GRANULO-10K: A LARGE-SCALE BENCHMARK DATASET FOR MULTIPLE-VIEW INDUSTRIAL GRANULOMETRY

Pasquale Coscia, Angelo Genovese, Vincenzo Piuri, Fabio Scotti
Department of Computer Science, Università degli Studi di Milano, Italy

ABSTRACT

Granulometric analysis of wood strands ensures structural integrity and production efficiency of Oriented Strand Board (OSB). Current vision-based studies have demonstrated the potential of automated image analysis for particle size estimation, but their progress is significantly hindered by the absence of public, high-quality, and domain-specific datasets. This paper introduces *Granulo-10k*, the first curated, open dataset of multiple-view wood-strand imagery designed specifically for research on OSB strand segmentation and multiple-view granulometry, covering height, width, and thickness. The dataset includes about $10,000$ high-resolution images of 200 wood strands captured under controlled acquisition conditions, along with granulometric ground truth. We also provide 3D point clouds to capture three-dimensional information. We describe the acquisition protocol and annotation methodology used to ensure representativeness across real production variability. Baseline evaluations using modern deep neural networks and foundation models demonstrate the dataset's utility for benchmarking while revealing open research challenges such as precise thickness estimation.

***Index Terms*—** Oriented Strand Board (OSB), wood, granulometry, vision foundation models.

1. INTRODUCTION

Oriented Strand Board (OSB) is a widely used engineered wood composite whose mechanical properties strongly depend on the quality, size distribution, and orientation of the individual wood strands forming its multilayer structure. During production, thousands of elongated strands are deposited onto a conveyor to form a mat, which is then pressed into structural panels [1]. The geometric characteristics of these strands directly influence panel strength, glue usage, and optimization of the manufacturing process [2]. OSB is also used to improve specific mechanical properties of Cross-Laminated Timber (CLT) compared to conventional all-lumber CLT. In particular, when OSB is used selectively, for example in the transverse layer, it can enhance shear performance and reduce long-term bending creep [3].

Within OSB production, granulometric measurement remains essential for detecting deviations in strand size and distribution that may compromise product quality [4]. Early studies based on image analysis of laboratory-manufactured panels investigated the influence of strand geometry on OSB mechanical behavior. Nishimura *et al*. [5] quantified the effects of strand height, width, and aspect ratio on bending stiffness and strength, showing that larger, high-aspect-ratio strands improve in-plane properties at the expense of increased variability. Several parametric investigations extended these findings by introducing descriptors such as slenderness and thinness and by explicitly varying strand thickness. In particular, Zhuang *et al*. [6] demonstrated strong interactions between strand dimensions, density profile, and global properties including bending strength, internal bond, and thickness swelling, highlighting the need for accurate control of all three geometric dimensions.

Compared to traditional granulometry methods, such as vibrating-sifter analysis [7], vision-based systems provide a more efficient and consistent alternative, enabling non-contact, in-line or off-line inspection [8]. For example, Donida *et al*. [4] introduced a granulometry-inspired framework combining color-based segmentation and multiscale analysis to estimate strand height and width distributions directly on the production line. Similarly, Verheyen *et al*. [9] present an industrial vision-based sorting system. Visual features related to shape (*e.g*., circularity, anisometry, bulkiness), texture (fiber structure), and color are extracted and fed into a multi-layer perceptron neural network trained on thousands of labeled samples for classification. However, while effective, these approaches are inherently limited to 2-D projections and typically assume strand thickness as a fixed process parameter or obtain it through sparse manual sampling. More recently, Hong *et al*. [8] combine deep learning with high-speed imaging to perform real-time instance segmentation and tracking of OSB strands, estimating height,

This work was supported in part by the EC under project EdgeAI (101097300). Project EdgeAI is supported by the Chips Joint Undertaking and its members including top-up funding by Austria, Belgium, France, Greece, Italy, Latvia, Netherlands, and Norway under grant agreement No. 101097300. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union, the Chips Joint Undertaking, or the Italian MUR. Neither the European Union, nor the granting authority, nor Italian MUR can be held responsible for them.
The authors would like to thank IMAL s.r.l., San Damaso (Modena), Italy, for their valuable cooperation in the data provided and the careful classification of samples. The authors would also like to acknowledge the support of Prof. Ruggero Donida Labati for his valuable contribution to the data collection process.

width, and orientation on an industrial line. However, even in this state-of-the-art system, strand thickness is not measured at the individual level.

To allow thickness estimation, a recent approach acquires images of particles in free fall rather than resting on a conveyor belt, thus reducing overlap and enabling access to additional geometric cues, such as depth [10]. Furthermore, to explicitly measure thickness, several studies in related wood-based processes explore 3D sensing solutions. For example, Lopez *et al*. [11] present a laser-scanner system for estimating chip height, width, and thickness in pulp-chip quality control.

Despite the well-established importance of full three-dimensional strand geometry for OSB performance, the lack of openly accessible datasets and inline measurement systems capable of jointly estimating strand height, width, and thickness remains a significant limitation. This gap motivates the development of new data acquisition strategies and learning-based methods capable of capturing complete strand geometry at scale.

To address these limitations, our paper introduces Granulo-10k[1], a large-scale dataset for multiple-view industrial granulometry and OSB strand research. By providing the community with an open, standardized dataset, Granulo-10k aims to enable reproducible experiments, accelerate the development of robust granulometry algorithms, and facilitate comparisons between methodologies. Our contributions are threefold:

- A large-scale OSB dataset captured using a multiple-view acquisition system under controlled laboratory conditions, for a total of about $10,000$ images.
- High-quality annotations, including granulometric measurements and segmentation masks. 3D point clouds are also included.
- Baseline analyses using data-driven (*e.g*., deep learning-based) approaches to establish performance reference points and highlight research challenges.

The remainder of the paper is structured as follows. Section 2 describes the dataset and the acquisition procedure. Section 3 describes the deep learning-based methodology for granulometry estimation, and Section 4 presents the experimental evaluation. Lastly, Section 5 concludes the work.

2. THE DATASET

Granulo-10k contains images of thinly chopped pieces of wood, known as strands, used in the production of OSB wood panels. We manually labeled and measured 200 strands using a caliper. Since the strands are irregularly shaped, we considered the maximum height and width of the strand. To measure thickness, which can have small variations throughout the strand, multiple measurements were taken in different parts of the strand and then averaged. After measurement, we divided the strands into two categories: *i)* 100 strands compliant with the reference size provided by the manufacturer, with an average size of $h \times w \times t = 115 \times 20 \times 0.70\,\mathrm{mm}$; *ii)* 100 strands not compliant with the reference size, with an average size $h \times w \times t = 91 \times 9 \times 0.65\,\mathrm{mm}$. Non-compliant strands may be produced by errors or wear in the machinery.

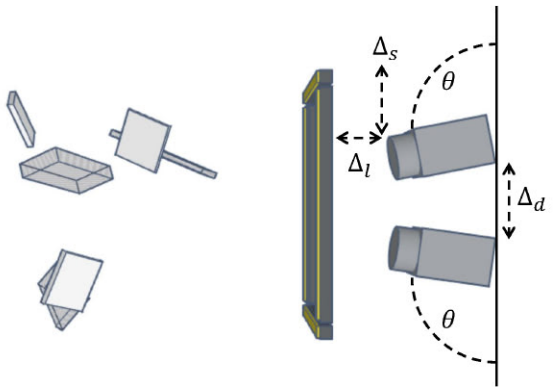

Fig. 1. Outline of the multiple-view acquisition setup (top view).

We collected the dataset using a calibrated, multiple-view acquisition system, composed of two Sony SX90CR color cameras, synchronized using a trigger mechanism. The cameras were placed at the same height and oriented at an angle $\theta = 85°$ with respect to the support. The distance between them was $\Delta_d = 125\,\mathrm{mm}$. The cameras were calibrated using the algorithms described in [12, 13]. For illumination, we placed four LED bars arranged in a rectangular fashion, at distances $\Delta_s = \Delta_l = 90\,\mathrm{mm}$ from the cameras, so as to provide approximately uniform illumination over the field of view [10] (see Fig. 1). To ensure that the two cameras captured images simultaneously, we used a trigger mechanism connected to a photocell.

To capture the images, we dropped the strands from random positions above the cameras, ensuring only that each strand fell within the intersection of the fields of view of the two cameras. To increase acquisition variability, for each strand the acquisition process was repeated 24 times. To ensure uniformity in the orientation at the moment of the acquisition, the strand was dropped 8 times starting from a frontal position, 8 times from a sideways position, and 8 times from an intermediate position.

In total, Granulo-10k includes $200 \times 24 \times 2 = 9,600$ RGB images with size 1280×960 pixels. For each paired acquisition, we include the corresponding segmentation masks and the three-dimensional point cloud (Fig. 2), computed using the algorithms described in [10]. Although the dataset contains 200 unique strands, it captures substantial variability through repeated acquisitions and multi-modal observations, supporting robust evaluation of geometric estimation methods.

[1] https://github.com/AngeloUNIMI/Granulo-10k

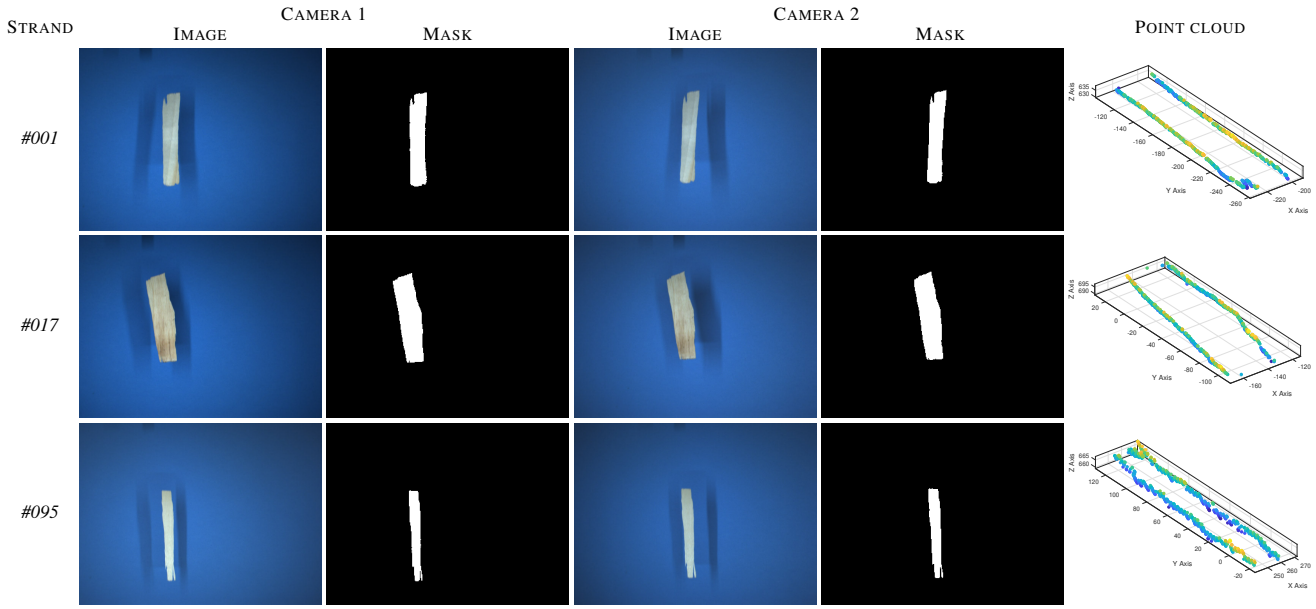

Fig. 2. Granulo-10k: examples of three strands *(#001, #017, #095)* captured using the multiple-view acquisition system. Each paired acquisition is associated with the corresponding segmentation masks and three-dimensional point cloud.

3. METHOD

Network architecture. The proposed architecture integrates multi-view visual information and 3D geometry into a unified multi-task regression framework for estimating height (h), width (w), and thickness (t). Our model is depicted in Fig. 3. Two image views are processed by separate encoders to produce visual embeddings, while the associated point cloud is encoded using a PointNet++ network [14], followed by a lightweight trainable MLP adapter that projects the point-cloud features to the same dimensionality as the image embeddings. The three resulting feature vectors are then stacked and aggregated via max pooling. We adopt a simple fusion strategy to isolate modality contributions without introducing additional architectural complexity. From this max-pooled shared multi-modal representation, a Multi-gate Mixture-of-Experts (MMoE) decoder is employed to model the three regression tasks jointly. The decoder first processes the shared input through a set of E shared expert networks, each implemented as a two-layer multilayer perceptron with 256 hidden units, ReLU activations, Layer Normalization, and dropout. These experts learn complementary representations that are shared across all tasks.

For each task, a task-specific gating module is defined as a linear layer mapping the shared max-pooled input to E gating coefficients, followed by a softmax activation. The resulting gate weights determine a task-dependent convex combination of the expert outputs, yielding a task-specific mixed expert representation. Each mixed representation is then passed to a task-specific module consisting of a linear layer with 64 hidden units and a ReLU activation, producing a compact task embedding of dimension 64. Finally, each task embedding is mapped to a scalar regression output through an independent linear prediction head. Thus, the decoder enables shared learning and task-specific specialization.

Loss function. Since the three predicted quantities are derived from a shared latent encoding, each of them may exhibit a different level of intrinsic uncertainty. A naïve approach would consist in assigning fixed weights to balance the corresponding regression losses; however, this strategy requires manual tuning of additional hyperparameters and makes the optimization procedure sensitive to their choice. Moreover, different noise scales make uniform weighting suboptimal. To address this issue, we adopt a multi-task weighting strategy based on task-dependent homoscedastic uncertainty [15], which allows the relative importance of each task to be learned directly from the data.

For each sample x_n, we consider three scalar regression targets $y_{n,1}, y_{n,2}, y_{n,3} \in \mathbb{R}^+$, corresponding to height, width, and thickness, respectively. Each regression target is modeled with homoscedastic (i.e., task-dependent and input-independent) Gaussian observation noise with standard deviations $\sigma_1 > 0$, $\sigma_2 > 0$, and $\sigma_3 > 0$, collected in $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$. The observation model associated with each task is defined as

$$y_{n,k} \sim \mathcal{N}\big(f_{\theta,k}(x_n),\, \sigma_k^2\big)\,, \quad k \in \{1,2,3\}, \tag{1}$$

where $f_{\theta,k}(x_n)$ denotes the model prediction for task k on sample n.

Defining the vector of task targets for sample n as $\mathbf{y}_n = (y_{n,1}, y_{n,2}, y_{n,3})$, and assuming conditional independence across tasks, the joint likelihood factorizes as

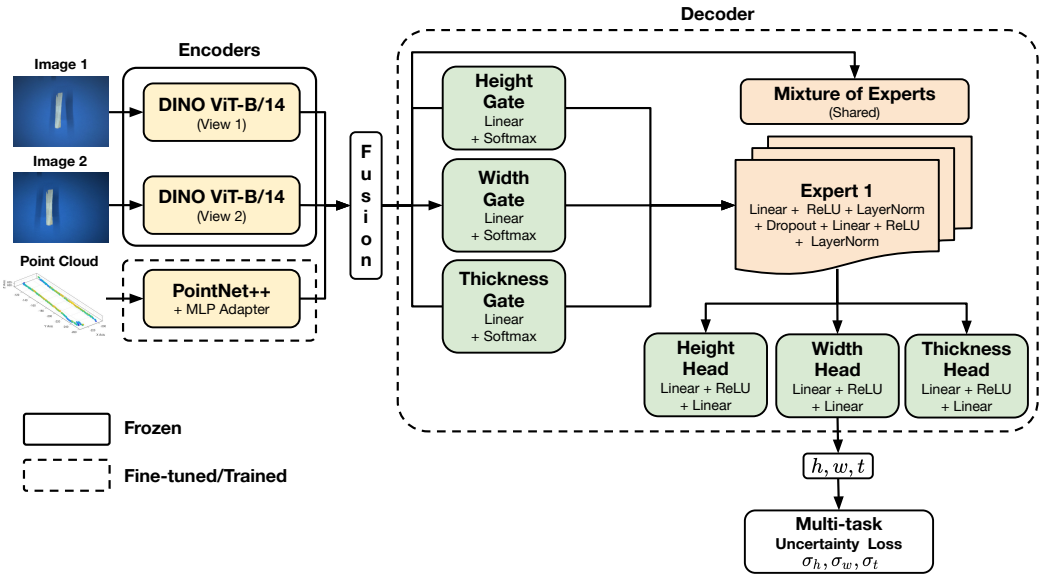

Fig. 3. Overview of the proposed multi-modal architecture, which integrates three encoders (image encoders are frozen during training) with a multi-gate mixture-of-experts (MMoE) decoder to jointly estimate strand height (h), width (w), and thickness (t).

$$p(\mathbf{y}_n \mid x_n, \theta, \boldsymbol{\sigma}) = \prod_{k=1}^{3} p(y_{n,k} \mid x_n, \theta, \sigma_k). \tag{2}$$

The training objective is defined as the negative log-likelihood of the joint model:

$$\mathcal{L}_n(\theta, \boldsymbol{\sigma}) := -\log p(\mathbf{y}_n \mid x_n, \theta, \boldsymbol{\sigma}). \tag{3}$$

By discarding additive constants that do not affect optimization, the resulting loss per sample can be written as

$$\mathcal{L}_n(\theta, \boldsymbol{\sigma}) = \sum_{k=1}^{3} \left(\frac{1}{2\sigma_k^2} \left(y_{n,k} - f_{\theta,k}(x_n) \right)^2 + \log \sigma_k \right), \tag{4}$$

where the first term corresponds to the regression error of task k, while the second term acts as a regularizer preventing degenerate solutions in which the estimated task uncertainty grows arbitrarily large[2]. Under this formulation, the task uncertainties are learned jointly with the network parameters and implicitly balance the contribution of each task during training.

To quantitatively evaluate the regression performance, we compute both absolute and relative error measures for each predicted quantity. Specifically, we report the Mean Absolute Error (MAE) and, to account for scale differences, the Mean Absolute Percentage Error (MAPE), defined as follows:

$$\begin{aligned} \mathrm{MAE}_k &= \frac{1}{N} \sum_{n=1}^{N} \left| f_{\theta,k}(x_n) - y_{n,k} \right|, \\ \mathrm{MAPE}_k &= 100 \times \frac{1}{N} \sum_{n=1}^{N} \frac{\left| f_{\theta,k}(x_n) - y_{n,k} \right|}{y_{n,k}}, \end{aligned} \tag{5}$$

where N denotes the total number of samples. Both metrics are computed independently for each regressed quantity.

4. RESULTS

For all experiments, we use a learning rate of 1.2×10^{-3} and perform 5-fold cross-validation, training each fold for 50 epochs. The image backbones are kept frozen during training. Point-cloud inputs are processed by a PointNet++ encoder [14], initialized from a ModelNet40-pretrained checkpoint [16] and fine-tuned, followed by an MLP adapter trained from scratch. Additionally, we adopt strand-disjoint splits, ensuring that all views of a given strand belong exclusively to a single split, preventing any training-test leakage.

Table 1 reports the performance of selected state-of-the-art image backbone architectures, evaluated both in an image-only configuration and with the inclusion of point cloud (PC) information. The considered backbones include transformer-based and hybrid architectures, DINO ViT-B/14 [17], CLIP ViT-L/14 [18], ConvNeXtV2 [19], and EVA02-CLIP ViT-L/14 [20], which reflect diverse pre-training paradigms and inductive biases.

In the image-only setting, all backbones significantly outperform the mean baseline across all target dimensions.

[2]We refer the reader to [15] for a detailed derivation and discussion of this loss formulation.

Among them, DINO ViT-B/14 achieves the lowest errors, yielding the best MAE and MAPE for height and width, and competitive performance for thickness. ConvNeXtV2 and EVA02-CLIP ViT-L/14 follow closely, with similar accuracy across dimensions, demonstrating that both convolutional and hybrid architectures can effectively capture geometric information from images alone. CLIP ViT-L/14 shows comparatively higher errors, particularly for height and thickness, suggesting that its large-scale multi-modal pre-training does not directly translate to optimal geometric regression performance in the absence of explicit 3D cues.

When point cloud information is included, performance trends become more differentiated across backbones. DINO ViT-B/14 presents the most pronounced and consistent improvements, achieving the lowest MAE and MAPE for both height and width, while also maintaining stable thickness estimation. This suggests that DINO better exploits complementary 3D cues. In contrast, CLIP ViT-L/14 does not benefit from the addition of PC input, showing higher errors compared to DINO and no systematic improvement over its image-only configuration. This suggests that multi-modal fusion may interfere with representations learned through large-scale vision-language pre-training if not explicitly optimized for geometric tasks. For ConvNeXtV2 and EVA02-CLIP ViT-L/14, the inclusion of point cloud data results in moderate and dimension-dependent effects, with small improvements in some cases and marginal degradations in others. This suggests that these architectures already encode substantial geometric structure in their image representations, diminishing returns from explicit 3D input.

Thickness estimation is consistently the most challenging dimension, with higher relative errors and smaller improvements from point cloud integration than height and width. Including point cloud data mainly reduces absolute errors without substantially altering relative error distributions. Furthermore, the effectiveness of additional geometric information is highly backbone-dependent: DINO ViT-B/14 shows the strongest and most consistent gains for multi-modal geometric estimation, underscoring the need to align backbone architecture and pre-training with the demands of multi-modal regression.

Number of experts. In Fig. 4 we depict an ablation study on the number of experts E in the MMoE decoder when using the DINO ViT-B/14 backbone (detailed results are reported in the supplementary material). As E increases, performance progressively improves, reflecting the greater representational capacity of the mixture to model task-dependent feature interactions. This effect is particularly noticeable for width and thickness estimation, which are inherently more challenging to predict due to higher variability and weaker visual cues compared to height, and therefore benefit more from increased expert specialization.

For our experiments, the optimal configuration is reached at $E = 64$, which provides the best overall trade-off between accuracy and stability. Beyond this point, further increasing the number of experts does not lead to additional improvements and, in some cases, slightly degrades performance, likely due to over-parameterization and more challenging optimization. These findings suggest that a sufficiently large but controlled number of experts provides the best trade-off between expressiveness and generalization.

Backbone	**PC**	**Height**		**Width**		**Thickness**	
		MAE [mm]	MAPE [%]	MAE [mm]	MAPE [%]	MAE [mm]	MAPE [%]
Mean Value	-	$17.88_{1.45}$	$21.58_{2.86}$	$6.24_{0.50}$	$56.22_{3.56}$	$0.18_{0.03}$	$29.29_{2.32}$
DINO ViT-B/14 [17]	✗	$2.70_{0.34}$	$2.99_{0.34}$	$1.65_{0.21}$	$12.13_{1.67}$	$0.09_{0.01}$	$\mathbf{13.70_{0.67}}$
ConvNeXtV2 [19]	✗	$2.95_{0.31}$	$3.30_{0.37}$	$1.64_{0.13}$	$11.67_{1.16}$	$0.09_{0.01}$	$14.91_{1.52}$
EVA02-CLIP ViT-L/14 [20]	✗	$2.82_{0.34}$	$3.19_{0.35}$	$1.73_{0.10}$	$12.11_{0.70}$	$\mathbf{0.09_{0.00}}$	$14.63_{1.33}$
CLIP ViT-L/14 [18]	✗	$2.99_{0.24}$	$3.34_{0.16}$	$1.84_{0.17}$	$13.38_{1.69}$	$0.11_{0.01}$	$17.77_{1.67}$
DINO ViT-B/14 [17]	✓	$\mathbf{2.53_{0.08}}$	$\mathbf{2.77_{0.09}}$	$\mathbf{1.59_{0.02}}$	$\mathbf{11.94_{0.35}}$	$0.10_{0.00}$	$14.58_{0.57}$
ConvNeXtV2 [19]	✓	$2.93_{0.16}$	$3.24_{0.16}$	$1.79_{0.06}$	$13.94_{0.38}$	$0.10_{0.00}$	$15.13_{0.13}$
EVA02-CLIP ViT-L/14 [20]	✓	$3.01_{0.07}$	$3.31_{0.08}$	$1.78_{0.07}$	$13.47_{0.93}$	$0.11_{0.00}$	$17.44_{0.79}$
CLIP ViT-L/14 [18]	✓	$3.41_{0.14}$	$3.80_{0.12}$	$2.06_{0.17}$	$15.93_{2.02}$	$0.13_{0.01}$	$19.96_{1.19}$

Table 1. Error metrics (MAE and MAPE) for different image backbone networks evaluated without (✗) and with (✓) the point cloud (PC) input. Bold values correspond to the minimum value of $\mu + \sigma$, where μ and σ denote the mean and standard deviation, respectively.

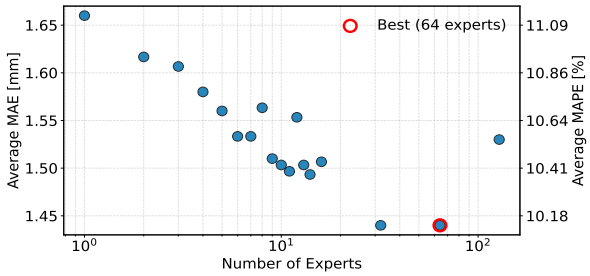

Fig. 4. Ablation study on the number of experts (E) for the DINO ViT-B/14 backbone.

5. CONCLUSIONS

In this work, we introduce a new dataset (Granulo-10k), designed to support the study of fine-grained geometric estimation in industrial scenarios. By providing annotated measurements for height, width, and thickness, the dataset addresses a critical gap in existing benchmarks, which typically lack multi-modal data suitable for evaluating three-dimensional geometric reasoning at scale. Building on this dataset, we proposed a multi-task learning framework based on a MMoE decoder, capable of jointly estimating multiple geometric properties while dynamically balancing task-specific representations. Our results demonstrate that an appropriate expert configuration reduces negative task interference and improves performance, while point cloud integration benefits certain backbones. Future work will expand dataset variability and improve thickness estimation through better geometric representations and fusion strategies.

6. REFERENCES

[1] G. Standfest, A. Petutschnigg, and M. Dunky, "2D pore size distribution in particleboard and oriented strand board," in *Proc. of the 6th Int. Symposium on Image and Signal Processing and Analysis (ISPA)*, 2009, pp. 359–364.

[2] H. Wan, X.-M. Wang, and K. Groves, "Manufacturing high strength and dimensional stable strand panels via optimizing panel manufacturing conditions," in *Proc. of the 2012 Int. Conf. on Biobase Material Science and Engineering (BMSE)*, 2012, pp. 242–245.

[3] H. Song, Z. Wang, and L. Li, "Experimental investigation on bending creep properties of hybrid cross-laminated timber fabricated from lumber and OSB," in *Proc. of the 2021 4th Int. Symposium on Traffic Transportation and Civil Architecture (ISTTCA)*, 2021, pp. 605–610.

[4] R. Donida Labati, A. Genovese, E. Muñoz, V. Piuri, F. Scotti, and G. Sforza, "Improving OSB wood panel production by vision-based systems for granulometric estimation," in *Proc. of the 2015 IEEE 1st Int. Forum on Research and Technologies for Society and Industry (RTSI)*, 2015, pp. 557–562.

[5] T. Nishimura, J. Amin, and M. P. Ansell, "Image analysis and bending properties of model OSB panels as a function of strand distribution, shape and size," *Wood Science and Technology*, vol. 38, no. 4, pp. 297–309, 2004.

[6] B. Zhuang, A. Cloutier, and A. Koubaa, "Effects of strands geometry on the physical and mechanical properties of oriented strand boards (OSBs) made from black spruce and trembling aspen," *BioResources*, vol. 17, no. 3, pp. 3929–3943, 2022.

[7] D. Krug, M. Direske, S. Tobisch, A. Weber, and C. Wenderdel, *Particle-Based Materials*, pp. 1409–1490, Springer International Publishing, 2023.

[8] W. Hong, Y. Shi, Z. Huo, W. Li, and C. Mei, "Real-time tracking of the characteristics of strands in OSB production lines," *Wood Science and Technology*, vol. 59, no. 1, pp. 1–16, 2024.

[9] M. Verheyen, W. Beckers, E. Claesen, G. Moonen, and E. Demeester, "Vision-based sorting of medium density fibreboard and grade a wood waste," in *Proc. of the 2016 IEEE 21st Int. Conf. on Emerging Technologies and Factory Automation (ETFA)*, 2016, pp. 1–6.

[10] R. Donida Labati, A. Genovese, E. Muñoz, V. Piuri, and F. Scotti, "3-D granulometry using image processing," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1251–1264, 2019.

[11] M. López, J. A. Vilán, C. Casqueiro, and J. M. Matías, "3D laser scanner: A new method for estimating the dimensions of wood pulp chips," *Nordic Pulp and Paper Research Journal*, vol. 21, no. 3, pp. 342–348, 2006.

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2nd edition, 2004.

[13] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[14] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: deep hierarchical feature learning on point sets in a metric space," in *Proc. of the 31st Int. Conf. on Neural Information Processing Systems (NeurIPS)*, 2017, p. 5105–5114.

[15] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7482–7491.

[16] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. of the 2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1912–1920.

[17] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *2021 IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2021, pp. 9630–9640.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. of the 38th Int. Conf. on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[19] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders," in *Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16133–16142.

[20] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, and Y. Cao, "EVA-02: A visual representation for neon genesis," *Image and Vision Computing*, vol. 149, pp. 105171, 2024.