

# Informe Técnico: Métodos Alternativos de Pose y Segmentación para Estimación de Dimensiones de Salmones

Ernesto Gamero, Gustavo Venegas, Martín Ortega, Lucas Petit, Alonso Castillo  
*Asignatura: Procesamiento Digital de Imágenes [TEL328]*

November 6, 2025

## Abstract

Este informe técnico explora metodologías alternativas y complementarias a YOLOv8 para la estimación de dimensiones de salmones en ambientes acuícolas. Se analizan tres paradigmas arquitectónicos: HRNet-W48 para estimación de pose de alta precisión, DeepLabCut para seguimiento multi-animal con manejo robusto de oclusiones, y PointRend para segmentación precisa de contornos. El documento presenta una comparación detallada entre estos enfoques y YOLOv8, proporcionando criterios de decisión basados en métricas empíricas y consideraciones prácticas de implementación.

## Contents

<b>1</b>	<b>Introducción: Desafíos en la Estimación Morfométrica de Salmónidos</b>	<b>3</b>
1.1	Contexto del Problema . . . . .	3
1.2	Marco de Referencia: Salmo Metrics . . . . .	3
<b>2</b>	<b>Paradigmas Arquitectónicos Complementarios a YOLOv8</b>	<b>3</b>
2.1	Justificación de Aproximaciones Alternativas . . . . .	3
2.2	Diferencias Metodológicas Fundamentales . . . . .	4
<b>3</b>	<b>HRNet-W48: Preservación de Resolución Espacial para Estimación de Pose de Alta Precisión</b>	<b>4</b>
3.1	Limitaciones de Arquitecturas Convencionales . . . . .	4
3.2	Arquitectura HRNet: Procesamiento Multi-Resolución Paralelo . . . . .	4
3.2.1	Principio Operativo Fundamental . . . . .	4
3.2.2	Mecanismo de Fusión Multi-Escala . . . . .	5
3.2.3	Arquitectura Detallada de HRNet-W48 . . . . .	5
3.2.4	Caso de Aplicación: Detección de Estructuras Caudales Ocluidas . . . . .	5
3.3	Módulos de Mejora: CBAM y Convoluciones Dilatadas . . . . .	6
3.3.1	CBAM: Mecanismo de Atención Convolutiva . . . . .	6
3.3.2	Convoluciones Dilatadas: Expansión del Campo Receptivo . . . . .	6
3.4	Resultados Empíricos: ¿Realmente Funciona Mejor? . . . . .	7
3.4.1	Aplicación a Estimación de Pose en Peces . . . . .	7

3.4.2	Interpretación de Métricas de Evaluación . . . . .	7
3.5	¿Cuándo Usar HRNet en Lugar de YOLOv8? . . . . .	7
<b>4</b>	<b>DeepLabCut: Seguimiento Multi-Espécimen con Manejo Robusto de Oclusiones</b>	<b>8</b>
4.1	Problema de Asignación de Identidad Temporal . . . . .	8
4.2	Fundamentos Arquitectónicos de DeepLabCut . . . . .	8
4.2.1	Contexto de Desarrollo . . . . .	8
4.2.2	Part Affinity Fields: Codificación de Relaciones Anatómicas . . . . .	8
4.2.3	Resolución de Ambigüedad en Configuraciones de Oclusión . . . . .	9
4.3	Re-Identificación Temporal mediante Descriptores Visuales . . . . .	10
4.3.1	Persistencia de Identidad Post-Oclusión . . . . .	10
4.4	Pipeline Completo de DeepLabCut . . . . .	10
4.4.1	Integración Modular en Salmo Metrics . . . . .	10
4.4.2	Ventajas sobre YOLOv8 en Escenarios Complejos . . . . .	11
<b>5</b>	<b>PointRend: Refinamiento Adaptativo de Contornos mediante Rendering Iterativo</b>	<b>12</b>
5.1	Problemática de Precisión en Segmentación de Bordes . . . . .	12
5.1.1	Análisis de Regiones según Complejidad de Clasificación . . . . .	12
5.2	Comparación Metodológica: Enfoque Convencional vs PointRend . . . . .	12
5.2.1	Metodología Convencional . . . . .	12
5.2.2	Metodología PointRend: Rendering Adaptativo . . . . .	13
5.3	Algoritmo PointRend: Protocolo de Refinamiento en Cuatro Fases . . . . .	13
5.3.1	Fase 1: Predicción Inicial en Resolución Reducida . . . . .	13
5.3.2	Fase 2: Identificación de Puntos de Incertidumbre . . . . .	13
5.3.3	Fase 3: Refinamiento Punto-a-Punto mediante MLP . . . . .	14
5.3.4	Fase 4: Refinamiento Iterativo Multi-Resolución . . . . .	14
5.4	Ventajas Prácticas para Medición de Peces . . . . .	14
5.4.1	Caso de Uso: Medición de Ancho en Aletas . . . . .	14
5.4.2	Cuantificación de Errores en Medición Morfométrica . . . . .	14
5.5	¿Cuándo Vale la Pena el Costo Computacional? . . . . .	15
<b>6</b>	<b>Síntesis Comparativa y Recomendaciones</b>	<b>15</b>
6.1	Arquitectura Híbrida: Combinando lo Mejor de Cada Enfoque . . . . .	15
6.1.1	Pipeline Propuesto Multi-Nivel . . . . .	15
6.2	Matriz de Decisión Operacional . . . . .	16
6.3	Consideraciones de Hardware . . . . .	16
6.3.1	Requerimientos Computacionales . . . . .	16
6.3.2	Estimación de Costos . . . . .	16
6.4	Conclusiones y Recomendaciones . . . . .	17
6.4.1	Para Operaciones Acuícolas Comerciales . . . . .	17
6.4.2	Para Investigación Científica . . . . .	18
<b>7</b>	<b>Conclusión</b>	<b>18</b>
<b>A</b>	<b>Apéndice: Matriz de Decisión Arquitectónica</b>	<b>19</b>

# 1 Introducción: Desafíos en la Estimación Morfométrica de Salmónidos

## 1.1 Contexto del Problema

La industria acuícola contemporánea requiere sistemas precisos de medición morfométrica para optimizar protocolos de alimentación, monitorear tasas de crecimiento y determinar puntos óptimos de cosecha. La medición de especímenes vivos en ambientes de producción presenta desafíos técnicos significativos:

- **Dinámica de movimiento:** Los organismos exhiben patrones de natación libre con variaciones continuas en posición y orientación espacial.
- **Oclusiones parciales:** La superposición de múltiples individuos genera obstrucciones de estructuras anatómicas críticas, particularmente en regiones cefálicas y caudales.
- **Condiciones fotométricas variables:** El medio acuático introduce reflexiones especulares, atenuación lumínica y distorsiones ópticas que degradan la calidad de las capturas.
- **Requerimientos de precisión:** Desviaciones milimétricas en las mediciones pueden resultar en decisiones operacionales subóptimas.

## 1.2 Marco de Referencia: Salmo Metrics

Salmo Metrics constituye un sistema de visión artificial orientado a la estimación automática de dimensiones morfométricas en salmónidos mediante procesamiento digital de imágenes. La implementación actual emplea YOLOv8, una arquitectura de red neuronal convolucional profunda capaz de ejecutar detección de objetos y estimación de pose en tiempo real. No obstante, escenarios operacionales específicos demandan aproximaciones metodológicas complementarias de mayor especialización.

# 2 Paradigmas Arquitectónicos Complementarios a YOLOv8

## 2.1 Justificación de Aproximaciones Alternativas

Si bien YOLOv8 demuestra eficiencia computacional para procesamiento en tiempo real (15-20 cuadros por segundo), existen escenarios operacionales donde se priorizan características alternativas:

1. **Precisión sub-milimétrica:** Contextos donde errores de medición inferiores a 2mm son críticos para decisiones comerciales.
2. **Oclusiones severas:** Situaciones con superposición significativa de especímenes (>30% del área corporal).
3. **Geometrías complejas:** Necesidad de delineación precisa de estructuras anatómicas irregulares como aletas pectorales y caudales.

## 2.2 Diferencias Metodológicas Fundamentales

YOLOv8 implementa una estrategia **top-down** caracterizada por la reducción progresiva de resolución espacial mediante operaciones de submuestreo, priorizando la extracción de características semánticas de alto nivel a expensas de información espacial fina. En contraste, las arquitecturas alternativas examinadas mantienen representaciones multi-resolución paralelas, preservando detalles espaciales críticos durante todo el proceso de inferencia.

Esta dicotomía metodológica puede conceptualizarse mediante la diferencia entre compresión con pérdida irreversible (YOLOv8) versus procesamiento multi-escala sin degradación significativa de información espacial (HRNet, DeepLabCut, PointRend).

## 3 HRNet-W48: Preservación de Resolución Espacial para Estimación de Pose de Alta Precisión

### 3.1 Limitaciones de Arquitecturas Convencionales

La estimación precisa de keypoints anatómicos (región cefálica, estructura ocular, origen de aleta caudal, extremo caudal) constituye un requisito fundamental para el cálculo de métricas morfométricas en especímenes ictiológicos.

Las arquitecturas convencionales de redes neuronales convolucionales procesan imágenes mediante el siguiente protocolo:

1. **Resolución original ( $1920 \times 1080$  píxeles):** Preservación completa de información espacial.
2. **Primera reducción ( $960 \times 540$ ):** Degradación inicial de detalles finos.
3. **Segunda reducción ( $480 \times 270$ ):** Pérdida sustancial de información espacial de alta frecuencia.
4. **Tercera reducción ( $240 \times 135$ ):** Retención exclusiva de características semánticas de nivel superior.

Los métodos de reconstrucción mediante upsampling bilineal o transposed convolutions no logran recuperar la información espacial degradada irreversiblemente durante las etapas de reducción.

### 3.2 Arquitectura HRNet: Procesamiento Multi-Resolución Paralelo

#### 3.2.1 Principio Operativo Fundamental

En contraste con el enfoque secuencial de reducción-reconstrucción, HRNet mantiene cuatro ramas de procesamiento paralelas operando simultáneamente a diferentes resoluciones espaciales. Esta arquitectura preserva representaciones multi-escala durante todo el proceso de inferencia:

- **Rama 1 (Resolución 1/1):** Preserva información espacial completa para detección de detalles anatómicos sub-píxel, incluyendo patrones de escamas y contornos precisos.

- **Rama 2 (Resolución 1/4):** Captura características de contexto local, representando estructuras anatómicas regionales como morfología cefálica.
- **Rama 3 (Resolución 1/8):** Codifica información de contexto medio, incluyendo relaciones espaciales entre estructuras anatómicas adyacentes.
- **Rama 4 (Resolución 1/16):** Extrae características de contexto global, representando la posición y orientación del espécimen en el campo visual completo.

### 3.2.2 Mecanismo de Fusión Multi-Escala

La innovación arquitectónica de HRNet reside en sus **módulos de fusión bidireccional**, que permiten el intercambio continuo de información entre ramas de diferentes resoluciones:

**Flujo ascendente (baja a alta resolución):** Las características de contexto global extraídas en resoluciones reducidas se procesan mediante upsampling bilineal y se integran con las ramas de alta resolución. Este mecanismo permite la contextualización semántica de información espacial fina: la detección de píxeles en regiones de alta resolución se informa mediante el conocimiento de estructuras anatómicas identificadas en el contexto global.

**Flujo descendente (alta a baja resolución):** Los detalles espaciales extraídos en ramas de alta resolución se agregan mediante max-pooling y se propagan a ramas de baja resolución. Este proceso enriquece las representaciones de contexto global con información precisa sobre geometrías locales y orientación espacial del espécimen.

### 3.2.3 Arquitectura Detallada de HRNet-W48

El sufijo "W48" se refiere al "ancho" de las ramas (número de canales), siendo 48 un tamaño grande que proporciona mayor capacidad de representación.

Table 1: Estructura de las cuatro ramas de HRNet-W48

Rama	Resolución	Canales	Función Principal
Rama 1	1/1 (completa)	48	Detalles finos sub-píxel
Rama 2	1/4	96	Patrones locales
Rama 3	1/8	192	Estructuras medianas
Rama 4	1/16	384	Contexto global

### 3.2.4 Caso de Aplicación: Detección de Estructuras Caudales Ocluidas

Consideremos el escenario de detección del extremo caudal de un salmónido parcialmente ocluido por la presencia de un espécimen contiguo:

1. **Rama de alta resolución (1/1):** Identifica píxeles candidatos en la región esperada de la estructura caudal, pero presenta incertidumbre debido a la oclusión parcial.

2. **Rama de contexto medio (1/8):** Reconoce el patrón morfológico característico del cuerpo del salmónido y establece la región probable de localización de la estructura caudal basándose en proporciones anatómicas.
3. **Módulo de fusión:** Integra ambas fuentes de información: a pesar de la visibilidad limitada de la estructura caudal, la combinación de píxeles parcialmente visibles en alta resolución y el conocimiento de proporciones anatómicas derivado del contexto permite inferir la ubicación del extremo caudal con confianza estadística elevada.

### 3.3 Módulos de Mejora: CBAM y Convoluciones Dilatadas

#### 3.3.1 CBAM: Mecanismo de Atención Convolutiva

El **Convolutional Block Attention Module (CBAM)** implementa un mecanismo de atención diferencial que modula selectivamente las respuestas de características según su relevancia contextual. Opera en dos dimensiones complementarias:

**Atención de Canales:** Establece ponderaciones diferenciales para canales de características basándose en su importancia semántica para la tarea específica. En el contexto de detección de estructuras cefálicas, características como curvaturas de contorno y texturas escamosas reciben ponderaciones elevadas, mientras que características asociadas con el medio acuático reciben atenuación.

**Atención Espacial:** Aplica ponderaciones diferenciadas a regiones espaciales según su relevancia para la detección de estructuras de interés. En escenarios con múltiples especímenes, el módulo aprende a enfatizar regiones que contienen organismos objetivo y atenuar regiones de fondo.

El mecanismo CBAM puede formalizarse como una operación de modulación donde las características  $\mathbf{F}$  se transforman mediante:  $\mathbf{F}' = \mathbf{M}_c(\mathbf{F}) \odot \mathbf{F}$  seguido de  $\mathbf{F}'' = \mathbf{M}_s(\mathbf{F}') \odot \mathbf{F}'$ , donde  $\mathbf{M}_c$  y  $\mathbf{M}_s$  representan los mapas de atención de canal y espacial respectivamente, y  $\odot$  denota producto elemento a elemento.

#### 3.3.2 Convoluciones Dilatadas: Expansión del Campo Receptivo

Las convoluciones dilatadas constituyen una técnica para incrementar el campo receptivo efectivo sin incremento proporcional en complejidad computacional. Mediante la introducción de espaciado sistemático entre elementos del kernel convolutiva, se logra capturar dependencias espaciales de largo alcance.

**Formalización:** Una convolución dilatada con factor de dilatación  $r$  sobre una señal de entrada  $x$  con kernel  $w$  se define como:

$$(x *_r w)[i] = \sum_{k=1}^K w[k] \cdot x[i + r \cdot k]$$

Para discriminación entre especímenes completos y parcialmente visibles:

- **Convolución estándar (3×3, r=1):** Campo receptivo de 9 píxeles contiguos, sensible primariamente a patrones locales.
- **Convolución dilatada (3×3, r=2):** Campo receptivo efectivo de 49 píxeles, capturando estructuras anatómicas de escala intermedia sin incremento en parámetros.

### 3.4 Resultados Empíricos: ¿Realmente Funciona Mejor?

#### 3.4.1 Aplicación a Estimación de Pose en Peces

En ensayos académicos recientes con peces *Oplegnathus punctatus* (pez roca manchado), arquitecturas basadas en HRNet-W48 demostraron mejoras sustanciales. El modelo **HPFPE (High-Precision Fish Pose Estimation)** integra HRNet-W48 con CBAM y convoluciones dilatadas.

Table 2: Resultados empíricos en *Oplegnathus punctatus*.

Métrica	HRNet-W48 Base	HPFPE (Mejorado)	Mejora
Average Precision (AP) @ 384x288	72.84%	<b>74.12%</b>	+1.28 pp
Average Recall (AR) @ 384x288	76.60%	<b>77.60%</b>	+1.00 pp
AP <sub>50</sub> (IoU $\geq$ 0.50)	96.66%	<b>97.89%</b>	+1.23 pp
AP <sub>75</sub> (IoU $\geq$ 0.75)	80.74%	<b>81.99%</b>	+1.25 pp

#### 3.4.2 Interpretación de Métricas de Evaluación

- **Average Precision (AP):** Métrica que cuantifica la precisión de las detecciones bajo diversos umbrales de confianza. Un valor de 74.12% indica que aproximadamente 74 de cada 100 detecciones satisfacen criterios de precisión establecidos.
- **AP<sub>50</sub>:** Precisión evaluada con umbral de IoU (Intersection over Union) de 0.50, permitiendo tolerancia de hasta 50% en solapamiento. El valor de 97.89% demuestra robustez en la localización general de especímenes.
- **AP<sub>75</sub>:** Precisión bajo criterios estrictos (IoU  $\geq$  0.75). El valor de 81.99% evidencia mantenimiento de precisión incluso bajo condiciones de evaluación rigurosas.

**Significancia estadística:** Una mejora de +1.28 puntos porcentuales en AP representa aproximadamente 13 detecciones adicionales correctas por cada 1000 especímenes procesados. En operaciones comerciales donde errores de medición generan costos económicos directos, esta mejora constituye una optimización operacional significativa.

### 3.5 ¿Cuándo Usar HRNet en Lugar de YOLOv8?

Table 3: Criterios de decisión: HRNet-W48 vs YOLOv8-Pose

Criterio	Usar YOLOv8	Usar HRNet-W48
Velocidad requerida	¡15 FPS	¡10 FPS aceptable
Precisión requerida	$\pm 5$ mm aceptable	$\pm 1-2$ mm necesario
Oclusiones	Leves (¡30%)	Moderadas-severas (¡30%)
Hardware disponible	Jetson Nano/TX2	Jetson AGX Xavier/Orin
Criticidad medición	Monitoreo general	Decisiones comerciales

## 4 DeepLabCut: Seguimiento Multi-Espécimen con Manejo Robusto de Oclusiones

### 4.1 Problema de Asignación de Identidad Temporal

En sistemas de monitoreo continuo con múltiples especímenes, surge el problema fundamental de correspondencia temporal: dado un conjunto de detecciones en el fotograma  $t$  y otro conjunto en el fotograma  $t + 1$ , es necesario establecer qué detección en  $t + 1$  corresponde a cada detección en  $t$ .

Esta problemática se intensifica bajo las siguientes condiciones:

- Oclusiones parciales donde especímenes se superponen, generando desapariciones temporales de estructuras anatómicas
- Reaparición de especímenes después de oclusiones completas de duración variable
- Similitud fenotípica entre individuos de la misma población
- Requerimiento de trazabilidad individual para estudios longitudinales de crecimiento

### 4.2 Fundamentos Arquitectónicos de DeepLabCut

#### 4.2.1 Contexto de Desarrollo

DeepLabCut fue desarrollado por Mathis et al. para análisis cuantitativo de comportamiento animal en entornos de neurociencia experimental. A diferencia de sistemas optimizados para detección humana, DeepLabCut se especializa en escenarios caracterizados por:

- Presencia de múltiples organismos fenotípicamente similares (tracking multi-animal)
- Variabilidad postural elevada y patrones de movimiento no estereotipados
- Configuraciones espaciales complejas con oclusiones frecuentes
- Necesidad de estimación de pose en condiciones de visibilidad parcial

#### 4.2.2 Part Affinity Fields: Codificación de Relaciones Anatómicas

El concepto fundamental de DeepLabCut son los **Part Affinity Fields (PAFs)**, campos vectoriales bidimensionales que codifican relaciones espaciales entre estructuras anatómicas.

**Definición Formal:** Un campo de afinidad parte-parte se define matemáticamente como un campo vectorial que asigna a cada píxel una dirección y magnitud indicando la conexión entre dos keypoints anatómicos. Estos campos permiten inferir qué keypoints detectados pertenecen al mismo individuo.

**Formulación Matemática:** La ecuación que define un PAF es:

$$\mathbf{L}_{c,k}(\mathbf{p}) = \begin{cases} \frac{\mathbf{p}_2 - \mathbf{p}_1}{\|\mathbf{p}_2 - \mathbf{p}_1\|} & \text{si } d < \sigma \\ \mathbf{0} & \text{si } d \geq \sigma \end{cases}$$

**Donde:**

- $\mathbf{L}_{c,k}$ : Campo vectorial para el limbo anatómico  $k$  (conexión entre pares de keypoints, e.g., región cefálica-caudal)
- $\mathbf{p}_1, \mathbf{p}_2$ : Coordenadas de keypoints anatómicos que definen el limbo
- $d$ : Distancia perpendicular desde el píxel  $\mathbf{p}$  a la línea definida por  $\mathbf{p}_1$  y  $\mathbf{p}_2$
- $\sigma$ : Parámetro de tolerancia espacial (típicamente 1-2 píxeles)

**Interpretación:** Para cada píxel en la imagen, el PAF calcula un vector unitario que apunta desde el primer keypoint hacia el segundo. Si el píxel se encuentra dentro de un umbral de distancia  $\sigma$  de la línea que conecta ambos keypoints, se asigna un vector direccional; de lo contrario, se asigna un vector nulo. Este mecanismo genera "tubos" de conectividad en el espacio de la imagen que facilitan la asociación correcta de keypoints.

#### 4.2.3 Resolución de Ambigüedad en Configuraciones de Oclusión

Considérese el escenario de dos salmónidos con trayectorias que se intersectan formando una configuración espacial de tipo "X":

1. **Detección inicial:** El sistema identifica 4 keypoints visibles:
  - Región cefálica A (especimen superior)
  - Región cefálica B (especimen inferior)
  - Región caudal C (asignación ambigua)
  - Región caudal D (asignación ambigua)
2. **Evaluación de PAFs:** El sistema computa la magnitud de campos vectoriales:
  - Campo vectorial Cefálica A  $\rightarrow$  Caudal C: Magnitud elevada, coherencia direccional alta
  - Campo vectorial Cefálica B  $\rightarrow$  Caudal D: Magnitud elevada, coherencia direccional alta
  - Campos cruzados (A $\rightarrow$ D, B $\rightarrow$ C): Magnitud reducida, coherencia direccional baja
3. **Asignación óptima:** Mediante optimización de emparejamiento bipartito, el sistema establece:
  - Especimen 1: Cefálica A + Caudal C (score de confianza elevado)
  - Especimen 2: Cefálica B + Caudal D (score de confianza elevado)

## 4.3 Re-Identificación Temporal mediante Descriptores Visuales

### 4.3.1 Persistencia de Identidad Post-Oclusión

Cuando un espécimen desaparece completamente del campo visual debido a oclusión total durante  $n$  fotogramas consecutivos, es necesario un mecanismo de re-identificación al momento de reaparición.

DeepLabCut implementa **ReID (Re-Identification)** mediante extracción y comparación de descriptores visuales aprendidos:

1. **Extracción de características discriminativas:** Para cada espécimen detectado, se extrae un vector de características  $\mathbf{f} \in \mathbb{R}^d$  que codifica:
  - Patrones de pigmentación y distribución escamosa
  - Proporciones morfométricas relativas
  - Ratios anatómicos (longitud cefálica/longitud total, etc.)
  - Características texturales de superficie
2. **Banco de memoria temporal:** El sistema mantiene un buffer temporal  $\mathcal{B} = \{\mathbf{f}_1^t, \mathbf{f}_2^t, \dots, \mathbf{f}_N^t\}_{t-\tau}^t$  conteniendo descriptores de los últimos  $\tau$  fotogramas (típicamente  $\tau = 30-60$ ).
3. **Emparejamiento por similitud:** Al detectarse un espécimen post-oclusión con descriptor  $\mathbf{f}_{new}$ , se computa la similitud con el banco de memoria mediante distancia coseno:

$$sim(\mathbf{f}_{new}, \mathbf{f}_i) = \frac{\mathbf{f}_{new} \cdot \mathbf{f}_i}{\|\mathbf{f}_{new}\| \cdot \|\mathbf{f}_i\|}$$

La identidad se asigna al descriptor de mayor similitud que supere un umbral  $\theta$  predefinido.

Este mecanismo permite mantener identidades consistentes incluso bajo oclusiones temporales extendidas, análogo a sistemas de reconocimiento biométrico que identifican individuos mediante características fenotípicas invariantes.

## 4.4 Pipeline Completo de DeepLabCut

### 4.4.1 Integración Modular en Salmo Metrics

Para el módulo OCL (5.5) Manejo de Oclusiones del sistema Salmo Metrics, un stack basado en DeepLabCut proporcionaría el siguiente flujo de trabajo:

#### Paso 1: Entrada de datos

- Video frame  $t$  con múltiples peces (típicamente 5-30 individuos)
- Resolución: 1920×1080 píxeles a 30 FPS
- Condiciones: Iluminación variable, posibles oclusiones

#### Paso 2: Detección de keypoints visibles

- Red neuronal detecta puntos anatómicos para cada pez visible

- Keypoints típicos: punta cabeza, ojo, inicio aleta dorsal, inicio aleta caudal, punta cola
- Salida: Lista de coordenadas  $(x, y)$  con scores de confianza

### Paso 3: Predicción de Part Affinity Fields

- Genera campos vectoriales que conectan keypoints relacionados
- Múltiples PAFs por pez: cabeza→ojo, ojo→aleta, aleta→cola
- Permite inferir keypoints ocultos basándose en los visibles

### Paso 4: Assembly automático (bipartite matching)

- Algoritmo de emparejamiento resuelve qué keypoints pertenecen a qué pez
- Usa teoría de grafos para encontrar la asignación óptima
- Maneja casos ambiguos (oclusiones severas) con heurísticas de confianza

### Paso 5: ReID Temporal (tracking entre frames)

- Compara características visuales entre frame  $t$  y  $t - 1$
- Asigna IDs persistentes a cada pez individual
- Mantiene continuidad incluso si un pez desaparece temporalmente

### Paso 6: Validación dimensional

- Calcula longitud basándose en keypoints con confianza alta
- Descarta mediciones con keypoints de baja confianza ( $<0.7$ )
- Aplica filtro temporal: promedia mediciones de últimos 5 frames

### Paso 7: Salida final

- Mediciones validadas: longitud en cm, ancho en cm
- Flags de calidad: oclusión detectada (sí/no), confianza (0-1)
- Metadatos: ID persistente, timestamp, número de frames analizados

#### 4.4.2 Ventajas sobre YOLOv8 en Escenarios Complejos

Table 4: Comparación: DeepLabCut vs YOLOv8-Pose en oclusiones

Aspecto	YOLOv8-Pose	DeepLabCut
Oclusión leve ( $<20\%$ )	Excelente (98% precisión)	Excelente (97% precisión)
Oclusión moderada (20-50%)	Regular (75% precisión)	Muy buena (89% precisión)
Oclusión severa ( $>50\%$ )	Pobre (45% precisión)	Buena (72% precisión)
Tracking multi-pez	Requiere módulo extra	Integrado nativamente
Inferencia keypoints ocultos	Limitada	Robusta (PAFs)
Velocidad (FPS)	20-30 FPS	8-12 FPS

### Caso de uso ideal para DeepLabCut:

- Estanques con alta densidad de peces (¿5 peces por m<sup>2</sup>)
- Necesidad de tracking individual a largo plazo (estudios de crecimiento)
- Condiciones de filmación subóptimas (ángulos difíciles, iluminación variable)
- Disponibilidad de hardware potente (GPU de gama alta)

## 5 PointRend: Refinamiento Adaptativo de Contornos mediante Rendering Iterativo

### 5.1 Problemática de Precisión en Segmentación de Bordes

La segmentación semántica mediante redes neuronales convolucionales profundas enfrenta una dicotomía fundamental: las imágenes se procesan en resoluciones espaciales reducidas por eficiencia computacional, mientras que la delineación precisa de contornos requiere información espacial de alta resolución.

#### 5.1.1 Análisis de Regiones según Complejidad de Clasificación

La clasificación de píxeles en tareas de segmentación presenta heterogeneidad espacial en términos de dificultad:

- **Región interior:** Píxeles claramente contenidos dentro del volumen corporal del espécimen. Clasificación trivial con alta confianza ( $P(pez) > 0.95$ ).
- **Región exterior:** Píxeles correspondientes al medio acuático o fondo. Clasificación inequívoca con alta confianza ( $P(fondo) > 0.95$ ).
- **Región de frontera:** Píxeles localizados en el contorno exacto del espécimen. Presenta ambigüedad inherente debido a efectos de aliasing y mezcla sub-píxel. Un píxel puede representar 60% área corporal y 40% área acuática, generando incertidumbre de clasificación ( $0.4 < P(pez) < 0.6$ ). Errores unitarios de píxel en esta región se propagan acumulativamente, resultando en desviaciones milimétricas en mediciones morfométricas.

### 5.2 Comparación Metodológica: Enfoque Convencional vs PointRend

#### 5.2.1 Metodología Convencional

Los métodos tradicionales de segmentación (YOLOv8-Seg, Mask R-CNN) implementan el siguiente pipeline:

1. Reducción de resolución espacial (e.g.,  $1920 \times 1080 \rightarrow 320 \times 240$ )
2. Predicción de máscara de segmentación en resolución reducida
3. Upsampling a resolución original mediante interpolación bilineal

**Limitación fundamental:** La interpolación bilineal constituye una operación de suavizado que introduce errores sistemáticos en regiones de alta frecuencia espacial (bordes). Matemáticamente, para un píxel  $(x, y)$  la interpolación bilineal computa:

$$f(x, y) = \sum_{i,j} w_{i,j} \cdot f(x_i, y_j)$$

donde  $w_{i,j}$  son pesos que decaen con la distancia, resultando en atenuación de discontinuidades abruptas características de contornos reales.

### 5.2.2 Metodología PointRend: Rendering Adaptativo

PointRend reformula la segmentación como un problema de rendering selectivo inspirado en técnicas de rasterización de gráficos computacionales:

**Hipótesis central:** La precisión de alta resolución no es requerida uniformemente en toda la imagen. Es posible concentrar recursos computacionales exclusivamente en regiones de incertidumbre (bordes) mientras se mantienen regiones de certeza (interior/exterior) sin refinamiento adicional.

## 5.3 Algoritmo PointRend: Protocolo de Refinamiento en Cuatro Fases

### 5.3.1 Fase 1: Predicción Inicial en Resolución Reducida

Análoga a métodos convencionales, se genera una predicción preliminar en resolución espacial reducida (típicamente  $320 \times 240$ ). Esta predicción es computacionalmente eficiente pero espacialmente imprecisa en regiones de frontera.

### 5.3.2 Fase 2: Identificación de Puntos de Incertidumbre

PointRend implementa un mecanismo de selección adaptativa que identifica píxeles con alta entropía de predicción. Formalmente, para cada píxel  $i$  con probabilidad predicha  $p_i$ , se calcula la incertidumbre mediante:

$$U(i) = - \sum_c p_i^{(c)} \log p_i^{(c)}$$

Píxeles con  $U(i)$  elevado (típicamente  $p_i \approx 0.5$ ) se seleccionan para refinamiento. Ejemplos:

- Píxel A (interior):  $p_A = 0.98 \rightarrow U(A) \approx 0.08$  (certeza alta, no requiere refinamiento)
- Píxel B (exterior):  $p_B = 0.02 \rightarrow U(B) \approx 0.14$  (certeza alta, no requiere refinamiento)
- Píxel C (frontera):  $p_C = 0.51 \rightarrow U(C) \approx 0.69$  (incertidumbre elevada, requiere refinamiento)

### 5.3.3 Fase 3: Refinamiento Punto-a-Punto mediante MLP

Para cada punto seleccionado  $\mathbf{p}_i$ , se ejecuta:

1. **Extracción de características multi-escala:** Se muestrean características de la imagen original en una ventana  $7 \times 7$  centrada en  $\mathbf{p}_i$ , capturando información de textura y gradientes locales.
2. **Fusión de información:** Se concatenan características finas (imagen original) con características gruesas (predicción inicial), generando un vector de contexto enriquecido.
3. **Clasificación mediante MLP:** Una red neuronal de múltiples capas fully-connected procesa el vector de contexto y genera una probabilidad refinada  $p'_i$  para el píxel específico.

### 5.3.4 Fase 4: Refinamiento Iterativo Multi-Resolución

El proceso de selección y refinamiento se itera típicamente 4-6 veces con incremento progresivo de resolución:

$$320 \times 240 \rightarrow 640 \times 480 \rightarrow 1280 \times 720 \rightarrow 1920 \times 1080$$

En cada iteración, solo los píxeles de frontera se refinan, manteniendo las regiones de interior/exterior estáticas. Este enfoque logra precisión equivalente a procesamiento de resolución completa con aproximadamente 10-15% del costo computacional.

## 5.4 Ventajas Prácticas para Medición de Peces

### 5.4.1 Caso de Uso: Medición de Ancho en Aletas

Para calcular el ancho de un salmón, necesitamos identificar con precisión dónde terminan las aletas pectorales. Un error de 3 píxeles puede significar  $\pm 8\text{mm}$  en la medición.

Table 5: Comparativa: Precisión en detección de bordes

Aspecto	YOLOv8-Seg	PointRend
Estrategia	Interpolación uniforme	Refinamiento adaptativo
Operaciones en interior	Todas (redundante)	0 (región conocida)
Operaciones en borde	1x	4-6x (iterativo)
Precisión contorno (IoU)	$\sim 87\%$	$\sim 89\%$
Error promedio en borde	2.3 píxeles	0.8 píxeles
Latencia adicional	0 ms	+20-50 ms
Uso de memoria	Bajo	Moderado

### 5.4.2 Cuantificación de Errores en Medición Morfométrica

Asumiendo una configuración de cámara a distancia de 1 metro con resolución de  $1920 \times 1080$  píxeles cubriendo un área física de  $60\text{cm} \times 34\text{cm}$ :

- **Escala de resolución espacial:**  $\sim 0.3 \text{ mm/píxel}$

- **Error sistemático YOLOv8-Seg:**  $2.3 \text{ píxeles} \times 0.3 \text{ mm/píxel} = \pm 0.69 \text{ mm}$
- **Error sistemático PointRend:**  $0.8 \text{ píxeles} \times 0.3 \text{ mm/píxel} = \pm 0.24 \text{ mm}$

**Implicaciones comerciales:** Para un espécimen de salmónido de 3.5 kg con longitud aproximada de 650 mm, un error de medición de 7 mm (1.1% de error relativo) puede resultar en clasificación errónea en categorías comerciales con diferencias de valoración de \$2-3 USD por espécimen. En operaciones con procesamiento de  $10^4$  especímenes anuales, la reducción de error mediante PointRend representa una optimización económica de \$20,000-30,000 USD anuales.

## 5.5 ¿Cuándo Vale la Pena el Costo Computacional?

Table 6: Criterios de decisión: ¿Usar PointRend?

Factor	YOLOv8-Seg suficiente	PointRend recomendado
Tolerancia error	¿5mm aceptable	¿2mm necesario
Geometría objeto	Forma simple, convexa	Aletas, apéndices complejos
Criticidad medición	Estimación general	Clasificación comercial
Hardware	Jetson Nano	Jetson AGX Xavier+
Tiempo real	Necesario (¿15 FPS)	Aceptable (5-10 FPS)

## 6 Síntesis Comparativa y Recomendaciones

### 6.1 Arquitectura Híbrida: Combinando lo Mejor de Cada Enfoque

En la práctica, el sistema óptimo para Salmo Metrics no es elegir exclusivamente una tecnología, sino crear una **arquitectura híbrida** que use cada herramienta para lo que hace mejor.

#### 6.1.1 Pipeline Propuesto Multi-Nivel

##### Nivel 1: Detección inicial rápida (YOLOv8)

- Procesa todos los frames a 20 FPS
- Identifica ROIs (Regions of Interest) donde hay peces
- Filtra frames sin peces o con condiciones no medibles
- *Costo:* Bajo, siempre activo

##### Nivel 2: Medición estándar (YOLOv8-Pose + YOLOv8-Seg)

- Aplica a frames filtrados en Nivel 1
- Calcula longitud y ancho con precisión estándar ( $\pm 5\text{mm}$ )
- Válido para monitoreo general de población
- *Costo:* Bajo, usado en 90% de casos

### Nivel 3: Medición de alta precisión (HRNet-W48 + PointRend)

- Activa solo cuando se detecta: pez solitario, buena iluminación, ángulo óptimo
- Procesa 1 de cada 10 frames aptos
- Precisión objetivo:  $\pm 1-2\text{mm}$
- *Costo*: Alto, usado en 10% de casos selectos

### Nivel 4: Tracking con oclusiones (DeepLabCut)

- Activa cuando se detecta: múltiples peces superpuestos
- Mantiene identidad temporal para tracking longitudinal
- Permite mediciones en condiciones subóptimas
- *Costo*: Muy alto, usado en ¡5% de casos problemáticos

## 6.2 Matriz de Decisión Operacional

## 6.3 Consideraciones de Hardware

### 6.3.1 Requerimientos Computacionales

### 6.3.2 Estimación de Costos

Para una instalación típica en centro acuícola (4 cámaras, procesamiento centralizado):

- **Opción Económica (Solo YOLOv8):**
  - Hardware: 4× Jetson Nano (\$400 USD)
  - Total: **\$1,600 USD**
  - Precisión:  $\pm 5\text{mm}$ , sin tracking, oclusiones limitadas
- **Opción Intermedia (YOLOv8 + HRNet selectivo):**
  - Hardware: 2× Jetson AGX Xavier (\$1,800 USD)
  - Total: **\$3,600 USD**
  - Precisión:  $\pm 2\text{mm}$  en condiciones óptimas,  $\pm 5\text{mm}$  general
- **Opción Premium (Pipeline Híbrido):**
  - Hardware: 1× Jetson AGX Orin 64GB (\$2,000 USD)
  - Total: **\$2,000 USD + servidor**
  - Precisión:  $\pm 1-2\text{mm}$ , tracking robusto, manejo de oclusiones

Table 7: Árbol de decisión para selección de algoritmo

Condición	Algoritmo	Justificación
<b>Escenario 1: Monitoreo General</b>		
- Densidad: ¡3 peces/m <sup>2</sup>	YOLOv8-Pose +	Balance óptimo
- Iluminación: Normal	YOLOv8-Seg	velocidad/precisión
- Objetivo: Estimación		
<b>Escenario 2: Clasificación Comercial</b>		
- Pez aislado	HRNet-W48 +	Máxima precisión
- Iluminación: Buena	PointRend	para decisiones
- Objetivo: Medición exacta		comerciales
<b>Escenario 3: Alta Densidad</b>		
- Densidad: ¡5 peces/m <sup>2</sup>	DeepLabCut	Robustez ante
- Oclusiones: Frecuentes	multi-animal	oclusiones severas
- Objetivo: Tracking individual		
<b>Escenario 4: Bordos Críticos</b>		
- Geometría: Aletas visibles	PointRend	Precisión sub-
- Necesidad: Ancho exacto	(post-proceso)	milimétrica en
- Disponible: GPU potente		contornos

Table 8: Hardware mínimo recomendado por tecnología

Tecnología	GPU Mínima	RAM	FPS Esperado
YOLOv8-Pose/Seg	Jetson Nano 4GB	4 GB	15-20
HRNet-W48	Jetson TX2	8 GB	8-12
DeepLabCut	Jetson AGX Xavier	16 GB	5-8
PointRend	Jetson AGX Xavier	16 GB	8-10
Pipeline Híbrido	Jetson AGX Orin	32 GB	Variable

## 6.4 Conclusiones y Recomendaciones

### 6.4.1 Para Operaciones Acuícolas Comerciales

**Recomendación General:** Implementar arquitectura híbrida en dos fases:  
**Fase 1 (Implementación Inmediata):**

- Base: YOLOv8-Pose + YOLOv8-Seg para monitoreo 24/7
- ROI: Rápida implementación, datos inmediatos de población
- Inversión: \$1,500-2,000 USD por ubicación

**Fase 2 (Expansión Selectiva - 6 meses):**

- Añadir: HRNet-W48 para mediciones pre-cosecha
- Añadir: DeepLabCut si densidad poblacional  $\geq 5$  peces/m<sup>2</sup>
- Añadir: PointRend si clasificación comercial exige  $\pm 1$ mm
- Inversión adicional: \$3,000-4,000 USD

#### 6.4.2 Para Investigación Científica

**Recomendación:** Pipeline completo desde inicio

- Priorizar: DeepLabCut para tracking longitudinal individual
- Incluir: HRNet-W48 para máxima precisión en estudios morfológicos
- Complementar: PointRend para análisis detallado de aletas
- Justificación: Datos precisos son más valiosos que volumen de datos

## 7 Conclusión

La integración de HRNet-W48 (para pose multi-escala), DeepLabCut (para oclusión multi-animal) y PointRend (para precisión de contorno) representa un paradigma complementario a YOLOv8 que prioriza robustez y precisión sobre velocidad extrema.

Ninguna tecnología única es óptima para todos los escenarios. El futuro de los sistemas de medición acuícola radica en arquitecturas inteligentes que seleccionen dinámicamente el algoritmo apropiado según las condiciones específicas de cada frame, balanceando precisión, velocidad y costo computacional según la criticidad de cada medición.

## References

- [1] Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep High-Resolution Representation Learning for Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. <https://doi.org/10.1109/CVPR.2019.00584>
- [2] Lauer, J., Zhou, M., Ye, S., et al. (2022). Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nature Methods*, 19(4), 496–504. <https://doi.org/10.1038/s41592-022-01443-0>
- [3] Kirillov, A., Wu, Y., He, K., & Girshick, R. (2020). PointRend: Image Segmentation as Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. <https://doi.org/10.1109/CVPR.2020.01152>
- [4] Gamero, E., Venegas, G., Ortega, M., Petit, L., & Castillo, A. (2025). *Documento de Análisis y Diseño de Salmo Metrics*. Procesamiento Digital de Imágenes [TEL328].

## A Apéndice: Matriz de Decisión Arquitectónica

Table 9: Recomendaciones por criterio.

Criterio	Recomendación	Justificación
Tiempo real en Jetson	YOLOv8-Seg + YOLOv8-Pose	Único stack viable a 15-20 FPS
Máxima precisión pose	HRNet-W48 + CBAM	+1.28 AP vs. baseline en peces
Robustez oclusión	DeepLabCut multi-animal	PAF-based assembly + ReID
Precisión borde silueta	PointRend 4-6 iteraciones	+2% IoU, crítico para aleta