Mallas de cuerpo como puntos

Jianfeng Zhang1 Dongdong Yu2 Jun Hao Liew1 Xuecheng Nie3 Jiashi Feng1

1Universidad Nacional de Singapur 2ByteDance AI Lab 3Tecnología Yitu

{zhangjianfeng,liewjunhao}@u.nus.edu yudongdong@bytedance.com elefjia@nus.edu.sg

Resumen

Consideramos la desafiante malla corporal 3D para varias personas tarea de estimación en este trabajo. Los métodos existentes son en su mayoría basado en dos etapas: una etapa para la localización de personas y la otra etapa para la estimación de malla de cuerpo individual, lo que lleva a tuberías redundantes con alto costo de cálculo y rendimiento degradado para escenas complejas (por ejemplo, ocluido instancias de persona). En este trabajo, presentamos un modelo de una sola etapa, Body Meshes as Points (BMP), para simplificar la canalización y elevar tanto la eficiencia como el rendimiento. En particular, BMP adopta un nuevo método que representa múltiples instancias de persona como puntos en el espacio de profundidad espacial donde cada punto está asociado con una malla de cuerpo. Aferrado a tales representaciones, BMP puede predecir directamente las mallas del cuerpo para múltiples personas en una sola etapa localizando simultáneamente puntos de instancia de persona y estimando las mallas corporales correspondientes. Para razonar mejor sobre el orden de profundidad de todas las personas dentro de una misma escena, BMP diseña un pérdida de profundidad ordinal entre instancias simple pero efectiva para obtener una estimación de malla de cuerpo coherente con la profundidad. BMP también presenta un novedoso aumento con reconocimiento de puntos clave para mejorar Robustez del modelo para instancias de personas ocluidas. Experimentos integrales en los puntos de referencia Panoptic, MuPoTS 3D y 3DPW demuestran claramente la eficiencia de vanguardia de BMP para la estimación de la malla corporal de varias personas, en conjunto con una precisión sobresaliente. El código se puede encontrar en: https://github.com/jfzhang95/BMP.

1. Introducción

La recuperación de la malla del cuerpo humano en 3D tiene como objetivo reconstruir el Malla 3D de cuerpo completo de la instancia de persona a partir de imágenes o vídeos. Como tarea fundamental pero desafiante, ha ha sido ampliamente aplicado para reconocimiento de acción [63], prueba virtual [41], retargeting de movimiento [35] , etc.], un escenario más realista y desafiante ha atraído una atención creciente, es decir, para estimar el cuerpo mallas para múltiples personas a partir de una sola imagen.

Los métodos existentes para la recuperación de malla de varias personas son principalmente soluciones de dos etapas, incluidas las de arriba hacia abajo [20] y enfoques ascendentes [69] . El primero primero localiza las instancias de personas a través de un detector de personas, en base al cual luego recupera las mallas 3D individualmente; el enfoque de abajo hacia arriba estima los puntos clave de la persona al principio, y luego conjuntamente reconstruye múltiples cuerpos humanos en 3D en la imagen a través de optimización restringida [69]. Aunque con una precisión notable, los paradigmas anteriores son ineficientes con redundancia computacional. Por ejemplo, el primero estima malla corporal para cada persona por separado, y en consecuencia la El costo total de computación crece linealmente con el número de personas en la imagen, mientras que la segunda requiere agrupar las puntos clave en personas correspondientes e inferir el cuerpo mallas iterativamente, lo que lleva a un alto costo computacional.

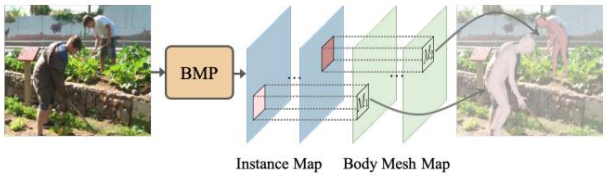

Figura 1. **Nuestra solución de una etapa.** El modelo propuesto representa cada instancia de persona como el punto central de su cuerpo. Instancia la localización y la recuperación de la malla corporal se predicen directamente desde las características del punto central, lo que permite la reconstrucción simultánea de varias personas en una sola etapa. Se ve mejor en color.

Dirigidos a una tubería más eficiente y compacta, considere explorar una solución de una sola etapa. A pesar de la reciente popularidad y el desempeño prometedor de los monoetapas métodos en la estimación de puntos clave 2D [43] y tareas de detección de objetos [75, 59], una canalización de una sola etapa para varias personas la recuperación de malla apenas se explora, ya que no está claro cómo para integrar de manera efectiva tanto la localización de personas como la malla pasos de recuperación dentro de una sola etapa. En este trabajo, proponemos una nueva representación de instancia para el cuerpo de varias personas. recuperación de malla que representa instancias de múltiples personas como puntos en el espacio de profundidad espacial donde cada punto está asociado con una malla de cuerpo. Tal representación permite paralelismo efectivo de localización de personas y malla corporal recuperación. En base a él, desarrollamos una nueva arquitectura de modelo que aprovecha las características compartibles tanto para la localización y recuperación de mallas y así lograr una solución en una sola etapa.

En particular, el modelo tiene dos ramas paralelas, una

por ejemplo la localización y la otra para la recuperación de la malla corporal. En la rama de localización, modelamos cada persona en stance como un punto único en un espacio tridimensional, es decir, espacial. (2D) y profundidad (1D), donde cada punto localizado (detectado persona) está asociado con una malla corporal en la malla corporal rama representada por el modelo paramétrico SMPL [36]. Esto a su vez convierte la recuperación de malla de varias personas en un problema de regresión de tiro único (Fig. 1). Específicamente, el la ubicación espacial está representada por coordenadas discretas wrt rejillas regulares sobre la imagen de entrada. Del mismo modo, discretizamos profundidad en varios niveles para obtener la representación de la profundidad. Para aprender una mejor representación de características para diferenciar instancias a diferente profundidad, motivadas por el fenómeno que una persona más cerca de la cámara tiende a parecer más grande en la imagen, adoptamos la red piramidal de características (FPN) [30] para extraer características multiescala y usar características de la parte inferior escalas para representar las instancias más cercanas (y más grandes). En esto manera, cada instancia se representa como un punto, cuyas características asociadas (extraídas de su correspondiente espacial ubicación y escala FPN) se utilizan para estimar efectivamente su malla corporal. A este cuerpo lo llamamos mallas como puntos **(BMP).**

Aplicación del modelo BMP para la estimación de varias personas body mesh enfrenta simultáneamente dos desafíos en escenarios realistas: cómo reconstruir coherentemente instancias con ordenar la profundidad correcta y cómo manejar el problema común de oclusión (p. ej., instancias superpuestas y observaciones parciales). Para el primer desafío, consideramos usar explícitamente las relaciones ordinales entre todas las personas en la escena. para supervisar el modelo para aprender a generar mallas corporales con orden de profundidad correcto. Sin embargo, obtener tales relaciones ordinales no es trivial para las escenas capturadas en la naturaleza, ya que no hay ninguna anotación 3D disponible. Inspirado en el reciente éxito de la estimación de profundidad para las articulaciones del cuerpo humano [42, 73], proponemos tomar la profundidad de cada persona (punto central) predicho por un modelo entrenado previamente en conjuntos de datos 3D con profundidad anotaciones como la relación pseudoordinal para el entrenamiento del modelo en los datos en estado salvaje, lo cual se prueba experimentalmente beneficioso para la reconstrucción con malla corporal coherente en profundidad.

Además, para abordar el problema de la oclusión común, proponemos una novedosa estrategia de aumento de la oclusión consciente de los puntos clave para mejorar la robustez del modelo para instancias de personas ocluidas. A diferencia del método anterior [55] que simula aleatoriamente la oclusión en imágenes, generamos una oclusión sintética basado en la posición de los puntos clave del esqueleto. Tal oclusión consciente del punto clave fuerza explícitamente al modelo a enfocarse en estructura corporal, haciéndola más robusta a la oclusión.

Experimentos completos sobre puntos de referencia de poses en 3D Panoptic [23], MuPoTS-3D [38] y 3DPW [64] evidentemente demostrar la alta eficiencia del modelo propuesto. Además, logra un nuevo estado del arte en Panoptic y Conjuntos de datos MuPoTS-3D y rendimiento competitivo en Conjunto de datos 3DPW. Nuestras contribuciones se resumen como sigue:

1) Hasta donde sabemos, estamos entre los primeros para explorar la solución de una sola etapa para la malla de varias personas recuperación. Presentamos una nueva representación de instancia de persona que permite la localización simultánea de personas y el cuerpo. recuperación de malla para todas las instancias de personas en una imagen dentro de un una sola etapa, y diseñar una arquitectura modelo novedosa en consecuencia. 2) Proponemos una inter-instancia simple pero efectiva supervisión de relación ordinal para fomentar la reconstrucción coherente en profundidad. 3) Proponemos una oclusión consciente del punto clave estrategia de aumento que tiene en cuenta la estructura corporal para mejorar la solidez del modelo a la oclusión.

2. Trabajo relacionado

Pose y forma 3D de una sola persona Estimación de trabajos anteriores Poses 3D en forma de esqueleto corporal [37, 40, 60, 74, 49, 47, 58, 71, 15] o forma 3D no paramétrica [13, 56, 62]. En este trabajo, usamos la malla 3D para representar el cuerpo completo pose y forma, y adopte el modelo paramétrico SMPL [36] para la recuperación de la malla corporal. En la literatura, Bogo et al. [5] propuso SMPLify, el primer método basado en la optimización para adaptarse SMPL en las uniones 2D detectadas iterativamente. Trabajos posteriores amplían SMPLify ya sea usando puntos de referencia más densos para reemplazar puntos clave dispersos como siluetas y cuadrículas de ocupación de vóxeles para ajuste SMPL [28, 62], o ajuste más modelo expresivo (por ejemplo, SMPL-X) que SMPL [46].

Algunos trabajos recientes hacen retroceder directamente los parámetros SMPL de las imágenes a través de redes neuronales profundas en dos etapas. manera. Primero estiman la representación intermedia (por ejemplo, puntos clave, siluetas, etc.) a partir de imágenes y luego mapear a los parámetros SMPL [48, 44, 61, 27]. Algunos otros estiman directamente los parámetros SMPL a partir de imágenes, ya sea utilizando estrategias de entrenamiento de modelos complejos [24, 16] o aprovechando la información temporal [3, 25]. Aunque alta precisión se logra en casos de una sola persona, no está claro cómo extenderlos a los casos más generales de pluripersonales.

Postura y forma 3D de varias personas Para 3D de varias personas estimación de pose, la mayoría de los métodos existentes adoptan un enfoque de arriba hacia abajo paradigma [53, 9, 54]. Primero detectan cada instancia de persona y luego hacer una regresión de las ubicaciones de las articulaciones del cuerpo. Hacer un seguimiento las mejoras se realizan mediante la estimación de valores absolutos adicionales profundidad [42], considerando la interacción de varias personas [17, 29] o extendiéndose a la estimación de la pose de todo el cuerpo [66]. Alternativamente, algunos enfoques también exploran el enfoque de abajo hacia arriba. paradigma. SSMP3D [39] y SMAP [73] estiman 3D poses a partir de mapas de poses conscientes de la oclusión y utilice Part Affin ity Fields [7] para inferir su asociación. LoCO [12] mapas la imagen a los mapas de calor volumétricos y luego estima poses 3D de varias personas de ellos por un codificador-decodificador estructura. PandaNet [4] es un modelo basado en anclas donde Las poses 3D se retroceden para cada posición de anclaje.

En contraste con la prosperidad de la estimación de poses 3D de varias personas, hay un número limitado de obras denotadas para el cuerpo. recuperación de malla para varias personas. Zanfir et al. [69] primero estime las uniones 3D de las personas en la imagen y luego optimice

sus formas 3D conjuntamente con múltiples restricciones. Ellos también proponer un esquema basado en una regresión de dos etapas que primero estima las articulaciones 3D para todas las personas y luego hace una regresión de sus Formas 3D basadas en estas uniones 3D [70]. En lugar de hacer una regresión de los parámetros SMPL desde una representación intermedia (p. ej., juntas 3D), Jiang et al. [20] adjunte un cabezal SMPL al Marco R-CNN más rápido [51] para estimar parámetros SMPL directamente desde la imagen de entrada de manera descendente.

A pesar de los resultados alentadores, estos métodos se basan en el marco indirecto de múltiples etapas y sufren de baja eficiencia. A diferencia de todos los métodos anteriores que se basan en una tubería de múltiples etapas con redundancia de cómputo, nuestro El método unifica la localización de personas y la malla corporal, y permite una etapa única sin optimización (ad hoc) y sin cajas.

solución para la recuperación de malla corporal de varias personas.

Representación basada en puntos Los métodos basados en puntos [10, 75, 59] representan instancias por un solo punto en su centro. Este enfoque se considera como un simple reemplazo del representación basada en anclas, que ha sido ampliamente utilizada en muchas tareas, incluida la detección de objetos [10, 75, 59], 2D estimación de puntos clave [43] y segmentación de instancias [65]. Sin embargo, estos métodos no se pueden aplicar directamente al cuerpo. recuperación de mallas. En este trabajo, extendemos la representación basada en puntos a la recuperación de mallas corporales de varias personas. Un trabajo concurrente [68] adopta una solución similar a la recuperación de malla corporal. Nuestro modelo se diferencia de él en dos aspectos significativos:

1) BMP apunta a una reconstrucción más coherente de las personas en las escenas. Maneja arreglos espaciales desafiantes y problemas de oclusión mediante la explotación de la pérdida de profundidad ordinal y la estrategia de aumento consciente de puntos clave, que no son considerado en [68]. 2) BMP adopta un novedoso 3D basado en puntos representación para diferenciar instancias a diferentes profundidades, por lo tanto, es más robusto para instancias superpuestas; mientras que [68] usa solo representación 2D y fallaría en tales casos.

3. Mallas de cuerpo como puntos

3.1. Solución propuesta de una sola etapa

Dada una imagen I, la recuperación de mallas corporales de varias personas se dirige a la recuperación de mallas corporales de todas las instancias de personas en I. Los enfoques existentes [70, 69, 20] resuelven esta tarea a través de localizar secuencialmente y estimar la malla del cuerpo en una manera de múltiples etapas, lo que lleva a la redundancia de cálculo. De manera diferente, este trabajo tiene como objetivo unificar la localización de instancias y la recuperación de malla corporal en una solución de una sola etapa para permitir un marco más eficiente y conciso.

Representamos cada instancia de persona como un único punto **(i, j,** k) en un espacio tridimensional (atravesado por espacio 2D y dimensiones de profundidad 1D). Al dividir la imagen de entrada de manera uniforme en cuadrículas **G × G ,** su dimensión espacial puede ser fácilmente representada dentro de dicha cuadrícula de coordenadas. Si el centro del cuerpo de una persona cae en la celda de la cuadrícula **(i,** j), se le asigna una coordenada espacial **(i,** j). De manera similar, para la dimensión de profundidad, discretizar el valor de profundidad a **K** niveles y obtener el valor **k** para cada instancia según su profundidad. tan discretizado El valor de profundidad es beneficioso para manejar instancias de oclusión, especialmente cuando los centros del cuerpo de instancias múltiples caen en la misma coordenada de cuadrícula espacial.

Dada esta representación, reformulamos pluripersonal recuperación de malla como dos tareas de predicción simultáneas: 1) localización de instancia y 2) recuperación de malla corporal.

Localización de instancias Para la primera tarea, empleamos el mapa de instancias **C** = **{C1, . . . , CK},** donde **Ck ÿ** R **G×G×1** , a ubique cada instancia de persona en la imagen, donde **G** denota el número de celdas de cuadrícula a lo largo de un lado, mientras que **K** se refiere a el número de niveles de profundidad total. Para cada nivel de profundidad, el La red está entrenada para retroceder un escalar que indica la probabilidad de que cada celda de la cuadrícula contenga una persona.

Para construir la realidad del terreno (GT) para el entrenamiento, primero determinamos el valor de profundidad **k** para cada instancia. observamos que una persona tiende a parecer más grande (más pequeña) en la imagen cuando de pie más cerca (lejos) de la cámara. En otras palabras, la profundidad de una instancia es aproximadamente inversamente proporcional a su escala. Inspirándonos en él, empleamos una pirámide de características Red (FPN) [30] con niveles de pirámide **K** para capturar **K** diferentes escalas, cada una de las cuales se utiliza para representar la instancia con la profundidad correspondiente. Más específicamente, para cada instancia, calculamos su escala **s** = **ÿ hw** donde **(h,** w) denota el tamaño del cuerpo del GT, y lo asocia al nivel k de la pirámide correspondiente, de acuerdo con la Tabla 1.

Pirámide	**P2**	**P3**	**P4**	**P5**	**P6**
zancada 8		8	16	32	32
Número de cuadrícula **G** 40		36	24	16	12
Escala de instancia **s** <64 32ÿ128 64ÿ256 **128ÿ512** ÿ256					

Tabla 1. Empleamos FPN con cinco niveles de pirámide. **Pk+1** se utiliza para predecir la instancia Ck y los mapas de malla corporal Pk, donde **k** = **1, . . . ,** 5.

A continuación, ubicamos la celda de la cuadrícula **(i,** j) en **Ck** donde se encuentra la región central de esa persona. Inspirándose en [75, 10], la región central se define de la siguiente manera: dado el centro del cuerpo GT **(x c , yc),** el tamaño del cuerpo **(h,** w) de cada persona y un controlable factor de escala ÿ, la posición y el tamaño de la región central son definido como **(x c , yc , ÿw,** ÿh). En este trabajo establecemos la posición de la pelvis como centro del cuerpo y **ÿ** = 0,2. Una vez identificado, el celda de cuadrícula **(i,** j) del k-ésimo nivel de la pirámide, es decir, Ck (i, j) se etiqueta como positiva (etiqueta 1). Los pasos anteriores se repiten para todas las instancias en la imagen.

Representación de malla de cuerpo Paralelamente a la localización de instancias, usamos el mapa de malla de cuerpo **P** = **{P1, . . . , PK},** para la recuperación de la malla del cuerpo, donde **Pk ÿ** R**G×G×S** y **S** es el dimensión de la representación de la malla del cuerpo. Concretamente, dado una respuesta positiva en **C** que indica la presencia de una persona, hacemos una regresión de la representación de la malla del cuerpo utilizando las características de la celda de cuadrícula correspondiente, como se muestra en la Fig. 2.

En este trabajo, utilizamos el modelo paramétrico SMPL [36] para representación de malla de cuerpo, que representa una malla de cuerpo usando los parámetros de pose ÿ **ÿ** R **72** y parámetros de forma ÿ **ÿ** R **10.** Para mejorar la estabilidad del entrenamiento, adoptamos el

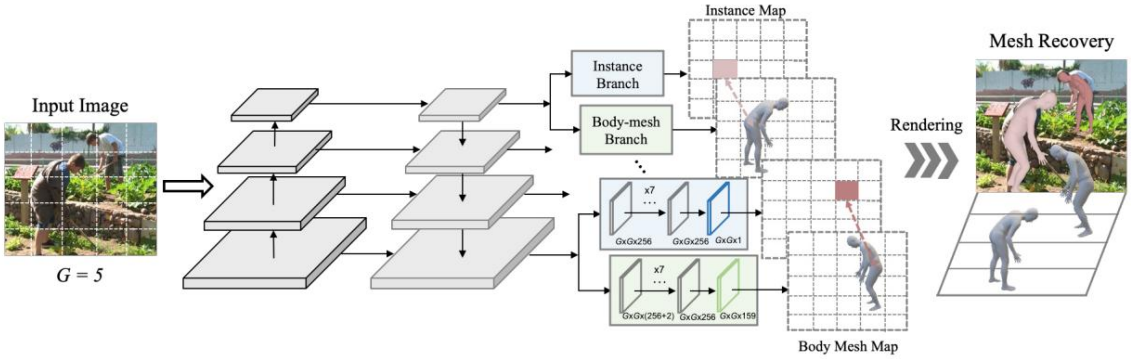

Figura 2. **Ilustración de nuestro marco BMP.** Una imagen de entrada se divide uniformemente en cuadrículas **G** × **G** con **G** = 5 en este ejemplo. El modelo adopta un FPN con **K** niveles **(K** = 4 aquí). Cada instancia de persona está representada por su celda de cuadrícula residente y su nivel FPN asociado (según su profundidad). BMP usa las características de la celda de cuadrícula y el nivel FPN para localizar a la persona contenida (arriba) y estimar la malla del cuerpo (abajo) simultáneamente.

Representación de rotación 6D [76] para los parámetros de pose con ÿ **ÿ** R **144.** El mapa de malla del cuerpo también predice un parámetro de cámara ÿ = **{s, tx, ty}** **ÿ** R para proyectar las articulaciones 3D del cuerpo **3** desde el espacio 3D hasta la al 2D, lo que permite entrenamiento en conjuntos de datos de pose 2D en estado salvaje [21, 32, 2] para mejorar la generalización del modelo [24]. Además, presentamos una puntuación de confianza escalar **c** definida como el OKS [14] entre los puntos clave proyectados y GT 2D, para reflejar el nivel de confianza de la predicción SMPL; y también proponemos una variable de profundidad absoluta **d** para la instancia de persona correspondiente que se usará para penalizar las estimaciones de malla corporal con profundidad o dering incoherente (ver Sec. 3.2 para más detalles). Por lo tanto, el número total de canales del mapa de malla corporal **S** es 159.

Arquitectura de red Empleamos ResNet-50 [18] como nuestra columna vertebral. FPN se construye sobre la columna vertebral para extraer una pirámide de mapas de características (256-d). Para realizar la recuperación de la malla del cuerpo, adjuntamos dos encabezados específicos de la tarea a cada nivel de la pirámide de funciones, uno para la localización de instancias y el otro para la recuperación de la malla del cuerpo correspondiente, responsabilidad ble para obtener el mapa de instancia **Ck** ÿ R mapa **G×G×1** y el de malla de cuerpo **Pk** ÿ R **G×G×159,** respectivamente. Como se muestra en la Fig. 2, cada cabeza consta de 7 circunvoluciones **de 3×3** apiladas y una capa de predicción específica de la tarea. Sin embargo, estimar directamente el parámetro de la cámara a partir de la imagen completa no es trivial, ya que es sensible a la posición de la instancia. Inspirándonos en CoordConv [34], concatenamos las coordenadas de píxeles normalizados al mapa de características de entrada al comienzo del cabezal de recuperación de malla para codificar la información de posición en la red para estimar mejor el parámetro de la cámara. Además, la normalización de grupos [67] se utiliza en ambos cabezales de predicción para facilitar el entrenamiento del modelo. Para hacer coincidir las características del tamaño **H × W** con **G × G,** aplicamos una interpolación bilineal antes de la instancia y la rama de recuperación de la malla del cuerpo.

3.2. Supervisión de profundidad ordinal entre instancias

La recuperación de la malla corporal de varias personas está inherentemente mal planteada, ya que múltiples predicciones 3D pueden corresponder al mismo 2D proyección. Por lo tanto, el modelo entrenado produciría estimaciones de malla corporal ambiguas con un orden de profundidad incorrecto debido a la falta de priorizaciones. Para aliviar este problema, usamos relaciones de profundidad ordinal entre todas las personas en la entrada como supervisión para guiar el razonamiento sobre el orden de profundidad durante el proceso de entrenamiento.

Más concretamente, dadas dos personas cualesquiera **(pm, pn)** en la imagen, definimos la relación de profundidad ordinal entre ellas como **R(pm,** pn), tomando el valor:

$$\mathbf{R(pm, pn)} = \begin{cases} \text{ÿ } +1, \text{ si } \mathbf{dn} \text{ ÿ } \mathbf{dm} > \mathbf{T}, \\ \text{ÿ1, si } \mathbf{dm} \text{ ÿ } \mathbf{dn} > \mathbf{T}, \text{ si } |\mathbf{dm} \text{ ÿ } \mathbf{dn}| \text{ ÿ } \mathbf{T}, \\ \mathbf{0}, \end{cases} \quad (1)$$

donde **dm ÿ** R **1** denota la profundidad de la persona **pm** y **T** es un umbral predefinido para determinar la relación ordinal. La relación ordinal **R(pm, pn)** = 0 significa que ambas instancias se encuentran aproximadamente a la misma profundidad; de lo contrario, uno de ellos está más cerca de la cámara que el otro. Con la relación ordinal de **(pm,** pn), definimos la pérdida de profundidad ordinal para este par como

$$\mathbf{L (pm, pn)} = \begin{cases} \text{ÿ } \log(1 + \exp \mathbf{(zm} \text{ ÿ } \mathbf{zn)}), \text{ si } \mathbf{R} = +1, \\ \log(1 + \exp \mathbf{(zn} \text{ ÿ } \mathbf{zm)}), \text{ si } \mathbf{R} = \text{ÿ}1, \text{ ÿÿÿ } \mathbf{(zm} \text{ ÿ } \mathbf{zn)} \\ \text{si } \mathbf{R} = \mathbf{0}, \end{cases}^{2}, \quad (2)$$

donde **zm** = denota la $\frac{2f}{szy}$ persona calculada a partir del parámetro de la cámara predicho con distancia focal f, escala **sm** y ancho del borde largo de las imágenes ÿ. La pérdida de profundidad ordinal impone un gran margen entre **zm** y **zn** si **R(pm, pn)** 6= 0, es decir, uno de ellos se mide más cerca que el otro y, de lo contrario, obliga a que sean iguales.

Sin embargo, en la práctica, estas relaciones de profundidad ordinal rara vez están disponibles para las escenas capturadas en la naturaleza debido a la falta de anotaciones en 3D. Para resolver este problema, proponemos utilizar relaciones pseudoordinales para el entrenamiento de modelos en los datos salvajes. Específicamente, primero entrenamos el modelo en conjuntos de datos 3D [19, 38] con anotaciones de profundidad para aprender a estimar

la profundidad de cada persona en las imágenes. Definimos la profundidad **d** de cada persona como la profundidad del centro del cuerpo (es decir, articulación de la pelvis). El modelo se entrena minimizando una pérdida de profundidad **Lprofundidad,** que se define como los errores cuadráticos medios (MSE) entre las profundidades previstas y GT. Después de eso, dados los datos sin etiquetar, primero aprovechamos el modelo previamente entrenado para estimar la profundidad que luego se usa para obtener las relaciones pseudoordinales para todas las personas en la imagen. Finalmente, dadas las relaciones pseudoordinales, adoptamos una pérdida de profundidad ordinal ponderada por puntaje OKS para supervisar el entrenamiento del modelo para imágenes en la naturaleza. La pérdida total de la imagen **I** se calcula como la pérdida promedio de todos los pares de instancias:

$$\mathbf{Lrank} = \frac{1}{\mathbf{N}} \sum_{\mathbf{cm,cn}} \mathbf{L\,(pm,pn)} \cdot \mathbf{N\,(pm,pn)} \quad (3)$$

donde **N** indica el número de instancias emparejadas en la imagen, **cm** indica la puntuación OKS de la m-ésima persona. Intuitivamente, entrenar el modelo con tal supervisión de profundidad ordinal entre instancias puede ayudar al modelo a construir una comprensión global del diseño de profundidad en la escena de entrada y así asegurar reconstrucciones más coherentes.

3.3. Aumento de oclusión con reconocimiento de punto clave

La recuperación de malla corporal basada en SMPL es muy sensible a la oclusión (parcial) (p. ej., superposición de personas, truncamiento) [72, 52]. Para mejorar la solidez del modelo a la oclusión sin requerir anotaciones y datos de entrenamiento adicionales, proponemos una estrategia de aumento de la oclusión consciente de los puntos clave durante el proceso de entrenamiento. La estrategia de aumento propuesta tiene como objetivo generar una oclusión sintética para sintetizar casos reales desafiantes, como la observación parcial para el entrenamiento del modelo. En comparación con trabajos anteriores [55] que simulan aleatoriamente la oclusión en las imágenes, lo que puede producir muestras de entrenamiento fáciles que son menos útiles para mejorar el rendimiento del modelo, nuestro método genera directamente una oclusión sintética basada en las posiciones de los puntos clave del esqueleto, lo que puede forzar el modelo. a prestar más atención a la estructura del cuerpo, lo que lleva a una mejora notable. Más concretamente, dado un conjunto de **J** puntos clave **{j1 , . . . , jJ }** de una persona en la imagen, primero elegimos al azar un punto clave **ji .** Luego tomamos una muestra aleatoria de un objeto no humano del conjunto de datos PASCAL VOC [11] y lo compartimos en la ubicación del punto clave seleccionado **ji .** Redimensionamos aleatoriamente el objeto muestreado al rango de **[0.1 ÿ A, 0.2 ÿ** A] antes de la composición, donde **A** = **wh** denota el área de esa persona. Además, cambiamos aleatoriamente la posición del punto clave por un desplazamiento ÿ para evitar un ajuste excesivo. Durante el entrenamiento, establecemos la probabilidad del aumento de la oclusión en 0,5.

3.4. Entrenamiento e inferencia

Entrenamiento Para entrenar nuestro modelo BMP propuesto, definimos la función de pérdida **L** de la siguiente manera:

$$\mathbf{L} = \mathbf{Linst} + \mathbf{Lmalla} + \mathbf{Lprofundidad} + \mathbf{0{,}1}\ \ddot{\mathbf{y}}\ \mathbf{Lrank,} \quad (4)$$

donde **Linst** es una pérdida focal modificada de dos clases [31] para la localización de instancias; **Lprofundidad** es la pérdida de profundidad (Sec. 3.2); **Lmesh** es la pérdida para la estimación de la malla del cuerpo. Los detalles de entrenamiento de la rama cuerpo-malla son similares a los de HMR [24]. Específicamente, formulamos **Lmesh** como

$$\mathbf{Lmalla} = \mathbf{Lpose} + \mathbf{Lvert} + \ddot{\mathbf{y}}\mathbf{3DL3D} + \ddot{\mathbf{y}}\mathbf{2DL2D} + \ddot{\mathbf{y}}\mathbf{formaLforma} + \ddot{\mathbf{y}}\mathbf{confLconf} + \ddot{\mathbf{y}}\mathbf{advLadv.} \quad (5)$$

Aquí , **Lpose, Lshape, L3D, Lvert** denotan MSE entre los parámetros de forma y pose predichos y GT, así como puntos clave y vértices 3D, respectivamente. **L2D** es la pérdida de puntos clave 2D que minimiza la distancia entre la proyección 2D desde los puntos clave 3D y los puntos clave GT 2D. **Lconf** es el MSE de las confianzas prevista y GT, donde la confianza GT se calcula como el OKS [14] entre los puntos clave proyectados y GT 2D. Además, usamos un discriminador y aplicamos una pérdida de adversario **Ladv** en los parámetros de forma y pose regresivos, para alentar las salidas a estar en las distribuciones de cuerpos humanos reales. **ÿ3D** = 4, **ÿ2D** = 4, **ÿshape** = 0,01, **ÿconf** = 1 y **ÿadv** = 0,01 son los pesos de los términos de pérdida correspondientes. La pérdida **Lmesh** se aplica de forma independiente a cada celda de cuadrícula positiva. La pérdida de profundidad **ordinal Lrank** ilustrada en la ecuación. (3) se adopta cuando la imagen contiene más de una instancia.

Inferencia El procedimiento general de inferencia para BMP se ilustra en la figura 2. Dada una imagen, BMP primero obtiene el mapa de instancia **C** y el mapa de malla corporal **P** de los cabezales de predicción. Luego realiza la operación de agrupación máxima para encontrar el máximo local en **C** para obtener las posiciones del punto central **N^)} kc** {(x **ci , yc**$_{\mathbf{i}}$**)}**$_{\mathbf{i=1,}}$ donde **k c**$_{\mathbf{i}}$ y **(x ci , yc) i** denotan la pirámide nivel y ubicación del centro del cuerpo para la i-ésima persona, respectivamente, y N^ es el número de personas estimadas. Después de eso, BMP extrae los parámetros de malla corporal de cada persona los pa[illegible] Tomamos la multiplicación de la puntuación OKS predicha y la puntuación de probabilidad del mapa de instancias como la puntuación de confianza para NMS.

3.5. Detalles de implementacion

Implementamos BMP con PyTorch [45] y la biblioteca mmdetection [8] y utilizamos Rectified Adam [33] como optimizador con una tasa de aprendizaje inicial de **1eÿ4.** Cambiamos el tamaño de todas las imágenes a **832 × 512** manteniendo la misma relación de aspecto siguiendo el esquema de entrenamiento original de COCO [57, 65, 20]. Durante el entrenamiento, aumentamos las muestras con volteo horizontal y oclusión consciente de los puntos clave (Sec. 3.3). El aumento de volteo se lleva a cabo durante la prueba. Además, dado que el modelo BMP extrae directamente características a nivel de imagen para estimaciones en lugar de características de cuadros delimitadores recortados, puede tomar imágenes con una resolución menor **(512×512)** como entradas. Denotamos una configuración como BMP-Lite. Otros entrenamientos y pruebas

Los ajustes son los mismos entre BMP-Lite y BMP. Por favor
consulte el suplemento para obtener más detalles.

4. Experimentos

En esta sección, nuestro objetivo es responder las siguientes preguntas. 1)
¿Puede BMP proporcionar datos multipersona eficientes y precisos?
recuperación de malla? 2) ¿Es BMP capaz de dar mallas coherentes para
varias personas con orden de profundidad correcto? 3) es BMP
robusta para los casos en los que las instancias de persona están ocluidas o se
observan parcialmente? Con este fin, llevamos a cabo extensos experimentos en
varios puntos de referencia a gran escala.

4.1. conjuntos de datos

Human3.6M [19] es el 3D para una sola persona más utilizado
pose de referencia recogida en un ambiente interior. Contiene 3,6 millones de
poses en 3D y los videos correspondientes para 15
asignaturas. Debido a sus anotaciones de alta calidad, lo usamos siguiendo [20]
tanto para entrenamiento como para prueba.

Panoptic [23] es un conjunto de datos a gran escala capturado en el estudio
Panoptic, que ofrece anotaciones de poses en 3D para varias personas.
participar en diversas actividades sociales. Utilizamos este conjunto de datos para
evaluación con el mismo protocolo que [69].

MuPoTS-3D [38] es un conjunto de datos de varias personas con pose 3D
anotaciones para escenas interiores y en la naturaleza. Seguimos
low [38] y utilícelo para la evaluación.

3DPW [64] es un conjunto de datos en estado salvaje de varias personas, que
Presenta diversos movimientos y escenas. Contiene 60 videos.
secuencias (24 trenes, 24 pruebas, 12 validaciones) con cuerpo completo
anotaciones de malla. Para verificar la generalización del modelo propuesto a
escenarios desafiantes en la naturaleza, usamos su
conjunto de prueba para evaluación, siguiendo el mismo protocolo que [25].

MPI-INF-3DHP [40] es un 3D multivista para una sola persona
posar conjunto de datos. Contiene 8 actores realizando 8 actividades,
Capturado por 14 cámaras. Mehta et al., [38] generan un
Conjunto de datos de varias personas llamado MuCo-3DHP, de MPI-INF 3DHP a
través de la mezcla de apariencia humana segmentada en primer plano. Usamos
ambos conjuntos de datos para el entrenamiento.

COCO [32], **LSP** [21], **LSP Extended** [22], **Pose Track** [1], **MPII** [2] son conjuntos
de datos en estado salvaje con anotaciones para articulaciones 2D. Los usamos
para entrenar con el
estrategia de entrenamiento débilmente supervisada [24] (ecuación (5)).

4.2. Comparación con el estado de la técnica

Entorno unipersonal Primero evaluamos nuestro BMP propuesto
modelo en el entorno de una sola persona para validar la estrategia
de BMP sobre la factorización de la localización de instancias y la malla
la recuperación no sacrifica el rendimiento. Concretamente,
evaluamos y comparamos el rendimiento de BMP en el
Conjunto de datos Human3.6M a gran escala con los enfoques más competitivos
[24, 20] que comparten el objetivo de regresión similar y
estrategia de aprendizaje Los resultados se muestran en la Tabla 2. Podemos
observe que BMP supera todos estos métodos.

Método HMR [24] CRMH [20] BMP			
PA-MPJPE	56,8	52.7	51.3

Tabla 2. **Resultados en Human3.6M.** Usamos la media por posición conjunta
errores en mm después de la alineación de Procrustes (PA-MPJPE) como métrica.

Configuraciones de varias personas Luego evaluamos nuestro modelo BMP
para la recuperación de malla corporal de varias personas. primero lo evaluamos
en el conjunto de datos de varias personas capturado en el interior Panop tic
Studio [23] y compárelo con los enfoques más competitivos [69, 70, 20]. Como se
muestra en la Tabla 3, nuestro BMP
El modelo logra el mejor rendimiento en todos los escenarios. En general, mejora
el modelo de arriba hacia abajo de última generación.
CRMH [20] en un 5,4 % (135,4 mm frente a 143,2 mm en MPJPE),
mientras ofrece una velocidad de inferencia más rápida1 . Además, supera
significativamente a CRMH para escenarios de ultimátum y pizza con escenas
abarrotadas y oclusión severa, verificando
su robustez a los casos de oclusión. Además, su versión lite, BMP-Lite, es aún
más rápida, ya que solo requiere 0,038 s.
para procesar una imagen, aproximadamente **2 veces** más rápido que CRMH mientras que
lograr un rendimiento comparable. Estos resultados demuestran tanto la eficacia
como la eficiencia de BMP para estimar mallas corporales de varias personas en
una sola etapa.

Método	Regatear Mafia Última. Hora media de pizza[s]						
Zanfir et al. [69] 140.0 165.9 150.7 156.0 153.4 MubyNet [70] 141.4 152.3							-
145.0 162.5 150.3 CRMH [20] 129.6 133.5 153.0 156.7 143.2 0.077							-
BMP-Lite 124,2 138,1 155,2 157,3 143,7 **0,038**							
PMB **120,4 132,7 140,9 147,5 135,4** 0,056							

Tabla 3. **Resultados en el Panóptico.** Usamos MPJPE como evaluación
métrico. Cuanto más bajo, mejor. Mejor en **negrita.**

Otro punto de referencia de estimación de pose 3D popular es el
Conjunto de datos MuPoTS-3D [40]. Comparamos nuestro método con
dos líneas de base sólidas, 1) la combinación de OpenPose [6]
con métodos de recuperación de malla de una sola persona (SMPLify X [46] y
HMR [24]), y 2) el estado del arte de arriba hacia abajo
enfoque CRMH [20]. Reportamos los resultados en la Tabla 4. Como
Como podemos ver, BMP supera significativamente a los métodos anteriores en
ambos protocolos de evaluación.

Método		Emparejado	Veces]
SMPLify-X [46]	Todos	68,04	6.4
HMR [24]	62,84	70,90	0.26
CRMH [20]	66,09	72,22	0.083
BMP-Lite	69,12	71,92	**0.038**
BMP	68,63	**75,34**	0.056

73,83 Tabla 4. **Resultados en MuPoTS-3D.** Los números son 3DPCK. Nosotros
informe la precisión general (Todos) y la precisión solo para la persona
anotaciones coincidentes con una predicción (Matched). Mejor en **negrita.**

Por último, comparamos nuestro modelo BMP con enfoques de última
generación en el desafiante 3DPW en estado salvaje
conjunto de datos Algunos enfoques utilizan la estrategia de autoformación

1Contamos el tiempo de inferencia por imagen en segundos. Para todos los métodos, el tiempo se cuenta en GPU Tesla P100 y CPU Intel E5-2650 v2@ 2,60 GHz, sin utilizar aumento de tiempo de prueba.

(es decir, SPIN [26]) o información temporal (es decir, VIBE [25]),
y confían en detectores de personas listos para usar [6, 50]. Como
como se muestra en la Tabla 5, nuestro BMP supera a CRMH [20] y
SPIN [26] en términos de 3DPCK mientras mantiene una eficiencia atractiva y logra
resultados comparables con
VIBE [25] sin depender de ninguna información temporal.
Además, BMP-Lite obtiene aproximadamente el mismo rendimiento que el modelo
CRMH de última generación al mismo tiempo que logra
Velocidad de inferencia 2,1 veces más rápida. Los resultados confirman aún más
la efectividad de nuestra solución de una etapa sobre las existentes
estrategias multietapa, con eficiencia muy competitiva.

Método	PCK	AUC	MPJPE	PA-MPJPE	PVE	Tiempo[s]
GIRO [26]	30,8	53,4	VIBE [25] **33,9**	99,4	68,1	-	0.31
56,6 CRMH [20]	25,3	51,6	102,2	**94,7**	66,1	**112.7**	-
51,3	108,5	BMP	32,1	54,5	104,1	**62,3**	122,2	0,09
						64,0	126,2	**0,038**
						63,8	119,3	0,056

Tabla 5. **Resultados en 3DPW.** Usamos 3DPCK, AUC, MPJPE, PA MPJPE y error por vértice (PVE) como métricas de evaluación.

Resultados cualitativos Visualizamos algunas reconstrucciones de malla corporal de BMP en el desafiante PoseTrack, MPII y
Conjuntos de datos COCO, como se muestra en la Fig. 3. Se puede observar que BMP es resistente a la oclusión severa y escenas concurridas y
puede reconstruir cuerpos humanos con el orden de profundidad correcto.

4.3. Estudios ablativos

Realizamos análisis de ablación en Panoptic, 3DPW y
Conjuntos de datos MuPoTS-3D tanto cualitativa como cuantitativamente
para justificar nuestras elecciones de diseño. El análisis cualitativo de la
El método propuesto se ilustra en la Fig. 4.
Representación de instancias de personas Primero evaluamos la representación
basada en puntos 3D propuesta para instancias de personas.
La principal diferencia entre la representación propuesta
y la representación espacial 2D anterior [43, 75, 68] es que
usamos una dimensión de profundidad adicional para diferenciar a la persona
instancias en el espacio de profundidad discretizado a través de FPN. Nosotros
luego compare BMP con un modelo de referencia (es decir, BMP usando
representación espacial 2D). Para una comparación justa, agregamos funciones
de puerta de todos los niveles de la pirámide FPN en la línea de base
modelo para obtener un resultado único tanto para la localización de instancias
como para la recuperación de la malla del cuerpo. En concreto, estudiamos tres
métodos para la agregación: cambiamos el tamaño de todas las pirámides de
características a una escala de 1/8 y luego las agregamos por 1) suma de
elementos (Baseline-Add), 2) concatenación (Baseline Concat), o 3) adopción de
una capa convolucional después de la concatenación ellos (Baseline-Conv). Los
resultados se muestran en la Tabla 6.
Podemos ver que nuestro modelo BMP mejora con respecto a la línea de base
modelos por un amplio margen en todos los conjuntos de datos, lo que demuestra su eficacia
para la recuperación de la malla corporal. Además, de la Fig. 4 (primera fila),
observamos BMP con la representación propuesta es más
robusto en el manejo de instancias de oclusión, especialmente cuando el
los centros del cuerpo de múltiples instancias caen en el mismo espacio
coordenadas de cuadrícula, mientras que la representación basada en 2D sería
suele fallar. Consulte el suplemento para el análisis.
en el nivel de la pirámide K.

Método	Panóptico (ÿ)	3DPW (ÿ)	MuPoTS-3D (ÿ)
Línea de base-Agregar	159,1	120,4	68,03
Baseline-Concat	150,3	114,6	68,52
Conv. de referencia	145,6	110,8	69,34
BMP	135,4	104,1	73,83

Tabla 6. **Ablación para representación de instancia de persona.** informamos
MPJPE para Panoptic y 3DPW, y 3DPCK para MuPoTS-3D.

Pérdida de profundidad ordinal Para investigar si la profundidad ordinal
Loss **Lrank** puede ayudar a producir resultados más coherentes con
correcto orden de profundidad, llevamos a cabo experimentos en el
Conjunto de datos MuPoTS-3D. Específicamente, evaluamos el ordinal
relaciones de profundidad de todos los pares de instancias en la escena y el informe
el porcentaje de relaciones de profundidad ordinal estimadas correctamente
en la Tabla 7. El modelo entrenado con **Lrank mejora** significativamente sobre la
línea de base (BMP entrenado sin **Lrank)** (de
91,42% a 94,50%). Estas mejoras también se pueden observar en la Fig. 4
(segunda fila). Además, al comparar nuestro método con Moon et al. [42] y CRMH
[20] ,
observe que BMP logra una mayor precisión con la profundidad relativa
ordenando que CRMH que solo considera la pérdida ordinal para
pares superpuestos (94,50% vs. 93,68%). Esto demuestra
nuestra pérdida ordinal completa por pares puede proporcionar una supervisión
más completa en el diseño de profundidad de la escena y
así entrenar el modelo para dar resultados más coherentes.

Método	Moon [42]	CRMH [20]	BMP sin Lrank	BMP
Precisión	90,85%	93,68%	91,42%	94,50%

Tabla 7. **Ablación por pérdida de profundidad ordinal** Orden de profundidad relativa
se muestran los resultados en MuPoTS-3D. Evaluamos la profundidad ordinal
relaciones de todos los pares de instancias en la escena e informar el porcentaje
de relaciones de profundidad ordinales estimadas correctamente.

Aumento de la oclusión con reconocimiento de puntos clave Finalmente,
estudiar el impacto de la estrategia propuesta de aumento de la oclusión consciente
de los puntos clave. Comparamos nuestro modelo BMP
con los modelos entrenados sin aumento de oclusión
(BMP-NoAug) y entrenado usando oclusión sintética aleatoria [55] (BMP-RandOcc)
en la Tabla 8. Podemos ver BMP
supera a ambos por un amplio margen en todos los conjuntos de datos.
En particular, trae respectivamente mejoras del 9,1 % y el 17,3 % sobre BMP-
NoAug en conjuntos de datos Panoptic y 3DPW,
que presentan escenas llenas de gente con severa superposición y observación
parcial. Por el contrario, el aumento aleatorio duele
rendimiento del modelo en MuPoTS-3D (71,71 frente a 70,78). Este
verifica que nuestro aumento de oclusión propuesto puede forzar
el modelo para centrarse en la estructura del cuerpo y así mejorar su
robustez a la oclusión.

5. Conclusiones

En este trabajo presentamos el primer modelo de una sola etapa,
Body Meshes as Points (BMP), para body mesh de varias personas

Figura 3. **Resultados cualitativos.** Visualizamos las reconstrucciones de nuestro enfoque en PoseTrack (1ra fila), MPII (2da fila) y COCO (3ra fila) desde diferentes puntos de vista: frontal (fondo verde), superior (fondo azul) y lateral (fondo rojo), respectivamente. Por favor refiérase a suplementario para resultados más cualitativos.

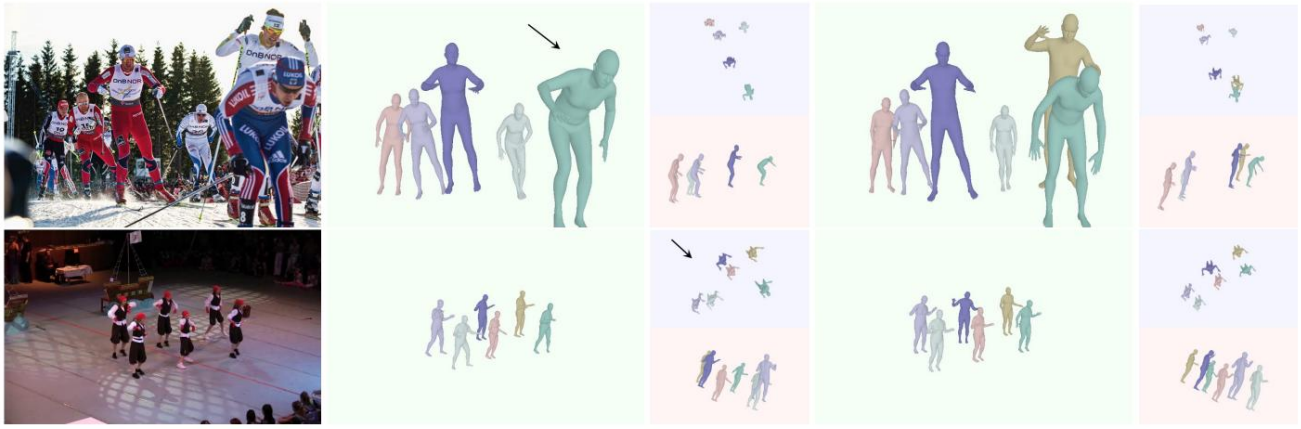

Figura 4. **Efecto cualitativo del método propuesto.** Resultados de la línea de base 1 (BMP usando representación 2D) (primera fila central), línea de base 2 (BMP entrenado sin Lrank) (segunda fila central) y BMP (derecha). Los errores se resaltan con flechas negras. Como era de esperar, los métodos propuestos toman efecto en la producción de mejores resultados (es decir, robusto a instancias superpuestas, ordenación de profundidad más consistente para mallas de cuerpo estimadas).

Método	Panóptico (ÿ)	3DPW (ÿ)	MuPoTS-3D (ÿ)
BMP-NoAgo	148,9	125,9	71,71
BMP-RandOcc	144,6	110,3	70,78
BMP	135,4	104,1	73,83

Tabla 8. **Ablación para aumento de oclusión.** Usamos MPJPE para los dos primeros y 3DPCK para el último conjuntos de datos como métricas.

recuperación. BMP introduce un nuevo método de representación para habilitar una canalización tan compacta: cada instancia de persona es representado como un punto en el espacio de profundidad espacial que es asociado con una malla de cuerpo parametrizada. con tal representación, BMP puede explotar completamente las características compartidas y realizar la localización de personas y la recuperación de malla corporal simultáneamente taneosamente. BMP mejora significativamente sobre convencional paradigmas de dos etapas, y ofrece una excelente eficiencia y precisión, validada por extensos experimentos en múltiples puntos de referencia. Además, BMP desarrolla varias técnicas nuevas para mejorar aún más la coherencia y robustez de mallas corporales recuperadas, que son de amplio interés para otras aplicaciones como estimación y detección de poses humanas. En el futuro, exploraremos cómo hacer el modelo. más compacto y mejorar aún más su eficiencia, así como extenderse al modelado de interacciones entre personas.

Agradecimientos Esta investigación fue financiada parcialmente por AISG-100E-2019-035, MOE2017-T2-2-151, NUS ECRA FY17 P08 y CRP20-2017-0006.

Referencias

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall y Bernt Schiele. Posetrack: un punto de referencia para la estimación y el seguimiento de poses humanas. En CVPR, 2018.

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler y Bernt Schiele. Estimación de la pose humana en 2D: un nuevo punto de referencia y un análisis de vanguardia. En CVPR , 2014 .

[3] Anurag Arnab, Carl Doersch y Andrew Zisserman. Explotación del contexto temporal para la estimación de la pose humana en 3D en la naturaleza. En CVPR, 2019.

[4] Abdallah Benzine, Florian Chabot, Bertrand Luvison, Quoc Cuong Pham y Catherine Achard. Pandanet: Estimación de poses 3D multipersona de un solo disparo basada en ancla.
En CVPR, 2020.

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero y Michael J Black. Keep it smpl: Estimación automática de la pose y la forma humana en 3D a partir de una sola imagen. En ECCV, 2016.

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei y Yaser Sheikh. Openpose: estimación de poses en 2D para varias personas en tiempo real utilizando campos de afinidad de piezas. arXiv, 2018.

[7] Zhe Cao, Tomas Simon, Shih-En Wei y Yaser Sheikh.
Estimación de pose 2d de varias personas en tiempo real utilizando campos de afinidad parcial. En CVPR, 2017.

[8] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tian heng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy y Dahua Lin. MMDetection: caja de herramientas de detección abierta de mmlab y evaluación comparativa. arXiv, 2019.

[9] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma y Arjun Jain. Aprendiendo la pose del hombre humano 3d a partir de la estructura y el movimiento. En ECCV, 2018.

[10] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qing ming Huang y Qi Tian. Centernet: trillizos de puntos clave para la detección de objetos. En ICCV, 2019.

[11] Mark Everingham y John Winn. El kit de desarrollo Pascal Visual Object Classes Challenge 2012 (voc2012). Trans. IEEE. sobre análisis de patrones e inteligencia artificial, 8, 2011.

[12] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Albed y Rita Cucchiara. Mapas de calor volumétricos comprimidos para la estimación de poses 3D de varias personas. En CVPR, 2020.

[13] Valentin Gabeur, Jean-Sebastien Franco, Xavier Martin, Cordelia Schmid y Gregory Rogez. Moldeo de seres humanos: Estimación no paramétrica de la forma humana en 3D a partir de imágenes individuales. En CVPR, 2019.

[14] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri y Du Tran. Detección y seguimiento: estimación eficiente de poses en videos. En CVPR, 2018.

[15] Kehong Gong, Jianfeng Zhang y Jiashi Feng. Poseaug: Un marco diferenciable de aumento de poses para la estimación de poses humanas en 3D en la naturaleza. En CVPR, 2021.

[16] Riza Alp Guler y Iasonas Kokkinos. Holopose: Reconstrucción humana holística en 3D en la naturaleza. En CVPR, 2019.

[17] Wen Guo, Enrique Corona, Francisco Moreno-Noguer, y Xavier Alameda-Pineda. Pino-net: Pose interacting network for multi-persona monocular 3d pose estimation. arXiv, 2020.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren y Jian Sun.
Aprendizaje residual profundo para el reconocimiento de imágenes. En CVPR, 2016.

[19] Catalin Ionescu, Dragos Papava, Vlad Olaru y Cristian Sminchisescu. Humano3. 6m: conjuntos de datos a gran escala y métodos predictivos para la detección humana en 3D en entornos naturales. Trans. IEEE. on Pattern Analysis and Machine Intelligence, 36(7):1325–1339, 2014.

[20] Wen Jiang, Nikos Kolotouros, George Pavlakos, Xiaowei Zhou y Costas Daniilidis. Reconstrucción coherente de múltiples humanos a partir de una sola imagen. En CVPR,

[21] Sam Johnson y Mark Everingham. Poses agrupadas y modelos de apariencia no lineal para la estimación de poses humanas. En BMVC, 2010.

[22] Sam Johnson y Mark Everingham. Aprender una estimación efectiva de la pose humana a partir de una anotación inexacta. En CVPR, 2011.

[23] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara y Yaser Sheikh. Panoptic studio: un sistema masivo de múltiples vistas para la captura de movimiento social. En ICCV, 2015.

[24] Angjoo Kanazawa, Michael J. Black, David W. Jacobs y Jitendra Malik. Recuperación integral de la forma y la pose humanas. En CVPR, 2018.

[25] Muhammed Kocabas, Nikos Athanasiou y Michael J Black. Vibe: inferencia de video para la estimación de la postura y la forma del cuerpo humano. En CVPR, 2020.

[26] Nikos Kolotouros, Georgios Pavlakos, Michael J Black y Kostas Daniilidis. Aprender a reconstruir la pose y la forma humana en 3D mediante el ajuste de modelos en el bucle. En ICCV, 2019.

[27] Nikos Kolotouros, George Pavlakos y Costas Dani ilidis. Regresión de malla convolucional para la reconstrucción de formas humanas de una sola imagen. En CVPR,

[28] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black y Peter V Gehler. Une a la gente: Cerrando el círculo entre las representaciones humanas en 3D y 2D. En CVPR, 2017.

[29] Jiefeng Li, Can Wang, Wentao Liu, Chen Qian y Cewu Lu. Hmor: relaciones ordinales jerárquicas de varias personas para la estimación monocular de poses en 3D de varias personas. En ECCV, 2020.

[30] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan y Serge Belongie. Cuenta con redes piramidales para la detección de objetos. En CVPR, 2017.

[31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He y Piotr Dollar. Pérdida focal para la detección de objetos densos. En ICCV, 2017.

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar y C Lawrence Zitnick. Microsoft coco: objetos comunes en contexto. En ECCV, 2014.

[33] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao y Jiawei Han. Sobre la variación de la tasa de aprendizaje adaptativo y más. arXiv, 2019.

[34] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev y Jason Yosinski. Una falla intrigante de las redes neuronales convolucionales y la solución coordconv. En NeurIPS, 2018.

[35] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma y Shenghua Gao. Liquid warping gan: un marco unificado para la imitación del movimiento humano, la transferencia de apariencia y la síntesis de vistas novedosas. En ICCV, 2019.

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll y Michael J Black. Smpl: un modelo lineal de varias personas con piel. Transacciones de ACM en gráficos (TOG), 34(6):1–16, 2015.

[37] Julieta Martínez, Rayat Hossain, Javier Romero y James J Little. Una línea de base simple pero efectiva para la estimación de la pose humana en 3D. En ICCV, 2017.

[38] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll y Christian Theobalt. Estimación de pose 3d multipersona de disparo único a partir de rgb monocular. En 3DV, 2018.

[39] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll y Christian Theobalt. Estimación de pose 3d multipersona de disparo único a partir de rgb monocular. En 3DV, 2018.

[40] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas y Christian Theobalt. Vnect: Estimación de la pose humana en 3D en tiempo real con una sola cámara RGB. ACM Trans. en Gráficos, 36(4):44, 2017.

[41] Aymen Mir, Thiemo Alldieck y Gerard Pons-Moll. Aprender a transferir texturas de imágenes de ropa a humanos en 3D. En CVPR, 2020.

[42] Gyeongsik Moon, Ju Yong Chang y Kyoung Mu Lee. Enfoque de arriba hacia abajo con reconocimiento de la distancia de la cámara para la estimación de la pose de varias personas en 3D a partir de una sola imagen RGB. En ICCV, 2019.

[43] Xuecheng Nie, Jianfeng Zhang, Shuicheng Yan y Jiashi Feng. Máquinas de pose para varias personas de una sola etapa. En ICCV, 2019.

[44] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler y Bernt Schiele. Adaptación neuronal del cuerpo: unificación del aprendizaje profundo y la estimación de la forma y la postura humana basada en modelos. En 3DV, 2018.

[45] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Al Ban Desmaison, Luca Antiga y Adam Lerer. Diferenciación Automática en Pytorch. En NeurIPSw , 2017 .

[46] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas y Michael J Black. Captura expresiva del cuerpo: manos, rostro y cuerpo en 3D a partir de una sola imagen. En CVPR, 2019.

[47] Georgios Pavlakos, Xiaowei Zhou y Kostas Daniilidis. Supervisión de profundidad ordinal para la estimación de la pose humana en 3D. En CVPR, 2018.

[48] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou y Kostas Daniilidis. Aprender a estimar la pose y la forma humana en 3D a partir de una sola imagen en color. En CVPR, 2018.

[49] Alin-Ionut Popa, Mihai Zanfir y Cristian Sminchisescu. Arquitectura multitarea profunda para detección humana 2D y 3D integrada. En CVPR, 2017.

[50] Joseph Redmon y Ali Farhadi. Yolov3: Una mejora incremental. arXiv, 2018.

[51] Shaoqing Ren, Kaiming He, Ross Girshick y Jian Sun. R-cnn más rápido: hacia la detección de objetos en tiempo real con redes de propuesta de región. En NeurIPS, 2015.

[52] Chris Rockwell y David F. Fouhey. Conciencia de cuerpo completo a partir de observaciones parciales. En ECCV, 2020.

[53] Gregory Rogez, Philippe Weinzaepfel y Cordelia Schmid. Lcr-net: Localización-clasificación-regresión para pose humana. En CVPR, 2017.

[54] Gregory Rogez, Philippe Weinzaepfel y Cordelia Schmid. Lcr-net++: Detección de poses 2D y 3D de varias personas en imágenes naturales. Trans. IEEE. on Pattern Analysis and Machine Intel ligence, 42(5):1146–1161, 2019.

[55] Istvan Sar´andi, Timm Linder, Kai O Arras y Bastian Leibe. ¿Qué tan robusta es la estimación de la pose humana en 3D para la oclusión? En ICRAw, 2018.

[56] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis y Javier Romero. Facsímil: escaneos rápidos y precisos de una imagen en menos de un segundo. En ICCV, 2019.

[57] Kai Su, Dongdong Yu, Zhenqi Xu, Xin Geng y Changhu Wang. Estimación de posturas de varias personas con información espacial y de canal mejorada. En CVPR, 2019.

[58] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang y Yichen Wei. Regresión integral de la pose humana. En ECCV, 2018.

[59] Zhi Tian, Chunhua Shen, Hao Chen y Tong He. Fcos: Detección de objetos de una etapa totalmente convolucional. En ICCV, 2019.

[60] Denis Tomé, Chris Russell y Lourdes Agapito. Levantamiento desde lo profundo: Estimación de pose 3d convolucional a partir de una sola imagen. En CVPR, 2017.

[61] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer y Katerina Fragkiadaki. Aprendizaje autosupervisado de captura de movimiento. En NeurIPS, 2017.

[62] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev y Cordelia Schmid. Bodynet: inferencia volumétrica de formas del cuerpo humano en 3D. En ECCV, 2018.

[63] Gul Varol, Ivan Laptev, Cordelia Schmid y Andrew Zisserman. Humanos sintéticos para el reconocimiento de acciones desde puntos de vista no vistos. arXiv, 2019.

[64] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn y Gerard Pons-Moll. Recuperando la pose humana 3d precisa en la naturaleza usando imus y una cámara en movimiento. En ECCV, 2018.

[65] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang y Lei Li. Solo: segmentación de objetos por ubicación. En ECCV, 2019.

[66] Philippe Weinzaepfel, Romain Bregier, Hadrien Combaluzier, Vincent Leroy y Gregory Rogez. Dope: Destilación de expertos parciales para la estimación de poses 3D de cuerpo entero en la naturaleza. En ECCV, 2020.

[67] Yuxin Wu y Kaiming He. Normalización del grupo. En ECCV, 2018.

[68] Sun Yu, Bao Qian, Liu Wu, Fu Yili y Mei Tao. Centerhmr: un método de disparo único de abajo hacia arriba para la recuperación de malla 3D de varias personas a partir de una sola imagen. arXiv, 2020.

[69] Andrei Zanfir, Elisabeta Marinoiu y Cristian Sminchis escu. Pose monocular 3d y estimación de la forma de múltiples personas en escenas naturales: la importancia de las restricciones de múltiples escenas. En CVPR, 2018.

[70] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa y Cristian Sminchisescu. Red profunda para la detección 3D integrada de múltiples personas en imágenes naturales. En NeuroIPS, 2018.

[71] Jianfeng Zhang, Xuecheng Nie y Jiashi Feng. Optimización de la etapa de inferencia para la estimación de la pose humana en 3D entre escenarios. En NeurIPS, 2020.

[72] Tianshu Zhang, Buzhen Huang y Yangang Wang. Objeto ocluido forma humana y estimación de pose a partir de una sola imagen en color. En CVPR, 2020.

[73] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao y Xiaowei Zhou. Smap: Estimación de pose 3D absoluta multipersona de un solo disparo. En ECCV, 2020.

[74] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue y Yichen Wei. Hacia la estimación de la pose humana en 3D en la naturaleza: un enfoque poco supervisado. En ICCV, 2017.

[75] Xingyi Zhou, Dequan Wang y Philipp Krahenb ¨uhl.Ob-¨ objetos como puntos. arXiv, 2019.

[76] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang y Hao Li. Sobre la continuidad de las representaciones de rotación en redes neuronales. En CVPR, 2019.