# ConvNeXt and Swin Transformer for Microrobot Pose and Depth Estimation

Andrés Godoy
*Department of Engineering*
*Imperial College London*
daniel.godoy-ortiz@imperial.ac.uk

***Abstract*—This coursework evaluates modern deep learning architectures for two key perception tasks: (i) pose classification into discrete pitch and roll angles, and (ii) continuous depth estimation. Using a fraction of the OTMR dataset, the analysis benchmarks a Simple CNN, ConvNeXt-V2, and Swin Transformer V2, complemented by a hyperparameter-optimised lightweight CNN. Experimental results show that ConvNeXt achieves near-perfect pose accuracy and state-of-the-art depth regression, while the optimised CNN provides competitive performance with a substantially smaller computational footprint.**

***Index Terms*—Microrobotics, Deep Learning, Pose Estimation, Depth Regression, Optical Microscopy, ConvNeXt, Swin Transformer, Lightweight CNNs**

## I. Introduction

Microscale imaging has become an essential component in the development of microrobotic systems for biomedical manipulation, micro-assembly, and lab-on-a-chip technologies. In such environments, optical tweezers and high-magnification microscopy are commonly employed to enable non-contact actuation and monitoring of micro-objects. The images produced at this scale are, however, strongly affected by low contrast, non-uniform illumination, optical diffraction and depth-dependent blur, all of which complicate the extraction of geometric and structural cues necessary for accurate pose and depth estimation.

Acquiring reliable labels for microscale perception is also challenging, as variations in out-of-plane pose or axial depth often produce only subtle visual changes, necessitating precise instrumentation for ground-truth annotation. Consequently, datasets in this area have traditionally been small or restricted to specific robot geometries. The OTMR dataset [1] mitigates these limitations by providing large-scale annotated microscope images covering multiple microrobot designs, a wide range of out-of-plane orientations, and continuous depth trajectories, enabling systematic evaluation of data-driven models under realistic microscopic conditions.

The original OTMR study evaluated a range of established convolutional architectures, including EfficientNet,ViT, ResNet, and MobileNet, for pose classification and depth regression. More recently, vision architectures such as ConvNeXt and Swin Transformers have introduced hierarchical designs that integrate convolutional inductive biases with transformer-inspired components, achieving strong performance on large-scale visual benchmarks. This motivates examining their suitability in the microscopy domain. However, their behaviour in the strongly physics-constrained setting of optical microscopy, where information is encoded through diffraction patterns and defocus rather than macroscopic texture, remains largely unexplored.

This project studies the suitability of these architectures for two tasks central to microrobot perception: (i) classification of out-of-plane pose, expressed through discrete pitch and roll configurations, and (ii) depth estimation, formulated as a continuous regression problem. Using a fraction of the OTMR dataset as a testbed, ConvNeXt and Swin Transformer models are evaluated alongside carefully optimised lightweight CNNs, and their learned representations are analysed in the presence of low texture, structural symmetry, and depth-dependent distortions. The aim is to provide a principled assessment of the trade-offs between accuracy, robustness, and computational cost in this microscopic domain.

## II. Methodology

### A. Exploratory Data Analysis

*1) Label Distributions:* Fig. 1 presents the distributions of pitch, roll, and depth, together with a 2D density map of available pose combinations. Although the marginal distributions of pitch and roll are nearly uniform, the lower heatmap reveals that a substantial fraction of the theoretically possible pitch–roll configurations are entirely absent from the dataset. This sparsity has direct implications for model design. Treating each pitch–roll pair as a single categorical class would restrict the model to learning only those combinations observed during training, thereby preventing meaningful generalisation to unseen orientations and offering no mechanism for extrapolating to novel poses. Such a formulation would effectively hard-code the incompleteness of the dataset into the model itself.

By contrast, modelling pitch and roll as two independent outputs in a dual-head architecture preserves their geometric separability and enables the model to predict valid combinations beyond those explicitly present in the training set. This approach not only accommodates the sparse coverage of the joint pose space but also provides a principled path for generalisation: new pitch–roll pairs may be approximated without retraining, or adapted with minimal fine-tuning.

Depth values, shown in the rightmost histogram, follow a continuous and non-uniform distribution.

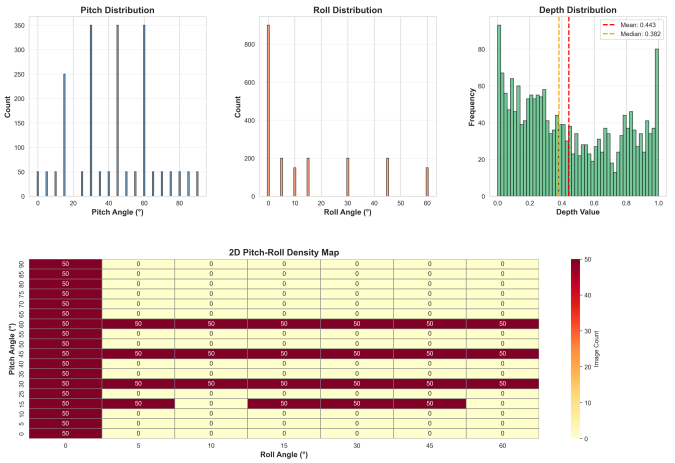


Fig. 1. Summary of label distributions for pitch, roll, and depth.

*2) Low-Dimensional Manifold Structure:* Fig. 2 shows PCA and t-SNE embeddings computed from high-level image features. PCA reveals broad, overlapping clusters, indicating that the principal axes of variance in pixel space do not align linearly with pitch, roll, or depth. Nonetheless, samples from the same pose neighbourhood remain locally coherent even in the principal-component space.

Nonlinear projections provided by t-SNE expose a more structured latent geometry. Both methods separate the dataset into a small number of well-defined clusters within which pitch varies smoothly as a continuous gradient. Roll also displays ordered variation, though its organisation is cluster-dependent, suggesting that roll interacts with broader appearance regimes (e.g., diffraction pattern families). Depth, however, does not shape the global geometry of the embeddings and appears as a locally varying attribute within clusters, reflecting that depth information is encoded at a finer visual level rather than through large-scale geometric changes.

*3) Sharpness–Depth Relationship:* Fig. 3 examines how optical sharpness varies with depth using two complementary measures: variance of the Laplacian and mean Sobel magnitude. Both metrics increase monotonically with axial distance, reflecting the physics of defocus in high-magnification microscopy.

## B. Data Processing

The data-processing was designed to ensure geometric consistency across images, preserve the physical cues relevant for pose and depth estimation, and prepare the dataset for robust model training. Each component of the process was informed by the exploratory analyses and by the optical characteristics of high-magnification microscopy.

*1) Orientation Standardisation:* The first step was to correct orientation inconsistencies arising from EXIF metadata. All images, from non-zero roll configurations, contained embedded rotation tags. Because pose annotations are defined

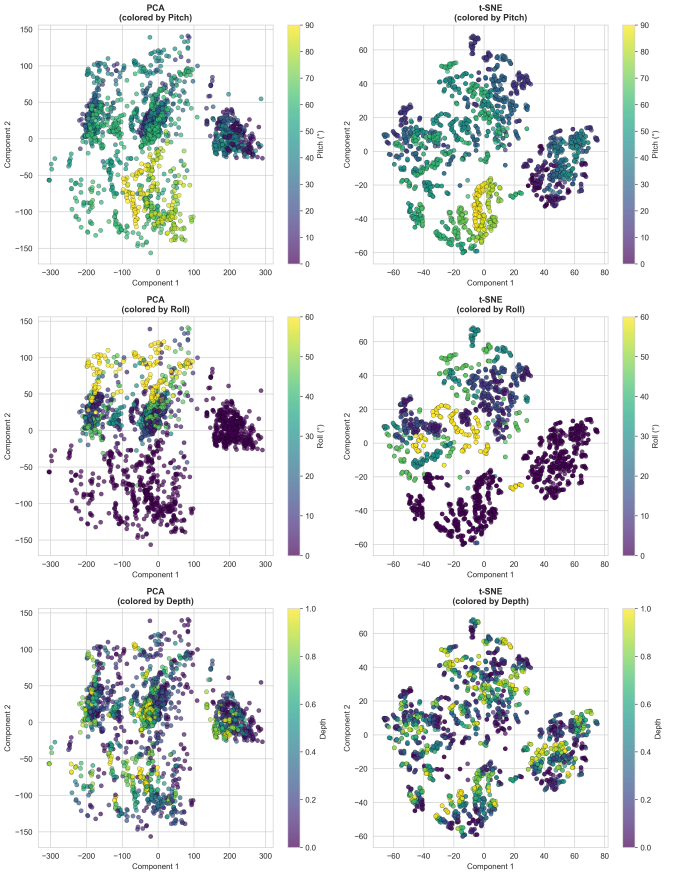


Fig. 2. PCA, t-SNE, and UMAP embeddings of image features, coloured by pitch, roll, and depth.

with respect to the physical orientation of the microrobot, all EXIF metadata were stripped and images were reloaded in a standardised reference frame. This ensured that subsequent transformations, augmentations, and model predictions operated on geometrically aligned data.

*2) Image Normalisation and Preprocessing:* All images were resized to a fixed spatial resolution of $224 \times 224$ (except for the transformer model that requires $256 \times 256$). Pixel intensities were normalised symmetrically to stabilise training across architectures.

*3) Augmentation Strategy:* Augmentation was applied conservatively to mitigate overfitting while preserving the physical interpretability of pose and depth. Since orientation defines the target in the pose classification task, any transformation that would artificially alter the underlying geometry—such as rotations, horizontal flips, or anisotropic scaling—was excluded. Instead, the augmentation strategy focused on appearance-level perturbations that replicate natural variation in microscopic imaging: small translations, controlled cropping that preserves microrobot structure, and mild adjustments to brightness or contrast.

Importantly, no blurring operations were used. As shown in the sharpness–depth analysis, blur is the primary visual indicator of axial displacement; artificially modifying blur

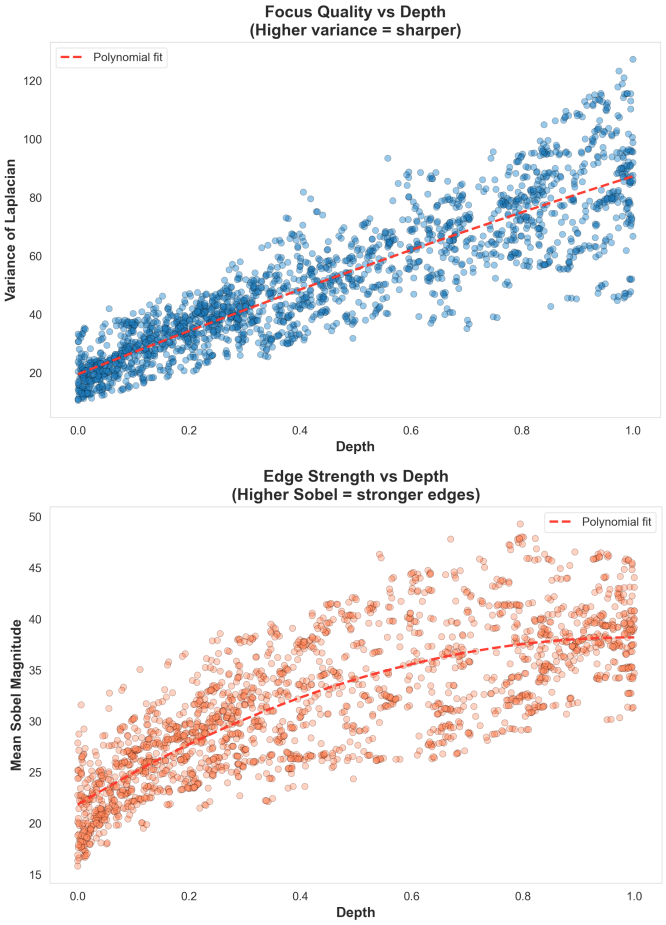


Fig. 3. Sharpness–depth analysis

would distort the mapping between image appearance and true depth. Thus, all augmentations were limited to transformations that preserve the fine-scale structural cues required for both tasks.

*4) Label Handling and Partitioning:* For pose classification, pitch and roll were treated as two independent categorical outputs rather than combined into a single joint label. This design follows from the sparsity of the pitch–roll grid observed during EDA: many theoretically possible combinations do not appear in the dataset. The dual-head formulation preserves geometric continuity and allows the model to represent unseen pitch–roll combinations through interpolation or small-scale fine-tuning. For depth estimation, each image was paired with a continuous depth value corresponding to its axial displacement relative to the focal plane. The dataset was partitioned into training, validation, and test sets following a 60–20–20 split, ensuring balanced pose coverage and preventing depth-sequence leakage across partitions.

## III. Pitch and Roll Classification Modeling

*1) Model Architectures:* To benchmark pose classification performance under microscopic imaging constraints, three representative deep-learning architectures were evaluated: a Simple CNN baseline, a ConvNeXt-V2 model, and a Swin Transformer V2. These models were chosen to reflect increasing levels of architectural sophistication, from traditional convolutional pipelines to modern hierarchical and attention-based frameworks.

**Simple CNN.** The first model is a lightweight convolutional network composed of 5 sequential convolution–ReLU–pooling blocks, followed by fully connected layers for pitch and roll prediction. Although limited in depth and representational capacity, such architectures provide a meaningful baseline by capturing local texture and edge information.

**ConvNeXt V2** [2] represents a modern evolution of convolutional networks, integrating design principles inspired by Transformers (e.g., large kernel convolutions, inverted bottlenecks, and improved normalization) while retaining the inductive biases of CNNs. Its hierarchical multi-stage structure facilitates both local texture modelling and progressively broader spatial receptive fields. Compared to ResNet architectures, ConvNeXt variants typically offer greater representational flexibility with fewer optimization difficulties. Their ability to model mid- and long-range spatial dependencies is particularly advantageous in microscopic images, where pose is encoded not by object silhouette but by distributed diffraction signatures.

**Swin Transformer V2** [3] extends self-attention mechanisms to vision via shifted local attention windows and hierarchical feature maps. This design preserves computational efficiency while enabling the model to capture global context. Standard Vision Transformers (ViT) operate on fixed-size token grids with global self-attention, which can lead to inefficiencies and reduced performance on small scientific datasets. Swin's locality and hierarchical scaling make it better suited for microscopy, where the relevant features are subtle, spatially structured, and often repeated across the field of view. Its windowed attention can capture interactions between diffraction rings and fine-scale topology that may indicate orientation.

*2) Classification Results:* Tables I and II summarise the classification performance of each model on pitch and roll, respectively. Table III aggregates the average across both heads for a holistic comparison.

TABLE I
Pitch Classification Performance (Test Set)

| Model | Acc | Prec | Rec | F1 |
|---|---|---|---|---|
| 5 block CNN | 0.8925 | 0.8965 | 0.8925 | 0.8895 |
| ConvNeXt | 0.9975 | 0.9977 | 0.9975 | 0.9976 |
| Swin Transformer | 0.9950 | 0.9955 | 0.9950 | 0.9950 |

TABLE II
Roll Classification Performance (Test Set)

| Model | Acc | Prec | Rec | F1 |
|---|---|---|---|---|
| 5 block CNN | 0.9600 | 0.9642 | 0.9600 | 0.9600 |
| ConvNeXt | 0.9975 | 0.9976 | 0.9975 | 0.9975 |
| Swin Transformer | 0.9975 | 0.9976 | 0.9975 | 0.9975 |

TABLE III
OVERALL CLASSIFICATION PERFORMANCE (AVERAGE ACROSS HEADS)

| Model | Acc | Prec | Rec | F1 |
|---|---|---|---|---|
| 5 block CNN | 0.9263 | 0.9304 | 0.9263 | 0.9247 |
| ConvNeXt | 0.9975 | 0.9976 | 0.9975 | 0.9975 |
| Swin Transformer | 0.9963 | 0.9965 | 0.9963 | 0.9962 |

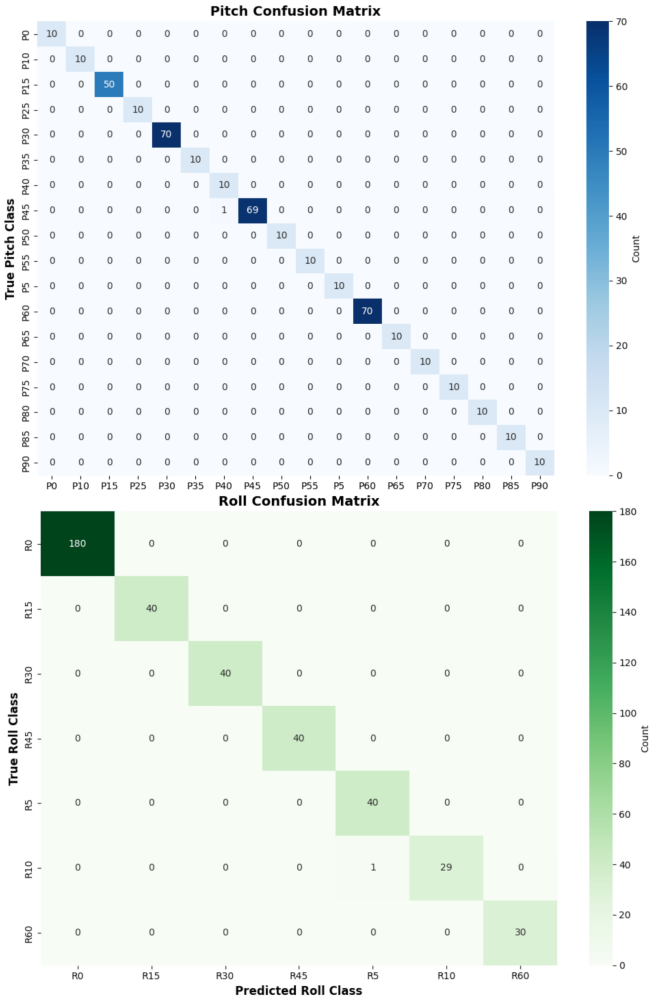


Fig. 4. ConvNeXt V2 confusion matrices for pitch (top) and roll (bottom).

The results reveal a clear hierarchy of performance. Modern hierarchical models (ConvNeXt and Swin Transformer) achieve near-perfect accuracy on both pitch and roll. ConvNeXt achieves marginally higher accuracy than Swin Transformer, consistent with its strong inductive biases for local–global texture modelling in structured scientific images.

**Confusion Matrix Analysis.** Fig. 4 presents the pitch and roll confusion matrices for the ConvNeXt V2 baseline, the strongest-performing model. Predictions lie almost entirely on the diagonal, with misclassifications limited to two neighbouring values.

**Embedding Structure via t-SNE.** To better understand the separability learned by the classifier, t-SNE embeddings of the ConvNeXt V2 latent representations are shown in Fig. 5.

The ordered, low-overlap cluster structure is consistent with the extremely high test accuracy and suggests that the latent space learned by ConvNeXt V2 provides a faithful parameterisation of the underlying orientation manifold.

## IV. DEPTH ESTIMATION (REGRESSION)

*1) Model Architectures and Training Strategy:* Depth estimation was approached using the same architectural families employed in pose classification: a 5-block CNN baseline, ConvNeXt-V2, and Swin Transformer V2. However, depth regression differs fundamentally from classification. Rather than mapping to discrete labels, the model must learn a continuous function relating image appearance to axial displacement—a relationship dominated by smooth changes in defocus, edge attenuation, and diffraction patterns. This distinction has architectural implications.

**ConvNeXt for Depth Regression** is particularly well suited for microscopic depth estimation because its hierarchical convolutional blocks offer strong inductive biases for modelling continuous spatial gradients. Defocus blur manifests as gradually varying local texture changes; large-kernel convolutions and residual connections allow ConvNeXt to capture these smooth transitions more naturally than window-based attention. Unlike Swin, whose local attention windows may fragment subtle blur gradients, ConvNeXt processes the entire receptive field with consistent filters, enabling more stable regression performance.

**Swin Transformer V2.** Although highly effective for structured pose estimation, Swin Transformers face inherent challenges in depth regression. Depth cues are low-frequency, globally coherent signals; partitioning the image into attention windows may disrupt these continuous patterns, requiring deeper layers to reassemble them. As a result, Swin tends to underperform ConvNeXt unless extensive fine-tuning is applied.

**Frozen vs. Unfrozen Backbones.** Two training regimes were explored for the pretrained backbones:

- **Frozen backbone:** Only the regression head is trained, allowing us to evaluate whether the backbone, already trained for pose classification, can be effectively repurposed for depth estimation without further adaptation.
- **Full fine-tuning:** All backbone layers are updated.

### A. Regression Results

ConvNeXt clearly achieves the best performance, reducing RMSE by half relative to the CNN baseline and outperforming Swin in both frozen and fine-tuned settings. Swin improves substantially with fine-tuning but still trails ConvNeXt, highlighting the advantages of convolutional inductive biases for smooth depth regression. On the other hand, the negligible performance difference between the frozen and unfrozen ConvNeXt backbones indicates that the features learned during pose classification already encode the depth-relevant variations in defocus and diffraction.

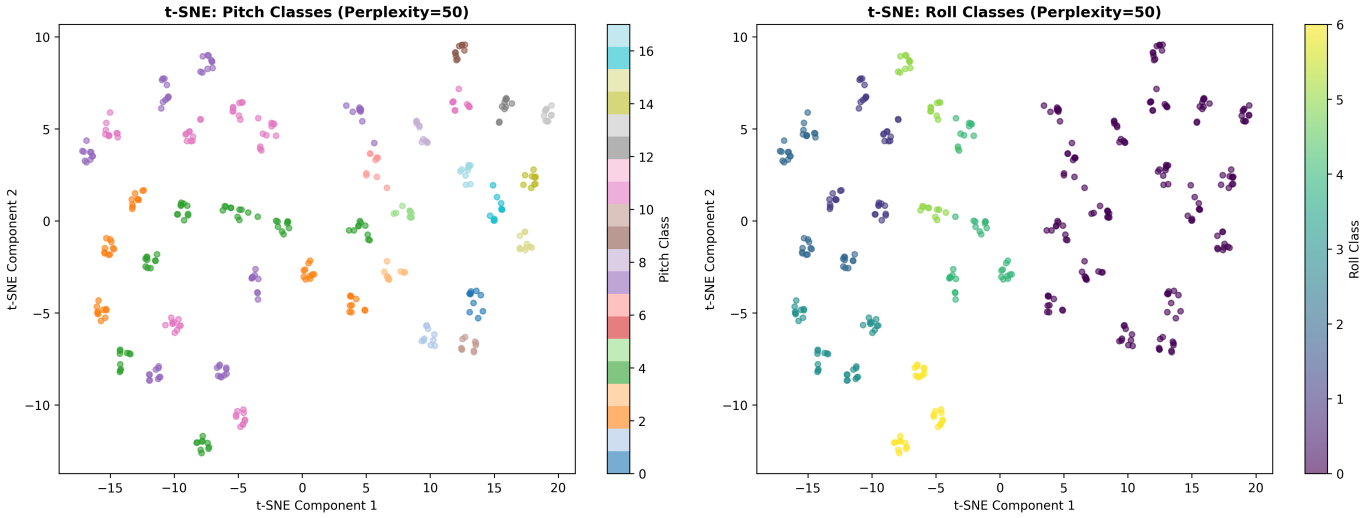


Fig. 5. t-SNE embeddings of ConvNeXt V2 latent features for pitch (left) and roll (right).

TABLE IV
DEPTH REGRESSION RESULTS (TEST SET)

| Model | RMSE | $R^2$ |
|---|---|---|
| 5-Block CNN Baseline | 0.08367 | 0.93012 |
| ConvNeXt (Frozen Backbone) | 0.04199 | 0.98240 |
| ConvNeXt (Fine-tuned, Unfrozen) | **0.04173** | **0.98262** |
| Swin Transformer (Frozen) | 0.07717 | 0.94056 |
| Swin Transformer (Fine-tuned) | 0.05039 | 0.97465 |

### B. Feature-Space Structure via 3D t-SNE

Fig. 6 visualises the learned representation from the ConvNeXt depth regressor using a 3D t-SNE embedding. Samples arrange themselves along a smooth manifold that closely follows the depth continuum. Neighbouring depth values map to neighbouring points, forming an ordered trajectory rather than discrete clusters.

### C. Error Distribution Analysis

Fig. 7 shows the absolute error distribution for the best-performing model. Most predictions exhibit extremely low error: the median absolute error is approximately 0.033, and the majority of values lie below 0.05. This confirms that the model captures the fine-scale optical cues associated with depth. A small set of higher-error cases (0.10–0.20) arises primarily where defocus cues are ambiguous and diffraction rings are weak.

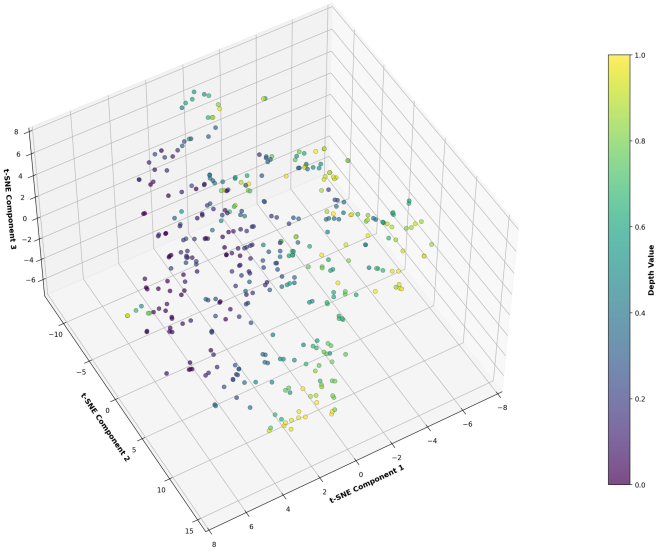


Fig. 6. 3D t-SNE embedding of ConvNeXt-V2 feature space, coloured by depth.

## V. CNN ARCHITECTURE OPTIMIZATION AND COMPUTATIONAL COST

Beyond benchmarking large backbones, the study pursued the design of a lightweight CNN capable of approaching ConvNeXt performance while remaining substantially cheaper to train and deploy. Hyperparameter optimisation of the baseline CNN was conducted using Optuna, jointly tuning the number of convolutional blocks, per-stage channel widths, fully connected dimensionality, learning rate, dropout, and the use of batch normalisation. The optimisation objective was the average validation accuracy across pitch and roll. For the classification task, the best-performing configuration converged to a five-block architecture with channel widths $[48, 32, 128, 128, 64]$, a 384-dimensional fully connected layer, and batch normalisation. This architecture achieved higher accuracy while reducing the overall parameter count relative to the original baseline. For depth estimation, the same optimised backbone was reused: the classification heads were replaced by a single regression head, and the model was retrained for 20 epochs.

The optimised CNN substantially narrows the gap to ConvNeXt in both tasks. For classification, it increases average accuracy from 0.93 to 0.98 while remaining within 1.4 percentage points of ConvNeXt. For depth regression, RMSE is

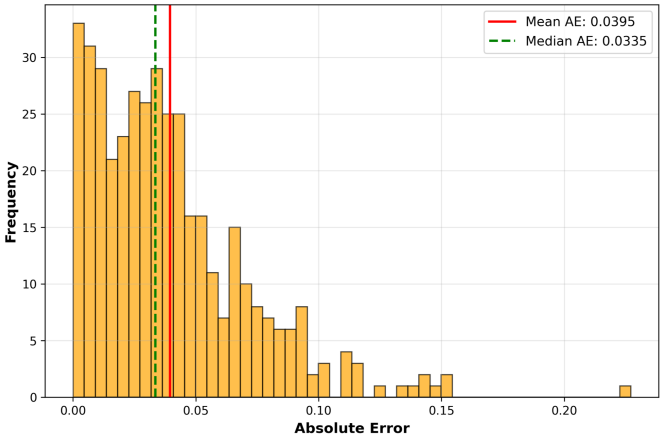


Fig. 7. Absolute error distribution for ConvNeXt depth regression.

TABLE V
CLASSIFICATION: AVERAGE ACCURACY COMPARISON

| Model | Avg Accuracy |
|---|---|
| Vanilla CNN Baseline | 0.9263 |
| Vanilla CNN Optimized | 0.9838 |
| ConvNeXt V2 Baseline | 0.9975 |

TABLE VI
DEPTH REGRESSION: KEY METRICS (TEST SET)

| Model | RMSE | $R^2$ |
|---|---|---|
| Vanilla CNN Depth Baseline | 0.0837 | 0.9301 |
| Vanilla CNN Optimized Depth | 0.0605 | 0.9635 |
| ConvNeXt V2 Depth (Fine-tuned) | **0.0417** | **0.9826** |

reduced by roughly 28% relative to the baseline (from 0.0837 to 0.0605), capturing most of the performance gains offered by the much larger ConvNeXt model.

From a computational perspective, the differences are more pronounced. As summarised in Fig. 8, the optimised CNN uses roughly an order of magnitude fewer parameters than ConvNeXt (about 0.3M vs. 28M) and offers a 7× reduction in inference time per image, while delivering accuracy and depth estimation quality that are close to the state-of-the-art backbone.

## VI. CONCLUSIONS AND FUTURE WORK

This study evaluated a range of deep learning architectures for pose classification and depth estimation of microrobots imaged under high-magnification optical microscopy. Through systematic analysis of the OTMR dataset, including its label structure, optical characteristics, and latent geometric organisation, the results demonstrate that modern hierarchical convolutional networks such as ConvNeXt achieve near-perfect classification accuracy and state-of-the-art depth regression performance. Swin Transformers perform competitively in classification but exhibit limitations in modelling the smooth, spatially coherent cues required for depth estimation. Furthermore, the development of an optimised lightweight CNN showed that much of the performance of larger backbones

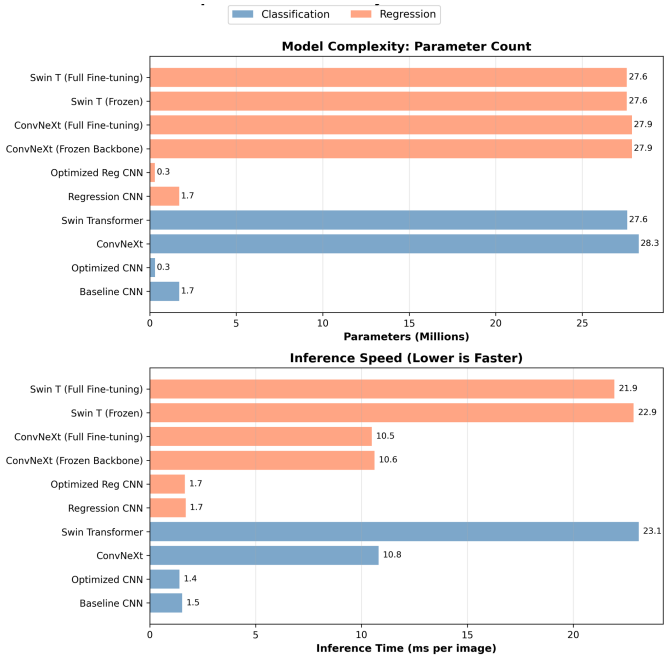


Fig. 8. Model complexity (top) and inference time (bottom).

can be retained at a fraction of the computational cost, making it a compelling alternative for embedded or closed-loop microrobotic applications.

Future work may focus on exploiting the shared structure between pose and depth through multi-task learning or self-supervised pretraining. Since both tasks arise from the same underlying optical phenomena, jointly learning representations or leveraging unlabelled microscopy data could reduce the dependence on large annotated datasets and further improve generalisation. Such approaches may enable feature extractors that are more robust to variations in imaging conditions and that transfer effectively across microrobot geometries and experimental setups.

## REFERENCES

[1] L. Wei and D. Zhang, "A dataset and benchmarks for deep learning-based optical microrobot pose and depth perception," 2025. [Online]. Available: https://arxiv.org/abs/2505.18303

[2] 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," 2023. [Online]. Available: https://arxiv.org/abs/2301.00808

[3] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: Scaling up capacity and resolution," 2022. [Online]. Available: https://arxiv.org/abs/2111.09883