encounters for the first time.
  - Use Cases: Dynamic environments, consumer applications.

FoundationPose

Unified Framework

Pipeline

Language-aided Synthetic Data Generation at Scale (§3.1)
Objaverse
Tag
<thermos bottle>
Original model
Noisy texture
Shape
Diffusion model
"Here is a <thermos bottle>. Please describe its possible appearance including color and styles"
CHATGPT
Repeat
"The thermos bottle is a modern matte gray with a geometric pattern of orange and white"
Augmented model
Physics engine
Path tracing
Neural Unknown Object Modeling (§3.2)
Reference images
Neural object field
Neural RGBD rendering
Geometry Network Ω(·)
Appearance Network Φ(·)
Pose Hypothesis Generation (§3.3)
Global pose sampling
Rendered coarse pose
Encoder
Shared weights
Encoder
H x W x C
Pose conditioned input crop
Conv ResBlock
Tokenization
Position embedding
Transformer encoder
Linear
Rotation update
Transformer encoder
Linear
Translation update
Refined pose
Pose Selection (§3.4)
K hypotheses
Hypothesis 0
Pose conditioned rendering
Pose conditioned input crop
Pose Ranking Encoder
Encoder
Shared weights
Encoder
H x W x 128
Conv ResBlock
Tokenization
Position embedding
P x 512
Multi-head
Self attention
P x 512
AvgPool
512
K x 512
Multi-head
Self attention
K x 512
Linear
Sort
Rank 0
Rank 1
Rank K-1

Neural Implicit Representation

Neural Implicit Representation

- Geometry function (Ω):
  - This function takes a 3D point (x) as input and outputs a signed distance value (s).
  - The signed distance indicates how far the point is from the object's surface.
    - A value of zero signifies the object's surface,
    - Positive and negative values represent points outside and inside the object, respectively.
- Appearance function (Φ):
  - This function takes an intermediate feature vector ($f_\Omega(x)$) from the geometry network, along with the point normal (n) and view direction (d), and outputs the color (c) of the object at that point.

Large-Scale Synthetic Data Generation

Large-Scale Synthetic Data Generation

- **3D Model Databases**:
  - **Objaverse** (Objects Universe): 800,000 3D objects
  - **GSO** (Google Scanned Objects): 1,030 3D objects
- **Large Language Models (LLMs)**:
  - Large Language Model (LLM) refers to a type of AI model designed to understand and generate human language.
  - **ChatGPT**: to create descriptions of the objects and their interactions with light and materials.
- **Diffusion Models**:
  - Diffusion models are a class of generative models that can progressively transform noise into realistic data.
  - **TexFusion**: a novel method for synthesizing textures for 3D geometries using large-scale text-guided image diffusion models.

Pose Hypothesis Generation

Pose Hypothesis Generation

Initial Pose Sampling: A set of viewpoints surrounding the detected object is uniformly sampled. Each viewpoint represents a potential pose for the object.

Pose Hypothesis Generation

Pose Hypothesis Generation

Pose Refinement Network:

- Extract feature maps from the two RGBD input branches.
- The feature maps are concatenated and tokenized.
- Predict the translation update and rotation update.
- This process can be iterative, where the updated pose is used to generate a new rendering, which is then compared with the cropped region for further refinement.

Pose Selection

- After refinement, multiple candidate poses with their corresponding adjustments are available. A pose selection module is tasked with selecting the most accurate pose from this set.
- The pose with the highest score is chosen as the final estimated 6D pose of the object.

THE UNIVERSITY OF TEXAS AT DALLAS

Thank You