

# Estimación de Profundidad Monocular Online con Transformers Eficientes

*Guillermo Sánchez Brizuela*

Universidad de Valladolid, curso 2020-2021

# Índice

<b>1. Introducción</b>	<b>4</b>
1.1. Motivación . . . . .	5
1.2. Objetivos . . . . .	6
1.3. Estructura del documento . . . . .	6
<b>2. Marco Teórico y Estado del Arte</b>	<b>7</b>
2.1. Inteligencia Artificial, Aprendizaje automático, y Aprendizaje Profundo . . . . .	7
2.2. Redes neuronales . . . . .	7
2.2.1. Funciones de activación . . . . .	8
2.2.2. Funciones de pérdida . . . . .	9
2.2.3. Optimizadores . . . . .	9
2.3. MLP . . . . .	11
2.4. RNN . . . . .	12
2.5. CNN . . . . .	13
2.5.1. Conexiones residuales y ResNets . . . . .	15
2.5.2. Efficientnet . . . . .	16
2.6. Transformers . . . . .	17
2.6.1. Arquitectura . . . . .	17
2.6.2. Mecanismos de atención . . . . .	18
2.6.3. Atención eficiente . . . . .	19
2.6.3.1. Patrones fijos - Fixed Patterns (FP) . . . . .	19
2.6.3.2. Combinación de patrones - Combination of Patterns (CP) . . . . .	19
2.6.3.3. Aprendizaje de patrones - Learnable Patterns (LP) . . . . .	19
2.6.3.4. Disminución de rango . . . . .	20
2.6.4. Performer . . . . .	20
2.7. Transformers para visión artificial - ViT . . . . .	21
2.8. Estimación de profundidades . . . . .	22
2.8.1. Técnicas de estimación de profundidades . . . . .	22
2.8.2. Aprendizaje automático no supervisado . . . . .	23
2.8.3. Aprendizaje automático semisupervisado . . . . .	23
2.8.4. Aprendizaje automático supervisado . . . . .	24
2.8.5. DPT . . . . .	24
2.9. Optimización de modelos . . . . .	26
<b>3. Material y Métodos</b>	<b>30</b>
3.1. Software . . . . .	30
3.2. Hardware . . . . .	32
3.3. Datasets . . . . .	32
3.3.1. ImageNet . . . . .	32
3.3.2. MIX6 . . . . .	33
3.3.3. KITTI . . . . .	34
3.3.3.1. Datos . . . . .	35
3.3.3.2. Conjuntos de entrenamiento, validación y evaluación . . . . .	36
3.4. Definición de modelos a entrenar . . . . .	37
3.5. Evaluación . . . . .	37
3.5.1. Métricas . . . . .	37
3.5.1.1. <i>Accuracy under a threshold</i> . . . . .	38
3.5.1.2. <i>Mean Absolute Value of the Relative Error (Abs Rel)</i> . . . . .	38
3.5.1.3. <i>Mean Squared Relative Error (Sq Rel)</i> . . . . .	38
3.5.1.4. <i>Linear Root Mean Squared Error (RMSE)</i> . . . . .	38

3.5.1.5.	<i>Logarithmic Root Mean Squared Error (RMSElog)</i>	38
3.5.1.6.	<i>Scale Invariant Logarithmic Error (SIlog)</i>	39
3.5.1.7.	<i>Mean Logarithmic Error (Log10)</i>	39
3.5.1.8.	Velocidad de procesamiento	39
<b>4.</b>	<b>Desarrollo y Modificaciones de la Arquitectura</b>	<b>41</b>
4.1.	Cloud	41
4.2.	Warmstart	42
4.3.	Entrenamiento	43
4.3.1.	Función de pérdida	44
4.4.	Reducción de tamaño de la entrada	44
4.5.	Capas de atención eficiente	45
4.6.	Número de cabezas	45
4.7.	Hooks del transformer y bloques de atención posteriores	45
4.8.	Backbone convolucional	46
<b>5.</b>	<b>Resultados</b>	<b>48</b>
5.1.	Resultados cuantitativos	48
5.1.1.	Reducción de tamaño de la entrada	48
5.1.2.	Número de cabezas	48
5.1.3.	Capas de atención eficiente	49
5.1.4.	Cambio en los hooks del transformer y eliminación de las capas de atención posteriores	50
5.1.5.	Cambio del backbone convolucional	50
5.2.	Resultados cualitativos	51
<b>6.</b>	<b>Discusión</b>	<b>52</b>
<b>7.</b>	<b>Conclusiones y Lineas Futuras</b>	<b>53</b>
<b>Bibliografía</b>		<b>54</b>

Backbone	Hooks	$\delta_1$	Abs.	Rel.	RMSE log	SI log
EfficientNet-B0	<b>0,1</b>	0,9107 ± 0,0057	0,0880 ± 0,0036	0,1372 ± 0,0057	12,2777 ± 0,4610	
	<b>2,5</b>	0,9229 ± 0,0159	0,0825 ± 0,0076	0,1277 ± 0,0121	11,4931 ± 0,8577	
	<b>8,11</b>	0,9280 ± 0,0101	0,0803 ± 0,0041	0,1255 ± 0,0080	11,4357 ± 0,5824	
ResNet50	<b>0,1</b>	0,9944 ± 0,0002	0,0336 ± 0,0007	0,0535 ± 0,0006	5,2550 ± 0,0687	
	<b>2,5</b>	0,9951 ± 0,0003	0,0316 ± 0,0004	0,0510 ± 0,0005	4,9825 ± 0,0519	
	<b>8,11</b>	0,9957 ± 0,0002	0,0301 ± 0,0007	0,0493 ± 0,0011	4,8452 ± 0,1162	

Tabla 1

## 1. Introducción

Cuando usamos una cámara para capturar una imagen o un vídeo, creamos una representación bidimensional de lo que es en realidad una escena tridimensional. Para conseguir esta reducción de dimensiones, se proyectan en un plano, cada uno de los puntos visibles. Al realizar esta proyección, se pierde una gran cantidad de información relacionada con la profundidad. Esto es debido a que los puntos ahora representados en el plano bidimensional podían encontrarse a cualquier distancia siempre y cuando estuviesen situados en la recta que atraviesa el punto verdadero y el centro de la cámara ([Figura 1](#)). Es decir, existen infinitas escenas tridimensionales con la misma proyección perspectiva.

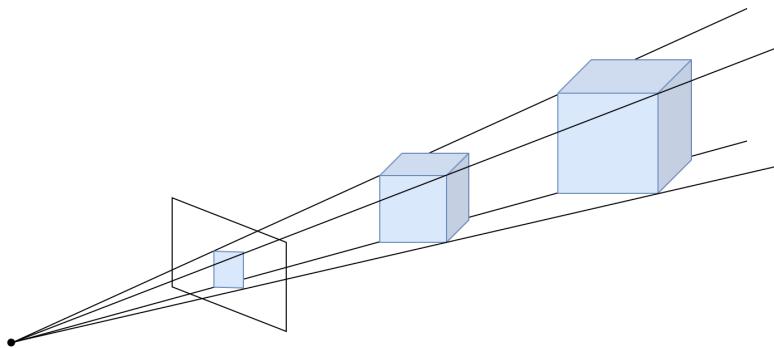


Figura 1: Proyección perspectiva y ambigüedad de escala.

Existen soluciones *hardware* capaces de capturar escenas tridimensionales sin esta perdida de información, por ejemplo: sensores LiDAR, cámaras de tiempo de vuelo, o conjuntos de cámaras para llevar a cabo estereovisión<sup>1</sup>[1]. Sin embargo, estas soluciones requieren de sensores adicionales, con el incremento de material, coste y peso que esto conlleva. Es por esto por lo que recuperar la profundidad (o una estimación de esta) a partir de una imagen obtenida con una cámara corriente sería de gran utilidad, por ejemplo:

- En diferentes tareas dentro de múltiples campos de aplicación: Navegación robótica y conducción autónoma, empleando la profundidad para reconstruir mapas o calcular el cambio en la posición del agente (odometría visual, VSLAM); detección de superficies y capacidad de incluir oclusión en aplicaciones de realidad aumentada; generación de modelos 3D a través de fotografías (fotogrametría); efectos fotográficos en aplicaciones móviles (efecto *Bokeh*); etc.
- Como información adicional o etapa intermedia en otras tareas típicas en visión artificial: detección de objetos, segmentación, clasificación, etc.

Si observamos, nunca mejor dicho, el sistema visual de los humanos, podemos comprobar como este es un sistema estereoscópico compuesto por dos “cámaras” (los ojos) y un cerebro que interpreta la disparidad entre estas imágenes para obtener una estimación de la profundidad a la que se encuentra cada objeto que vemos. No obstante, si nos tapamos un ojo, aunque con peores resultados, somos capaces de estimar la distancia a la que se encuentran los elementos que están dentro de nuestro campo visual, manteniendo, en una gran mayoría de las ocasiones, la capacidad de distinguir cuáles están más cerca y cuáles más lejos. Esto se debe en gran parte a una serie de sesgos cognitivos que los humanos hemos aprendido a medida que crecemos, conocidos como pistas monoculares (pueden ser dinámicas o estáticas, en función de si consideran la información a lo largo del tiempo, por ejemplo, objetos en movimiento), y que no solo se emplean cuando nos

<sup>1</sup>Con una pareja de cámaras de posiciones conocidas, es posible estimar la profundidad a partir de la disparidad geométrica entre las dos imágenes capturadas.

tapamos un ojo, si no que también los emplea el cerebro cuando vemos con los dos ojos. Algunas de las principales pistas monoculares (estáticas) [2] son: el tamaño relativo con el que observamos un objeto en función de la distancia a la que se encuentre; la oclusión de los elementos que están más próximos que otros; la convergencia de líneas paralelas a medida que se alejan, por ejemplo, en carreteras o vías; el cambio en el tono del color de los objetos lejanos debido a la dispersión de la luz; o la forma de los reflejos y las sombras que producen los elementos, originados por las fuentes de luz de la escena.

No obstante, realizar este análisis de las imágenes monoculares que tan eficientemente llevamos a cabo los humanos en un ordenador de forma automática empleando técnicas de visión artificial tradicional roza lo imposible.

Las limitaciones de este tipo de técnicas no solo aparecen en la estimación de profundidades, si no que están presentes en una gran parte de las tareas propias de la visión artificial. Buscando dejar atrás estas limitaciones, en los últimos años se han desarrollado un gran número de sistemas de aprendizaje automático profundo diseñados para trabajar con imágenes. Estos sistemas, pese a tener ciertos inconvenientes, han conseguido ofrecer unos muy buenos resultados junto a una mucho mayor robustez y capacidad de generalización ante modificaciones en las entradas (cambios de entorno, color, luz, orientación de los elementos, etc.), muy frecuentes en los entornos reales no controlados.

Dentro de estas técnicas, en general, y especialmente para la estimación de profundidades, las redes neuronales convolucionales han prevalecido como las arquitecturas que mejores resultados aportaban. No obstante, en los últimos años han surgido otro tipo de arquitecturas, los *transformers* [3], que presentan resultados muy competitivos. En vista de esto, este trabajo revisa el estado del arte actual en estimación de profundidad monocular y explora una de las arquitecturas que mejores resultados consigue, DPT [4], reproduciendo su entrenamiento y modificándola para disminuir su tamaño y acelerar la inferencia de resultados, es decir, reducir el tiempo necesario para estimar la profundidad en una imagen dada.

## 1.1. Motivación

Este trabajo tiene dos motivaciones claramente diferenciadas, la primera, académica y la segunda más aplicada. En primer lugar, para poder trabajar en el campo de la estimación de profundidades, es necesario explorar y conocer las distintas técnicas que han surgido a lo largo de los años, tanto las bases y los primeros enfoques para resolver el problema, como las soluciones especializadas que conforman el estado del arte. Por lo tanto, en este trabajo se pretende resumir y agrupar las principales técnicas para poder, posteriormente, revisar y estudiar los temas que se tratan de una manera dirigida y ordenada.

Por otro lado, los modelos del estado del arte tienden a ser (con excepciones) cada vez más complejos, tienen un mayor número de parámetros, y precisan de grandes cantidades de datos para ser entrenados. Esto conlleva una perdida de accesibilidad al desarrollo y experimentación con dichas arquitecturas, que necesitan una infraestructura costosa para ejecutarse. Además, este incremento en tamaño de los modelos hace que sus velocidades de ejecución e inferencia de resultados se vea afectada. En muchas de las aplicaciones mencionadas en el apartado anterior, el tiempo de inferencia es un factor crítico, ya que muchas veces el procesamiento de las imágenes debe llevarse a cabo en entornos con recursos computacionales limitados y de forma *online*, es decir, procesar las imágenes a medida que están disponibles (sin considerar las restricciones de un entorno de tiempo real). En el caso de que la inferencia de los modelos no se lleve a cabo en dispositivos embebidos y recaiga en servidores a los que los clientes hacen peticiones, un mayor tiempo de ejecución se traduce directamente en un incremento de costes, por lo que tampoco es despreciable. Por estas razones, en este trabajo se busca modificar una de los modelos del estado del arte en estimación de profundidades a partir de imágenes monoculares para reducir

su tamaño y tiempo de inferencia reduciendo lo mínimo posible la calidad de los resultados.

## 1.2. Objetivos

Los objetivos principales de este Trabajo Fin de Máster, son:

1. Llevar a cabo una revisión del estado del arte relacionado con la estimación de profundidades en imágenes monoculares. Más concretamente, en aquellas técnicas que empleen aprendizaje automático, prestando especial atención a las arquitecturas basadas en *transformers* y sus variantes eficientes.
2. Estudiar una arquitectura del estado del arte de estimación de profundidades y modificar su estructura para acelerar la inferencia hasta obtener modelos capaces de procesar imágenes de forma *online*. Además, una vez definidos los modelos con las variaciones planteadas, comparar sus resultados tras ajustarlos a un *dataset* concreto.
3. Explorar diferentes técnicas generales para acelerar el entrenamiento e inferencia de los modelos de aprendizaje automático profundo.
4. Diseñar una solución en la nube que permita desplegar de forma automática instancias que ejecuten los experimentos necesarios, es decir, entrenando los distintos modelos planteados.

Checkear  
esto

## 1.3. Estructura del documento

Este trabajo Fin de Máster está organizado de la siguiente manera: Primero, se han expuesto en esta sección de Introducción tanto la motivación detrás del proyecto como los objetivos que se plantearon al comenzarlo; a continuación, en el segundo capítulo se contextualiza el trabajo repasando las bases teóricas sobre las que se apoya su desarrollo, revisando también el Estado del Arte de estos campos; en el tercer capítulo, se presentan y justifican tanto los materiales empleados para el desarrollo del trabajo como la metodología que se ha seguido; después, en el cuarto capítulo, se definen y explican aquellos desarrollos especialmente significativos dentro del trabajo para continuar en el quinto capítulo con los resultados que se han obtenido, haciendo especial hincapié en la influencia en estos resultados de cada uno de los desarrollos llevados a cabo. Por último, se incluye en el capítulo seis una discusión de los resultados y en el capítulo siete las conclusiones del documento junto a una serie de líneas de investigación futuras que podrían explorarse para continuar trabajando en el contexto de este proyecto.

Poner algo de los anexos?

## 2. Marco Teórico y Estado del Arte

### 2.1. Inteligencia Artificial, Aprendizaje automático, y Aprendizaje Profundo

La Inteligencia Artificial (IA), engloba el estudio de agentes que perciben un entorno y actúan en consecuencia. Este campo, comprende multitud de disciplinas y técnicas aplicables en incontables escenarios. Dentro de los subcampos que conforman la IA están por ejemplo los sistemas expertos, los modelos probabilísticos, la lógica difusa, o, de especial importancia en este trabajo, el aprendizaje automático (*Machine Learning - ML*).

El aprendizaje automático estudia diferentes modelos que son capaces de extraer información de forma automática a partir de un conjunto de datos. En función de qué información tengan estos datos y qué se busque hacer con ella, es posible distinguir entre aprendizaje supervisado (donde los datos están etiquetados con lo que se quiere aprender), aprendizaje no supervisado (donde los datos no tienen anotaciones y se busca extraer información o encontrar relaciones entre ellos), aprendizaje semisupervisado (si se usan datos etiquetados y no etiquetados) o aprendizaje por refuerzo (si se emplea una función de recompensa y se deja al modelo elegir las acciones que lleva a cabo).

Este proyecto, no obstante, se centra más aún en el aprendizaje profundo (*Deep Learning - DL*), un subcampo del aprendizaje automático donde se emplean modelos y técnicas con una complejidad computacional mayor que, en consecuencia, requieren de una cantidad de datos acorde para poder aprender de forma satisfactoria.

Cabe mencionar, que pese a que muchos de los conceptos explicados a continuación son perfectamente validos en otras modalidades, al centrar este trabajo en el aprendizaje profundo supervisado, el marco teórico se ajusta y limita a este.

### 2.2. Redes neuronales

Las redes neuronales (en inglés, *Neural Networks - NN*), son sistemas computacionales inspirados en una versión simplificada de las neuronas biológicas. Estas neuronas artificiales, están definidas por una serie de elementos: entradas, **pesos**, **bias**, **función de activación** y salida. El valor que toma esta salida, siguiendo la notación de la [Figura 2](#), viene dado por la ecuación  $y = f(Wx + b)$ .

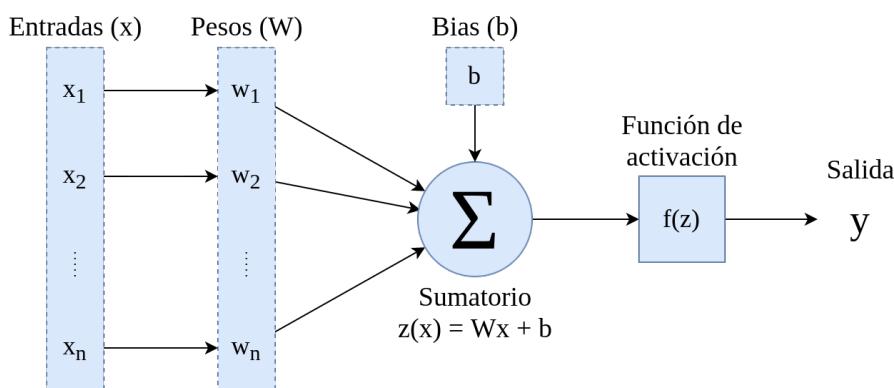


Figura 2: Estructura de una neurona artificial.

Una de las bases teóricas más importantes de las NN, es que con un conjunto de neuronas lo suficientemente grande y funciones de activación no lineales, es posible aproximar cualquier función continua (*Universal approximation theorem [5]*). Por desgracia, ajustar los parámetros (pesos y bias) de cada una de las neuronas de dicha red para aproximar la función deseada no es un problema trivial.

Una de las soluciones más comunes a este problema, es, de forma iterativa, modificar los parámetros de forma que minimicen la distancia entre la salida de la red y la función que se busca aproximar. Para esto, se emplea la propagación hacia atrás (***backpropagation***) [6], que necesita: un **optimizador** (en los que se profundiza más adelante), un **conjunto de entradas con salidas conocidas** y una **función de perdida** que se emplea como aproximación optimizable de la distancia previamente mencionada. Este algoritmo, de forma simplificada, consiste en:

1. Calcular la salida de la red a partir de las entradas de las cuales conocemos la salida esperada (***forward pass***).
2. Emplear una **función de pérdida** para calcular una medida del error entre la salida obtenida y la salida esperada.
3. Calcular las derivadas parciales del error respecto de cada uno de los parámetros (**pesos** y ***bias***) que se quieren ajustar.
4. Ajustar los valores de los parámetros en función de su influencia en el error empleando un optimizador concreto.

Estos pasos, normalmente se repiten varias veces para cada ejemplo disponible (entendiendo como ejemplo las parejas entrada-salida conocidas). En el contexto de las redes neuronales, este proceso de ajuste se conoce como **entrenamiento** y cada una de las iteraciones que la red realiza sobre el conjunto de ejemplos se conoce como **época** (*epoch*, en inglés).

El proceso de entrenamiento, sin embargo, hay que controlarlo y validarla de alguna forma, ya que existe el riesgo de sobreajustar los parámetros de la red neuronal a los datos de entrenamiento, siendo el modelo incapaz de generalizar a datos no antes vistos. Este sobreajuste se conoce en inglés como ***overfitting***. Para controlar si el modelo se está sobreajustando, es común dividir el conjunto de datos en tres subconjuntos: entrenamiento, validación, y evaluación. El conjunto de entrenamiento, se emplea para ajustar los parámetros del modelo; el de evaluación, para comprobar si hay *overfitting* y para elegir entre distintas configuraciones de un mismo modelo o distintos modelos; y el de evaluación, para calcular las métricas finales del modelo elegido para un problema concreto. En el trabajo de Sebastian Raschka [7] se presentan distintas formas de hacer estas particiones, en este trabajo, no obstante, se emplearán particiones ya establecidas en la literatura existente para los conjuntos de datos empleados ([Sección 3.3.3.2](#)).

A continuación, se profundiza en algunos de los conceptos mencionados prestando especial atención a los elementos que aparecen en este trabajo.

### 2.2.1. Funciones de activación

Tal y como se ha visto en la [Figura 2](#), la función de activación transforma la suma del *bias* y el producto de los pesos y las entradas para obtener el valor de salida. En la [Figura 3](#), es posible observar graficadas varias de estas funciones junto con sus ecuaciones.

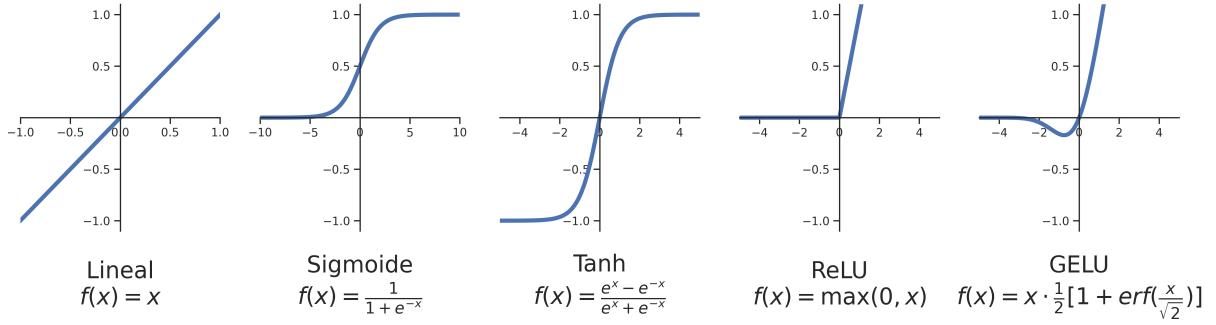


Figura 3: Funciones de activación.

La primera de estas funciones de activación es la función de activación lineal. Esta función, no suele emplearse, ya que al concatenar múltiples neuronas con activaciones lineales, se obtiene otra función lineal que podría representarse con una sola neurona. Por esta razón, son funciones no lineales como la sigmoide, la tangente hiperbólica, ReLU (*Rectified Linear Unit*) [8] o GELU (*Gaussian Error Linear Unit*) [9] las que suelen aparecer en las arquitecturas actuales.

Cada una de estas funciones de activación tiene características propias que las hacen diferentes entre si. La tangente hiperbólica, aún siendo similar a la sigmoide, es simétrica y tiene unos gradientes más fuertes en las zonas cercanas al cero, lo que origina que sus salidas sean valores pequeños centrados en cero y que el entrenamiento sea más rápido [10]. La función ReLU, por otro lado, pese a ser parcialmente lineal, puede aproximar funciones no lineales gracias a su definición a trozos, además, el gradiente para los valores positivos es constante (1), reduciendo en gran medida el desvanecimiento de los gradientes cuando se definen redes neuronales con un gran número de capas. Por último, la función GELU, es comúnmente empleada en las arquitecturas basadas en *transformers* y mantiene el gradiente constante en valores positivos a la vez que añade una ligera ponderación a los valores negativos cercanos al cero.

### 2.2.2. Funciones de pérdida

Las funciones de pérdida, tal y como se ha mencionado previamente, son aproximaciones optimizables de la distancia o error entre el valor real a predecir y la predicción realizada. Dentro de estas funciones, cuyo valor se busca minimizar, existen diferentes situaciones que pueden dificultar el aprendizaje de los modelos, por ejemplo, los mínimos locales, donde los valores de la función de pérdida son menores que los valores de su entorno, originando que los gradientes se desvanezcan y detengan el proceso de optimización, o los puntos de silla (*saddle points*), donde, pese a no haber un mínimo local, los gradientes también se desvanecen, deteniendo la minimización.

### 2.2.3. Optimizadores

Los optimizadores son algoritmos que minimizan el valor de la función de pérdida ( $f(\theta)$ ) ajustando los parámetros ( $\theta$ ) de los modelos, entre otras cosas, en base a la influencia de estos parámetros en el error obtenido en dicha función. Esta influencia se obtiene calculando las derivadas parciales del error respecto de cada parámetro (gradiente:  $\nabla_{\theta} f_t(\theta_{t-1})$ ), y, en función del algoritmo que se elija, puede verse afectada por distintos factores. Algunos de los optimizadores más empleados son SGD, AdaGrad [11], RMSProp [?] TODO, Adam [12] o AdamW [13].

El descenso de gradiente tradicional promedia los gradientes del error en todos los ejemplos disponibles para actualizar los parámetros. La modificación que introduce el **SGD** (*Stochastic Gradient Descent*), es actualizar los parámetros tras ver cada uno de los ejemplos. De esta forma, se acelera el entrenamiento, ya que aunque cada ajuste no es óptimo, los parámetros

se actualizan mucho más frecuentemente. Además, el ruido intrínseco de cada ejemplo tiene un efecto regularizador que puede ayudar a salir de mínimos locales.

No obstante, gracias a la capacidad de procesamiento paralelo de las tarjetas gráficas es muy común promediar el cálculo del error en pequeños subconjuntos (**batch**) de todos los datos disponibles. Esto se conoce como *Mini Batch Gradient Descent* y el tamaño (**batch size**) que se elija influye directamente en diferentes factores como son la velocidad de entrenamiento, la rapidez para converger, o el punto al que se converge [14].

Estos tres algoritmos, y en general la mayoría de los optimizadores, controlan el cambio de los parámetros mediante un hiperparámetro conocido como tasa de aprendizaje (*learning rate*:  $\gamma$ ), que se multiplica por el valor del gradiente de la función de pérdida respecto de los parámetros ([Ecuación \(1\)](#)).

$$\theta_t = \theta_{t-1} - \gamma(\nabla_\theta f_t(\theta_{t-1})) \quad (1)$$

En el caso del *Mini Batch Gradient Descent*, al promediar los gradientes de varios ejemplos se reduce el ruido del SGD. Inspirado por el concepto físico de **Momento**, es posible añadir a estos optimizadores una modificación para que se consideren también los gradientes de los pasos anteriores del optimizador. De esta forma, se reduce el ruido y se hace más robusta la optimización frente a mínimos locales. El momento se controla con un hiperparámetro  $\mu$  siguiendo la [Ecuación \(2\)](#), donde  $v_t$  es el Momento en un paso concreto y  $g_{t,t-1}$  los gradientes correspondientes a este paso.

$$v_t = \mu v_{t-1} + g_{t,t-1} \quad (2)$$

Este cálculo multiplica los momentos anteriores de forma recursiva por un valor menor que uno, de forma que influyen todos los gradientes anteriores, pero cada vez con menos fuerza. Esto se puede ver comprobar matemáticamente desarrollando la [Ecuación \(2\)](#) ([Ecuación \(3\)](#)). Finalmente, este momento toma la posición del gradiente en la actualización de los parámetros ([Ecuación \(4\)](#)).

$$v_t = \mu^2 v_{t-2} + \mu g_{t-1,t-2} + g_{t,t-1} = \sum_{i=0}^{t-1} \mu^i g_{t-i,t-i-1} \quad (3)$$

$$\theta_t = \theta_{t-1} - \gamma(v_t) \quad (4)$$

**AdaGrad** (*Adaptative Gradients*) es otro optimizador que también modifica el descenso del gradiente. En este caso, para reducir las tasas de aprendizaje en aquellos parámetros que han tenido mayores derivadas parciales, dividiéndolas entre la suma de los gradientes ( $g$ ) anteriores ([Ecuación \(5\)](#) y [Ecuación \(6\)](#)). Estos gradientes se acumulan durante toda la optimización, por lo que es especialmente efectivo en datos dispersos donde no en todos los ejemplos se dispone de todas las características, ya que las características que no se hayan visto tan frecuentemente, tendrán actualizaciones mayores. Esto, sin embargo, también ralentiza el algoritmo, ya que según avanza el proceso, las actualizaciones de los parámetros son menores y pueden llegar a detener el proceso, impidiendo que se alcance a un mínimo aceptable.

$$acumulación_t = acumulación_{t-1} + g_t^2 \quad (5)$$

$$\theta_t = \theta_{t-1} - \gamma \frac{g_t}{\sqrt{acumulación_t}} \quad (6)$$

**RMSProp** (*Root Mean Square Propagation*) soluciona el problema de la velocidad de AdaGrad añadiendo un factor de decaimiento. Este factor de decaimiento (*decay rate*:  $\alpha$ ), escala de forma recursiva la suma acumulada de los gradientes anteriores (media móvil exponencialmente ponderada - *Exponential weighted moving averages*), por lo que a medida que el optimizador avanza, la influencia de los gradientes anteriores disminuye. De esta forma, se mantienen los beneficios de AdaGrad evitando reducir indefinidamente la magnitud de la actualización de los parámetros ([Ecuación \(7\)](#) y [Ecuación \(8\)](#)).

$$\text{acumulación}_t = (\alpha)\text{acumulación}_{t-1} + (1 - \alpha)g_t^2 \quad (7)$$

$$\theta_t = \theta_{t-1} - \gamma \frac{g_t}{\sqrt{\text{acumulación}_t}} \quad (8)$$

Por último, **Adam** y **AdamW** son dos optimizadores que fusionan las ideas y ventajas detrás del momento y de RMSProp. Para conseguir esto, se vuelve a aplicar la media móvil exponencialmente ponderada, esta vez tanto para el primer como el segundo momento, controlado por dos parámetros  $\beta$  ([Ecuación \(9\)](#) y [Ecuación \(10\)](#)).

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1)g_t \quad (9)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2)g_t^2 \quad (10)$$

Además, para reducir la influencia del valor inicial (igual a cero) de  $m_0$  y  $v_0$ , se añade una normalización que sigue la [Ecuación \(11\)](#) y la [Ecuación \(12\)](#).

$$\hat{m}_t = \frac{m_t}{(1 - \beta_1^t)} \quad (11)$$

$$\hat{v}_t = \frac{v_t}{(1 - \beta_2^t)} \quad (12)$$

Una vez calculados estos dos términos, se actualizan los parámetros de forma similar a como se hacía en RMSProp, con la diferencia de que esta vez se multiplica por el momento acumulado y no por los gradientes ([Ecuación \(13\)](#)).

$$\theta_t = \theta_{t-1} - \gamma \frac{\hat{m}_t}{\sqrt{\hat{v}_t}} \quad (13)$$

En cuanto a la diferencia entre Adam y AdamW, esta reside en cómo se implementa el *weight decay*, un tipo de regularización que penaliza los pesos mayores ya que pueden introducir mayor varianza en el modelo (*overfitting*). Tal y como se indica en el artículo *Decoupled Weight Decay Regularization* [13], la implementación del *weight decay* en los optimizadores (Adam entre ellos) aplica este decaimiento en los gradientes. Sin embargo, al añadir las operaciones relacionadas con los gradientes adaptativos, aplicar el decaimiento en los gradientes deja de ser equivalente a aplicarlo realmente en los pesos. AdamW soluciona este problema implementando el método de regularización correctamente.

### 2.3. MLP

El Perceptrón Multicapa, o por sus siglas en inglés, **MLP** (*Multi Layer Perceptron*), es una de las arquitecturas de red neuronal más básicas y está formado por una capa de entrada, una capa de salida, y un bloque de capas intermedias u ocultas. Estas capas intermedias, son agrupaciones de neuronas donde cada una de las neuronas (normalmente) se conecta con todas las neuronas

de la capa anterior y con todas las neuronas de la capa posterior (Figura 4). De esta forma, se crea una sucesión donde las entradas de una neurona son las salidas de las neuronas de la capa anterior y la salida de dicha neurona es una de las entradas de todas las neuronas de la capa siguiente.

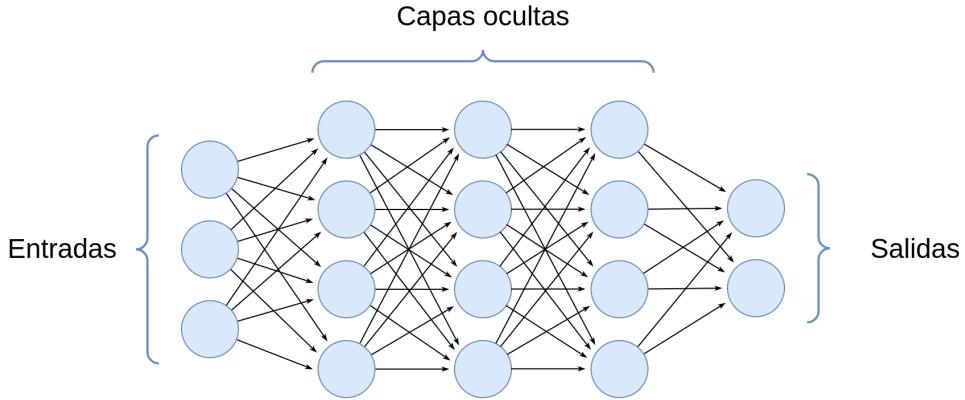


Figura 4: Perceptrón Multicapa.

## 2.4. RNN

Las redes neuronales recurrentes (*Recurrent Neural Networks* - RNN) son un grupo de redes neuronales cuya arquitectura se fundamenta en la utilización repetida de la salida de la propia red en un instante como entrada adicional de la red en el instante siguiente. Para conseguir esto, se mantiene la información en forma de estado oculto (*hidden state*). Gracias a esta transmisión recurrente, es posible transferir información entre entradas sucesivas de la red.

Este tipo de redes está especialmente diseñado para trabajar con entradas secuenciales de valores (series temporales, texto, señales de audio y video, etc.) ya que su arquitectura aprovecha la información de los datos previos y no tiene limitaciones en el tamaño de la secuencia de entrada.

Sin embargo, la ejecución y entrenamiento de este tipo de redes requiere ejecutarlas individualmente con cada uno de los valores de la entrada, haciendo que el proceso sea temporalmente costoso. Además, pese a que se han presentado distintas modificaciones como las celdas GRU (*Gated Recurrent Units*) [15] o LSTM (*Long Short-Term Memory*) [16] que los mitigan, también sufren tanto de desvanecimientos de gradientes por las activaciones sucesivas como de pérdida de información en cadenas de gran tamaño (al codificar toda la información anterior en un estado oculto, llega un punto en el que la información almacenada en instantes alejados temporalmente deja de ser representativa).

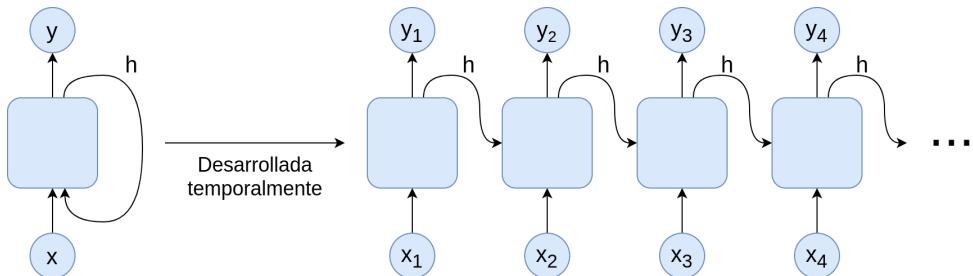


Figura 5: Red recurrente.

## 2.5. CNN

A la hora de trabajar con imágenes (matrices de dos o tres dimensiones), el número de conexiones, y por lo tanto, parámetros, que tendrían que existir para conectar cada valor de entrada con cada neurona crece enormemente. Las redes convolucionales, no solo lidian con este problema, si no además lo hacen proporcionando muy buenos resultados. Este tipo de redes, se basan en el concepto de núcleo (*kernel*) proveniente del procesamiento digital de imágenes y la operación de convolución, y su origen se atribuye al trabajo de LeCun et al. [17].

Los filtros o *kernels* (Figura 6) son pequeñas matrices (cuyos valores son equivalentes a los pesos vistos anteriormente, y por lo tanto son ajustados durante el entrenamiento) que se convolucionan sobre la imagen de entrada, extrayendo de esta forma mapas de características (también conocidos como mapas de activaciones) que son a su vez la entrada de las siguientes capas convolucionales tras aplicar una función de activación y puede que otras operaciones. En cuanto a la convolución (o correlación cruzada<sup>2</sup>), es una operación cuya salida se obtiene desplazando una de las dos matrices y calculando su producto escalar con la otra matriz (Figura 6). Estas capas, están definidas principalmente por el tamaño del *kernel*, el número de *kernels* que se aplican y la zancada o *stride*, que es el desplazamiento del *kernel* para calcular cada valor de salida y por lo tanto afecta directamente en su tamaño.

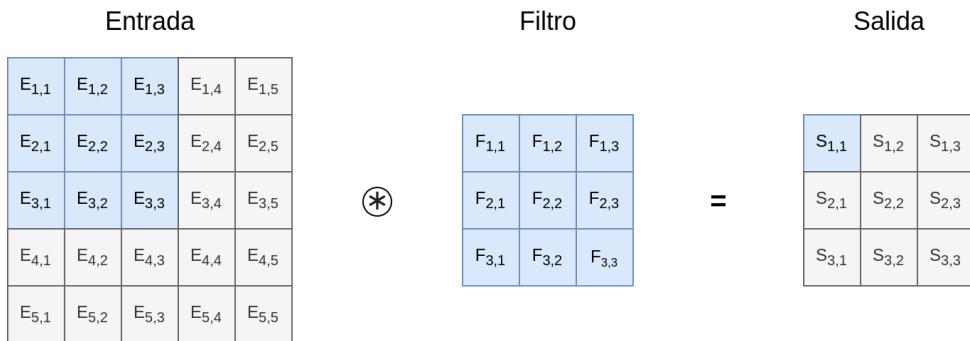


Figura 6: Operación de convolución con un *kernel* de  $3 \times 3$  y un *stride* de 1 para uno de los valores de salida.

Los *kernels*, normalmente tienden a extraer características de más alto nivel cuanto más adelante están en la red. Esto significa que los mapas de características resultantes en las capas iniciales son características de muy bajo nivel (líneas verticales, horizontales, curvas, esquinas, etc.), mientras que las de las últimas capas extraen características más complejas propias de cada dominio de aplicación [18].

En las redes convolucionales, es común encontrar otras operaciones como son:

- Pooling: Las capas de *Pooling* se emplean principalmente para reducir el tamaño de los mapas de características resultantes de las capas convolucionales. Para conseguirlo, agrupan un conjunto de valores en un solo valor aplicando una función que normalmente es el máximo (*MaxPooling* - se escoge el máximo valor) o la media (*AveragePooling* - se calcula la media de los valores). La operación de *Pooling* está definida principalmente por dos valores similares a los de la convolución: el tamaño del *kernel*, que define el tamaño del subconjunto de valores; y la zancada o *stride*, que define el número de valores que se traslada

<sup>2</sup>La operación matemática de convolución entre dos funciones incluye voltear una de las funciones. Sin embargo, en el campo del Deep Learning, voltear la entrada o el filtro requeriría una operación adicional que solo conllevaría que los pesos obtenidos tras el descenso de gradiente estuvieran también volteados. Por lo tanto, la operación de convolución en redes neuronales realmente es una operación de correlación cruzada donde no se voltean ninguna de las funciones.

el *kernel* cada paso. En la [Figura 7](#), se puede observar una operación de *Pooling* con un *kernel* y un *stride* de  $2 \times 2$  donde cada uno de los valores de la salida se calcula a partir de los valores de la entrada con el mismo color.

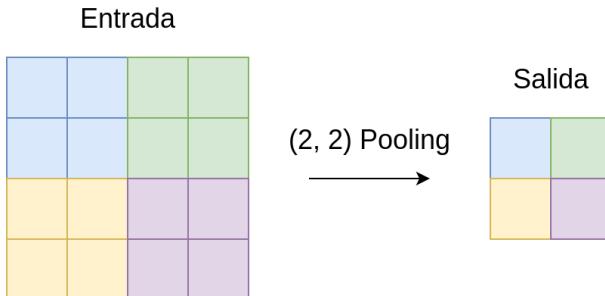


Figura 7: Operación de Pooling2D.

- **Convolución transpuesta:** La operación de convolución puede, dependiendo del tamaño del *kernel*, producir un resultado de dimensiones espaciales (altura y anchura) iguales o menores que las de la entrada. Sin embargo, para obtener una salida con mayores dimensiones que la entrada es necesario recurrir a la convolución transpuesta, una operación similar a la convolución donde para obtener los valores de la salida, se multiplican todos los valores del *kernel* por solo uno de los valores de la entrada, que se recorren individualmente. Al hacer el cálculo de esta forma, dependiendo de la zancada (*stride*) que se elija, pueden solaparse valores de la salida, por lo que se suman aquellos que hayan obtenido un valor en más de una operación ([Figura 8](#)).

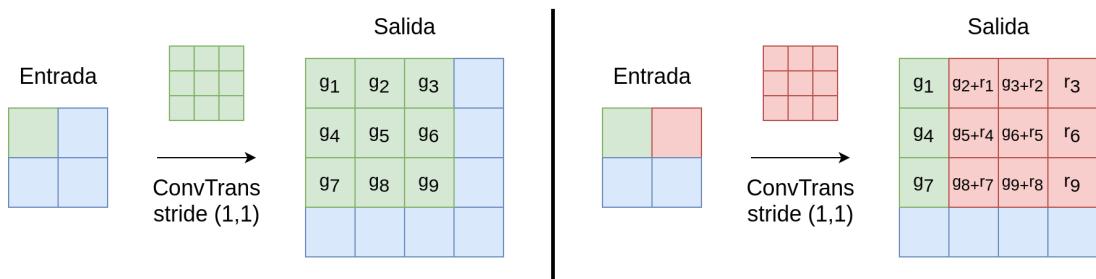


Figura 8: Primeros pasos de una convolución transpuesta.

- **Convolución 1x1 (*Pointwise Convolution*):** Pese a que no es un tipo de capa distinta por sí, esta aplicación particular de la convolución suele emplearse junto a otro tipo de capas o como proyección para modificar el número de canales. Su principio de funcionamiento es el uso de *kernels* de tamaño  $1 \times 1$  y tantos canales como mapas de características tenga la entrada. De esta forma, cada uno de los filtros produce un solo canal de igual altura y anchura que la entrada en la salida, permitiendo por lo tanto controlar con el número de filtros de la capa convolucional el nuevo número de canales en el que se proyecta la entrada ([Figura 9](#)).

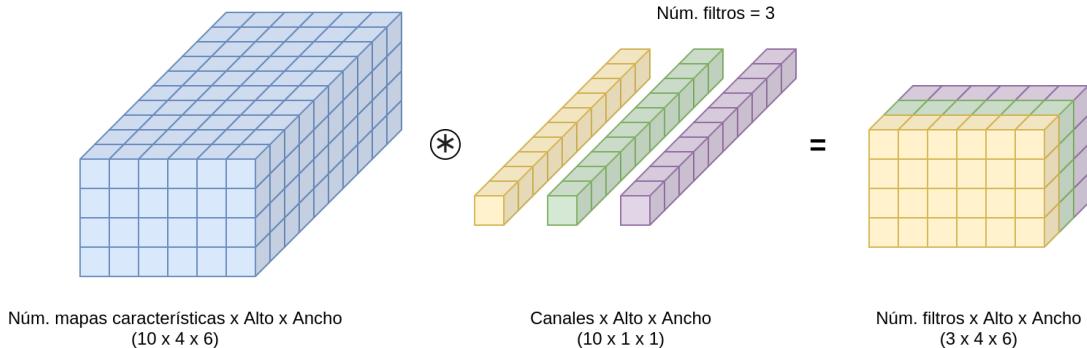


Figura 9: *Pointwise convolution.*

- Convolución por profundidad separable (*Depthwise Separable Convolution*): Similar al caso anterior, este cálculo es mas bien un caso especial de la capa de convolución. La operación de convolución por profundidad separable se compone de dos partes:
  - Una convolución por profundidad donde los *kernels* solamente tienen un canal y se aplican de forma individual en solo uno de los canales (el número de *kernels* tiene que ser igual al número de canales en la entrada) para luego agrupar los resultados (Figura 10).
  - Una convolución con un solo *kernel* de tamaño  $1 \times 1$ , fusionando los canales individuales resultantes de la operación anterior en uno solo. De esta forma, la operación se asemeja a una convolución corriente en la que el *kernel* tendría tantos canales como la entrada, pero reduciendo muy considerablemente el número de operaciones.

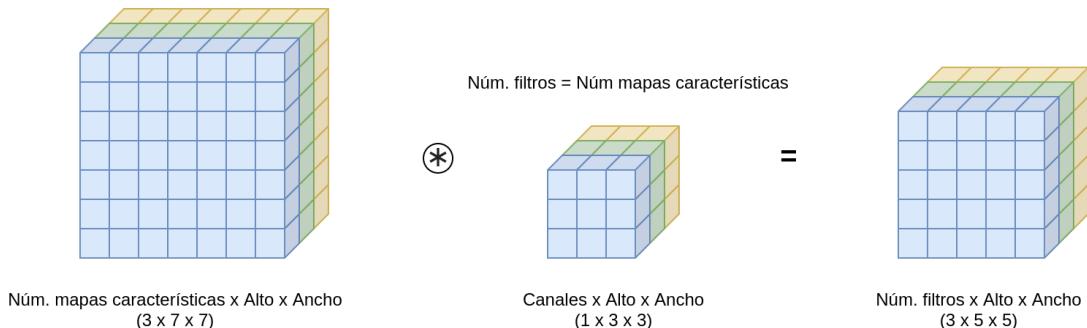


Figura 10: *Depthwise convolution.*

### 2.5.1. Conexiones residuales y ResNets

En el artículo publicado por He et al. [19] se presentan las conexiones residuales y la familia de redes ResNet, compuesta por distintos modelos en función de su número de capas. En el momento en el que se publicó dicho estudio, gracias a las conexiones residuales, se consiguió multiplicar por aproximadamente 7 el número de capas en los modelos del estado del arte (de las 22 capas de GoogleLeNet [20] a las 152 capas de ResNet152). Para conseguir esto, las ResNet se apoyan en las conexiones residuales (Figura 11), que permiten aumentar tanto la capacidad de aprendizaje de las redes como la calidad de sus resultados.

Las conexiones residuales están motivadas por la idea de que dada una red con una profundidad óptima para una tarea, en teoría, por muchas capas que se incluyesen, el descenso del gradiente sería capaz de hacer que las capas adicionales convergiesen en una función identidad y no perjudicasen los resultados. No obstante, empíricamente se ha comprobado que con las redes neuronales llega un punto en el que añadir más profundidad hace que los resultados empeoren. Aquí es donde entran las conexiones residuales, ya que al pasar hacia adelante la entrada de un bloque de capas y sumarlo a su salida ([Figura 11](#)), se facilita que el descenso del gradiente reste importancia a la señal que ha atravesado dicho bloque ( $F(x)$ ) en caso de que no mejore la salida final de la red.

Además, para reducir el uso de recursos computacionales al incrementar el número de capas, las ResNet utilizan conexiones residuales con cuellos de botella (*bottleneck*) para reducir el tiempo de ejecución. Este sistema, en vez de usar como  $F(x)$  ([Figura 11](#)) dos capas convolucionales, emplea tres, siendo la primera y la tercera convoluciones de filtros  $1 \times 1$  que se encargan de reducir el número de canales y posteriormente incrementarlo para que se pueda sumar con la entrada. De esta forma, se reducen en gran medida el número de operaciones a realizar en la capa convolucional central, con filtros de tamaño  $3 \times 3$ .

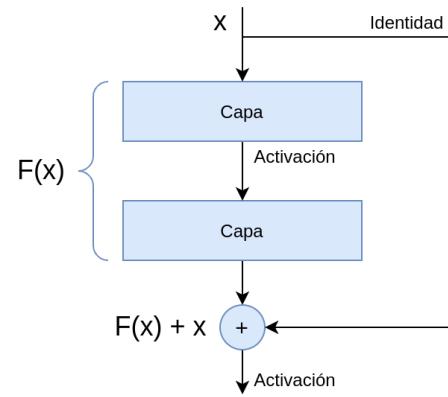
### 2.5.2. Efficientnet

Posteriormente, el artículo *EfficientNet: Rethinking Model Scaling for Convolutional Neuronal Networks* [[21](#)] propone un método para escalar de forma proporcional la profundidad (número de capas), anchura (número de filtros) y resolución de entrada de las redes convolucionales de forma que se puedan adaptar los modelos a los recursos disponibles. Además, esta publicación también presenta la red **EfficientNet-B0** (inicialmente diseñada para ilustrar el escalado proporcional propuesto) y su familia de redes, que se obtienen escalando dicho modelo. Sin embargo, más allá de ilustrar el escalado en anchura, profundidad y resolución, la arquitectura de la familia de EfficientNets obtiene resultados muy competitivos con un número de parámetros muy reducido en comparación con el de sus competidores.

Para diseñar esta red se extiende el trabajo de Tan et al. [[22](#)], y se emplea *Neural Architecture Search - NAS* (una técnica que automatiza la búsqueda y configuración de arquitecturas de redes neuronales) buscando optimizar el *accuracy* y el número de operaciones en coma flotante (FLOPS). Como resultado, se obtiene una arquitectura basada en bloques de capas MBConv (*Mobile Inverted Bottleneck Residual Convolution*) [[23](#)] y *Squeeze-and-Excitation* [[24](#)].

La primera de estas dos capas, (MBConv) se propone en la publicación de **MobileNetV2** [[23](#)] y modifica los bloques con conexiones residuales y cuello de botella invirtiéndolos. En vez de utilizar filtros de  $1 \times 1$  para reducir el número de mapas de características y después aumentarlo, este bloque emplea dichos filtros para aumentar y luego reducir el número de canales. Sin embargo, emplea convoluciones por profundidad separables (*Depthwise Separable Convolutions*) para reducir enormemente el número de operaciones al aumentar el número de canales.

La segunda, es uno de los bloques principales de la **SENet** [[24](#)] y basa su funcionamiento en dos operaciones que realmente pueden añadirse a cualquier otra capa: el mecanismo de *squeeze*, que crea un descriptor de cada canal de la entrada agregando sus características espacialmente; y el mecanismo de *excitation*, que toma como entrada los descriptores anteriores y los utiliza para modular los pesos de sus canales correspondientes.



[Figura 11: Conexión residual básica.](#)

En español exactitud me suena feo y precisión me suena a precision

## 2.6. Transformers

La arquitectura inicial del *transformer* (Figura 12), propuesta en *Attention is All You Need* [3], se basa en una estructura *encoder-decoder*. Es decir, un conjunto de capas (*encoder*) que codifica la entrada en una representación latente, que después es tomada como entrada del *decoder*, otro conjunto de capas que decodifica esta representación latente en una salida útil. En la propuesta inicial, destinada a procesamiento de lenguaje natural, el *encoder* se encarga de trasladar una secuencia de entrada  $(x_1, \dots, x_n)$  - una frase - en una secuencia de representación  $(z_1, \dots, z_n)$ . Esta secuencia  $z$ , es la entrada del *decoder*, que la convierte en una secuencia de salida  $(y_1, \dots, y_m)$  - otra frase -.

Una de las principales diferencias frente a los modelos *encoder-decoder* basados en redes recurrentes, es que mientras que en las redes recurrentes la entrada tiene que procesarse de forma secuencial para poder obtener los estados ocultos de las entradas anteriores, en los *transformers* la secuencia de entrada no está alineada temporalmente con la ejecución del modelo, ya que se procesa toda la entrada simultáneamente. De esta forma, se acelera muy significativamente el entrenamiento y la inferencia del modelo.

### 2.6.1. Arquitectura

En el **encoder** (parte izquierda de la Figura 12), se encuentra un *stack* de 6 capas. Cada una de estas, está compuesta por dos subcapas: una capa de *Multi-Head Self-Attention* (un mecanismo de atención que se explicará más adelante); y una capa que contiene una red *feed-forward* totalmente conectada. Cada una de estas subcapas, cuenta además con una conexión residual, que conecta la entrada de la subcapa con su salida de forma que puedan ser sumadas y normalizadas. Para facilitar la suma y normalización de entradas y salidas, todas las capas del modelo producen elementos de dimensión  $d = 512$ . Antes de estas 6 capas, cada uno de los *tokens* - elementos de la secuencia - de entrada (en la propuesta inicial, palabras), se convierten a vectores de dimensión  $d$  a través de un *embedding*<sup>3</sup> previamente entrenado y se les añade una codificación posicional (en esta propuesta, generada a partir de funciones seno y coseno de distintas frecuencias) que aporta al modelo información sobre la posición de cada *token* dentro de la secuencia inicial.

Por otro lado, en el **decoder** (parte derecha de la Figura 12), se vuelve a encontrar un *embedding* previamente entrenado que transforma las salidas del modelo desplazadas una posición. Al resultado de este *embedding*, se le añade una codificación posicional similar a la del *encoder*. A continuación, hay otro *stack* de 6 capas, que esta vez está compuesto por las dos subcapas que

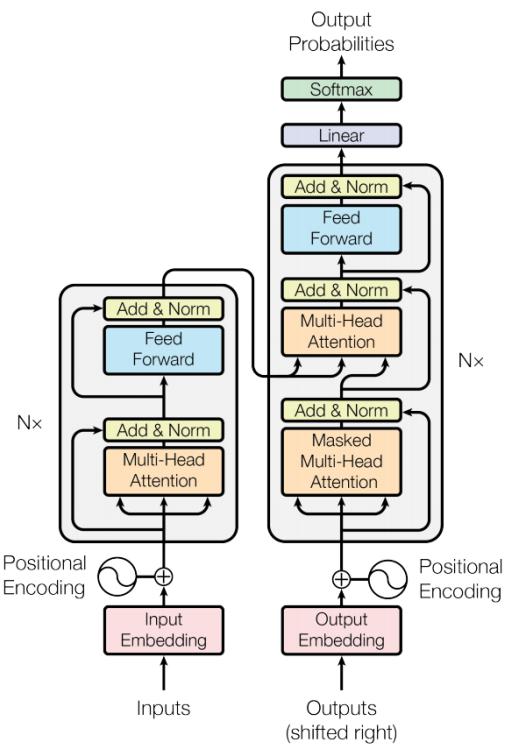


Figura 12: Arquitectura del *transformer*.

Fuente: [3]

<sup>3</sup>Operación que transforma, en el caso de la publicación original, palabras, en una representación numérica en un espacio vectorial donde las palabras con significado similar se encuentran próximas entre sí

están presentes en el *encoder* (en este caso la capa de *Multi-Head Self-Attention* es en realidad *Multi-Head Masked Self-Attention* ya que se aplica una máscara para evitar que influyan en la red los *tokens* siguientes al *token* que se va a predecir), y una subcapa adicional de *Multi-Head Cross-Attention*, situada entre las otras dos subcapas, donde las salidas de la subcapa de *masked self-attention* del *decoder* pueden acceder a las salidas del conjunto de capas del *encoder*. (**La entrada de esta última capa de atención proviene de la última capa del encoder, no de sus capas intermedias**). Por último, a la salida del *stack* de capas del *decoder*, se encuentra una transformación lineal (entrenada) y una función softmax para predecir la salida de la red.

## 2.6.2. Mecanismos de atención

Los mecanismos de atención, presentados por primera vez en [25], buscan simular la atención cognitiva y han sido previamente empleados en redes recurrentes [26] y convolucionales [27, 28] para aprender qué partes de la entrada son más relevantes en la tarea a completar. Sin embargo, en [3], con los *transformers*, se propone por primera vez una arquitectura basada solamente en estos mecanismos.

Las funciones de atención más empleadas son la atención aditiva [25] y multiplicativa, siendo esta última la empleada en los *transformers*, donde la atención se consigue a través de un producto escalar dentro de un bloque con múltiples cabezas llamado *Multi-Head Attention*, elementos principales de los *transformers*, que aparecen de dos formas distintas:

- Bloques de *Self-Attention*, en el *encoder* y en el *decoder*, con todas las entradas dentro de sus respectivas capas.
- Bloques de *Cross-Attention*, en las capas del *decoder*, con entradas provenientes del final de la pila de capas del *encoder* y de la subcapa anterior del *decoder*.

Cada una de las cabezas que componen estos bloques basan su funcionamiento en multiplicar sus entradas por una serie de matrices  $W^V$ ,  $W^K$  y  $W^Q$  que son aprendidas durante el entrenamiento, de donde se obtienen, respectivamente, vectores *Value* (V), *Key* (K) y *Query* (Q). Estos vectores, permiten que cada uno de los elementos de la secuencia de entrada, con el cálculo asociado a la atención (Figura 13) soliciten a través de su vector *Query* la información que determinen más importante de la secuencia. Esto se consigue al calcular el producto escalar de todos los vectores Q con todos los vectores K, que resultará mayor cuanto más alineados estén ambos vectores - mayor similitud entre las *Queries* (consultas) y las *Keys* (claves) -. A los resultados de estos productos escalares, se les aplica una función *SoftMax* para asegurar que sumen una unidad y finalmente se multiplican por los vectores V para obtener el resultado final de la atención. Los resultados de todas las cabezas, se concatenan en una sola matriz para aprovechar al máximo el procesamiento en paralelo y atravesan una última proyección lineal.

Al usar solamente mecanismos de atención y cruzar toda la secuencia de entrada con sí misma, la longitud de esta deja de ser relevante, ya que cualquier elemento de la entrada puede obtener información de otro elemento, independientemente de lo alejado que esté (a diferencia de las redes recurrentes, donde la información de los elementos alejados perdía significado).

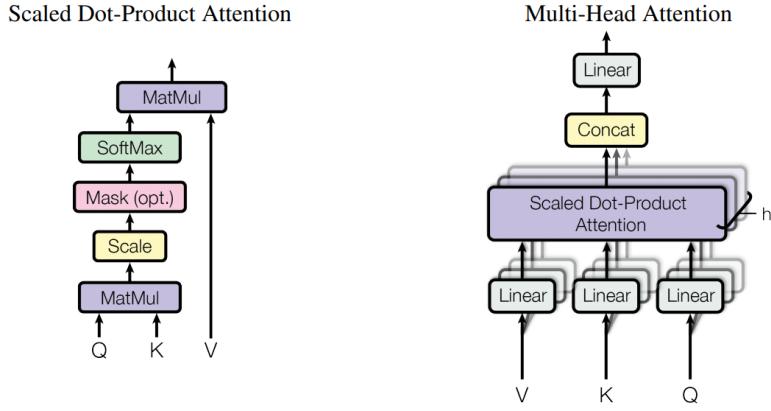


Figura 13: Producto escalar para el cálculo de la atención (izquierda) y bloque de *Multi-Head Attention* (derecha). Fuente: [3]

No obstante, estos mecanismos de atención son considerablemente costosos, ya que tienen una complejidad  $O(n^2)$  tanto en tiempo como en memoria, siendo  $n$  el número de elementos de la secuencia de entrada (al bloque de atención). Es por esta razón por lo que han ido surgiendo una serie de propuestas para reducir dicha complejidad computacional, algunas de las cuales se exponen en la siguiente sección.

### 2.6.3. Atención eficiente

Para reducir los requisitos de memoria y el coste computacional de los *transformers*, que tal y como se ha mencionado anteriormente, obtienen sus resultados gracias a los mecanismos de atención, pero que sin embargo son muy costosos computacionalmente, surgen distintas técnicas que buscan aproximar el resultado de la multiplicación de matrices que se lleva a cabo en los bloques de atención. Siguiendo el esquema propuesto en [29], estas modificaciones de las capas de atención pueden agruparse en:

**2.6.3.1. Patrones fijos - Fixed Patterns (FP)** : En los patrones fijos, la longitud de la secuencia de entrada a los mecanismos de atención se reduce, por ejemplo: accediendo a ella en bloques de un tamaño determinado, en esto se basan *Blockwise Attention* [30] y *Local Attention* [31]; accediendo a la secuencia en intervalos previamente definidos, como en *Sparse Transformer* [32] o *Longformer* [33] donde se emplean ventanas dilatadas o con un determinado *stride* (zancada); o también empleando operaciones de *pooling* para reducir la longitud de la entrada, en *Compressed Attention* [34].

**2.6.3.2. Combinación de patrones - Combination of Patterns (CP)** : Este grupo de métodos, se basa principalmente en la combinación de diferentes patrones de acceso a los elementos que componen la secuencia de entrada y aparece en *Sparse Transformer* [32] donde se combinan *Local Attention* y *strided attention* asignando la mitad de las cabezas del bloque de *multi-head attention* a cada uno de los métodos. También aparece en *Axial Transformer* [35], donde la atención se aplica de forma independiente en cada uno de los ejes de la entrada (en este caso, el tensor de entrada debería ser multidimensional).

**2.6.3.3. Aprendizaje de patrones - Learnable Patterns (LP)** : Pese a ser similar a los dos casos anteriores ya que siguen basándose en diferentes formas de acceder a la secuencia de entrada para hacerla más dispersa, estos métodos son capaces de aprender en la etapa de entrenamiento del modelo qué patrones de acceso son más adecuados. Algunas de las propuestas

que emplean este tipo de patrones son **Reformer** [36], que agrupa los elementos de la secuencia de entrada (*tokens*) empleando una medida de similitud en grupos de elementos (para posteriormente aplicar el mecanismo de atención de forma independiente en cada grupo) o **Routing Transformer** [37] que emplea k-medias para agrupar los *tokens*, ambos modelos, reducen la complejidad a  $O(n \log n)$ . Dentro de estos modelos también destaca **ResT** [38], que está enfocado a imágenes y emplea convoluciones separables para reducir las dimensiones de las entradas al mecanismo de atención.

**2.6.3.4. Disminución de rango** : Este conjunto de arquitecturas, incluyen una proyección para conseguir una aproximación de la matriz resultante del cálculo de la atención, esta aproximación, pese a tener el mismo número de elementos (filas), obtiene una representación de los vectores menor (columnas), por lo que la dimensión de la matriz pasa de ser  $n \times n$  a ser  $n \times k$ , con la consecuente disminución de coste computacional. El principal ejemplo de este tipo de arquitectura es **Linformer**, [39] que presenta una complejidad  $O(n)$

#### 2.6.4. Performer

En este trabajo, se estudia la influencia de un mecanismo de atención eficiente en concreto al sustituir la atención estándar de DPT. El mecanismo elegido, y por tanto, el que se explica con mayor profundidad en esta sección es el propuesto por Choromanski et al. en el artículo *Rethinking Attention with Performers*. Este mecanismo de atención eficiente, reduce la complejidad de la atención de  $O(n^2)$  a  $O(n)$  aproximando la multiplicación de matrices empleando *kernelización* y por lo tanto podría entrar en el grupo ya mencionado de **Disminución de rango**.

Al multiplicar primero  $K \times V$  y después  $Q \times Ans$ , se pasa de  $2L^2d$  multiplicaciones a  $2d^2L$  multiplicaciones, por lo que si  $d \ll L$  la complejidad pasa a ser lineal.

Figura similar a la del paper

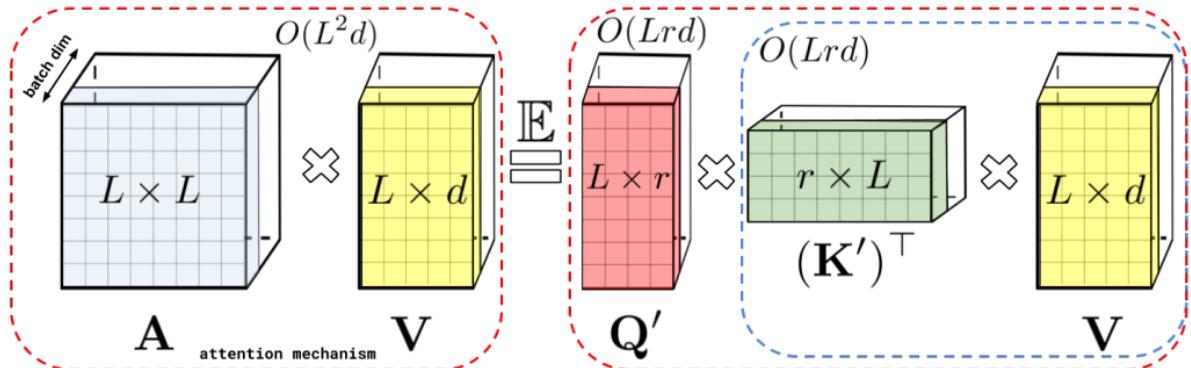


Figura 14: Temp TODO

Paper performer: <https://arxiv.org/pdf/2009.14794.pdf>

Video Kernels: [https://www.youtube.com/watch?v=y\\_RjsDH15Y4](https://www.youtube.com/watch?v=y_RjsDH15Y4)

Random Fourier Features

Video Performer: <https://www.youtube.com/watch?v=xJrKIPwVwGM&t=2s>

## 2.7. Transformers para visión artificial - ViT

A la hora de aplicar la arquitectura de los *transformers* a procesamiento de imágenes, surge un problema importante. La complejidad del mecanismo original de atención es  $O(n^2)$ , siendo  $n$  el número de elementos en la secuencia de entrada, por lo que para una imagen de dimensiones  $lado \times lado$ , el número de píxeles que conformarían la secuencia de entrada al mecanismo de atención es  $n = l^2$ , disparando la complejidad de la atención a  $O(l^4)$ . Para lidiar con este problema, se han propuesto distintas soluciones como limitar el mecanismo de atención al entorno de cada píxel (presentado en *Image Transformers* [40]) o aplicar convoluciones para reducir el tamaño de la secuencia de entrada [41].

Sin embargo, la solución propuesta en *An Image is Worth 16x16 Words* con el *Vision Transformer* [42] es la que mejores resultados ha obtenido y por lo tanto aquella que más popularidad ha ganado como base de arquitecturas para otros problemas de visión artificial [4, 43, 44, 45]. Esta solución, consiste en dividir la imagen original en fragmentos de tamaño fijo y convertir con una proyección cada fragmento en un vector de valores (*embedding*). Estos vectores, son equivalentes al resultado del *embedding* de palabras en la arquitectura original y se introducen al modelo como tal, es decir, cada fragmento extraído de la imagen correspondería a una palabra de una frase ([Figura 15](#)). Esta arquitectura, al ser inicialmente propuesta para realizar clasificación, solamente emplea el *encoder* del *transformer*.



Figura 15: *Backbone* basado en fragmentos y proyección lineal del ViT.

Figura inspirada en [42]

En esta publicación, además, se presenta una variante del *Vision Transformer*, el ***Hybrid Vision Transformer***. La diferencia entre esta arquitectura y la original es la sustitución del *backbone* del ViT, es decir, de la separación en fragmentos y la proyección lineal, por una red convolucional de cuyos mapas de características se extraen los vectores que pasan a los bloques de atención ([Figura 16](#)). Para formar estos vectores, en la implementación usada en este trabajo, se agrupan los mapas de características en función de sus píxeles, por lo tanto, si hay  $n$  mapas de características, el píxel número 1 de cada uno de ellos formará un vector de dimensión  $n$ . De esta forma, la dimensión del vector viene dada por el número de mapas de características y el número de *tokens* será equivalente al número de píxeles de los mapas de características (imágenes mayores llevan un mayor número de *tokens*).

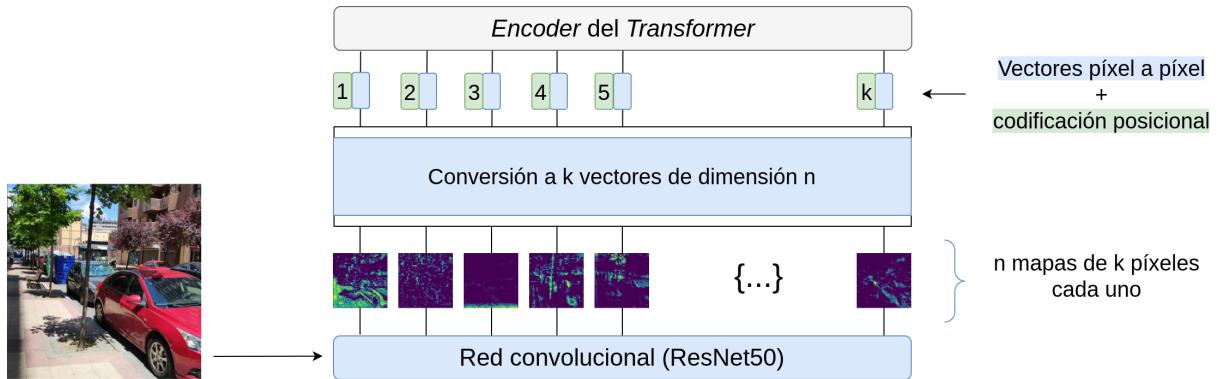


Figura 16: *Backbone* convolucional del ViT-Hybrid.

## 2.8. Estimación de profundidades

Otra de las bases fundamentales de este trabajo es la estimación de profundidades. Esta tarea, tanto cuando es llevada a cabo por humanos como por máquinas, consiste en detectar la distancia relativa entre todo aquello que se ve. Tal y como se ha mencionado en el capítulo de Introducción, el cerebro humano se apoya principalmente en la disparidad existente entre las imágenes que capturan cada uno de los ojos (estereovisión), ya que las cosas que están más lejanas ven su posición menos alterada entre la vista de un ojo y de otro que las cosas cercanas. Sin embargo, también capaz de estimar la profundidad en una sola imagen a partir de nuestro conocimiento previo del mundo y del entorno (pistas monoculares).

Las técnicas de visión artificial tradicional (basadas en dos cámaras - visión estereoscópica) no pueden lidiar con este problema, por lo que la estimación monocular (con una sola cámara) es prácticamente imposible empleando técnicas tradicionales. Para este tipo de estimación (monocular), entran en juego los modelos de aprendizaje automático, que han demostrado una gran capacidad para explotar el conocimiento sobre el entorno en todo tipo de tareas, no siendo la estimación de profundidades diferente.

A continuación, aunque haciendo especial hincapié en aquellos basados en aprendizaje automático, se enumeran los distintos enfoques existentes, incluyendo aquellos propios de la visión artificial tradicional, las soluciones *hardware*, y las técnicas de aprendizaje profundo.

### 2.8.1. Técnicas de estimación de profundidades

La estimación de profundidades se ha intentado resolver de múltiples maneras [46]. Dentro de estas metodologías, existen tres enfoques principales en función del tipo de *software* o *hardware* que se emplea.

- **Soluciones geométricas:** Este grupo de métodos extrae información de las restricciones geométricas que existen entre parejas de imágenes. Principalmente se agrupan en técnicas de *SfM* (*Structure from Motion*) donde se reconstruye la tercera dimensión a partir de imágenes tomadas por una sola cámara en movimiento, y en técnicas de estereovisión, donde la profundidad se obtiene de la disparidad entre las imágenes capturadas por una pareja de cámaras con posiciones relativas conocidas. Estos métodos, sin embargo, basan su funcionamiento en el emparejamiento de puntos claves (*feature points*) que deben encontrarse en ambas imágenes, requiriendo por lo tanto texturas o formas características que poder emparejar.
- **Soluciones hardware:** Por otro lado, existen una serie de soluciones basadas en distintos sensores como son los LIDAR, las cámaras de tiempo de vuelo (ToF) o los escáneres de luz estructurada. Estas soluciones, no obstante, cuentan con ciertas limitaciones como son la

densidad de sus capturas (representaciones de puntos dispersos en el caso del LIDAR<sup>4</sup>), la sensibilidad a la iluminación (en las cámaras ToF) o su rango de acción y la necesidad de un entorno controlado (luz estructurada).

- **Aprendizaje automático:** Debido a las restricciones de los métodos existentes y a los resultados obtenidos empleando aprendizaje automático en otros campos de la visión artificial, en los últimos años han surgido múltiples arquitecturas enfocadas a recuperar la profundidad a partir de imágenes. Este documento, pese a revisar superficialmente otras opciones, se centra en las **soluciones monoculares** debido al interés detrás de obtener una representación de la profundidad a partir de una sola cámara (coste, espacio, consumo, etc.).

### 2.8.2. Aprendizaje automático no supervisado

Estas propuestas, ofrecen soluciones que emplean datos sin etiquetar debido a la dificultad que entraña la obtención de este tipo de anotaciones. Normalmente, estas soluciones emplean secuencias (vídeos) de imágenes monoculares, extrayendo automáticamente a partir de estas una señal supervisora con distintos métodos. En [47] se propone una arquitectura que emplea una red neuronal para calcular la posición (y movimiento) de la cámara entre *frames* consecutivos (*ego-motion*), que a su vez se emplea para calcular la profundidad de la imagen. Sin embargo, da por hecho que ningún objeto ha cambiado de posición y que el entorno es estático.

Esto, no es aplicable a entornos reales, por lo que surgen diferentes arquitecturas que generan máscaras, tanto basadas en redes neuronales [47, 48] como en técnicas geométricas [49, 50], para lidiar con los objetos dinámicos. Otros enfoques, con la misma idea subyacente, sustituyen la red de estimación de posición con métodos de odometría visual<sup>5</sup> tradicional [51], que puede aportar posiciones más exactas y mejorar el funcionamiento final. Por último, algunos enfoques calculan elementos adicionales como por ejemplo el flujo óptico de la escena (cálculo de los patrones de movimiento para cada punto de la imagen) que aporta información relevante sobre la posición relativa de los objetos [48, 52].

### 2.8.3. Aprendizaje automático semisupervisado

Debido también a la escasez de datos etiquetados, también son populares los enfoques semi supervisados. Estas soluciones, emplean información parcial como señal supervisora, que complementan con información no etiquetada. Algunos ejemplos característicos de este tipo de aprendizaje son:

- Síntesis artificial de parejas a partir de imágenes monoculares para emplear técnicas de estereovisión, entrenando los modelos con parejas de imágenes obtenidas por conjuntos de cámaras preparados para estéreo [53, 54].
- Generación de los mapas de disparidad entre parejas de imágenes para estereovisión [55, 56], donde la señal supervisora es la disparidad obtenida mediante técnicas tradicionales de estereovisión.
- Utilización de datos etiquetados de forma dispersa (frecuentemente obtenidos con dispositivos LIDAR), que o bien emplean imágenes no etiquetadas para densificar dichas repre-

<sup>4</sup>Por representaciones dispersas se entiende la captura de la profundidad en forma de una nube de puntos discretos separados entre sí, mientras que una representación densa tiene una mayor cantidad de puntos de información.

<sup>5</sup>La odometría agrupa aquellas técnicas que estiman el cambio de posición a partir de las lecturas de sensores. En el caso de la odometría visual, estos cambios de posición se estiman empleando principalmente imágenes capturadas por una cámara.

sentaciones [57, 58, 59] o añaden la información del LIDAR en la función de perdidas para después inferir a partir de imagen monocular [60].

#### 2.8.4. Aprendizaje automático supervisado

Pese a la dificultad para obtener datos etiquetados correctamente, los enfoques supervisados siguen siendo los que mejores resultados ofrecen y por lo tanto han sido y siguen siendo extensamente estudiados. El primero de estos modelos fue propuesto en [61], y empleaba dos conjuntos de capas, uno para generar una estimación tosca de la profundidad y otro para refinar esa primera estimación. Enfoques posteriores propusieron modificaciones en la función de perdidas para fomentar la consistencia en las predicciones [62] a través del cálculo de los gradientes de la diferencia entre el resultado y el objetivo. Además de aquellos basados en arquitecturas *encoder-decoder*, también hay propuestas con arquitecturas de aprendizaje adversario [63] donde el discriminador trata de distinguir entre la profundidad generada y la real [64].

Todas estas soluciones, basan su funcionamiento en redes convolucionales, sin embargo, los *transformers* han presentado muy buenos resultados en los últimos años y también se han propuesto soluciones que emplean estas arquitecturas, en el momento de redactar este documento, predominan dos: los **Dense Prediction Transformers** (DPT) [4] y **AdaBins** [43].

El primero de estos dos modelos, al ser la arquitectura que se modifica a lo largo del proyecto, se explica en mayor profundidad en la siguiente sección. El segundo, *AdaBins*, está compuesto de un *encoder* convolucional y una etapa inspirada en los *Vision Transformers* [42] que recibe la salida de el *encoder* y clasifica cada píxel en un histograma de profundidades cuyas barras (*bins*) son parametrizadas (centro y rango) dinámicamente para cada imagen. El resultado final se consigue con la suma ponderada de la predicción de pertenencia a cada barra y el valor medio de dicha barra, consiguiendo así una estimación de profundidad suavizada.

#### 2.8.5. DPT

Tal y como se ha mencionado previamente, la gran mayoría de arquitecturas de estimación de profundidades se basan en redes convolucionales, normalmente de tipo *encoder-decoder*. Las líneas de investigación se centran en el *decoder* y sus estrategias de agregación de información. DPT (*Dense Prediction Transformer*) [4] también se estructura como un *encoder-decoder*, pero centra su estudio en la modificación del *encoder* debido a la gran influencia que tiene este en la información que llega a la segunda parte de la red, sustituyéndolo por una arquitectura basada en mecanismos de atención (*transformer*). De esta forma, gracias a las características propias de estos modelos ya comentadas, la arquitectura se beneficia de un campo receptivo que permite atender a la información de toda la imagen en las etapas de atención. Es decir, se pasa de un campo receptivo local mucho más limitado (en las redes convolucionales) a un campo receptivo global.

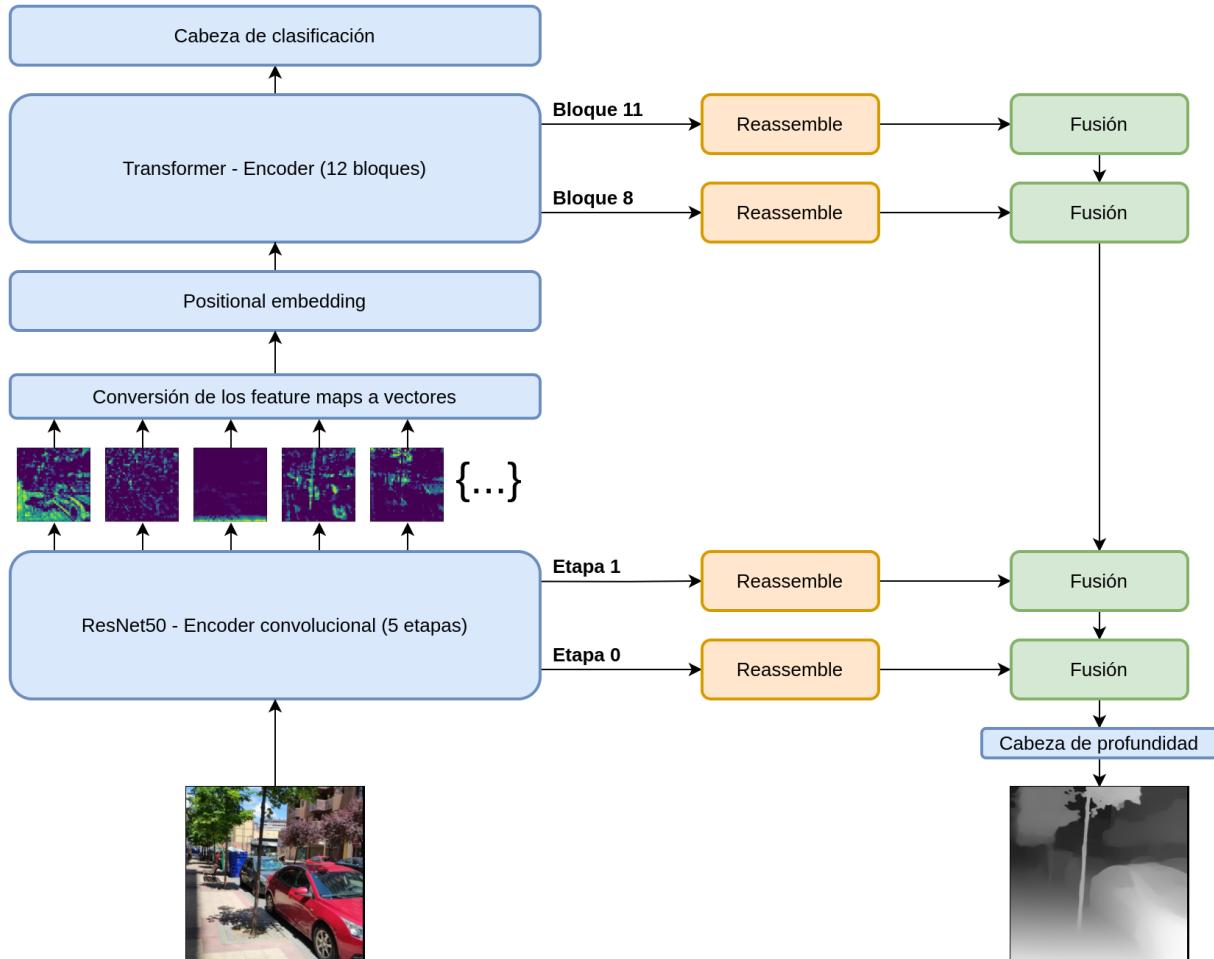


Figura 17: Esquema de la arquitectura de DPT-Hybrid.

En la publicación que presenta este modelo se utilizan tres variantes del *Vision Transformer* como **encoder** para llevar a cabo los experimentos: ViT-Base, con *embedding* de las imágenes a partir de fragmentos y 12 bloques de atención; ViT-Large, con el mismo método de *embedding* (esta vez con vectores de mayor dimensión para representar cada uno de los fragmentos codificados) y 24 bloques de atención, y por último, ViT-Hybrid, con el mecanismo de *embedding* de las entradas basados en una red convolucional (ResNet50) tal como se indica en la [Figura 17](#) seguido de 12 bloques de atención.

Por otro lado, el **decoder** de la arquitectura está compuesto por una serie de capas convolucionales que transforman representaciones extraídas de distintas etapas del *encoder*. En el caso del ViT-Base, las representaciones a la salida de los bloques de atención {3, 6, 9, 12}, del ViT-Large, las correspondientes a las salidas de los bloques de atención {5, 12, 18, 24}, por último, en el caso del ViT-Hybrid, son las representaciones obtenidas de los bloques de atención {9, 12} junto con las representaciones a la salida de la primera y segunda etapa de la ResNet50 encargada del *embedding*.

Estas representaciones, como están en forma de *tokens* en el caso de los bloques de atención, o en forma de mapas de características cuando se extraen de la ResNet50, atraviesan un bloque *Reassemble* donde la información se concatena para reconstruir la estructura de imagen ([Figura 18](#)), se proyecta en el número de canales deseado, y se escala a una resolución determinada en función de la altura del *encoder* de la cual provienen (cuanto antes se extraen las características, menos se reduce su tamaño, haciendo una especie de pirámide de resolución).

Una vez se dispone de los datos en un formato similar a imágenes (a la salida de los bloques de *Reassemble*), pasan a una etapa similar a la arquitectura RefineNet [65], que duplica el tamaño de cada una de las representaciones y las suma de forma escalonada (Figura 18 y Figura 17) para fusionar su información.

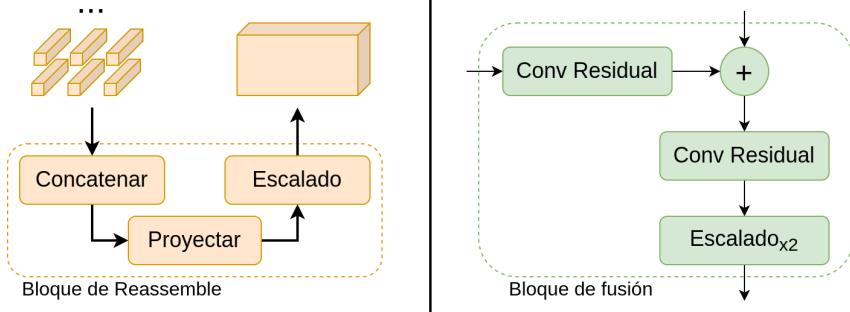


Figura 18: Bloques base del *decoder* de DPT. Izquierdo: Bloque de *Reassemble*. Derecho: Bloque de fusión. Figura adaptada de [4].

El resultado de este último bloque (una imagen con la mitad de la resolución de la entrada), pasa a una última etapa (cabeza especializada<sup>6</sup>) para proporcionar la predicción final.

Antes de pasar a la metodología empleada en el trabajo original de DPT para entrenar los modelos, cabe mencionar que la red, tal y como se comenta a continuación, se entrena con una función de pérdida invariante a la escala y al desplazamiento. Esto significa que la profundidad obtenida del modelo es relativa y no un valor métrico medible. No obstante, los autores extraen valores de desviación y escala (ajustando por mínimos cuadrados las predicciones y las anotaciones en el conjunto de entrenamiento) con los que es posible transformar las profundidades relativas obtenidas en entradas nuevas a profundidades métricas (siempre y cuando la naturaleza de las entradas sea la misma que la de los ejemplos utilizados para obtener los valores de desplazamiento y escala).

## Entrenamiento

Los modelos proporcionados junto a la publicación, y en concreto, el que se emplea como base de este trabajo, fueron preentrenados en MIX6 (Sección 3.3.2) y después ajustados a otros *datasets* para su evaluación. El entrenamiento en MIX6 se lleva a cabo empleando Adam como optimizador con un *learning rate* de  $1e-5$  para el *encoder* y de  $1e-4$  para el *decoder* y se parte de parámetros preentrenados en Imagenet (*encoder*) y inicializados aleatoriamente (*decoder*). En cuanto a la duración del entrenamiento, el modelo se entrena durante 60 épocas, cada una con 72000 pasos y un *batch size* de 16 imágenes. Además, se voltean horizontalmente las imágenes de manera aleatoria como forma de *data augmentation*. Como función de coste, se utiliza una función de pérdida [66] invariante a la escala (relativa) y al desplazamiento de la salida junto a otra función de pérdida que compara los gradientes<sup>7</sup> [67] en los valores de profundidad de la salida y la anotación.

## 2.9. Optimización de modelos

Los modelos de aprendizaje automático del estado del arte tienden a ser cada vez más grandes. Esto generalmente conlleva un tiempo de entrenamiento y de inferencia mayor, así como unos

<sup>6</sup>En la publicación original, se adapta la red para segmentación semántica y para estimación de profundidades monocular. Dado que escapa del alcance de este proyecto, no se discuten en este documento las capacidades de DPT para segmentar objetos.

<sup>7</sup>A la hora de entrenar en otros *datasets*, desactivan esta función de pérdida si la anotación no es profundidad densa, es decir, si hay píxeles sin anotar como ocurre por ejemplo en KITTI.

requisitos de memoria y consumos de energía mayores. Para solucionar algunos de los problemas asociados, como por ejemplo la necesidad de proporcionar buenos resultados con restricciones temporales o restricciones de *hardware* (modelos embebidos, dispositivos móviles, etc.) han ido surgiendo a lo largo de los años una serie de soluciones que tratan estos aspectos, buscando siempre perjudicar lo mínimo posible los resultados de los modelos originales. A continuación, se presentan algunas de las más generales, aplicables en una gran variedad de arquitecturas.

Aquí puede que se profundice demasiado en algunos temas (poda, por ejemplo)

- **Poda / Pruning:** Consiste en eliminar de los modelos aquellas conexiones o neuronas que son redundantes o menos relevantes para la red, con el objetivo principal de reducir el tamaño del modelo, ya que las redes neuronales suelen estar sobredimensionadas y son redundantes. Al reducir el tamaño del modelo, aumentar la velocidad con la que se realiza la inferencia<sup>8</sup> sin sacrificar la exactitud del modelo en exceso. Siguiendo la clasificación presentada en [68], los métodos más empleados de poda se pueden agrupar en dos grandes grupos: Basados en magnitud y basados en sensibilidad.
  - **Métodos basados en magnitud:** Estas técnicas, eliminan los parámetros basándose en su valor o en la influencia que tienen en la siguiente capa. Por ejemplo, un peso con un valor muy próximo a cero apenas influirá en la capa siguiente, y por lo tanto puede eliminarse. Estas técnicas, suelen llevarse a cabo eliminando parámetros y reentrenando la red de forma iterativa, repitiendo este proceso hasta encontrar un equilibrio entre reducción de tamaño y pérdida de exactitud. Han et al. [69] presentaron, empleando este entrenamiento recursivo, resultados donde se eliminan más de un 90 % de los parámetros aumentando solamente en unas décimas el porcentaje de error en comparación con los modelos sin podar. Además de la poda de conexiones y neuronas, existen también distintas técnicas destinadas a podar con una menor grano de detallado, por ejemplo, mapas de características y filtros. Algunas de estas técnicas incluyen las basadas en la varianza entre canales [70] o en el número medio de ceros que tienen los mapas de características [71] entre otras.
  - **Métodos basados en sensibilidad:** Estos métodos, a diferencia de los basados en magnitud, buscan analizar el efecto de la modificación de los pesos en la función de perdida. Para ello, suelen centrarse en aproximar los cambios en la función de perdida a través de una serie de Taylor. Esta serie de Taylor, incluye una matriz hessiana, que se obtiene a partir de las segundas derivadas de la perdida respecto de los pesos. Dado que el cálculo de la matriz hessiana es computacionalmente costoso, Lecun et al. en *Optimal Brain Damage (OBD)* [72] ignora los elementos que no están situados en la diagonal de la matriz, reduciendo la complejidad del cálculo considerablemente. Posteriormente, Hassibi et al. plantearon no descartar dichos valores en *Optimal Brain Surgeon (OBS)* [73] pero sus cálculos son prohibitivos con el número de parámetros de las arquitecturas actuales. Estas series de Taylor, también se han empleado para la poda de canales y mapas de características en redes convolucionales, tanto con aproximaciones de primer orden [74] como de segundo orden [75].
- **Quantization:** La cuantificación tiene como objetivo convertir los parámetros de las redes almacenados en 32 bits en representaciones más pequeñas como son los números enteros (normalmente en 8 bits). Esta transformación, conlleva una pérdida de calidad en los resultados de los modelos, pero reduce sus requisitos de memoria y acelera la inferencia de resultados. Esta aceleración viene dada por la velocidad a la que se pueden realizar

<sup>8</sup> Dependiendo de las técnicas utilizadas para llevar a cabo la poda, pueden aparecer limitaciones en el hardware de uso general para trabajar con matrices dispersas, pero existen soluciones tanto hardware como software para trabajar con estos datos.

operaciones con números enteros comparado con las operaciones con números en coma flotante. Existen principalmente tres tipos de cuantificación, dinámica, estática (estas dos se aplican sobre un modelo ya entrenado) y durante el entrenamiento (*Quantization Aware Training*).

- **Cuantificación dinámica:** En este caso, no solo se convierten los pesos del modelo ya entrenado a enteros, sino que también se transforman las activaciones buscando los parámetros de dichas conversiones de forma dinámica durante la inferencia. Este tipo de cuantificación no requiere de datos pero sin embargo no es tan rápida como las otras dos al tener que realizar las transformaciones durante la inferencia.
- **Cuantificación estática:** De forma similar al caso dinámico, las activaciones de las capas se cuantifican durante la inferencia, sin embargo, en este caso una vez se ha entrenado la red se le pasan bloques adicionales de datos con los que se estiman los parámetros de estos procesos de cuantificación para acelerar la inferencia cuando se haya desplegado el modelo. De esta forma, pese a que es necesario tener datos adicionales (no hace falta que estén etiquetados), se alcanza una mayor velocidad de inferencia.
- ***Quantization Aware Training*:** Por último, esta opción tiene en cuenta la cuantificación durante todo el proceso de entrenamiento simulando el efecto de la cuantificación en los pesos y activaciones de forma que influyan en la función de perdidas (las operaciones durante el entrenamiento siguen haciéndose en coma flotante). Este método, debido a la consideración de la cuantificación durante el entrenamiento, resulta en una inferencia más rápida y resultados superiores a los de los métodos anteriores. Sin embargo, no siempre es aplicable al requerir el entrenamiento del modelo.
- **Weight clustering:** El *clustering* de pesos, o *weight sharing*, agrupa los pesos del modelo en un número determinado de *clusters* para asignar a cada peso el valor del centroide de su grupo correspondiente. De esta forma, se reducen los requisitos de memoria del modelo, ya que solamente es necesario almacenar los índices que apuntan al vector de centroides, que al ser números enteros se pueden representar con un número de bits mucho menor (por ejemplo, en 8 bits, reduciendo el tamaño de la matriz de pesos original (cada uno 32 bits) a un cuarto de su tamaño).
- **Mixed-precision training:** Propuesto por primera vez en [76], el entrenamiento con precisión mixta almacena los pesos, activaciones y gradientes en formato de coma flotante de media precisión (16bits - IEEE 754) en vez de simple precisión (32 bits). De esta forma, sin perder precisión, se reducen a cerca de la mitad los requisitos de memoria en el entrenamiento, que además se ve acelerado en las últimas arquitecturas de GPUs. Para conseguirlo se llevan a cabo tres estrategias: Mantener una copia maestra de los pesos almacenada en FP32, escalar el resultado de la función de perdida y acumular ciertos resultados en FP32.
  1. **Copia maestra de los pesos en FP32:** A pesar de que en el *forward* y *backward pass* del entrenamiento se usan valores FP16, se debe almacenar una copia maestra en FP32 de los pesos, que son los empleados en el optimizador para multiplicar por gradientes y *learning rates*. Si bien es cierto que esto aumenta los requisitos de memoria de los pesos en un 50 %, el impacto en el conjunto global es mucho menor ya que el consumo de memoria durante el entrenamiento está dominado por las activaciones de cada capa.
  2. **Escalado de la perdida:** Por otro lado, al calcular los gradientes de los pesos, también podrían resultar demasiado pequeños como para representarse correctamente en FP16. Esto se soluciona escalando el resultado de la función de perdidas al finalizar el *forward pass* y antes de empezar el *backpropagation* para que tengan una magnitud

mayor. De esta forma, todos los gradientes resultan escalados por la misma magnitud al derivar y basta con reescalarlos a FP32 una vez finalizado el *backpropagation*, antes de la etapa del optimizador.

3. **Precisión aritmética:** Para poder asegurar los mismos resultados que con FP32, las multiplicaciones parciales de los productos escalares y las reducciones (sumas de todos los elementos) de vectores FP16 tienen que acumularse en valores FP32, que antes de escribirse en memoria se convierten a FP16.

### 3. Material y Métodos

En este tercer capítulo se introducen los materiales empleados para la realización del Trabajo Fin de Máster, tanto a nivel informático (*software* y *hardware*) como a nivel de datos. Después, se detalla las metodología seguida durante el proyecto para definir los distintos modelos de que se entrenarán, y por último se describen las herramientas empleadas durante el proceso de evaluación.

#### 3.1. Software

##### Lenguaje de programación y librerías

Para el desarrollo de este proyecto, se ha elegido como lenguaje de programación **Python 3.7.10** debido a su ecosistema de librerías y código abierto orientado al aprendizaje profundo. Junto con Python, se han empleado principalmente una serie de librerías y paquetes que pueden distinguirse en dos grupos:

- Uso de CPU: **Numpy 1.20** [77], para trabajar con matrices y acelerar operaciones matemáticas; **OpenCV 4.5.2** [78], para la carga y manipulación de imágenes antes de convertirlas en tensores, y **Matplotlib 3.4** [79] / **Seaborn 0.11** [80], para graficar resultados y otras figuras de este documento.
- Uso de GPU: **PyTorch 1.9.0** [81], para la creación, modificación, entrenamiento y evaluación de modelos de aprendizaje profundo acelerados por *hardware* (es decir, ejecutados en GPUs); **timm (PyTorch Image Models) 0.4.9** [82], desarrollado y mantenido por Ross Whigtman, que pone a disposición del usuario un gran número modelos del estado del arte preentrenados e implementados en PyTorch; el **repositorio de DPT** [4], modelo que se modifica a lo largo del proyecto; y por último, el repositorio **performer-pytorch 1.1.3** [83] de Phil Wang, que ofrece una implementación, también en PyTorch, de la arquitectura Performer y sus mecanismos de atención.

Algunas de estas librerías tienen alternativas que podrían haberse empleado perfectamente en este proyecto. La elección más importante es probablemente el uso de PyTorch frente a Tensorflow/Keras, ya que ambas librerías permiten construir y entrenar modelos de *Deep Learning* a partir de funciones y abstracciones que representan distintos tipos de capas, funciones de activación o procesos de transformación de datos, entre otras. Además, ambos paquetes de software ofrecen la posibilidad de ejecutar estos modelos, así como sus entrenamientos y evaluaciones en tarjetas gráficas dedicadas (GPU), reduciendo de forma drástica el tiempo necesario para completar entrenamiento e inferencia. Esta decisión, se ha tomado principalmente por la cada vez más frecuente elección de PyTorch en proyectos de investigación debido a su mayor flexibilidad. Consecuencia directa de esto, es que gran parte de los repositorios de código relacionados con publicaciones científicas recientes (por ejemplo, DPT) usan PyTorch para crear sus modelos.

##### Aceleración por *hardware*

En el párrafo anterior, se ha mencionado que PyTorch acelera por *hardware* el entrenamiento y la inferencia de los modelos de aprendizaje profundo. Para esto, se apoya principalmente en **CUDA** y **cuDNN**. El primero, es una plataforma de computación paralela desarrollada por NVIDIA para sus tarjetas gráficas dedicadas que permite desarrollar código para ejecutarlo en dichos dispositivos, aprovechando así el gran número de procesadores que tienen estos componentes. El segundo, también desarrollado por NVIDIA, es una librería de primitivas aceleradas por GPU preparadas para construir redes neuronales. Pese a que en este proyecto no se ha trabajado directamente con ninguna de estas herramientas, es necesario disponer de ellas ya que PyTorch las utiliza. Las versiones empleadas son, respectivamente, CUDA 11.1 y cuDNN 8.

## Gestión y seguimiento de experimentos

Dada la naturaleza del proyecto, era de esperar que el número de experimentos y de variaciones de modelos a entrenar fuese numeroso, por esta razón, se elige **Weights and Biases (wandb)** [?] para gestionar y monitorizar dichas pruebas, es decir, visualizar y controlar su evolución, registrando métricas y resultados para su posterior utilización. *Weight and Biases* es un servicio de seguimiento de experimentos, gratuito para uso académico y personal, que se ejecuta en la nube, dispone de una interfaz gráfica web (Figura 19) y permite registrar de forma sencilla variables y métricas durante las distintas ejecuciones que se lleven a cabo. Además, ofrece también un gestor de búsqueda de hiperparámetros, donde es posible configurar los valores que se quieren probar para que wandb se encargue de inicializar los scripts de entrenamiento con las configuraciones correspondientes de forma automática y coordinada en todas las máquinas en las que se ejecute su cliente. Ya que para el entrenamiento se han empleado varios equipos en paralelo, esta última característica se ha valorado muy positivamente al compararlo con otro *software* de monitorización como por ejemplo Tensorboard.



Figura 19: Interfaz web de *Weights and Biases*.

## Entorno de desarrollo

Para gestionar la instalación y ejecución de este conjunto de *software* en un entorno controlado, limitado, y fácilmente replicable, se ha elegido **Docker** junto al **NVIDIA Container Toolkit**. Docker proporciona una capa de abstracción virtualizando a nivel del sistema operativo. Esto significa que es capaz de utilizar el kernel de Linux de la máquina anfitrión, consiguiendo de esta forma ser mucho más rápido y eficiente que una máquina virtual. Por otro lado, el NVIDIA Container Toolkit envuelve el Docker Engine y mapea las primitivas de CUDA desde el interior del contenedor hasta el driver de la GPU del sistema anfitrión. De esta forma, la máquina anfitrión solo necesita tener actualizados los drivers de la(s) tarjeta gráfica para que puedan ser empleados de manera transparente por CUDA. Para el desarrollo, se parte de una de las imágenes proporcionadas por PyTorch con la versión de PyTorch y de CUDA necesarias donde se instalan todas las librerías requeridas.

Si bien es cierto, existen otras opciones para conseguir entornos de desarrollo funcionalmente similares: Conda, por ejemplo, también gestiona las dependencias de CUDA de las librerías de aprendizaje profundo, pero puede entrar en conflicto con las librerías instaladas usando pip (el

instalador de paquetes de Python) en su mismo entorno virtual, ya que no todas las librerías están disponibles en los repositorios de conda; otra opción que nos permite usar pip sin riesgo de dañar otras instalaciones en el equipo es el uso de entornos virtuales como venv, pero estos no gestionan correctamente el software y las dependencias de los paquetes relacionados con CUDA.

No obstante, Docker ofrece una ventaja más que es decisiva, la portabilidad que ofrece entre sistemas. En caso de querer ejecutar los scripts en cloud ([4.1](#)) o en dispositivos embebidos (p.e. los dispositivos Jetson de NVIDIA, que incluyen el NVIDIA Container Toolkit) sería suficiente con usar la misma imagen para tener un entorno idéntico.

Los ficheros necesarios para crear el entorno empleado en el proyecto están disponibles tanto en el repositorio del proyecto como en el Anexo TODO.

## Otros

Por último, para la redacción de este documento se ha empleado **LaTeX** como sistema de composición de texto, **diagrams.net** para el diseño de figuras e ilustraciones, y **BibTeX** para gestionar las referencias bibliográficas. Tanto esta memoria como el desarrollo del código relacionado con el proyecto se han llevado a cabo empleado **Git** como software de control de versiones y se pueden encontrar en los repositorios TODO y TODO respectivamente.

### 3.2. Hardware

Para el desarrollo de este proyecto, se ha utilizado principalmente el Equipo 1 de la [Tabla 2](#). No obstante, para el entrenamiento y la evaluación de los distintos modelos de aprendizaje profundo, se han empleado también instancias en la nube con la configuración del Equipo 2 en la [Tabla 2](#).

	Equipo 1 (Sobremesa)	Equipo 2 (Google Cloud)
Procesador	AMD Ryzen 7 3800x 8 núcleos @ 3.9 GHz	Intel Xeon 4 vCPU @ 2.30 GHz
GPU	NVIDIA RTX 3070 8 GB GDDR6 5888 CUDA cores 184 Tensor cores	NVIDIA Tesla T4 16 GB GDDR6 2560 CUDA cores 320 Tensor cores
Memoria	Arquitectura Ampere 32 GB DDR4	Arquitectura Turing 15 GB

Tabla 2: Especificaciones de los equipos empleados durante el trabajo de fin de máster.

### 3.3. Datasets

Durante este trabajo, de forma directa o indirecta, se emplean ciertos conjuntos de datos. Más concretamente, ImageNet [[84](#), [85](#)] y MIX6 [[4](#)] han sido empleados (no durante el desarrollo de este trabajo) para preentrenar los distintos modelos usados, mientras que KITTI [[86](#), [87](#), [88](#), [89](#)] se elige como conjunto de datos con el que comparar y evaluar las distintas modificaciones, implementadas y por lo tanto se utiliza para entrenar los modelos. A continuación se resumen las características de estos tres *datasets*.

#### 3.3.1. ImageNet

ImageNet [[84](#), [85](#)] es un *dataset* que proporciona un gran número de imágenes etiquetadas en función de la presencia o ausencia de una serie de conceptos definidos como *synsets*. Estos conceptos siguen la jerarquía propuesta por WordNet [[90](#)], donde se agrupan palabras y categorías

en función de sus relaciones semánticas. Para construir ImageNet, partiendo de una fracción de la ya mencionada estructura de WordNet, se buscaron imágenes de Internet para poblar cada una de las categorías. Estas imágenes, se filtraron y posteriormente fueron manualmente etiquetadas por humanos.

Dentro del proyecto de Imagenet, existen dos conjuntos: ImageNet21K e ImageNet1K (este último, normalmente llamado ImageNet). La principal diferencia entre estos dos conjuntos es que el primero, ImageNet21K suma más de 14 millones de imágenes clasificadas en más de 21 mil clases diferentes. Por otro lado, ImageNet1K es un subconjunto de ImageNet21K compuesto por cerca de 1.2 millones de imágenes clasificadas en 1000 categorías diferentes. Además de esto, también cuenta con anotaciones de localización de objetos (*bounding boxes*) en más de medio millón de imágenes. Debido a la gran cantidad de imágenes y la variedad de elementos que abarcan, ImageNet es normalmente empleado para entrenar las arquitecturas de aprendizaje automático profundo. De esta forma, los modelos preentrenados en ImageNet pueden ajustarse de una manera mucho más rápida y efectiva a tareas e imágenes nuevas con otros *datasets*, ya que al haber sido entrenados previamente los modelos han aprendido a extraer características generales (normalmente reutilizables) de las imágenes.

Una muestra de imágenes que conforman ImageNet está disponible en la [Figura 20](#).

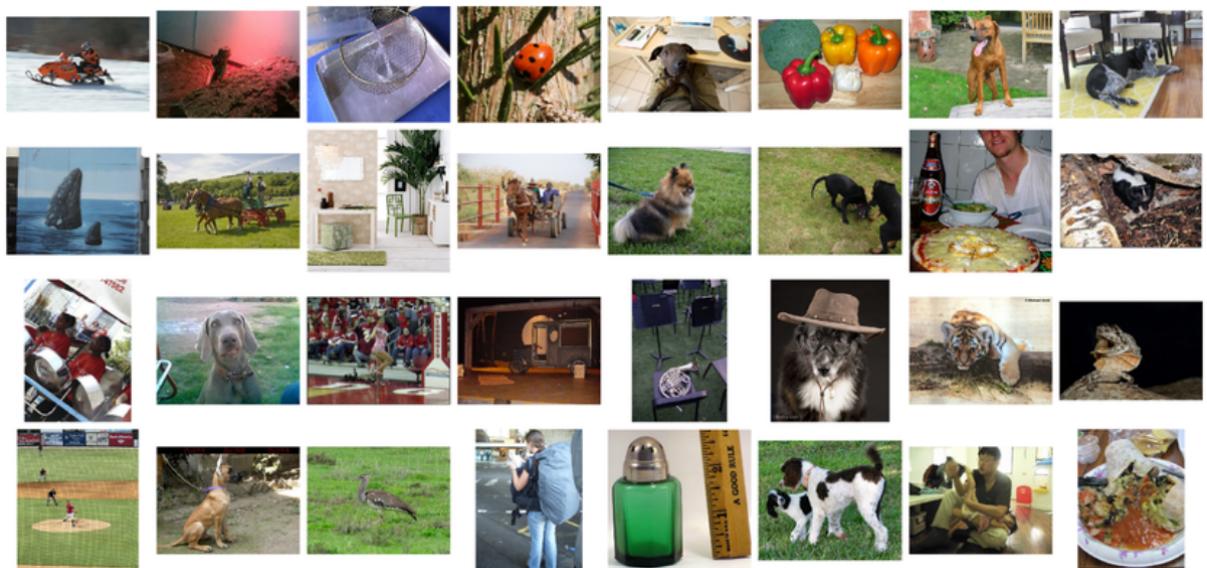


Figura 20: Muestra de las imágenes de ImageNet. Fuente: [\[85\]](#)

### 3.3.2. MIX6

MIX6 [\[4\]](#), una ampliación de MIX5 [\[66\]](#), es en realidad una agrupación de otros *datasets* que proporcionan anotaciones de profundidad de sus imágenes. Estas agrupaciones, consiguen dos características importantes: Primero, suman una cantidad de imágenes considerablemente alta; segundo, al tener datos de naturalezas tan distintas, existe una enorme variedad entre las imágenes, lo que permite entrenar modelos de estimación de profundidades generales, es decir, que no estén especializados en ningún entorno concreto.

Estas dos características, hacen que MIX6 sea una muy buena opción para entrenar arquitecturas basadas en *transformers* pero también dificultan el entrenamiento de modelos debido a la falta de homogeneidad entre los formatos de las imágenes, sus anotaciones, etc. Un desglose resumido de los *datasets* que componen MIX6 está disponible en la [Tabla 3](#).

Dataset	Descripción	Núm. de imágenes
Entrenamiento		
DIML Indoor [91]	Imágenes reales anotadas con cámara Kinect de Microsoft.	220K
MegaDepth [67]	Imágenes reales anotadas con MVS ( <i>Multi View Stereo</i> - Múltiples puntos de vista en diferentes fotografías)	130K
ReDWeb [92]	Imágenes reales anotadas a partir de estereovisión.	3.6K
WSVD [93]	Vídeos recuperados de YouTube en formato de estereovisión anotados a partir de dicha pareja de imágenes.	1.5M
3D Movies [66]	Películas 3D grabadas con cámaras estereoscópicas anotadas a partir de la pareja de imágenes.	75K
TartanAir [94]	Imágenes sintéticas.	1M
HRWSI [95]	Imágenes reales anotadas a partir de estereovisión.	21K
ApolloScape [96]	Imágenes reales anotadas con sensor LiDAR.	5.1K
BlendedMVS [97]	Imágenes sintéticas.	17K
IRS [98]	Imágenes sintéticas.	103K
Evaluación		
DIW [99]	Imágenes reales anotadas manualmente con la profundidad relativa entre pares de puntos aleatorios.	495K
ETH3D [100]	Imágenes reales anotadas con sensor LiDAR.	5.2K
Sintel [101]	Imágenes sintéticas.	1K
KITTI [86]	Imágenes reales anotadas con sensor LiDAR.	45K
NYUDepthV2 [102]	Imágenes reales anotadas con cámara Kinect de Microsoft.	407K
TUM [103]	Imágenes reales anotadas con cámara Kinect de Microsoft.	87K

Tabla 3: Datasets que conforman MIX6. Subrayados aquellos que no forman parte de MIX5.

### 3.3.3. KITTI

KITTI [86, 87, 88, 89] es un proyecto desarrollado por el *Karlsruhe Institute of Technology* y el *Toyota Technological Institute* que engloba un *dataset* y un conjunto de *benchmarks* enfocados a diferentes tareas relacionadas con la conducción autónoma. Los *benchmarks* que incluye este proyecto evalúan: estereovisión, flujo óptico (*optical flow*), flujo de la escena, **estimación de profundidades monocular**, *depth completion*, odometría visual/SLAM, localización de objetos (2D, 3D y cenital), seguimiento de objetos, segmentación de carreteras, y por último, segmentación de objetos general, tanto semántica como a nivel de instancia. Debido a la naturaleza de este trabajo, este apartado se centrará en la parte referente a la predicción de profundidad monocular.

Tipo de sensor	Modelo	Características
Cámara escala de grises (x2)	PointGray Flea2 grayscale	Sensor: FL2-14S3M 1.4 MP 1/2" Sony ICX267 CCD Global Shutter
Cámara color (x2)	PointGray Flea2 color	Sensor: FL2-14S3M 1.4 MP 1/2" Sony ICX267 CCD Global Shutter
Escáner láser (x1)	Velodyne HDL-64E	64 haces a 10Hz Resolución en distancia: 2cm Rango: 120m Posicionamiento GPS
Medida inercial (x1)	OXTS RT3003	Velocímetro Acelerómetro - Giroscopio

Tabla 4: Sensores equipados en el vehículo empleado para recoger los datos de KITTI.

Esta tabla no me convence mucho pero tampoco se me ocurre como resumirlo mejor

Esta tabla puede que sea un poco demasiado

Los datos disponibles en KITTI fueron capturados empleando un vehículo equipado con diferentes sensores ([Tabla 4](#)) para realizar diferentes recorridos en distintas zonas urbanas e interurbanas. De esta forma, se capturaron escenarios variados en múltiples condiciones de luz, hora, presencia de vehículos y peatones, etc. Dentro de los sensores equipados, son de especial interés para este trabajo las dos parejas de cámaras para estereovisión (un montaje con dos cámaras en escala de grises y otro montaje con dos cámaras en color) y el escáner láser rotatorio de 360°. Además de estos sensores, el automóvil también equipaba un sensor de medida inercial con sistema de navegación GPS para registrar información relacionada con la odometría que no se ha empleado durante el desarrollo del trabajo.

### 3.3.3.1. Datos

Si nos centramos en la información relevante para la estimación de profundidades monocular, KITTI ofrece los siguientes datos:

#### Datos en bruto

El *dataset* está compuesto por fotogramas muestreados y sincronizados a 10 Hz de los vídeos capturados por las cámaras en diferentes recorridos. Debido a las características del sistema óptico, para cada instante se disponen de cuatro imágenes, derecha e izquierda en escala de grises, y derecha e izquierda en color. Una muestra de estas imágenes puede observarse en la [Figura 21](#).

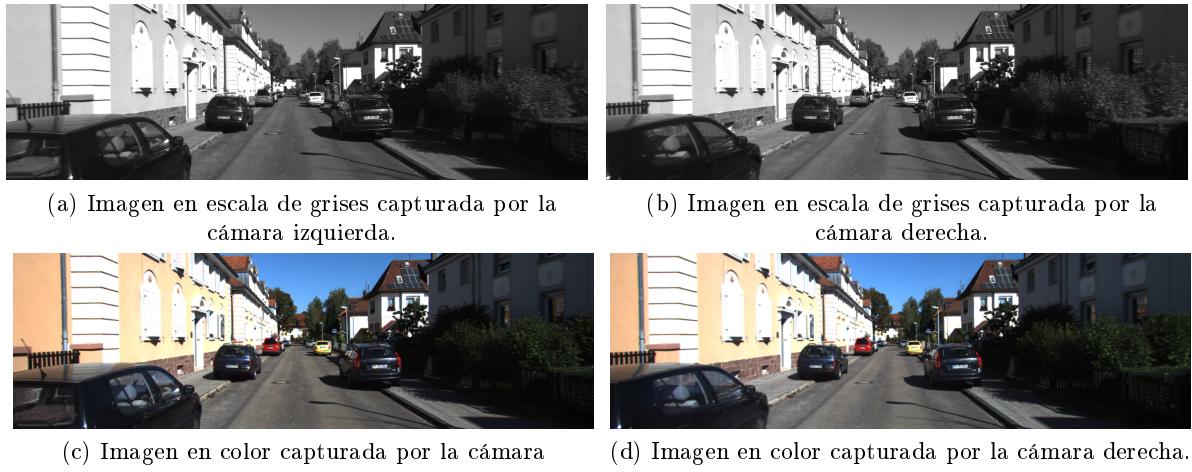


Figura 21: Muestra de las cuatro imágenes en bruto disponibles en KITTI para un instante dado.

En total, se disponen de 192760 imágenes ( $\sim 196$  GB) de tamaño 1242x375 píxeles, de las cuales 96430 (la mitad) corresponden a las cámaras a color. Como el objetivo es la estimación de profundidades monocular, solo se emplea una de las imágenes de cada pareja de imágenes producido por el sistema de estereovisión, por lo que realmente se emplean 48215 imágenes de los datos en bruto (una cuarta parte de la cantidad original).

#### Anotaciones

Por otro lado, KITTI proporciona también los valores numéricos de la profundidad para cada uno de los píxeles (de las imágenes presentadas previamente) almacenados como imágenes en formato PNG con un solo canal y 16 bits para cada valor (UINT16). Estos valores son los obtenidos por el escáner láser equipado en el vehículo y pueden considerarse una medida fiable de la profundidad

en cada imagen. Por lo tanto serán los datos que se emplearan como anotaciones para entrenar los modelos y evaluar sus capacidades de estimación de profundidades. Un punto importante a considerar sobre las medidas de estas anotaciones es que debido a la naturaleza del sensor con el que fueron tomadas, son anotaciones **dispersas** (*sparse*), no densas. Esto significa que no todos los píxeles de una imagen tienen anotación, y por lo tanto aquellos píxeles no anotados deberán ser ignorados tanto durante el entrenamiento como durante la evaluación. Una muestra de estas etiquetas y de las anotaciones dispersas puede observarse en la [Figura 22](#). Estas anotaciones están disponibles tanto como para las imágenes capturadas con las cámaras derechas como para las capturadas con las cámaras izquierdas.

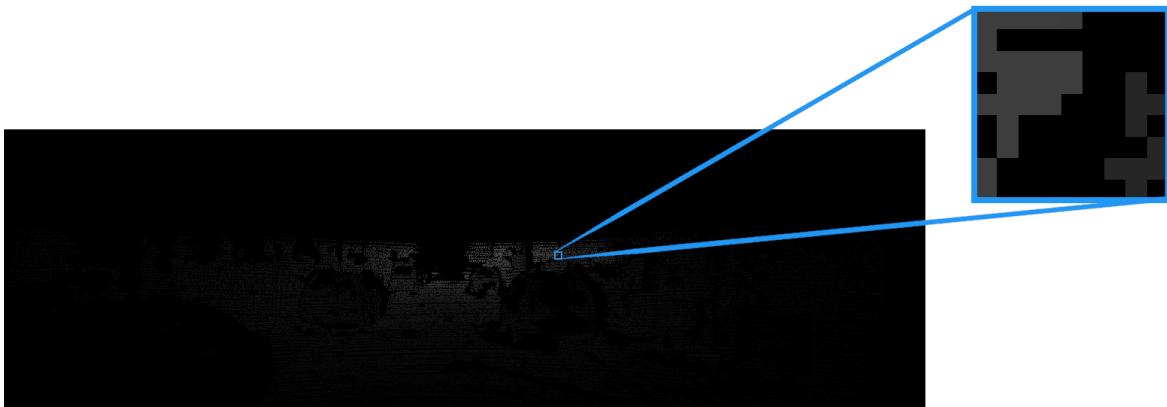


Figura 22: Anotación de KITTI y detalle de su carácter disperso para un instante dado.

### 3.3.3.2. Conjuntos de entrenamiento, validación y evaluación

Para el desarrollo de este trabajo es necesario disponer de un conjunto de entrenamiento con el que ajustar los parámetros de los modelos, un conjunto de validación con el que comprobar que no se está sobre ajustando el modelo al primer conjunto y para elegir la combinación de hiperparámetros óptima, y por último, un conjunto de evaluación (*test*) en el que calcular una serie de métricas que nos aportarán información del rendimiento real de los modelos finales.

El *dataset* de KITTI ya está dividido en entrenamiento y validación, e incluye una descarga adicional con el conjunto de test para el cual no están disponibles públicamente las anotaciones. No obstante, en las publicaciones sobre estimación de profundidad monocular [4, 50, 66, 104] es común encontrar los conjuntos de entrenamiento, validación y evaluación definidos por Eigen et al. [61] (conocidos como el *Eigen split*) que no respetan las particiones originales del *dataset* de KITTI.

Las listas con los archivos que pertenecen a cada uno de las particiones se han descargado desde el repositorio<sup>9</sup> del trabajo de Godard et al. [50]. En estas listas, hay nombres de archivos que no tienen ninguna anotación asociada, por lo que se eliminan de sus respectivas particiones. La distribución del número de imágenes, así como el número de imágenes eliminadas de cada conjunto se muestran en la [Tabla 5](#).

<sup>9</sup><https://github.com/nianticlabs/monodepth2/tree/master/splits> - (*Eigen full*)

Adjuntar  
el código  
usado  
para estas  
cosillas?

	Eigen split		
	Entrenamiento	Validación	Evaluación
Num. de archivos	45200	1776	697
Imágenes no encontradas	0 (0 %)	0 (0 %)	0 (0 %)
Anotaciones no encontradas	630 (1.39 %)	30 (1.69 %)	45 (6.46 %)
Num. imágenes útiles	44570	1746	652

Tabla 5: Distribución de las imágenes y número de imágenes no encontradas en el dataset.

Como comprobación adicional, se han cruzado las listas de archivos descargadas para asegurar que ninguno de los elementos de los conjuntos de entrenamiento y validación se encuentran en el conjunto de evaluación.

### 3.4. Definición de modelos a entrenar

Uno de los objetivos de este Trabajo Fin de Máster es explorar distintas modificaciones en la arquitectura DPT [4]. Para poder comparar de forma exhaustiva los efectos y la influencia en el rendimiento de cada una de estas modificaciones, se exploran todas las combinaciones posibles de los valores que se van a estudiar. Por lo tanto, no se detiene el entrenamiento de ningún modelo en caso de que su rendimiento sea peor que los modelos ya entrenados, ni se optimiza la búsqueda en función de la influencia de cada modificación en una métrica concreta.

Para definir tantos modelos como combinaciones sean posibles, se toman los conjuntos de valores elegidos para cada una de las modificaciones planteadas y se calcula su producto cartesiano (*grid search*). De esta forma, se obtienen las arquitecturas de los modelos que se entrenarán durante este trabajo y de los que se obtendrán los resultados finales.

### 3.5. Evaluación

Para la evaluación de los modelos presentados en este trabajo y sus modificaciones, se ha seguido la metodología propuesta en la publicación de Lee et al. [104] para evaluar los resultados en KITTI que consiste en recortar un pequeño marco alrededor de la imagen de salida y crear una máscara para los píxeles que no tienen una profundidad definida en la anotación. En la publicación de DPT [4], también se emplea el mismo procedimiento.

Al usar esta misma metodología se satisfacen dos objetivos: poder reproducir los resultados presentados en dicho artículo con el modelo original, y evaluar las modificaciones introducidas para comparar sus resultados con los del modelo sin modificar.

Como se mencionará más adelante, una de las modificaciones introducidas en este trabajo reduce el tamaño de la imagen en la entrada de las arquitecturas. El resultado de estos modelos, no obstante, se escala a su tamaño original antes de llevar a cabo la evaluación para asegurar que la magnitud de las métricas se ajuste a la del modelo original (con entradas de mayor resolución).

#### 3.5.1. Métricas

Una vez escaladas y enmascaradas las predicciones y las anotaciones, se calculan una serie de valores cuantitativos que permiten comparar y evaluar el rendimiento de los modelos. Las funciones que nos proporcionan estos valores son conocidas como métricas. Dentro del gran número de funciones que permiten evaluar los resultados de los modelos, se han elegido aquellas comúnmente empleadas en los modelos de aprendizaje profundo dedicados a la estimación de profundidad en imágenes monoculares [4, 43, 61, 66, 104, 105, 106, 107].

En las siguientes ecuaciones,  $d_p$  representa el valor del mapa de profundidad original (anotación) para cada pixel  $p$ , mientras que  $\hat{d}_p$  representa el valor de la profundidad estimada por el modelo para cada pixel  $p$ . Por otro lado,  $T$  denota el número de píxeles con información de profundidad disponibles en la anotación (al ser anotaciones dispersas, las imágenes no tienen información sobre la profundidad en todos los píxeles).

### 3.5.1.1. Accuracy under a threshold

La primera de estas métricas, el *accuracy under a threshold*, viene dada por la [Ecuación \(14\)](#) y cuantifica el porcentaje de píxeles a los que el modelo ha asignado una profundidad cuya relación de escala respecto de su valor real es menor que un determinado umbral. Los valores que se emplean para este umbral son 1,25,  $1,25^2$  y  $1,25^3$ .

$$\% \text{ de } p \in T : \max\left(\frac{\hat{d}_p}{d_p}, \frac{d_p}{\hat{d}_p}\right) = \delta < \text{umbral} \quad (14)$$

### 3.5.1.2. Mean Absolute Value of the Relative Error (Abs Rel)

Otra métrica usada habitualmente es el promedio del error relativo en todos los píxeles que disponen de valor de profundidad anotada. Para conseguir este error relativo, se calcula el error absoluto y se divide entre el valor real de la profundidad ([Ecuación \(15\)](#)).

$$\frac{1}{T} \sum_{p \in T} \frac{|d_p - \hat{d}_p|}{d_p} \quad (15)$$

### 3.5.1.3. Mean Squared Relative Error (Sq Rel)

Similar a la métrica anterior, en este caso el error absoluto se eleva al cuadrado antes de ser dividido entre el valor a estimar y de promediarlo con el resto de píxeles ([Ecuación \(16\)](#)). De esta forma, por la naturaleza cuadrática de la fórmula, se le da una mayor importancia a los errores mayores que a los menores.

$$\frac{1}{T} \sum_{p \in T} \frac{(d_p - \hat{d}_p)^2}{d_p} \quad (16)$$

### 3.5.1.4. Linear Root Mean Squared Error (RMSE)

El valor del error cuadrático medio proporciona una medida del promedio de la magnitud de la diferencia entre la profundidad predicha para cada uno de los píxeles y su profundidad real ([Ecuación \(17\)](#)). Una característica interesante de esta métrica es que sus unidades coinciden con las de la variable predicha, lo que facilita su interpretación. Como los errores se elevan al cuadrado antes de promediarse, estos tienen una importancia relativa directamente relacionada con su magnitud, es decir, cuanto mayor sea el error, más peso tendrá en el promedio. Es por esto por lo que es especialmente útil si se busca penalizar más los errores más grandes en las predicciones.

$$\sqrt{\frac{1}{T} \sum_{p \in T} (d_p - \hat{d}_p)^2} \quad (17)$$

### 3.5.1.5. Logarithmic Root Mean Squared Error (RMSElog)

Similar a la métrica anterior, en este caso el error cuadrático medio se calcula sobre los logaritmos naturales de las medidas a comparar ([Ecuación \(18\)](#)). Al realizar la resta de los logaritmos, la

operación es equivalente a calcular el logaritmo de la división del valor de profundidad estimado y el valor de profundidad anotado, restando de esta forma importancia a la escala del error y obteniendo una aproximación al error relativo de las medidas (frente al *RMSE*, que sería una medida del error absoluto). Además, debido al escalado que realizan los logaritmos, los *outliers* pierden importancia, por lo que es una métrica más robusta frente a este tipo de errores puntuales.

Otra característica a destacar de esta métrica es que está sesgada para penalizar aquellos casos en los que el valor predicho es menor que el valor real (subestimación). De esta forma, el error en dicha situación será mayor que si el valor predicho es mayor que el valor real (sobreestimación) aún cuando la diferencia entre ambos valores sea la misma.

$$\sqrt{\frac{1}{T} \sum_{p \in T} (\ln d_p - \ln \hat{d}_p)^2} \quad (18)$$

### 3.5.1.6. Scale Invariant Logarithmic Error (*SIlog*)

Esta métrica, es la raíz cuadrada de la función de pérdida propuesta por Eigen et al. [61] con  $\lambda = 1$  ([Ecuación \(19\)](#)). Al fijar el valor de  $\lambda$  en la unidad, se obtiene una medida totalmente independiente de la escala de la salida. De esta forma, se obtiene una medida de la calidad de los resultados de los modelos ignorando completamente la escala en la que se han producido las predicciones, que como ya se ha comentado es uno de los problemas fundamentales de la estimación de profundidades en imagen monocular.

Aquí no sé que hacer, he hecho la demostración matemática de que es invariante a la escala pero es un churro de ecuaciones, no sé si merece la pena poner un anexo solo para eso. La función de pérdida es esta misma así que más abajo cuando hablo de ella vuelvo a tener el mismo dilema

$$\sqrt{\frac{1}{T} \sum_{p \in T} (\ln d_p - \ln \hat{d}_p)^2 - \left( \frac{1}{T} \sum_{p \in T} \ln d_p - \ln \hat{d}_p \right)^2} * 100 \quad (19)$$

### 3.5.1.7. Mean Logarithmic Error (*Log10*)

Por último, se calculará también el promedio del error (en escala logarítmica) de las profundidades predichas respecto de las profundidades reales siguiendo la [Ecuación \(20\)](#).

$$\frac{1}{T} \sum_{p \in T} |\log_{10} d_p - \log_{10} \hat{d}_p| \quad (20)$$

### 3.5.1.8. Velocidad de procesamiento

Además de la calidad de los resultados, es de especial interés en este trabajo obtener medidas relacionadas con la velocidad de procesamiento que pueden alcanzar los modelos. Dentro de las medidas empleadas hay dos grupos: condicionadas por el *hardware* utilizado (Tiempo de inferencia y Tasa de transferencia efectiva) e independientes del *hardware* (Número de operaciones en coma flotante).

#### Tiempo de inferencia

Esta medida corresponderá al tiempo que tarda el modelo en procesar **una sola** imagen. Si suponemos que la aplicación de estos modelos es el procesamiento de vídeo de forma online, donde los fotogramas no pueden procesarse en lotes, esta medida es la inversa de los fotogramas

por segundo (*FPS*). Como se ha mencionado antes, esta métrica estará sujeta al *hardware* en el que se ejecute, y por lo tanto variará de un equipo a otro.

#### **Tasa de transferencia efectiva (*Throughput*)**

Por otro lado, en caso de que el procesamiento de imágenes se haga de forma offline y se disponga de todas las imágenes antes de comenzar el procesamiento, estas se podrían agrupar en lotes (*batches*) para paralelizar su inferencia. Al paralelizar el procesamiento de las entradas, aumenta el número de imágenes que se puede procesar por unidad de tiempo, que es lo que medirá esta métrica. Es decir, la tasa de transferencia efectiva es el número máximo de imágenes que puede procesar un modelo por unidad de tiempo. De nuevo, como se ha mencionado en el párrafo introductorio, este valor está ligado al equipo en el que se lleve a cabo la inferencia.

#### **Número de operaciones en coma flotante (*FLOPs*)**

Por último, esta vez independiente del *hardware* en el que se ejecuta el modelo, se empleará el número de operaciones en coma flotante (en inglés, FLOPs) que se requieren para procesar una sola entrada de la red como medida de la complejidad y coste computacional de los modelos.

## 4. Desarrollo y Modificaciones de la Arquitectura

En este capítulo, se presentan los desarrollos principales que se han llevado a cabo durante el trabajo, prestando especial atención a las modificaciones realizadas sobre la arquitectura DPT.

### 4.1. Cloud

Con el objetivo de reducir el tiempo necesario para entrenar los distintos modelos que se plantean en este trabajo, se ha recurrido al servicio de infraestructura (IaaS - *Infrastructure as a Service*) que ofrece la empresa Google: Google Cloud. Los servicios en la nube (*cloud*), permiten disponer de recursos informáticos de manera flexible, pagando únicamente por aquellos que estén activos. Los proveedores de IaaS, se encargan del mantenimiento y gestión de la infraestructura (redes, almacenamiento, servidores, virtualización), mientras que el usuario se encarga de la gestión del sistema operativo y todo lo que hay por encima.

Dentro de Google Cloud, se han empleado los servicios **Compute Engine**, para disponer de máquinas virtuales y **Cloud Storage**, para crear recursos de almacenamiento (*buckets*).

El flujo de trabajo que se ha seguido ha sido el siguiente ([Figura 23](#)):

1. Primero, se ha creado un *bucket* en el que se han dejado disponibles el conjunto de datos empleado durante el entrenamiento, el código necesario para ejecutar el entrenamiento, y una serie de scripts para facilitar la configuración del equipo. Al disponer de estos archivos en la nube, se desacopla totalmente la configuración de las máquinas virtuales y el ordenador local en el que se lleva a cabo el desarrollo (Equipo 1 en [Tabla 2](#)).
2. A continuación, se configura una máquina virtual con el *hardware* elegido<sup>10</sup> (Equipo 2 en [Tabla 2](#)). Una vez conectados a esta máquina virtual a través de SSH, se descarga del *bucket* creado el script de configuración (disponible en el Anexo TODO) y se ejecuta. Este script, se encarga de: descargar el resto de archivos disponibles en el *bucket*, instalar los drivers de NVIDIA necesarios para poder usar la GPU de la instancia, instalar Docker y el NVIDIA Container Toolkit, instalar Weights and Biases, y construir la imagen de Docker especificada en el Dockerfile descargado.
3. Una vez configurada la instancia con todos los archivos necesarios en su disco SSD, se crea una imagen de dicha instancia en Google Cloud de forma que sea fácilmente replicable. Posteriormente, se configura desde la consola de Google Cloud el inicio de estas instancias de forma que cada vez que se encienda una (nueva o ya existente), se cree dentro del equipo un contenedor de Docker a partir de la imagen ya construida, y se ejecute en este contenedor el cliente de Weight and Biases para entrenar modelos automáticamente. De esta forma, para añadir una nueva máquina al proceso de entrenamiento de experimentos, solamente hay que crear un nuevo equipo a partir de la imagen preconfigurada.
4. Por último, una vez finalizados los experimentos, se copian a través de SSH los parámetros de los modelos entrenados guardados en cada una de las máquinas virtuales empleadas.

<sup>10</sup>Para elegir el *hardware* de las máquinas virtuales, se ha elegido la GPU que mayor relación TFLOPS/euro ofrecía para minimizar el coste de los equipos. El resto de características se han elegido de forma que la limitación del equipo sea el procesamiento en GPU.

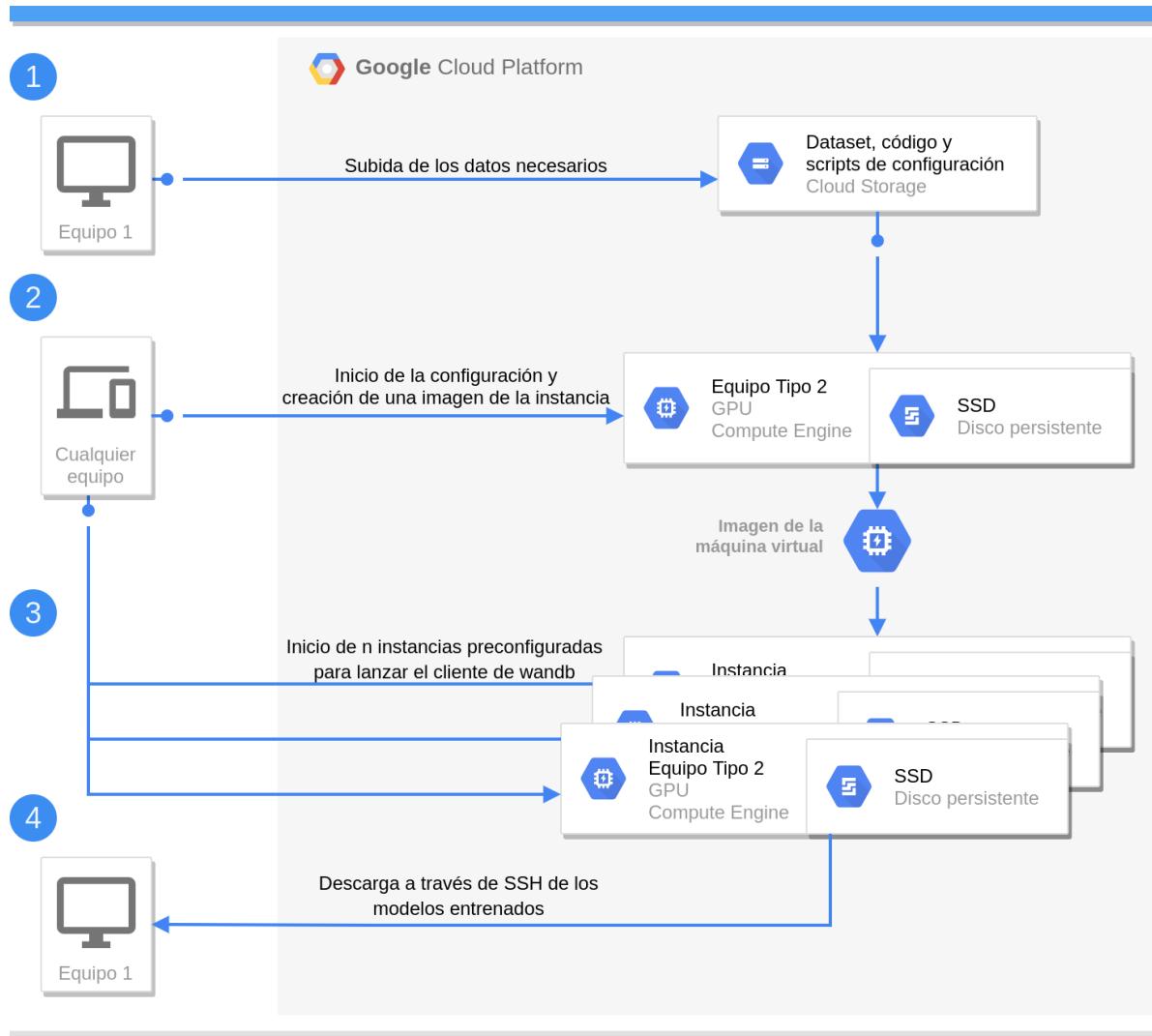


Figura 23: Esquema de la configuración llevada a cabo en la nube.

#### 4.2. Warmstart

Para acelerar lo máximo posible la convergencia del modelo durante el entrenamiento, se aprovechan en la medida de lo posible parámetros con sesgos inductivos ya aprendidos, es decir, parámetros de modelos ya entrenados. Más concretamente, los parámetros de los modelos a entrenar se inicializan con los valores de los parámetros del modelo DPT-Hybrid entrenado en MIX6, publicados en el artículo original de *Dense Prediction Transformers* [4].

Además, ya que en este trabajo se estudian diferentes modificaciones de dicho modelo, se emplea el método `load_state_dict()` de la clase `torch.nn.Module` con el parámetro `strict=False` para cargar los parámetros, evitando así que la carga falle si no coinciden una a una las capas en el modelo y las definidas en el archivo que se trata de cargar. En cuanto a las capas/conjuntos de capas que se han modificado y/o añadido, al no disponer del conjunto de datos MIX6 para preentrenarlas, se han inicializado con sus parámetros correspondientes tras ser entrenadas en ImageNet21K.

Comprobar esto entrenando el modelo de ResNet con los pesos de imagenet.

Esta forma de proceder, sin embargo, origina un problema que se comentará más adelante en

detalle, y es que como era de esperar, los modelos con menor porcentaje de su arquitectura modificada parten de una situación inicial ventajosa para la tarea de estimación de profundidad monocular (al estar preentrenados en MIX6).

### 4.3. Entrenamiento

Pese a que los autores del artículo de DPT [4] han publicado el código del modelo y sus pesos preentrenados en MIX6 y KITTI, a día de hoy no han hecho públicos los *scripts* de entrenamiento empleados. Por esto, ha sido necesario escribir el proceso de entrenamiento así como el **Dataset** de PyTorch con el que leer los datos de KITTI.

Para el **Dataset**, se crea una clase `KITTIDataset`<sup>11</sup> que hereda de la clase abstracta base para conjuntos de datos que ofrece PyTorch, `torch.utils.data.Dataset`, y sobreescribe los métodos `__len__()` y `__getitem__()` de forma que estos se adapten a la estructura de directorios y nombres de las imágenes y de sus anotaciones. Al sobreescribir estos métodos, es posible crear un `torch.utils.data.DataLoader` de forma directa, pasando las imágenes y las etiquetas a los modelos aprovechando las herramientas de PyTorch. En el método `__getitem__()`, además, se aplican las transformaciones necesarias a los datos, así como el *Data Augmentation*.

Por *Data Augmentation* se entiende el conjunto de operaciones y transformaciones que se pueden aplicar a los datos para modificar su apariencia. Esto normalmente favorece el aprendizaje y la capacidad de generalización de la red (tiene efecto regularizador y evita el sobreajuste al aumentar el número de ejemplos y la variedad entre ellos). En el entrenamiento llevado a cabo, siguiendo una vez más la metodología de DPT, se ha incluido un **reflejado horizontal aleatorio**, es decir, cada una de las imágenes (junto con las anotaciones) tiene un 50 % de posibilidades de ser reflejada horizontalmente antes de atravesar la red como ejemplo de entrenamiento.

En cuanto al *script* de entrenamiento<sup>12</sup>, se tienen en cuenta una serie de factores para acelerar el proceso lo máximo posible:

- **Número de trabajadores en el `Dataloader`:** Con el objetivo de asegurar que la limitación en la velocidad de entrenamiento sea el procesamiento en la GPU, se cambia el valor del parámetro `num_workers` del constructor del `Dataloader` de entrenamiento a 8. Este parámetro, controla el número de procesos que se lanzan en paralelo para leer los datos del disco y preprocesarlos.
- **pin\_memory:** También en el constructor del `Dataloader`, es posible activar `pin_memory`, parámetro desactivado por defecto. Este parámetro, acelera la transferencia a la memoria de la GPU de los datos cargados en memoria (RAM) por la CPU [108]. De forma resumida, lo que habilita este parámetro es que la carga de datos se haga en memoria no paginable (*pinned*) a la que la GPU accede directamente, evitando así cargar los datos en una zona de memoria paginable y transferir estos a una *pinned memory* temporal cada vez que la GPU quiere leerlos para transferirlos a su propia memoria.
- **`torch.backends.cudnn.enabled` y `torch.backends.cudnn.benchmark`:** Estas dos opciones, se activan para asegurar, respectivamente, que se use CuDNN en la ejecución del modelo y que se ejecuten al comienzo del *script* distintas implementaciones de algoritmos de convolución para emplear el más rápido en el sistema actual.
- **Mixed precision:** Ya mencionado en la Sección 2.9, pese a que finalmente se ha descartado su uso en el entrenamiento debido a la inestabilidad numérica que introducía y los fallos que ocasionaba, el *script* de entrenamiento incluye la opción de activar el uso de precisión mixta con el escalado pertinente.

<sup>11</sup>URL al repo?

<sup>12</sup>URL de `train.py` y `train_utils.py`?

Para ajustar los parámetros de los modelos, se ha empleando como optimizador AdamW con una tasa de aprendizaje  $lr = 1e-5$  y parámetros:  $\beta_1 = 0,9$ ,  $\beta_2 = 0,999$ ,  $\epsilon = 1e-8$  y *weight decay* = 0,01. El número de épocas se ha fijado en 20 tras analizar el comportamiento de la pérdida en distintas ejecuciones previas. En cuanto al tamaño de lote usado, se ha utilizado solamente una imagen para cada actualización de parámetros. Esta elección viene motivada por dos razones, la primera, la falta de recursos computacionales para entrenar algunas de las variaciones de los modelos con lotes de más imágenes, y la segunda y más importante, que tras implementar en el *script* de entrenamiento la posibilidad de acumular gradientes<sup>13</sup>, se comprobó que se obtenían mejores resultados con lotes de una sola imagen, probablemente por el efecto regularizador característico de un tamaño de lote tan reducido.

Además, se ha empleado la función `torch.nn.utils.clip_grad_value_`, que se encarga de limitar los valores de los gradientes en función de su valor con un `clip_value` de 0,5, limitando considerablemente las posibilidades de que explotasen los gradientes del modelo y se desestabilice su entrenamiento. No obstante, como precaución y para evitar el desperdicio de recursos, el entrenamiento se detiene en caso de que el valor de la función de pérdida se vuelva  $\pm\infty$ .

#### 4.3.1. Función de pérdida

La función de pérdida usada durante el proyecto, y por lo tanto implementada en el script de entrenamiento, es la función empleada en la publicación de DPT [4] para ajustar los modelos preentrenados en MIX6 a *datasets* más pequeños. Tal y como indica esta publicación, su función de pérdida se compone de la función de pérdida propuesta por Eigen et al. [61] y de otra función que calcula los gradientes de la profundidad obtenida para penalizar la falta de suavidad en píxeles contiguos, propuesta en el trabajo de Li et al. [67]. No obstante, como el *dataset* empleado en este trabajo es KITTI y sus anotaciones son dispersas, no es posible calcular dichos gradientes en las etiquetas y por lo tanto se elimina esa parte de la función.

De esta forma, la función de pérdida resultante es la definida por la Ecuación (21), donde  $\hat{d}_p$  es la profundidad obtenida de la red para cada píxel  $p$ ,  $d_p$  es la etiqueta con los valores de profundidad,  $n$  es el número de píxeles que tienen un dato de profundidad en la etiqueta correspondiente y  $\lambda$  es un hiperparámetro cuyo valor puede estar en el rango  $[0, 1]$  y controla la influencia de la escala, ya que con  $\lambda = 0$  el segundo término se anula (quedando la distancia de cuadrados en espacio logarítmico), y con  $\lambda = 1$  la función de pérdida es invariante a la escala. La demostración de la invariancia a la escala es equivalente a la de la métrica *SILog* (Sección TODO), ya que dicha métrica es la raíz cuadrada de esta función de pérdida. Para el entrenamiento de los modelos,  $\lambda$  se ha fijado en 0,5.

$$L(\hat{d}, d) = \frac{1}{n} \sum_p (\ln \hat{d}_p - \ln d_p)^2 - \frac{\lambda}{n^2} \left( \sum_p (\ln \hat{d}_p - \ln d_p) \right)^2 \quad (21)$$

#### 4.4. Reducción de tamaño de la entrada

Para acelerar DPT lo primero que se modificó fue el tamaño de la entrada, buscando estudiar la flexibilidad de la arquitectura para aprender a inferir profundidad a partir de imágenes con menor resolución. Los resultados de la publicación original calculan la profundidad en el conjunto de evaluación de KITTI con las imágenes en su tamaño original, 1216x352. Para reducir el consumo de memoria del modelo durante su entrenamiento, así como acelerar entrenamiento e inferencia,

---

<sup>13</sup>La acumulación de gradientes (*gradient accumulation*) es una técnica que promedia las pérdidas en distintos lotes para, aún sin poder aprovechar la mejora de rendimiento de un tamaño de lote mayor, conseguir que la actualización de los parámetros sea matemáticamente equivalente (con alguna limitación como la imposibilidad de usar *batch normalization*) a usar un tamaño de lote mayor.

se han añadido dos operaciones de cambio de tamaño: una al principio de la red que reduce el tamaño de las imágenes a 640x192 píxeles, y otra al final de la red que reescala la salida al tamaño original.

#### 4.5. Capas de atención eficiente

Tal y como se ha visto en la [Sección 2.6.3](#) de Marco Teórico y Estado del Arte, existen numerosas técnicas para reducir la complejidad computacional y requisitos de memoria de los mecanismos de atención característicos de los *transformers*. Sin embargo, estas técnicas ofrecen mayores incrementos en el rendimiento cuanto mayor es la longitud de la cadena de la entrada. Por lo tanto (especialmente tras reducir el tamaño de las imágenes) no se va a apreciar un incremento especialmente substancial en la velocidad de inferencia por modificar estos mecanismos. Aún así, se decide modificar el mecanismo de atención con el objetivo de estudiar la capacidad de estas nuevas técnicas para igualar los resultados de los mecanismos originales. En concreto, se compara el mecanismo de atención del *Perfomer* con la atención estándar.

Para poder llevar a cabo los experimentos relacionados con este aspecto de la arquitectura y poder entrenar los modelos necesarios, se modifica el código de DPT de forma que se puedan seleccionar distintos mecanismos de atención, entre ellos la implementación del mecanismo de atención del *Performer* desarrollada por Phil Wang [\[83\]](#).

#### 4.6. Número de cabezas

En la publicación *Are Sixteen Heads Really Better than One?* [\[109\]](#), se plantea y prueba la idea de que el número de cabezas en los bloques de atención está sobredimensionado, siendo posible eliminar un gran porcentaje de las cabezas sin afectar al rendimiento de forma significativa. Esta influencia en el rendimiento, se reduce aún más cuando se reentrena el modelo después de reducir el número de cabezas. Además, los autores señalan que reducir el número de cabezas degrada el rendimiento de forma más significativa en los bloques de *Cross-Attention* ([Apartado 2.6.2](#)) con elementos del *encoder* y *decoder* que en los bloques de *Self-Attention*.

En el caso de DPT, al emplear como *encoder* un ViT, no existen bloques de *Cross-Attention*, si no que son todos de *Self-Attention* (recordemos que el ViT solo se compone de la parte del *encoder* del *transformer* original). Por lo tanto, se plantea también como una de las modificaciones del proyecto estudiar la influencia del número de cabezas en la velocidad de inferencia y los resultados, adaptando el código del modelo para que sea posible cambiar la cantidad de cabezas en los bloques de atención (tanto de atención estándar como de atención eficiente). Los valores con los que se experimenta para este hiperparámetro son 1, 12 (el número de cabezas que tiene DPT por defecto) y 24.

#### 4.7. Hooks del transformer y bloques de atención posteriores

En la arquitectura de DPT ([Sección 2.8.5](#)), la transferencia de información desde el *encoder* al *decoder* tiene lugar a través de cuatro ganchos o *hooks* que toman las salidas de ciertas capas para que pasen a ser las entradas que se fusionan en el *decoder*. En la publicación original, si bien es cierto que se valoran diferentes tamaños de *Vision Transformers*, no se presenta ningún experimento modificando los bloques de atención de los cuales se cogen las salidas. Más concretamente, el modelo DPT-Hybrid coloca dos de estos *hooks* en el *backbone* convolucional y los otros dos en los bloques de atención número 8 y número 11 (empezando en cero). Por lo tanto, la imagen de entrada tiene que atravesar todos los bloques de atención del ViT-Hybrid para conseguir la salida del último bloque de atención.

Los valores que se han probado en este trabajo para estos *hooks* del ViT, es decir, los bloques de atención cuya salida se ha seleccionado como entrada del *decoder*, son: bloques número 0 y

1, bloques número 2 y 5, y por último, bloques número 8 y 11 (siendo estos últimos los de la publicación original).

Al modificar las capas del transformer de las que se cogen las activations para pasarlas a la etapa de fusión convolucional, se abre la posibilidad de eliminar aquellos bloques de atención que ya no se usan. Por lo tanto, se ha modificado el código de la arquitectura para incluir dos cambios en el momento de instanciar el modelo: el primero, capar la inferencia para que la ejecución del ViT se detenga cuando llega al bloque de atención más alto requerido por los *hooks*, y el segundo, eliminar las capas superiores no empleadas. De esta forma, además de acelerar el entrenamiento e inferencia del modelo, se reduce su tamaño considerablemente, tanto a la hora de almacenar sus parámetros como a la hora de cargarlo en memoria para desplegarlo en una aplicación real.

En la [Figura 24](#), se puede apreciar la modificación del ViT empleado en DPT: a la izquierda está el *encoder* tal y como se encuentra en el trabajo de DPT, con los *hooks* en los bloques 8 y 11, sin eliminar ninguno de los parámetros del *transformer*, y con la inferencia recorriendo los 12 bloques de atención; a la derecha, por otro lado, está una de las opciones valoradas para este hiperparámetro de la arquitectura, donde los *hooks* se sitúan en la salida de los bloques 0 y 1 del ViT. De esta forma, en el ejemplo, los parámetros a partir del segundo bloque de atención se eliminarían ya que la inferencia del modelo solo llega hasta este segundo bloque.

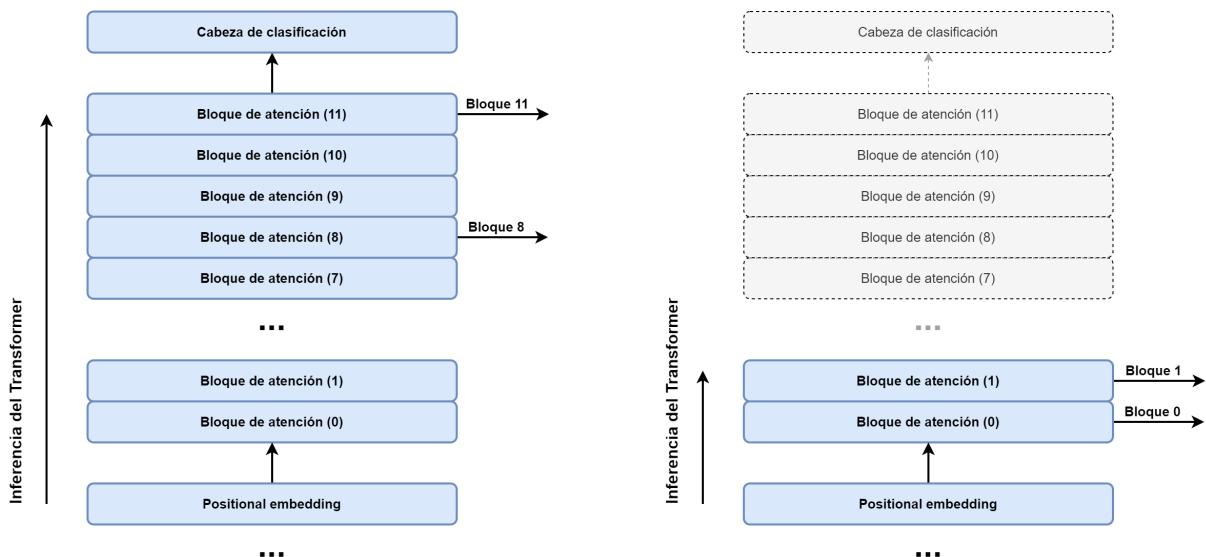


Figura 24: Cambio en los *hooks* y en el número de bloques de atención.

#### 4.8. Backbone convolucional

La última modificación de la arquitectura de DPT estudiada en este proyecto es la del *backbone* convolucional del ViT-Hybrid, encargado de extraer los mapas de características empleados para obtener los *tokens* que pasan a los bloques de atención.

En el modulo propuesto en la publicación original, se elige como *backbone* una ResNet50. De esta red, se extraen las activations en los bloques 0 y 1 para pasarlas al *decoder*. Por otro lado, la salida de la última capa de la ResNet50, es decir, la entrada de los bloques de atención, tiene forma  $[n, c, h, w]$ , donde: n es el número de imágenes en el lote (*batch size*), c es el número de canales (en este caso mapas de características), y h y w la altura y anchura de los mapas de características. En el caso de este *backbone*, la salida está compuesta por  $c = 1024$  mapas de características de dimensiones iguales a las de la imagen de entrada divididas 4 veces entre 2.

Después de la ResNet, hay una capa de proyección que no es más que una capa convolucional con *kernels* de tamaño  $1 \times 1$  y zancada  $1 \times 1$ , con 1024 canales de entrada y 768 canales de salida.

Este tipo de convoluciones  $1 \times 1$ , realmente se componen de 786 kernels de  $1 \times 1 \times 1024$ , por lo que al convolucionar cada uno de ellos la entrada, se obtienen 768 mapas de características del mismo tamaño que los de la entrada. Después, los 768 mapas de características resultantes se aplanan en tensores de forma  $[n, 768, \frac{h}{8} \times \frac{w}{8}]$  y se transponen de forma que la entrada sea de tipo  $[n, t, 768]$ , donde 768 es el tamaño de los *tokens* de los bloques de atención y t el número de *tokens* extraídos de la imagen. De esta forma, a mayor tamaño de imagen mayor será el número de *tokens* (secuencia más larga, aumenta el coste de los bloques de atención), pero la dimensión de estos permanece constante.

A modo de ejemplo, supongamos una entrada de una sola imagen de tamaño  $384 \times 384$ : la salida de la ResNet50 tendrá la forma  $[1, 1024, 24, 24]$ . Esta salida atraviesa la convolución  $1 \times 1$  de proyección y pasa a tener forma  $[1, 768, 24, 24]$ . Estos mapas de características se aplanan en un tensor  $[1, 768, 576]$  que se traspone para obtener la forma  $[1, 576, 768]$ , que equivale a 576 *tokens* de dimensión 768. Una vez llegados a este punto, el tensor está listo para pasar a los bloques de atención del *transformer*.

Buscando una alternativa que ofrezca resultados similares con un menor coste computacional, se adapta e incluye en las pruebas la arquitectura EfficientNet-B0 como *backbone* convolucional de DPT. Esta arquitectura, proporciona una salida de forma  $[n, c, h, w]$  (manteniendo la nomenclatura anterior) donde el número de mapas de características es ahora 1280 y la altura y anchura se corresponden con las de las imágenes de entrada divididas 5 veces entre 2. Es decir:  $[n, 1280, \frac{h}{16}, \frac{w}{16}]$ . Para sustituir el *backbone* convolucional del ViT-Hybrid sin tener que cambiar el tamaño de todas las capas de atención (y así aprovechar los pesos preentrenados), se sustituye la capa de proyección mencionada en el párrafo superior, que recordemos era una capa convolucional con kernels de tamaño  $1 \times 1$  por una capa de convolución transpuesta. Esta capa de convolución transpuesta, cumple dos funciones fundamentales: la primera, proyectar los 1280 mapas de características en 768 gracias al número de *kernels* empleados; y la segunda, al tener *kernels* de tamaño  $2 \times 2$  y una zancada también de  $2 \times 2$ , conseguir que las dimensiones de los mapas de características se multipliquen exactamente por 2 en ambas dimensiones, convirtiendo la entrada  $[n, 1280, \frac{h}{16}, \frac{w}{16}]$  en  $[n, 768, \frac{h}{8}, \frac{w}{8}]$ . Estas dimensiones, son las que espera la etapa de *embedding* que aplana los mapas de características y transpone el tensor para obtener la entrada de los bloques de atención, quedando a su salida un tensor con forma  $[n, t, 768]$  donde t vuelve a ser el número de *tokens* que pasan al *transformer*, cada uno de dimensión 768.

Dado que no se disponen de los parámetros ajustados en MIX6 para EfficientNet-B0, se ha modificado también el *script* del modelo para cargar en su inicialización los parámetros ajustados en ImageNet21K (proporcionados por la librería PyTorch Image Models [82], de donde también se ha utilizado la implementación de EfficientNet-B0) correspondientes a este nuevo módulo.

## 5. Resultados

En este capítulo se exponen los resultados obtenidos tras entrenar las arquitecturas con las modificaciones planteadas. Para poder llevar a cabo un estudio lo más completo posible, se han tomado una serie de valores candidatos para cada modificación (TODO contarles) y se ha llevado a cabo el producto cartesiano de dichos conjuntos, obteniendo así  $3 \times 2 \times 3 \times 2 = 36$  modelos distintos.

Meter comparación de los tiempos de inferencia con autocast y sin autocast ->Llevar la gráfica al apartado teórico?

### 5.1. Resultados cuantitativos

#### 5.1.1. Reducción de tamaño de la entrada

Dado que el entrenamiento de los modelos - por limitaciones materiales y temporales - se ha llevado a cabo con la entrada del modelo reducida, no es posible realizar una comparación exhaustiva de los resultados obtenidos para cada una de las modificaciones de la arquitectura respecto del modelo correspondiente entrenado con las imágenes en su tamaño original.

Hacer una prueba si cabe en la GPU entrenando solo un modelo

En cuanto a la medida de la velocidad de inferencia, que si que es posible obtener independientemente de que los pesos correspondan al modelo entrenado o no, en la Figura 25 se puede apreciar la diferencia en función del tamaño de entrada de los FPS medios de todas las modificaciones realizadas sobre el modelo DPT. El valor de la izquierda, corresponde al tamaño con el que se evalúa KITTI en la publicación original [?], mientras que el de la derecha es la reducción de tamaño establecida en este trabajo. Dado que este cambio es significativo y es, en términos de magnitud, el que mayor aceleración media conlleva ( $\times 2.47$ ), en las siguientes Figuras que representan los FPS en función de los otros cambios mencionados en la sección TODO se presentará la influencia de dichos cambios en la inferencia tanto con la entrada original como con la entrada reducida para poder así valorar también la mejora de rendimiento que conllevaría la modificación si no se hubiese modificado el tamaño de la entrada.

#### 5.1.2. Número de cabezas

Completar el párrafo de abajo cuando estén los resultados de las otras métricas, a lo mejor algo de la explicación de los resultados llevarlo a un apartado de discusión?

Respecto de la influencia de este parámetro en la velocidad de inferencia, en la Figura 26 se puede observar, agrupado por tamaño de entrada, backbone, y número de cabezas, los promedios de los FPS de los modelos. Teniendo en cuenta que las capas de atención del modelo original tienen 12

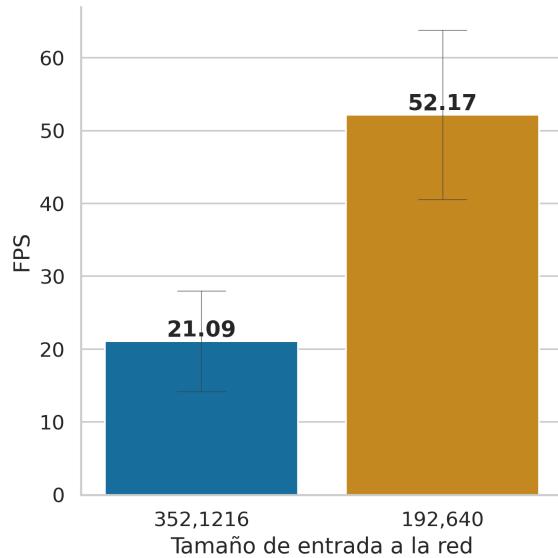


Figura 25: FPS promedios de los modelos en función del tamaño de su entrada. Las barras grises representan la desviación estándar de las medidas.

cabezas, era de esperar que aquellos modelos con una sola cabeza fueran más rápidos y los que cuentan con 24 fuesen más lentos. Además, este cambio en la velocidad de inferencia relativa es mayor en el caso de la entrada sin escalar que en el caso de la entrada escalada. Por ejemplo, en el paso de 12 a una cabeza con la entrada sin escalar, el incremento en el número de FPS es de  $\times 1,2$  y  $\times 1,26$  dependiendo del backbone empleado, mientras que al escalar la salida este cambio pasa a ser  $\times 1,04$  y  $\times 1,03$  respectivamente. Esto se debe a la implementación previamente comentada en la sección TODO, ya que TODO reescribir esto

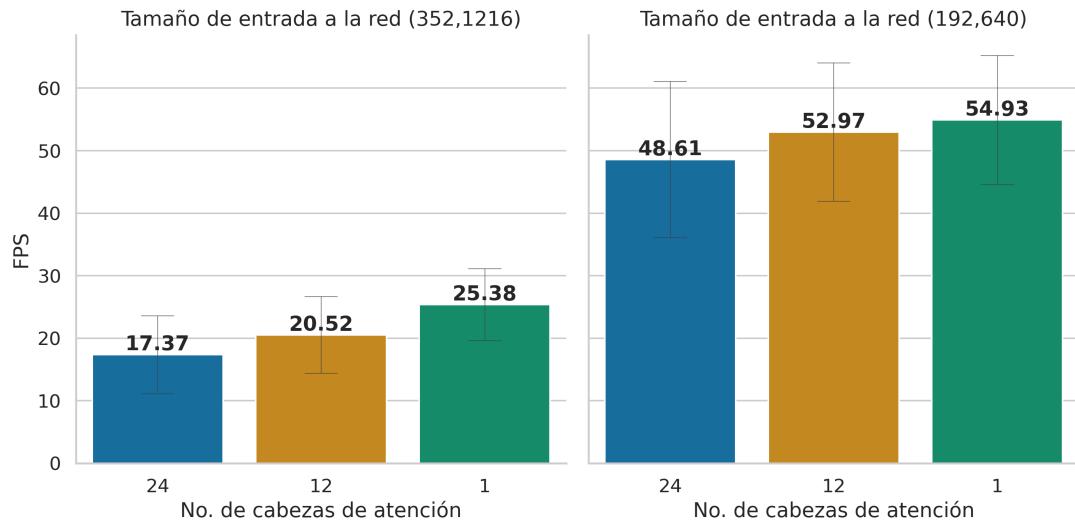


Figura 26: Resultados de.

### 5.1.3. Capas de atención eficiente

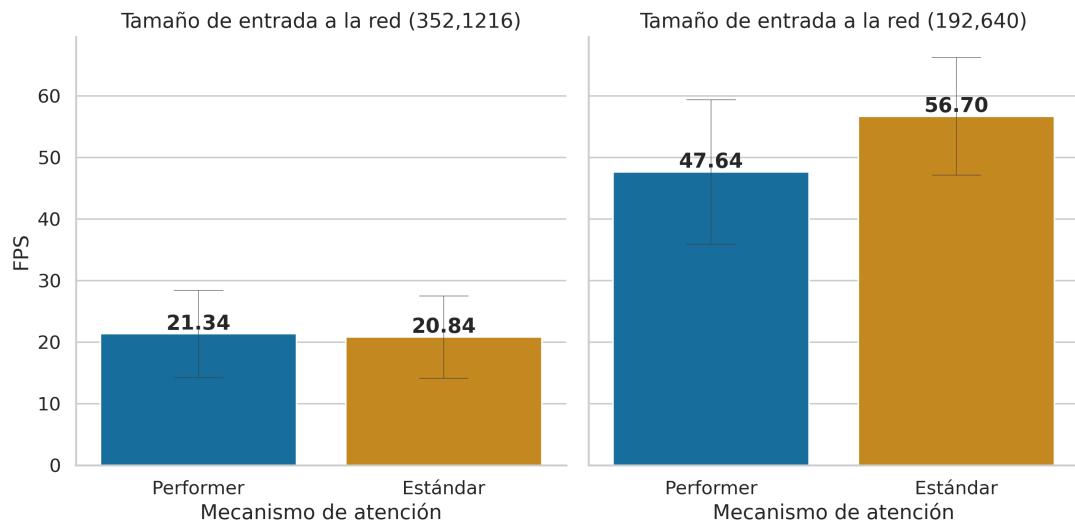


Figura 27: Resultados de.

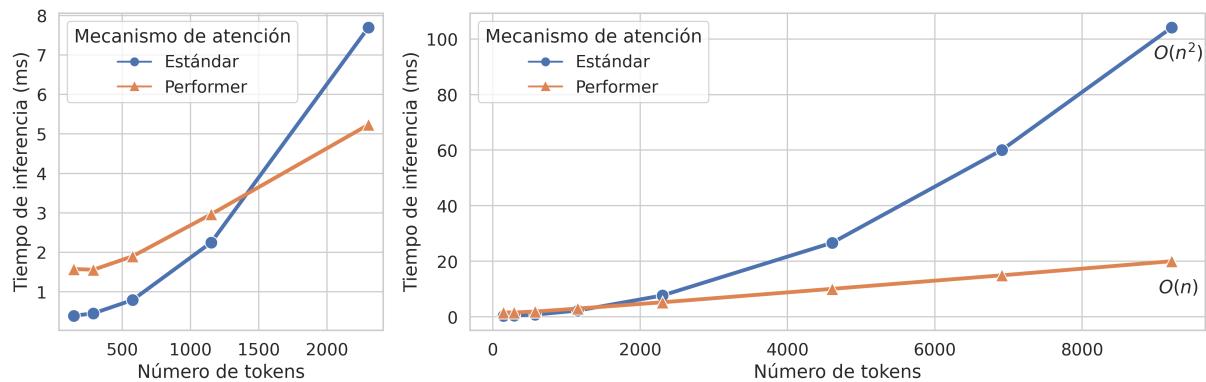


Figura 28: Resultados de.

#### 5.1.4. Cambio en los hooks del transformer y eliminación de las capas de atención posteriores

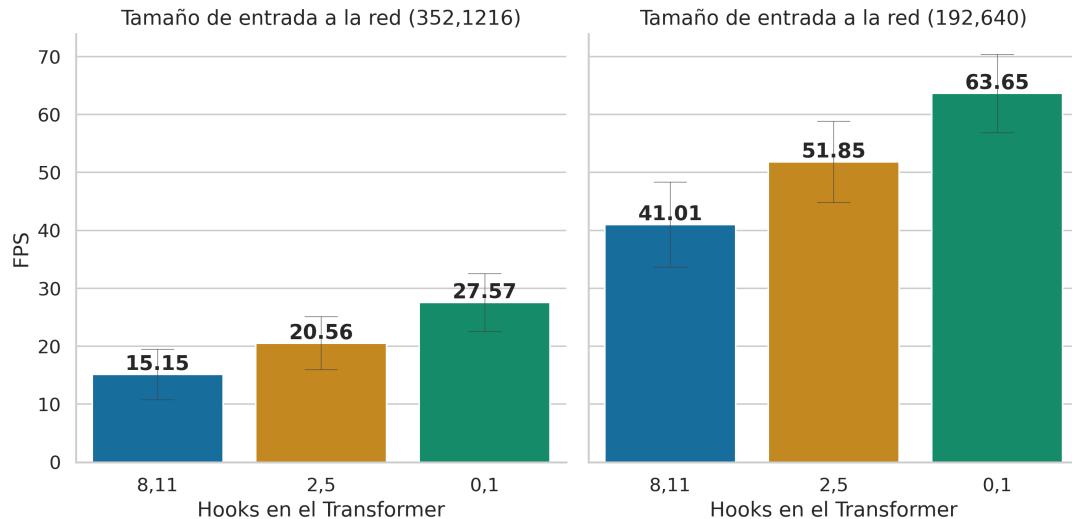


Figura 29: Resultados de.

#### 5.1.5. Cambio del backbone convolucional

Comparar los feature maps que saca uno y otro?

Justificar que al no estar preentrenado en mix6 los resultados son peores con un entrenamiento de un resnet a partir de imagenet

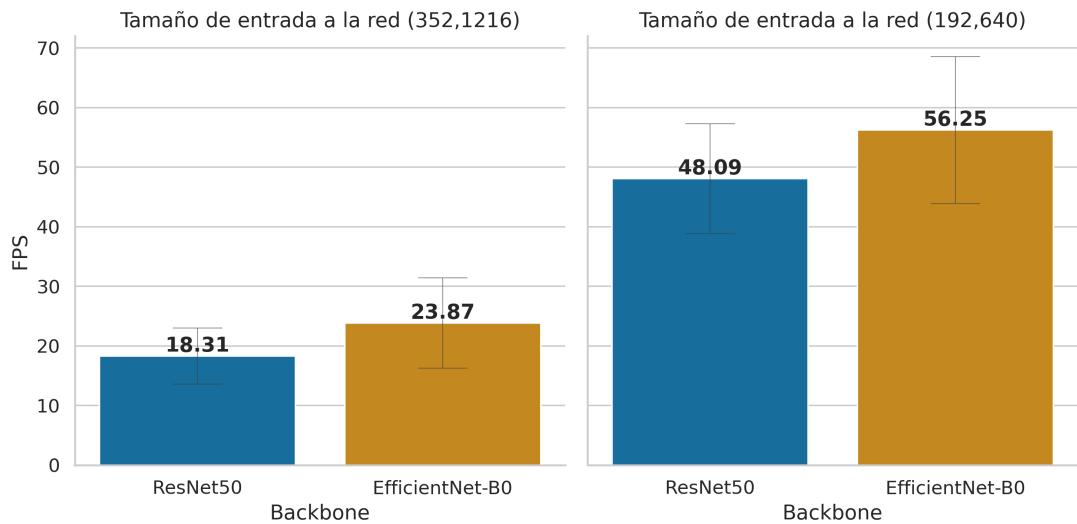


Figura 30: Resultados de.

## 5.2. Resultados cualitativos



Figura 31: Comparación cualitativa de resultados en el conjunto de evaluación entre el modelo DPT original y el modelo DPT con las modificaciones desarrolladas en este trabajo. Para facilitar su visualización, el rango de las profundidades predichas se ha ajustado a toda la escala de grises.

## 6. Discusión

Hablar y comentar los resultados obtenidos con otros resultados (convolucionales? y transformers), si los resultados en la jetson son lentos y no se puede acceder a otra más potente, buscar referencias para ver el incremento de rendimiento que presentan con otros algoritmos para poder dar una estimación de cuanto más rápido podría ir en otro hardware embebido más potente.

Mencionar que las capas eficientes no afectan muchisimo debido a que el tamaño de las cadenas de tokens no son demasiado grandes (estamos trabajando con imágenes pequeñas y no con larguisimas cadenas de texto para las que fueron diseñadas).

Mencionar que la atención eficiente no se nota hasta imágenes más grandes pero atender también a los requisitos de memoria.

## 7. Conclusiones y Lineas Futuras

Overview del trabajo y de los resultados obtenidos, hacer hincapié en los modelos producidos y su utilidad, hablar del interés de la gente en optimizar este tipo de modelos (github)...

Decir que queda pendiente probar en otro hardware empotrado, entrenar con más datos, reentrenar desde cero sin hacer warmstart, entrenar en imagenet antes las transformers con las capas de atención modificadas, explorar más técnicas de optimización que no se han empleado, más atenciones eficaces, aplicarlo a imágenes mucho mayores. Por otro lado, desarrollar aplicaciones con los modelos producidos como hacer un slam monocular, conducción autónoma,

$$\begin{aligned}
L(\hat{d}, d) &= \frac{1}{n} \sum_p (\ln \hat{d}_p - \ln d_p)^2 - \frac{1}{n^2} \left( \sum_p (\ln \hat{d}_p - \ln d_p) \right)^2 \\
&= \frac{1}{n} \sum_p (\ln \alpha \hat{d}_p - \ln d_p)^2 - \frac{1}{n^2} \left( \sum_p (\ln \alpha \hat{d}_p - \ln d_p) \right)^2 \\
&= \frac{1}{n} \sum_p (\ln \alpha + \ln \frac{\hat{d}_p}{d_p})^2 - \frac{1}{n^2} \left( \sum_p (\ln \alpha + \ln \frac{\hat{d}_p}{d_p}) \right)^2 \\
&= \frac{1}{n} \sum_p (\beta + \ln \frac{\hat{d}_p}{d_p})^2 - \frac{1}{n^2} \left( \sum_p (\beta + \ln \frac{\hat{d}_p}{d_p}) \right)^2 \\
&= \frac{1}{n} \sum_p (\beta^2 + \ln \frac{\hat{d}_p}{d_p}^2 + 2\beta \ln \frac{\hat{d}_p}{d_p}) - \frac{1}{n^2} \left( \beta n + \sum_p (\ln \frac{\hat{d}_p}{d_p}) \right)^2 \\
&= \frac{1}{n} (\beta^2 n + \sum_p (\ln \frac{\hat{d}_p}{d_p})^2 + \sum_p (2\beta \ln \frac{\hat{d}_p}{d_p})) - \frac{1}{n^2} \left( \beta^2 n^2 + (\sum_p (\ln \frac{\hat{d}_p}{d_p}))^2 + 2\beta n \sum_p (\ln \frac{\hat{d}_p}{d_p}) \right) \\
&= \beta^2 + \frac{\sum_p (\ln \frac{\hat{d}_p}{d_p})^2}{n} + \frac{\sum_p (2\beta \ln \frac{\hat{d}_p}{d_p})}{n} - \left( \beta^2 + \frac{(\sum_p (\ln \frac{\hat{d}_p}{d_p}))^2}{n^2} + \frac{2\beta n \sum_p (\ln \frac{\hat{d}_p}{d_p})}{n^2} \right) \\
&= \beta^2 - \beta^2 + \frac{\sum_p (2\beta \ln \frac{\hat{d}_p}{d_p})}{n} - \frac{2\beta \sum_p (\ln \frac{\hat{d}_p}{d_p})}{n} + \frac{\sum_p (\ln \frac{\hat{d}_p}{d_p})^2}{n} - \frac{(\sum_p (\ln \frac{\hat{d}_p}{d_p}))^2}{n^2} \\
&= \frac{\sum_p (\ln \frac{\hat{d}_p}{d_p})^2}{n} - \frac{(\sum_p (\ln \frac{\hat{d}_p}{d_p}))^2}{n^2} \\
&= \frac{1}{n} \sum_p (\ln \hat{d}_p - \ln d_p)^2 - \frac{1}{n^2} \left( \sum_p (\ln \hat{d}_p - \ln d_p) \right)^2
\end{aligned}$$

## Bibliografía

- [1] R. Hartley and A. Zisserman, *Two-View Geometry*, p. 237–238. Cambridge University Press, 2 ed., 2004.
- [2] M. Kalloniatis and C. Luu, *Webvision: The Organization of the Retina and Visual System*. 2005.
- [3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
- [4] R. Ranftl, A. Bochkovskiy, and V. Koltun, “Vision transformers for dense prediction,” 2021.
- [5] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.
- [6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning Representations by Back-propagating Errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
- [7] S. Raschka, “Model evaluation, model selection, and algorithm selection in machine learning,” *CoRR*, vol. abs/1811.12808, 2018.
- [8] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML’10, (Madison, WI, USA), p. 807–814, Omnipress, 2010.
- [9] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” *arXiv preprint arXiv:1606.08415*, 2016.
- [10] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, “Efficient backprop,” in *Neural networks: Tricks of the trade*, pp. 9–48, Springer, 2012.
- [11] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *Journal of Machine Learning Research*, vol. 12, no. 61, pp. 2121–2159, 2011.
- [12] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR (Poster)*, 2015.
- [13] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *International Conference on Learning Representations*, 2019.
- [14] N. Keskar, J. Nocedal, P. Tang, D. Mudigere, and M. Smelyanskiy, “On large-batch training for deep learning: Generalization gap and sharp minima,” 2017. 5th International Conference on Learning Representations, ICLR 2017 ; Conference date: 24-04-2017 Through 26-04-2017.
- [15] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1724–1734, Association for Computational Linguistics, Oct. 2014.
- [16] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [17] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, *Object Recognition with Gradient-Based Learning*, pp. 319–345. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999.

- [18] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision*, pp. 818–833, Springer, 2014.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
- [20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.
- [21] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114, PMLR, 09–15 Jun 2019.
- [22] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, “Mnasnet: Platform-aware neural architecture search for mobile,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2815–2823, 2019.
- [23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.
- [24] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.
- [25] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” Jan. 2015. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.
- [26] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 2048–2057, PMLR, 07–09 Jul 2015.
- [27] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, “The application of two-level attention models in deep convolutional neural network for fine-grained image classification,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 842–850, 2015.
- [28] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, D. Ramanan, and T. S. Huang, “Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2956–2964, 2015.
- [29] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, “Efficient Transformers: A Survey,” *arXiv e-prints*, p. arXiv:2009.06732, Sept. 2020.
- [30] J. Qiu, H. Ma, O. Levy, W.-t. Yih, S. Wang, and J. Tang, “Blockwise self-attention for long document understanding,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, (Online), pp. 2555–2565, Association for Computational Linguistics, Nov. 2020.
- [31] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, “Stand-alone self-attention in vision models,” in *Advances in Neural Information Processing Sys-*

- tems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.
- [32] R. Child, S. Gray, A. Radford, and I. Sutskever, “Generating Long Sequences with Sparse Transformers,” *arXiv e-prints*, p. arXiv:1904.10509, Apr. 2019.
  - [33] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” 2020.
  - [34] P. J. Liu\*, M. Saleh\*, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, “Generating wikipedia by summarizing long sequences,” in *International Conference on Learning Representations*, 2018.
  - [35] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, “Axial attention in multidimensional transformers,” 2019.
  - [36] N. Kitaev, L. Kaiser, and A. Levskaya, “Reformer: The efficient transformer,” in *International Conference on Learning Representations*, 2020.
  - [37] A. Roy, M. Saffar, A. Vaswani, and D. Grangier, “Efficient Content-Based Sparse Attention with Routing Transformers,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 53–68, 02 2021.
  - [38] Q. Zhang and Y. Yang, “Rest: An efficient transformer for visual recognition,” *arXiv preprint arXiv:2105.13677v2*, 2021.
  - [39] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, “Lformer: Self-attention with linear complexity,” 2020.
  - [40] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, “Image transformer,” in *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, pp. 4055–4064, PMLR, 10–15 Jul 2018.
  - [41] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *Computer Vision – ECCV 2020* (A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds.), (Cham), pp. 213–229, Springer International Publishing, 2020.
  - [42] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021.
  - [43] S. F. Bhat, I. Alhashim, and P. Wonka, “Adabins: Depth estimation using adaptive bins,” 2020.
  - [44] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “Transunet: Transformers make strong encoders for medical image segmentation,” *arXiv preprint arXiv:2102.04306*, 2021.
  - [45] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” *arXiv preprint arXiv:2103.14030*, 2021.
  - [46] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F. Qian, “Monocular depth estimation based on deep learning: An overview,” *Science China Technological Sciences*, vol. 63, p. 1612–1627, Jun 2020.

- [47] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised learning of depth and ego-motion from video,” in *CVPR*, 2017.
- [48] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki, “Sfm-net: Learning of structure and motion from video,” 2017.
- [49] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid, “Unsupervised scale-consistent depth and ego-motion learning from monocular video,” in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.
- [50] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, “Digging into self-supervised monocular depth estimation,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3827–3837, 2019.
- [51] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, “Learning depth from monocular videos using direct methods,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2022–2030, 2018.
- [52] Z. Yin and J. Shi, “Geonet: Unsupervised learning of dense depth, optical flow and camera pose,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1983–1992, 2018.
- [53] F. Tosi, F. Aleotti, M. Poggi, and S. Mattoccia, “Learning monocular depth estimation infusing traditional stereo knowledge,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9791–9801, 2019.
- [54] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin, “Single view stereo matching,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 155–163, 2018.
- [55] N. Smolyanskiy, A. Kamenev, and S. Birchfield, “On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1120–11208, 2018.
- [56] J. Xie, R. B. Girshick, and A. Farhadi, “Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks,” in *ECCV*, 2016.
- [57] Y. Xu, X. Zhu, J. Shi, G. Zhang, H. Bao, and H. Li, “Depth completion from sparse lidar data with depth-normal constraints,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [58] F. Ma, G. V. Cavalheiro, and S. Karaman, “Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera,” 2019.
- [59] M. Hu, S. Wang, B. Li, S. Ning, L. Fan, and X. Gong, “Towards precise and efficient image guided depth completion,” 2021.
- [60] L. He, C. Chen, T. Zhang, H. Zhu, and S. Wan, “Wearable depth camera: Monocular depth estimation via sparse optimization under weak supervision,” *IEEE Access*, vol. 6, pp. 41337–41345, 2018.
- [61] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’14, (Cambridge, MA, USA), p. 2366–2374, MIT Press, 2014.

- [62] D. Eigen and R. Fergus, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2650–2658, 2015.
- [63] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.
- [64] H. Jung, Y. Kim, D. Min, C. Oh, and K. Sohn, “Depth prediction from a single image with conditional adversarial networks,” in *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 1717–1721, 2017.
- [65] G. Lin, A. Milan, C. Shen, and I. Reid, “Refinenet: Multi-path refinement networks for high-resolution semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [66] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, “Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.
- [67] Z. Li and N. Snavely, “Megadepth: Learning single-view depth prediction from internet photos,” in *Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [68] S. Vadera and S. Ameen, “Methods for pruning deep neural networks,” 2020.
- [69] S. Han, J. Pool, J. Tran, and W. Dally, “Learning both weights and connections for efficient neural network,” in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.
- [70] A. Polyak and L. Wolf, “Channel-level acceleration of deep face representations,” *IEEE Access*, vol. 3, pp. 2163–2175, 2015.
- [71] H. Hu, R. Peng, Y.-W. Tai, and C.-K. Tang, “Network Trimming: A Data-Driven Neuron Pruning Approach towards Efficient Deep Architectures,” *arXiv e-prints*, p. arXiv:1607.03250, July 2016.
- [72] Y. LeCun, J. Denker, and S. Solla, “Optimal brain damage,” in *Advances in Neural Information Processing Systems* (D. Touretzky, ed.), vol. 2, Morgan-Kaufmann, 1990.
- [73] B. Hassibi, D. Stork, and G. Wolff, “Optimal brain surgeon: Extensions and performance comparisons,” in *Advances in Neural Information Processing Systems* (J. Cowan, G. Tesauro, and J. Alspector, eds.), vol. 6, Morgan-Kaufmann, 1994.
- [74] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, “Pruning convolutional neural networks for resource efficient inference,” 2017.
- [75] C. Wang, R. Grosse, S. Fidler, and G. Zhang, “EigenDamage: Structured pruning in the Kronecker-factored eigenbasis,” in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 6566–6575, PMLR, 09–15 Jun 2019.
- [76] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, “Mixed precision training,” 2018.
- [77] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant,

- K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, “Array programming with NumPy,” *Nature*, vol. 585, pp. 357–362, Sept. 2020.
- [78] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, 2000.
- [79] J. D. Hunter, “Matplotlib: A 2d graphics environment,” *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.
- [80] M. L. Waskom, “seaborn: statistical data visualization,” *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.
- [81] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, eds.), pp. 8024–8035, Curran Associates, Inc., 2019.
- [82] R. Wightman, “Pytorch image models.” <https://github.com/rwightman/pytorch-image-models>, 2019.
- [83] P. Wang, “Performer pytorch.” <https://github.com/lucidrains/performer-pytorch>, 2013.
- [84] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *CVPR09*, 2009.
- [85] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
- [86] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset,” *International Journal of Robotics Research (IJRR)*, 2013.
- [87] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [88] J. Fritsch, T. Kuehnl, and A. Geiger, “A new performance measure and evaluation benchmark for road detection algorithms,” in *International Conference on Intelligent Transportation Systems (ITSC)*, 2013.
- [89] M. Menze and A. Geiger, “Object scene flow for autonomous vehicles,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [90] C. Fellbaum, ed., *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication, Cambridge, MA: MIT Press, 1998.
- [91] Y. Kim, H. Jung, D. Min, and K. Sohn, “Deep monocular depth estimation via integration of global and local predictions,” *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4131–4144, 2018.
- [92] K. Xian, C. Shen, Z. Cao, H. Lu, Y. Xiao, R. Li, and Z. Luo, “Monocular relative depth perception with web stereo data supervision,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [93] C. Wang, S. Lucey, F. Perazzi, and O. Wang, “Web stereo video supervision for depth prediction from dynamic scenes,” 2019.

- [94] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer, “Tartanair: A dataset to push the limits of visual slam,” 2020.
- [95] K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin, and Z. Cao, “Structure-guided ranking loss for single image depth prediction,” in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [96] P. Wang, X. Huang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, “The apolloscape open dataset for autonomous driving and its application,” *IEEE transactions on pattern analysis and machine intelligence*, 2019.
- [97] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou, T. Fang, and L. Quan, “Blendedmvs: A large-scale dataset for generalized multi-view stereo networks,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1787–1796, 2020.
- [98] Q. Wang, S. Zheng, Q. Yan, F. Deng, K. Zhao, and X. Chu, “Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation,” in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, (Los Alamitos, CA, USA), pp. 1–6, IEEE Computer Society, jul 2021.
- [99] W. Chen, Z. Fu, D. Yang, and J. Deng, “Single-image depth perception in the wild,” in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.
- [100] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, “A multi-view stereo benchmark with high-resolution images and multi-camera videos,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [101] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, “A naturalistic open source movie for optical flow evaluation,” in *European Conf. on Computer Vision (ECCV)* (A. Fitzgibbon et al. (Eds.), ed.), Part IV, LNCS 7577, pp. 611–625, Springer-Verlag, Oct. 2012.
- [102] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, “Indoor segmentation and support inference from rgbd images,” in *ECCV*, 2012.
- [103] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, “A benchmark for the evaluation of rgb-d slam systems,” in *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.
- [104] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, “From big to small: Multi-scale local planar guidance for monocular depth estimation,” *arXiv preprint arXiv:1907.10326*, 2019.
- [105] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, “Deep Ordinal Regression Network for Monocular Depth Estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [106] T. Koch, L. Liebel, F. Fraundorfer, and M. Korner, “Evaluation of cnn-based single-image depth estimation methods,” in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.
- [107] C. Cadena, Y. Latif, and I. D. Reid, “Measuring the performance of single image depth estimation methods,” in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4150–4157, 2016.
- [108] M. Harris, “How to optimize data transfers in cuda c/c++,” Dec 2012.
- [109] P. Michel, O. Levy, and G. Neubig, “Are sixteen heads really better than one?,” in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.