



Escuela de Másteres Oficiales

Reconstrucción 3D multivista y detección de *landmarks* faciales en CMU Panoptic

Memoria del Trabajo Fin de Máster

Máster Universitario en Visión Artificial

Autora:

Esther Vera Moreno

Tutor: José Miguel Buenaposada Biencinto

Abril 2024

Agradecimientos

Quiero agradecer primero de todo, a la Universidad Rey Juan Carlos por ofrecerme la oportunidad de realizar este proyecto. Concretamente, a mi tutor José Miguel Buenapospada, por su paciencia durante el desarrollo de este Trabajo Fin de Máster, gracias por todo el *feedback* y los sabios consejos.

A todos los compañeros de trabajo que he tenido la suerte de conocer durante mis prácticas y mi primer empleo, gracias por ayudarme, guiarme y hacerme mejorar haciendo lo que más me gusta.

En especial, agradecer a mis profesores de la Universidad de Alicante, mis profesores de la Universidad de Skövde y mis profesores de la Universidad Rey Juan Carlos, que me han llenado del conocimiento necesario para seguir mi camino y poder desarrollar esta tesis.

A todos mis amigos, los que siempre estuvieron allí, los que me acompañaron en las aventuras más increíbles, los que conocí por el camino, los que me acogieron en Skövde, los que me acogieron en Madrid y los que me acogen ahora en Barcelona. Gracias por alegrarme la vida.

Por último, a mi familia, por ser el apoyo fundamental en cada corriente y contracorriente. Nunca podré devolveros todas las cosas que me habéis dado, os debo todo.

Gracias a todos por haber confiado en mí y hacerme cada día un poquito más feliz.

Resumen

Con el avance de las nuevas tecnologías, cada vez existen más aplicaciones donde la creación y análisis de modelos faciales resulta indispensable. Tanto en el ámbito del ocio, utilizando estos sistemas para reproducir reconstrucciones faciales dentro de interfaces, como en el ámbito de la seguridad, para la detección temprana de incidencias, resulta un área en continuo desarrollo. Sin embargo, para el entrenamiento de estos modelos de redes neuronales es necesario el uso de grandes conjuntos de datos faciales 3D. A pesar de que existen otras bases de datos de laboratorio multivista, el conjunto de CMU Panoptic destaca por su gran diversidad demográfica y de movimiento.

El objetivo principal de este proyecto es, por tanto, anotar la base de datos de Panoptic para obtener la reconstrucción 3D completa de la cabeza de los sujetos mediante información multivista. En cada instante de tiempo de una escena en Panoptic, se obtienen las detecciones de los *landmarks* faciales con el modelo DAD-3DNet sobre las imágenes de las distintas cámaras. Para seleccionar las vistas de la cámara con predicciones más precisas, se estudian varios métodos basados en la ordenación de predicciones según su error de reprojeción frente a las anotaciones de Panoptic y un *ground-truth* elaborado manualmente. Tras seleccionar un método basado en la comparación de cámaras por pares mediante RANSAC, se obtienen las nuevas reconstrucciones 3D y los puntos proyectados sobre las imágenes anteriores, constituyendo la nueva base de datos sobre Panoptic.

Con estos datos, se procede a validar los resultados mediante su utilización en una tarea de análisis facial. Para ello, se ha seleccionado un detector de *landmarks* del rostro y se han realizado dos entrenamientos con diferentes funciones de pérdida. Un primer procedimiento donde todos los puntos característicos poseen el mismo valor en el error y otro donde se tiene en cuenta la condición de visibilidad de cada punto. Esta visibilidad se extrae según la orientación del sujeto en la imagen y algunas restricciones geométricas. Los resultados se evalúan frente a las detecciones de DAD-3DNet y las anotaciones refinadas, logrando unas predicciones adecuadas con ambos modelos.

En resumen, en este proyecto se alcanza con éxito el objetivo de proporcionar un método de anotación de la base de datos de Panoptic, obteniendo el modelo 3D de los rostros y la proyección de sus puntos en cada imagen. Además se consiguen utilizar estos datos en los entrenamientos de detectores de puntos de referencia faciales. El error de reprojeción obtenido por el método final de reconstrucción es menor de 15 píxeles para las escenas evaluadas, llegando a reducir el error desde unos pocos píxeles hasta casi 400 píxeles en comparación con una reconstrucción ordenada de forma aleatoria. También se comparan estos modelos 3D frente a las predicciones originales y un procedimiento de texturización facial, destacando el mejor ajuste sobre la parte lateral de la cabeza. Finalmente, para los detectores de *landmarks*, se implementa la característica de visibilidad en los entrenamientos consiguiendo una tendencia adecuada en la detección de puntos.

Índice general

1. Introducción	1
1.1. Motivación	1
1.2. Planteamiento del problema	3
1.3. Objetivos	4
1.4. Alcance y limitaciones	5
1.5. Estructura de la memoria	6
2. Estado del arte	7
2.1. Bases de datos de reconstrucciones faciales 3D	7
2.2. Reconstrucción 3D del rostro	9
3. Metodología	14
3.1. Análisis de la base de datos Panoptic	14
3.1.1. Estructura multicámara	14
3.1.2. Anotaciones disponibles	15
3.2. Modelo DAD-3DNet	17
3.3. Parámetros de la cámara	18
3.4. Método de reconstrucción	18
3.4.1. Reconstrucción triangular lineal	19
3.4.2. Refinamiento con <i>Bundle Adjustment</i>	20
3.4.3. Ordenación y selección de cámaras	21
3.5. Método de evaluación de las reconstrucciones	22
3.5.1. Selección de escenas de evaluación	22
3.5.2. Anotación manual de las imágenes	24
3.5.3. Evaluación de las reconstrucciones refinadas	24
3.6. Visualización de las reconstrucciones 3D	26
3.7. Entrenamiento de los detectores de <i>landmarks</i> 2D con las reconstrucciones 3D en Panoptic	26
3.7.1. Definición de visibilidad de los <i>landmarks</i> 2D anotados	26
3.7.2. Entrenamiento de los detectores de <i>landmarks</i> 2D con las nuevas anotaciones	27

4. Desarrollo	29
4.1. Preprocesamiento de las imágenes de Panoptic	30
4.2. Reconstrucción 3D y anotación de la base de datos de Panoptic	32
4.2.1. Predicciones de las vistas con DAD-3DNet	32
4.2.2. Reconstrucción facial 3D	33
4.2.3. Algoritmos de ordenación de cámaras	33
4.2.4. Selección de cámaras para la reconstrucción facial 3D final	35
4.2.5. Anotaciones faciales de la base de datos de Panoptic	36
4.3. Extracción de la visibilidad de los <i>landmarks</i> 2D anotados	36
4.4. Entrenamiento de los detectores de <i>landmarks</i> 2D	39
5. Resultados experimentales	41
5.1. Evaluación de los métodos de ordenación de vistas	41
5.1.1. Evaluación de los métodos frente a las anotaciones de Panoptic . .	42
5.1.2. Evaluación de los métodos frente a las anotaciones manuales . . .	45
5.2. Evaluación de la reconstrucción final	47
5.2.1. Evaluación de la reconstrucción final frente a Panoptic	48
5.2.2. Evaluación de la reconstrucción final frente a las anotaciones manuales	49
5.2.3. Evaluación visual frente a las predicciones de DAD-3DNet	52
5.2.4. Visualización de la reconstrucción 3D final	54
5.3. Resultados de los detectores de <i>landmarks</i> con las nuevas anotaciones . .	56
6. Conclusiones	60
7. Trabajo futuro	63

Índice de tablas

1.	Especificaciones de las escenas de evaluación seleccionadas.	23
2.	Tiempo medio de ordenación de las cámaras por cada método.	42
3.	Error de las reconstrucciones para cada método frente a Panoptic.	43
4.	Error de las reconstrucciones para cada método frente a la anotación manual.	45

Índice de figuras

1.	Error en mm de la reconstrucción 3D generada mediante anotaciones 2D [3].	8
2.	Procedimiento de captura de imágenes a modelos uniformes [20].	8
3.	Alineación de los modelos 3D multivista según la confianza de los puntos visibles y ocultos [25]	10
4.	Mallas 3D generadas por DAD-3DNet [3].	11
5.	Reconstrucciones con textura obtenidas en [34].	11
6.	Refinamiento de rasgos utilizando Shape-from-Shading (SfS) [34].	12
7.	Comparación de reconstrucciones faciales obtenidas con distintos métodos [29].	12
8.	Proceso iterativo para la creación del modelo facial [35].	12
9.	Estructura para la grabación de escenas de Panoptic [15].	15
10.	Anotaciones proporcionadas por Panoptic imágenes.	15
11.	Puntos seleccionados de Panoptic utilizados en el preprocesamiento.	16
12.	Puntos seleccionados de Panoptic utilizados en la evaluación.	16
13.	Comparación entre los puntos anotados por Panoptic con los anotados manualmente.	16
14.	Estructura del modelo DAD-3DNet [14].	17
15.	Predicciones ofrecidas por DAD-3DNet.	18
16.	Representación visual del Bundle Adjustment (BA)	20
17.	Escenas de evaluación seleccionadas.	23
18.	Estructura de los 27 <i>landmarks</i> en CVAT.	24
19.	Diagrama del proyecto.	29
20.	Poses válidas.	30
21.	Poses no válidas filtradas.	31
22.	Oclusiones entre una misma persona.	31
23.	Oclusiones entre personas filtradas.	31
24.	Detecciones de <i>landmarks</i> faciales con DAD-3DNet.	32

ÍNDICE DE FIGURAS

25. Detecciones de <i>landmarks</i> faciales con DAD-3DNet que presentan un error considerable	33
26. Orientación del sujeto en la imagen.	36
27. Vectores de orientación sobre la reconstrucción.	37
28. Sistema de referencia 3D.	37
29. Vector contenido en el plano.	38
30. Plano visibilidad de los <i>landmarks</i>	38
31. Extracción de los puntos visibles según la orientación de la cabeza.	39
32. Ejemplo de imágenes utilizadas en el entrenamiento con sus etiquetas.	40
33. Ejemplo de imágenes incorrectas utilizadas en el entrenamiento con sus etiquetas.	40
34. Error de reproyección para cada escena frente a Panoptic.	43
35. Error de reproyección para cada reconstrucción frente a Panoptic (escena 2). .	44
36. Error de reproyección para cada reconstrucción frente a Panoptic (escena 10).	44
37. Error de reproyección para cada reconstrucción frente a Panoptic (escena 11).	44
38. Error de reproyección para cada escena frente a las anotaciones manuales. .	46
39. Error de reproyección para cada reconstrucción frente a las anotaciones manuales (escena 2).	46
40. Error de reproyección para cada reconstrucción frente a las anotaciones manuales (escena 4).	47
41. Error de reproyección para cada reconstrucción frente a las anotaciones manuales (escena 6).	47
42. Error de reproyección de la reconstrucción final para cada escena frente a Panoptic.	48
43. Error de reproyección de la reconstrucción final para cada cámara frente a Panoptic (escena 11).	49
44. Error de reproyección de la reconstrucción final para cada cámara frente a Panoptic (escena 12).	49
45. Error de reproyección para cada <i>landmark</i> frente a Panoptic (escena 11). .	49
46. Error de reproyección de la reconstrucción final para cada escena frente a las anotaciones manuales.	50
47. Error de reproyección de la reconstrucción final para cada cámara frente a las anotaciones manuales (escena 2).	50
48. Error de reproyección de la reconstrucción final para cada cámara frente a las anotaciones manuales (escena 5).	50
49. Error de reproyección de la reconstrucción final para cada <i>landmark</i> frente a las anotaciones manuales (escena 2).	51

50.	Error de reproyección de la reconstrucción final para cada <i>landmark</i> frente a las anotaciones manuales (escena 5).	51
51.	Comparación de <i>landmarks</i> (escena 2).	51
52.	Comparación de <i>landmarks</i> (escena 5).	52
53.	En la primera fila se muestran las reproyecciones refinadas utilizando el algoritmo RANSAC y Panoptic (escena 2). En la segunda fila se presentan las reproyecciones predichas por DAD-3DNet (escena 2).	52
54.	En la primera fila se muestran las reproyecciones refinadas utilizando el algoritmo RANSAC y Panoptic (escena 12). En la segunda fila se presentan las reproyecciones predichas por DAD-3DNet (escena 12).	53
55.	En la primera fila se muestran las reproyecciones refinadas utilizando el algoritmo RANSAC y Panoptic (escena 11). En la segunda fila se presentan las reproyecciones predichas por DAD-3DNet (escena 11).	53
56.	En la primera fila se muestran las reproyecciones refinadas utilizando el algoritmo RANSAC y Panoptic (escena 6). En la segunda fila se presentan las reproyecciones predichas por DAD-3DNet (escena 6).	54
57.	Visualización texturizada (escena 1).	55
58.	Error de las reconstrucciones 3D según el modelo texturizado (escena 1).	55
59.	Visualización texturizada (escenas 3, 9, 10 y 15).	55
60.	Pérdidas de entrenamiento y validación.	56
61.	Ejemplo de las predicciones obtenidas.	57
62.	Ejemplo de ciertos errores sobre las predicciones obtenidas.	58

Glosario

3DMM Modelo 3D Deformable. [2](#), [7](#), [9–11](#), [17](#), [63](#)

BA Bundle Adjustment. [3–5](#), [19–21](#), [32–34](#), [42](#), [47](#), [48](#), [60](#)

DLT Transformación Lineal Directa. [5](#), [19](#), [20](#), [48](#)

GAN Red Neuronal Generativa Adversaria. [11](#)

LM Levenberg-Marquardt. [21](#)

LMM Modelo Local Deformable. [10](#)

MSE Error Medio Cuadrático. [27](#)

PCA Análisis de Componentes Principales. [10](#)

RANSAC RANdom SAMpled Consensus. [3](#), [4](#), [18](#), [22](#), [34](#), [35](#), [42–45](#), [47](#), [49–53](#), [60](#), [61](#)

RMSE Error de Raíz Cuadrada Media. [24](#), [25](#), [42](#), [56](#)

SfM Structure from Motion. [14](#)

SfS Shape-from-Shading. [9–12](#)

SVD Descomposición en Valores Singulares. [20](#)

Capítulo 1

Introducción

En este capítulo, se describe el contenido fundamental del presente Trabajo Fin de Máster. Se comienza contextualizando la motivación del proyecto escogido y el planteamiento del problema que será abordado. Después, se delinean los objetivos generales y específicos a conseguir en esta investigación junto con los alcances y limitaciones que influyen de forma inevitable en el desarrollo. Para concluir, se expone la descripción general de la estructura del proyecto.

1.1. Motivación

En los últimos años, ha habido un notable esfuerzo en la construcción de sistemas de visión artificial destinados a detectar puntos característicos de la cara y generar modelos faciales en 3D. De forma específica, las iniciativas más recientes se centran en desarrollar métodos que puedan interpretar la información bidimensional de una imagen y extrapolar estos puntos, conocidos como *landmarks* faciales, hacia reconstrucciones en el espacio tridimensional. Entre los métodos actuales se incluye el desarrollo de DECA (*Detailed Expression Capture Animation*) [1], EMOCA (*EMOtion Capture and Animation*) [2] o DAD-3DNet [3], el cual será utilizado en este proyecto. Estas nuevas capacidades están tomando terreno en el campo del análisis facial, permitiendo realizar estudios exhaustivos únicamente a partir de imágenes y aumentando la cantidad y calidad de las aplicaciones disponibles.

Las ventajas ofrecidas por la reconstrucción facial están estrechamente relacionadas con el gran número de áreas en las que puede aplicarse. Entre ellas destacan su utilidad en sistemas biométricos, de seguridad y vigilancia, control de interfaces, búsqueda virtual basada en rostros e incluso el análisis médico mediante rasgos faciales [4] [5]. En el ámbito del entretenimiento, se observa su empleo en la animación, la generación de avatares y la creación de modelos faciales realistas para su uso en la postproducción de todo tipo de contenido multimedia, videojuegos o realidad virtual [6].

Sin embargo, a pesar de que se han superado numerosos desafíos relacionados con la detección y reconstrucción facial a partir de imágenes en entornos con cámaras calibradas, aún queda mucho por avanzar en este campo, especialmente en lo que respecta a la robustez de los sistemas desarrollados en entornos menos controlados [7]. Esto se debe a la considerable variabilidad presente en las imágenes faciales. Concretamente, factores

1.1. MOTIVACIÓN

como la iluminación, la posición, las occlusiones o las características propias del individuo pueden ser determinantes para lograr una reconstrucción de calidad que se asemeje lo máximo posible a la realidad.

Asimismo, la calidad de los resultados se ve directamente influenciada por el método de captura utilizado para extraer estos datos. Por ejemplo, algunas aplicaciones con interés en reconstrucciones más precisas pueden utilizar sistemas multicámara de profundidad [8] o la combinación de cámaras con sensores LiDAR [9], incrementando así la resolución de los datos. No obstante, la obtención de estos datos presenta ciertas limitaciones debido a la complejidad y tiempo necesarios para utilizar estas herramientas, lo que dificulta la creación de bases de datos extensas de modelos faciales 3D. Con relación a esto, aunque hoy en día existen conjuntos de datos disponibles, aún poseen ciertas limitaciones. El problema subyace en los sesgos y limitaciones asociados con la creación de los conjuntos, como la falta de variabilidad en términos de edad, género o las características de las escenas analizadas. Este es el caso de la base de datos CMU Multi-PIE [10]. Ciertamente, esto también influye en la evaluación final de los modelos y los niveles de precisión alcanzados. A pesar de los desafíos asociados a esta tarea, se han logrado avances significativos en el desarrollo de modelos bajo distintas estrategias. Se puede destacar por ejemplo la evolución de los [Modelos 3D Deformables \(3DMMs\)](#) [11] en los últimos 20 años y el desarrollo de las redes neuronales profundas que han mejorado el rendimiento de las aplicaciones en términos de precisión [12].

Por esta razón, dada la facilidad de adquisición de imágenes 2D y a la mayor precisión visual proporcionada por los modelos faciales 3D, se puede observar una creciente atención en la traducción de 2D a 3D, lo que ofrece la oportunidad de aumentar la disponibilidad de modelos tridimensionales [4]. En particular, el escenario abordado en este proyecto consta del uso de secuencias de vídeo tomadas desde distintas perspectivas, donde se realiza una reconstrucción completa de la cabeza. De esta manera, se desea anotar la base de datos CMU Panoptic con las posiciones de los puntos 3D de la cabeza y sus reprojeciones.

Por otro lado, la detección de *landmarks* faciales es un paso crucial en las aplicaciones que hacen uso de estos modelos faciales. En los primeros intentos de detección de estos puntos de referencia, se utilizaban modelos estadísticos que estaban fuertemente influenciados por las condiciones del entorno. Sin embargo, gracias a la introducción de las redes neuronales, se han podido superar algunos de estos problemas, logrando la robustez necesaria frente a occlusiones faciales o cambios en la iluminación, y garantizando un rendimiento adecuado incluso en imágenes tomadas en entornos no controlados [13]. Por tanto, la ocultación de ciertas partes en las imágenes puede influenciar drásticamente en los resultados de detección obtenidos. Con el fin de validar las nuevas anotaciones de reconstrucción 3D, se pretende utilizar estos datos para el entrenamiento de dos detectores de *landmarks* faciales. Durante uno de estos entrenamientos se aprovechará también la información sobre la visibilidad de los puntos en la imagen, comprobando su influencia en los resultados.

En resumen, el objetivo principal del presente Trabajo fin de Máster consiste en implementar un algoritmo para obtener la reconstrucción 3D refinada utilizando información de diferentes vistas de la cámara, pudiendo utilizar las nuevas reprojeciones de los puntos sobre las imágenes para el entrenamiento de nuevos modelos de análisis facial.

1.2. Planteamiento del problema

Para realizar la reconstrucción 3D del rostro, se emplea el modelo desarrollado en [14], llamado DAD-3DNet, capaz de extraer los puntos característicos de la cabeza de una persona en una imagen. Las imágenes a anotar pertenecen a la base de datos Panoptic [15], que abarca un conjunto de escenas con personas realizando ciertos movimientos e interacciones sociales. Utilizando el modelo DAD-3DNet sobre las imágenes de Panoptic, se predicen los modelos faciales de las distintas vistas de la cámara para cada escena. Estos datos serán refinados realizando una selección de las mejores vistas con el algoritmo **RANdom SAmpled Consensus (RANSAC)** y el método **Bundle Adjustment (BA)** para obtener las anotaciones finales deseadas.

Una vez calculadas estas reconstrucciones refinadas, la segunda parte del proyecto consta de la reproyección de estos puntos 3D sobre las imágenes correspondientes para entrenar dos modelos de detección de *landmarks*. Para este propósito, se realizará un entrenamiento con una pérdida equivalente a todos los puntos y un segundo entrenamiento teniendo en cuenta la visibilidad de dichos puntos según la orientación del sujeto en la imagen.

De esta manera, se pueden entrever varios aspectos clave. En primer lugar, se debe realizar un primer filtrado de las imágenes de los puntos de vista de las escenas de Panoptic para cada fotograma. Para ello, se debe tener en cuenta si la cámara se encuentra activa en la secuencia actual, así como la posición del sujeto con respecto a esta vista y las posibles occlusiones con sus propios gestos o el resto de personas en la escena. De igual forma, para realizar el refinamiento posterior de las reconstrucciones se utilizará **BA** en el procedimiento, por lo que debe ser necesario comprender cómo integrar el funcionamiento de este algoritmo iterativo. Con el propósito de obtener la mejor reconstrucción posible, se estudiará la comparativa de los resultados obtenidos por varios métodos basados en las reconstrucciones por pares y el uso de **RANSAC**. Así se consiguen ordenar las cámaras y no utilizar aquellas cuyas predicciones ofrecidas por DAD-3DNet pueden resultar más erróneas con respecto al resto de vistas, evitando que sean usadas en la reconstrucción refinada.

En cuanto al entrenamiento del detector de *landmarks* faciales, se usará un modelo sencillo para dicho propósito. Como entrada a esta red neuronal, se emplean también las escenas de Panoptic junto con los puntos característicos 2D reprojectados de las nuevas anotaciones. En este caso, se desea analizar la influencia en el entrenamiento al tener en cuenta la visibilidad de los puntos en comparación con no utilizar esta información. Con este fin, se aplican métodos trigonométricos que ofrecen la condición de visibilidad de los *landmarks* según la orientación de la cabeza del sujeto en la imagen.

Es necesario concretar que este proyecto tiene la intención de presentar un método de anotación de reconstrucciones faciales en 3D sobre la base de datos de Panoptic, de manera que los datos puedan ser utilizados en el desarrollo de proyectos más complejos en un futuro. Por esta razón, se implementa también esta aplicación de detección de *landmarks* a modo de estudio sobre la utilidad de las nuevas anotaciones en un caso real. Sin embargo, cabe decir que la calidad de los datos depende inherentemente de las secuencias utilizadas y el correcto funcionamiento del modelo DAD-3DNet.

1.3. Objetivos

El principal objetivo de este proyecto trata de utilizar las predicciones ofrecidas por el modelo DAD-3DNet sobre las secuencias de Panoptic para obtener las anotaciones faciales correspondientes. Este procedimiento busca entonces mejorar las detecciones de las mallas 3D de la cabeza ofrecidas por DAD-3DNet combinando las predicciones de las distintas vistas de las cámaras. Como resultado se obtienen las anotaciones de las reconstrucciones y las reprojeciones de estos puntos sobre las imágenes, creando un nuevo conjunto de datos faciales.

Con el uso de los datos generados, se propone un segundo objetivo orientado a validar las anotaciones mediante el entrenamiento de dos modelos de *deep learning* para el análisis facial, en concreto, para la detección de *landmarks* faciales. En este proceso, se añade también información sobre la visibilidad de los puntos en las imágenes.

Por consiguiente, para lograr los objetivos establecidos, es fundamental el análisis de la literatura existente para entender los enfoques actuales y abordar el problema de manera eficiente. Además, se deberá comprender el funcionamiento del modelo DAD-3DNet para poder aplicarlo de forma correcta sobre las imágenes de Panoptic. Por otra parte, se estudiarán varios métodos para la selección de las mejores cámaras, siendo un paso clave en el desarrollo del proyecto debido a su dependencia directa con la calidad de los datos finales. Además, se establecerá una evaluación basada en la comparación con las anotaciones de Panoptic y un *ground-truth* creado manualmente, en conjunto con una visualización texturizada 3D. Todos estos procedimientos para generar las reconstrucciones refinadas serán integrados en un único proceso, capaz de realizar la anotación completa de la base de datos, de la cual se utilizará una parte para los entrenamientos finales.

De forma más específica, se exponen las tareas asociadas al cumplimiento de los objetivos planteados:

- Estudio de la literatura existente en relación las bases de datos de modelos faciales 3D y la detección y reconstrucción de los puntos característicos del rostro.
- Desarrollo de un método para la selección automática de las cámaras de cada fotograma en las secuencias de Panoptic, realizando un primer filtrado sobre las posiciones de la cara y evitando aquellas imágenes con occlusiones totales.
- Aplicación del modelo DAD-3DNet para la detección de las mallas 3D faciales sobre las imágenes seleccionadas para cada instante de tiempo en una escena.
- Construcción del modelo de la cabeza tridimensional mediante la inicialización con una reconstrucción triangular lineal y el refinamiento con [BA](#).
- Estudio de diferentes algoritmos basados en comparaciones por pares y [RANSAC](#) para la ordenación y selección de las cámaras con mejores predicciones para cada fotograma según el error de reprojeción.
- Anotación manual de las escenas seleccionadas para la evaluación de resultados.
- Evaluación de las reconstrucciones calculadas por los algoritmos según su comparación con las anotaciones de Panoptic y las anotaciones manuales.

- Desarrollo del algoritmo completo con los pasos anteriores para la anotación completa de la base de datos.
- Representación de las reconstrucciones finales realizadas mediante la texturización de los modelos.
- Extracción de puntos visibles y ocultos de las reconstrucciones 3D refinadas según la posición de la cabeza en las imágenes.
- Entrenamiento de dos modelos de detección de *landmarks* faciales utilizando las reproyecciones de los puntos refinados, utilizando una pérdida equivalente y distinta según la visibilidad calculada anteriormente.
- Evaluación de las detecciones obtenidas para ambos modelos frente a las predicciones de DAD-3DNet y las anotaciones.

En resumen, la consecución de los objetivos requerirá un enfoque integral que abarque desde el estudio del estado del arte hasta el desarrollo y la evaluación de los algoritmos propuestos, garantizando así la efectividad y la calidad de los resultados obtenidos.

1.4. Alcance y limitaciones

La detección de *landmarks* faciales y su reconstrucción en un modelo 3D puede ofrecer grandes ventajas en su extensión a otras aplicaciones. Debido a la propia naturaleza de la red neuronal empleada (DAD-3DNet), los puntos característicos utilizados se limitan a 5023, a partir de los cuales se extraen la malla completa de la cabeza.

A pesar de que las imágenes utilizadas se han obtenido en un entorno controlado y el sistema multicámera se encuentra calibrado y fijo, existen aún variaciones que pueden afectar a la calidad de las reconstrucciones. Entre estos, es significativo el problema de las occlusiones en las secuencias de Panoptic, por lo que se debe realizar un filtrado robusto para evitar utilizar estas imágenes en la reconstrucción final. Además, la variabilidad tanto étnica como numérica de la cantidad de personas que participan en los vídeos es un factor determinante a la hora de realizar las detecciones.

Un aspecto que limita el alcance del proyecto es el tiempo necesario de procesamiento para poder realizar la reconstrucción refinada de cada una de las escenas. Cada secuencia de vídeos está formada aproximadamente por 20000 fotogramas, y cada escena ha sido grabada por un conjunto de 31 cámaras en HD. Para realizar la reconstrucción en un instante determinado, se deben primero filtrar las imágenes de las cámaras donde se identifique la cara del sujeto correctamente y recortar esta zona para poder realizar la predicción de los *landmarks*. Después, se debe realizar la reconstrucción lineal inicial con la [Transformación Lineal Directa \(DLT\)](#) y aplicar [BA](#) durante una serie de iteraciones para poder ordenar y seleccionar las cámaras a usar en la reconstrucción final según el algoritmo de comparación desarrollado. Por tanto, el tiempo necesario para realizar la reconstrucción de un único fotograma con múltiples vistas es elevado (40 segundos), lo que conlleva no poder reanotar todo CMU Panoptic en el tiempo de desarrollo del Trabajo Fin de Máster.

Asimismo, en adición a los tiempos inherentes a la programación y la ejecución de cada algoritmo, se debe añadir el costo necesario para realizar la anotación manual para la evaluación de las reconstrucciones con el error de reproyección en diferentes vistas.

Tras obtener estas reconstrucciones, se calculan también sus reproyecciones sobre las imágenes originales, siendo estos puntos posteriormente clasificados según su visibilidad para el entrenamiento del nuevo modelo. En total, todos estos pasos acaban resultando en un proceso costoso computacionalmente.

Como se puede ver, el cumplimiento de las especificaciones del proyecto depende de tres grandes factores: la calidad del modelo utilizado en la detección de *landmarks*, el tiempo para realizar las tareas y la capacidad de cómputo disponible. Además, intervienen otras variables como el número de secuencias analizadas, las cámaras utilizadas en cada reconstrucción, el procesamiento de fotogramas limitados o la optimización de los algoritmos.

1.5. Estructura de la memoria

A modo de esquema general, se muestra un resumen de los capítulos destacando sus puntos principales:

Capítulo 1 ([Introducción](#)): muestra la introducción del trabajo, incluyendo la motivación, el planteamiento del problema, los objetivos a lograr y los alcances y limitaciones existentes.

Capítulo 2 ([Estado del arte](#)): ofrece información sobre las bases de datos de modelos faciales 3D disponibles y los modelos que realizan estas reconstrucciones mediante distintos métodos.

Capítulo 3 ([Metodología](#)): incluye explicaciones sobre la base datos y la red neuronal utilizada, así como las especificaciones sobre los algoritmos y programas empleados en las reconstrucciones, entrenamientos y evaluaciones.

Capítulo 4 ([Desarrollo](#)): expone información sobre el procesamiento de las imágenes, la programación de los algoritmos para la ordenación y selección de cámaras y la reconstrucción de las escenas. Además, contiene el desarrollo para la extracción de visibilidad de los *landmarks* y el entrenamiento de los nuevos modelos.

Capítulo 5 ([Resultados experimentales](#)): muestra los resultados obtenidos por los algoritmos de ordenación y reconstrucción de las escenas, la texturización de las reconstrucciones finales refinadas y la evaluación del modelo entrenado.

Capítulo 6 ([Conclusiones](#)): ofrece una conclusión sobre los resultados obtenidos, enfocando de forma constructiva los retos afrontados y dando una visión final sobre la calidad del estudio.

Capítulo 7 ([Trabajo futuro](#)): se plantean algunos de los pasos futuros a fin de mejorar el proyecto.

Capítulo 2

Estado del arte

En esta sección se proporciona información sobre algunas de las bases de datos de reconstrucciones faciales 3D más destacadas, detallando sus características principales. Además, se examinan ciertos estudios centrados en el desarrollo de sistemas para reconstrucción de modelos faciales 3D.

2.1. Bases de datos de reconstrucciones faciales 3D

El creciente uso de métodos de aprendizaje profundo en la reconstrucción de la geometría facial ha impulsado el desarrollo de bases de datos más extensas y detalladas [16]. Esta tendencia se refleja en la literatura, donde algunos de los primeros conjuntos de datos desarrollados contenían una cantidad baja de reconstrucciones 3D, con pocas variaciones entre sujetos y poses, mientras que en las nuevas bases de datos ya hay datos suficientes para poder ser utilizadas por redes neuronales profundas. A continuación, se describen algunas de bases de datos más relevantes a fin de comprender su alcance y características:

- **FG3D** [16]: este conjunto de datos se creó fusionando los datos RGB-D de tres *datasets* gracias al uso de texturas con un **3DMM** y modelos de iluminación, que permitieron registrar con precisión los rostros. Los conjuntos de datos utilizados incluyen el FRGC, que contiene casi 5000 ejemplos de escaneos faciales 3D, el BP4D extrayendo 3376 *frames* de 328 vídeos 2D y 3D de 41 sujetos y el CASIA-3D, formado por 4624 escaneos de 123 personas, donde se utilizaron únicamente las poses frontales. Estos tres *datasets* componen FG3D con más de 200000 muestras. Asimismo, se utilizó la base de datos de Florence para realizar una la validación cruzada.
- **DAD-3DHeads** [3]: se trata del *dataset* a partir del cual se ha construido el modelo que se utilizará en el proyecto. Es una de las bases de datos 3D densas para reconstrucción de estructuras faciales más grandes hasta la fecha y proporciona 3500 *landmarks* verificados. Los rostros presentan una gran cantidad de orientaciones diferentes, casos de occlusiones, expresiones o cambios de iluminación. Las anotaciones manuales se validaron comparándolas con la información de un escáner 3D (ver Figura 1).
- **BU-3DFE** [17] : este conjunto de datos está formado por 100 sujetos de entre 18 y 70 años, donde el 56 % son mujeres. Los participantes representan seis variedades

2.1. BASES DE DATOS DE RECONSTRUCCIONES FACIALES 3D

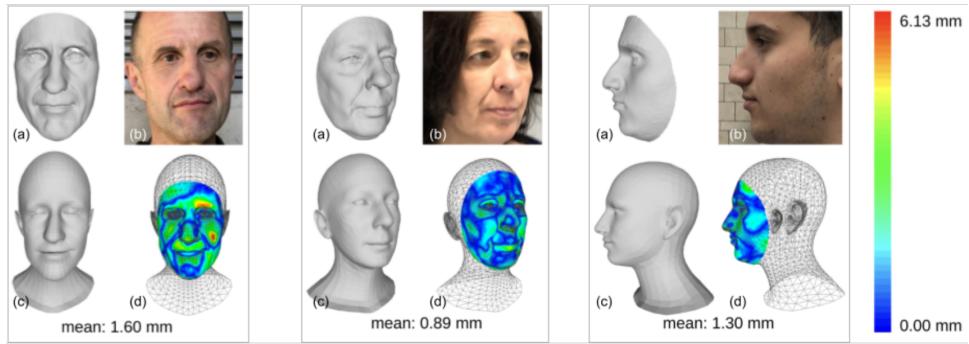


Figura 1: Error en mm de la reconstrucción 3D generada mediante anotaciones 2D [3].

éticas diferentes y cada uno de ellos muestra distintos grados de las expresiones protípicas: felicidad, asco, miedo, enfado, sorpresa y tristeza. Además de esto, se añaden dos vistas distintas, obteniendo 5000 imágenes y 2500 reconstrucciones 3D de cada expresión. Este *dataset* fue ampliado posteriormente con BU-4DFE (3D + time) [18], donde se incluyen secuencias de vídeo de 25 fps, conteniendo 100 *frames* por expresión. Participaron 101 sujetos con distinta procedencia y género y cada modelo 3D está formado por 35000 vértices. Posteriormente, se añadió la base de datos BP4D-Spontaneous [19], que tiene vídeos de diferentes tareas donde se realizan las expresiones de forma más espontánea. En este caso, 41 personas participaron en las secuencias.

- **FaceScape** [20]: en la construcción de este conjunto, se utilizaron 68 cámaras densas con iluminación controlada, consiguiendo definiciones detalladas del rostro. Para la recopilación de datos, participaron 938 personas entre 16 y 70 años, realizando 20 expresiones diferentes, constituyendo un total de 18760 modelos faciales 3D de gran calidad. En la Figura 2 se observa el procedimiento realizado para la captura y transformación de los datos a modelos 3D.

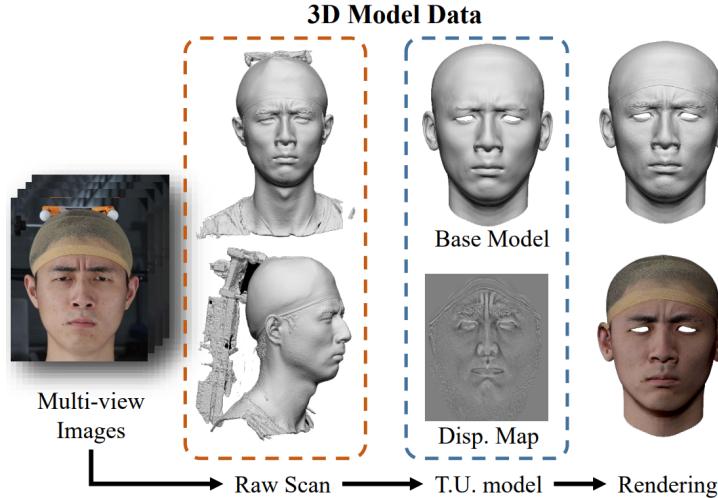


Figura 2: Procedimiento de captura de imágenes a modelos uniformes [20].

- **3DFAW-Video** [21]: Este conjunto de datos está formado por secuencias de vídeo con 66 sujetos diferentes, de entre 18 y 28 años. Las grabaciones se realizaron con vídeos de iPhone, mientras que para la generación del *ground-truth* de las mallas

3D se usaron sistemas de alta resolución 3D con pares estéreo, obteniendo reconstrucciones densas para cada *frame*.

En definitiva, se evidencia el interés en la construcción de bases de datos más detalladas y amplias, poniendo el enfoque en conseguir mejores modelos a través del incremento de los datos disponibles. Los datos de referencia para las reconstrucciones 3D pueden ser generados mediante anotaciones manuales, escaneos 3D o cámaras RGB-D. Es notable el uso recurrente de conjuntos de datos previos como punto de partida para ampliar o crear nuevos conjuntos con una mayor diversidad y tamaño.

2.2. Reconstrucción 3D del rostro

El campo del análisis, seguimiento, reconocimiento y reconstrucción facial tiene un gran desarrollo en investigación dentro de la visión por computador [6]. Como se comentó anteriormente, esta reconstrucción puede ser utilizada en multitud de aplicaciones, desde controles y seguridad hasta planificación de cirugías estéticas.

Además, aunque el análisis previo 2D de las imágenes puede ser necesario en algunos de estos procedimientos, el uso de información de profundidad resulta realmente positivo a la hora de obtener sistemas robustos y de calidad. Sin embargo, como se señala en [22], la adquisición de esta información puede estar limitada, lo que ha generado un creciente interés en el desarrollo de modelos capaces de predecir una malla facial 3D a partir de una imagen, evitando así la necesidad de usar escáneres en tiempo real. Estos escáneres 3D, aunque son ampliamente usados en otro tipo de aplicaciones, pueden presentar un obstáculo para esta tarea debido a los requisitos para su correcto funcionamiento. Se incluyen, por ejemplo, la necesidad de luz controlada, las limitaciones en su uso automático, sus costos asociados o los problemas potenciales relacionados con el uso de un láser cerca de los ojos.

Según los enfoques empleados en los estudios, [23] clasifica los métodos de reconstrucción 3D de rostros en diferentes grupos, aunque también pueden encontrarse de forma combinada en la literatura. Los métodos son los siguientes:

- Métodos de reconstrucción por ejemplos, donde se utiliza el modelo de una cabeza de referencia que se deforma.
- Reconstrucciones de la geometría facial a partir de la detección de *landmarks* faciales 2D.
- Mediante la utilización técnicas como **Shape-from-Shading (SfS)**.
- Con el uso de métodos estadísticos, destacando el ajuste de parámetros realizado con los **Modelos 3D Deformables (3DMM)**.
- Mediante redes neuronales profundas para la estimación de los detalles faciales 3D. Dentro de esta categoría se incluirían aquellas redes capaces de producir una reconstrucción facial a partir de una o varias imágenes de una persona.

2.2. RECONSTRUCCIÓN 3D DEL ROSTRO

En el caso de reconstrucción por ejemplos, se presenta el trabajo presentado en [24], que es capaz de realizar una estimación 3D de rostros manejando cierto grado de variabilidad en la apariencia facial. Para este propósito se desarrolla un proceso de optimización que busca maximizar la similaridad entre la apariencia y profundidad de la cara a partir de un modelo facial de referencia.

Centrando el análisis en el uso de *landmarks* faciales 2D para su utilización en reconstrucciones 3D, se destaca el estudio realizado en [25] que ha servido de inspiración para este proyecto. Específicamente, su investigación ha contribuido tanto al refinamiento de las reconstrucciones como al empleo de la visibilidad de *landmarks* en el entrenamiento de la red neuronal posterior. En este trabajo, desarrollan una optimización personalizada que se basa en la generación consistente espacial y temporalmente de mallas faciales 3D a partir de los *landmarks* extraídos de secuencias de imágenes en diversas vistas. Con el fin de alcanzar este objetivo, proponen una función de pérdida en tres partes. La primera está basada en el error de los *landmarks* 2D, la segunda en el error temporal de la reconstrucción en la secuencia, y la tercera en la consistencia multivista que considera las imágenes de las diferentes cámaras al mismo tiempo. En concreto, esta última función de pérdida tiene en cuenta la visibilidad de los puntos característicos según la orientación del sujeto en la imagen. Aquellos *landmarks* que no sufren occlusiones reciben una ponderación mayor, lo que aumenta su contribución en la optimización del modelo (ver Figura 3).

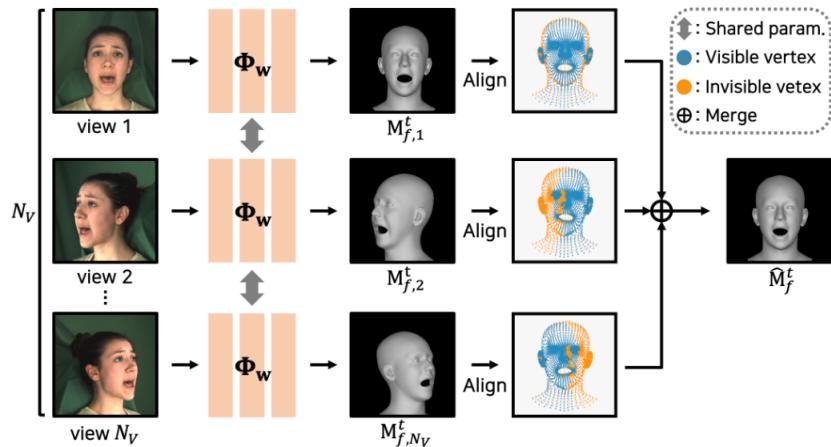


Figura 3: Alineación de los modelos 3D multivista según la confianza de los puntos visibles y ocultos [25]

De entre los métodos expuestos anteriormente, el **SfS**, desarrollado por [26], suele utilizarse generalmente como un refinamiento posterior para una reconstrucción ya realizada. Sin embargo, su uso puede resultar menos conveniente en ciertos casos, como describen los autores en [27], ya que asume una correspondencia directa entre la textura de la imagen y su geometría. A pesar de ello, son ampliamente usados en la literatura, como en los proyectos desarrollados por [28], [29] o incluso en conjunto con otros métodos como el uso de **Modelos Locales Deformables (LMMs)** [30]. De esta manera, se consiguen suprir algunas de sus limitaciones.

A diferencia del **SfS**, los **3DMMs**, introducidos por [11], se basan en modelos paramétricos construidos normalmente mediante el **Análisis de Componentes Principales (PCA)** de una base datos y son capaces de estimar y reducir el error entre la imagen 2D y la reconstrucción. Los **3DMMs** también son recurrentes en la literatura y se tienen numerosos

ejemplos como los desarrollados en los estudios de [31] o [32]. Estos poseen la ventaja de afrontar mejor las condiciones diferentes de iluminación y orientaciones de la cabeza, aunque presentan algunos problemas cuando los *landmarks* de la cara no son completamente visibles.

En el presente Trabajo Fin de Máster, se emplea el modelo DAD-3DNet, propuesto en [3]. Este predice la localización de los *landmarks* faciales y extrae los parámetros de la forma de la cabeza, su expresión y su orientación. Con esta información, consigue generar una reconstrucción completa en tres dimensiones. Su rendimiento se encuentra al nivel del estado del arte y ha sido entrenado teniendo en cuenta una multitud de orientaciones, occlusiones y situaciones. En la Figura 4 se puede visualizar como las mallas 3D generadas, basadas en la detección de *landmarks* faciales, son capaces de ajustarse a las formas de la cabeza de cada persona.



Figura 4: Mallas 3D generadas por DAD-3DNet [3].

Otro proyecto que realiza la generación de mallas faciales 3D a partir de una única imagen es el planteado en [33], que ofrece una solución mediante diferentes estructuras de **Redes Neuronales Generativas Adversarias (GANs)** y compara los resultados con otros métodos del estado del arte: los autoencoders y los algoritmos **SfS**. Esta combinación de métodos también se observa en [34], cuyos resultados se muestran en la Figura 5. En contraste con otros trabajos, el entrenamiento propuesto usa datos sintéticos, consiguiendo un modelo **3DMM** capaz de obtener una reconstrucción válida global sin el uso de *landmarks* intermedios. Finalmente, aplican la técnica **SfS**, logrando perfilar los rasgos más finos del rostro (ver Figura 6).



Figura 5: Reconstrucciones con textura obtenidas en [34].

En el caso de querer realizar reconstrucciones a partir de las imágenes recogidas con varias cámaras calibradas, se pueden utilizar sistemas de pares estéreo, para estimar mapas de disparidad refinados. Estos sistemas multicámara calibrados son realmente útiles si se

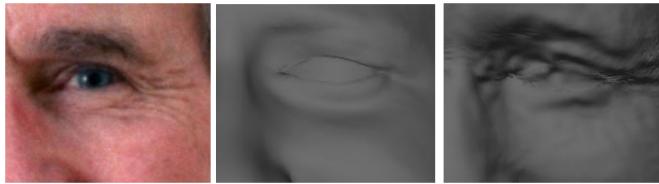


Figura 6: Refinamiento de rasgos utilizando SfS [34].

desea reconstruir una escena desde distintas perspectivas mediante las correspondencias entre las imágenes y la información de profundidad. Con este planteamiento, se presenta el artículo escrito por [29], que emplea de igual forma el refinamiento SfS después de generar las reconstrucciones con las cámaras estéreo. Este método no lo utilizan únicamente con rostros, sino que prueban distintos objetos e iluminaciones y consiguen obtener la estructura geométrica en algunos casos con más detalle que con escáneres láser. En la Figura 7 se observa el detalle del rostro y las reconstrucciones conseguidas con distintos métodos.

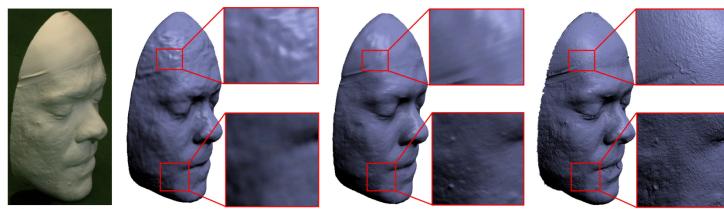


Figura 7: Comparación de reconstrucciones faciales obtenidas con distintos métodos [29].

Asimismo, es destacable el enfoque desarrollado por [7], donde separan las imágenes de una misma persona en grupos según su orientación. La mayoría de estas imágenes, tomadas de Internet, no están calibradas y se utilizan para llevar a cabo la reconstrucción densa en 3D. Además, desarrollan un algoritmo capaz de obtener las vistas laterales de las regiones ocultas de la cara junto con sus medidas y limitaciones de profundidad. En la misma línea se encuentra el proyecto de [35], que utiliza igualmente imágenes de Internet del mismo individuo en poses diferentes y cuya reconstrucción se hace de manera iterativa sobre el modelo 3D, tal como se ve en la Figura 8.

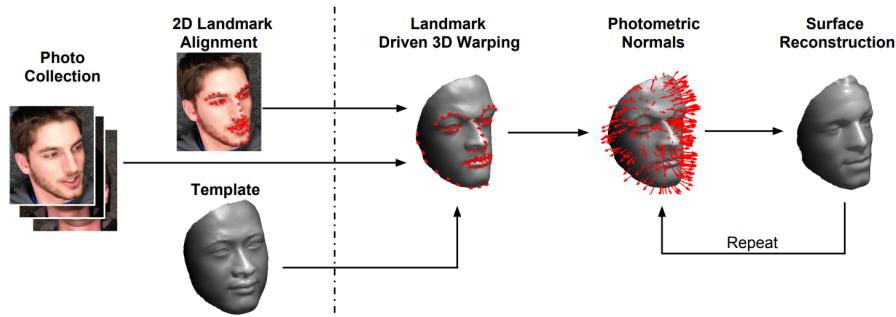


Figura 8: Proceso iterativo para la creación del modelo facial [35].

En resumen, existen multitud de enfoques estudiados con el propósito de obtener reconstrucciones del rostro fieles a la realidad. Se han podido ver algunos de los retos a los que se enfrentan los autores y las técnicas que han sido desarrolladas en consecuencia. Desde

el uso de métodos más tradicionales hasta la incorporación de las actuales redes neuronales, se observa cómo este ámbito continúa en desarrollo y se siguen consiguiendo avances realmente destacables. De esta manera, se está consiguiendo incrementar la calidad de las aplicaciones que utilizan sistemas faciales 3D.

Capítulo 3

Metodología

En las siguientes secciones se realiza un análisis sobre las herramientas y la metodología empleada. Se proporciona una descripción detallada de la base de datos de Panoptic y el modelo DAD-3DNet, así como del proceso de reconstrucción y refinamiento. Además, se presenta el método utilizado para evaluar y visualizar las reconstrucciones. Por último, se expone la estrategia de entrenamiento utilizando las nuevas anotaciones, junto con una descripción de la extracción de visibilidad de los puntos de referencia faciales.

3.1. Análisis de la base de datos Panoptic

Dado que el enfoque principal de este trabajo trata de la anotación del conjunto de escenas de Panoptic, se va a realizar un análisis de sus características. Esta base de datos está públicamente disponible y fue introducida en [15], donde se detalla tanto el desarrollo llevado a cabo para su construcción como las funcionalidades ofrecidas.

La razón principal de su creación reside en la detección visual de la estructura del movimiento de las personas durante ciertas interacciones sin la necesidad de utilizar marcadores. De esta forma, se buscaba construir un gran conjunto de datos visuales para utilizarlo en el análisis de posiciones corporales e interacciones sociales. Gracias al estudio desarrollado en [15], se consiguieron anotar las poses las personas en cada fotograma de forma automática, solventando las occlusiones presentes en las escenas gracias a los diversos puntos de vista. Después de esto, otros proyectos como [36], [37] y [38], han ido enriqueciendo el conjunto, aumentando los análisis y el contenido de los datos.

3.1.1. Estructura multicámaras

Con el propósito de capturar las escenas, se construyó la estructura mostrada en la Figura 9, que abarca un diámetro y altura de 5,49 y 4,15 metros respectivamente. Esta se encuentra formada por paneles pentagonales y hexagonales, donde se añaden un conjunto de 24 cámaras VGA por panel, 31 cámaras HD, 5 proyectores y 10 Kinects con sensores RGB+D en todo el conjunto, constituyendo un total de 480 cámaras. En el caso de este trabajo, se emplearon exclusivamente los videos capturados por las cámaras HD, que tienen una resolución de 1920x1080 y se calibraron utilizando [Structure from Motion \(SfM\)](#). Además, se sincronizaron con un reloj central para asegurar que las grabaciones se realizaran a la misma frecuencia.

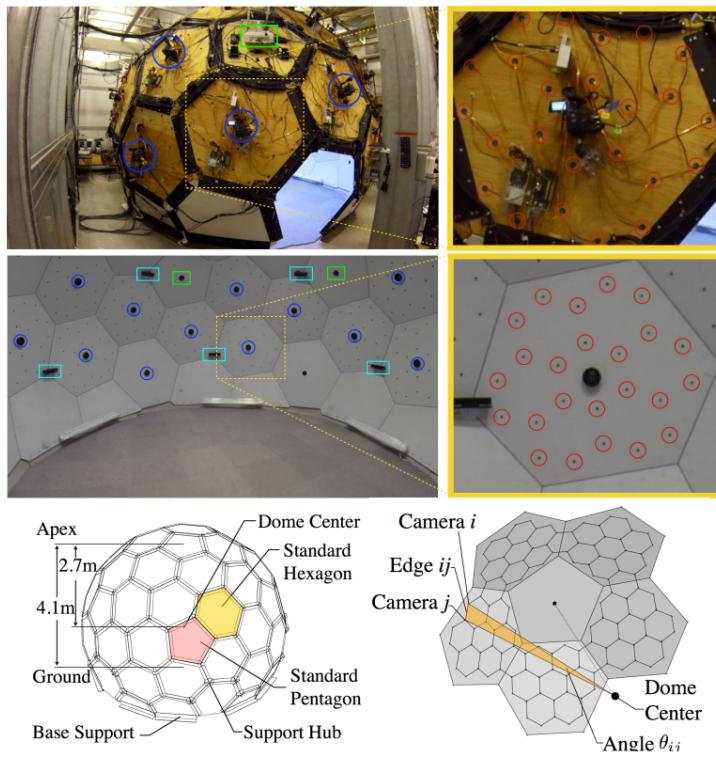


Figura 9: Estructura para la grabación de escenas de Panoptic [15].

3.1.2. Anotaciones disponibles

Panoptic proporciona una gran cantidad de anotaciones para cada imagen captada por las cámaras. En concreto, la base de datos ofrece los puntos característicos de cada persona en cada fotograma, marcando los *landmarks* de la cara y los del cuerpo, así como la posición de las manos y dedos con identificadores en cada caso (ver Figura 10).

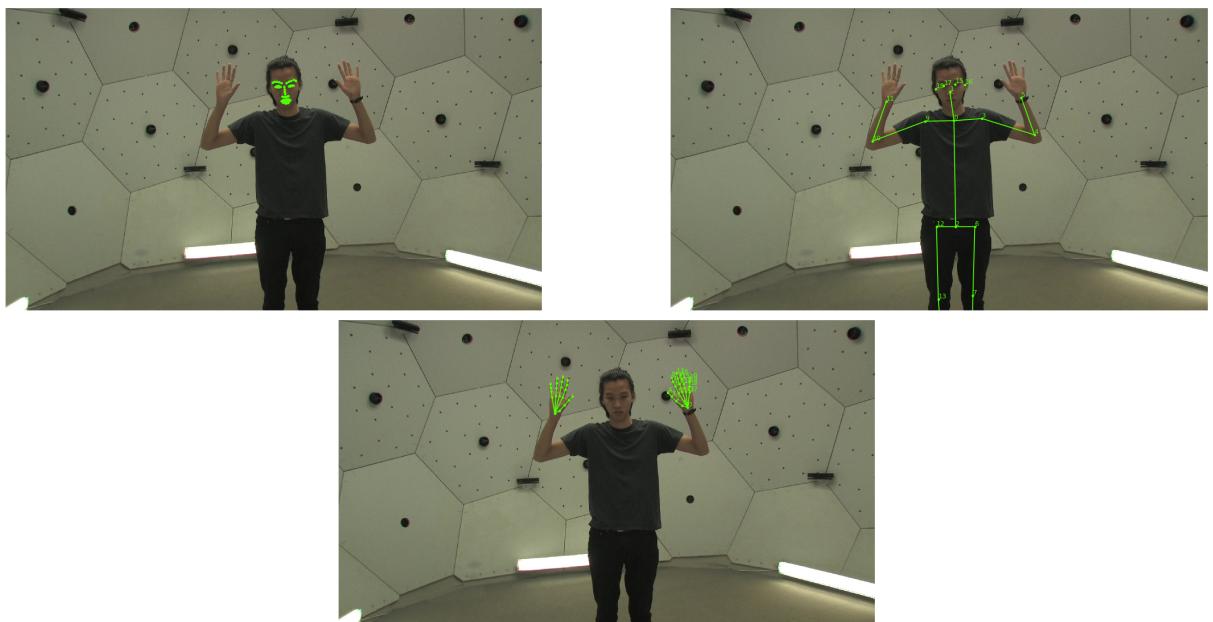


Figura 10: Anotaciones proporcionadas por Panoptic imágenes.

3.1. ANÁLISIS DE LA BASE DE DATOS PANOPTIC

De entre todas las anotaciones, se utilizan los puntos mostrados en la Figura 11 para el preprocesamiento de las imágenes. Estos sirven como referencia para evitar las situaciones de occlusiones entre la misma persona y con otras. En blanco se muestran los puntos de las manos, en verde los puntos que serán usados para detectar la posición del sujeto y otras occlusiones y en rosa el punto correspondiente a la posición de la boca.



Figura 11: Puntos seleccionados de Panoptic utilizados en el preprocesamiento.

Asimismo, se seleccionan 27 puntos característicos de la cara (Figura 12), que serán utilizados en el desarrollo del algoritmo de reconstrucción y la posterior evaluación de los resultados.

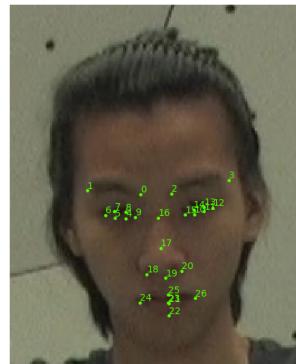


Figura 12: Puntos seleccionados de Panoptic utilizados en la evaluación.

Sin embargo, cabe tener en cuenta que algunas de estas anotaciones pueden no ser del todo correctas. Por ejemplo, en la Figura 13 se muestran algunos casos que comparan los puntos de Panoptic (en rojo) con anotaciones realizadas manualmente para este Trabajo Fin de Máster (en blanco).

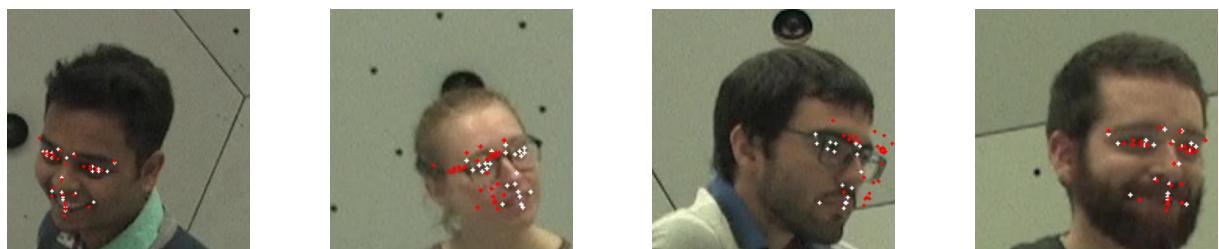


Figura 13: Comparación entre los puntos anotados por Panoptic con los anotados manualmente.

Se observa cómo difieren las posiciones de los *landmarks* en ciertas imágenes de forma

notable, por lo que este posible error debe ser considerado, ya que afectará a la calidad final de los resultados.

3.2. Modelo DAD-3DNet

La red neuronal DAD-3DNet[14] será la empleada para realizar las predicciones de los *landmarks* faciales que serán utilizadas para las nuevas reconstrucciones 3D. Este fue entrenado con la base de datos DAD-3DHeads utilizando imágenes de tamaño 256x256 píxeles. La arquitectura se muestra en la Figura 14 y está formada por los elementos que se describen a continuación:

- Un encoder convolucional para obtener las características de la imagen.
- Un estimador de *landmarks* faciales con mapas de calor.
- Un módulo para fusionar las dos anteriores.
- Un módulo de regresión para refinar la posición de los *landmarks* y los parámetros de FLAME [39], un modelo 3DMM del rostro.
- Una capa diferencial de FLAME [39] que genera los puntos 3D de la cara a partir de los parámetros del 3DMM estimado.

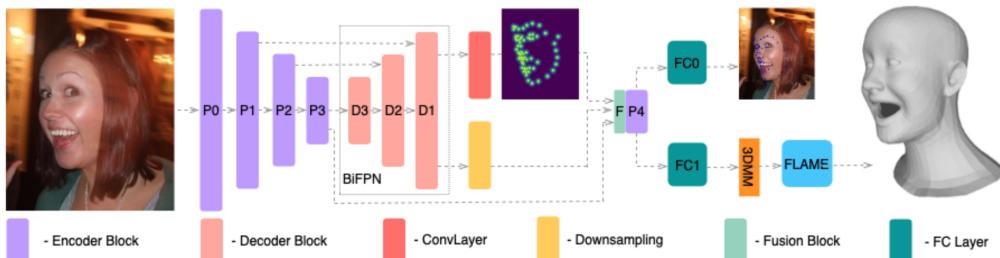


Figura 14: Estructura del modelo DAD-3DNet [14].

Este modelo, implementado en PyTorch e inicializado con los pesos de ImageNet, utiliza una función objetivo que se divide en varias partes: la función de pérdida de la forma y expresión, el error de reproyección y la pérdida gaussiana del mapa de calor. Estos errores calculados se encargan de ajustar la geometría de la cabeza, su orientación y la posición de los *landmarks*. Por otro lado, DAD-3DNet se evaluó calculando el error de reproyección de la malla 3D obtenida con FLAME en cada imagen con el fin de validar la posición de algunos *landmarks* de referencia y la orientación del sujeto predicha. Además, se compararon los resultados obtenidos con otros métodos del estado del arte y se evaluaron frente a situaciones específicas.

Tras la aplicación de DAD-3DNet sobre una imagen, se pueden obtener las predicciones mostradas en la Figura 15. En este proyecto, se emplearán los 5023 *landmarks* predichos. Además, se aprovechará la estimación de la orientación de la cabeza mostrada en la tercera imagen de la segunda fila, con el propósito de identificar y extraer las occlusiones de los *landmarks* en la imagen. Este elemento será fundamental en el proceso de entrenamiento de los nuevos modelos que utilizarán las anotaciones refinadas.

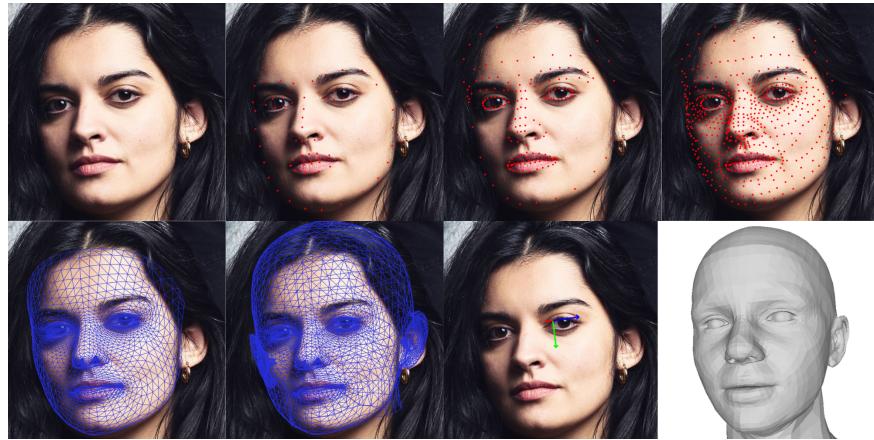


Figura 15: Predicciones ofrecidas por DAD-3DNet.

3.3. Parámetros de la cámara

La calibración de las cámaras es indispensable para los proyectos que incluyen transformaciones de coordenadas del espacio 2D al 3D y viceversa. En concreto, Panoptic ofrece los parámetros intrínsecos y extrínsecos de cada una de las 31 cámaras HD. Estos parámetros se presentan en forma de matrices, como se muestra en la Ecuación 1:

$$K = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}, \quad t = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \quad (1)$$

donde K es la matriz de parámetros intrínsecos y R y t los extrínsecos que representan la orientación y la traslación de la cámara con respecto al sistema de coordenadas de la escena. Estas matrices pueden combinarse a su vez para obtener la matriz de proyección de cada cámara como en la Ecuación 2.

$$P = K [R \mid t] \quad (2)$$

Esta matriz de tamaño 3x4 es útil tanto para el cálculo de la reconstrucción lineal como para obtener las reprojeciones de los puntos 3D de la reconstrucción sobre las imágenes.

3.4. Método de reconstrucción

En este apartado, se va a exponer el método de reconstrucción desarrollado. DAD-3DNet proporciona la proyección de 5023 puntos 3D sobre las imágenes de cada una de las 31 cámaras. Estas proyecciones se calculan imagen a imagen y son las que emplearemos para realizar la reconstrucción 3D de la cabeza. Sin embargo, las proyecciones a veces no son tan buenas que se desearía (ver Figura 25), por lo que hay que determinar en qué cámaras no hay errores para usarlas en la reconstrucción. Para realizar este proceso, se desarrollan varios algoritmos para la comparación entre predicciones según el error de las reconstrucciones y la utilización de [RANSAC](#). Estos métodos se encargan de encontrar las

cámaras donde la reconstrucción se proyecta donde se encuentran los *landmarks* anotados y seleccionar aquellas que serán utilizadas para obtener las coordenadas 3D finales.

Para la realización de cada reconstrucción se necesita una primera inicialización lineal y un posterior refinamiento con [Bundle Adjustment \(BA\)](#). Estos procedimientos serán explicados a continuación, así como los algoritmos desarrollados para la ordenación y selección de las mejores predicciones.

3.4.1. Reconstrucción triangular lineal

El algoritmo de reconstrucción triangular lineal está basado en la [DLT](#) (ver Ecuación 4) y se utiliza para obtener una primera aproximación de la reconstrucción 3D de la cabeza a partir de dos vistas. Sin embargo, este procedimiento puede ampliarse para más conjuntos. Siguiendo la explicación expuesta por Zisserman en [40], para cada imagen se tienen las correspondencias $u_i = P_i M$ y $u_j = P_j M$, que representan la transformación de un punto 3D, M , a las coordenadas cartesianas de cada imagen, siendo i y j dos cámaras diferentes y u_i y u_j las proyecciones correspondientes. Esto se muestra de manera más clara en la Ecuación 3.

$$\lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = P \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (3)$$

Donde:

- x y y serían las cartesianas de la proyección de M sobre la imagen.
- λ es el factor de escala.
- P es la matriz de proyección de la cámara,
- $M = [X, Y, Z, 1]^\top$ son las coordenadas homogéneas en el espacio tridimensional.

Para obtener el sistema de ecuaciones $AM = 0$, se calculan primero los productos cruzados $u \times PM = 0$ en la Ecuación 4, donde las p_i^\top representan las filas de la matriz de proyección.

$$\begin{aligned} x(p_3^\top M) - (p_1^\top M) &= 0 \\ y(p_3^\top M) - (p_2^\top M) &= 0 \\ x(p_3^\top M) - y(p_1^\top M) &= 0 \end{aligned} \quad (4)$$

De las tres ecuaciones, hay dos linealmente independientes y se puede construir la matriz del sistema, A , como en la Ecuación 5.

$$A = \begin{bmatrix} xp_{3,i}^\top - p_{1,i}^\top \\ yp_{3,i}^\top - p_{2,i}^\top \\ x'p_{3,j}^\top - p_{1,j}^\top \\ y'p_{3,j}^\top - p_{2,j}^\top \end{bmatrix} \quad (5)$$

3.4. MÉTODO DE RECONSTRUCCIÓN

Por tanto, para cada punto a reconstruir en una imagen y su determinada correspondencia en la otra, se construye la matriz A y se realiza la [Descomposición en Valores Singulares \(SVD\)](#), $A = UDV^T$, obteniendo el valor de la coordenada 3D en la última columna de V , como el mínimo valor singular. Al realizar esto para todas las correspondencias de puntos, se puede obtener la reconstrucción deseada de la cabeza del sujeto.

3.4.2. Refinamiento con *Bundle Adjustment*

El algoritmo de [BA](#) se emplea normalmente para refinar una reconstrucción en sucesiones iterativas que minimicen el error entre la reprojeción de los puntos 3D y los *landmarks* originales de cada imagen. Dada una primera reconstrucción 3D inicial como la obtenida con [DLT](#), este método permite recalcular tanto las matrices de proyección de las cámaras como los puntos 3D y reducir este error en consecuencia. Siguiendo la descripción del artículo [41], en la Figura 16, se representa el [BA](#), donde se muestran dos cámaras y la reconstrucción de 5 puntos. Los puntos u_{ij} corresponden a los *landmarks* originales detectados en cada imagen, donde i es el *landmark* actual y j la cámara. Por el contrario, u'_{ij} representa las reprojeciones de los puntos 3D, mostrados como M_i , de la reconstrucción en las imágenes.

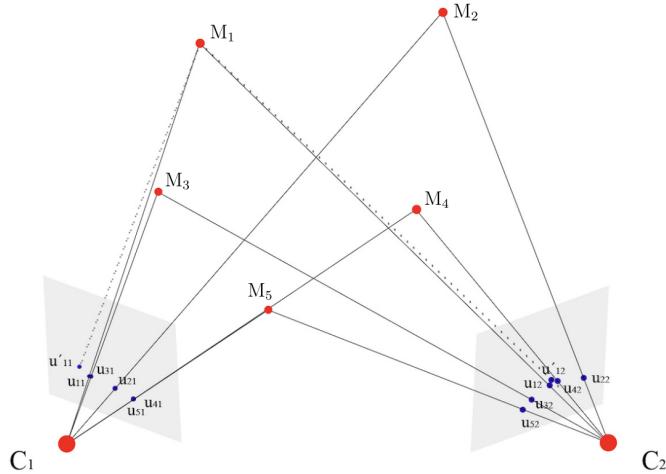


Figura 16: Representación visual del [BA](#)

Por tanto, la minimización se realiza sobre la diferencia entre los puntos clave y sus reprojeciones. En la Ecuación 6 se observa el cálculo del error de reprojeción donde la n sería el número de puntos, m el número de cámaras y θ los parámetros a ajustar.

$$S = \sum_{j=1}^m \sum_{i=1}^n \sqrt{(\vec{u}_{ij} - \vec{u}'_{ij}(\theta))^2} \quad (6)$$

Una gran ventaja de este algoritmo es que puede lidiar con la falta de datos en algunas imágenes. Además, su ejecución se puede plantear de tres formas distintas según las especificaciones de la reconstrucción:

- ***Bundle Adjustment Motion Structure***: considera que tanto la calibración como la primera inicialización de la reconstrucción pueden refinarse, por lo que se proponen nuevos parámetros y posiciones de los puntos 3D para la minimización del error.

- **Bunde Adjustment Motion:** este modo es útil cuando se establecen los puntos 3D como correctos y se desea estimar la calibración de las cámaras a partir de estos, de forma que se minimice el error con la escena. Debe haber una calibración inicial de las cámaras.
- **Bunde Adjustment Structure:** este es el caso usado en el TFM. La calibración de las cámaras se asume como correcta, por lo que no se modifica durante la optimización. El error se encuentra únicamente en los puntos 3D y se refinan las coordenadas con el fin de ajustarse al resto de imágenes y las proyecciones de sus *landmarks*.

Para poder realizar este refinamiento de los puntos 3D con **BA**, se ha utilizado el paquete CvSBA¹, que ofrece un recubrimiento software en python para el código desarrollado en C++ por [42]. Este utiliza el algoritmo de optimización **Levenberg-Marquardt (LM)** y ofrece las funcionalidades descritas anteriormente. Para su ejecución, se selecciona el *Bundle Adjustment Structure* y se proporcionan los datos de los parámetros intrínsecos y extrínsecos de la cámara, la reconstrucción inicial y las correspondencias de los puntos clave entre las cámaras. Al finalizar la ejecución para un conjunto, se obtienen los nuevos datos 3D refinados además de información sobre la progresión de las iteraciones y el error de minimización.

3.4.3. Ordenación y selección de cámaras

A partir de ahora hablaremos de la *reconstrucción en un fotograma o en un instante* para referirnos a la reconstrucción 3D obtenida a partir de procesar las imágenes provenientes de las cámaras HD. La reconstrucción de un rostro en un instante dado se realiza mediante las predicciones generadas por el modelo DAD-3DNet para cada vista, aplicando el procedimiento lineal y el refinamiento expuesto en los apartados anteriores. En este proceso, la selección adecuada de cámaras desempeña un papel crucial, ya que determina la calidad y precisión de la reconstrucción. Sin embargo, dado que el modelo DAD-3DNet realiza la predicción utilizando una única imagen, puede haber situaciones en las que las predicciones contengan errores significativos.

Por ende, el algoritmo propuesto para realizar la reconstrucción final debe ser capaz de organizar las cámaras para evitar aquellas vistas donde el error sea más pronunciado. El funcionamiento de este algoritmo se resume en dos objetivos:

1. Conseguir una aproximación a la mejor ordenación de las cámaras. Es decir, ordenar las cámaras de manera que aquellas que producen un menor error en la reconstrucción estarán al principio de la lista.
2. Seleccionar el número de cámaras con el cual el error de reconstrucción es menor.

Para afrontar este reto, se han desarrollado tres posibles estrategias algorítmicas y se ha llevado a cabo un estudio sobre el rendimiento de cada una para seleccionar la más adecuada para la tarea. El estudio realizado se centra en comprobar el impacto de la ordenación de las cámaras en la reconstrucción frente a una selección completamente aleatoria del conjunto de cámaras. Los algoritmos desarrollados son los siguientes:

¹<https://www.uco.es/investiga/grupos/ava/portfolio/cvsba/>

- **Algoritmo por pares con Panoptic:** evaluando las reconstrucciones por pares de cámaras frente a las anotaciones de Panoptic.
- **Algoritmo RANSAC con Panoptic:** realizando la comparación de las reconstrucciones de pares de cámaras frente a Panoptic mediante la implementación de **RANSAC**. De esta manera, no se evalúan todos los pares de cámaras, por lo que el tiempo de ejecución se reduce cuando el número de predicciones a comparar sea mayor de 7.
- **Algoritmo RANSAC con predicciones de DAD-3DNet:** sigue el mismo desarrollo que el caso anterior pero evaluando el error frente a las predicciones originales de DAD-3DNet.
- **Algoritmo aleatorio:** mediante la ordenación aleatoria de las cámaras.

El desarrollo y funcionamiento de estos métodos será descrito con más detalle en la sección [4.2.3](#). Asimismo, el método de evaluación propuesto para seleccionar el mejor algoritmo se exponen en la sección [3.5](#) y los resultados se verán en la sección [5.1](#).

3.5. Método de evaluación de las reconstrucciones

En esta sección, se presenta el método empleado para evaluar las reconstrucciones obtenidas en Panoptic. Esta evaluación resulta esencial para determinar la precisión y eficacia de los algoritmos de reconstrucción facial utilizados en este estudio. El proceso se divide en varias etapas que abarcan desde la selección de las escenas a evaluar hasta el cálculo de errores de reproyección para cada punto característico. A continuación, se detallan las fases clave de este método de evaluación.

3.5.1. Selección de escenas de evaluación

Con el objetivo de asegurar un análisis exhaustivo y diversificado, se seleccionaron 13 instantes de tiempo con sus 31 vistas de cada cámara provenientes de 6 vídeos distintos de la base de datos de Panoptic. Este conjunto de evaluación abarca un total de 15 individuos diferentes, siendo elegidas específicamente las vistas sin occlusiones mediante el filtrado que se detallará en la sección [4.1](#). En total, se anotaron manualmente un conjunto de 183 imágenes para llevar a cabo esta tarea.

Las escenas seleccionadas se muestran en la Figura [17](#) en la vista correspondiente a la cámara *00_00*. Se observa la variación de poses, características y tipos de occlusiones parciales presentes, como por ejemplo el uso de gafas. Se debe tener en cuenta que de algunos de estos fotogramas se utiliza la información de la posición de dos personas distintas.

En la Tabla [1](#) se añade información sobre cada una de las escenas y su identificador, que será utilizado posteriormente para hacer referencia a cada una. Igualmente, se muestra el fotograma seleccionado de cada secuencia, indicando el número del sujeto utilizado y el total de vistas filtradas en las que se puede apreciar claramente la cara de la persona.

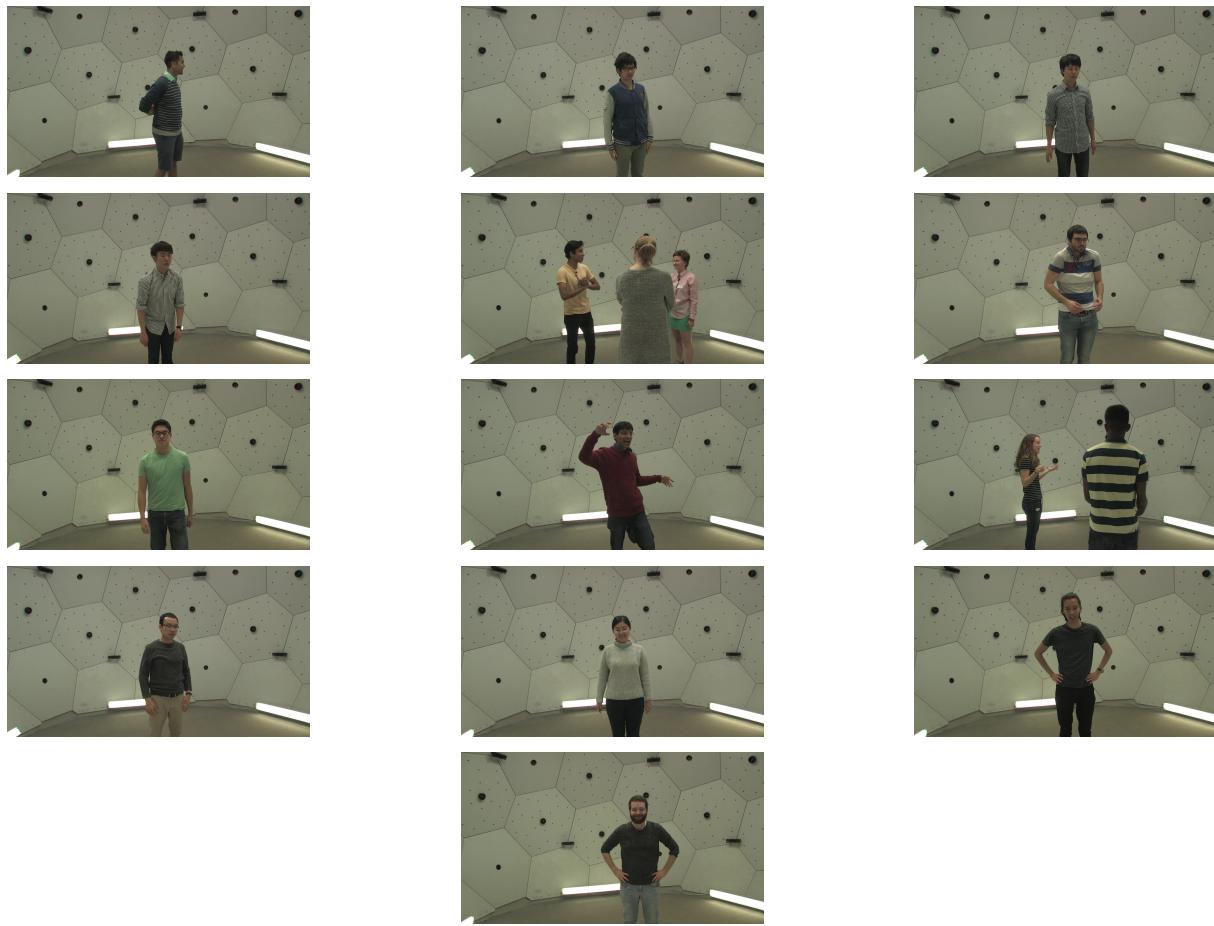


Figura 17: Escenas de evaluación seleccionadas.

Tabla 1: Especificaciones de las escenas de evaluación seleccionadas.

Identificador	Escena	Frame	Sujeto	Número de vistas
1	171204_pose5	153	0	13
2	171026_pose1	2008	0	14
3	171026_pose2	2595	0	14
4	171026_pose3	2616	0	12
5	170404_haggling_b3	3270	3	15
6	170404_haggling_b3	3270	4	10
7	171026_pose2	8799	1	10
8	171204_pose5	10808	2	11
9	171026_pose1	14790	1	9
10	170404_haggling_b1	15320	16	15
11	170404_haggling_b1	15320	17	15
12	171026_pose1	16132	2	13
13	171204_pose5	17524	3	11
14	171204_pose5	22633	4	10
15	171204_pose5	26918	5	11
Total				183

3.5.2. Anotación manual de las imágenes

Para evaluar las reconstrucciones de manera precisa, se propone la anotación manual de ciertos *landmarks* faciales sobre las escenas presentadas anteriormente. Para ello se utiliza la aplicación CVAT², ya que es un programa adecuado para etiquetar conjuntos de imágenes que presentan una estructura definida de puntos.

De entre los 5023 *landmarks* de la cara, se seleccionan los mismos 27 puntos correspondientes a los elegidos de las anotaciones de Panoptic (ver Figura 12), estableciendo una referencia común. De esta manera, se define una estructura común con la posición de cada uno de los puntos y se ajusta manualmente su localización en la imagen, tal como se ve en la Figura 18.

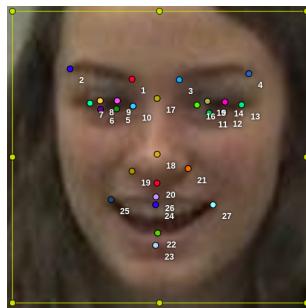


Figura 18: Estructura de los 27 *landmarks* en CVAT.

La anotación manual se lleva a cabo para cada vista de cada fotograma de las escenas de evaluación, considerando también los puntos ocultos en la imagen. En caso de que alguno de estos puntos no sea visible, se identifica como tal y se excluye posteriormente en el cálculo de la evaluación.

3.5.3. Evaluación de las reconstrucciones refinadas

La evaluación se realiza midiendo el error de reproyección de los puntos 3D de la reconstrucción refinada. Este error se calcula tanto contra las anotaciones provistas por CMU Panoptic como contra nuestras anotaciones manuales. Este procedimiento de evaluación se utilizará para seleccionar el algoritmo final de reconstrucción y validar las reconstrucciones finales. Se debe tener en cuenta que para realizar la evaluación de los métodos de ordenación se realizan un conjunto de reconstrucciones para cada escena utilizando una cantidad diferente de cámaras, por lo que los errores medios se obtienen teniendo en cuenta este conjunto de reconstrucciones.

El cálculo realizado para obtener el error de reproyección por píxeles está basado en la función de OpenCV *computeReprojectionErrors*³, la cual utiliza la norma \mathcal{L}_2 para los *landmarks* y el [Error de Raíz Cuadrada Media \(RMSE\)](#). Este error se desglosa en varias categorías según el objetivo de la medición:

1. **Error de los *landmarks*:** es el error de reproyección calculado para cada punto característico del conjunto de todas las vistas de una escena. Para obtener el error de reproyección de un *landmark* de una imagen se utiliza la siguiente Ecuación 7:

²<https://www.cvcat.ai/>

³<https://github.com/kipr/opencv/blob/master/samples/cpp/calibration.cpp>

$$\mathcal{E}_l = \|u_l - \hat{u}_l\|_2 \quad (7)$$

donde l sería el punto actual sobre el que se está calculando el error en una imagen concreta siendo u_l el *landmark* refinado y \hat{u}_l el del *ground-truth*. Si se desea el error de reprojeción de este *landmark* teniendo en cuenta todas las vistas de una escena se utiliza la Ecuación 8:

$$\mathcal{E}_{le} = \sqrt{\frac{\sum_{c=1}^C \mathcal{E}_{lc}^2}{L_{le}}} \quad (8)$$

donde C sería el total de imágenes y \mathcal{E}_{lc} el error de un *landmark* para la imagen actual, c , dividido entre el total de veces que aparece el *landmark*, l , en todas las imágenes de la escena e (L_{le}). En conjunto, se obtiene el RMSE de un *landmark* teniendo en cuenta todas las imágenes de una escena. Este cálculo de error se utilizará en la evaluación de la reconstrucción final obtenida.

2. **Error de una imagen:** ofrece el error de reprojeción para una imagen (\mathcal{E}_c) siguiendo la Ecuación 9:

$$\mathcal{E}_c = \sqrt{\frac{\|u_c - \hat{u}_c\|_2^2}{L_c}} \quad (9)$$

donde $u_c \in \mathcal{R}^{2L_c \times 1}$, agrupa la matriz de *landmarks* a evaluar de una imagen, c , y $\hat{u}_c \in \mathcal{R}^{2L_c \times 1}$, representa la matriz de *landmarks* de la imagen contra la que se evalúan los puntos. En este caso, L_c es el total de *landmarks* visibles de la imagen actual.

3. **Error de una reconstrucción:** para obtener el RMSE de los errores de reprojeción de una reconstrucción, se tiene en cuenta el error de estos puntos en las imágenes. Este cálculo será usado en la evaluación de los métodos de ordenación y las reconstrucciones finales. Se define la Ecuación 10:

$$\mathcal{E}_r = \sqrt{\frac{\sum_{c=1}^C \mathcal{E}_c^2}{L_e}} \quad (10)$$

que mantiene la misma notación con C para el total de cámaras y \mathcal{E}_c siendo el error calculado para cada vista, c . Esto se divide entre el total de puntos visibles de todas las cámaras de una escena (L_e).

4. **Error de una escena:** muestra el RMSE medio de los *landmarks* de todas las reconstrucciones realizadas para un método de ordenación siguiendo la Ecuación 11:

$$\mathcal{E}_e = \frac{1}{R} \sum_{r=1}^R \mathcal{E}_r \quad (11)$$

siendo \mathcal{E}_r el error de una reconstrucción, r , concreta del total de reconstrucciones R de la escena, e .

5. Error del método: muestra la media de los errores calculados anteriormente teniendo en cuenta todas las reconstrucciones realizadas para todas las escenas con un método de ordenación concreto. Se utiliza la la Ecuación 12:

$$\mathcal{E}_m = \frac{1}{E} \sum_{e=1}^E \mathcal{E}_e \quad (12)$$

donde E es el total de escenas y \mathcal{E}_e representa el error para una escena, e , calculado anteriormente.

Estas ecuaciones serán utilizadas en las secciones 5.1 y 5.2 para el cálculo del error de reproyección y ofrecen una visión detallada del rendimiento de los métodos de reconstrucción y la reconstrucción final ofrecida por el método seleccionado.

3.6. Visualización de las reconstrucciones 3D

Con el propósito de visualizar las reconstrucciones finales se utiliza el módulo VTK de Python, capaz de procesar una malla formada por triángulos en tiempo real y mostrarla con textura. Para obtener esta visualización, se desarrolla un algoritmo basado en la conectividad de los puntos para la generación de triángulos. Se emplea la información de conexión entre *landmarks* que DAD-3DNet utiliza para dibujar las mallas. Cuando dos puntos están definidos como conexos, se comprueba si se encuentran a su vez unidos con un tercero. A partir de estas conexiones, se construyen los triángulos para poder visualizar la cabeza completa con textura.

3.7. Entrenamiento de los detectores de *landmarks* 2D con las reconstrucciones 3D en Panoptic

En esta sección se presenta la metodología para el entrenamiento de los nuevos modelos de detección de *landmarks* utilizando las anotaciones provenientes de la proyección de las reconstrucciones 3D. Este enfoque tiene como objetivo integrar información sobre las occlusiones faciales derivadas de la pose de la cabeza, a fin de comparar la precisión final de los puntos 2D predichos.

3.7.1. Definición de visibilidad de los *landmarks* 2D anotados

Antes de comenzar el entrenamiento, se analiza la visibilidad de los *landmarks* según la orientación de la cabeza. Primero, se obtiene esta orientación predicha por el modelo DAD-3DNet para una imagen concreta (ver Figura 15). Para determinar si los *landmarks* son visibles o no en la imagen, se debe utilizar la reconstrucción 3D, por lo que el vector de orientación calculado sobre la imagen se traslada al modelo 3D. De esta manera, se pueden dividir los puntos 3D mediante un plano sobre la reconstrucción de la cabeza en los conjuntos de puntos visibles y ocultos. Este procedimiento será explicado con mayor detalle en la sección 4.3.

3.7.2. Entrenamiento de los detectores de *landmarks* 2D con las nuevas anotaciones

La realización de la aplicación final con las nuevas anotaciones se basa en el análisis del entrenamiento de un nuevo modelo con Tensorflow teniendo en cuenta las oclusiones de los *landmarks*. La red neuronal utilizada es la propuesta por Yin Guobing⁴ y resulta adecuada para realizar detecciones de *landmarks* faciales. La arquitectura de esta red neuronal consiste en seis capas convolucionales definidas por normalizaciones por lotes y filtros de convolución de diferentes tamaños. Además, se realiza una agrupación (o *pooling*) para cada una de ellas. Luego, la salida de las capas convolucionales se aplana y se conecta a una capa densa de tamaño 1024 activada por ReLU y normalizada. Finalmente, la capa de salida, también densa, genera las predicciones con una función de activación lineal. Este diseño permite aprender características de las imágenes de entrada y realizar detecciones de los puntos faciales de referencia. Como estas predicciones se basan en la posición de los *landmarks*, cabe mencionar que es un tipo de problema de regresión donde el objetivo es predecir valores numéricos continuos que definen las coordenadas de los puntos.

Las imágenes de entrada al modelo serán recortes de 256x256 píxeles que muestran únicamente las caras de los sujetos. Por otro lado, las etiquetas corresponden a los 5023 puntos refinados proyectados sobre la imagen.

El entrenamiento se realizará dos veces cambiando la función de pérdida utilizada. De este modo, se desea comparar el rendimiento en las predicciones al incluir información sobre la visibilidad y oclusión de los puntos frente a utilizar una pérdida equivalente para todos. Las dos funciones de pérdida utilizadas en cada entrenamiento son las siguientes:

1. Usando el [Error Medio Cuadrático \(MSE\)](#) según la Ecuación 13:

$$\mathcal{L}_1 = \frac{1}{n} \sum_{i=1}^n (y_{\text{true},i} - y_{\text{pred},i})^2 \quad (13)$$

donde n es el total de *landmarks*, $y_{\text{true},i}$ es la posición real de los *landmarks* y $y_{\text{pred},i}$ es la posición de los *landmarks* predicha por el modelo.

2. Utilizando también [MSE](#) pero reduciendo en un 70 % el valor de la pérdida de los *landmarks* no visibles. Para ello, se calcula la media de las diferencias correspondientes a los *landmarks* visibles y no visibles como en las Ecuaciones 14 y 15:

$$\mathcal{L}_{\mathcal{O}} = 0,3 \cdot \frac{1}{|\mathcal{O}|} \sum_{i \in \mathcal{O}} (y_{\text{true},i} - y_{\text{pred},i})^2 \quad (14)$$

$$\mathcal{L}_{\mathcal{V}} = \frac{1}{|\mathcal{V}|} \sum_{j \in \mathcal{V}} (y_{\text{true},j} - y_{\text{pred},j})^2 \quad (15)$$

donde \mathcal{O} es el conjunto de *landmarks* ocluidos y \mathcal{V} el de visibles. Además, $|\mathcal{O}|$ representa el cardinal del conjunto O y $|\mathcal{V}|$ el del conjunto \mathcal{V} . Finalmente, se suman ambos valores calculando el valor de pérdida final como en la Ecuación 16:

$$\mathcal{L}_2 = (\mathcal{L}_{\mathcal{O}} + \mathcal{L}_{\mathcal{V}}) \quad (16)$$

⁴<https://github.com/yinguobing/cnn-facial-landmark/tree/master>

3.7. ENTRENAMIENTO DE LOS DETECTORES DE *LANDMARKS* 2D CON LAS RECONSTRUCCIONES 3D EN PANOPTIC

De esta manera se consiguen definir las distintas funciones de pérdida dependientes de la visibilidad de los puntos que serán implementadas en los entrenamientos.

Capítulo 4

Desarrollo

En esta sección se exponen detalladamente cada uno de los pasos seguidos en la elaboración de este proyecto. En la Figura 19 se muestra el diagrama que resume visualmente esta estructura.

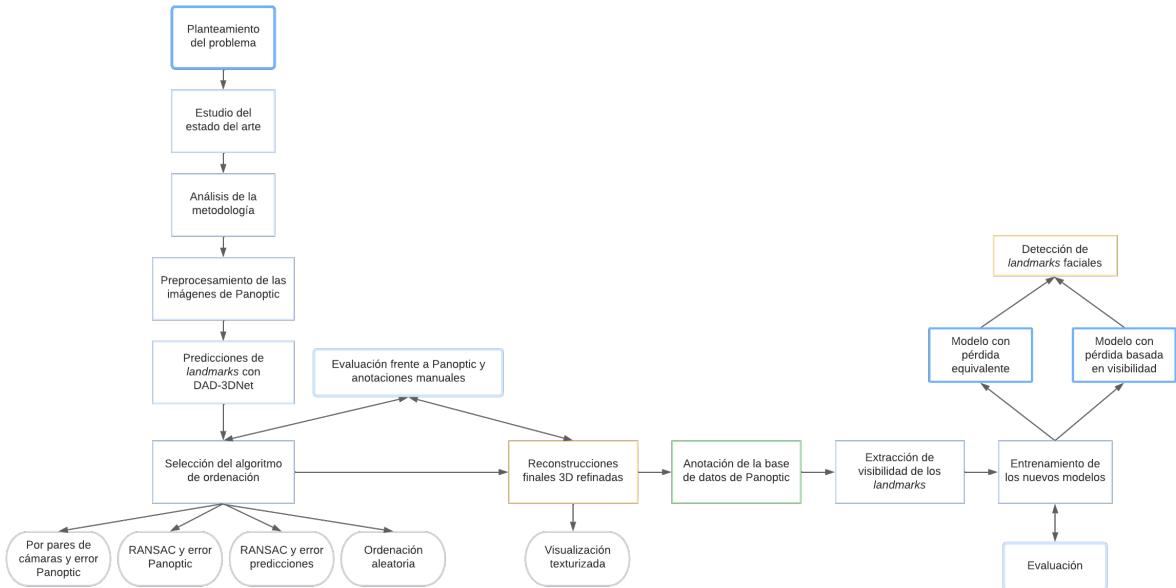


Figura 19: Diagrama del proyecto.

Tras el estudio de la propuesta planteada y sus características, se comienza con el desarrollo del proyecto. En esta sección se abarca en primer lugar, el procesamiento llevado a cabo sobre las imágenes de las escenas de Panoptic para filtrar y recortar las imágenes, de manera que puedan ser utilizadas con el modelo entrenado de DAD-3DNet.

Igualmente, se analizan los algoritmos utilizados para la ordenación y selección de cámaras empleadas en la reconstrucción final. Finalmente, se detalla el proceso de extracción de la visibilidad de los puntos refinados y el entrenamiento de las nuevas redes neuronales para la detección de *landmarks* 2D.

4.1. Preprocesamiento de las imágenes de Panoptic

Para poder realizar las predicciones con DAD-3DNet y procesar las reconstrucciones, es necesario obtener primero las imágenes de las escenas de Panoptic en el formato esperado. Con este objetivo se tratan las imágenes de las cámaras HD, identificando los rostros y las occlusiones para obtener el conjunto de datos a utilizar.

Debido a que el modelo de DAD-3DNet trabaja con imágenes donde el rostro de la persona ocupa prácticamente toda la imagen, se debe preparar Panoptic para obtener este tipo de visualización. Este proceso implica identificar las caras en la escena y recortar esas áreas considerando el tamaño y la orientación de la cabeza. Con la ayuda de algunas anotaciones proporcionadas por los datos de Panoptic (ver Figura 11), se establece un algoritmo que utiliza las posiciones de estos puntos para comprobar la validez de las anotaciones, manejar las occlusiones y extraer la pose de cada sujeto.

Si al observar las posiciones de los puntos característicos de las orejas, la nariz y los ojos del sujeto en la imagen, estos aparecen de manera ordenada según sus posiciones esperadas de izquierda a derecha, la persona tendrá una pose frontal. Si aparecen en orden contrario, la persona se encontrará de espaldas a la cámara y la imagen será eliminada dado que DAD-3DNet necesita que el rostro sea visible aunque sea parcialmente. En el caso de que la posición del *landmark* de la nariz esté más hacia la izquierda o derecha de la imagen que el resto, se establece que la pose de la cabeza es de perfil. Finalmente, se ajusta el recorte de la imagen original de acuerdo con la estimada de la pose de la cabeza (frontal, perfil izquierdo o perfil derecho).

Adicionalmente, se incorporan ciertas condiciones respecto a las distancias entre los ojos, orejas y nariz, con el propósito de prevenir posiciones de perfil excesivamente marcadas o situaciones en las que la cámara se encuentre ubicada sobre la cabeza, lo que podría generar casos más extremos de occlusión. Utilizando las coordenadas del punto característico del pecho, es posible verificar si la persona está mirando hacia abajo, lo que ayuda a distinguir entre una pose frontal y una pose donde solo se observa la parte superior de la cabeza en la cámara. En la Figura 20 se ilustran algunas posiciones que serían conservadas para la reconstrucción según estas condiciones. De forma contraria, en la Figura 21 se presentan tres ejemplos de vistas de cámara que serían filtradas (rechazadas) en este procesamiento.

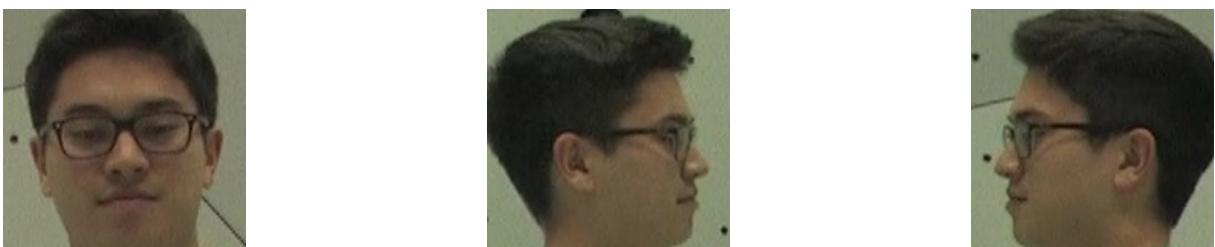


Figura 20: Poses válidas.

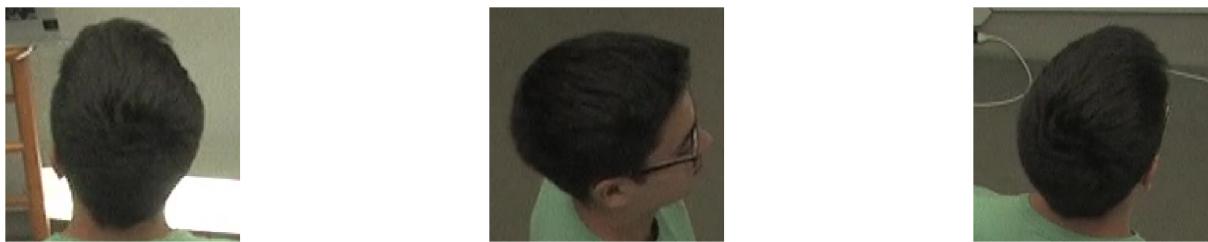


Figura 21: Poses no válidas filtradas.

En las Figuras 20 y 21 anteriores, se puede ver cómo cambia la pose de la persona según la posición de la cámara para un mismo fotograma, pudiendo realizar este filtrado según la relación entre los puntos comentados.

Después de realizar el primer procesamiento para determinar la pose de la persona, es crucial considerar las posibles occlusiones, ya sean causadas por el mismo individuo o por otros. Para una misma persona (ver Figura 22), se emplean anotaciones sobre las posiciones de sus manos y se verifica la ausencia de occlusiones causadas por estas. En el caso de la pose frontal, se examinan las posiciones de ambas manos, mientras que para el perfil izquierdo se evalúa la ocultación con la mano izquierda, y para el perfil derecho, con la mano derecha. A pesar de excluir ciertos casos, el modelo DAD-3DNet es capaz de lidiar con ciertas occlusiones parciales. Sin embargo, es posible que haya situaciones en las que no se pueda evitar una pérdida total de visibilidad al no considerar todas las anotaciones en el proceso, tal como se ve en la Figura 22.

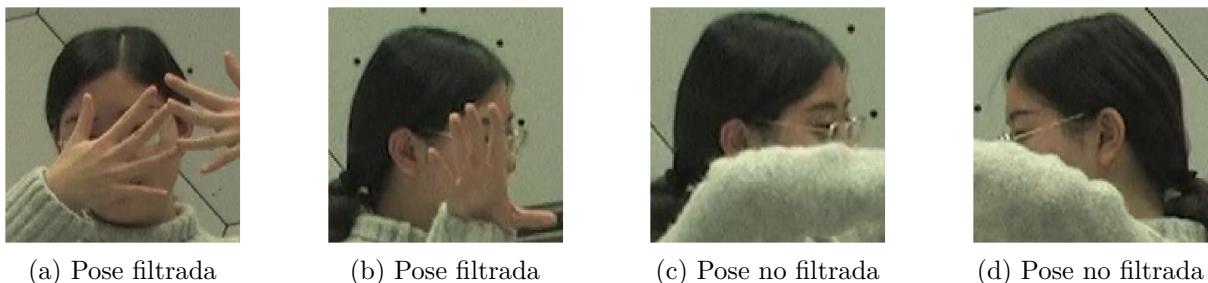


Figura 22: Oclusiones entre una misma persona.

Con relación a las occlusiones causadas entre individuos, es necesario verificar si las coordenadas de los puntos sobre la cara o las manos coinciden con el rostro de otra persona. Si existe una superposición entre caras, ambas se identificarán como occlusiones en esta vista con el fin de prevenir posibles errores del modelo al interpretar la presencia de dos caras juntas. En la Figura 23 se muestran dos ejemplos donde las imágenes serán identificadas como occlusiones y estas vistas de la cámara no serán utilizadas en la futura reconstrucción.



Figura 23: Oclusiones entre personas filtradas.

En caso de que se identifique una cara como válida y visible, se procede a recortarla de la imagen original, tomando en consideración tanto la orientación inicialmente detectada como la posición de los puntos de los ojos, la nariz y los labios. Aunque el tamaño de las imágenes extraídas puede variar para adaptarse al tamaño de la cabeza en la imagen, se mantendrá un formato cuadrado, ya que es el formato requerido por DAD-3DNet y proporciona una mejor adaptación ante distintas orientaciones del rostro. Aparte de este propósito, este filtrado también sirve para evitar utilizar las cámaras que no se encuentren en funcionamiento, ya que no tendrán anotaciones asociadas.

4.2. Reconstrucción 3D y anotación de la base de datos de Panoptic

En los siguientes apartados se detalla tanto el proceso para llevar a cabo la reconstrucción 3D utilizando múltiples vistas de las cámaras de un mismo fotograma como los algoritmos desarrollados para la ordenación y selección de estas vistas. El objetivo es investigar cómo la disposición y omisión selectiva de algunas cámaras afectan a la calidad de la reconstrucción final. Se proporciona información sobre los métodos utilizados para la evaluación de cámaras, el criterio de error utilizado y, por último, el proceso de obtención de las anotaciones de la base de datos.

4.2.1. Predicciones de las vistas con DAD-3DNet

Después de aplicar el preprocesamiento mencionado anteriormente para cada fotograma y habiendo filtrado las vistas de la cámara, las imágenes seleccionadas serán utilizadas para llevar a cabo la reconstrucción 3D. El primer paso consiste en obtener las predicciones de los 5023 *landmarks* de cada vista mediante DAD-3DNet. Estos datos son fundamentales para realizar la inicialización lineal y el refinamiento con [BA](#).

Por ejemplo, en el caso de la escena de evaluación número 8, de las 31 cámaras, se han seleccionado 11 imágenes donde se visualiza claramente el rostro del sujeto. Al aplicar DAD-3DNet a las tres imágenes de la Figura 20 correspondientes al mismo fotograma, se generan las mallas representadas en la Figura 24.

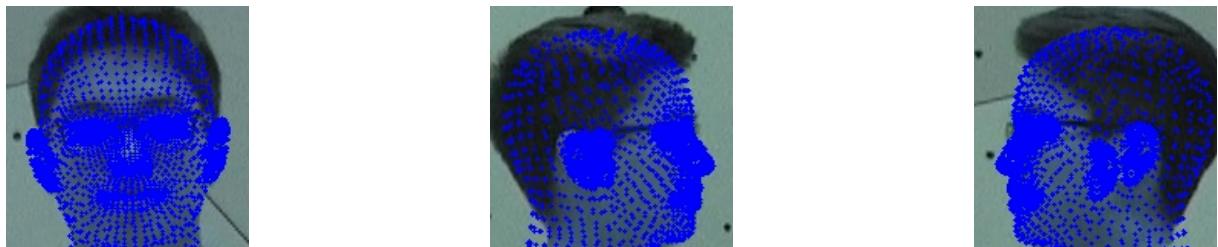


Figura 24: Detecciones de *landmarks* faciales con DAD-3DNet.

Sin embargo, es importante señalar que debido a que los datos utilizados son predicciones de una red neuronal a partir de una única imagen, algunos de los puntos de referencia pueden tener cierto error, tal como se ve en los ejemplos de la Figura 25 para diferentes escenas.

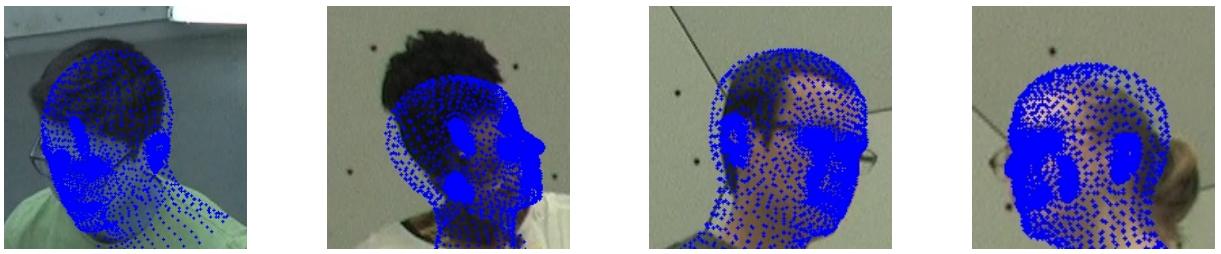


Figura 25: Detecciones de *landmarks* faciales con DAD-3DNet que presentan un error considerable

Debido a que estos son los datos utilizados para el proceso de reconstrucción de un fotograma, se debe prestar especial atención a su calidad en cada imagen, para evitar la utilización de predicciones incorrectas en la reconstrucción final. Este procedimiento de análisis será estudiando en detalle más adelante.

4.2.2. Reconstrucción facial 3D

El método de reconstrucción 3D, se basa en la utilización de los *landmarks* detectados por DAD-3DNet para obtener un modelo 3D refinado. Este algoritmo hace uso de este conjunto para construir el modelo de la cabeza mediante el método expuesto en la sección 3.4.1. Después, con el algoritmo [BA](#), desarrollado en la sección 3.4.2, se ajusta esta reconstrucción facial utilizando los parámetros de calibración de las cámaras provistos en Panoptic y modificando únicamente la posición de los puntos en el espacio.

4.2.3. Algoritmos de ordenación de cámaras

Con el fin de obtener reconstrucciones de mayor calidad se han estudiado varios métodos para ordenar las predicciones asociadas con cada cámara según su precisión con respecto a unos datos de referencia. Como se detallará en la sección 5.1, cuando se utilizan todas las cámaras en la reconstrucción, el error tiende a ser más elevado que si se realiza una selección previa. Para encontrar esta mejor disposición de las cámaras y evitar las predicciones incorrectas, se han desarrollado tres algoritmos que serán comparados junto con la ordenación aleatoria de las cámaras. Estos organizan las cámaras de las escenas de evaluación y sus resultados serán expuestos en la sección 5, justificando el algoritmo utilizado en la obtención de las reconstrucciones 3D finales.

A continuación se va a describir el funcionamiento de cada uno de los tres métodos propuestos, detallando sus características y funcionamiento. Para cada uno de ellos es necesario realizar una reconstrucción lineal como inicialización de [BA](#).

Ordenación de las cámaras por pares frente a Panoptic

En este método se ordenan las vistas (o cámaras) utilizando el error de la reconstrucción (Ecuación 10) por cada par de cámaras. Como criterio de ordenación se utiliza el promedio de los errores anteriores para cada vista en comparación con el resto. Para cada cámara se calcula su reconstrucción lineal con otra cámara y se refina con [BA](#). Tras este procedimiento, se obtiene su error frente a los 27 *landmarks* obtenidos de las anotaciones

de Panoptic y se guarda este valor de error para las dos cámaras involucradas. Finalmente, se calcula el promedio de los errores de reconstrucción para cada cámara y se ordenan las vistas de menor a mayor según estos valores. En este algoritmo se realiza una reconstrucción por cada par de cámaras con lo que el tiempo de ejecución aumenta según $\mathcal{O}(n \cdot (n - 1))$, siendo n el número total de vistas.

Ordenación de las cámaras usando **RANSAC** frente a Panoptic

Este procedimiento es el algoritmo final seleccionado para la ordenación y selección de las cámaras. En este caso se utiliza la misma idea que en el método anterior, obteniendo el error frente a las anotaciones de Panoptic, pero aplicando un algoritmo de muestreo aleatorio (a la **RANSAC**) a fin de reducir el tiempo de ejecución. El funcionamiento es el siguiente:

1. Se calcula el error mínimo para considerar dos cámaras *inliers*. Para obtener este valor, se construye y refina la reconstrucción utilizando todas las cámaras ordenadas de forma aleatoria. Como se verá en la sección 5.1, el valor del error se estabiliza cuando se utilizan todas las cámaras sin importar el orden utilizado y este valor suele ser más alto que cuando se usan dos cámaras únicamente.
2. Durante 50 iteraciones, se eligen dos cámaras al azar y se obtiene su reconstrucción lineal y refinada con **BA**. Se debe tener en cuenta que con el preprocesamiento de las cámaras anterior suele eliminarse la mitad de las 31 cámaras HD debido a su disposición en Panoptic (Figura 9) u otras occlusiones. Este número de iteraciones fue comparado frente a usar 30 y 40 iteraciones, viendo que el error podía subir considerablemente cuando había un mayor número de cámaras. Al usar un mayor número de iteraciones, como 60, 80 o 100, este error se estabiliza y el tiempo de ejecución no compensa con el posible error reducido. Esto es debido a que las mejores predicciones suelen producir un error menor que el error mínimo en conjunto con otras cámaras gracias al ajuste con **BA**, por lo que son consideradas igualmente como *inliers*. Sin embargo, en combinaciones de cámaras con mayor error entre sí, el error en conjunto suele ser superior al error mínimo. Es cierto que podrían ocurrir errores debido a la naturaleza aleatoria del algoritmo, pero se considera que este número de iteraciones es suficiente para el análisis realizado.
3. Si el error de esta reconstrucción frente a los *landmarks* seleccionados en Panoptic es menor que el error mínimo, las cámaras son consideradas *inliers* para esta iteración. Así va aumentando su contador propio que indica cuántas veces han cumplido esta condición.

Tras finalizar las iteraciones, se obtiene el número de veces que cada cámara ha sido *inlier*, pudiendo ordenar las vistas según este valor de mayor a menor.

Ordenación de las cámaras usando **RANSAC** frente a las predicciones de DAD-3DNet

En este caso, el procedimiento es el mismo que el anterior, pero obteniendo el error en función de los 5023 *landmarks* de las predicciones de DAD-3DNet. Esta comparación tiene

la ventaja de que se tienen en cuenta todos los puntos de la cabeza y no solo unos pocos de la parte frontal de la cara. Sin embargo, con respecto a la aplicación de los métodos, es necesario recordar que tanto las predicciones de DAD-3DNet como las anotaciones de Panoptic pueden contener errores en los puntos característicos, por lo que estos criterios pueden llevar a resultados erróneos al no tratarse de anotaciones realizadas manualmente.

4.2.4. Selección de cámaras para la reconstrucción facial 3D final

Tras la evaluación de los métodos mencionados, cuyo análisis se presentará en la sección 5, se selecciona el método [RANSAC](#) sobre las anotaciones de Panoptic. Esta elección se basa en que obtiene menores errores en general y su tiempo de ejecución se ve reducido frente a comprobar todas las vistas a partir de un número mayor a 7 cámaras. En comparación con el método que utiliza las predicciones y el procedimiento de ordenación aleatoria de cámaras, esta diferencia en el error es aún más notable en la mayoría de los casos.

Hasta este punto, estos métodos se limitaban a la ordenación de cámaras o más bien, predicciones de cada cámara. Sin embargo, también se desea evitar el uso de ciertas vistas que pueden perjudicar el modelo final. Para esta selección de cámaras, se utiliza la lista del número de veces que cada cámara permite obtener una reconstrucción con un error bajo según las siguientes premisas:

- Si ninguna cámara ha sido *inlier*, es decir, ningún par de cámaras ha conseguido un error menor que utilizar todas las cámaras, se usan todas las vistas en la reconstrucción según como estaban ordenadas inicialmente.
- En el caso de que solo 2 o 3 cámaras hayan sido *inliers*, nótese que nunca habrá el caso de una sola cámara lo sea, se usarán únicamente estas predicciones en la reconstrucción.
- Si hay 4 o más cámaras que han sido *inliers*, se cuentan cuántas de estas han sido *inliers* más de 3 veces. Si hay al menos 4 cámaras o más que han sido *inliers* 3 o más veces, se seleccionan solo las predicciones de estas vistas para la reconstrucción final.
- Si de las cámaras seleccionadas hay menos de 4 cámaras que han sido *inliers* 3 veces o más, entonces la reconstrucción se realizará con todas las cámaras que hayan sido *inliers* alguna vez.

Lo que se trata de obtener con esta selección es el conjunto de las cámaras que han sido *inliers* varias veces durante las iteraciones. No obstante, si hay un número muy reducido de cámaras consideradas como *inliers*, o si estas solo han sido identificadas en pocas ocasiones, se optará por utilizar todas las cámaras que alguna vez han sido *inlier*. Esto ocurre cuando hay un conjunto de cámaras que tienen un error considerablemente mayor, que evita que en su emparejamiento con mejores vistas el error sea menor que el mínimo. De esta manera se tiene en consideración que estas cámaras *inliers* ciertamente lo son cuando han sido emparejadas entre sí. Por otro lado, se aplicará un filtro más estricto seleccionando menos cámaras cuando estas destacan obteniendo errores significativamente menores en las reconstrucciones que generan, siendo mayor el número de veces que han sido *inliers* junto a otras cámaras.

4.2.5. Anotaciones faciales de la base de datos de Panoptic

Los puntos XYZ del espacio 3D que forman las reconstrucciones finales refinadas, son guardados en formato TXT, pudiendo ser utilizados para el entrenamiento de nuevos modelos. De igual forma, se guarda la reproyección de estos puntos sobre las imágenes correspondientes en formato JSON de manera normalizada, obteniendo las etiquetas correspondientes a los *landmarks* faciales 2D. Esta normalización se realiza entre 0 y 1 y sirve para poder ajustar los *landmarks* cuando el tamaño de la imagen cambia. Este procedimiento resulta necesario debido a que no todas las imágenes recortadas poseen el mismo tamaño y deben ajustarse para los entrenamientos de los detectores de *landmarks*, por lo que los puntos pueden ser relacionados junto con este nuevo tamaño. Para las reproyecciones utilizadas en las secciones 4.3 y 4.4 se selecciona una parte de estas anotaciones. Aun así, se debe tener en cuenta que a pesar de que los datos han sido procesados y refinados, estos no son tan precisos como si se hubiese realizado una anotación manual o mediante escáneres láser.

4.3. Extracción de la visibilidad de los *landmarks* 2D anotados

Se desea utilizar las reconstrucciones 3D de la cabeza para entrenar modelos de análisis facial. En el TFM se va a validar estas anotaciones con dos detectores de *landmarks*, uno de los cuales considerará la visibilidad de los *landmarks* 2D durante el entrenamiento. Este concepto de visibilidad de los *landmarks* es dependiente de la pose de la cabeza de cada sujeto. Por ejemplo, en una imagen frontal, las occlusiones estarán en la parte posterior y superior de la cabeza. Sin embargo, en una imagen de perfil, estas occlusiones pertenecerán al lateral contrario de la cabeza mostrada en la imagen. Por tanto, para extraer estos puntos, el procedimiento seguido es el siguiente:

1. Se extraen los ángulos *yaw*, *pitch* y *roll* correspondientes a la orientación de la cabeza ofrecida por el modelo DAD-3DNet y se calcula la matriz de rotación correspondiente. En la Figura 26 se muestra la detección de orientación del sujeto en la imagen.

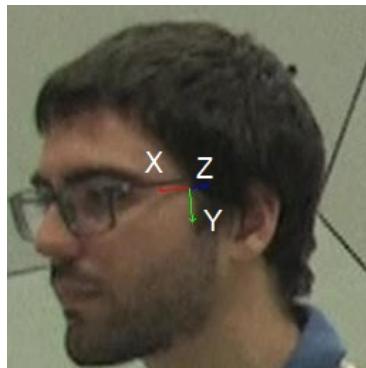


Figura 26: Orientación del sujeto en la imagen.

2. Se obtiene el punto medio de los 5023 puntos de la reconstrucción, representado como m en la Figura 27.

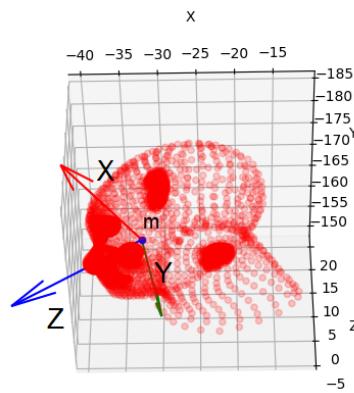


Figura 27: Vectores de orientación sobre la reconstrucción.

3. Se calculan la matriz de rotación R , entre el sistema de referencia del modelo 3D de la cabeza y el de la cámara. Los ejes del sistema de referencia de la cabeza, columnas de R , se muestran en este modelo desde el punto medio de la reconstrucción. En la Figura 27 se ven estos ejes detectados de orientación referenciados al modelo 3D. El color azul representa el eje Z , el verde el eje Y y el rojo el X , que indica la dirección hacia la que mira el sujeto.
4. Para construir el sistema de referencia de la reconstrucción 3D en el espacio, se obtiene primero el vector H que define el eje de la cabeza 3D. Mediante el punto medio y la selección del *landmark* número 3565, se configura este vector, mostrado en color negro en la Figura 28, que ofrece la información de la inclinación de la cabeza en el espacio.

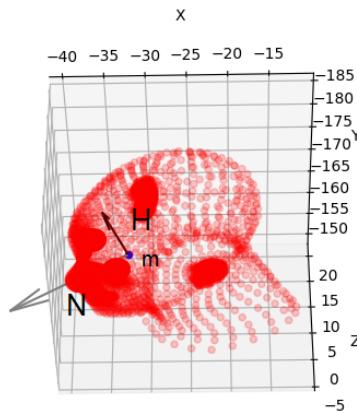


Figura 28: Sistema de referencia 3D.

5. Después, se obtiene el eje N que indica hacia donde está orientado el modelo 3D, es decir, sería nuestro eje X de la reconstrucción. Para ello, se selecciona el punto número 3526, que está situado en la nariz y se calcula el vector del centro medio a la nariz, constituyendo el vector N . De esta manera se define el sistema de referencia en el espacio 3D como se ve en la Figura 28
6. Se calcula la diferencia en grados entre el eje X de la imagen frente al correspondiente eje N .

7. Se obtiene el vector P que pertenece al plano que divide la visibilidad de los *landmarks*. Para conseguir este vector, se gira el vector X sobre el eje H el doble de la diferencia calculada anteriormente. Este giro será en sentido horario o antihorario según si el eje X se encuentra en la parte derecha o izquierda del modelo 3D. De esta manera se obtiene un vector análogo hacia el lado contrario. Después, se suman o restan 90° adicionales, obteniendo el vector P que forma parte del plano que divide los puntos en visibles o no visibles. Este se muestra en la siguiente Figura 29 de color magenta.

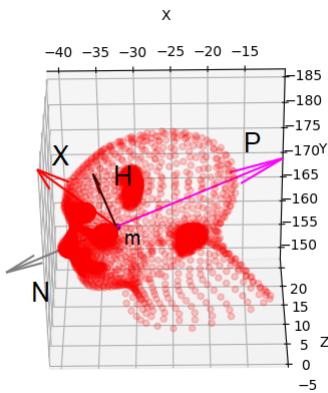


Figura 29: Vector contenido en el plano.

8. El plano se construye mediante el eje H de la reconstrucción 3D y el vector P del plano. Este plano se dibuja en la Figura 30

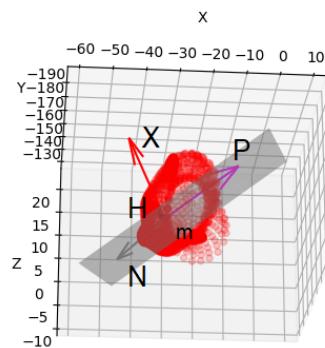


Figura 30: Plano visibilidad de los *landmarks*.

9. Para cada punto, se calcula en qué parte de la imagen del plano se encuentra, definiendo así su visibilidad.

Tras realizar los pasos anteriores, se puede proceder a extraer visibilidad u oclusión de cada punto según el plano anterior. En la Figura 31 se pueden observar los puntos visibles en rojo y los puntos ocultos en negro según cada imagen, así como el vector X de orientación, el vector P contenido en el plano y los ejes N y H .

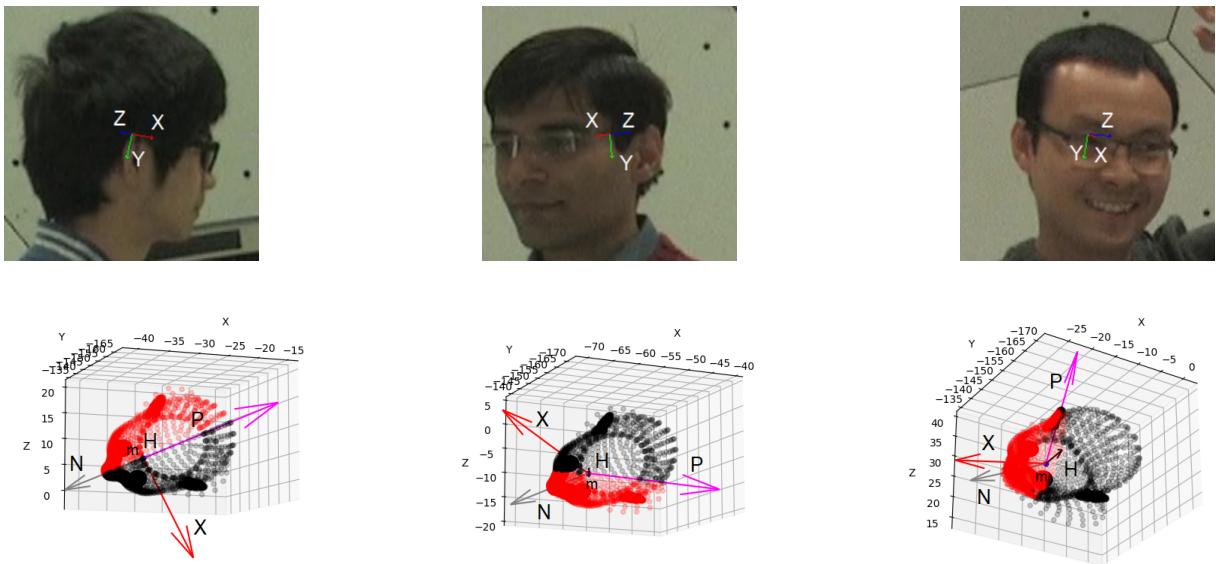


Figura 31: Extracción de los puntos visibles según la orientación de la cabeza.

Es importante considerar que la división se lleva a cabo siempre con respecto al plano a lo largo del eje central, sin tener en cuenta si el sujeto tiene la cabeza inclinada. En consecuencia, puede haber errores en la declaración de ciertos puntos como no visibles en algunos casos. Además, este cálculo está condicionado por la exactitud de la matriz de rotación predicha por DAD-3DNet.

4.4. Entrenamiento de los detectores de *landmarks* 2D

Una vez se han obtenido la clasificación de los *landmarks* de la cabeza en visibles y ocultos, se procede al entrenamiento de un nuevo modelo definido en Keras. La entrada a este modelo será una imagen recortada con la cabeza de un sujeto de tamaño 256x256, mientras que la salida esperada corresponderá a los 5023 puntos característicos que definen la forma de la cabeza, tal como se obtenía con DAD-3DNet.

Como se expuso en la subsección 3.7.2, se realizan dos entrenamientos, uno donde la pérdida es igual para todos los puntos y otro donde la pérdida de los puntos con oclusión equivale a un 30 % de la diferencia de los puntos (Ecuación 15), reduciendo su impacto sobre la predicción final.

Para realizar esta parte se utilizaron un total de 1568 imágenes de las escenas *170404_hagging_b1*, *171026_pose1* y *171026_pose2* con diferentes personas y vistas de cámara. El 80 % fueron utilizadas el conjunto de entrenamiento y el 20 % restante para test, siendo un total de 314 imágenes. El conjunto de entrenamiento se dividió en entrenamiento y validación en la proporción 80/20, empleando un total de 1003 y 251 imágenes respectivamente. Estas fueron divididas en *batches* de tamaño 4 y se utilizó la condición de parada temprana (o *EarlyStopping*) para terminar automáticamente el entrenamiento cuando el error de validación hubiese aumentado durante 15 épocas de forma consecutiva, lo que evita un sobreentrenamiento de la red.

En la Figura 32 se muestran algunas imágenes utilizadas en el entrenamiento junto con sus etiquetas, donde los *landmarks* visibles se marcan en rojo y los ocultos en negro.

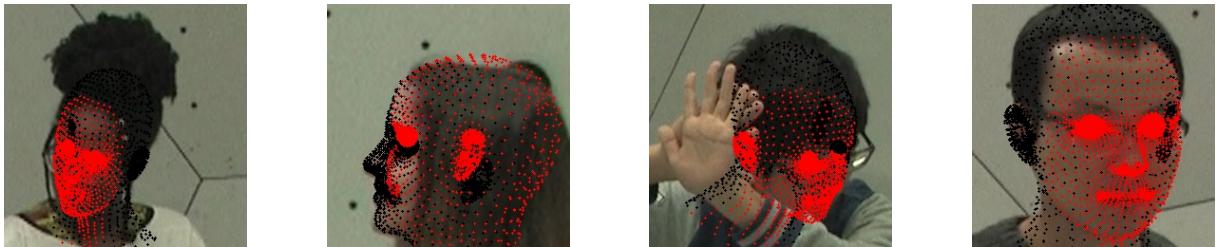


Figura 32: Ejemplo de imágenes utilizadas en el entrenamiento con sus etiquetas.

No obstante, es necesario recordar que debido a la propia naturaleza de las etiquetas utilizadas, ya que se trata de anotaciones procesadas automáticamente, pueden contener errores. En la Figura 33 se presentan algunos ejemplos de estos casos, donde los errores pueden ser debidos a una mala reconstrucción procesada o una orientación incorrecta obtenida por DAD-3DNet. Estos errores claramente afectarán al rendimiento de los modelos, condicionando en más medida a la función de pérdida calculada según la visibilidad de los puntos.

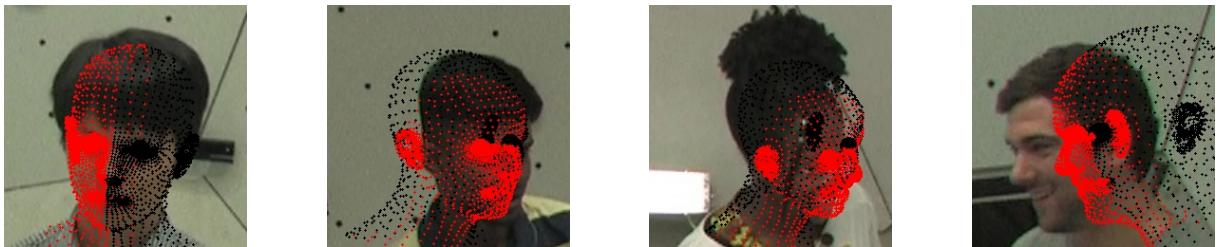


Figura 33: Ejemplo de imágenes incorrectas utilizadas en el entrenamiento con sus etiquetas.

Por otra parte, para poder realizar el entrenamiento con la pérdida personalizada teniendo en cuenta los *landmarks* ocultos en la imagen y el tamaño de los *batches*, se sigue el siguiente procedimiento:

1. Se declaran cuatro variables usando el *backend* de Keras que almacenarán los índices de los *landmarks* visibles y ocultos para el entrenamiento.
2. Una de las variables guarda la lista de los índices de los *landmarks* visibles para el *batch* actual, mientras que la otra guarda la lista de los índices ocultos. En relación con las otras dos variables, se almacenan únicamente los índices visibles y ocultos correspondientes al último *batch*. Este proceso se realiza debido a que el último *batch* podría tener un tamaño diferente al anterior y es necesario manejar estos cambios de tamaño durante el entrenamiento con Tensorflow.
3. Mediante una función de *callback* del entrenamiento, se actualizan los índices de los *landmarks* visibles y ocultos de las variables de Keras al final de cada *batch* con los valores para el nuevo conjunto de imágenes.

El proceso anterior se realiza para aplicar las funciones de pérdida en los diferentes conjuntos de puntos visibles y ocultos durante el entrenamiento, de manera que no haya que alterar la estructura de las predicciones. Además de esto, cabe decir que durante la fase validación, la pérdida se calcula con la Ecuación 13 para ambos métodos. Esto se debe a la imposibilidad de crear esta actualización en la lista de índices durante este estado del entrenamiento.

Capítulo 5

Resultados experimentales

En este capítulo se van a presentar los resultados obtenidos durante la evaluación de los algoritmos de ordenación y selección de cámaras en las reconstrucciones 3D finales. También se analizan los resultados de los entrenamientos de los detectores de *landmarks* llevados a cabo con las nuevas anotaciones.

Para la primera parte, esta sección se ha dividido en una evaluación general de los métodos y una evaluación específica del algoritmo seleccionado. En esta evaluación general de las reconstrucciones, se analiza la variación del error con el grupo de cámaras utilizado en este proceso, tanto frente a las anotaciones ofrecidas por Panoptic como frente a las anotaciones manuales. Es importante remarcar que esta evaluación se realiza sobre las escenas de evaluación frente a las anotaciones de CMU Panoptic y las anotaciones realizadas manualmente. Aunque se tienen todas las anotaciones de la base de datos de Panoptic, la evaluación manual únicamente sobre un conjunto de 15 escenas seleccionadas (ver Figura 17) con sus respectivas vistas, abarcando un total de 183 imágenes. Estas serán referenciadas según fue expuesto en la Tabla 1.

Una vez elegido el mejor algoritmo de ordenación de cámaras, se expone una evaluación específica sobre las reconstrucciones finales refinadas mediante un análisis más en detalle de las escenas de evaluación. Finalmente, se muestran estos resultados mediante la texturización de los modelos 3D.

En el apartado de los entrenamientos de los detectores de *landmarks*, se muestran las gráficas del valor de la función de pérdida de los nuevos modelos y se comparan con las predicciones de la red neuronal DAD-3DNet y las anotaciones finales utilizadas.

5.1. Evaluación de los métodos de ordenación de vistas

En esta evaluación general se presentan los resultados del estudio llevado a cabo para la selección del algoritmo de ordenación de cámaras desarrollado. Como se detalló en la sección 4.2.3, se ejecutaron tres métodos evaluados en comparación con el uso de una ordenación aleatoria de cámaras para determinar la mejor estrategia de reconstrucción. En primer lugar, uno de los factores clave de estos algoritmos es el tiempo de ejecución asociado a cada uno para obtener la ordenación de las predicciones. En la Tabla 2 se muestran los tiempos medios en segundos para obtener el orden de cámaras en cada caso.

Tabla 2: Tiempo medio de ordenación de las cámaras por cada método.

	Tiempo medio (s)
Ordenación con comparación por pares y Panoptic	88.76
Ordenación con RANSAC y Panoptic	30.36
Ordenación con RANSAC y predicciones	42.17
Ordenación aleatoria	2.61e-05

Debido a que el algoritmo de ordenación aleatoria utiliza únicamente una función, su ejecución es inmediata. También se observa una diferencia de más de 40 segundos de los métodos de **RANSAC** frente a comprobar todas las cámaras por pares. En este caso, se necesita una ejecución de minuto y medio para concluir la ordenación de las cámaras (o vistas).

Además de este estudio, se van a analizar los errores obtenidos tras realizar varias reconstrucciones con cada método en comparación con las anotaciones de Panoptic y las escenas de evaluación anotadas manualmente. Por tanto, es importante destacar que la evaluación realizada en esta sección no corresponde a la de la reconstrucción final seleccionada. El propósito de este apartado es exponer de manera objetiva por qué tanto el orden, como la elección del número de cámaras, son parámetros clave en la elección de la reconstrucción final.

Para realizar este proceso, primero se emplean los algoritmos propuestos para ordenar las cámaras según el criterio de cada uno. Posteriormente, para cada conjunto de cámaras en orden, se lleva a cabo su refinamiento mediante **BA** añadiendo las demás cámaras en orden y obteniendo el error correspondiente para cada caso. Es decir, para una inicialización lineal con dos cámaras, se realiza el refinamiento con 2, 3, 4... cámaras, luego utilizando tres cámaras en la reconstrucción inicial lineal, se refina igualmente comenzando con 2, 3, 4... cámaras hasta el total. Esto se realiza para todas las vistas de cámara disponibles de un mismo fotograma.

De esta manera, se estudia la influencia de la ordenación de las mejores predicciones en el modelo 3D y cómo la variación en el número total de vistas utilizadas en la inicialización o el refinamiento puede influir en la calidad de los resultados finales.

5.1.1. Evaluación de los métodos frente a las anotaciones de Panoptic

Para evaluar las reconstrucciones en comparación con las anotaciones de Panoptic, es necesario comparar las posiciones de los 27 *landmarks* originalmente anotados en Panoptic (ver Figura 12) con los correspondientes puntos reproyectados de una reconstrucción. El error se determina como el promedio de todos los **RMSE** de reprojeción obtenidos para las 15 escenas de evaluación tras las reconstrucciones realizadas con distinto número de cámaras en la inicialización y refinamiento según fue explicado anteriormente. Por tanto, se compara la diferencia de posición entre los *landmarks* de las anotaciones y los *landmarks* refinados y se calcula el error de reprojeción por píxeles de cada método según la Ecuación 12. En la Tabla 3, se presenta este error obtenido de todas las combinaciones de reconstrucciones para todas las escenas evaluadas en cada algoritmo.

Tabla 3: Error de las reconstrucciones para cada método frente a Panoptic.

	\mathcal{E}_m (píxel)
Reconstrucciones ordenadas con pares de cámaras y Panoptic	37.16
Reconstrucciones ordenadas con RANSAC y Panoptic	26.65
Reconstrucciones ordenadas con RANSAC y predicciones	56.02
Reconstrucciones ordenadas aleatoriamente	51.55

En general se observa un menor error cuando se ordenan las cámaras basándose en el error de las cámaras frente a Panoptic. Esto tiene sentido debido a que la evaluación actual se está realizando efectivamente frente a las anotaciones de Panoptic. Se observa que utilizar un orden aleatorio puede resultar incluso beneficioso en vez de usar las comparaciones con las predicciones de DAD-3DNet. El menor error se obtiene para el método **RANSAC** con Panoptic, teniendo además un tiempo de ejecución menor.

Para obtener el error calculado para cada escena individual se utiliza la Ecuación 11. En la gráfica de la Figura 34 se muestran los \mathcal{E}_e de cada escena, e .

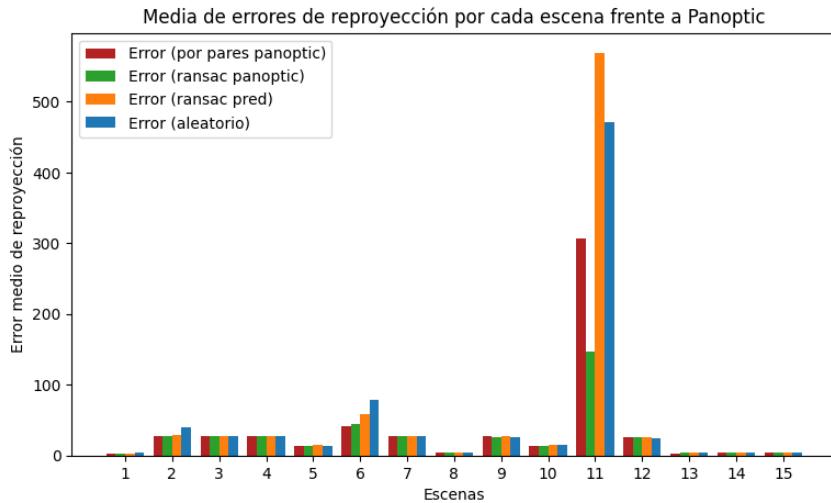


Figura 34: Error de reproyección para cada escena frente a Panoptic.

Se puede ver una tendencia parecida en las escenas, donde tanto el método aleatorio como el de predicciones suelen obtener errores mayores. Cabe destacar aun así que el error tiende a estabilizarse cuando se calcula la media de muchas combinaciones. Se puede ver una diferencia para la escena 2, gracias a que se evita utilizar una cámara que produce un gran error en la mayoría de las reconstrucciones. Sin embargo, esta cámara sí es utilizada con el método de ordenación aleatorio. Se debe recordar que el refinamiento de los puntos está condicionado por las propias predicciones de DAD-3DNet, por lo que cuando estas predicciones tienen errores mayores, igualmente ocurrirá en el refinamiento, como es el caso de la escena 11. Además, como estas evaluaciones se realizan únicamente sobre 27 *landmarks*, tampoco se puede observar completamente toda la influencia de la ordenación de cámaras.

Con el fin de entender exactamente de dónde provienen estos errores medios, se selecciona la escena 171026_pose1, correspondiente al identificador 2, con el sujeto 0, y la escena 170404_haggling_b1 con los sujetos 16 y 17 y los identificadores 10 y 11 respectivamente.

5.1. EVALUACIÓN DE LOS MÉTODOS DE ORDENACIÓN DE VISTAS

Estas escenas representan adecuadamente las diferencias entre los métodos en el valor del error según las reconstrucciones. Para obtener este cálculo se aplica la Ecuación 10.

En las gráficas mostradas en las Figuras 35, 36 y 37, se pueden ver estos errores según el número de cámaras usadas en la inicialización lineal y su refinamiento posterior. Se distingue la variación para una misma reconstrucción inicial al añadir más cámaras en el proceso de refinamiento según el orden específico obtenido por cada método. Se presentan los errores correspondientes a ordenar las cámaras por pares (color rojo), los errores con RANSAC y Panoptic (verde), los correspondientes a RANSAC comparado con las predicciones de DAD-3DNet (naranja) y los errores de las combinaciones aleatorias (azul). A medida que se añaden más cámaras, el error suele aumentar para los algoritmos que ordenan las cámaras mientras que suele reducirse para el método de ordenación aleatoria, llegando a un error parecido.

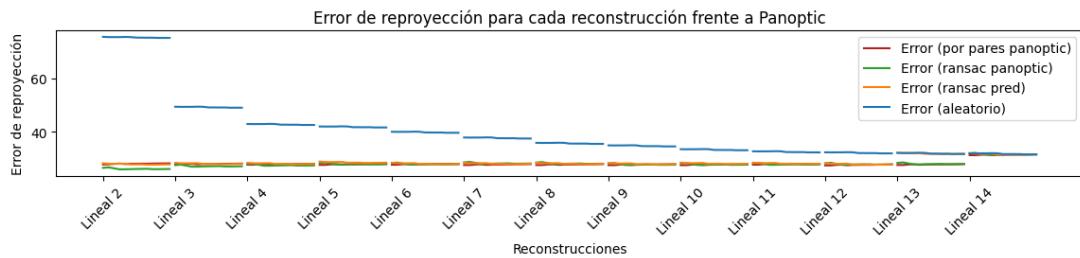


Figura 35: Error de reproyección para cada reconstrucción frente a Panoptic (escena 2).

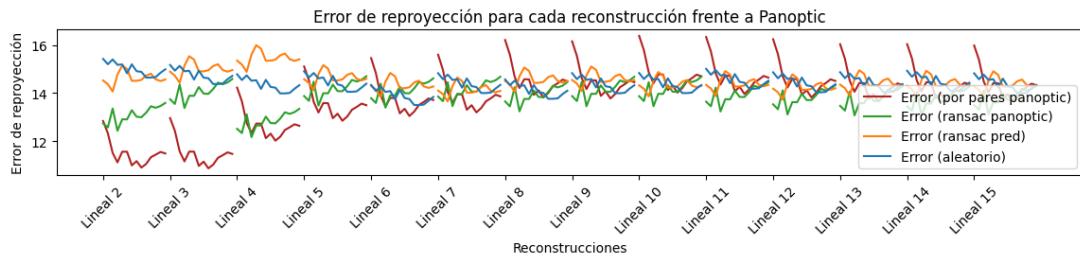


Figura 36: Error de reproyección para cada reconstrucción frente a Panoptic (escena 10).

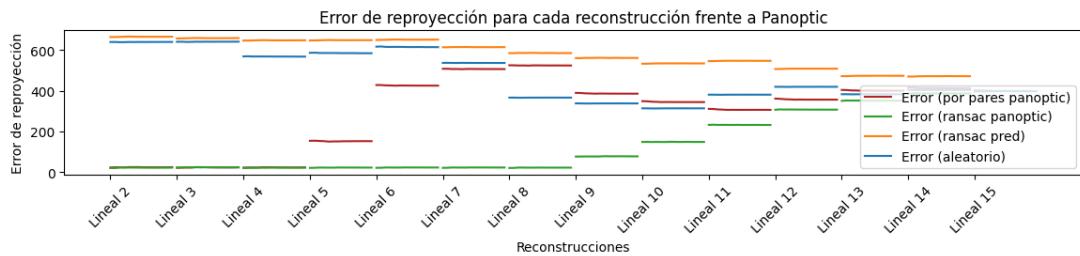


Figura 37: Error de reproyección para cada reconstrucción frente a Panoptic (escena 11).

Se observa claramente que el menor error es ofrecido en la mayoría de los casos al ordenar las cámaras según su error comparado con Panoptic y los mayores errores se encuentran en el algoritmo de ordenación aleatoria y el método de comparación con DAD-3DNet. Estas diferencias pueden ser de más de 50 píxeles como en la Figura 35, de apenas unos píxeles como en la Figura 36 o de cientos de píxeles (Figura 37). En el caso de la escena 2 este error está condicionado con el uso de la cámara número 13 en las reconstrucciones, cuyo

error en las predicciones es mayor. Esta cámara no es filtrada por el método aleatorio y se utiliza desde el principio en las reconstrucciones. En el caso de las escenas 10 y 11, el método por pares tiene errores menores al inicio. Sin embargo, a partir de añadir cinco cámaras se añade una predicción con valores más incorrectos. Esta situación consigue evitarse con el error mínimo y las iteraciones de [RANSAC](#).

Cabe decir que efectivamente influye el orden y el número de cámaras utilizado en cada caso, obteniendo errores menores cuando se usan menos cámaras en un orden determinado. Además, el error de las reconstrucciones se estabiliza cuando se usan más cámaras independientemente del orden, aunque obteniendo errores más altos que seleccionando menos cámaras. Es por ello que este valor de error es el utilizado durante los algoritmos de [RANSAC](#) como el error mínimo para considerar una cámara *inlier* en la iteración.

De igual manera, esta es la comparación frente a las anotaciones ofrecidas por Panoptic, por lo que pueden ser incorrectas en algunos casos y dar lugar a situaciones donde puntos peor posicionados tienen un menor error en comparación.

5.1.2. Evaluación de los métodos frente a las anotaciones manuales

A fin de obtener una evaluación más correcta y fiable, se utiliza el mismo planteamiento anterior pero obteniendo el error frente a las anotaciones realizadas manualmente. En la siguiente Tabla 4 se expone el error medio calculado con la Ecuación 12 (\mathcal{E}_m) de todas las reconstrucciones para todos los métodos.

Tabla 4: Error de las reconstrucciones para cada método frente a la anotación manual.

	\mathcal{E}_m (píxel)
Reconstrucciones ordenadas con pares de cámaras y Panoptic	24.82
Reconstrucciones ordenadas con RANSAC y Panoptic	14.77
Reconstrucciones ordenadas con RANSAC y predicciones	43.62
Reconstrucciones ordenadas aleatoriamente	39.17

Lo más destacable de esta Tabla 4 en comparación con la Tabla 3 es que el error medio de los conjuntos es mucho menor, lo cual es beneficioso dado que se trata del error frente al *ground-truth* real. Asimismo, se consiguen errores menores con los métodos que comparan frente a Panoptic que usando las predicciones o un orden aleatorio de cámaras, indicando que la tendencia se mantiene y existe una ventaja al ordenar las cámaras con estas anotaciones.

Si dividimos el error medio por escenas como en la Figura 38, se observan los distintos valores entre los fotogramas.

5.1. EVALUACIÓN DE LOS MÉTODOS DE ORDENACIÓN DE VISTAS

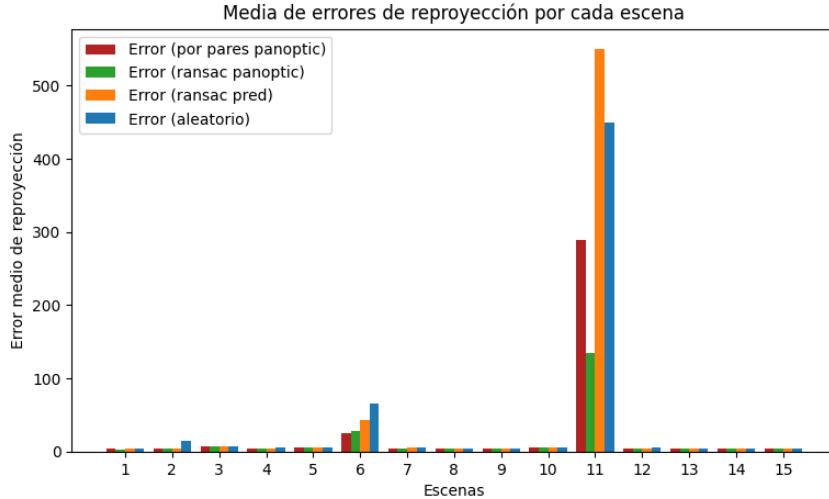


Figura 38: Error de reproyección para cada escena frente a las anotaciones manuales.

En la mayoría de los casos, los errores de los métodos se mantienen para las escenas aunque sí existen algunas diferencias, obteniendo menores errores igualmente con los métodos frente a Panoptic. El error en la escena 2 también muestra la característica comentada anteriormente. Para las escenas 6 y 11 se obtienen mayores errores debido a que las predicciones ofrecidas por DAD-3DNet contienen un error mayor frente al *ground-truth* manual que otras escenas. Es necesario destacar que en ocasiones el uso de los métodos de comparación por pares puede resultar en datos sesgados, ya que se tiene en cuenta únicamente el error ofrecido por dos cámaras para obtener una reconstrucción, por lo que algunas vistas con peores predicciones pueden alcanzar errores menores en combinación y ser posicionadas antes en la lista de ordenación. Asimismo, las anotaciones de Panoptic también pueden presentar errores y afectar a estos resultados significativamente.

Para comprender mejor el error de reproyección de cada una de las reconstrucciones de cada método, se van a desglosar los errores \mathcal{E}_r para las escenas 2, 4 y 6. Siguiendo la misma leyenda de colores, para las Figuras 39, 40 y 41 se pueden extraer algunas conclusiones. En la Figura 39 se ve claramente que la adición de la cámara 13 hace que el error aumente considerablemente en la reconstrucción. Al usar los algoritmos de ordenación se consigue que esta cámara sea posicionada al final, que es cuando se observa este incremento en el error. Sin embargo, al usar un método aleatorio, las predicciones ofrecidas por la cámara 13 son posicionadas al principio. En el resto de las gráficas, la tendencia es la misma, siendo el error de la ordenación aleatoria mayor en general, ya que las cámaras con peores predicciones no son posicionadas al final de la lista mientras que los métodos propuestos obtienen un menor error con menos cámaras.

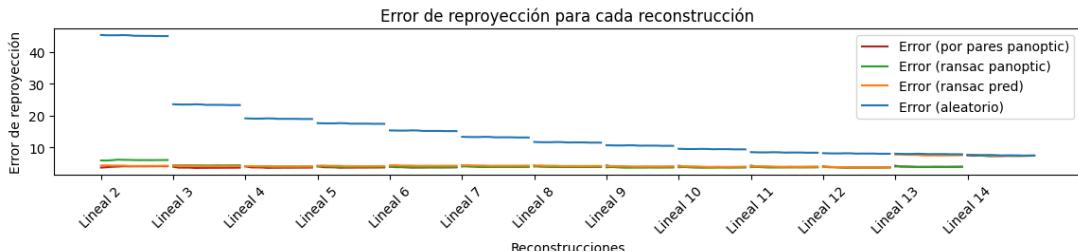


Figura 39: Error de reproyección para cada reconstrucción frente a las anotaciones manuales (escena 2).

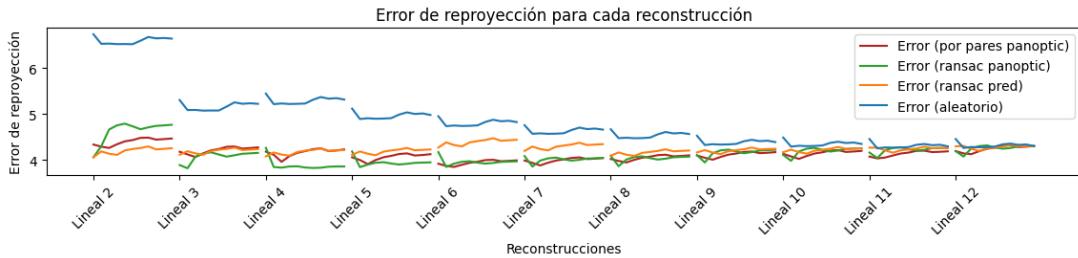


Figura 40: Error de reproyección para cada reconstrucción frente a las anotaciones manuales (escena 4).

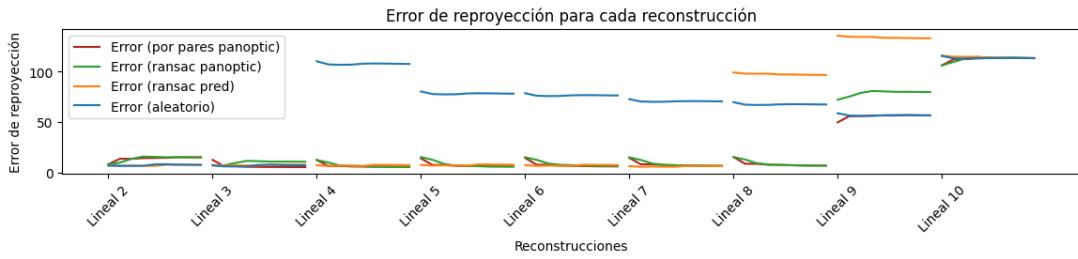


Figura 41: Error de reproyección para cada reconstrucción frente a las anotaciones manuales (escena 6).

Por tanto, se destaca que las mejores reconstrucciones se encuentran para un número de cámaras de entre 4 y 7 aproximadamente, que es donde se tienen los errores más pequeños para todos los métodos. El error obtenido aumenta de manera progresiva con el uso de más imágenes en la reconstrucción inicial. Igualmente, existe una diferencia al utilizar un número diferente de cámaras en el refinamiento con BA con la misma inicialización, aunque se mantiene la tendencia para todos los algoritmos, igual que pasaba en las comparaciones anteriores.

En conclusión, se selecciona el método que utiliza RANSAC evaluado con Panoptic durante la reconstrucción debido a que se consigue un error menor en general para las reconstrucciones. Aunque en algunos casos pueda obtener errores mayores, realmente apenas varía unos píxeles frente al resto de métodos, tal como se pudo ver en las anteriores gráficas. Este posible efecto no es perceptible en las reconstrucciones finales y el tiempo necesario para su ejecución es considerablemente menor frente a utilizar la comparación por pares si el fotograma actual contiene más de 7 cámaras.

Cabe decir que las conclusiones descritas a lo largo de este análisis sirvieron de base para la selección u omisión de las cámaras según la lista de *inliers* ofrecida por RANSAC. Concretamente, este estudio sirvió para desarrollar el procedimiento explicado en la sección 4.2.4, que debe seleccionar al menos 4 cámaras para realizar la reconstrucción final. De esta manera, se ha tenido en cuenta el número de cámaras cuando el error es menor y las cámaras *inliers* asociadas a estas reconstrucciones con el algoritmo de RANSAC con Panoptic desarrollado.

5.2. Evaluación de la reconstrucción final

Tras la decisión de emplear el método de RANSAC en combinación con las anotaciones de Panoptic según la evaluación anterior, se sigue el procedimiento explicado en la sección 4.2.4, generando una única reconstrucción por cada fotograma de una escena. Esta

5.2. EVALUACIÓN DE LA RECONSTRUCCIÓN FINAL

reconstrucción utilizará el mismo número de cámaras y en el mismo orden tanto para la inicialización con [DLT](#) como en el método de refinamiento con [BA](#).

En este apartado, se analiza en detalle la calidad del modelo 3D en comparación con el uso de todas las vistas ordenadas de forma aleatoria, ofreciendo una evaluación similar a la que se realizó en la sección anterior.

5.2.1. Evaluación de la reconstrucción final frente a Panoptic

Primeramente, en la Figura 42 se muestran los errores de cada escena (\mathcal{E}_e) de la reconstrucción final frente a la reconstrucción aleatoria según el error con las anotaciones de Panoptic.

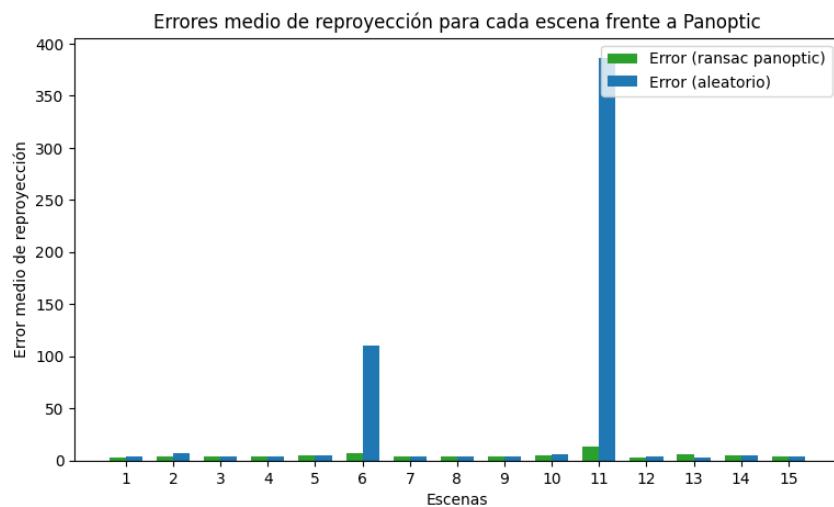


Figura 42: Error de reproyección de la reconstrucción final para cada escena frente a Panoptic.

Se puede ver que en general el error suele ser mayor al usar el orden aleatorio de las cámaras, aunque la diferencia normalmente no es grande. Para las escenas 6 y 11 se puede ver también que el error en píxeles es mucho mayor en comparación con otras escenas. Esto ocurre debido a que las predicciones de DAD-3DNet contienen un error mayor para algunas de ellas. El error para los puntos faciales se reduce en consecuencia con el método de ordenación aunque la reconstrucción se vea afectada también.

Este error de reproyección puede ser analizado para cada vista de una escena concreta usando la Ecuación 9. Se seleccionan las escenas 11 y 12 donde se pueden observar claramente los distintos valores entre imágenes. Estos errores (\mathcal{E}_c) se muestran para cada cámara en las Figuras 43 y 44.

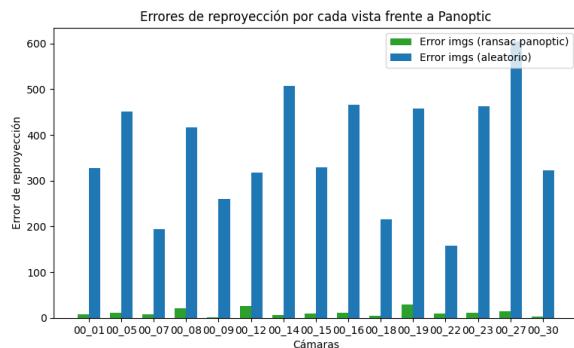


Figura 43: Error de reproyección de la reconstrucción final para cada cámara frente a Panoptic (escena 11).

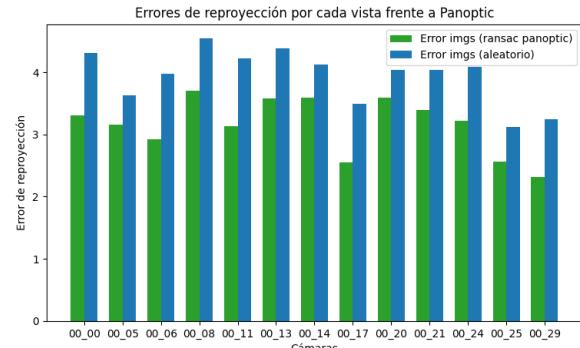


Figura 44: Error de reproyección de la reconstrucción final para cada cámara frente a Panoptic (escena 12).

Se aprecia que la gran diferencia en la escena 11 entre ambos métodos. Para la escena 12 también se mantiene una menor tasa de error con RANSAC y Panoptic pero apenas hay una diferencia de 2 píxeles como máximo, por lo que no es un error significativo. Continuando con el análisis de la escena 11, se puede estudiar el error por *landmarks* si se aplica la Ecuación 7. Este error (\mathcal{E}_{le}) de los puntos característicos de la escena 11 se presenta en la Figura 45 con el fin de comprender cuáles son los puntos que se detectan con mayor diferencia.

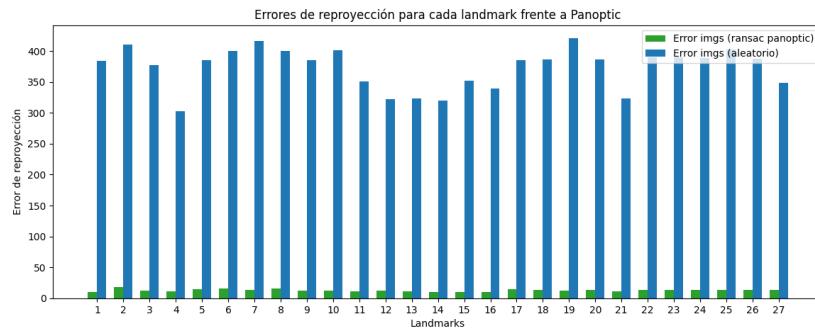


Figura 45: Error de reproyección para cada *landmark* frente a Panoptic (escena 11).

Por ejemplo, en el caso de esta escena concreta, los puntos característicos con mayor error son el 2, 7 y 19, correspondientes a un punto de la ceja, el ojo derecho y la nariz (ver Figura 18).

5.2.2. Evaluación de la reconstrucción final frente a las anotaciones manuales

Para comenzar la evaluación comparativa frente a las anotaciones realizadas manualmente, se muestra igualmente el error calculado para cada escena según la Ecuación 11 utilizando RANSAC y Panoptic y la ordenación aleatoria (Figura 46). Este error también es inferior en comparación con las anotaciones de Panoptic vistas en la Figura 42. Además, se destaca un patrón similar en la tendencia de los resultados para todas las escenas.

5.2. EVALUACIÓN DE LA RECONSTRUCCIÓN FINAL

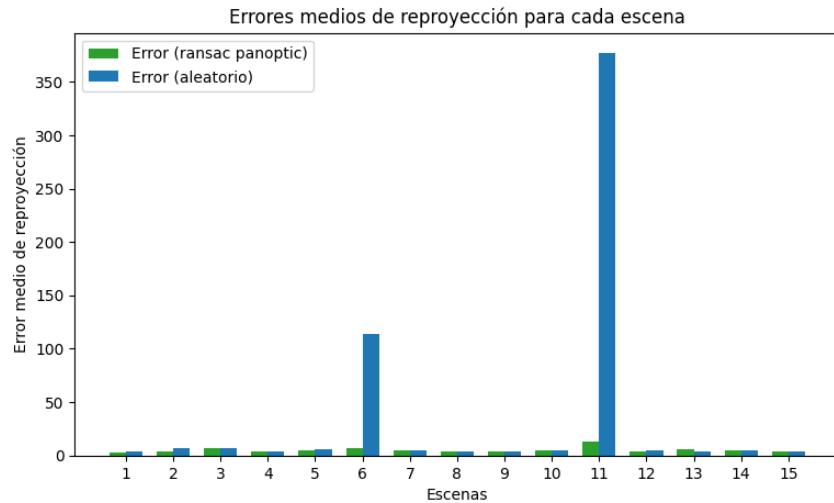


Figura 46: Error de reproyección de la reconstrucción final para cada escena frente a las anotaciones manuales.

En la Figura 46 anterior, la mayor discrepancia de error se observa en las escenas 6 y 11, ya comentadas anteriormente.

En este apartado se van a analizar las escenas 2 y 5 para observar más claramente los errores cuando las diferencias no son tan grandes. En las Figuras 47 y 48 se obtienen los errores de reproyección para estas escenas por cada vista.

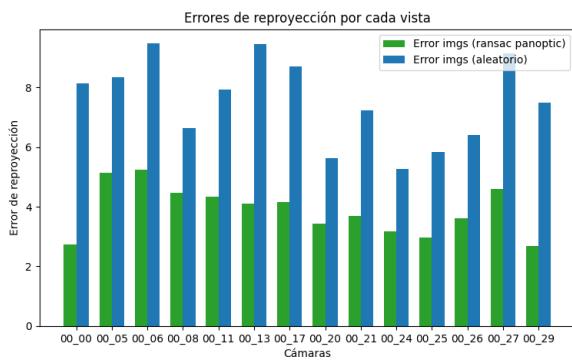


Figura 47: Error de reproyección de la reconstrucción final para cada cámara frente a las anotaciones manuales (escena 2).

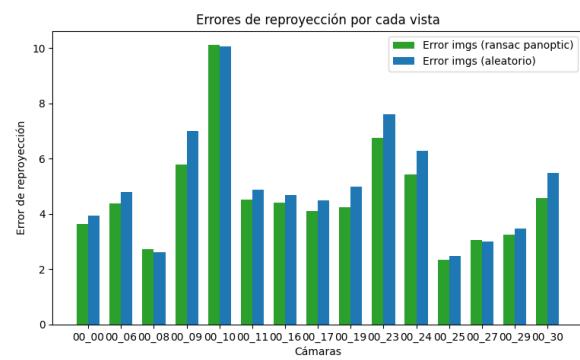


Figura 48: Error de reproyección de la reconstrucción final para cada cámara frente a las anotaciones manuales (escena 5).

Para la escena 2 se observan mayores diferencias de errores en todas las vistas en comparación con la 5, aunque siendo igualmente menor para RANSAC en casi todos los casos. Si mostramos el error de cada punto característico para ambas, se obtienen las gráficas de las Figuras 49 y 50.

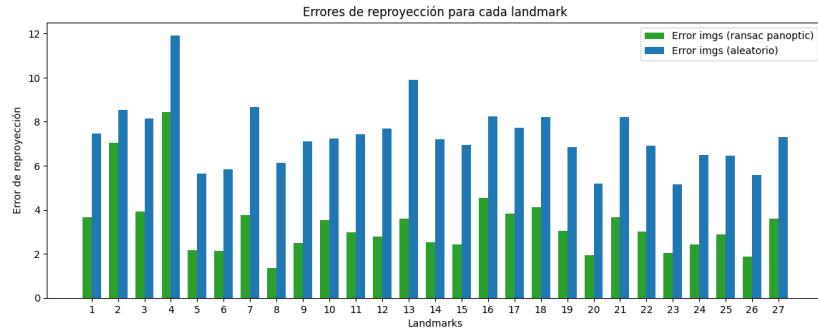


Figura 49: Error de reproyección de la reconstrucción final para cada *landmark* frente a las anotaciones manuales (escena 2).

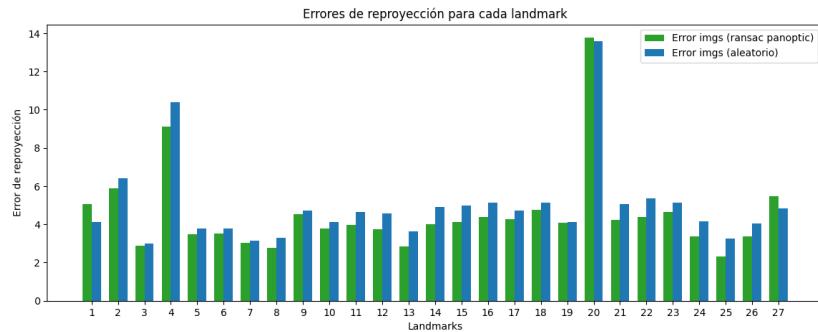


Figura 50: Error de reproducción de la reconstrucción final para cada *landmark* frente a las anotaciones manuales (escena 5).

Viendo los resultados para los *landmarks*, se puede concluir que el punto característico número 4, correspondiente a la ceja izquierda (Figura 18), contiene un error grande en ambos casos. En la escena 5 se obtiene también un gran error para el punto 20, situado en la nariz.

A continuación, en las Figuras 51 y 52, se muestran varias perspectivas de las escenas 2 y 5, que proporcionan una visualización más clara de estas disparidades. Los puntos verdes corresponden al algoritmo de **RANSAC** seleccionado, los puntos azules a la ordenación aleatoria y los blancos al *ground-truth* manual.

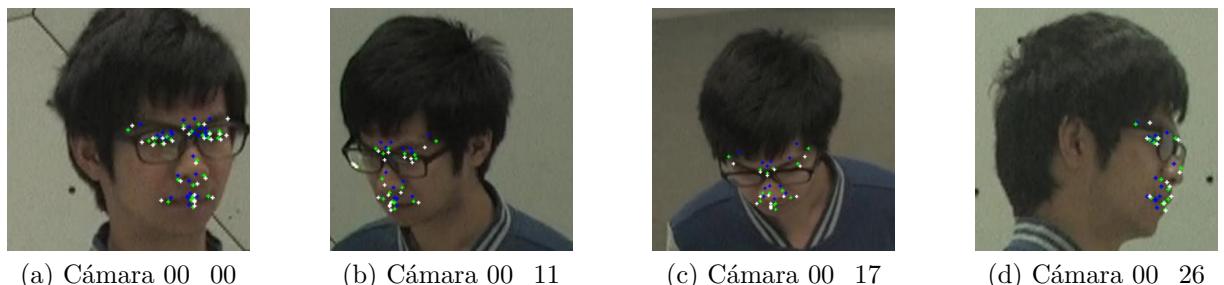


Figura 51: Comparación de *landmarks* (escena 2).

5.2. EVALUACIÓN DE LA RECONSTRUCCIÓN FINAL

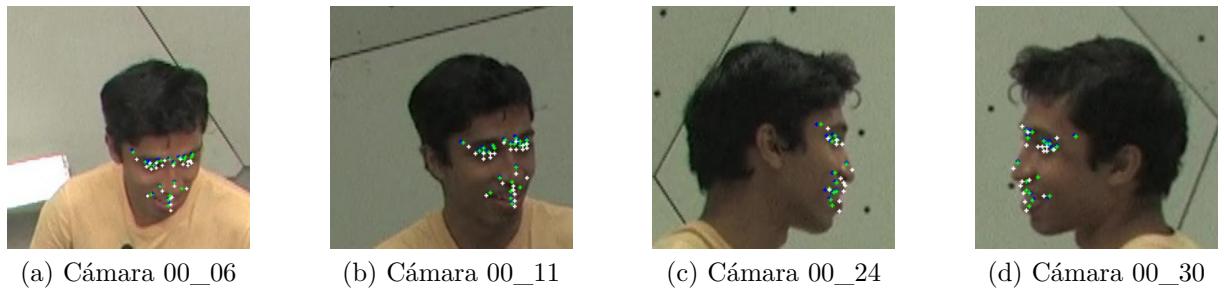


Figura 52: Comparación de *landmarks* (escena 5).

En la Figura 51 se ven grandes diferencias en la posición de los *landmarks*, pudiendo resaltar la importancia de seleccionar y ordenar las cámaras. Por otra parte, en el caso de la Figura 52 existe una menor diferencia en los puntos verdes y azules. Igualmente, cuando existen errores en los *landmarks* de estas imágenes, ambos conjuntos suelen desviarse de manera similar, aunque aquellos obtenidos por RANSAC tienden a estar más cerca de los puntos del *ground-truth*, lo que conduce a errores menores.

5.2.3. Evaluación visual frente a las predicciones de DAD-3DNet

En las secciones anteriores, se ha podido estudiar el impacto de seleccionar las cámaras, evaluando los resultados frente a un conjunto de 27 *landmarks* de la cara. No obstante, las predicciones finales realmente ofrecen un total de 5023 puntos, por lo que quedan muchos *landmarks* sin poder evaluarse correctamente. Considerando esto, se desea mostrar las reprojeciones obtenidas de todos los puntos formando una malla para poder contrastar su forma con las predicciones originales predichas por el modelo de DAD-3DNet.

Siguiendo con la escena 2, donde se observaba una diferencia en el error, se presentan las mallas que forman los puntos refinados en comparación con los resultados originales ofrecidos por DAD-3DNet en la Figura 53. Estas imágenes se muestran además la cámara 13 en la tercera columna, que es aquella que aumentaba considerablemente el error en la reconstrucción. Se destaca su predicción errónea por DAD-3DNet y cómo este caso es solventado por la reconstrucción multivista procesada.



Figura 53: En la primera fila se muestran las reprojeciones refinadas utilizando el algoritmo RANSAC y Panoptic (escena 2). En la segunda fila se presentan las reprojeciones predichas por DAD-3DNet (escena 2).

En la Figura 54 se comparan las mallas 3D de la escena 12. Se puede ver la mejora en los ajustes a la forma de la cabeza para todas las imágenes seleccionadas de esta escena.



Figura 54: En la primera fila se muestran las reproyecciones refinadas utilizando el algoritmo RANSAC y Panoptic (escena 12). En la segunda fila se presentan las reproyecciones predichas por DAD-3DNet (escena 12).

Entre todas las figuras presentadas se debe tener en cuenta que las reproyecciones del modelo refinado provienen de un único modelo 3D, mientras que la malla predicha por DAD-3DNet se calcula de manera independiente en cada vista. Por lo tanto, el ajuste de **RANSAC** permite realizar una reconstrucción utilizando simultáneamente la información de todas las vistas. Se observa una mejora al considerar la predicción de varias cámaras, especialmente resolviendo algunas de las desviaciones en las partes periféricas de la cabeza.

Sin embargo, si las predicciones de partida de DAD-3DNet se desvían demasiado de la verdadera posición de los *landmarks* la reconstrucción obtenida por el método propuesto puede perjudicar algunas de las vistas. Este es el caso para la escena 11, mostrado en la Figura 55, que ya presentaba un gran error en las predicciones originales.



Figura 55: En la primera fila se muestran las reproyecciones refinadas utilizando el algoritmo RANSAC y Panoptic (escena 11). En la segunda fila se presentan las reproyecciones predichas por DAD-3DNet (escena 11).

5.2. EVALUACIÓN DE LA RECONSTRUCCIÓN FINAL

En la anterior escena 11, los *landmarks* faciales estaban mejor situados con el método de ordenación, pero se observa una peor disposición general de la malla. Para el caso de la escena 6 se ajustan algunas de las predicciones pero pueden surgir pequeñas desviaciones que no estaban presentes en la predicción individual de las imágenes (ver Figura 56).



Figura 56: En la primera fila se muestran las reprojeciones refinadas utilizando el algoritmo RANSAC y Panoptic (escena 6). En la segunda fila se presentan las reprojeciones predichas por DAD-3DNet (escena 6).

Esto puede ocurrir debido también a desviaciones en la anotación ofrecida por CMU Panoptic (ver Figura 13), ya que el método de ordenación utiliza estos puntos para seleccionar las cámaras para la reconstrucción final.

En esta parte se destaca la importancia de comprobar los resultados con todas las detecciones siendo la ordenación y selección de estas predicciones un punto clave a la hora de realizar las reconstrucciones. Sin embargo, debido a que no se utiliza una gran cantidad de *landmarks* de Panoptic y estos pueden ser incorrectos en algunos casos, es probable que alguna predicción errónea pueda ser incluida en la reconstrucción. A pesar de esto, si la mayoría de las vistas elegidas poseen buenas predicciones, este problema puede llegar a solventarse en las imágenes donde las detecciones son de peor calidad porque la pose de la persona es complicada.

5.2.4. Visualización de la reconstrucción 3D final

Con el fin de visualizar el modelo facial completo, se utiliza la librería VTK, pudiendo comprobar la calidad de los puntos en el espacio de forma tridimensional. En la Figura 57 se enseñan varios ejemplos de las imágenes del sujeto 0 en la escena 171204_pose5.

Se observa que el modelo 3D sigue fielmente la expresión de la cara, viéndose claramente en el movimiento de los labios. No obstante, en algunos de los *frames* que poseen un error mayor puede haber desviaciones en la forma de la cabeza a pesar de que sí extraigan bien la expresión de la cara. Esta influencia se ve, por ejemplo, en los modelos extraídos de la siguiente Figura 58, que corresponde también al sujeto de la escena 1 para fotogramas parecidos a los anteriores.

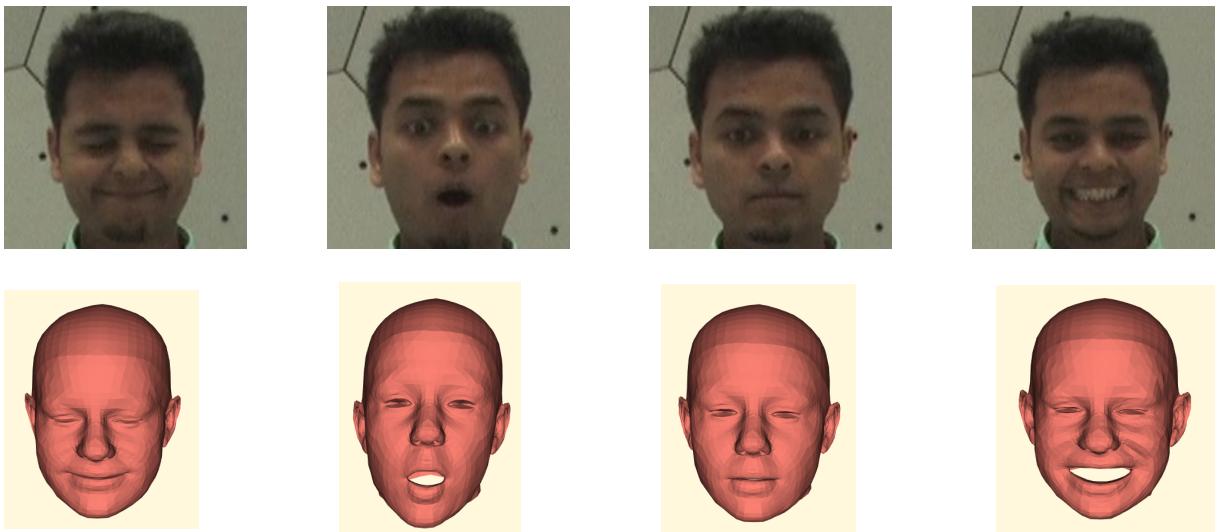


Figura 57: Visualización texturizada (escena 1).

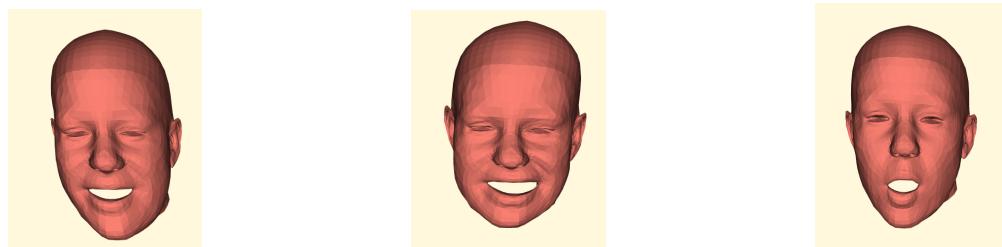


Figura 58: Error de las reconstrucciones 3D según el modelo texturizado (escena 1).

Aplicando esta texturización a algunas imágenes de las escenas 3, 9, 10 y 15 de evaluación, en la Figura 59 se puede apreciar cómo la reconstrucción se ajusta a la forma característica de cada rostro y su expresión facial.

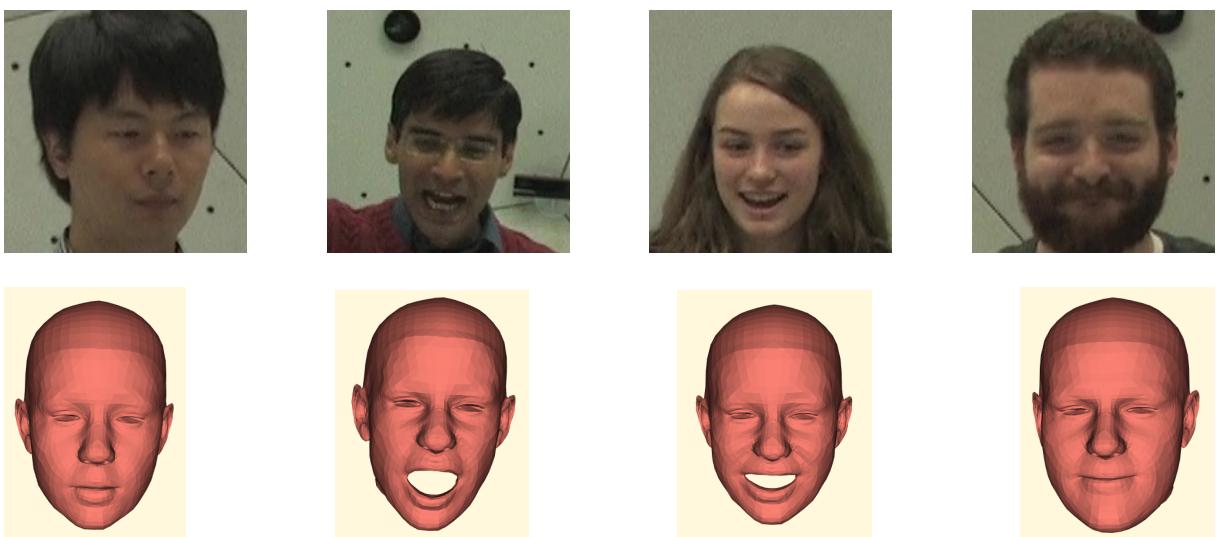


Figura 59: Visualización texturizada (escenas 3, 9, 10 y 15).

Mediante esta herramienta de visualización, se consigue entender de forma más completa los resultados de la reconstrucción facial, pudiendo comparar los resultados entre sí y extraer conclusiones sobre las mejoras y fallos asociados. Se destaca que la forma, tamaño

y expresión de la cabeza suele estar en concordancia con las imágenes aunque sí pueden aparecer ciertas deformaciones en algunos fotogramas.

5.3. Resultados de los detectores de *landmarks* con las nuevas anotaciones

Tras realizar los entrenamientos especificados en la sección 4.4, se obtienen dos redes neuronales capaces de detectar puntos característicos faciales. Las gráficas de la pérdida en ambos casos se presentan en la Figura 60. En primer lugar para el entrenamiento que utiliza la Ecuación 13 como función de pérdida y después para el entrenamiento que emplea la Ecuación 16 según la visibilidad de los puntos. Cabe decir que debido a esta característica, no se pueden comparar las pérdidas del entrenamiento, ya que los cálculos son distintos. Sin embargo, la pérdida utilizada en la parte validación fue obtenida de igual manera, por lo que estas sí son comparables.

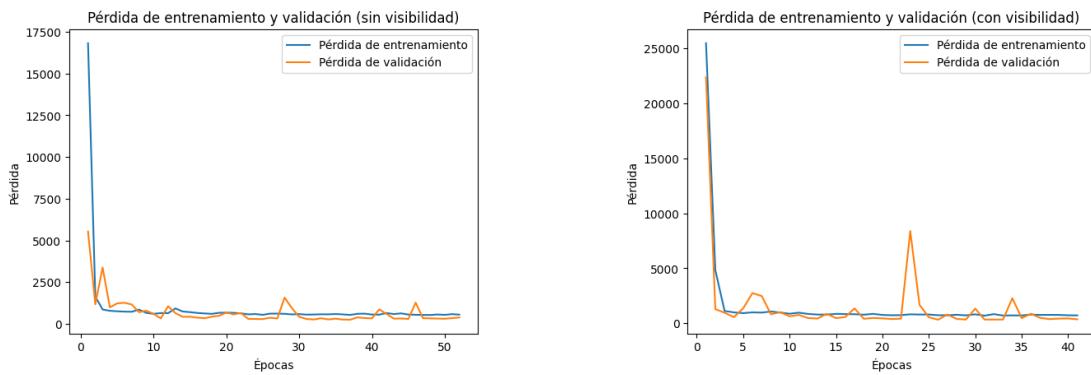


Figura 60: Pérdidas de entrenamiento y validación.

Al observar la Figura 60, se nota en ambos escenarios que la pérdida es elevada en las primeras épocas pero disminuye drásticamente al comenzar el entrenamiento. Una vez que este valor aumenta de forma consecutiva durante 15 épocas, el procesamiento se detiene automáticamente, utilizando esta medida para prevenir el sobreentrenamiento. En el primer caso, se observa un total de 52 épocas con un valor de la función de pérdida final de 380.29, mientras que en el segundo caso se realizan un total de 41 épocas con una diferencia final de 353.33. Si se calcula el RMSE calculando la raíz cuadrada sobre estos valores, podemos obtener la medida en píxeles, resultando un error de reprojeción de 19.50 y 18.80 respectivamente. No obstante, a pesar de que el segundo entrenamiento concluye con una pérdida final mayor, se obtiene una pérdida mínima inferior en el primero.

Tras calcular la inferencia sobre las imágenes de test, se calcula el error medio obtenido entre estas predicciones y las anotaciones para ambos modelos. En el caso de la red neuronal entrenada con una pérdida equivalente para todos los puntos se obtiene un error de 304.42, es decir, 17.44 píxeles, a diferencia del modelo con pérdida basada en la visibilidad que consigue un error medio cuadrático de 346.84 (18.62 píxeles). Por tanto, el error sigue siendo inferior para el modelo que utiliza la información de posición de todos los *landmarks* por igual.

Para facilitar la comprensión de los resultados obtenidos, se presentan las inferencias para algunas imágenes del conjunto de test en la Figura 61. En color verde se muestran los

resultados del entrenamiento que no tienen en cuenta la visibilidad, en rojo los puntos predichos por la red neuronal con la función de pérdida según la visibilidad, en color azul los puntos ofrecidos por DAD-3DNet y en blanco los correspondientes a las reconstrucciones 3D refinadas.

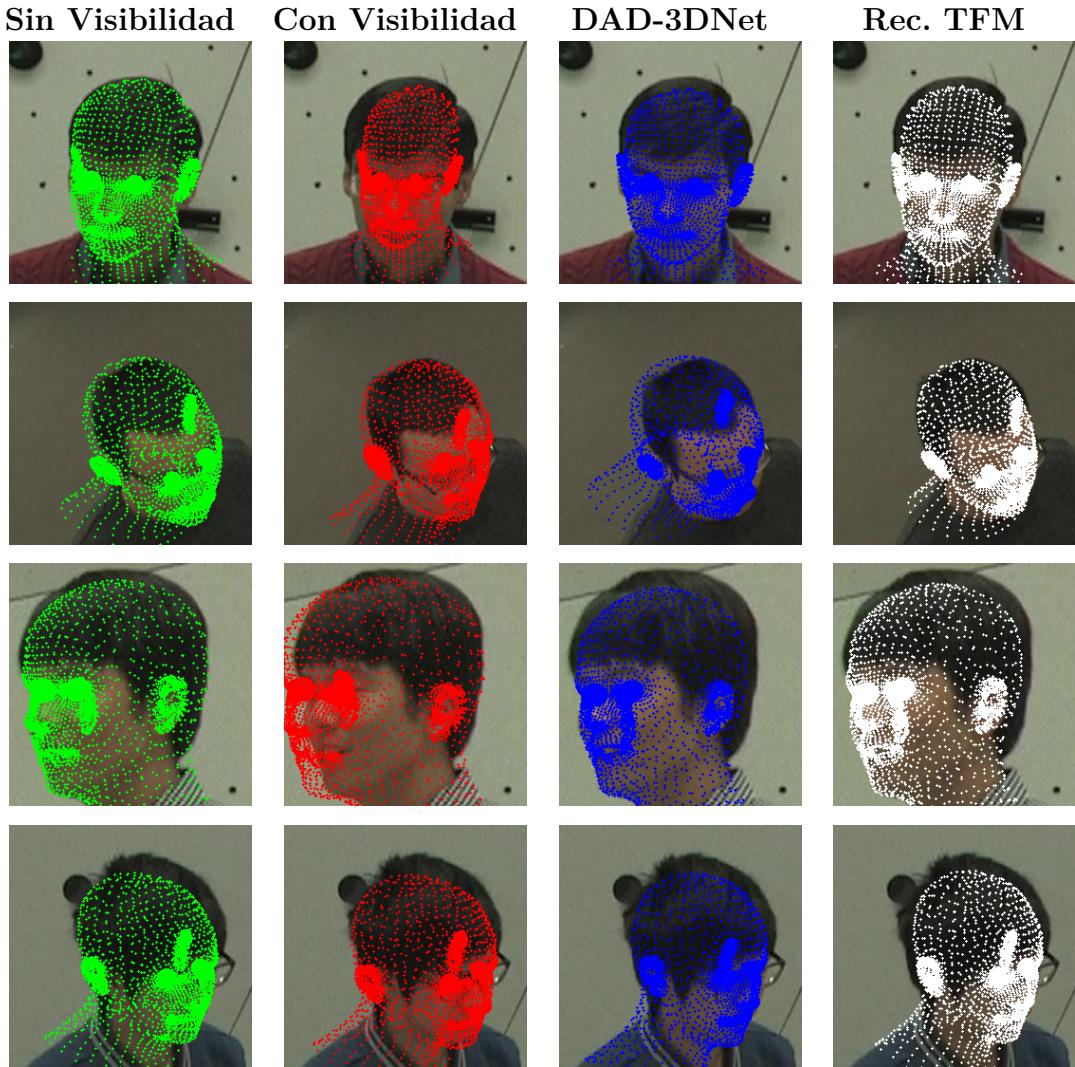


Figura 61: Ejemplo de las predicciones obtenidas.

Se destaca que el detector de *landmarks* que no tiene en cuenta la visibilidad de los puntos suele conseguir mejores resultados en la localización de los puntos sobre la cabeza en comparación con el detector con visibilidad. Además, el tamaño general de la cabeza se ajusta de mejor manera que en los ejemplos presentados. Por otra parte, se aprecia que DAD-3DNet y las reconstrucciones refinadas derivadas de esta misma red neuronal muestran una notable consistencia en los resultados.

En la Figura 62 se muestran algunos ejemplos donde los detectores de *landmarks* tienen cierto problema al realizar las predicciones sobre las imágenes. En estos se puede ver que los modelos son capaces de orientar de forma más o menos correcta los puntos de la cabeza, sin embargo, no consiguen ajustarse adecuadamente a la forma del rostro para ninguno de los detectores. A pesar de que los propios entrenamientos de los detectores se han llevado a cabo con imágenes similares de Panoptic, tanto DAD-3DNet como las reconstrucciones realizadas en este TFM consiguen mejores aproximaciones. Además, esta afirmación se respalda por la inversión de recursos y el estudio dedicado en la construcción

5.3. RESULTADOS DE LOS DETECTORES DE *LANDMARKS* CON LAS NUEVAS ANOTACIONES

de la red DAD-3DNet, en comparación con el modelo entrenado, el cual posee una menor complejidad y número de parámetros. Asimismo, se debe tener en cuenta que el entrenamiento de DAD-3DNet fue realizado utilizando un *ground-truth* elaborado manualmente, mientras que las etiquetas de este entrenamiento presentan errores en algunos casos, lo que puede resultar en un rendimiento inferior.

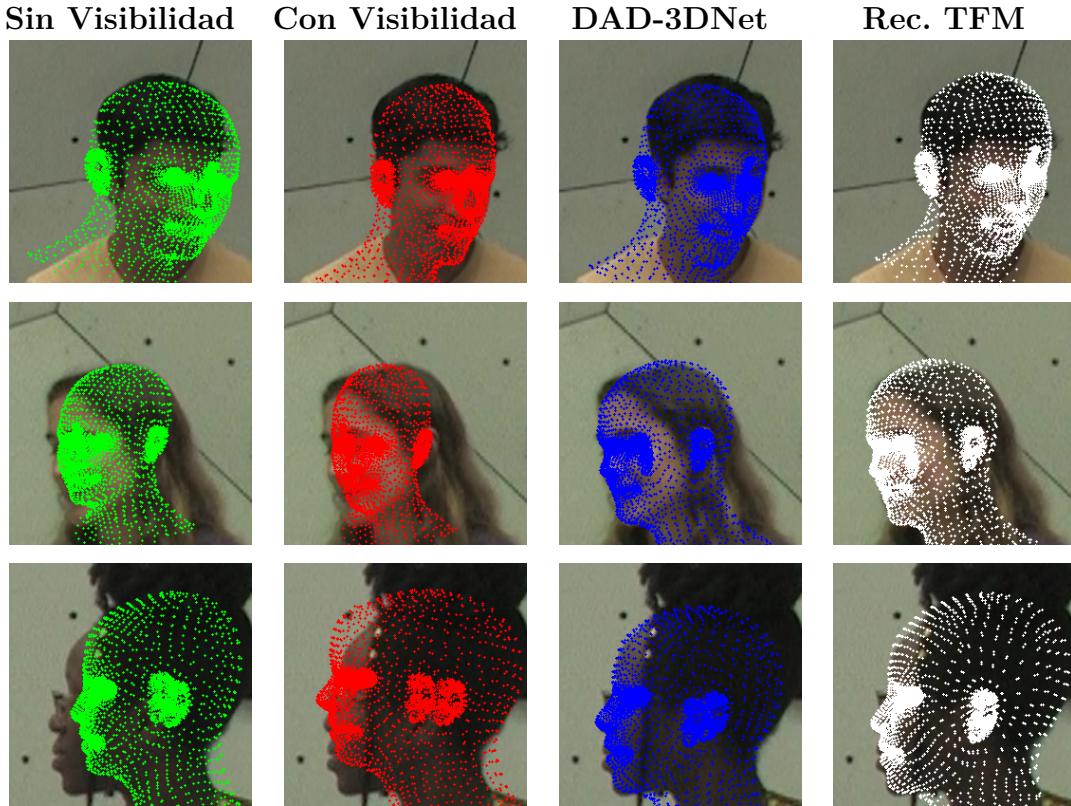


Figura 62: Ejemplo de ciertos errores sobre las predicciones obtenidas.

Comparando los resultados de los entrenamientos ejecutados de forma general, se puede ver que la inclusión de la información de visibilidad en este proceso no ofrece una ventaja en comparación con usar una pérdida equitativa a todos los puntos. Aunque podría asumirse en un principio que la red con la pérdida basada en visibilidad ajusta mejor la parte observable de la imagen, realmente esta premisa perjudica a la calidad de los resultados. Observando las detecciones obtenidas por ambas redes neuronales, se concluye que el modelo que no tiene en cuenta la visibilidad ofrece una calidad mejor de los resultados. Esto se valida también por la menor pérdida obtenida durante la fase de validación y test, lo que se traduce en un mejor ajuste de los 5023 puntos. Sin embargo, no alcanza el nivel de madurez necesario para ser un detector de *landmarks* preciso. No obstante, estos resultados son dependientes del proceso de entrenamiento y el número de épocas.

Por tanto, es evidente que para abordar estos fallos se necesitaría de un entrenamiento más riguroso, con un mayor número de muestras, redes neuronales más complejas y un mejor filtrado de anotaciones incorrectas para alcanzar la calidad de las predicciones de DAD-3DNet. Aun así, es importante destacar el enfoque utilizado considerando la oclusión de los puntos y el uso de las escenas de Panoptic anotadas, lo que arroja resultados preliminares sobre su utilidad en el posterior desarrollo de nuevos modelos.

A pesar de esto, la tendencia en las predicciones por ambos modelos es la correcta y se observa que si la pose del sujeto no es comprometida, se pueden obtener algunas buenas

detecciones de los puntos de referencia. Se puede afirmar que las nuevas anotaciones se pueden utilizar en entrenamientos de detectores de *landmarks*, validando su utilidad y potenciando el desarrollo de nuevos sistemas de análisis facial.

Capítulo 6

Conclusiones

Como se ha explicado a lo largo del proyecto, los modelos faciales 3D pueden ser utilizados en una amplia variedad de aplicaciones, abarcando desde el ámbito del entretenimiento, la creación de modelos virtuales, el análisis facial para la realización de diagnósticos y otros enfoques basados en reconocimiento facial para la seguridad. Sin embargo, para cualquiera de estas áreas, existe una falta de bases de datos completas de las reconstrucciones 3D que puedan ser utilizadas en el desarrollo de nuevas aplicaciones con modelos más precisos.

El trabajo realizado se ha desarrollado teniendo en cuenta dos propósitos diferentes interrelacionados entre sí. En primer lugar, llevar a cabo una anotación de las reconstrucciones faciales de la base de datos Panoptic aplicando un refinamiento a las predicciones de DAD-3DNet y, por otra parte, la utilización de estos nuevos datos para un entrenamiento teniendo en cuenta las occlusiones de los puntos.

Mediante el desarrollo del método para la obtención de las reconstrucciones 3D con información multivista, se ha remarcado la importancia en usar datos de calidad que puedan garantizar la precisión necesaria para este tipo de modelos tridimensionales. Uno de los puntos fundamentales de este desarrollo ha sido el primer preprocesamiento realizado sobre las imágenes de Panoptic. En este caso, se pudo conseguir un filtrado preliminar para evitar realizar predicciones sobre imágenes de la cámara donde no se puede ver correctamente el rostro del sujeto.

Asimismo, tras obtener las predicciones de DAD-3DNet sobre estas imágenes, se consiguió comprender e implementar el algoritmo para realizar reconstrucciones lineales con diferente número de cámaras, así como aplicar el refinamiento con **BA** mediante la instalación de una librería externa.

De igual manera, se exploraron distintas estrategias para poder obtener reconstrucciones faciales precisas, las cuales fueron comparadas con las anotaciones de Panoptic y los puntos característicos anotados manualmente en las escenas de evaluación. En este contexto, se estudió cómo el uso de una disposición o un número diferente de cámaras puede afectar en el error final de la reconstrucción total. Para realizar este análisis, se propusieron 3 métodos comparando la calidad de sus resultados frente a una ordenación aleatoria de las cámaras. De esta manera, se comprobó que los dos métodos basados en ordenar las cámaras según su error frente a Panoptic ofrecían los resultados globales con menor error. No obstante, el método por pares sin **RANSAC** presenta una gran desventaja si se desea procesar un gran número de reconstrucciones y el número de cámaras a analizar aumenta progresivamente. Su coste computacional junto con el tiempo necesario, así como un mayor

error para algunos casos, resulta en que el método de **RANSAC** con Panoptic sea más adecuado para esta tarea, a pesar de que pueda omitir algunas comparaciones durante las iteraciones. Con este algoritmo se pudieron obtener datos de mejor calidad limitando el número de iteraciones a 50. Es por ello que se seleccionó este método para llevar a cabo las anotaciones finales. Igualmente, este procedimiento también puede ajustarse y reducir el tiempo de ejecución realizando menos iteraciones, aunque aumentando el error en consecuencia. Sobre el algoritmo de **RANSAC** que comprueba los resultados frente a las predicciones de DAD-3DNet se destaca su comparación frente a los 5023 *landmarks*, por lo que tiene en cuenta la cabeza entera en la ordenación de las cámaras. A pesar de ello, el error obtenido para los 27 puntos anotados manualmente es superior y el tiempo necesario de ejecución no se ve reducido. Por otra parte, se pudo ver que al ordenar las cámaras de forma aleatoria, el error de reproyección aumentó en comparación con utilizar uno de los algoritmos de ordenación y selección mencionados. Principalmente, estos errores se vuelven más significativos cuando se utiliza un número muy reducido de cámaras. Aun así, si en la reconstrucción final se utilizan el total de las cámaras, el error final suele estabilizarse para todos los métodos.

Es importante señalar que, si bien la organización y selección de las cámaras que ofrecen las mejores predicciones sobre los puntos de referencia de la cabeza puede influir en los resultados, en términos generales, si se busca un procesamiento realmente rápido, se podrían emplear todas las cámaras en la reconstrucción para obtener resultados válidos.

Asimismo, se debe mencionar también la visualización ofrecida por VTK mediante el procesamiento de polígonos de los puntos para crear texturas. De esta manera, se ha podido visualizar en un espacio 3D algunos fallos que podrían pasar desapercibidos en la evaluación realizada. Además, esta herramienta es realmente eficaz para explorar los resultados a lo largo de una secuencia de fotogramas.

Tras este primer procesamiento, se utilizaron las anotaciones realizadas sobre un conjunto de secuencias de Panoptic para poder llevar a cabo los entrenamientos para obtener modelos detectores de *landmarks*. Para ello, se reproyectaron los resultados de la reconstrucción 3D del espacio XYZ sobre cada una de las imágenes utilizadas, obteniendo los nuevos puntos característicos refinados en 2D. A pesar de que inicialmente se había considerado la posibilidad de volver a entrenar DAD-3DNet, se concluyó que la complejidad de la red neuronal y los datos superaban los límites de alcance de este proyecto. Por ello, para intentar obtener los mejores resultados posibles usando los nuevos datos, se propuso el uso de dos pérdidas distintas.

En primer lugar, se ejecutó el entrenamiento de la red neuronal con una pérdida en la cual todos los puntos de referencia faciales tenían el mismo peso en el cálculo del error. Luego, como segundo enfoque, se configuró el entrenamiento para considerar la visibilidad de estos puntos. Para poder obtener esta característica inherente a la orientación de la cabeza en las imágenes, se aplicaron procedimientos trigonométricos sobre los puntos 3D de la reconstrucción, filtrando aquellos *landmarks* visibles según la orientación de la cabeza en la imagen 2D. De esta manera, se pudo aplicar una pérdida personalizada centrada en dar más valor a los *landmarks* visibles en la imagen, con el fin de aumentar la calidad del ajuste global según la parte visible de la imagen. Además, para poder realizar este entrenamiento, se debió establecer un protocolo de actualización de los datos para obtener esta propiedad de visibilidad del *batch* correspondiente durante el entrenamiento.

Mediante la evaluación de dichas técnicas de entrenamiento se apreciaron ciertas características en los resultados. Por una parte, el error de validación y test resulta ser menor para

el entrenamiento con la pérdida equivalente para todos los puntos. Este detector consigue mejores resultados que el modelo entrenado con la característica de visibilidad, ajustando mejor los *landmarks* faciales en las imágenes. A pesar de que los detectores todavía presentan algunos errores, los resultados tienen una tendencia adecuada. Con el uso de un modelo sencillo y las nuevas anotaciones, se ha conseguido una primera aproximación en el desarrollo de redes neuronales de detección, consiguiendo predicciones de acuerdo con las expectativas y cumpliendo con el objetivo de validar las nuevas anotaciones. Cabe decir que utilizando la base de datos anotada completa, junto con redes neuronales más complejas, se podrían realizar entrenamientos de mejor calidad y conseguir el desarrollo de aplicaciones más precisas.

Por tanto, un factor limitante de este trabajo ha sido esencialmente la calidad de las anotaciones utilizadas para evaluar el refinamiento de las reconstrucciones 3D. Tanto las predicciones de DAD-3DNet como las anotaciones de la base de datos de Panoptic, afectan al resultado final ya que el *ground-truth* utilizado no fue realizado manualmente. A pesar de esto, los resultados obtenidos son prometedores y se encuentran en el objetivo principal del proyecto que es poder crear nuevas bases de datos más precisas para la generación de modelos 3D.

A modo de conclusión, es importante resaltar que los objetivos del proyecto han sido logrados gracias a la aplicación de la metodología propuesta. De esta manera, se ha conseguido anotar el conjunto de datos de Panoptic mediante el uso de métodos de filtrado y refinamiento. En este proceso es crucial resaltar la importancia de emplear estrategias para reconstrucción 3D que permitan fusionar datos de distintos puntos de vista, mejorando la calidad de resultados. Además, los modelos entrenados consiguen realizar predicciones cercanas a lo previsto y cumplen con el propósito final que radica esencialmente en poder aplicar las anotaciones refinadas a nuevos procesamientos.

Capítulo 7

Trabajo futuro

Con el objetivo de continuar con la labor de este trabajo, se proponen ciertas mejoras en los siguientes apartados:

- Las reconstrucciones obtenidas en el TFM dependen de la calidad de un sistema de estimación de parámetros de FLAME, es decir, un modelo **3DMM**. Una mejora podría venir de refinar el modelo de DAD-3DNet (*fine-tuning*) para que utilizase la información multivista y temporal del vídeo de forma coordinada.
- Aunque anotar manualmente Panoptic de manera completa es inviable, sería beneficioso anotar esta base de datos de manera parcial. Con este enfoque, se podría lograr un mejor resultado con la selección del algoritmo de ordenación de cámaras, obteniendo una mayor precisión y fiabilidad de las reconstrucciones.
- En lo respectivo a los entrenamientos, el uso de redes neuronales complejas junto con el total de los datos de Panoptic procesados podría dar lugar al desarrollo de sistemas de mayor calidad y robustez ante distintas perspectivas. Con esta información, se podrían obtener modelos tanto para detectar *landmarks* faciales 2D como para obtener reconstrucciones 3D de la cabeza a partir de una imagen. Asimismo, la inclusión de información multivista podría resultar realmente útil en este tipo de entrenamientos.

Al incorporar las mejoras mencionadas, se puede concluir que se logaría un incremento significativo en la calidad de los resultados, lo que llevaría a la generación de modelos más robustos y eficientes. Estas modificaciones no solo apuntan a la optimización de la calidad, sino que también potencian la integración de estas anotaciones en proyectos futuros.

Bibliografía

- [1] Yao Feng y col. “Learning an Animatable Detailed 3D Face Model from In-the-Wild Images”. En: *ACM Transactions on Graphics (ToG), Proc. SIGGRAPH* 40.4 (ago. de 2021), 88:1-88:13 (vid. pág. 1).
- [2] Radek Danecek, Michael J. Black y Timo Bolkart. “EMOCA: Emotion Driven Monocular Face Capture and Animation”. En: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, págs. 20311-20322 (vid. pág. 1).
- [3] Tetiana Martyniuk y col. “DAD-3DHeads: A Large-scale Dense, Accurate and Diverse Dataset for 3D Head Alignment from a Single Image”. En: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2022 (vid. págs. 1, 7, 8, 11).
- [4] Araceli Morales, Gemma Piella y Federico M Sukno. “Survey on 3D face reconstruction from uncalibrated images”. En: *Computer Science Review* 40 (2021), pág. 100400 (vid. págs. 1, 2).
- [5] Marwa Chendeb EL Rai y col. “Using facial images for the diagnosis of genetic syndromes: A survey”. En: *2015 International Conference on Communications, Signal Processing, and their Applications (ICCSPA ’15)*. 2015, págs. 1-6. DOI: [10.1109/ICCSPA.2015.7081271](https://doi.org/10.1109/ICCSPA.2015.7081271) (vid. pág. 1).
- [6] Michael Zollhöfer y col. “State of the art on monocular 3D face reconstruction, tracking, and applications”. En: *Computer graphics forum*. Vol. 37. 2. Wiley Online Library. 2018, págs. 523-550 (vid. págs. 1, 9).
- [7] Shu Liang, Linda G Shapiro e Ira Kemelmacher-Shlizerman. “Head reconstruction from internet photos”. En: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II* 14. Springer. 2016, págs. 360-374 (vid. págs. 1, 12).
- [8] Oscar Meruvia-Pastor. “Enhancing 3D capture with multiple depth camera systems: A state-of-the-art report”. En: *RGB-D image analysis and processing* (2019), págs. 145-166 (vid. pág. 2).
- [9] Zimo Li, Prakruti C Gogia y Michael Kaess. “Dense surface reconstruction from monocular vision and LiDAR”. En: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, págs. 6905-6911 (vid. pág. 2).
- [10] Ralph Gross y col. “Multi-pie”. En: *Image and vision computing* 28.5 (2010), págs. 807-813 (vid. pág. 2).
- [11] Volker Blanz y Thomas Vetter. “A morphable model for the synthesis of 3D faces”. En: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 2023, págs. 157-164 (vid. págs. 2, 10).

- [12] Sahil Sharma y Vijay Kumar. “3d face reconstruction in deep learning era: A survey”. En: *Archives of Computational Methods in Engineering* 29.5 (2022), págs. 3475-3507 (vid. pág. 2).
- [13] K Konstantin y K Larysa. “Fast facial landmark detection and applications: a survey”. En: DOI: <https://doi.org/10.13140/RG.2.32735.07847> (2020), pág. 1 (vid. pág. 2).
- [14] Tetiana Martyniuk y col. “Dad-3dheads: A large-scale dense, accurate and diverse dataset for 3d head alignment from a single image”. En: *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 2022, págs. 20942-20952 (vid. págs. 3, 17).
- [15] Hanbyul Joo y col. “Panoptic Studio: A Massively Multiview System for Social Motion Capture”. En: *The IEEE International Conference on Computer Vision (ICCV)*. 2015 (vid. págs. 3, 14, 15).
- [16] Xiangyu Zhu y col. “Beyond 3dmm space: Towards fine-grained 3d face reconstruction”. En: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*. Springer. 2020, págs. 343-358 (vid. pág. 7).
- [17] Xing Zhang y col. “A high-resolution spontaneous 3d dynamic facial expression database”. En: *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE. 2013, págs. 1-6 (vid. pág. 7).
- [18] Lijun Yin y col. “A 3D facial expression database for facial behavior research”. En: *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE. 2006, págs. 211-216 (vid. pág. 8).
- [19] Xing Zhang y col. “Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database”. En: *Image and Vision Computing* 32.10 (2014), págs. 692-706 (vid. pág. 8).
- [20] Haotian Yang y col. “Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction”. En: *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 2020, págs. 601-610 (vid. pág. 8).
- [21] Rohith Krishnan Pillai y col. “The 2nd 3d face alignment in the wild challenge (3dfaw-video): Dense reconstruction from video”. En: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019, págs. 0-0 (vid. pág. 8).
- [22] Georgios Stylianou y Andreas Lanitis. “Image based 3d face reconstruction: a survey”. En: *International Journal of Image and Graphics* 9.02 (2009), págs. 217-250 (vid. pág. 9).
- [23] Anh Tuân Trần y col. “Extreme 3d face reconstruction: Seeing through occlusions”. En: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, págs. 3935-3944 (vid. pág. 9).
- [24] Tal Hassner. “Viewing real-world faces in 3D”. En: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, págs. 3607-3614 (vid. pág. 10).
- [25] Kim Youwang y col. “A Large-Scale 3D Face Mesh Video Dataset via Neural Reparameterized Optimization”. En: *arXiv preprint arXiv:2310.03205* (2023) (vid. pág. 10).

- [26] Emmanuel Prados y Olivier Faugeras. “Shape from shading”. En: *Handbook of mathematical models in computer vision*. Springer, 2006, págs. 375-388 (vid. pág. 10).
- [27] Feng Liu y col. “Joint face alignment and 3d face reconstruction”. En: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*. Springer. 2016, págs. 545-560 (vid. pág. 10).
- [28] HM Rehan Afzal y col. “3D face reconstruction from single 2D image using distinctive features”. En: *IEEE Access* 8 (2020), págs. 180681-180689 (vid. pág. 10).
- [29] Chenglei Wu y col. “High-quality shape from multi-view stereo and shading under general illumination”. En: *CVPR 2011*. IEEE. 2011, págs. 969-976 (vid. págs. 10, 12).
- [30] Hai-bin Liao y col. “Rapid 3D face reconstruction by fusion of SFS and Local Morphable Model”. En: *Journal of Visual Communication and Image Representation* 23.6 (2012), págs. 924-931 (vid. pág. 10).
- [31] Luan Tran y Xiaoming Liu. “Nonlinear 3d face morphable model”. En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, págs. 7346-7355 (vid. pág. 11).
- [32] Luan Tran y Xiaoming Liu. “On learning 3d face morphable model from in-the-wild images”. En: *IEEE transactions on pattern analysis and machine intelligence* 43.1 (2019), págs. 157-171 (vid. pág. 11).
- [33] Abdullah Taha Arslan y Erol Seke. “Face depth estimation with conditional generative adversarial networks”. En: *IEEE Access* 7 (2019), págs. 23222-23231 (vid. pág. 11).
- [34] Elad Richardson, Matan Sela y Ron Kimmel. “3D face reconstruction by learning from synthetic data”. En: *2016 fourth international conference on 3D vision (3DV)*. IEEE. 2016, págs. 460-469 (vid. págs. 11, 12).
- [35] Joseph Roth, Yiying Tong y Xiaoming Liu. “Unconstrained 3D face reconstruction”. En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, págs. 2606-2615 (vid. pág. 12).
- [36] Tomas Simon, Hanbyul Joo y Yaser Sheikh. “Hand Keypoint Detection in Single Images using Multiview Bootstrapping”. En: *CVPR* (2017) (vid. pág. 14).
- [37] Hanbyul Joo y col. “Panoptic Studio: A Massively Multiview System for Social Interaction Capture”. En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017) (vid. pág. 14).
- [38] Donglai Xiang, Hanbyul Joo y Yaser Sheikh. “Monocular total capture: Posing face, body, and hands in the wild”. En: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, págs. 10965-10974 (vid. pág. 14).
- [39] Tianye Li y col. “Learning a model of facial shape and expression from 4D scans.” En: *ACM Trans. Graph.* 36.6 (2017), págs. 194-1 (vid. pág. 17).
- [40] Richard Hartley y Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003 (vid. pág. 19).
- [41] Luca Bernecker y Andrea Idini. “Quantum Levenberg–Marquardt Algorithm for optimization in Bundle Adjustment”. En: *arXiv preprint arXiv:2203.02311* (2022) (vid. pág. 20).

- [42] Manolis IA Lourakis y Antonis A Argyros. “SBA: A software package for generic sparse bundle adjustment”. En: *ACM Transactions on Mathematical Software (TOMS)* 36.1 (2009), págs. 1-30 (vid. pág. [21](#)).