Privacy-Preserving Computer Vision: Leveraging Depth Images for Image Classification

Dario Furlan
dario.furlan.2@studenti.unipd.it

Dawit Andargachew
dawitandargachew.asmare@studenti.unipd.it

Abstract

As privacy concerns grow, computer vision must move beyond RGB, but can depth-only data compete? While most research focuses on encrypted color images or custom architectures, we ask: Can off-the-shelf models classify objects using just depth?

We evaluate standard CNNs on depth-only inputs from two datasets: depth estimations from ImageNet using the Marigold model, and real-world depth scans from the Washington RGB-D dataset. Fine-tuning these models yielded promising results, with accuracies exceeding 70% (Top-5) on ImageNet subsets and 90% (Top-1) on Washington RGB-D.

These findings challenge the reliance on RGB data and specialized architectures for privacy-preserving vision. Depth data, whether approximated or raw, can enable strong classification performance using widely available models, opening doors to more complex applications like object detection and anomaly detection.

1. Introduction

Today, computer vision is highly advanced and even able to surpass human-level cognitive performance in a wide range of tasks, ranging from image classification to object detection, image segmentation, and more. However, most of these models rely on RGB images, which put privacy at risk in settings such as hospitals, retirement homes, and child or elderly care centers, where the protection of vulnerable individuals is of critical importance.

To address this concern, we explored a promising alternative: depth images; which are captured by 3D cameras, preserving the spatial structure needed for visual recognition without sacrificing privacy.

In this paper, our experiments solely focus on image classification, which we regard as the building block to the broader spectrum of computer vision challenges, thereby laying a foundation for future work and extensions.

2. Related Work

Recent research has increasingly focused on integrating computer vision with privacy-preserving techniques, especially in sensitive environments such as hospitals, nursing homes, and eldercare centers. Depth images have gained traction in the research community as a compelling privacy-aware alternative to conventional RGB images. They inherently lack fine-grained visual details—such as facial features or clothing patterns—while still capturing the spatial and structural cues necessary for recognition tasks. Studies such as [**?**, **?**] have demonstrated the effectiveness of low-resolution or depth-only data across tasks such as action recognition, pose estimation, and fall detection, highlighting its potential in real-world healthcare and surveillance settings.

Encrypted image-based classification techniques, such as block-wise scrambling, permutation encryption, and homomorphic encryption, have been used to protect sensitive data during both training and inference [**?**, **?**]. When combined with modern architectures like Vision Transformers or ConvMixers, these methods have been shown to effectively preserve privacy without compromising accuracy.

Hardware-based approaches have also been proposed to enforce privacy directly at the sensing stage. For example, optical filters like phase masks [**?**] can limit facial detail capture, which allows privacy-preserving depth estimation at the point of image acquisition. Some approaches go further by anonymizing individuals in real time through avatars or abstract representations [**?**], addressing privacy concerns by system-level architecture rather than relying solely on data manipulation or model design.

Regarding the use of depth data for classification, several papers have proposed adaptations of models originally trained on RGB images. For example, lightweight encoders [**?**] can take a single-channel depth input and transform it into a three-channel RGB-like input, thereby making it compatible with pre-trained CNNs such as VGG-16. Other methods augment VGG-16 with a single 3D convolution layer [**?**] to enable depth-only inference without retraining on depth images. Similarly, applying simple color

maps (e.g., Jet) to depth images allows standard models like ResNet-101 and ResNet-18 to achieve high classification accuracy in fruit sorting and animal posture recognition [**?**, **?**].

Depth-based methods have also proven effective in privacy-sensitive healthcare environments. Privacy-aware in-home monitoring systems using depth sensors for tasks like fall detection and gait analysis in eldercare [**?**] offer a less privacy-invasive alternative to RGB-based solutions. Beyond human monitoring, depth data has been applied to classify animal postures [**?**], illustrating the broader versatility of this modality where texture is unnecessary or unavailable.

Collectively, these studies reflect a growing consensus on the technical viability and societal urgency of privacy-preserving computer vision. However, many focus on specialized architectures, encrypted RGB inputs, or hardware solutions. By contrast, relatively few explore the baseline performance of classification models on raw depth images. In this paper, we address that gap by evaluating widely used image classification models (i.e., AlexNet [**?**], VGG19 [**?**], ResNet50 [**?**], and Inception-v3 [**?**]) on depth images, providing a foundation for future privacy-preserving systems.

3. Dataset

Our experiments are conducted using depth images from three sources: the ImageNet, a 200-class subset of it, and the Washington RGB-D Object Dataset. Due to computational constraints, we focus exclusively on the validation portions of each dataset. The main focus is exclusively on depth images, either obtained synthetically via estimation or captured directly using depth sensors. Below, we describe each dataset, its source, and the preprocessing steps applied.

3.1. ImageNet 1k-class depth

ImageNet [**?**] is a large-scale dataset originally developed for visual object recognition, containing over 1.4 million images across 1,000 classes. Due to its size, we limited our experiments to the 50,000-image validation set.

Since ImageNet contains only RGB images, we used Marigold [**?**] (a diffusion-based model for depth estimation) to generate corresponding depth maps, as seen in Figure 1. Marigold produces single-channel, 16-bit PNGs, where each pixel has a value between 0 (near plane) and 65,535 (far plane), with both planes determined by the model during inference. These estimated depth maps retain the original width and height of the RGB images.

We called the resulting dataset *ImageNet 1k-class depth*. These synthesized depth maps enable us to evaluate the feasibility of using depth-only inputs for image classification with standard architectures.

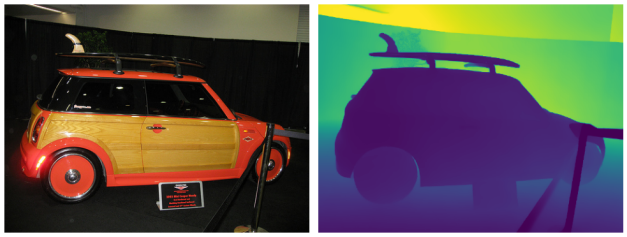

Figure 1. Side-by-side comparison between an original ImageNet RGB image and its estimated depth map using Marigold.

3.2. ImageNet 200-class depth

To further reduce computational overhead while maintaining diversity in object categories, we constructed a 200-class subset of previously mentioned dataset (Imagenet 1k-class depth) by randomly sampling 200 classes from the full 1,000-class set. This subset contains 10,000 images, with 50 images per class. The list of sampled classes can be found in the Supplementary Material. This smaller subset is more balanced and allows us to conduct more detailed evaluations by reducing label space complexity while also preserving the large-scale nature of ImageNet.

3.3. Washington-RGB-D

The Washington RGB-D Object Dataset [**?**] is a real-world RGB-D dataset widely used in robotics and object recognition research. It was captured by a Microsoft Kinect sensor and contains 51 categories of common indoor objects (e.g., scissors, cereal boxes, keyboards) with 300 distinct object instances.

The training set includes 207,920 RGB-D frames, and the validation set (which we use in our study) has 41,877 images. Each instance is recorded from three video sequences, with one frame extracted every fifth frame. For each image, the dataset provides aligned RGB, raw depth (in millimeters), segmentation masks, and cropping metadata.

We specifically use the *depthcrop.png* files, which are single-channel, 16-bit depth images. These real depth images were used as-is, with only minor preprocessing. While the dataset was originally intended for object recognition and pose estimation, we repurpose it for our depth-only classification task.

3.4. Preprocessing

Convolutional Neural Networks (CNNs) are typically designed to operate on 3-channel, 8-bit RGB images of fixed dimensions. However, our experiments focus on depth images, which differ significantly in structure and format. Therefore, several preprocessing steps, such as resizing and normalization to standardize input dimensions and pixel value ranges, are necessary to adapt the datasets to the input

requirements of standard classification architectures.

3.4.1 Color Mapping Depth Maps for CNNs

Depth images are single-channel and thus incompatible with CNNs pretrained on RGB data like AlexNet, VGG19, ResNet50 and Inception-v3. To bridge this gap, we convert them into three-channel inputs using perceptually uniform colormaps and a custom stacked encoding. Unlike the commonly used jet (rainbow) colormap, these mappings preserve perceptual ordering and reduce visual artifacts that may impair representation learning [**?**]. This process is composed of two main steps:

- Normalization: The 16-bit depth values (range: 0-65,535) are scaled down to 8-bit values (range: 0-255) using min-max normalization.
- Channel Expansion: We convert the single-channel 8-bit image into a 3-channel image using two main strategies:
  - Channel Duplication (Stacking): Simply by replicating the depth channel across all three channels.
  - Colormap Mapping: We apply perceptually uniform colormaps (Viridis, Plasma, Magma and Grayscale) to produce pseudo-RGB images, allowing us to investigate whether color encoding affects feature extraction as shown in Figure 2.

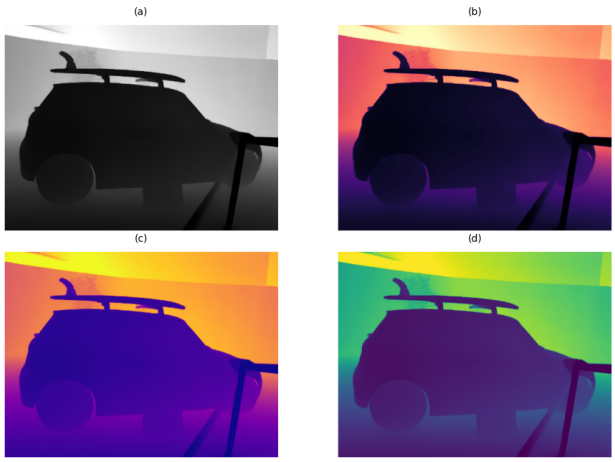

Figure 2. Examples of 8-bit 3-channel depth images obtained through different preprocessing methods: (a) stacked and grayscale, (b) Magma, (c) Plasma, and (d) Viridis colormaps. *Note that grayscale and stacked images are visually similar, as they both represent depth information without color encoding.*

3.4.2 Preprocessing the Washington RGB-D Dataset

The Washington RGB-D dataset already includes 16-bit depth maps captured via Kinect. We applied the same post-processing steps used for ImageNet: normalizing 16-bit depth values to 8-bit, and conversion to 3-channel format via channel duplication and colormap mapping. This ensures consistency in input structure across both synthetic and real-world depth data.

4. Methodology

We investigate whether standard convolutional neural networks (CNNs) can effectively classify objects using depth-only inputs, a question motivated by the need for privacy-preserving computer vision.

4.1. Architectures

To assess performance across different model families and depths, we evaluate four well-known CNN architectures, originally developed for RGB image classification:

- **AlexNet**: a relatively shallow model with fewer parameters, representing the early generation of deep learning-based classifiers [**?**].
- **VGG19**: a deeper model with a straightforward design based on stacked 3×3 convolutions [**?**].
- **ResNet50**: a residual network with 50 layers and identity shortcut connections, allowing for effective training of deeper models [**?**].
- **Inception-v3**: an architecture that uses parallel convolutional paths and factorized filters to balance efficiency and performance [**?**].

These models were chosen to cover a range of architectural complexity and to observe whether trends in RGB image classification carry over to depth-only data.

4.2. Training Regimes

Each model is evaluated under three distinct training configurations:

- **Baseline**: The model is evaluated using its original pre-trained weights, without any adaptation to the depth modality.
- **Partial Fine-tuning**: Only the final classification layers are made trainable, allowing limited adaptation to the depth input.
- **Complete Fine-tuning**: All parameters are made trainable, enabling full adaptation of the network to the new data modality.

4.3. Evaluation Protocol

Not all training regimes are applied to all datasets due to dataset-specific constraints. Baseline evaluation is performed only on the *ImageNet 1k-class depth* dataset, where pre-trained models can be meaningfully assessed without retraining. Both partial and complete fine-tuning are applied where sufficient training data is available (*ImageNet 1k-class and 200-class depth*), while the smaller *Washington RGB-D* dataset is used only with partial fine-tuning to avoid overfitting.

We use classification **accuracy** as our main evaluation metric, and report both **Top-1** accuracy (the correct label is the highest-ranked prediction) and **Top-5** accuracy (the correct label is among the five highest-ranked predictions). For baseline evaluations, the entire dataset is used for testing. In fine-tuning settings, we adopt a standard 80/20 split for training and validation.

5. Experiments

In this section, we present three experiments conducted to evaluate the performance of pretrained models on depth images. First, we establish a baseline by comparing model accuracy across different colormaps applied to the depth images. Second, we explore the impact of fine-tuning on model performance using the stacked colormap representation. Finally, we assess the models on subsets of ImageNet and the Washington RGB-D dataset to understand their behavior on tasks with reduced label spaces and lower complexity.

5.1. ImageNet 1k-class Depth

We initiated our experiments by evaluating the performance of the models "as-is", without any fine-tuning applied to the depth images.

The models employed in this study (AlexNet, VGG19, InceptionV3, ResNet50) were conveniently pretrained on the same 1,000 classes as our dataset. This alignment was achieved by carefully curating the dataset, thereby obviating the need to fine-tune the final layer of the classifier, as it already produced the correct labels.

To establish a baseline and investigate whether the models exhibited varying performance under different colormaps of the depth images, we experimented with a stacked approach, grayscale, viridis, plasma and magma colormaps as shown in Figure 3 and summarized in Table 1.

This experiment demonstrated that the grayscale approach yielded results most familiar to the models. Furthermore, we established that a straightforward approach, such as duplicating the original single-channel depth image to the three channels required by these models, is as robust as applying a colormap function to the grayscale image.

Further experimentation involving fine-tuning these

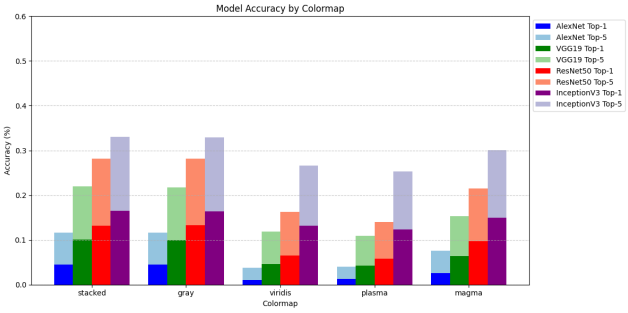

Figure 3. Baseline accuracy comparison across colormaps for ImageNet 1k-class depth.

models resulted in a significant increase in accuracy. Given the strong performance of the "stacked" colormap, we focused on this approach for fine-tuning and excluded other colormaps from further analysis.

The partial fine-tuning procedure involved freezing all model parameters except for a small subset of layers that were specifically selected for training. In the case of **AlexNet**, only the final `classifier` layer was unfrozen and trained. For **VGG19**, we applied the same strategy, unfreezing only the `classifier` module. In the case of **ResNet50**, we unfroze both the last residual block (`layer 4`) and the final fully connected layer (`layer 5`). Finally, for **InceptionV3**, the fine-tuning was performed by unfreezing the last inception block (`Mixed_7c`), the auxiliary classifier (`AuxLogits`), and the final fully connected layer (layer 5).

The complete fine-tuning procedure involved unfreezing all model parameters, allowing the models to adapt fully to the depth images. This approach was computationally more expensive but yielded higher accuracy.

As reported in Table 2, both partial and complete fine-tuning substantially improved model performance on depth images. The stacked colormap representation proved effective, with InceptionV3 achieving the highest Top-1 accuracy (48.53%) and Top-5 accuracy (74.53%). These results suggest that models originally developed for RGB image classification retain their relative performance ranking when applied to depth data. For instance, architectures that achieve higher accuracy on RGB tasks (e.g., InceptionV3 vs. AlexNet) also perform better on depth images, indicating that the improvements are architecture-driven rather than modality-specific.

While both fine-tuning strategies enhanced the models' ability to interpret depth images, the limited number of examples per class in our dataset constrained the overall performance. Nonetheless, the improvements observed with complete fine-tuning are notable: on average, we recorded a nearly **3x** increase in accuracy compared to the baseline (computed as the mean ratio between complete fine-tuning and baseline results for Top-1 and Top-5 across all models).

Model	Stacked		Grayscale		Viridis		Plasma		Magma	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
AlexNet	**04.53**	**11.63**	04.50	11.62	01.04	03.83	01.24	03.98	02.54	07.59
VGG19	**10.04**	**22.00**	09.97	21.79	04.64	11.87	04.29	10.95	06.45	15.31
ResNet50	13.18	**28.16**	**13.27**	28.16	06.56	16.23	05.83	14.00	09.71	21.44
InceptionV3	**16.48**	**33.05**	16.40	32.94	13.16	26.66	12.36	25.34	15.01	30.07

Table 1. Baseline accuracy comparison across colormaps for ImageNet 1k-class depth.

Model	Baseline		Partial Fine-Tuning		Fine-Tuning	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
AlexNet	04.53	11.63	18.05	36.41	**23.62**	**45.76**
VGG19	10.04	22.00	25.40	48.18	**34.20**	**60.96**
ResNet50	13.18	28.16	32.36	57.87	**44.95**	**71.90**
InceptionV3	16.48	33.05	22.31	44.57	**48.53**	**74.53**

Table 2. Fine-tuning accuracy comparison for ImageNet 1k-class depth.

5.2. ImageNet Subset (200 Classes)

As shown in Table 3, complete fine-tuning consistently outperforms partial fine-tuning across all models on the 200-class subset. InceptionV3 achieves the highest Top-1 accuracy (47.95%, 72.80%), while ResNet50 leads in Top-5 accuracy (47.75%, 73.20%). The relative performance ranking of the models aligns with the results from the 1000-class subset, indicating stable model behavior despite the smaller dataset size. Both fine-tuning approaches improve performance; however, deeper networks benefit more substantially from complete fine-tuning, highlighting the importance of end-to-end adaptation when working with limited data.

5.3. Washington RGB-D Dataset

Following the same approach we used again the stacked colormap representation for the Washington RGB-D dataset. In contrast, the Washington RGB-D dataset comprises only 51 classes, with an average of 821 images per class, unlike ImageNet, which has only 50, representing a significantly lower task complexity and substantially larger training data.

As reported in Table 4, all models achieved near 90% Top-1 accuracy, with VGG19 reaching up to 92.23%. These results were obtained using partial fine-tuning alone. Given this high baseline, we opted not to undertake complete fine-tuning, concluding that the incremental improvements would not justify the higher computational expense.

Overall, for sufficiently large datasets, transfer learning tends to be highly effective, eliminating the need for complete fine-tuning.

6. Conclusion and Future Work

This work has demonstrated that depth-only images can achieve competitive performance in image classification tasks using standard convolutional neural networks (CNNs) pretrained on RGB data. Our results indicate that, even without fine-grained texture or color, depth information alone can provide sufficient structure for effective classification across both synthetic and real-world datasets.

- **High accuracy on real-world data:** On the Washington RGB-D dataset, models such as VGG19 surpassed 92% Top-1 accuracy, confirming the viability of depth-only inputs in practical, privacy-sensitive environments.
- **Use of existing architectures:** By applying minimal preprocessing—normalization, channel stacking, and colormap mapping—we successfully adapted standard architectures (e.g., ResNet50, InceptionV3, VGG19) to operate on depth images without requiring modifications to the model structure.
- **Effectiveness of fine-tuning:** Fine-tuning, even when limited to later layers, consistently improved classification performance, especially in complex scenarios such as the full 1,000-class ImageNet subset, where baseline accuracies more than doubled after training.
- **Privacy-aware vision:** Depth images inherently lack identifying features such as facial detail or clothing patterns, making them particularly well-suited for computer vision applications in domains requiring privacy, such as healthcare, education, and public surveillance.

6.1. Future Work

A key limitation of this study is the limited number of depth samples per class, due to the scarcity of large-scale depth-only image classification datasets. To address this, we plan to scale our experiments to the full ImageNet dataset with estimated depth maps, enabling evaluation in

Model	Partial FT		Complete FT	
	Top-1	Top-5	Top-1	Top-5
AlexNet	26.25	**52.90**	**26.40**	51.25
VGG19	32.50	59.95	**36.70**	**65.25**
ResNet50	44.30	71.15	**47.75**	**73.20**
InceptionV3	38.70	64.15	**47.95**	**72.80**

Table 3. Accuracy on 200-class ImageNet subset.

a higher-data regime and testing the scalability of our approach.

Additionally, adapting single-channel depth maps to standard three-channel CNN inputs poses a challenge. To this end, future work will explore learned depth encodings via lightweight autoencoder networks, as demonstrated in [**?**], which proposes a deep learning approach to transform depth images into RGB format.

Moreover, we intend to extend our exploration beyond classification to the task of *semantic segmentation*. While high-quality depth-based datasets for image classification remain scarce, numerous segmentation datasets include depth information. This shift will enable a broader and more realistic evaluation of depth-only models in structured prediction tasks, and may provide further evidence for the suitability of depth data in privacy-sensitive vision applications.

Ultimately, this line of research seeks to develop practical and privacy-conscious computer vision systems that do not compromise on performance, while avoiding reliance on RGB data.

Model	Top-1	Top-5
AlexNet	87.51	98.79
VGG19	92.23	99.61
ResNet50	88.48	99.31
InceptionV3	89.45	99.08

Table 4. Partial fine-tuning accuracy on Washington RGB-D.