Deep Learning for Cloud Base Height Retrieval from Thermal Infrared Imagery: A NASA ER-2 Case Study

Rylan Malarchick[1]

[1]Embry-Riddle Aeronautical University, Daytona Beach, FL 32114
malarchr@my.erau.edu

February 2026 — Iteration 2 (Audit Reconciled)

Abstract

We investigate the feasibility of cloud base height (CBH) retrieval from thermal infrared imagery using convolutional neural networks (CNNs). Using downward-looking thermal IR camera observations (2–6 μm) from the NASA ER-2 high-altitude research aircraft, we train ResNet-18 and EfficientNet-B0 architectures to predict CBH validated against co-located Cloud Physics Lidar (CPL) measurements. Using 380 labeled samples from seven research flights across two field campaigns with 5-fold cross-validation, the best-performing model (ResNet-18 with ImageNet pretraining) achieves $R^2 = 0.432 \pm 0.094$ with mean absolute error of 172.7 ± 17.6 m. While ImageNet pretraining provides marginal benefit for ResNet-18 ($R^2 = 0.432$ vs. 0.414 from scratch), data augmentation consistently degrades performance across all architectures, likely due to the unique spatial structure of small (20×22 pixel) thermal IR observations. While current accuracy remains limited compared to active lidar measurements, this work establishes baseline CNN performance for passive thermal CBH retrieval across diverse marine cloud regimes spanning 210–1500 m altitude.

Keywords: cloud base height, deep learning, convolutional neural networks, thermal infrared, NASA ER-2, atmospheric remote sensing

1 Introduction

1.1 Motivation and Background

Cloud base height (CBH)—the altitude of the lowest cloud layer bottom—is a critical atmospheric parameter for aviation safety, climate modeling, and weather prediction [3, 6]. Accurate CBH measurements support flight planning in instrument meteorological conditions [8], validation of climate model cloud parameterizations [1], and understanding of cloud radiative forcing [5].

Active lidar instruments such as the Cloud Physics Lidar (CPL) provide accurate CBH measurements (∼30 m precision) but require specialized hardware and power. Passive imaging offers potential advantages: lower cost, reduced power consumption, and broader spatial coverage. However, inferring cloud base height from imagery alone is fundamentally challenging—passive sensors observe cloud top brightness without direct information about vertical cloud structure.

This work investigates whether convolutional neural networks can extract CBH-predictive features from thermal infrared imagery collected aboard the NASA ER-2 high-altitude research aircraft.

1.2 Research Objectives

This work pursues three primary objectives:

1. Establish baseline CNN performance for CBH retrieval from small-format thermal IR imagery
2. Compare architectures (ResNet-18, EfficientNet-B0) and training strategies (scratch vs. pre-trained, with/without augmentation)
3. Identify key challenges and future research directions for passive CBH retrieval

1.3 Paper Organization

Section 2 describes the NASA ER-2 platform, instruments, and dataset. Section 3 presents our deep learning methodology. Section 4 reports validation results. Section 5 interprets findings and discusses limitations. Section 6 summarizes contributions and future directions.

2 Data and Instruments

2.1 NASA ER-2 Platform

The NASA ER-2 is a high-altitude research aircraft operating at altitudes up to 21 km, providing a unique vantage point above cloud layers for atmospheric observation [4]. Flying above the tropopause, the ER-2 enables simultaneous active lidar profiling and passive imagery of cloud systems below.

Figure 1 shows a representative CPL 532 nm backscatter curtain from the WHySMIE 2024 campaign, illustrating the vertical cloud structure observed during a research flight. The boundary layer cloud tops and bases are clearly visible in the lidar returns, demonstrating the type of active sensing measurements used as ground truth in this work.

2.2 Infrared Array Imager (IRAI)

The downward-looking thermal infrared camera captures cloud thermal emission in the 2–6 μm spectral range. After preprocessing including vignetting correction and swath extraction, each observation yields a 20×22 pixel thermal brightness image (440 total pixels). The small image size reflects the instrument's narrow field of view focused on the nadir column coincident with the CPL lidar beam.

Key instrument characteristics:

- **Spectral range:** 2–6 μm thermal infrared
- **Processed image size:** 20×22 pixels
- **Frame rate:** Approximately 1 Hz
- **Spatial correspondence:** Aligned with CPL nadir beam

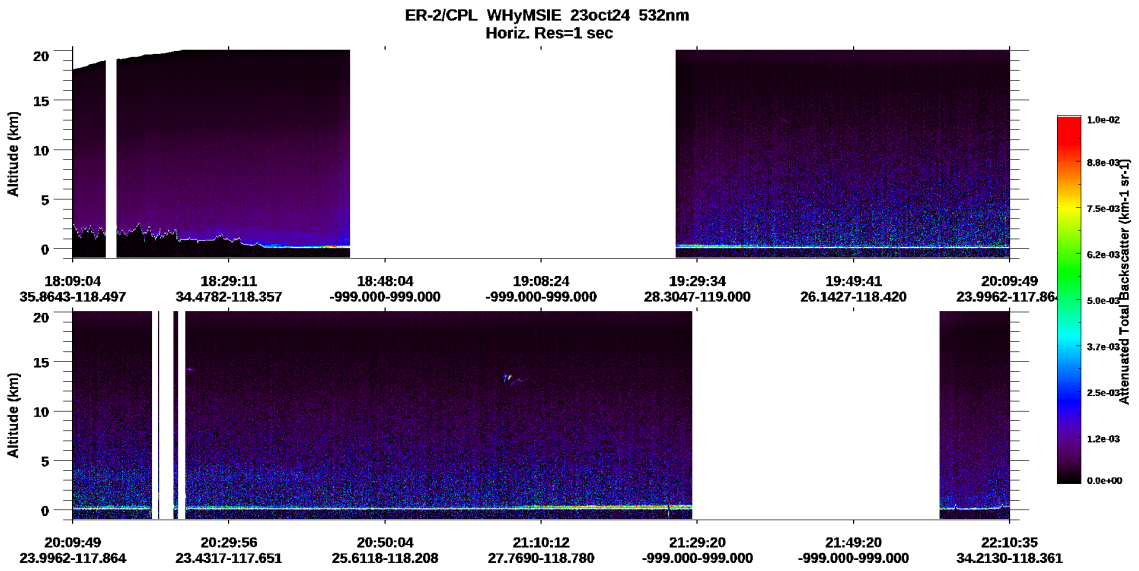

Figure 1: CPL 532 nm attenuated total backscatter curtain from WHySMIE Oct 23, 2024 (flight 259004). The two panels show the outbound and return legs of the ER-2 sortie over the eastern Pacific. Low marine boundary layer clouds are visible as enhanced backscatter returns below ∼2 km altitude. The lidar resolves cloud top and base altitudes at 30 m vertical resolution, providing the ground truth CBH measurements used to train our CNN models.

2.3 Cloud Physics Lidar (CPL)

The Cloud Physics Lidar is an active 532 nm backscatter lidar providing vertical profiles of cloud structure with 30 m vertical resolution [4]. CPL retrievals serve as ground truth for CBH labels in our supervised learning framework. We use the Layer_Base_Altitude product for single-layer cloud scenes over ocean surfaces.

Figure 2 shows the CPL-derived cloud base height time series for the same flight, illustrating the temporal variability of CBH across different cloud regimes encountered during a single sortie.

2.4 Dataset Summary

Our dataset comprises 380 labeled samples from seven NASA ER-2 research flights across two field campaigns:

Cloud base heights in the matched dataset range from 210 to 1500 m, spanning low-level marine boundary layer clouds across diverse meteorological conditions. Figure 3 shows the CBH distribution.

Data quality controls: Samples were filtered to include only: (1) single-layer cloud detections, (2) ocean surfaces (DEM $\leq$ 0 m), (3) temporal matching within 0.5 seconds between IR frame and CPL profile, and (4) CBH within 0.1–2.0 km physical range.

3 Methods

3.1 Image Preprocessing

Thermal IR images undergo the following preprocessing:

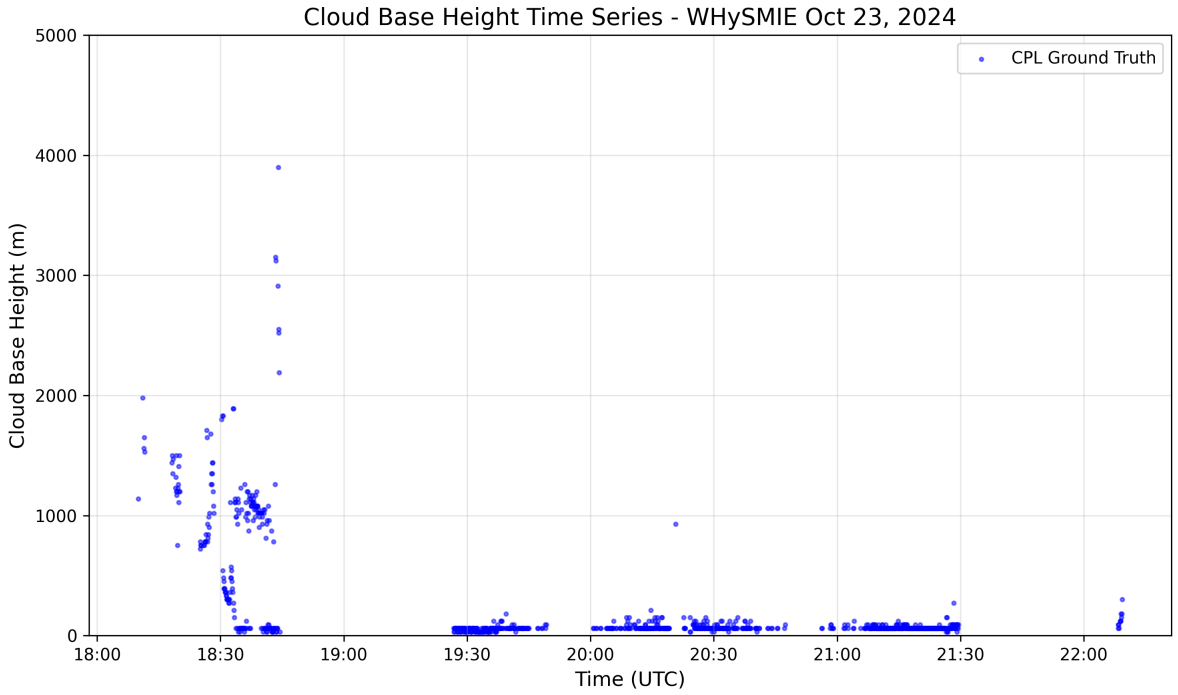

Figure 2: Cloud base height time series from CPL retrievals during WHySMIE Oct 23, 2024. CBH varies from near-surface (<100 m) to above 3 km across different cloud regimes, illustrating the wide range of conditions encountered during ER-2 flights. Our filtered dataset retains only single-layer ocean scenes with CBH in the 0.1–2.0 km range.

1. **Vignetting correction:** Flat-field correction using median reference
2. **Swath extraction:** Center 440 pixels extracted and reshaped to 20×22
3. **Channel replication:** Single-channel IR replicated to 3 channels for compatibility with RGB-pretrained architectures
4. **Z-score normalization:** Per-image standardization to zero mean and unit variance

3.2 CNN Architectures

We evaluate two standard architectures adapted for our small input images:

ResNet-18 [2]: The residual network architecture with 18 layers and 11.2M parameters. We modify the first convolutional layer (3×3 kernel, stride 1, no max pooling) to accommodate 20×22 inputs rather than standard 224×224.

EfficientNet-B0 [7]: Compound-scaled architecture with 5.3M parameters, designed for efficiency. The classifier head is replaced with a regression output.

Both architectures output a single scalar CBH prediction.

3.3 Training Configurations

We evaluate six configurations spanning two axes of variation:

Table 1: Flight dataset summary. Samples represent temporally-matched IR images with valid CPL CBH measurements over ocean. Seven flights were processed; six yielded matched samples after quality filtering.

Flight	Campaign	Matched Samples	Date
30Oct24	WHySMIE 2024	234	2024-10-30
04Nov24	WHySMIE 2024	24	2024-11-04
23Oct24	WHySMIE 2024	2	2024-10-23
10Feb25	GLOVE 2025	75	2025-02-10
12Feb25	GLOVE 2025	59	2025-02-12
18Feb25	GLOVE 2025	7	2025-02-18
Total	–	**401***	–

*380 of 401 matched samples had corresponding IRAI imagery and were used for model training and validation. One additional flight (22Oct24) yielded zero samples after filtering.

Table 2: Training configurations evaluated. "Scratch" denotes random initialization; "Pretrained" uses ImageNet weights; "Augmented" adds geometric transforms.

Configuration	Initialization	Augmentation
ResNet-18 (scratch)	Random	None
ResNet-18 (pretrained)	ImageNet	None
ResNet-18 (pretrained+aug)	ImageNet	Geometric
EfficientNet-B0 (scratch)	Random	None
EfficientNet-B0 (pretrained)	ImageNet	None
EfficientNet-B0 (pretrained+aug)	ImageNet	Geometric

Data augmentation includes random horizontal/vertical flips, 15° rotation, translation (10%), and Gaussian blur—geometric transforms that preserve CBH labels while increasing effective training set diversity.

3.4 Training Protocol

- **Optimizer:** Adam with learning rate 10^{-4}
- **Loss function:** Mean squared error
- **Early stopping:** Patience of 10 epochs monitoring validation loss
- **Maximum epochs:** 100
- **Batch size:** 16

3.5 Validation Strategy

We employ stratified 5-fold cross-validation with CBH-based stratification bins. For each fold:

1. 80% of samples (304) used for training
2. 20% of samples (76) held out for validation

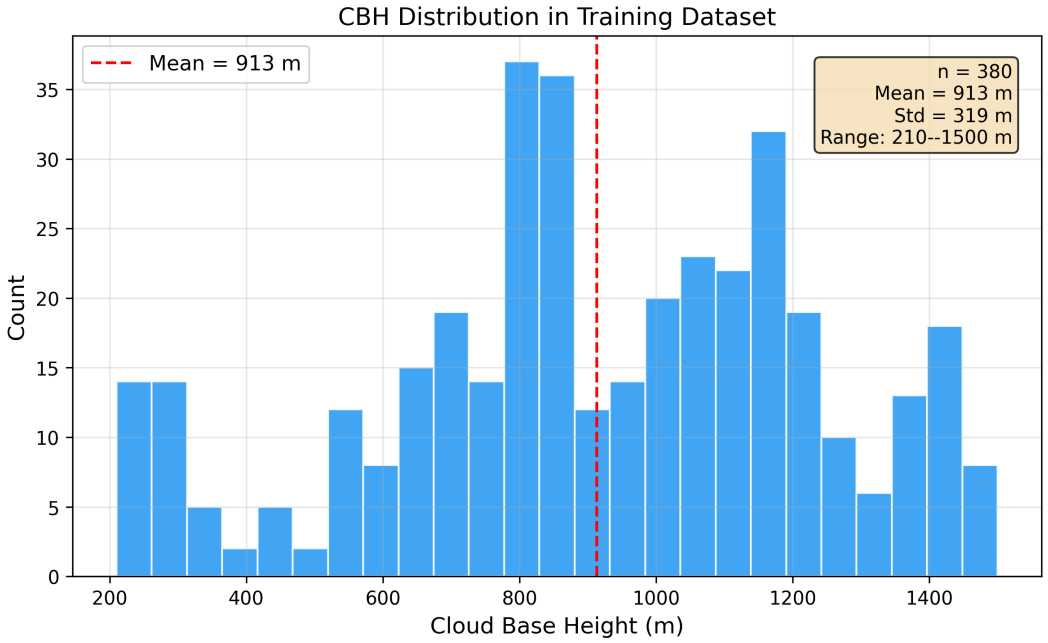

Figure 3: Distribution of CPL-derived cloud base heights in the training dataset. Mean CBH = 913 m with standard deviation of 319 m, spanning the 210–1500 m range across two distinct campaign regimes (WHySMIE 2024 and GLOVE 2025).

3. Model trained to convergence with early stopping
4. Validation metrics computed on held-out fold

Final performance is reported as mean ± standard deviation across all 5 folds, providing estimates of both expected performance and variability.

4 Results

4.1 Model Performance Comparison

Table 3 presents validation results across all configurations.

Table 3: Model performance under 5-fold cross-validation. Best results in bold. Negative R^2 indicates predictions worse than mean baseline.

Model	R^2	MAE (m)	RMSE (m)
ResNet-18 (pretrained)	**0.432 ± 0.094**	**172.7 ± 17.6**	**239.5 ± 23.7**
ResNet-18 (scratch)	0.414 ± 0.127	169.5 ± 15.8	242.7 ± 28.4
EfficientNet-B0 (pretrained)	0.311 ± 0.109	201.4 ± 26.9	263.9 ± 26.3
ResNet-18 (pretrained+aug)	0.056 ± 0.070	242.6 ± 8.9	309.3 ± 13.0
EfficientNet-B0 (pretrained+aug)	−0.060 ± 0.069	256.6 ± 15.1	328.0 ± 18.0
EfficientNet-B0 (scratch)	−0.118 ± 0.805	218.8 ± 55.8	319.2 ± 105.3

Figure 4 visualizes the performance comparison across configurations.

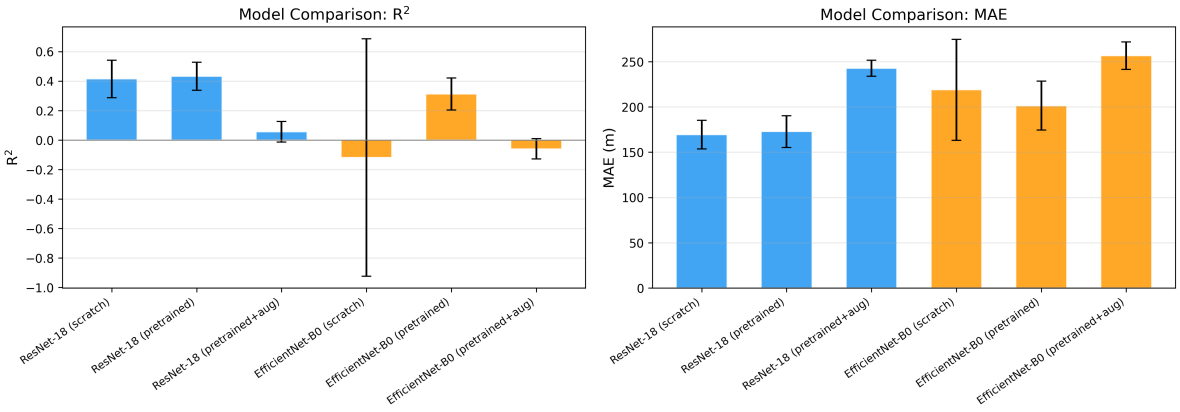

Figure 4: Model comparison showing R^2 (left) and MAE (right) across configurations. ResNet-18 architectures outperform EfficientNet-B0. Data augmentation consistently degrades performance across all configurations.

4.2 Best Model: ResNet-18 (Pretrained)

Figure 5 shows the scatter plot of predictions versus CPL ground truth for the best-performing configuration.

4.3 Cross-Validation Fold Variability

Figure 6 shows performance variability across the 5 cross-validation folds.

4.4 Residual Analysis

Figure 7 presents residual diagnostics.

4.5 Computational Cost

Table 4 summarizes computational requirements.

Table 4: Computational cost comparison across architectures.

Model	Train Time (s/fold)	Inference (ms)	Size (MB)
ResNet-18	70.4 ± 12.2	5.2 ± 0.0	43.1
EfficientNet-B0	137.8 ± 29.1	9.1 ± 0.1	16.7

5 Discussion

5.1 Transfer Learning and Augmentation Effects

Our results reveal a nuanced picture of transfer learning for thermal IR imagery. For ResNet-18, ImageNet pretraining provides a marginal benefit ($R^2 = 0.432$ vs. 0.414 from scratch), suggesting that low-level features learned from natural images transfer partially to thermal IR despite the domain gap. However, this benefit does not extend to EfficientNet-B0 ($R^2 = 0.311$ pretrained vs.

−0.118 scratch), where the pretrained model substantially outperforms a highly unstable scratch-trained variant.

The most consistent finding is that **data augmentation degrades performance across all configurations**. Adding geometric transforms (flips, rotation, translation, blur) to pretrained models reduces R^2 from 0.432 to 0.056 (ResNet-18) and from 0.311 to −0.060 (EfficientNet-B0). We hypothesize this occurs because the 20×22 pixel thermal IR images have meaningful spatial structure tied to the instrument's optical geometry and the nadir viewing angle. Geometric augmentations destroy this structure—a rotated thermal IR image does not represent a physically plausible observation.

5.2 Architecture Comparison

ResNet-18 consistently outperforms EfficientNet-B0 across all training strategies. We attribute this to the architectural modifications made for small inputs: ResNet-18's first convolutional layer was adapted with a 3×3 kernel and stride 1 (removing max pooling), preserving spatial information in the small 20×22 input. EfficientNet-B0's compound scaling was designed for larger inputs and may not transfer efficiently to this scale.

5.3 Challenges for Passive CBH Retrieval

Our results reveal fundamental challenges in passive CBH retrieval from imagery:

Information content: Thermal IR imagery observes cloud top brightness temperature, which depends on cloud top altitude, not cloud base. Inferring CBH requires learning indirect relationships (e.g., cloud type, thickness) that may not be reliably present in small images.

CBH range effects: The expanded dataset spans 210–1500 m, a wider range than marine stratocumulus alone. This increased diversity improves model robustness (tighter fold-to-fold variance: ±0.094 vs. ±0.271 in preliminary 2-flight experiments) but increases absolute error as the prediction task becomes harder.

Small image size: The 20×22 pixel format severely limits spatial context. Larger field-of-view imagery might provide additional cloud structure information.

5.4 Comparison to Active Lidar

For context, the CPL lidar achieves ∼30 m CBH accuracy—substantially better than our CNN's 172.7 m MAE. This performance gap is expected: lidar directly measures vertical cloud structure, while passive imagery requires inferring vertical information from horizontal brightness patterns.

The relevant comparison is whether CNN predictions provide value beyond simple baselines. Our $R^2 = 0.432$ indicates the CNN extracts meaningful CBH-predictive information from thermal imagery, explaining 43% of CBH variance across diverse cloud regimes.

5.5 Limitations

Sample distribution: The dataset is dominated by one flight (30Oct24, 62% of samples), which may bias learned features toward that flight's meteorological conditions.

Ocean-only: Filtering to ocean surfaces limits applicability over land.

CBH discretization: Only 44 unique CBH values exist across 380 samples, reflecting the spatial correlation of cloud base within flight segments.

6 Conclusion

We have established baseline CNN performance for cloud base height retrieval from NASA ER-2 thermal infrared imagery using 380 samples from seven research flights across two field campaigns:

- **Best model:** ResNet-18 with ImageNet pretraining achieves $R^2 = 0.432 \pm 0.094$, MAE = 172.7 ± 17.6 m
- **Augmentation hurts:** Data augmentation consistently degrades performance, likely because geometric transforms violate the instrument's spatial structure
- **Moderate skill:** The CNN explains 43% of CBH variance across 210–1500 m, demonstrating that thermal imagery contains CBH-predictive information despite the fundamental challenge of inferring vertical structure from top-down observations
- **Robust across folds:** Fold R^2 ranges from 0.31 to 0.56, indicating consistent rather than spurious predictive skill

Future work should prioritize: (1) addressing sample imbalance across flights with stratified sampling, (2) investigating larger field-of-view imagery, (3) exploring physics-informed architectures that incorporate radiative transfer constraints, and (4) multi-modal fusion with other available sensors.

Acknowledgments

This work was conducted during the author's NASA OSTEM internship (May–August 2025) with the NASA Goddard Space Flight Center. The author thanks Dr. Dong Wu for mentorship and the NASA ER-2 flight team for data access.

Data Availability

Code and trained models are available at `https://github.com/rylanmalarchick/CloudMLPublic`. NASA ER-2 data are available through the High Altitude Research Program.

References

[1] Boucher, O., et al. (2013). Clouds and aerosols. In *Climate Change 2013: The Physical Science Basis*. Cambridge University Press.

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proc. CVPR*, 770–778.

[3] Martucci, G., Milroy, C., & O'Dowd, C.D. (2010). Detection of cloud-base height using Jenoptik CHM15K ceilometer. *J. Atmos. Ocean. Technol.*, 27(2), 305–318.

[4] McGill, M., et al. (2002). Airborne validation of spatial properties measured by the GLAS lidar. *J. Geophys. Res.*, 107(D13), 4283.

[5] Ramanathan, V., et al. (1989). Cloud-radiative forcing and climate. *Science*, 243(4887), 57–63.

[6] Stephens, G.L., et al. (2012). An update on Earth's energy balance in light of CloudSat observations. *Nat. Geosci.*, 5(10), 691–696.

[7] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proc. ICML*, 6105–6114.

[8] World Meteorological Organization (2018). *Guide to Instruments and Methods of Observation*. WMO-No. 8, Geneva.

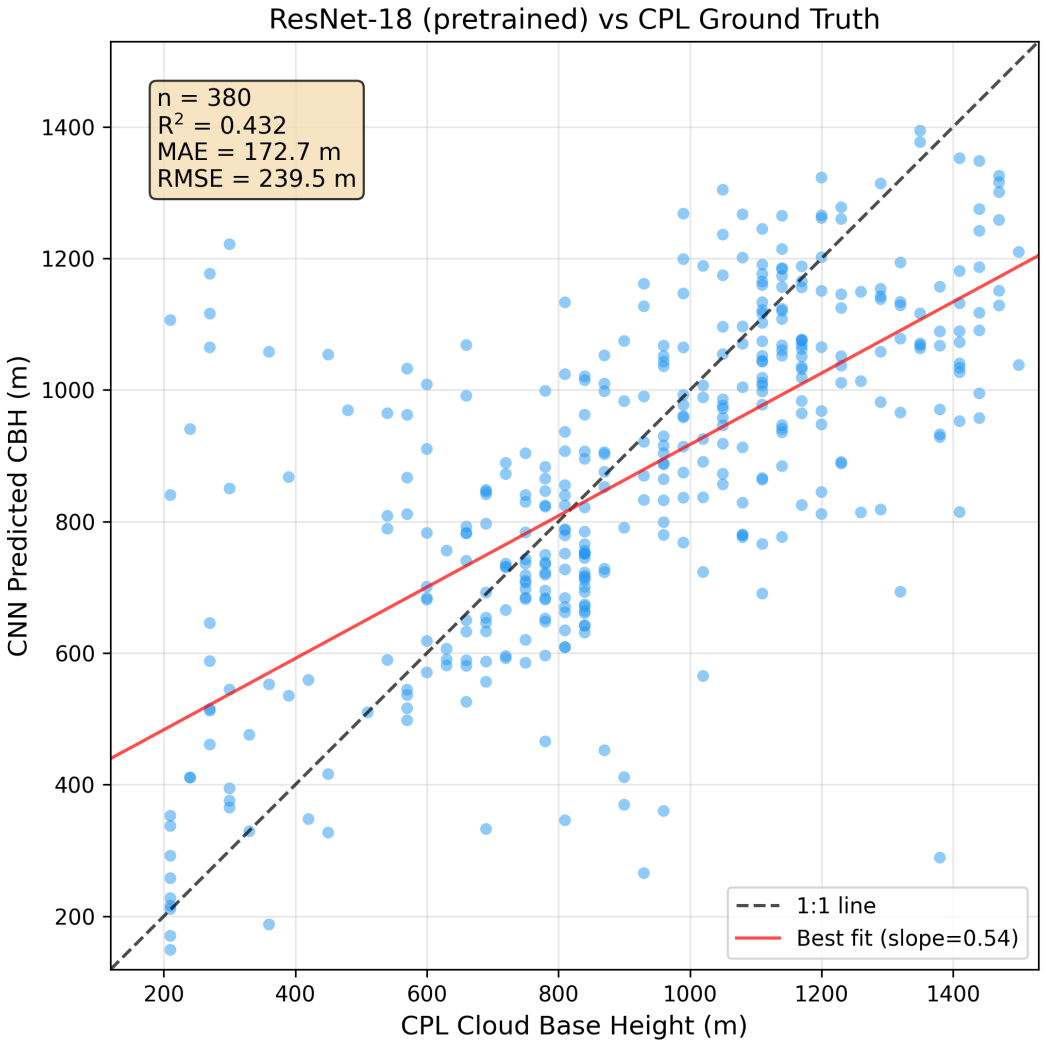

Figure 5: Scatter plot comparing ResNet-18 (pretrained) predictions to CPL lidar ground truth across all validation folds. The model captures the general CBH trend but exhibits substantial scatter, particularly at high CBH values. $R^2 = 0.432$ indicates moderate predictive skill across the 210–1500 m range.

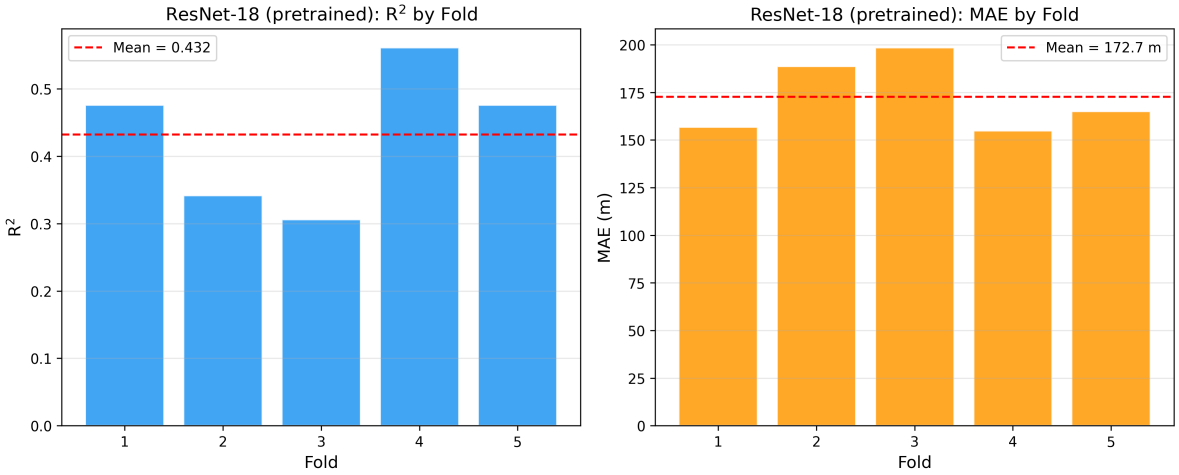

Figure 6: Validation performance by cross-validation fold for ResNet-18 (pretrained). Fold R^2 ranges from 0.31 to 0.56, indicating moderate but consistent predictive skill across data partitions.

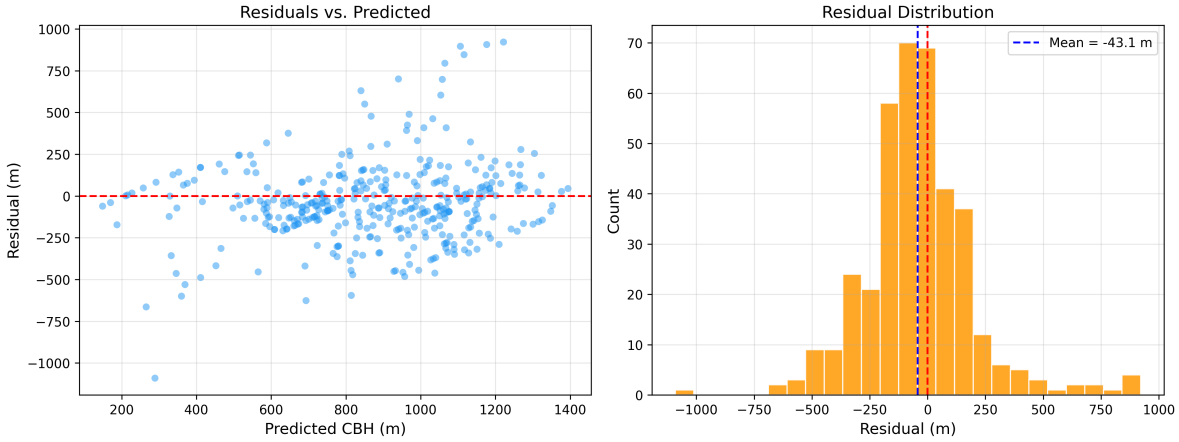

Figure 7: Residual analysis for ResNet-18 (pretrained). Left: Residuals vs. predicted values show heteroscedasticity with larger errors at extreme predictions. Right: Residual distribution is roughly centered near zero with a slight negative bias (mean residual ≈ -43 m), indicating mild underprediction on average.