 DOI: 10.15203/99106-150-2-33

Solving Microorganism Enumeration through Weakly-Supervised Counting with Vision Transformers - A comparative study

Javier Ureña Santiago
Intelligent and Interactive Systems
University of Innsbruck
Innsbruck, Austria
javier.urena@uibk.ac.at

Antonio Rodriguez-Sanchez
Intelligent and Interactive Systems
University of Innsbruck
Innsbruck, Austria
antonio.rodriguez-sanchez@uibk.ac.at

Abstract

Microorganism enumeration, critical for evaluating contamination levels and ensuring safety in various fields, traditionally relies on labor-intensive methods that usually involve manual counting. This has been tackled with computer vision and machine learning methods to automate this process commonly through techniques like instance segmentation or density estimation. But these techniques rely on data annotations that can be omissible for the task, or may be difficult to obtain. This makes weakly-supervised counting an end-to-end, task-oriented approach to solve the microorganism enumeration problem, by omitting the use of any spatial data rather than the global count. This study explores the integration of weakly-supervised counting with Vision Transformers (ViTs), aiming to enhance microorganism counting efficiency by focusing on aggregate counts without needing spatial details. We compared ViTs against traditional models like ResNet and analyzed ViT-based models such as TransCrowd and other ViT back-boned architectures, for their accuracy in microorganism enumeration across three microbiological datasets. The findings reveal that ViTs not only compete well enough with traditional methods but also open research lines on their exploration in weakly-supervised counting of microbial imagery.

1 Introduction

There are numerous tasks where the counting of microorganisms is necessary, such as quality control in the pharmaceutical [24], food industry [22] or environmental monitoring [13]. Traditional approaches to enumerating microbes involves methods like manual counting on agar plates, that is subjective, error-prone, tedious, and labor-intensive [1]. As result, image analysis methods for efficient microbial enumeration appear and increasingly improve in efficiency thanks to machine learning. Most of these methods depend on the use of mask estimates [11, 20] or density maps regression [9, 10], both dependent of ground-truth annotations regarding spatial information in the image regarding location and/or size of the instances. This kind of data is very present in a variety of datasets oriented for instance counting [28, 12, 23] which has led to the resolution of this problem through the use of novel deep learning methodologies.

Conversely, estimating the density map or applying instance segmentation is not always the best approach for every use case. An example would be the use of microbial control (swab tests) for bacterial analysis that are able to count the number of colonies and estimate a number of bacteria for surface cleanliness assertion [3]. Positional information is trivial, and an overall count of the bacteria is obtained, which allows estimation of the level of infection of the surface.

The First Austrian Symposium on AI, Robotics, and Vision (AIROV24).

This points to the need of generating spatial-aware datasets that can be a complex and time-consuming task compared to the use of simple count annotations. Thus, a direct end-to-end regression counting method on images to predict the overall number of instances can be an efficient alternative to deal with use cases where spatial information is redundant. This approach is called weakly-supervised counting and the use of CNNs is prominent as a solution to microbial enumeration [27, 4].

However, CNNs face difficulties in modelling the global context and interactions between image patches due to the limitations of their reception fields, hence, their performance in weakly-supervised counting applications may be limited. In contrast, Vision Transformers (ViT) [5] emerge as a potential solution. The inherent self-attention characteristics of ViTs model the global context of the image effectively and capture long context dependency thanks to their wide reception field, making them some of the best tools in tasks like image classification or segmentation [16].

ViTs competence to achieve weakly-supervised counting of instances [18, 7] or, in most cases, people in crowds [25, 19, 17] has shown great results, claiming to be an approach that surpass in the State-Of-The-Art benchmark imposed by the CNNs and competing with more precise solutions like instance-segmentation or density estimation solutions. Yet, the use of ViTs in weakly-supervised counting in the paradigm of microorganism enumeration remains unexplored.

This study aims to highlight the effectiveness of Vision Transformers (ViTs) in weakly-supervised microorganism counting, presenting them as a promising solution compared to the current present approaches. We conducted a thorough analysis of State-Of-The-Art ViT-based models, including TransCrowd and our custom ViT back-boned regression architectures, comparing them with traditional ResNet models. Our focus was on identifying an efficient approach to microorganism enumeration that bypasses the need for spatial data, directly predicting aggregate counts. The findings demonstrate ViTs' competence in regression counting tasks for weakly-supervised counting in the paradigm of microbial enumeration.

2 Methodology

Architectures We compare three architectural categories for weakly supervised counting: ViT-based regression architectures employing various ViT iterations for feature extraction, leading-edge models with strong performance in weakly supervised counting, and conventional deep learning computer vision architectures. TransCrowd [17], a top-performing model originally for crowd counting, serves as our benchmark for integrating ViTs in weakly supervised counting scenarios. In the second category, we explore ViT backbones, including the original Vision Transformer [5] and a modified variant, *ViT**, which uses Euclidean distance instead of dot product for exploratory model performance analysis. The third category examines the Residual Network (ResNet) [8], a staple in computer vision tasks, through an ablation study to evaluate its applicability to microorganism counting, drawing on its previous application in cancer cell enumeration [14]. This structured comparison aims to identify optimal strategies for enhancing the accuracy and efficiency of microorganism counting in a weakly supervised context.

Datasets Three different datasets were used to train the models from scratch. The datasets represent different scenarios where bacteria or cell enumeration can be achieved. The datasets have different characteristics: number of images, application of data augmentation, variability between images. Dataset 1: yellow fluorescent cells Dataset [21] consists of 10k training images with counts per image ranging from 0 to 20. Dataset 2: blue fluorescent cells dataset from "Learning To Count Objects" [15] is composed of 19.3k training images with quantities per image ranging from 1 to 317 cells. And dataset 3: an artificial dataset of fluorescent *bacillus subtilis* counting with 12k training images, with quantities ranging from 0 to 1900 bacteria per image. The models are evaluated under the same conditions, studying the ability to predict the total number of microorganisms and the ability to analyze the images in different quantities scenarios, as seen in Figure 1. The metrics used to compare the architectures are Mean Absolute Error and Root Mean Squared Error. All models were trained on an NVIDIA RTX 3090 with different batch sizes depending on the computational load of each architecture.

Table 1: Quantitative results of different architectures performing weakly supervised counting on three microorganism datasets. TransCrowd-G and TransCrowd-T refer to the two variants of the TransCrowd architecture, Gap and Token, respectively. ViT refers to the traditional Vision Transformer architecture, while ViT* refers to the Vision Transformer using Euclidean distance instead of point operation. The average rank shows the performance ranking of the architectures on the three datasets.

	Fluorescent Yellow Cells		LTCO Blue Cells		Fluorescent Artificial Bacteria		
Architectures	MAE	RMSE	MAE	RMSE	MAE	RMSE	Average Rank
ResNet34	0,46	0,78	**2,64**	**5,18**	19,75	26,93	1º
ResNet50	0,51	0,81	2,79	5,49	42,92	50,48	5º
ResNet101	0,51	0,82	3,11	6,02	20,24	27,81	3º
TransCrowd-G	0,76	1,131	20,97	33,89	38,98	65,34	7º
TransCrowd-T	**0,38**	**0,67**	4,23	7,67	22,01	30,18	2º
ViT	0,58	0,96	6,04	10,59	15,09	21,77	6º
ViT*	0,57	0,91	5,21	9,25	**14,75**	**20,897**	4º

3 Results

The results show us that each approach can achieve different performances depending on the dataset being evaluated. As seen in the table 1, the benchmark architecture for weakly supervised counting used for crowd counting TransCrowd ("token" variant) outperforms the rest in dataset 1 (yellow fluorescent cells). In dataset 2 (blue fluorescent cells), the simpler ResNet34 is able to predict very similarly to its more complex variant ResNet50, but both significantly outperform the rest of the architectures. Finally, in dataset 3 (artificial bacteria), the modified Euclidean distance ViT* as backbone outperforms the rest, even the original configuration of ViT. An explanation of this differentiation in between evaluation for each model can be explained by the differences of the datasets used, as for some datasets the amount of data is more scarce, or the use of data augmentation has not been done to all of them. For example, ResNets can achieve better convergence with smaller datasets and generalize better between a dataset with many differentiations between its images (property of the second dataset). TransCrowd might even be better at discriminating between the yellow cells of dataset 1, as it is likely to be better at generalizing between larger, more salient regions of an image. None of these characteristics are observed in Data Set 3, where the best performance is achieved by the ViT* backbone regression architecture, which could explain why ViT is actually able to better capture the global context of an image in a more homogeneous scenario (as opposed to Data Sets 1 and 2).

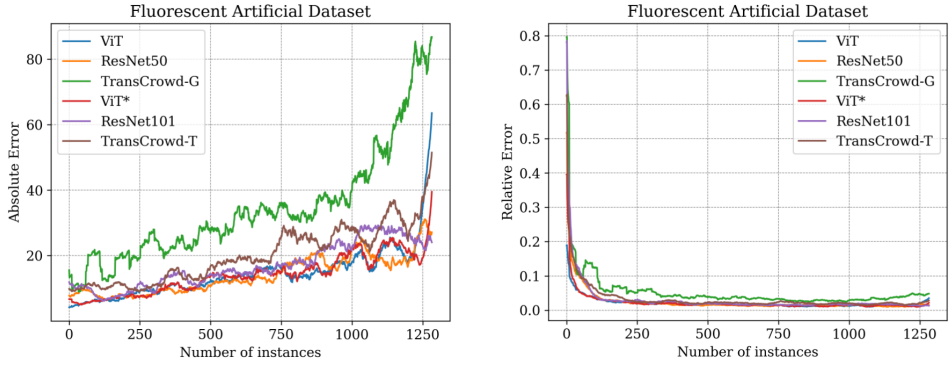

Figure 1: Results evaluating amount of cells per image with the fluorescent artificial dataset. Left: Absolute error. Right: Relative Error. Here we can evaluate how the models behave not only predicting an overall estimate of the number of bacteria but also analyze the capability of the model to capture visual features in all counting scenarios.

In figure 1 its represented as qualitative results the distribution of absolute error and relative error in the range of instances of the dataset 3. This visual analysis helps to evaluate the competence of the

architectures in terms of understanding the general characteristics of the images as more instances appear in them. In the case of dataset 3, ViT* is the one that has the lowest MAE overall, but its ResNet34 is the one that actually maintains a more consistent slope, meaning that the MAE actually doesn't vary much along the range of bacteria per image.

These results are based on end-to-end training of the architectures, without the use of pre-trained weights. The training for each architecture varied depending on the computational needs of each architecture and dataset, affecting training hyper-parameters such as batch size, epochs, or learning rate. They show that training from scratch yields quite similar results, and that further fine-tuning experiments should continue.

4 Discussion

The use of ViTs is common in tasks where the self-attention mechanism understand better the features within image regions. It's use in regression tasks is not well explored, and hence in this paper they have shown promise in such a task like weakly-supervised counting. This success suggests potential in microorganism counting and beyond, including crowd and instance counting. Exploring ViTs further could unravel new methods for these tasks, highlighting their versatility and expanding their application scope in image-based problem-solving.

The methodology is being scrutinized in the current research. Current considerations consider more implementations of the architectures, like the use of pre-trained models (ResNet, TransCrowd, of pre-trained ViT back-bones for feature extraction) to analyze new approaches in regard of capturing the characteristics of the images in lower dimensions. The study could be also improved by the measurement of the performance of the architectures at inference time (time and floating-point operations "FLOPs"), e.g., ViTs usually have high amount of parameters which could mean that at inference they would not be as efficient as other architectures. Other means of comparison in between architectures can be the use of cross-dataset evaluation to measure extrapolation outside of their training dataset.

Other SOTA architectures are also considered to be applied in weakly-supervised microorganism counting, like CCTrans [25], and more in depth ablation study of the ViT backbone architecture, with the use of other iterations like the DeepViT [29], parallel ViT [26], XCiT [6] and CrossViT [2].

Future work might comprise the conclusions extracted from this analytic comparison into assembling a bacteria/cell oriented weakly-supervised counting model, and implementation in real systems for bacterial carrier tests. This can be useful for practitioners and biologists for solving such task by automatizing the counting of microbes, or for re-designing standardized quality tests with automatic systems that would make the process of quality assurance faster.

5 Conclusion

In this study, we conduct a comparative study of the ViT architectures with the intention of discerning their capabilities to achieve weakly-supervised microorganism enumeration. The results show that these architectures can all solve the problem of weakly supervised microbial enumeration, but some considerations must be taken into account. We show that, depending on the type of data, some approaches can be better than others under the same evaluation conditions, but at the same time we prove that the use of ViTs for weakly supervised counting of microorganisms is feasible and confirms the hypothesis that it can be used as an effective solution. Furthermore, the use of larger datasets is something that greatly affects the training of these architectures, so benchmarking with larger datasets should be considered. Overall, traditional architectures such as ResNet can generalize well for such a task, and they are not as computationally expensive as ViTs, although ViTs show potential in solving such a task as it has been demonstrated in this study and in previous works such as [17, 25, 18, 7].

6 Acknowledgments

This paper is part of the research and development project DesDet in collaboration with the department of Analytical Chemistry and Radiochemistry, Hollu Systemhygiene GmbH and Planlicht GmbH & Co KG. This project is funded by Standortagentur Tirol.

References

[1] Ching-Wen Chang, Yaw-Huei Hwang, Sergey A Grinshpun, Janet M Macher, and Klaus Willeke. Evaluation of counting error due to colony masking in bioaerosol sampling. *Applied and Environmental Microbiology*, 60(10):3732–3738, 1994.

[2] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification, 2021.

[3] C. A. Davidson, Christopher J. Griffith, Adrian C. Peters, and Louise Fielding. Evaluation of two methods for monitoring surface cleanliness-atp bioluminescence and traditional hygiene swabbing. *Luminescence : the journal of biological and chemical luminescence*, 14 1:33–8, 1999.

[4] Xin Ding, Qiong Zhang, and William J. Welch. Classification Beats Regression: Counting of Cells from Greyscale Microscopic Images based on Annotation-free Training Samples, October 2020. arXiv:2010.14782 [cs, eess].

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs].

[6] Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jegou. Xcit: Cross-covariance image transformers, 2021.

[7] Hairui Fang, Jin Deng, Yaoxu Bai, Bo Feng, Sheng Li, Siyu Shao, and Dongsheng Chen. Clformer: A lightweight transformer based on convolutional embedding and linear self-attention with strong robustness for bearing fault diagnosis under limited sample conditions. *IEEE Transactions on Instrumentation and Measurement*, 71:1–8, 2022.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[9] Shenghua He, Kyaw Thu Minn, Lilianna Solnica-Krezel, Mark Anastasio, and Hua Li. Automatic microscopic cell counting by use of deeply-supervised density regression model. In *Medical Imaging 2019: Digital Pathology*, page 19, March 2019. arXiv:1903.01084 [cs].

[10] Shenghua He, Kyaw Thu Minn, Lilianna Solnica-Krezel, Mark A. Anastasio, and Hua Li. Deeply-Supervised Density Regression for Automatic Cell Counting in Microscopy Images, November 2020. arXiv:2011.03683 [cs, eess].

[11] Fatemeh Hoorali, Hossein Khosravi, and Bagher Moradi. Automatic Bacillus anthracis bacteria detection and segmentation in microscopic images using UNet++. *Journal of Microbiological Methods*, 177:106056, October 2020.

[12] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds, 2018.

[13] Raymond L Kepner Jr and James R Pratt. Use of fluorochromes for direct enumeration of total bacteria in environmental samples: past and present. *Microbiological reviews*, 58(4):603–615, 1994.

[14] Falko Lavitt, Demi J. Rijlaarsdam, Dennet van der Linden, Ewelina Weglarz-Tomczak, and Jakub M. Tomczak. Deep learning and transfer learning for automatic cell counting in microscope images of human cancer cell lines. *Applied Sciences*, 11(1111):4912, January 2021.

[15] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[16] Xiangtai Li, Henghui Ding, Haobo Yuan, Wenwei Zhang, Jiangmiao Pang, Guangliang Cheng, Kai Chen, Ziwei Liu, and Chen Change Loy. Transformer-based visual segmentation: A survey. (arXiv:2304.09854), December 2023. arXiv:2304.09854 [cs].

[17] Dingkang Liang, Xiwu Chen, Wei Xu, Yu Zhou, and Xiang Bai. TransCrowd: weakly-supervised crowd counting with transformers. *Science China Information Sciences*, 65(6):160104, June 2022. arXiv:2104.09116 [cs].

[18] Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. CounTR: Transformer-based Generalised Visual Counting, June 2023. arXiv:2208.13721 [cs].

[19] Zhuangzhuang Miao, Yong Zhang, Yuan Peng, Haocheng Peng, and Baocai Yin. DTCC: Multi-level dilated convolution with transformer for weakly-supervised crowd counting. *Computational Visual Media*, April 2023.

[20] Roberto Morelli, Luca Clissa, Roberto Amici, Matteo Cerri, Timna Hitrec, Marco Luppi, Lorenzo Rinaldi, Fabio Squarcio, and Antonio Zoccoli. Automating cell counting in fluorescent microscopy through deep learning with c-ResUnet. *Scientific Reports*, 11(1):22920, November 2021.

[21] Roberto Morelli, Luca Clissa, Roberto Amici, Matteo Cerri, Timna Hitrec, Marco Luppi, Lorenzo Rinaldi, Fabio Squarcio, and Antonio Zoccoli. Automating cell counting in fluorescent microscopy through deep learning with c-resunet. *Scientific Reports*, 11:22920, 11 2021.

[22] Mahboob Nemati, Aliasghar Hamidi, Solmaz Maleki Dizaj, Vahid Javaherzadeh, and Farzaneh Lotfipour. An overview on novel microbial determination methods in pharmaceutical and food quality control. *Advanced pharmaceutical bulletin*, 6 3:301–308, 2016.

[23] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything, 2021.

[24] Michael Riepl, Sonja Schauer, Sonja Knetsch, Elisabeth Holzhammer, Andreas Farnleitner, Regina Sommer, and Alexander Kirschner. Applicability of solid-phase cytometry and epifluorescence microscopy for rapid assessment of the microbiological quality of dialysis water. *Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association*, 26:3640–5, 09 2011.

[25] Ye Tian, Xiangxiang Chu, and Hongpeng Wang. CCTrans: Simplifying and Improving Crowd Counting with Transformer, September 2021. arXiv:2109.14483 [cs].

[26] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Hervé Jégou. Three things everyone should know about vision transformers, 2022.

[27] Yao Xue, Nilanjan Ray, Judith Hugh, and Gilbert Bigras. Cell Counting by Regression Using Convolutional Neural Network. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, volume 9913, pages 274–290. Springer International Publishing, Cham, 2016. Series Title: Lecture Notes in Computer Science.

[28] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 589–597, 2016.

[29] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer, 2021.