Simposio CEA de Robótica, Bioingeniería, Visión por Computador y Automática Marina 2026

Ajuste probabilístico de primitivas geométricas con inicialización guiada por VLM

Menéndez, E.*, Martínez, S., Balaguer, C.

RoboticsLab, Dpto. de Ingeniería de Sistemas y Automática, Universidad Carlos III de Madrid, Av. De la Universidad 30, 28911 Leganés, España.

Resumen

La interacción física segura con objetos cotidianos a partir de una única vista requiere representaciones compactas y coherentes con la geometría de la escena. Este trabajo aborda el modelado de objetos parcialmente visibles mediante el ajuste probabilístico de primitivas geométricas (superelipsoides, supertoroides y superparaboloides de base desplazada) sobre nubes de puntos incompletas. Utilizamos un modelo Gaussiana–Uniforme con asignación probabilística de inliers y regularizaciones físicas, extendido mediante un ajuste secuencial que permite recuperar múltiples primitivas por objeto. Nuestra contribución principal es una inicialización geométrica guiada por un modelo multimodal visión–lenguaje (VLM), que a partir de una única imagen RGB propone una descomposición en partes y, para cada una de ellas, el tipo de primitiva, su orientación y su región de influencia. Estas estimaciones se integran como priors espaciales y condiciones iniciales en la optimización local. Una evaluación cualitativa en escenas de sobremesa muestra mejoras en fidelidad geométrica y consistencia física frente a un ajuste secuencial basado únicamente en inicializaciones por PCA.

Palabras clave: Percepción y sensado, Robots manipuladores, Robótica inteligente, Aprendizaje automático, Fusión de información y sensores

Probabilistic Fitting of Geometric Primitives with VLM-Guided Initialization

Abstract

Safe physical interaction with everyday objects from a single viewpoint requires compact and physically consistent geometric representations. This work addresses the modeling of partially visible objects through the probabilistic fitting of geometric primitives (superellipsoids, supertoroids, and tapered superparaboloids) to incomplete point clouds. We rely on a Gaussian–Uniform model with probabilistic inlier assignment and physical regularizations, extended with a sequential fitting scheme that recovers multiple primitives per object. Our main contribution is a VLM-guided geometric initialization strategy that, from a single RGB image, proposes a part decomposition together with an initial configuration for the primitive type, orientation, and region of influence. These estimates are integrated as spatial priors and initial conditions in the local optimization. A qualitative evaluation on tabletop scenes with partial views shows improved geometric fidelity and physical consistency compared to a sequential fitting pipeline relying solely on PCA-based initialization.

Keywords: Perception and sensing, Robot manipulators, Intelligent robotics, Machine learning, Information and sensor fusion

1. Introducción

La robótica asistencial y de servicio requiere que los robots razonen sobre objetos cotidianos a partir de observaciones parciales, normalmente obtenidas desde un único punto de vista. Estas nubes de puntos incompletas y ruidosas dificultan la recuperación de modelos 3D detallados y limitan la aplicabilidad de muchos métodos orientados a la manipulación (Jiang et al., 2021; Lu et al., 2022). En contraste, los humanos pueden inferir la estructura geométrica de un objeto desde una vista parcial recurriendo a formas geométricas simples (Biederman, 1987).

Este principio ha motivado el uso de primitivas geométricas

*Autor para correspondencia: emenende@ing.uc3m.es

como representación compacta. Mientras que modelos simples como cuboides (Tulsiani et al., 2017) son demasiado restrictivos, las superformas permiten capturar regiones convexas, cavidades y estructuras anulares con pocos parámetros (Paschalidou et al., 2019). Dentro de esta familia, los superelipsoides han destacado especialmente por su uso en trabajos de ajuste geométrico sobre nubes de puntos parciales. Diversos trabajos han explorado su ajuste a nubes de puntos obtenidas desde una sola vista, aprovechando simetrías y regularidades geométricas para inferir partes no observadas (Makhal et al., 2018; Menendez et al., 2024a,b). Sin embargo, la mayoría de estos métodos representan cada clúster segmentado mediante una única primitiva, lo que limita su capacidad para modelar objetos compuestos.

El ajuste de superelipsoides se ha abordado tradicionalmente mediante mínimos cuadrados (Solina and Bajcsy, 2002; Gross and Boult, 1988), aunque estos métodos son sensibles al ruido, a nubes parciales y a mínimos locales. Para mejorar la robustez, Liu et al. (2022) propusieron un modelo probabilístico Gaussiana–Uniforme optimizado con un proceso de estimación y maximización (EM) y complementado con una fase de *switching* que explora parametrizaciones alternativas. Sin embargo, este mecanismo asume superficies casi completas y resulta poco fiable en escenarios de vista única, donde distintas parametrizaciones compatibles con la región visible pueden diferir drásticamente en las zonas no vistas, violar restricciones físicas (como invadir espacio libre) o inducir extrapolaciones inconsistentes.

En nuestro trabajo previo (Menendez et al., 2025) extendimos el enfoque probabilístico de Liu et al. (2022) al caso de observaciones parciales de una sola vista, adaptando el modelo Gaussiana–Uniforme e incorporando regularizaciones físicas como evitar la invasión del espacio libre y respetar el plano de apoyo. También introdujimos un ajuste secuencial de múltiples superformas por objeto mediante la eliminación progresiva de puntos explicados. No obstante, el proceso sigue siendo una optimización local y depende fuertemente de la inicialización de los parámetros de la primitiva: en objetos compuestos o con partes poco separadas, las inicializaciones basadas en PCA pueden ser insuficientes o ambiguas, especialmente cuando la nube no revela claramente qué tipo de primitiva corresponde a cada región.

En este trabajo proponemos una estrategia de ajuste asistida por un modelo multimodal visión–lenguaje que, a partir de una única imagen RGB, infiere la estructura geométrica del objeto y proporciona una inicialización semánticamente informada para cada primitiva. El modelo predice las partes principales, el tipo de primitiva adecuado, su orientación y una región aproximada en la imagen. Estas sugerencias actúan como guía estructural que complementa la nube de puntos parcial, permitiendo iniciar el ajuste en configuraciones más coherentes con la geometría real y reduciendo significativamente el riesgo de converger hacia mínimos locales no deseados.

La combinación de esta inicialización guiada con nuestro ajuste probabilístico permite reconstruir objetos compuestos desde una sola vista de manera más estable y físicamente coherente. En conjunto, el trabajo muestra que integrar un modelo generativo multimodal con un método de optimización geométrica robusta mejora significativamente la reconstrucción de objetos complejos a partir de observaciones parciales.

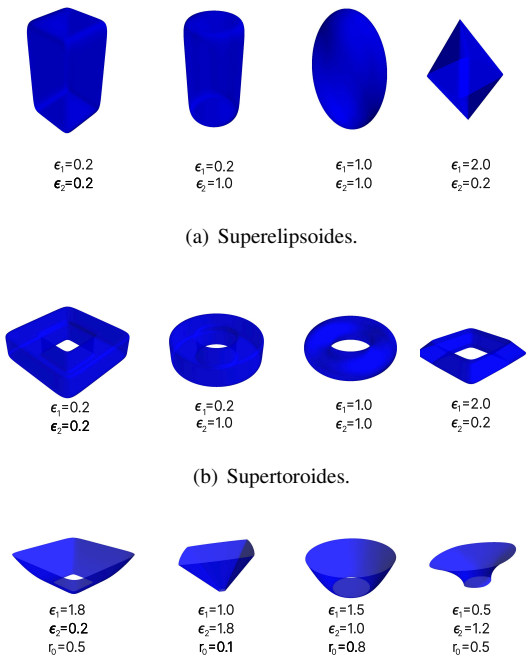
2. Método propuesto

En esta sección se describe el método propuesto para ajustar múltiples primitivas geométricas a nubes de puntos parciales a partir de una única vista RGB. El enfoque combina (i) una representación basada en primitivas geométricas, (ii) un modelo probabilístico que integra un prior espacial por punto, (iii) un esquema de optimización EM suavizado con restricciones físicas, (iv) una inicialización guiada por un modelo visión–lenguaje (VLM), y (v) un ajuste secuencial que permite recuperar múltiples componentes por objeto. A continuación se detallan cada uno de estos elementos.

2.1. Representación geométrica mediante primitivas

En este trabajo representamos las distintas partes del objeto mediante primitivas geométricas, una descripción compacta capaz de modelar diversas formas con pocos parámetros. Consideramos tres familias: superelipsoides para regiones convexas, supertoroides para geometrías anulares y superparaboloides de base desplazada para aperturas y cavidades abiertas.

Cada primitiva se define mediante una función implícita en su sistema de referencia local, cuyos parámetros controlan tamaño y forma. Esta parametrización es lo suficientemente flexible para aproximar diversas geometrías con pocos grados de libertad. La Figura 1 muestra ejemplos representativos de las tres familias y cómo varían según sus parámetros.



(a) Superelipsoides.

(b) Supertoroides.

(c) Superparaboloides de base desplazada.

Figura 1: Modelos generados a partir de superelipsoides, supertoroides y superparaboloides de base desplazada. Cada fila muestra cómo sus parámetros de forma influyen en la geometría resultante.

Superelipsoides. Los superelipsoides son superficies paramétricas capaces de representar formas convexas con distintos grados de redondez. Se definen mediante la función implícita de (1), basada en (Barr, 1981), donde $(x, y, z) = \mathbf{x}_i^{local}$ es el punto en el sistema local de la primitiva. Los parámetros

a_1, a_2, a_3 fijan los semiejes, y los exponentes ϵ_1 y ϵ_2 controlan la curvatura, ambos restringidos al intervalo $(0, 2]$.

$$F(\mathbf{x}_i^{local}; \boldsymbol{\theta}) = \left(\left| \frac{x}{a_1} \right|^{\frac{2}{\epsilon_2}} + \left| \frac{y}{a_2} \right|^{\frac{2}{\epsilon_2}} \right)^{\frac{\epsilon_2}{\epsilon_1}} + \left| \frac{z}{a_3} \right|^{\frac{2}{\epsilon_1}} \tag{1}$$

Supertoroides. Los supertoroides modelan geometrías anulares y superficies con un hueco central mediante la función implícita de (2), basada en (Barr, 1981). Aquí $(x, y, z) = \mathbf{x}_i^{local}$ es el punto en el sistema local, R define el radio mayor y a_r, a_z el tamaño de la sección transversal. Los exponentes ϵ_1 y ϵ_2 controlan la curvatura del contorno y se restringen al intervalo $(0, 2]$.

$$F(\mathbf{x}_i^{local}; \boldsymbol{\theta}) = \left(\left| \frac{\rho}{a_r} \right|^{\frac{2}{\epsilon_1}} + \left| \frac{z}{a_z} \right|^{\frac{2}{\epsilon_1}} \right)^{\frac{\epsilon_1}{2}}, \qquad \rho = \left(|x|^{\frac{2}{\epsilon_2}} + |y|^{\frac{2}{\epsilon_2}} \right)^{\frac{\epsilon_2}{2}} - R \tag{2}$$

Superparaboloide de base desplazada. Los superparaboloides de base desplazada permiten modelar cavidades abiertas cuyo radio aumenta con la altura. A diferencia del modelo estándar, incluyen un término r_0 que introduce un radio inicial distinto de cero en la base (Ferreira, 2018). Su formulación implícita se muestra en (3), donde $(x, y, z) = \mathbf{x}_i^{local}$. Los parámetros a_1, a_2 controlan la escala lateral, a_3 la variación en altura y los exponentes ϵ_1, ϵ_2 determinan la curvatura, con r_0 fijando el radio en la base. Ambos exponentes se restringen al intervalo $(0, 2]$.

$$\begin{aligned} F(\mathbf{x}_i^{local}, \boldsymbol{\theta}) &= \left(\left| \frac{x}{a_1 r(z)} \right|^{\frac{2}{\epsilon_2}} + \left| \frac{y}{a_2 r(z)} \right|^{\frac{2}{\epsilon_2}} \right)^{\frac{\epsilon_2}{\epsilon_1}}, \\ r(z) &= r_0 + \left(\frac{z}{a_3} \right)^{1/\epsilon_1} \end{aligned} \tag{3}$$

La función implícita $F(\mathbf{x}_i^{local}, \boldsymbol{\theta})$ permite determinar si un punto está sobre, dentro o fuera de una primitiva: $F(\mathbf{x}_i^{local}, \boldsymbol{\theta}) = 1$ corresponde a la superficie, $F(\mathbf{x}_i^{local}, \boldsymbol{\theta}) < 1$ al interior y $F(\mathbf{x}_i^{local}, \boldsymbol{\theta}) > 1$ al exterior. En el caso del superparaboloide, la función solo es válida para $z \geq 0$. A partir de los parámetros de escala (a_1, a_2, a_3) y de forma ϵ_1, ϵ_2, y en el caso del superparaboloide, del radio inicial r_0 estas primitivas pueden generar diversas geometrías.

Como las nubes de puntos están en el sistema global, cada primitiva requiere una transformación rígida para expresar los puntos en su sistema de referencia local. Esta transformación combina una rotación $\mathbf{R}$ y una traslación $\mathbf{t}$, según (4).

$$\mathbf{x}_i^{local} = \mathbf{R}^\top (\mathbf{x}_i^{global} - \mathbf{t}) \tag{4}$$

El conjunto completo de parámetros queda entonces definido por $\boldsymbol{\theta} = (\epsilon_1, \epsilon_2, a_1, a_2, a_3, \theta_z, \theta_y, \theta_x, t_x, t_y, t_z)$, incluyendo además r_0 en el caso de los superparaboloides.

2.2. Modelo probabilístico de primitivas geométricas

Para ajustar una primitiva a un conjunto de puntos 3D, modelamos cada observación $\mathbf{x}_i$ mediante una variable latente $z_i \in \{0, 1\}$ que indica si el punto es inlier ($z_i = 1$) u outlier ($z_i = 0$), siguiendo un esquema similar a (Liu et al., 2022; Menendez et al., 2025). La desviación respecto a la superficie se cuantifica mediante un residuo $d_i(\boldsymbol{\theta})$, calculado en el sistema local de la primitiva.

2.2.1. Verosimilitud con prior por punto

Las distancias de los inliers se modelan con una gaussiana isotrópica siguiendo (5).

$$\begin{aligned} p(\mathbf{x}_i \mid z_i = 1, \boldsymbol{\theta}) &= \mathcal{N}(d_i(\boldsymbol{\theta}); 0, \sigma^2) \\ &= \frac{1}{(2\pi\sigma^2)^{3/2}} \exp\left(-\frac{d_i^2(\boldsymbol{\theta})}{2\sigma^2} \right) \end{aligned} \tag{5}$$

Los outliers se asumen uniformes en un volumen delimitador V, tal como se define en (6):

$$p(\mathbf{x}_i \mid z_i = 0) = \frac{1}{V} \tag{6}$$

A diferencia de un modelo con un único peso global, introducimos ahora un *prior de pertenencia por punto* $w_i \in (0, 1)$ que codifica información previa sobre qué puntos es más probable que pertenezcan a la primitiva (por ejemplo, aquellos situados en una región de interés del objeto). La verosimilitud marginal queda entonces definida en (7).

$$p(\mathbf{x}_i \mid \boldsymbol{\theta}) = w_i \, \mathcal{N}(d_i(\boldsymbol{\theta}); 0, \sigma^2) + (1 - w_i) \, \frac{1}{V} \tag{7}$$

En la práctica, descomponemos $w_i = w \, \pi_i$, donde w es un factor global que se ajusta durante la optimización y $\pi_i \in [0, 1]$ es un prior espacial fijo asociado a cada punto.

2.2.2. Probabilidad posterior de pertenencia

Aplicando la regla de Bayes, la probabilidad posterior de que $\mathbf{x}_i$ sea un inlier viene dada por (8).

$$p(z_i = 1 \mid \mathbf{x}_i, \boldsymbol{\theta}) = \frac{w_i \, \mathcal{N}(d_i(\boldsymbol{\theta}); 0, \sigma^2)}{w_i \, \mathcal{N}(d_i(\boldsymbol{\theta}); 0, \sigma^2) + (1 - w_i) \, \frac{1}{V}} \tag{8}$$

Estas probabilidades $p_i = p(z_i = 1 \mid \mathbf{x}_i, \boldsymbol{\theta})$ ponderan la contribución de cada punto en la función de ajuste, reforzando los puntos con prior alto y descartando gradualmente outliers.

2.2.3. Función de distancia y actualización de la varianza

La distancia se define siguiendo la ecuación (9):

$$d_i(\boldsymbol{\theta}) = \left\| \mathbf{x}_i^{\text{local}} \right\| \left| F(\mathbf{x}_i^{\text{local}}, \boldsymbol{\theta}) - 1 \right| \tag{9}$$

donde $\mathbf{x}_i^{\text{local}}$ es el punto expresado en el sistema local de la primitiva y $F(\cdot)$ es la función implícita correspondiente. Esta distancia gobierna la verosimilitud de cada punto y determina su probabilidad posterior de pertenencia.

La varianza σ^2 se estima de forma adaptativa a partir de los residuos actuales y las probabilidades de pertenencia, tal y como se define en (10).

$$\sigma^2 \leftarrow \frac{2 \sum_i p_i \, d_i^2(\boldsymbol{\theta})}{3 \sum_i p_i} \tag{10}$$

2.3. Función objetivo y términos de pérdida

El ajuste de cada primitiva geométrica se formula como la minimización de una función de pérdida que combina la coherencia con los datos observados y varias restricciones físicas derivadas del entorno. El modelo sigue la formulación probabilística descrita por (Liu et al., 2022), donde cada punto posee una probabilidad de pertenencia p_i que pondera su contribución durante la optimización.

La pérdida total combina un término de ajuste y varios términos de regularización, tal y como se define en 11.

$$\begin{aligned}\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{fit}}(\theta) + \lambda_{\text{free}}\mathcal{L}_{\text{free}}(\theta) + \lambda_{\text{table}}\mathcal{L}_{\text{table}}(\theta) \\ + \lambda_{\text{drop}}\mathcal{L}_{\text{drop}}(\theta) + \lambda_{\text{extent}}\mathcal{L}_{\text{extent}}(\theta)\end{aligned} \tag{11}$$

Los superelipsoides y supertoroides utilizan únicamente $\mathcal{L}_{\text{fit}}$, $\mathcal{L}_{\text{free}}$ y $\mathcal{L}_{\text{table}}$; los superparaboloides, al ser formas abiertas, requieren además $\mathcal{L}_{\text{drop}}$ y $\mathcal{L}_{\text{extent}}$.

De forma resumida:

- Data fitting loss ($\mathcal{L}_{\text{fit}}$): asegura que la superficie estimada se ajusta a los puntos con alta probabilidad de pertenencia.
- Free-space loss ($\mathcal{L}_{\text{free}}$): evita que la primitiva invada el espacio libre entre la cámara y los puntos visibles.
- Table-contact loss ($\mathcal{L}_{\text{table}}$): impide que la primitiva atraviese el plano de apoyo estimado.
- Drop loss ($\mathcal{L}_{\text{drop}}$): en formas abiertas, penaliza que la base de la primitiva caiga por debajo de los puntos observados.
- Extent loss ($\mathcal{L}_{\text{extent}}$): restringe la altura del superparaboloide para evitar dilataciones no justificadas por los datos.

Para las expresiones completas de cada término, nos remitimos al trabajo original de (Menendez et al., 2025).

2.4. Procedimiento de optimización

El ajuste de primitivas se formula como una optimización local no convexa. Por tanto, el proceso requiere una inicialización adecuada de los parámetros $\boldsymbol{\theta}$, ya que distintas configuraciones iniciales pueden conducir a mínimos diferentes. A partir de esta inicialización, la optimización alterna entre actualizar la probabilidad de pertenencia de cada punto y refinar los parámetros de la primitiva. El proceso sigue un esquema EM suavizado:

- **Actualización de pertenencia (E-step).** A partir de los residuos actuales y del modelo probabilístico, se recalculan las probabilidades p_i de que cada punto sea un inlier. Estas probabilidades incorporan tanto la distancia geométrica como el prior espacial derivado de la región proyectada de la primitiva.
- **Actualización de parámetros (M-step).** Se realiza un paso de descenso por gradiente sobre la pérdida total, actualizando los parámetros de forma, orientación y escala. Periódicamente, la varianza σ^2 se recalibra usando los inliers estimados, estabilizando la convergencia.

Este proceso continúa hasta la convergencia o hasta alcanzar un número máximo de iteraciones.

2.5. Inicialización guiada por VLM

El proceso de optimización descrito anteriormente es de naturaleza local y, por tanto, requiere una inicialización adecuada de los parámetros $\boldsymbol{\theta}$ para converger hacia soluciones coherentes con la geometría real. Para este propósito utilizamos un modelo multimodal visión–lenguaje (VLM) que, a partir de una imagen RGB del objeto, identifica sus principales componentes geométricos y propone para cada uno de ellos una primitiva adecuada. El modelo opera a partir de un *prompt* específico que define su tarea, las entradas requeridas y el formato estructurado de salida que debe generar.

Antes de solicitar la descomposición geométrica, se proporciona al modelo una descripción explícita de los sistemas de referencia utilizados. El sistema de referencia de la cámara es un sistema 3D en el que *x_cam* apunta a la derecha, *y_cam* hacia abajo y *z_cam* hacia delante, con la cámara inclinada un ángulo predeterminado respecto al plano horizontal, lo que genera una vista ligeramente cenital del objeto. Asimismo, las coordenadas de imagen se expresan en un sistema 2D en el que *x_img* crece hacia la derecha y *y_img* hacia abajo.

A partir de la imagen RGB asociada al objeto, el modelo genera para cada parte una descripción estructural que incluye: (i) el tipo de primitiva a utilizar (superelipsoide, supertoroide o superparaboloide), (ii) un vector que define la orientación del eje principal en el sistema de referencia de la cámara (*axis_cam*), (iii) un recuadro delimitador aproximado en coordenadas normalizadas que cubre la región correspondiente a la parte (*bbox_uv*), (iv) una estimación del tamaño proyectado de la parte en la imagen (*axis_extent_px*), y (v) valores iniciales para los parámetros de forma, concretamente los exponentes ϵ_1, ϵ_2 y, en el caso de los superparaboloides, el radio inicial r_0. El eje principal proporcionado por el modelo siempre corresponde al eje geométrico propio de la primitiva: el eje de simetría en superelipsoides y superparaboloides, o el eje que atraviesa el orificio en supertoroides. En ningún caso representa el eje global del objeto completo, sino únicamente el de la parte descrita.

Estos elementos no actúan como una segmentación estricta, sino como una guía semántica inicial. En particular, el recuadro delimitador se utiliza para definir un prior espacial w_i sobre los puntos de la nube: los puntos cuya proyección cae dentro del *bbox* reciben una probabilidad inicial de pertenencia mayor, mientras que los demás comienzan con un peso inferior. Este prior se combina con el esquema probabilístico del ajuste durante las iteraciones de optimización, lo que permite que puntos inicialmente descartados puedan incorporarse si la forma lo justifica, y viceversa.

El centro 3D inicial de cada primitiva se calcula directamente a partir de la nube de puntos: se proyecta el *bbox* de la imagen, se seleccionan los puntos cuya proyección cae en esa región y se toma el centroide 3D de dicho subconjunto. Este procedimiento produce una inicialización geométricamente consistente y evita depender exclusivamente de estimaciones 2D que no contienen información de profundidad.

A partir del vector que define el eje principal en el sistema de referencia de la cámara, *axis_cam*, construimos la rotación inicial R_0 alineando el eje local x de la primitiva con dicho vector. Primero normalizamos *axis_cam* y lo tomamos como $\mathbf{x}_{\text{loc}}$, escogemos un vector auxiliar $\mathbf{t}$ lo menos alineado posible con él

y completamos una base ortonormal mediante productos vectoriales.

$$\mathbf{z}_{\text{loc}} = \frac{\mathbf{x}_{\text{loc}} \times \mathbf{t}}{\|\mathbf{x}_{\text{loc}} \times \mathbf{t}\|}, \mathbf{y}_{\text{loc}} \quad = \frac{\mathbf{z}_{\text{loc}} \times \mathbf{x}_{\text{loc}}}{\|\mathbf{z}_{\text{loc}} \times \mathbf{x}_{\text{loc}}\|} \tag{12}$$

y finalmente la rotación inicial queda definida según la ecuación 13.

$$R_0 = [\, \mathbf{x}_{\text{loc}} \;\; \mathbf{y}_{\text{loc}} \;\; \mathbf{z}_{\text{loc}}] \tag{13}$$

Si el determinante de R_0 es negativo, invertimos $\mathbf{y}_{\text{loc}}$ para obtener una base de mano derecha.

Las escalas iniciales se obtienen combinando la orientación propuesta con las extensiones en píxeles (*axis_extent_px*) y la distribución 3D de los puntos proyectados en la región correspondiente. De esta forma, los parámetros de tamaño a_1, a_2, a_3 se inicializan de manera coherente con la forma aparente sin imponer restricciones rígidas sobre la geometría real del objeto.

2.6. Ajuste secuencial de primitivas

Una vez obtenidas las propuestas de primitivas para las distintas partes del objeto, el ajuste se realiza de forma secuencial sobre la nube de puntos asociada al objeto completo. El objetivo es ajustar cada primitiva a la región sugerida y retirar progresivamente los puntos explicados, permitiendo el ajuste de las siguientes partes.

El procedimiento seguido es el siguiente:

1. **Inicialización por parte propuesta.** Se inicializan los parámetros $\boldsymbol{\theta}$ utilizando la orientación, la extensión aproximada y la región asociada a la primitiva. Los puntos cuya proyección cae dentro de dicha región reciben un peso inicial mayor en el modelo probabilístico. Esta asignación no constituye una segmentación estricta: cualquier punto del clúster puede acabar siendo un inlier si así lo determina la optimización.
2. **Optimización local mediante EM suavizado.** A partir de esta inicialización se aplica el procedimiento iterativo de ajuste descrito previamente. Este algoritmo estima de forma iterativa los parámetros $\boldsymbol{\theta}$ y las probabilidades de pertenencia de los puntos, incorporando las restricciones físicas y geométricas del entorno.
3. **Eliminación de puntos explicados.** Tras la convergencia del ajuste, los puntos con alta probabilidad de pertenencia a la primitiva se retiran del conjunto de puntos disponible. Los puntos restantes, aún no explicados por ninguna forma previa, se utilizan como entrada para el ajuste de la siguiente primitiva propuesta.

Este proceso continúa hasta agotar todas las primitivas sugeridas o hasta que no queden suficientes puntos para justificar nuevas formas. De este modo, la propuesta inicial proporcionada por el modelo multimodal se combina con un refinamiento geométrico local basado exclusivamente en los datos 3D, permitiendo descomponer objetos complejos en múltiples componentes ajustados de manera coherente.

3. Experimentos y resultados

El método propuesto se implementó en PyTorch y utiliza un modelo VLM de OpenAI para generar las inicializaciones geométricas. Tanto el método basado en PCA como el método propuesto comparten la misma configuración de optimización: descenso por gradiente con Adam (2000 iteraciones por ajuste de primitiva, $\eta = 10^{-3}$), actualización de la varianza cada dos iteraciones y pesos de regularización $\lambda_{\text{free}} = 120$ y $\lambda_{\text{table}} = 10$. En el caso de los superparaboloides, se añadieron $\lambda_{\text{drop}} = 20$ y $\lambda_{\text{extent}} = 10$. En el método basado en PCA, la subsegmentación posterior al ajuste se realizó con una distancia máxima de 10 mm y un mínimo de 30 puntos por subclúster.

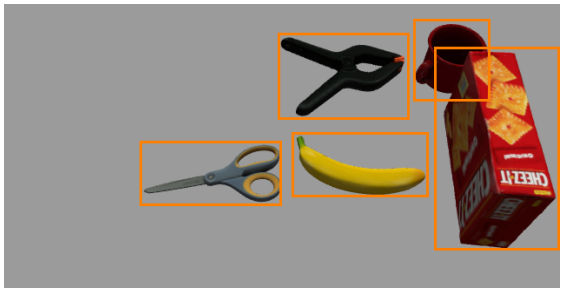
Con esta configuración común, comparamos cualitativamente el ajuste secuencial inicializado mediante PCA (Menendez et al., 2025) con el método guiado por VLM. El método basado en PCA no dispone de información previa sobre cuántas partes componen el objeto, su localización ni el tipo de primitiva más adecuado para cada región. En este caso, la primera primitiva se inicializa aplicando PCA al clúster completo y, tras su ajuste, los puntos con baja probabilidad de pertenencia se resegmentan por distancia. A continuación, se prueban sucesivamente ajustes con superelipsoides, supertoroides y superparaboloides, seleccionando la primitiva con menor pérdida. Este procedimiento puede funcionar cuando la primera primitiva queda bien orientada, pero en escenas parcialmente visibles o con geometrías compuestas suele generar inicializaciones frágiles, y sus errores se propagan a las fases posteriores del ajuste secuencial.

En contraste, el método propuesto recibe del VLM una descomposición en partes que incluye, para cada una de ellas, el tipo de primitiva, una orientación aproximada y una región de influencia asociada. Esta información proporciona una estructura inicial para el proceso de ajuste, evitando segmentaciones iterativas y pruebas sucesivas de distintas primitivas. Además, las regiones propuestas se incorporan como un prior espacial que guía la asignación probabilística de inliers desde la primera iteración, dando lugar a inicializaciones más estables en escenas parcialmente visibles.

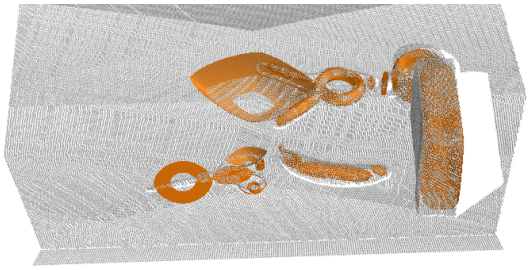
La Figura 2 muestra un ejemplo representativo de la diferencia entre ambos enfoques. En la Figura 2(a) se presentan los recortes de imagen RGB utilizados como entrada al VLM. En la inicialización basada en PCA (Figura 2(b)), la descomposición resultante es menos coherente y aparecen primitivas con orientaciones y extensiones menos ajustadas a la estructura observada. En cambio, la inicialización guiada por el modelo visión–lenguaje (Figura 2(c)) produce una descomposición más acorde con las distintas partes visibles del objeto, dando lugar a un ajuste global más consistente en la escena mostrada.

4. Conclusiones

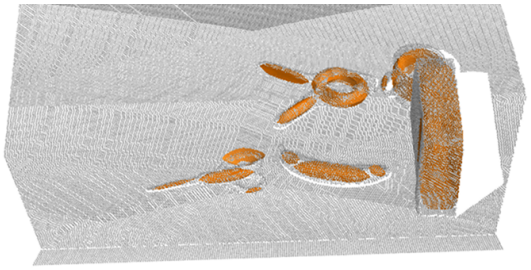
En este trabajo se ha presentado un ajuste probabilístico de primitivas geométricas sobre nubes de puntos parciales que combina un modelo Gaussiana–Uniforme con regularizaciones físicas y un ajuste secuencial de múltiples primitivas por objeto. La inicialización guiada por un modelo multimodal visión–lenguaje, obtenida a partir de una única imagen RGB, proporciona una descomposición en partes y parámetros iniciales más



(a) Recortes de imagen RGB



(b) Inicialización secuencial basada en PCA.



(c) Inicialización geométrica guiada por VLM (propuesta).

Figura 2: Comparación cualitativa del ajuste secuencial en una escena de sobremesa parcialmente visible. (a) Recortes de imagen RGB utilizados como entrada al VLM para proponer las primitivas iniciales. La nube de puntos observada se muestra en gris y las primitivas ajustadas en naranja. En la inicialización basada en PCA (b), la descomposición resultante es menos coherente y aparecen primitivas peor ajustadas a la estructura observada. En cambio, la inicialización geométrica guiada por VLM (c) produce una descomposición más coherente de las partes observadas.

coherentes que las aproximaciones puramente geométricas basadas en PCA, mejorando la fidelidad de la reconstrucción.

Como trabajo futuro, se plantea una evaluación más exhaustiva en múltiples escenas y categorías de objetos, analizando cuantitativamente el impacto de distintos *prompts* y modelos multimodales en la calidad de la inicialización. Además, resulta prometedor emplear el VLM para proponer regiones funcionales de agarre y restringir el cálculo de configuraciones a las primitivas presentes en dichas zonas, acercando el método a un uso directo en tareas de manipulación.

Agradecimientos

La investigación que ha conducido a estos resultados ha recibido financiación del proyecto iRoboCity2030-CM, Robótica inteligente para ciudades sostenibles (TEC-2024/TEC-62), financiado por Programas de Actividades I+D en tecnologías de la Comunidad de Madrid.

Referencias

Barr, A. H., 1981. Superquadrics and angle-preserving transformations. IEEE Computer Graphics and Applications 1 (1), 11–23.
DOI: 10.1109/MCG.1981.1673799

Biederman, I., 1987. Recognition-by-components: a theory of human image understanding. Psychological review 94 (2), 115.
DOI: 10.1037/0033-295X.94.2.115

Ferreira, P., 2018. Sampling superquadric point clouds with normals. arXiv preprint arXiv:1802.05176.
DOI: 10.48550/arXiv.1802.05176

Gross, A. D., Boult, T. E., 1988. Error of fit measures for recovering parametric solids. In: 1988 Second International Conference on Computer Vision. IEEE Computer Society, pp. 690–691.
DOI: 10.1109/CCV.1988.590052

Jiang, H., Liu, S., Wang, J., Wang, X., 2021. Hand-object contact consistency reasoning for human grasps generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11107–11116.
DOI: 10.1109/ICCV48922.2021.01092

Liu, W., Wu, Y., Ruan, S., Chirikjian, G. S., 2022. Robust and accurate superquadric recovery: A probabilistic approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2676–2685.
DOI: 10.1109/CVPR52688.2022.00270

Lu, S., Wang, R., Miao, Y., Mitash, C., Bekris, K., 2022. Online object model reconstruction and reuse for lifelong improvement of robot manipulation. In: 2022 International Conference on Robotics and Automation (ICRA). IEEE, pp. 1540–1546.
DOI: 10.1109/ICRA46639.2022.9812440

Makhal, A., Thomas, F., Gracia, A. P., 2018. Grasping unknown objects in clutter by superquadric representation. In: 2018 Second IEEE international conference on robotic computing (IRC). IEEE, pp. 292–299.
DOI: 10.1109/IRC.2018.00062

Menendez, E., Martínez, S., Balaguer, C., 2024a. Selección y agarre robótico de objetos basada en el seguimiento de la mirada. In: Actas del Simposio de Robótica, Bioingeniería y Visión por Computador: Badajoz, 29 a 31 de mayo de 2024. Servicio de Publicaciones, pp. 127–132.

Menendez, E., Martínez, S., Balaguer, C., 2025. Ajuste probabilístico de múltiples primitivas cóncavas y convexas a nubes de puntos parciales. Simposios del Comité Español de Automática 1 (1), 8.
DOI: 10.64117/simposioscea.v1i1.58

Menendez, E., Martínez, S., Díaz-de María, F., Balaguer, C., 2024b. Integrating egocentric and robotic vision for object identification using siamese networks and superquadric estimations in partial occlusion scenarios. Biomimetics 9 (2), 100.
DOI: 10.3390/biomimetics9020100

Paschalidou, D., Ulusoy, A. O., Geiger, A., 2019. Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10344–10353.
DOI: 10.1109/CVPR.2019.01059

Solina, F., Bajcsy, R., 2002. Recovery of parametric models from range images: The case for superquadrics with global deformations. IEEE transactions on pattern analysis and machine intelligence 12 (2), 131–147.
DOI: 10.1109/34.44401

Tulsiani, S., Su, H., Guibas, L. J., Efros, A. A., Malik, J., 2017. Learning shape abstractions by assembling volumetric primitives. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2635–2643.
DOI: 10.1109/CVPR.2017.160