

## NOISE: Nuclei-Aware Osteoclast Instance Segmentation for Mouse-to-Human Domain Transfer

Sai Kumar Reddy Manne<sup>1,\*</sup> Brendan Martin<sup>1,\*</sup> Tyler Roy<sup>2</sup> Ryan Neilson<sup>2</sup>  
Rebecca Peters<sup>2,3</sup> Meghana Chillara<sup>1</sup> Christine W. Lary<sup>1</sup> Katherine J. Motyl<sup>2,3,4</sup>  
Michael Wan<sup>1,†</sup>

<sup>1</sup>Northeastern University <sup>2</sup>MaineHealth Institute for Research

<sup>3</sup>University of Maine <sup>4</sup>Tufts University School of Medicine

\*equal contribution <sup>†</sup>mi.wan@northeastern.edu

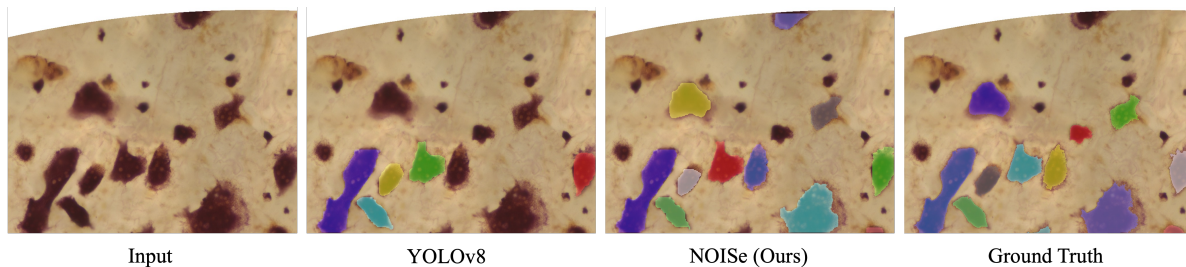


Figure 1. Human osteoclast cell segmentations from our nuclei-aware osteoclast instance segmentation (NOISE) model, compared to a baseline YOLOv8 model. Since both models are trained only on mouse osteoclast data, this highlights the effectiveness of our nuclei-aware pretraining strategy for transfer learning to the human domain, where osteoclast microscope samples are harder to obtain and annotate.

### Abstract

Osteoclast cell image analysis plays a key role in osteoporosis research, but it typically involves extensive manual image processing and hand annotations by a trained expert. In the last few years, a handful of machine learning approaches for osteoclast image analysis have been developed, but none have addressed the full instance segmentation task required to produce the same output as that of the human expert led process. Furthermore, none of the prior, fully automated algorithms have publicly available code, pretrained models, or annotated datasets, inhibiting reproduction and extension of their work. We present a new dataset with  $\sim 2 \times 10^5$  expert annotated mouse osteoclast masks, together with a deep learning instance segmentation method which works for both *in vitro* mouse osteoclast cells on plastic tissue culture plates and human osteoclast cells on bone chips. To our knowledge, this is the first work to automate the full osteoclast instance segmentation task. Our method achieves a performance of 0.82 mAP<sub>0.5</sub> (mean average precision at intersection-over-union threshold of 0.5) in cross validation for mouse osteoclasts. We present a novel **nuclei-aware osteoclast**

**instance segmentation training strategy (NOISE)** based on the unique biology of osteoclasts, to improve the model’s generalizability and boost the mAP<sub>0.5</sub> from 0.60 to 0.82 on human osteoclasts. We publish our annotated mouse osteoclast image dataset, instance segmentation models, and code at [github.com/michaelwwan/noise](https://github.com/michaelwwan/noise) to enable reproducibility and to provide a public tool to accelerate osteoporosis research<sup>1</sup>.

### 1. Introduction

Osteoporosis is a widespread, debilitating bone structure disease, affecting an estimated 200 million people worldwide, especially older women. Discovery of treatments for osteoporosis hinges around advances in basic and translational research in bone biology, and in particular, in the delicate balance between bone-synthesizing osteoblast cells and bone-resorbing *osteoclast* cells, which is disrupted in osteoporosis [5, 23]. Many drugs, including alendronate

<sup>1</sup>We gratefully acknowledge support from the National Institutes of Health under award numbers R01AR076349, R01AR081040, and P20GM121301; the Roux Institute at Northeastern University; and AI + Health, Institute for Experiential AI, Northeastern University.

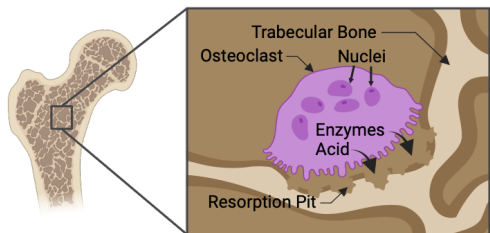


Figure 2. An illustration of an osteoclast cell in purple acting on bone structure in light brown. Osteoclasts are responsible for bone-resorption, and are characterized by having three or more nuclei and being positive for purple TRAP enzyme stain.

and denosumab, target a reduction of bone resorption by modulating the activity of osteoclasts [3, 7]. The potential effectiveness of candidate treatments are gauged by applying them to osteoclast cell cultures and observing their effects on the differentiation and function of those cells, via cellular microscopy. These effects are quantified by a laborious, low-throughput, manual procedure, in which cell shapes are carefully annotated on microscope images by human researchers, with the help of image processing tools that require extensive finetuning and manual adjustment. In addition, high measurement variability necessitates many replicates, further exacerbating the problem. Even a small translational treatment study can require over 1,000 hours of skilled manual labor to analyze.

Fortunately, initial strides have been made in automating osteoclast microscope image characterization with machine learning, with some work targeting osteoclast cell *object detection* and some attempting osteoclast *semantic segmentation* [6, 8, 10, 14, 25]. Critically, though, none of these efforts replicate the full annotation task required by osteoporosis treatment research, namely, that of osteoclast *instance segmentation*, which entails both osteoclast detection and *individual* cell segmentation. Furthermore, prior approaches are variously limited by (i) requiring manual image processing and thereby not being truly automated, (ii) considering the more readily available mouse osteoclast cells and ignoring the human osteoclast cells needed for treatment research, and (iii) requiring specific cellular stains or lab preparations. Lastly, but no less critically, none of these earlier groups addressing object detection or semantic segmentation publish their code, models, or annotated datasets, preventing reproduction and validation of their work, and implementation by osteoporosis researchers.

To address these barriers, we present the first public osteoclast dataset with instance segmentation masks meticulously annotated for  $\sim 2 \times 10^5$  osteoclasts, and a fully-automated algorithm which solves the osteoclast instance segmentation task. Our deep learning method uses a state of the art instance segmentation architecture based on convolu-

tional neural networks, further improved by a nuclei-aware training strategy, injecting domain knowledge from cellular biology into the learning process. We test the model’s generalizability by testing on human osteoclast cells imaged on a bone chip and show a performance improvement of  $mAP_{0.5}$  from 0.60 to 0.82. In summary, our key contributions are as follows.

1. We publish the first public dataset for osteoclast instance segmentation, consisting of microscope images of mouse osteoclast cultures derived from osteoporosis treatment experiments, fully expert annotated with individual osteoclast locations and shape masks.
2. We publish our osteoclast instance segmentation models. To our knowledge, these are the first fully-automatic osteoclast detection or segmentation algorithms made publicly available.
3. We describe a novel **nuclei-aware osteoclast instance segmentation strategy (NOISE)** aimed at improving the generalizability of our instance segmentation model for mouse osteoclasts on plastic tissue culture plates and human osteoclasts on bone chips.

## 2. Background

### 2.1. Bone Biology and Research

Healthy, homeostatic bone remodeling is mediated by a delicate balance between osteoblasts, which are derived from mesenchymal stem cells and responsible for bone formation, and *osteoclasts*, which are multi-nucleated cells derived from the macrophage and monocyte lineages, and are responsible for bone resorption [5, 23]. Osteoclasts are dynamical cells which, through their lifetime, can undergo fusion, fission, recycling, and apoptosis [20]. Osteoclast dynamics and function are disrupted in diseases of the bone such as rheumatoid arthritis and osteoporosis, spurring intense research interest in the life cycle and morphology of osteoclasts. Osteoclasts can be studied via microscopy: tartrate-resistant acid phosphatase (TRAP) enzyme stain highlights certain cells in purple, and these are classified as osteoclasts if they have three or more nuclei, and pre-osteoclasts if they have less than three nuclei. See Fig. 2 for an illustration of an osteoclast cell under TRAP staining.

Traditionally, bone and osteoporosis researchers have carried out osteoclast microscope image analysis with a combination of manual annotation and manually-developed image processing pipelines, using specialized software for biological image analysis such as ImageJ or CellProfiler [4, 20]. This process requires training and experimentation, and manual annotations are unavoidable. Each cell culture well takes 2–3 hours of labor to process, after training. We typically produce at least 3 biological replicates for each manipulation (*e.g.*, genotype or drug treatment), themselves with 5 technical replicates each to account for variability,

Table 1. Prior Algorithmic Approaches to Osteoclast Microscope Image Analysis

Paper	Year	Dataset Size	Fully Automated?	Object Detection?	Semantic Segmentation?	Instance Segmentation?	Human Cells?	Standard Stains?	Data Public?	Code Public?
Emmanuel <i>et al.</i> [10]	2021	Slides from 20 rat tibiae	✗	✓	✗	✗	✗	✓	✗	✗
Cohen-Karlik <i>et al.</i> [6]	2021	$\sim 10^4$ mouse osteoclasts	✓	✓	✗	✗	✗	✓	✗	✗
Wang <i>et al.</i> [25]	2022	$\sim 1.4 \times 10^4$ mouse osteoclasts	✗	✗	✓	✗	✗	✓	✓	✓
Kohtala <i>et al.</i> [14]	2022	$\sim 10^5$ mouse osteoclasts	✓	✓	✗	✗	✓	✗	✗	✗
Davies <i>et al.</i> [8]	2023	$\sim 10^5$ mouse osteoclasts	✗	✗	✓	✗	✗	✓	✗	✗
Ours	2024	$\sim 2 \times 10^5$ mouse, $\sim 4 \times 10^4$ human osteoclasts	✓	✓	✓	✓	✓	✓	✓	✓

resulting in at least 15 wells per experimental group. As a result, even a small translational treatment study can require over 1,000 hours of labor to process.

## 2.2. Osteoclast Image Analysis Algorithms

The need for an automated solution to osteoclast microscope image analysis is reflected by the flurry of recent research activity applying machine learning tools to this task, summarized in Tab. 1. However, while some methods tackle osteoclast detection or semantic segmentation, none of them address the full instance segmentation task that would provide the same detailed output that the manual process does, namely, the individual locations, shapes, and areas of each osteoclast cell. Secondly, and equally critically, no other group has published model code or software that would allow other researchers to use their osteoclast detection or semantic segmentation methods, nor annotated datasets that would allow others to reproduce their results or benchmark detection or segmentation methods. We will address both of these critical shortcomings in our work.

The first machine learning paper for osteoclast analysis, by Emmanuel *et al.* [10], provides a step-by-step guide on how to use a piece of commercial machine learning software to segment osteoclasts *in vivo* from slides featuring thin rat tibiae bone slices. However, many steps involve manual choices in response to intermediate results, so the method is far from automated. Furthermore, the nature of the dataset used and the corresponding train–test split are not specified, and validation is only provided by comparing software and human estimations of a single quantity derived from the segmentations over the entire dataset, so verification of the method’s performance on osteoclast identification is not possible.

Cohen-Karlik *et al.* [6] take a more standard approach, studying *in vitro* slides featuring primary osteoclasts differentiated from bone marrow cells extracted from female mice femurs and tibias and stained with TRAP. They train a convolutional neural network (VGG16) on osteoclast object detection using image patches extracted from 10 culture wells, bolstered by extensive data augmentation, and test on one held-out well, reporting high correlation scores between

human and machine-predicted osteoclast counts and areas. However, high correlation does not rule out systematic errors in the predictions—for instance, they use the area of the estimated box bounding the osteoclast as a proxy for the cell’s area, which even in ideal conditions will yield a consistent overestimate of the true area (*e.g.*, if osteoclasts were circular, by a factor of  $\frac{4}{\pi} \cong 1.3$ ). Standard machine learning metrics such as mean average precision ( $\text{mAP}_t$ ) could help detect and quantify such prediction biases, but the authors do not publish their model or dataset, preventing independent evaluation.

Wang *et al.* [25] develop a two-step method: they first apply a non-learning-based image processing and size filtration step to obtain a set of candidate cell locations and shapes, and then second apply a machine learning classifier to determine which of the candidate cells are osteoclasts. Since the first step requires manual tuning, the algorithm is not fully automated. Furthermore, while the authors do publish part of their dataset and algorithm for public use, the data only supports training and verification of the second step. Specifically, they only provide the coordinates of the approximate centers of osteoclast cells, and only those successfully detected in the first image processing step. Thus, their method does not automate osteoclast detection or instance segmentation, and likewise, their data does not directly support training or testing of osteoclast detection or instance segmentation methods.

Kohtala *et al.* [14] published the first method algorithmic method specifically designed for human osteoclasts. They use 307 wells differentiated from both CD14+ and macrophage cells, which are split into various train–test configurations, and train a YOLOv4-based DarkNet model optimized for object detection. Detection results are reported using  $\text{mAP}_t$  and performance reaches  $\text{mAP}_{0.1} = 0.85$  and  $\text{mAP}_{0.5} = 0.76$ ; overall counts show high concordance with human annotations. This work offers an approach in line with standard machine learning practices, yielding interpretable, robust, and comparable metrics and reproducible methods. However, unique among algorithmic approaches, this method uses cell samples obtained from a fluorescent nuclear staining process, which is intended aid

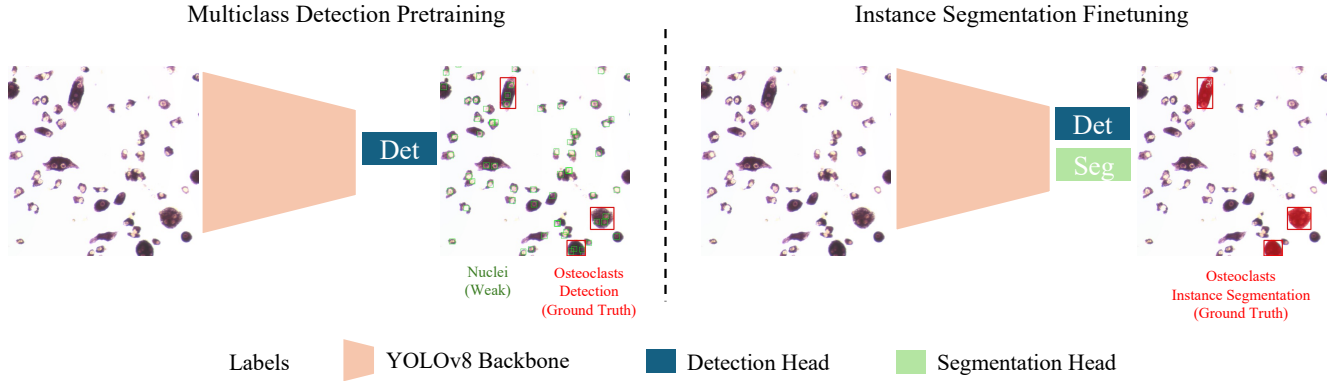


Figure 3. An overview of NOISe, our nuclei-aware osteoclast instance segmentation training pipeline. A two-stage training process features a pretraining stage for multiclass detection weakly supervised by nuclei location information. The pretraining significantly boosts subsequent performance of the overall osteoclast instance segmentation model, especially in the data-scarce human domain.

manual image analysis and therefore likely makes the object detection task easier. As such, there is no reason to assume that the method will work with cell samples obtained using the standard TRAP stain, but we cannot verify this since the code and data are not made public.

The most recent method, by Davies *et al.* [8], is largely interested in analyzing changes to osteoclasts in response to treatments, using results obtained from standardized algorithms. The latter consists of an open-sourced deep learning image segmentation tool, ilastik, followed by semi-manual processing with another open-sourced tool, ImageJ, to obtain desired osteoclast characteristics. Their evaluation metrics are fairly unique, in this case relying on correlation scores on aggregate metrics like total osteoclast cell count and area. This again precludes reproducibility and robust comparison with other methods. Data and models are not provided publicly.

### 3. Method

#### 3.1. End-to-End Instance Segmentation

Instance segmentation identifies unique instances of objects in an image and predicts their segmentation masks. It can be thought of as the combination of object detection, which locates and distinguishes instances of objects of interest with minimal bounding boxes, and semantic segmentation, which classifies each pixel or region by object type (or as belonging to the background). Sometimes, the methods for accomplishing both tasks can overlap, while in other contexts, the two can be addressed separately, as is often the case in cytological and histological imaging. As detailed in Sec. 2, all previous approaches to algorithmic osteoclast image analysis have attempted either object detection, or semantic segmentation, or their own unique task, but none have directly tackled instance segmentation.

The core of our osteoclast instance segmentation is the YOLOv8 model [13], which follows a long line of You Only Look Once (YOLO) object detection models [21]. YOLO models are based on the idea of directly regressing bounding boxes in an end-to-end manner, supplanting earlier multi-stage methods [11]. YOLOv8 predicts a segmentation mask, a bounding box, a classification label, and an “objectness” score for each instance. The model architecture employs a version of the CSPDarknet53 backbone first introduced in YOLOv4 [1] and output heads for bounding box regression, class prediction, and two heads for segmentation [24]. The segmentation heads are similar to the You Only Look At Coefficients (YOLACT) model [2]: one head predicts 32 mask prototypes, and the other head predicts 32 mask coefficients. The final instance segmentation predictions come from a linear combination of the prototypes weighted by the coefficients, followed by score thresholding and non-maximum suppression (NMS).

YOLOv8 is trained with a combination of four loss functions: (i) a classification loss  $L_{cls}$ , which is a simple binary cross-entropy (BCE) calculated separately for each class; (ii) a complete intersection over union (CIoU) [27] loss  $L_{box}$ , which compares the central points and aspect ratios of the two boxes along with IoU for the bounding box regression; (iii) a distribution focal loss (DFL) [16]  $L_{dff}$ , which distinguishes itself from classic bounding box regression by targeting a distribution of correct bounding boxes as opposed to a single correct target; and (iv) a segmentation loss  $L_{seg}$ , which is a pixel-wise BCE similar to that of YOLACT. The final loss function is a weighted combination of the four components,

$$L = \lambda_{box}L_{box} + \lambda_{cls}L_{cls} + \lambda_{dff}L_{dff} + \lambda_{seg}L_{seg}, \quad (1)$$

with corresponding weights  $\lambda_*$  for each component  $L_*$ .

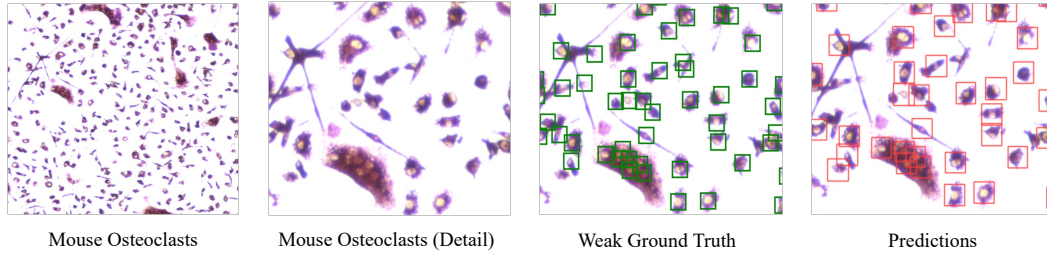


Figure 4. Detail from a mouse osteoclast microscope image, together with weak ground truth nuclei labels (note the uniform size and shape of the boxes), and YOLOv8 object detection predictions of the same. Our nuclei-aware training method exploits this weak ground truth information to improve the generalizability of osteoclast instance segmentation.

### 3.2. Nuclei-Aware Osteoclast Instance Segmentation (NOISe)

Most osteoclast image analysis algorithms address the domain of *in vitro* mouse (or generally murine) osteoclast cells in clear plastic plates, due to availability and controllability of mouse subjects for research. However, translational osteoporosis researchers must also analyze the less commonly available human osteoclast cells, which in the case of our dataset are hosted and imaged on thin translucent bone chips to study the effects of treatments on osteoclast bone resorption. In our experimentalists’ experience, human osteoclasts have more complex shapes and take longer to annotate manually. Since we have enough mouse osteoclast data to support strong mouse osteoclast instance segmentation, we explore the possibility of improving human domain performance using only a small amount of human osteoclast data for training and testing. The one prior machine learning osteoclast image analysis work that considers human osteoclasts (on transparent plates) [14] uses fluorescent Hoechst staining to highlight the nuclei. Given the importance of nuclei in the definition and identification of osteoclast cells, this fluorescent staining presumably makes osteoclast detection easier, and at the same time, less applicable to the more common plain light setting. Here, we describe a machine learning alternative to fluorescent staining for highlighting nuclei.

Our nuclei-aware osteoclast instance segmentation (NOISe) pretraining–finetuning strategy, shown in Fig. 3, leverages weak or pseudo nuclei information for efficient transfer learning. First, we curate a small, informal nuclei detection dataset as described in Sec. 4.2. We train a YOLOv8 object detection model on the curated nuclei dataset, which achieves an  $mAP_{0.5}$  of 0.85, and demonstrates highly precise nuclei detection, as shown in Fig. 4. We next apply this detector to our entire mouse osteoclast dataset, providing a set of weak ground truth nuclei bounding box labels, on top of the existing ground truth osteoclast instance segmentation labels. Then, in order to inject the nuclei information in our osteoclast segmentation model, we pretrain it on a multiclass object *detection* task

with both osteoclast labels and weak nuclei labels. This encourages the model to learn priors about nuclei and their relationship with osteoclasts. Finally, we finetune the pre-trained detection model for instance segmentation task with osteoclast segmentation labels. Since the nuclei bounding boxes can only be considered as weak supervision signals, their use must be justified by improvements on downstream tasks that are validated with accurate ground truth instance masks. We show experimentally, in Sec. 5.3, that the use of the nuclei in the NOISe method does indeed lead to segmentation improvements over a baseline YOLOv8 model.

## 4. Datasets

### 4.1. Expert-Annotated Human and Mouse Osteoclast Instance Segmentation Dataset

Microscope images with osteoclast cells were captured from two types of experiments, those with mouse osteoclasts plated on 96-well plastic plates to measure differentiation of osteoclasts in response to treatments, and those with human osteoclasts plated on bone chips (devitalized bovine bone slices within a 96-well plastic plate) to measure bone resorption and differentiation in a more realistic setting. After one week of differentiation for mouse cells, and two weeks for human cells, wells were fixed and treated with tartrate-resistant acid phosphatase (TRAP). Treated and control plates were photographed under the microscope at approximately  $10^5 \times 10^5$  resolution (10 billion pixels, or 10 gigapixels). Individual osteoclast cell instance masks were annotated by human researchers with at least two years of experience in osteoclast culturing, imaging, and osteoclast image analysis, after passing an osteoclast cell count validation test. Using ImageJ, a purple color filter was applied and selected objects with areas of at least 2,000 pixels (the rough mean size of a 3-nuclei osteoclast) were marked as osteoclast candidates. Researchers then manually analyzed the images to remove non-osteoclasts and add manual masks for osteoclasts that were missed. This process yielded 150 well images with  $\sim 2 \times 10^5$  mouse osteoclast masks and 24 slide images with  $\sim 4 \times 10^4$  mouse

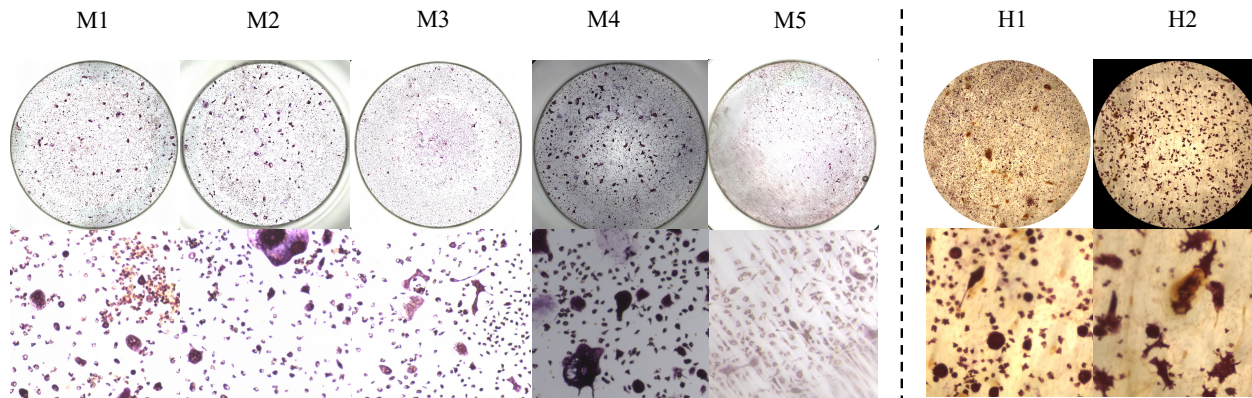


Figure 5. Cell images and cropped patches from mouse experiments M1 through M5, and human experiments H1 and H2, illustrating the diversity in slide lighting, background, and appearance, and in osteoclast size, shape, and density in our data.

human osteoclasts masks.

Mouse osteoclast images were split into five batches, M1, M2, M3, M4, and M5; and the human osteoclast images into two batches, H1 and H2; based on the experimental procedures, each consisting of tens of images. See Fig. 5 for samples. Since each batch potentially shares qualitative characteristics, we used these divisions for training and testing splits (*e.g.*, we report mouse results under 5-fold cross-validation in Sec. 5) and in our published dataset. However, since the scientific assays are ongoing, we did not indicate which batch represents which experimental procedure, and we did not conduct our computational experiments or analyses based on those distinctions.

## 4.2. Osteoclast-Specific Nuclei Dataset

We experimented with various nuclei instance segmentation datasets [12, 15, 19, 22, 26] in our osteoclast instance segmentation pretraining paradigm, but found them to be ineffective due to several notable differences between the biology and imaging of those nuclei and the nuclei in our osteoclast image datasets. Existing nuclei datasets use hematoxylin and eosin (H&E) staining to improve the contrast of cell visibility under a microscope—it stains nuclei in a dark color, against a lighter cytoplasm, easing nuclei identification. By contrast, our TRAP-stained images show cells with light colored nuclei on a darker cell that is set against a light background. Consequently, nuclei detection methods developed for existing H&E stained datasets did not perform well on our osteoclast images, prompting us to develop our own osteoclast-specific nuclei dataset.

Nuclei far outnumber osteoclasts and annotating each with segmentation masks would be infeasible for our purposes. Instead, we randomly selected 20 images from each of our five mouse experiments to curate a smaller nuclei dataset for manual annotations. In each image, we marked the approximate center-of-mass of each nucleus’s shape

with a point annotation tool VGG Image Annotator [9]. These annotations were converted to bounding boxes with width and height uniformly set to 36 pixels, chosen empirically to cover larger nuclei in the dataset. A total of 20,784 nuclei were marked in this way, with an average of 200 annotations per image, illustrating the density of the information contained in the small dataset. By comparison, the well-known MoNuSeg dataset [15] has a similar count of 21,000 instances, albeit with full instance segmentation masks.

## 5. Experiments & Results

### 5.1. Experimental Implementation

Throughout, we used the large variant in the YOLOv8 family of models, YOLOv8l-seg, which has 45.9 M parameters in the backbone, detection, segmentation heads collectively. Each model was trained for 100 epochs, stopping early when validation split performance did not improve. Since the whole slide images are of gigapixel resolutions, following [14], we extracted smaller patches of size  $832 \times 832$  with 50% overlap for our experiments. Images were down-sampled to  $416 \times 416$  for training and inference during pre-processing to speed-up inference, and outputs were scaled back to full resolution in post-processing. Other parameters were set to their default values [13]. Performance was evaluated using standard precision, recall, and mean average precision ( $mAP_t$  at various thresholds  $t$ ) metrics from the Common Objects in Context (COCO) object detection dataset [17, 18].

We recall from our discussion in Sec. 2 that direct performance comparisons with earlier osteoclast image algorithms are difficult because groups addressing standard tasks like detection and semantic segmentation have not published their data and code. We hope our work can establish a public framework and benchmark for future machine

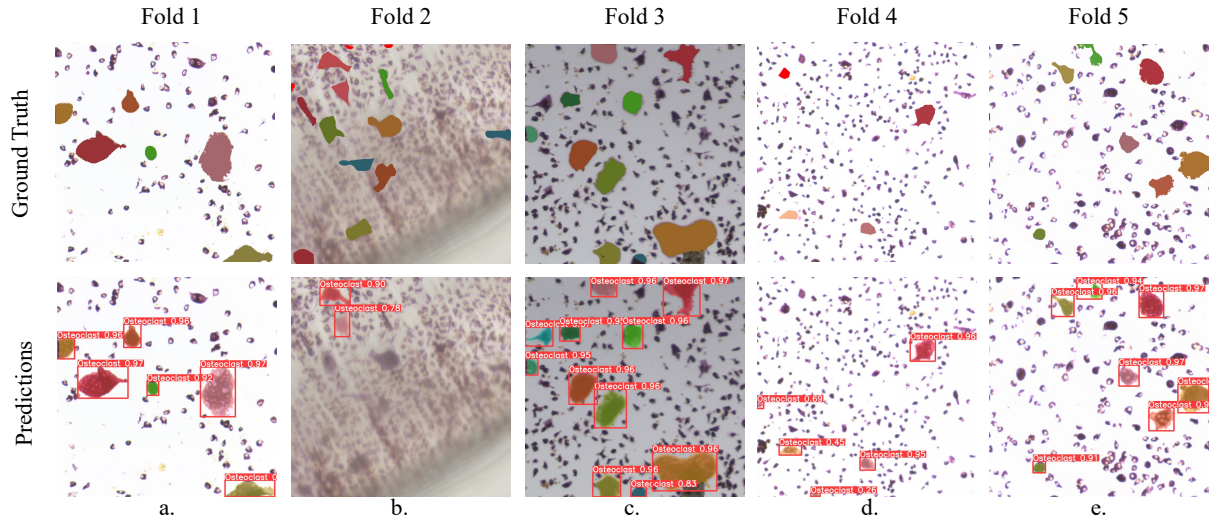


Figure 6. Predictions of our mouse osteoclast instance segmentation models under five-fold cross-validation. Ground truth osteoclast shape masks are carefully produced by experts. Our model’s predictions of these masks are generally faithful, but we highlight some mistakes and challenges. In (b), image blur seems to impair the model’s overall ability; in (c), a pre-osteoclast on the top left is mistakenly identified as an osteoclast; and in (d), an osteoclast on the top left is not detected, while two pre-osteoclasts below are mis-identified as osteoclasts.

Table 2. Mouse Osteoclast Instance Segmentation Model Cross-Validation (YOLOv8 M→M)

Model	IoU $t$	0.10			0.25			0.50			0.75			0.90		
		P	R	mAP	P	R	mAP	P	R	mAP	P	R	mAP	P	R	mAP
Fold 1	M1	0.871	0.838	0.906	0.868	0.835	0.902	0.869	0.809	0.885	0.816	0.714	0.784	0.609	0.434	0.437
Fold 2	M5	0.884	0.524	0.639	0.882	0.518	0.630	0.851	0.487	0.583	0.555	0.297	0.299	0.063	0.023	0.010
Fold 3	M4	0.894	0.850	0.911	0.892	0.845	0.906	0.884	0.818	0.881	0.831	0.722	0.779	0.569	0.436	0.407
Fold 4	M3	0.862	0.826	0.888	0.858	0.820	0.882	0.851	0.787	0.851	0.805	0.677	0.735	0.537	0.392	0.353
Fold 5	M2	0.908	0.849	0.921	0.907	0.845	0.917	0.896	0.825	0.898	0.852	0.736	0.812	0.654	0.494	0.512
Average		0.884	0.777	<b>0.853</b>	0.881	0.773	<b>0.847</b>	0.870	0.745	<b>0.820</b>	0.772	0.629	<b>0.682</b>	0.486	0.356	<b>0.344</b>
Standard Deviation		0.018	0.142	0.120	0.019	0.143	0.122	0.020	0.145	0.133	0.122	0.187	0.216	0.241	0.190	0.195

learning work on osteoclast image analysis.

## 5.2. Mouse Osteoclast Instance Segmentation

We tested the model on the mouse datasets via five-fold cross-validation. For each fold, one dataset was set aside for testing, and of the remaining training data, a random 20% was reserved for validation. Cross-validation result are reported in Tab. 2. At the default threshold,  $t = 0.5$ , the model maintained an mAP score above 80% on all test sets except M5. Performance is weaker under the most severe threshold,  $t = 0.9$ , but is strong overall. Illustrated examples of predictions on each cross-validation fold can be found in Fig. 6. In general, the model does an excellent job of predicting the masks of identified objects, but it can sometimes struggle to distinguish between osteoclasts and pre-osteoclasts (which are similar to osteoclasts but have fewer than three nuclei), or to detect osteoclasts when the image is blurry. The weak performance on the M5 test set in particular could be caused by blurring artifacts present in some images, such as the one shown in Fig. 6.

## 5.3. Transfer Learning for Human Osteoclasts

In Tab. 3, we compare the instance segmentation performance of our NOISE model against the baseline YOLOv8 model, under various mouse–human and train–test configurations to quantify both models’ generalizability. This also serves as an ablation study for the NOISE method. The configurations are indicated by a MODEL TRAIN→TEST naming scheme which we expand on below.

In our notation, the baseline configuration employed in Tab. 2 would be denoted YOLOv8 M→M, with the mouse-to-mouse domain indicator M→M implying a cross-validation paradigm. The YOLOv8 M→H configuration denotes a baseline YOLOv8 model trained on the entire mouse data and tested on all human data. The performance drop in YOLOv8 M→H, compared to YOLOv8 M→M, shows that YOLOv8 trained on mouse data does not generalize well to the human domain. By contrast, performance is much stronger under the NOISE M→H configuration, where YOLOv8 is trained with all of the mouse data under the NOISE strategy and tested on all of the human data. This demonstrates effectiveness of NOISE in improv-

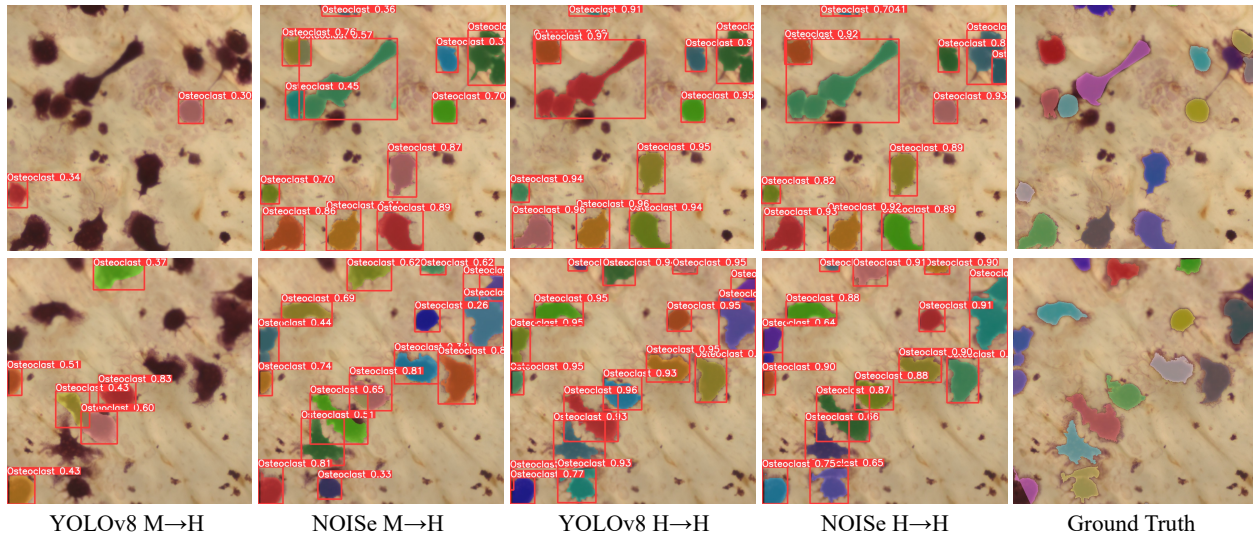


Figure 7. Qualitative results from different training configurations of YOLOv8 and NOISE models tested on human osteoclast data, showing that the NOISE strategy can improve performance on human osteoclasts even when the model is trained only on mouse data (NOISE M→H).

Table 3. Human Osteoclast Instance Segmentation Performance under Various Train–Test Configurations

IoU <i>t</i>	0.10			0.25			0.50			0.75			0.90		
	P	R	mAP	P	R	mAP	P	R	mAP	P	R	mAP	P	R	mAP
YOLOv8 M→H	0.839	0.605	0.777	0.826	0.595	0.755	0.695	0.666	0.595	0.425	0.309	0.285	0.084	0.098	0.046
NOISE M→H	0.868	0.783	<b>0.878</b>	0.861	0.775	<b>0.865</b>	0.812	0.729	<b>0.796</b>	0.677	0.586	<b>0.592</b>	0.324	0.233	<b>0.151</b>
YOLOv8 H→H	0.849	0.826	0.884	0.843	0.816	0.872	0.799	0.766	0.805	0.702	0.609	0.626	0.506	0.339	<b>0.296</b>
NOISE H→H	0.860	0.810	<b>0.891</b>	0.857	0.800	<b>0.880</b>	0.807	0.757	<b>0.820</b>	0.722	0.613	<b>0.650</b>	0.478	0.320	0.265

ing generalization to unseen human data.

The bottom half of Tab. 3 shows results from both YOLOv8 and the NOISE model under the H→H configuration of two-fold cross-validation on the human data (with performance on both folds averaged). Contrasting the top and bottom halves of Tab. 3 shows that training on some human data improves test performance on human data, as expected. But more interestingly, focusing on the bottom half of Tab. 3, we see that when some human data is available for training, the NOISE strategy can still offer further, small gains—yielding the strongest overall performance on human models—but these gains are nowhere as dramatic as those provided by NOISE in M→H setting.

We show some qualitative results in Fig. 7. These illustrate the improved osteoclast detection ability (recall) of NOISE M→H over YOLOv8 M→H, and closer inspection shows that when predicting correct instances that both agree on, NOISE does so with higher confidence. The YOLOv8 H→H and NOISE H→H predictions are more comparable to each other. They are also comparable to those of NOISE M→H, which is again remarkable, since the latter has never seen any human osteoclast data (nor human nuclei data). These numerical and qualitative results show that the NOISE strategy significantly improves mouse-to-human domain transfer, and suggests that it might generalize effectively to other, new osteoclast image domains. Thus,

we recommend NOISE M→H and NOISE H→H as starting points in such settings.

## 6. Conclusion

We have presented the first osteoclast instance segmentation dataset and algorithm, fully automating the laborious cellular image analysis task performed by osteoporosis researchers. We introduced a novel nuclei-aware pre-training strategy which significantly improves the performance of models trained only on mouse osteoclast data, when tested on human osteoclast images, which are harder to obtain and annotate. To support further algorithm development and accelerate osteoporosis research, we publish our annotated mouse osteoclast dataset and the code and pretrained weights for our NOISE model, which appears to exhibit strong generalizability and is recommended for both familiar and new osteoclast image settings. Future work could explore gains in performance and generalizability with more robust nuclei supervision, or by incorporating methods like cell stain style transfer to address variability in lab preparation or image appearance. Continual “human-in-the-loop” learning could enable more efficient gains in segmentation performance, and improve domain transferability on a larger scale.



## References

- [1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020. 4
- [2] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact++ better real-time instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):1108–1121, 2022. 4
- [3] H. G. Bone, D. Hosking, J. P. Devogelaer, J. R. Tucci, R. D. Emkey, R. P. Tonino, J. A. Rodriguez-Portales, R. W. Downs, J. Gupta, A. C. Santora, U. A. Liberman, and I. I. I. Osteoporosis Treatment Study Group Alendronate Phase. Ten years' experience with alendronate for osteoporosis in postmenopausal women. *N Engl J Med*, 350(12):1189–99, 2004. Type: Journal Article. 2
- [4] Mikkel Bo Brent and Thomas Emmanuel. Contemporary Advances in Computer-Assisted Bone Histomorphometry and Identification of Bone Cells in Culture. *Calcified Tissue International*, 112(1):1–12, 2022. 2
- [5] E. R. Burns, J. A. Stevens, and R. Lee. The direct costs of fatal and non-fatal falls among older adults - United States. *J Safety Res*, 58:99–103, 2016. Type: Journal Article. 1, 2
- [6] Edo Cohen-Karlik, Zamzam Awida, Ayelet Bergman, Shahar Eshed, Omer Nestor, Michelle Kadashev, Sapir Ben Yosef, Hussam Saed, Yishay Mansour, Amir Globerson, Drorit Neumann, and Yankel Gabet. Quantification of Osteoclasts in Culture, Powered by Machine Learning. *Frontiers in Cell and Developmental Biology*, 9:674710, 2021. Type: Journal Article. 2, 3
- [7] S. R. Cummings, J. San Martin, M. R. McClung, E. S. Siris, R. Eastell, I. R. Reid, P. Delmas, H. B. Zoog, M. Austin, A. Wang, S. Kutilek, S. Adami, J. Zanchetta, C. Libanati, S. Siddhanti, C. Christiansen, and Freedom Trial. Denosumab for prevention of fractures in postmenopausal women with osteoporosis. *N Engl J Med*, 361(8):756–65, 2009. Type: Journal Article. 2
- [8] Bethan K. Davies, Andrew P. Hibbert, Scott J. Roberts, Helen C. Roberts, Jennifer C. Tickner, Gill Holdsworth, Timothy R. Arnett, and Isabel R. Orriss. A Machine Learning-Based Image Segmentation Method to Quantify In Vitro Osteoclast Culture Endpoints. *Calcified Tissue International*, 2023. Type: Journal Article. 2, 3, 4
- [9] Abhishek Dutta and Andrew Zisserman. The via annotation software for images, audio and video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2276–2279, 2019. 6
- [10] Thomas Emmanuel, Annemarie Br  el, Jesper Skovhus Thomsen, Torben Steiniche, and Mikkel Bo Brent. Artificial intelligence-assisted identification and quantification of osteoclasts. *MethodsX*, 8:101272, 2021. Type: Journal Article. 2, 3
- [11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 4
- [12] Simon Graham, Mostafa Jahanifar, Ayesha Azam, Mohammed Nimir, Yee-Wah Tsang, Katherine Dodd, Emily Hero, Harvir Sahota, Atisha Tank, Ksenija Benes, et al. Lizard: a large-scale dataset for colonic nuclear instance segmentation and classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 684–693, 2021. 6
- [13] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, 2023. 4, 6
- [14] Sampsa Kohtala, Tonje Marie Vikene Nedal, Carlo Kriesi, Siv Helen Moen, Qianli Ma, Kristin Sirnes   degaard, Therese Standal, and Martin Steinert. Automated Quantification of Human Osteoclasts Using Object Detection. *Frontiers in Cell and Developmental Biology*, 10:941542, 2022. Type: Journal Article. 2, 3, 5, 6
- [15] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7):1550–1560, 2017. 6
- [16] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 4
- [17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll  r, and C. Lawrence Zitnick. Coco dataset, detection evaluation, 2014. 6
- [18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll  r, and C. Lawrence Zitnick. COCO API, 2020. 6
- [19] Amirreza Mahbod, Christine Polak, Katharina Feldmann, Rumsha Khan, Katharina Gelles, Georg Dorffner, Ramona Woitek, Sepideh Hatamikia, and Isabella Ellinger. Nuinsseg: a fully annotated dataset for nuclei instance segmentation in h&e-stained histological images. *Scientific Data*, 11(1):295, 2024. 6
- [20] M. M. McDonald, A. S. Kim, B. S. Mulholland, and M. Rauner. New Insights Into Osteoclast Biology. *JBMR Plus*, 5(9):e10539, 2021. Type: Journal Article. 2
- [21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 4
- [22] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, 35(5):1196–1206, 2016. 6
- [23] D. H. Solomon, S. S. Johnston, N. N. Boytsov, D. McMorro, J. M. Lane, and K. D. Krohn. Osteoporosis medication use after hip fracture in U.S. patients between 2002 and 2011. *J Bone Miner Res*, 29(9):1929–37, 2014. Type: Journal Article. 1, 2
- [24] Juan Terven, Diana-Margarita C  rdova-Esparza, and Julio-Alejandro Romero-Gonz  lez. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine Learning and Knowledge Extraction*, 5(4):1680–1716, 2023. 4

- [25] Xiao Wang, Mizuho Kittaka, Yilin He, Yiwei Zhang, Yasuyoshi Ueki, and Daisuke Kihara. OC\_finder: Osteoclast Segmentation, Counting, and Classification Using Watershed and Deep Learning. *Frontiers in Bioinformatics*, 2: 819570, 2022. Type: Journal Article. [2](#), [3](#)
- [26] Stephan Wienert, Daniel Heim, Kai Saeger, Albrecht Stenzinger, Michael Beil, Peter Hufnagl, Manfred Dietel, Carsten Denkert, and Frederick Klauschen. Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach. *Scientific reports*, 2(1):503, 2012. [6](#)
- [27] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12993–13000, 2020. [4](#)