Title

Expanding Vision in Tree Counting: Novel Ground Truth Generation and Deep Learning Model

Authors

Ton-That, Minh Nhat

Le, Tin Viet

Truong, Nhien Hao

et al.

Publication Date

2024-08-02

DOI

10.1109/icce62051.2024.10634677

Expanding Vision in Tree Counting: Novel Ground Truth Generation and Deep Learning Model

Minh Nhat Ton That *
Department of Computer Science and Engineering
Vietnamese German University
Binh Duong, Vietnam
10422050@student.vgu.edu.vn

Tin Viet Le *
Department of Computer Science and Engineering
Vietnamese German University)
Binh Duong, Vietnam
10422078@student.vgu.edu.vn

Nhien Hao Truong *
Department of Computer Science and Engineering
Vietnamese German University
Binh Duong, Vietnam
10422062@student.vgu.edu.vn

An Dinh Le
Department of Computer Science and Engineering
University of California San Diego
La Jolla, USA
d0le@ucsd.edu

Anh Duy Pham
Joint Lab Artificial Intelligence and Data Science
Universität Osnabrück
Osnabrück, Germany
apham@uni-osnabrueck.de

Hien Bich Vo
Department of Computer Science and Engineering
Vietnamese German University
Binh Duong, Vietnam
hien.vb@vgu.edu.vn

***Abstract*—Manual labor in tree management is expensive and not efficient enough to track changes in the number and distribution of trees in a large area. Recent methods focus on automatic counting using deep-learning networks, in particular density map-based methods, in aerial images. This research introduces two contributions: a new ground truth generation practice to boost the training effectiveness by minimizing the loss of labeled trees in the target density maps; and a combination of a Vision Transformer (ViT) and a dilated convolutional neural network (CNN), termed TreeVision, to generate high-quality density maps. The combination of these two methodologies provides us with high-quality ground truths to extract rich contextual information by the global self-attention performed by ViT and a large receptive field of dilated convolutions. Empirically, both of our proposals yield positive results, in particular, better performance of models trained in updated ground truth and state-of-the-art results when evaluating TreeVision on two benchmark tree-counting datasets (Yosemite and KCL-London). Our TreeVision delivers a 9.29% lower Mean Absolute Error (MAE) in the KCL-London dataset and competitive results in the Yosemite dataset in comparison with previous leading methods.**

I. Introduction

Trees are vital biological resources that serve many purposes relating to human well-being, such as agriculture [1], [2], human health [3], [4], [5], or carbon sequestration and greenhouse effect alleviation [3], [5]. Additionally, since manual labor was barely effective for large areas, a clamor for automatic counting or estimating the number of trees in a specific area has arisen to meet the need for statistical analysis in management. Many counting methods have been introduced with a variety of input sources, which are primarily remote sensing imagery [2], [6], [7], retrieved by various manned or unmanned devices such as satellites [8] or UAVs [2]. Along with the exponential applications of DNNs and Deep Learning, many Convolutional Neural Network (CNN) models have been introduced to solve this problem efficiently which can be classified into two main types. The first one utilizes bounding box annotations and outputs the number of trees detected as in a tree detection task [11]. The second one applies emerging and trending point-based annotations and density maps [6], [7], [8], [12] which is recently ubiquitous for counting objects in high-density images such as in [14], [15], [13]. Approaches deployed for this kind of annotation are not prone to typical errors of bounding box annotations such as overlapped areas or missing details. However, density map-based methods rely heavily on the quality of the ground truths to well extract features and estimate dense values at the pixel level.

In this paper, we propose a new ground truth generation practice that is improved from [39] for density map-based methods in the tree counting task. Our method produces more accurate density maps and better performance of models trained with the proposed ground truths. We also propose a new density map-based model for the tree counting task inspired by CSRNet [15], originally introduced in the crowd counting field, with a new Vision Transformer backbone. Firstly introduced in [16], attention-based methods have appealed to researchers by their power of global context consideration and better suitability for domain generalization than the traditional CNN model.

II. Literature review

A. Density maps in tree counting

Counting trees in a dense forest canopy is considerably difficult because trees can appear continuously and overlap each other in aerial images. Many early research focuses on detection-based approaches by identifying and localizing individual trees with bounding boxes using Mask-RCNN [18]

*These three authors contributed equally.

YOLO family [19], or Faster R-CNN [20]. However, in extremely thick tree regions, the detection-based approaches perform unsatisfactorily due to occlusion and cluttered context [12], [15], [20], [21].

Therefore, the density map-based methods, which utilize spatial information to increase performance, have been introduced to tackle the tree counting task. A density map is a form of visualization in computer vision and spatial analysis, generated by applying a Gaussian filter, which aggregates and smooths individual data points represented as dots into continuous intensity values [21]. In the context of images, density maps can be generated from dot maps where each dot represents an object or point of interest in the image [22]. The integral of this density map provides a quantitative representation of the total number of objects in the image. Cheng and Shang [8] combine CNN, a transformer encoder, and a Density Map Generator to predict the density of trees. In [12], a semi-supervised framework for tree counting is built upon a pyramid vision transformer to extract multi-scale features. In [23], the authors compiled a tree-counting dataset comprising 24 GF-II images. Their approach involved rasterizing point labels to match the spatial resolution and subsequently applying discrete Gaussian kernel blurring to create density maps.

B. Vision Transformer

The success of Transformer [16] in Natural Language Processing has motivated researchers to apply attention-based methods to Computer Vision tasks to which CNN-based models are the dominant approach. In [38], a novel network called Pyramid Attention Network is introduced with Feature Pyramid Attention and Global Attention Upsample modules to extract global contextual information for semantic segmentation. Vision Transformer (ViT) [24] also employs an attention-based approach with a pure Transformer architecture tackling classification tasks. It is kept as close to the original Transformer as possible with no image-specific inductive biases and a standard Transformer encoder. The encoder comprises alternating layers of Multi-head Attention and Multi-Layer Perception blocks whose inputs are embedded fixed-size patches added with learnable position embeddings and an extra [class] token [25]. ViT delivers modest results when training in mid-sized datasets but superior performance in large datasets compared with state-of-the-art CNNs. Following the success of ViT, many variants have been introduced, for example, SimpleViT [26], T2TViT [27], CrossViT [28], etc. improving the performance of Transformer-based models in vision tasks. Meanwhile, in [29], the authors combine Vision Transformer with Mask-RCNN [30], a convolutional neural network, to create ViTDet. ViTDet explores two main ideas to maintain a simple design while still benefiting from high-resolution inputs including a Simple Feature Pyramid instead of a conventional hierarchical Feature Pyramid Network (FPN) [31] and a backbone adaption with a pre-trained model from Masked Autoencoder (MAE) [32].

III. Methods

A. Proposed Ground Truth Generation Method

1) Density map generation: The ground truth tree density map is derived from the keypoint annotations of the trees. To accomplish this, we adopt the method outlined in [8] and [39], which utilizes geometric-adaptive Gaussian functions to account for the varying size of each tree. This approach dynamically adjusts the spread of the Gaussian filter based on the average distance to the k nearest neighbors, ensuring that tree spreads do not overlap each other. The Gaussian filter is set to operate in constant mode, extending its effect beyond the image boundaries. The function is defined as:

$$F(x) = \sum_{i=1}^{N} \delta\left(x - x_i\right) \times G_{\sigma_i}, \text{ with } \sigma_i = \beta \overline{d_i} \qquad (1)$$

In this case, d_i denotes the average distance of the k nearest neighbors, and δ is the Dirac delta function for each targeted item x_i in the ground truth. The density map is produced by convolving $\sigma(x - x_i)$ with a Gaussian kernel that is parameterized by σ_i (standard deviation), where x denotes a pixel's location in the picture. We conduct our experiment using the same methods as described in [39], with $k = 3$ and $\beta = 0.3$.

2) Contribution on density map generation: Aerial photogrammetry in forestry must cover a vast region which significantly increases the image resolution to retain as much information as possible. For example, on the Yosemite dataset [8], a tree counting dataset, the image has an enormous dimension of 19200 x 38400 pixels which necessitates cropping to a smaller size. However, because the cropping process divides the tree area into smaller images and breaks the continuity of the area across each image, the calculation for geometric-adaptive Gaussian filters may be affected, especially for trees at the edges where the nearest neighbor could be cropped into the adjacent image. Moreover, because the values spread beyond the edges are lost, if we apply the Gaussian Filter to a small-size cropped image, the number of objects along the edges will increase leading to a bigger loss of target trees than applying the Gaussian Filter to a bigger image.

To address these issues and reduce information loss, we explore applying the Gaussian filter to larger images (1536x1536 pixels, three times the required size) and subsequently splitting them into nine sections to achieve the desired size. The sample density maps are provided in Fig. 1. It can be observed that the cropped density map contains more information on the right side and bottom edge compared to the original left-side map. This discrepancy can be attributed to the fact that the larger image retains the relationship among trees which affects the spreading of the Gaussian filter and the number of trees at the edges of a 1536x1536 image is fewer than in nine 512x512 images leading to a smaller loss of objects. As a result, the density map generated from the larger image captures more information near the edges, which boosts the learning of models.

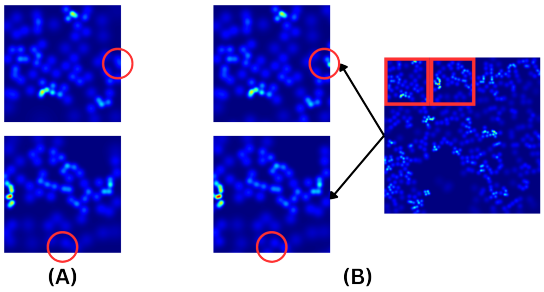

Fig. 1. Visualization of different ways of generating density maps. In (A), density maps are generated directly in size 512x512. In (B), density maps are constructed initially in size 1536x1536 and then cropped into 9 sections of size 512x512. Red circles indicate the difference in two ways to generate density maps in the required size.

B. TreeVision architecture

We draw inspiration from the pioneering work in [15] on CSRNet, which has set a remarkable benchmark in leveraging deep learning for crowd density estimation. CSRNet designed with VGG16 at the backbone and dilated convolutions follow enables efficient handling of variations in crowd density and occlusions surpassing traditional CNN architectures. Based on this foundation, we adopt Vision Transformer, in particular ViT-Base as the backbone of our model, combining its global attention mechanism with the large receptive field of dilated convolutions to capture comprehensive contextual information in images. The resulting architecture, termed TreeVision, is depicted in Fig. 2.

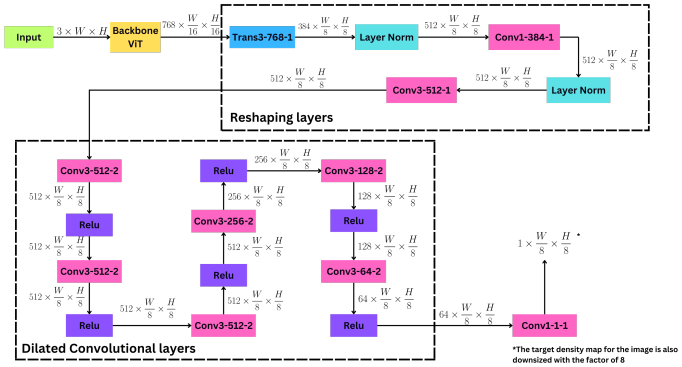

Fig. 2. Configuration of TreeVision, in which the parameters are denoted as (layer type)(kernel size)-(number of filters)-(dilation rate). Conv stands for a convolutional layer and Trans stands for a transposed convolutional layer. If the dilation rate is equal to 1, the corresponding dilated convolution is equivalent to a standard convolution.

1) Dilated Convolution: Density estimation requires extracting contextual information locally and globally to predict at a pixel level [34]. Conventional layers of pooling and standard convolution can be implemented in a deep network to look at context in a large field but they consequently increase the number of trainable parameters as well as calculations and lose details during processing [15]. A solution suggested by Yu and Koltun [35] is to use dilated convolutions which can exponentially expand the receptive field without loss of resolution and maintaining more details. Yu and Koltun [35] define a 2D Dilated Convolution as follows:

$$(F *_l k)(p) = \sum_{s+lt=p} F(s)k(t) \tag{2}$$

where $*_l$ is referred to as the dilated convolution, ($F*_lk$)(p) is the output for input F(s) and discreet filter k(t). The parameter l is the dilation rate which determines the size of gaps in the filter. As shown in Fig. 3, the higher the dilation rate is, the more sparse the kernel used in the dilated convolution becomes. In [15], [36], [37], [9] dilated convolutional layers have been implemented and proved effective in improving the performance of neural networks by capturing more high-level contextual information without increasing the number of parameters or computations. Inspired by the model CSRNet [15], we implement our dilated convolutional networks with 3x3 kernels and dilation rates equal to 2. An extra 1x1 standard conventional layer is added at the end of the architecture to estimate a density map.

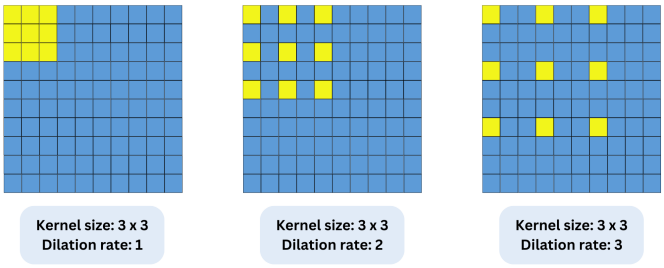

Fig. 3. Visualization of dilated convolutions with dilation rates ranging from 1 to 3. This figure is adapted from [15].

2) Vision Transformer as the backbone: We choose ViT-Base from [24] with a 16x16 input patch size as the backbone of our model following the configurations of ViTDet [29]. The primary components of ViTDet's backbone remain close to the original work in [24] with positional and patch embeddings and Transformer blocks. Additionally, ViTDet proposes two new features to the backbone including a Simple Feature Pyramid (SFP) and backbone propagation. The SFP is a simple hierarchical design having the last feature map of ViT passed through sets of convolutions and deconvolutions in parallel to extract multi-scale information. This strategy is introduced to replace the conventional Feature Pyramid Network to eliminate the hierarchical constraints of the ViT backbone. Another contribution of ViTDet is backbone propagation with the help of window attention. In this component, ViTDet aims to move the backbone toward a task-agnostic approach with fewer inductive biases, which lets the backbone be pre-trained with irrelevant large datasets before fine-tuning with more specific but scarce data. In our work, we remove the Simple Feature Pyramid from our backbone to reduce the computation cost but still retain the other features of ViTDet. Moreover, while the last feature map of Vision Transformer has 768 channels and is 1/16 the input resolution, the input of the dilated convolutions requires 512 channels and the size of 1/8 the input. Hence, to

TABLE I
THE CHARACTERISTICS OF IMAGES AND TREE SAMPLES OF KCL-LONDON AND YOSEMITE DATASETS

Dataset	Ratio ($Train : Test$)	Image size	Number of images	Minimum number of trees	Maximum number of trees	Average density (tree/images)	Total
KCL-London	452 : 161	1024×1024	613	4	332	155	95,067
Yosemite	1350 : 1350	512×512	2700	0	113	36	98,949

maintain an appropriate shape, we add, after the backbone, an extra convolutional block that consists of a transposed and two standard convolutions. Switching from the VGG16 backbone in the original work to the Vision Transformer helps us utilize the global attention features of the new backbone to yield better results and mitigate the reliance on the spatial biases of convolutions which makes our model less vulnerable to the changes in the spatial structure of the images.

IV. EXPERIMENT SETUP

A. Dataset

1) KCL-London dataset: This tree-counting dataset contains 613 annotated and 308 unannotated images each having a size of 1024x1024 [12]. The area is focused on London, the United Kingdom. Within the labeled set, 452 samples are for training and 161 samples are reserved for testing with 95,067 annotated trees in total. Because our model is trained in a supervised manner, the unlabeled set is purposefully set aside.

2) Yosemite dataset: The study area is located at Yosemite National Park, California, United States of America [8]. The original image covers a 2262.5m x 4525.1m rectangular area in reality(19200 x 38400 pixels in image). The total number of labeled trees is 98,949. The data are collected from Google Maps and divided equally into 4 regions, of which regions B and D are used for training, while regions A and C are used for testing. To compare with a state-of-the-art model [12], we crop the images to 512x512 pixels and update the labels. The total number of images in this size is 2700 images.

The features of the two datasets are presented in Table I.

B. Metrics

To evaluate the performance of models, we take into consideration the ground truth and predicted counts of trees which are produced by taking the total value of all pixels in the corresponding density maps. Following the published works for crowd estimation [15] and tree counting [38], Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are chosen as benchmarking metrics. Moreover, based on a comprehensive comparison of the tree counting task [7], R-squared (R^2) is also adopted in this experiment. The three metrics are defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i^{GT}| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|y_i - y_i^{GT}\right\|^2} \tag{4}$$

$$R^2 = \frac{\sum_{i=1}^{N} y_i - y_i^{GT}}{\sum_{i=1}^{N} y_i - \overline{y}^{GT}} \tag{5}$$

where N denotes the total number of samples, y_i and y_i^{GT} is the predicted count and ground truth count for the $i-th$ sample respectively, and $\overline{y}^{GT}$ represents the mean tree number of all samples. Overall, lower MAE and RMSE and higher R^2 correlate with better performance.

C. Training setup

We implement TreeVision and CSRNet on Google Colab using a T4 GPU, batch size equal to 1, and trained for 20 epochs for the former and 25 epochs for the latter on each dataset. Following the approach outlined in [15], the dataset is multiplied by four and randomly shuffled for each epoch. Additionally, we initialize our dilated convolutions with the weights of CSRNet trained in dataset ShanghaiB [39] and our Vision Transformer backbone with the weights of ViTDet [29] to expedite training. For optimization, we utilize the AdamW optimizer with a learning rate of 10^{-4} and a weight decay of $5\text{x}10^{-4}$.

D. Comparisons of different density map generation method

To validate the effects of different ground truths on the training effectiveness, we train and evaluate TreeVision and CSRNet with two different sizes of density map generation for the Yosemite dataset [8]. From Table III, although the number of trees in the proposed density maps is higher after applying the Gaussian filter with 45666 objects compared to 43324 in the current method, both models make fewer errors when trained with our proposed ground truth. Similar to the experiments in [10], because of the reliance on spatial biases of CNN, in particular, the VGG16 backbone of CSRNet, amplified by the noises of batch size equal to 1, CSRNet witnesses a more significant drop in the MAE of 2.33 compared to only a slight decrease of 0.2 of TreeVision. The positive effect of our proposed method is also reflected in Fig. 4. The TreeVision model trained with the normal method tends to predict fewer trees than the actual number with the graph widening at the lower half. Meanwhile, the results produced in the proposed method form a taller graph with higher similarity to the targets.

TABLE II
COMPARISONS WITH STATE-OF-THE-ART METHODS ON KCL-LONDON AND YOSEMITE DATASETS.
LOWER MAE AND RMSE AND HIGHER R^2 CORRELATE WITH BETTER PERFORMANCE.
THE BEST AND SECOND-BEST RESULTS ARE WRITTEN IN BOLD AND ITALIC RESPECTIVELY.

Dataset	**KCL-London**			**Yosemite**		
Method	$MAE \downarrow$	$RMSE \downarrow$	$R^2 \uparrow$	$MAE \downarrow$	$RMSE \downarrow$	$R^2 \uparrow$
MCNN	25.87	34.12	0.45	10.44	12.45	0.61
CSRNet	20.21	25.69	0.69	7.25	9.35	0.76
SASNet	24.33	30.12	0.56	6.33	8.46	0.81
EDNet	26.18	32.02	0.52	9.92	12.39	0.60
S-TreeFormer	<u>18.52</u>	<u>24.32</u>	<u>0.72</u>	**4.29**	**5.85**	**0.91**
TreeVision (ours)	**16.80**	**21.98**	**0.76**	<u>4.58</u>	<u>6.17</u>	<u>0.89</u>

TABLE III
COMPARISON OF MODELS ON DIFFERENT GROUND TRUTHS OF YOSEMITE DATASET.
LOWER MAE AND RMSE AND HIGHER R^2 CORRELATE WITH BETTER PERFORMANCE.

	Yosemite (512x512)			**Yosemite (1536x1536)**		
Method	$MAE \downarrow$	$RMSE \downarrow$	$R^2 \uparrow$	$MAE \downarrow$	$RMSE \downarrow$	$R^2 \uparrow$
CSRNet	7.25	9.35	0.76	4.92	6.48	0.89
TreeVision	4.58	6.17	0.89	4.32	5.76	0.91

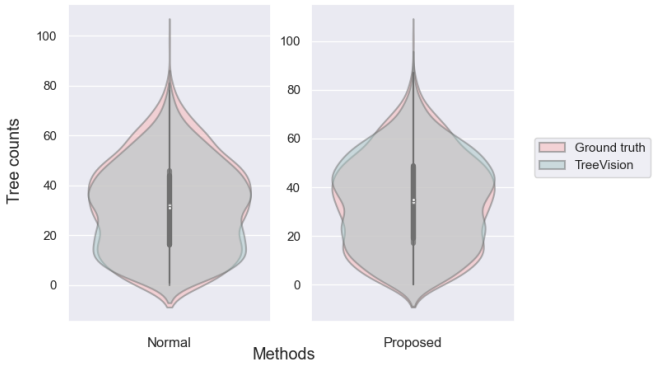

Fig. 4. Comparison of the target and predicted number of trees in the normal and proposed test set of the Yosemite dataset. The inference results are produced by TreeVision trained on the corresponding ground truth.

E. Comparisons with state-of-the-art Tree-counting models

To evaluate the effectiveness of TreeVision, we compare it with models reported in [12] including MCNN [39], CSRNet [15], SASNet [40], EDNet [23], and S-TreeFormer [12], among which MCNN, CSRNet and SASNet are state-of-the-art models in crowd counting using density estimation and have been implemented for the tree counting task in [12]. The results shown in Table II reveal that our model delivers competitive performance to the previous leading models in both KCL-London and Yosemite datasets. In the experiment of the KCL-London dataset, TreeVision achieves the best accuracy with 9.29% lower MAE and 9.63% lower RMSE than S-Treeformer. Meanwhile, on the Yosemite dataset, our model ranks second in all three metrics with small gaps from S-Treeformer, however, we still outperform the other models by a large margin.

V. CONCLUSION

In this paper, we propose a new ground truth generation practice involving applying a Gaussian Filter in a bigger region before cropping the images down to the required size. This method is proven to boost the training effectiveness and discovers the potential of using bigger original ground truth images to improve the performance of tree-counting models. We also introduce a novel framework called TreeVision based on CSRNet architecture that achieved competitive results to

its predecessors in the tree counting task. Combining layers of global self-attention and dilated convolutions, the model attains a more detailed look at the images to generate high-quality pixel-wise density estimation.

REFERENCES

[1] F. Santoro, E. Tarantino, B. Figorito, S. Gualano, and A. M. D'Onghia, "A tree counting algorithm for precision agriculture tasks," Int. J. Digit. Earth, vol. 6, no. 1, pp. 94–102, Jan. 2013.

[2] C. Donmez, O. Villi, S. Berberoglu, and A. Cilek, "Computer vision-based citrus tree detection in a cultivated environment using UAV imagery," Comput. Electron. Agric., vol. 187, p. 106273, Aug. 2021.

[3] D. J. Nowak and D. E. Crane, "Carbon storage and sequestration by urban trees in the USA," Environ. Pollut., vol. 116, no. 3, pp. 381–389, Mar. 2002.

[4] J. A. Salmond et al., "Health and climate related ecosystem services provided by street trees in the urban environment," Environ. Health, vol. 15, no. S1, p. S36, Dec. 2016.

[5] K. L. Wolf, S. T. Lam, J. K. McKeen, G. R. A. Richardson, M. Van Den Bosch, and A. C. Bardekjian, "Urban Trees and Human Health: A Scoping Review," Int. J. Environ. Res. Public. Health, vol. 17, no. 12, p. 4371, Jun. 2020.

[6] K. Djerriri, M. Ghabi, M. S. Karoui, and R. Adjoudj, "Palm Trees Counting in Remote Sensing Imagery Using Regression Convolutional Neural Network," in IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia: IEEE, Jul. 2018.

[7] L. Yao, T. Liu, J. Qin, N. Lu, and C. Zhou, "Tree counting with high spatial-resolution satellite imagery based on deep neural networks," Ecol. Indic., vol. 125, p. 107591, Jun. 2021.

[8] G. Chen and Y. Shang, "Transformer for Tree Counting in Aerial Images," Remote Sens., vol. 14, no. 3, p. 476, Jan. 2022.

[9] N. P. L. Le, H. N. Do, V. T. D. Huynh and L. Mai, "Image Super Resolution Using Deep Learning," 2022 IEEE Ninth International Conference on Communications and Electronics (ICCE), Nha Trang, Vietnam, 2022, pp. 369-374.

[10] B. R. Mitchell, "The Spatial Inductive Bias of Deep Learning".

[11] P. N. Chowdhury, P. Shivakumara, L. Nandanwar, F. Samiron, U. Pal, and T. Lu, "Oil palm tree counting in drone images," Pattern Recognit. Lett., vol. 153, pp. 1–9, Jan. 2022.

[12] H. A. Amirkolaee, M. Shi, and M. Mulligan, "TreeFormer: a Semi-Supervised Transformer-based Framework for Tree Counting from a Single High Resolution Image," IEEE Trans. Geosci. Remote Sens., vol. 61, pp. 1–15, 2023.

[13] H. C. Nguyen et al., "A method for automatic honey bees detection and counting from images with high density of bees," 2022 IEEE Ninth International Conference on Communications and Electronics (ICCE), Nha Trang, Vietnam, 2022, pp. 406-411.

[14] Q. Song et al., "Rethinking Counting and Localization in Crowds: A Purely Point-Based Framework," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada: IEEE, Oct. 2021, pp. 3345–3354.

[15] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT: IEEE, Jun. 2018, pp. 1091–1100.

[16] A. Vaswani et al., "Attention is All you Need," in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017.

[17] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." in 3rd International Conference on Learning Representations (ICLR 2015), Computational and Biological Learning Society, 2015, pp. 1–14.

[18] M. Machefer, F. Lemarchand, V. Bonnefond, A. Hitchins, and P. Sidiropoulos, "Mask R-CNN Refitting Strategy for Plant Counting and Sizing in UAV Imagery," Remote Sens., vol. 12, no. 18, Art. no. 18, Jan. 2020.

[19] K. Itakura and F. Hosoi, "Automatic Tree Detection from Three-Dimensional Images Reconstructed from 360° Spherical Camera Using YOLO v2," Remote Sens., vol. 12, no. 6, Art. no. 6, Jan. 2020.

[20] J. Zheng, W. Li, M. Xia, R. Dong, H. Fu, and S. Yuan, "Large-Scale Oil Palm Tree Detection from High-Resolution Remote Sensing Images Using Faster-RCNN," in IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Jul. 2019, pp. 1422–1425.

[21] J. Wan, Q. Wang, and A. B. Chan, "Kernel-Based Density Map Generation for Dense Object Counting," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 3, pp. 1357–1370, Mar. 2022.

[22] V. Lempitsky and A. Zisserman, "Learning To Count Objects in Images," in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2010.

[23] L. Yao, T. Liu, J. Qin, N. Lu, and C. Zhou, "Tree counting with high spatial-resolution satellite imagery based on deep neural networks," Ecol. Indic., vol. 125, p. 107591, Jun. 2021.

[24] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," presented at the International Conference on Learning Representations, Oct. 2020.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[26] L. Beyer, X. Zhai, and A. Kolesnikov, "Better plain ViT baselines for ImageNet-1k." arXiv, May 03, 2022. Accessed: Feb. 19, 2024.

[27] L. Yuan et al., "Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada: IEEE, Oct. 2021, pp. 538–547.

[28] C.-F. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada: IEEE, Oct. 2021, pp. 347–356.

[29] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring Plain Vision Transformer Backbones for Object Detection," in Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX, Berlin, Heidelberg: Springer-Verlag, Oct. 2022, pp. 280–296.

[30] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2980-2988.

[31] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 936-944.

[32] K. He, X. Chen, S. Xie, Y. Li, P. Dollár and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 15979-15988.

[33] N. Park and S. Kim, "How Do Vision Transformers Work?" arXiv, Jun. 08, 2022. Accessed: Feb. 21, 2024.

[34] N. Takahashi and Y. Mitsufuji, "Densely connected multidilated convolutional networks for dense prediction tasks," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA: IEEE, Jun. 2021, pp. 993–1002.

[35] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions." arXiv, Apr. 30, 2016. Accessed: Feb. 22, 2024.

[36] I. Bakour, H. N. Bouchali, S. Allali, and H. Lacheheb, "Soft-CSRNet: Real-time Dilated Convolutional Neural Networks for Crowd Counting with Drones," in 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH), Feb. 2021, pp. 28–33.

[37] A. Salehi and M. Balasubramanian, "DDCNet: Deep dilated convolutional neural network for dense prediction," Neurocomputing, vol. 523, pp. 116–129, Feb. 2023.

[38] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid Attention Network for Semantic Segmentation." arXiv, Nov. 25, 2018.

[39] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 589–597.

[40] Q. Song et al., "To Choose or to Fuse? Scale Selection for Crowd Counting," Proc. AAAI Conf. Artif. Intell., vol. 35, no. 3, Art. no. 3, May 2021